Statistical Journal of the IAOS 32 (2016) 413–420
DOI 10.3233/SJI-160995
IOS Press

New proposal for linkage error estimation

Tiziana Tuoto
Italian National Statistical Institute – Istat, Via Cesare Balbo 16, 00184 Rome, Italy
Tel.: +39 06 4673 6372; E-mail: tuoto@istat.it

Abstract. The combined use of data from different sources is an opportunity that the National Statistical Institutes exploit more and more frequently. In a context where huge amount of information, produced by different actors, can be integrated and compared, it becomes even more necessary to provide quality assessments of methods and techniques that have allowed to achieve integration results. When considering data integration at the micro level, record linkage procedures are widely used and generally produce good results (when strong identifying variables are available), although rarely are these procedures provided with associated quality indicators. However, especially in official statistics, quality indicators need to be used in subsequent statistical analyses to guarantee and assess data accuracy and reliability. This paper proposes a method for linkage error estimation. The method enriches the Fellegi and Sunter model for probabistic record linkage: as well known, the Fellegi and Sunter decision rule is very effective for link identification but generally less reliable for result evaluation. The proposal aims at predicting the linkage quality in the Fellegi and Sunter framework, introducing a supervised step.

Keywords: Probabilistic record linkage, linkage errors, linkage quality assessment

1. Introduction

The combined use of data from different sources is an opportunity that the National Statistical Institutes exploit more and more frequently. In a context where huge amount of information, produced by different actors, can be integrated and compared, it becomes ever more urgent the need of providing quality assessments of methods and techniques that have allowed to achieve integration results. When considering data integration at the micro level, record linkage procedures are widely used and generally produce good results (when strong identifying variables are available), although rarely, these procedures are complemented by quality indicators. However, especially in official statistics, quality indicators need to be used in subsequent statistical analyses to guarantee and assess data accuracy and reliability.

Record linkage evaluation is particularly important in order to perform further statistical analyses on linked data avoiding biased and less efficient results related to the integration procedures. In recent years, a lot of work has been done in the field of statistical analyses on linked data, starting from the seminal paper [3]. Linkage quality measurements should be provided together with the linkage results, as well as other linkage information which guarantee the applicability of proper statistical methodologies and ensure high quality results from the integrated data.

From the perspective of increased use of integration procedures, e.g. due to the increased availability and use of administrative data, National Statistical Institutes need record linkage procedures which are efficient enough to run in a production environment with suitably low levels of error.

This paper proposes a method to enrich the Fellegi and Sunter model [5] for probabilistic record linkage in order to achieve good linkage errors estimates. As well known, Fellegi and Sunter decision rule is very effective for link identification but generally less reliable for result evaluation. This proposal aims at predicting linkage errors in the Fellegi and Sunter framework, introducing a supervised step. As in [2] a training sample is required where the true matching status is known, but differently from [2] no data transformation is required, the proposed method exploits further auxiliary information, instead.

Section 2 defines linkage errors, Section 3 contains a short description of the Fellegi and Sunter model with special regards to error evaluation task. In Section 4 the

new proposal for estimating linkage errors is described and Section 5 presents an application on fictitious data. Finally, Section 6 describes our conclusions.

2. Record linkage and linkage error measurement

Record linkage is a multidisciplinary set of methods aiming to recognise the same real world entity, which may be represented differently in different data sources, through appropriate unit identifiers. As is well known, record linkage can be seen as a classification problem in which all the pairs generated by the Cartesian product of all the records coming from two (or more) files to be integrated must be classified into two disjoint sets, the set of Matches and the set of Unmatches. In this paper, following a widespread convention, the term 'match' refers to the true match status of the pair of records, whereas the term 'link' refers to the match status assigned to the pair by the automated matching process The integration quality indicators should represent the errors that can result from record linkage: therefore, they are essentially false match and missing match errors. In fact, at the end of the linkage procedure, one could observe records that are not linked even if they actually represent the same real world entity; so the two units remain distinctly assigned to the No-links set. In a similar way, one can expect false matches when the integration procedure assigns to the Links set a pair of units that do not actually refer to the same real-world entity.

The record linkage strategies often try to reach a compromise between the two types of errors: in fact, generally if one wants to decrease the number of false positives (false matches), a greater number of false negatives (missed matches) has to be accepted and vice versa. The two errors types may have different importance depending on the objectives of the specific problem of integration.

If the true linkage status were known, the measurement of these errors could be obtained on the basis of the table that compares it with the integration results: see Table 1.

From Table 1, the following ratios can be obtained:

- The false match rate: $b/(a+b)$
- The false un-match (missed match) rate: $c/(c+a)$

While false match rate and false un-match rate are the most common indicators in official statistics, in other fields linkage quality indicators are referred to as sensitivity and specificity, the proportion of matches that were correctly accepted as links and the proportion of un-matches that were correctly not accepted as links, respectively; in formulas:

- Sensitivity: $a/(a+c)$
- Specificity: $d/(d+b)$

Of course, in real life applications, the true linkage status is unknown. Anyway, providing measures of the linkage quality is the first step in order to improve the statistical process involving integrated data and the linkage process itself. However, testing quality of linkage outcomes is often very difficult, sometimes it is so difficult that quality is not measured at all. Nevertheless, the quality assessment of linkage results should be explicitly included in the linkage process, in order to allow comparison of different linkage procedures and to perform further statistical analyses on linked data avoiding biased and less efficient results. In fact, the standard statistical analysis of linked data, i.e. treating a linked file as perfectly linked and ignoring the integration procedure used to create the file, can lead to biased estimates, increased variability and errors on causality relationships.

In the following section, we discuss linkage error evaluation as an outcome of the Fellegi and Sunter model for probabilistic record linkage, highlighting weaknesses and strengths.

3. Linkage error estimation in Fellegi and Sunter model

A class of methods for the assessment of linkage errors mainly follow the approach of mixture models with latent variables. The variable (not observed) representing the real matching status is precisely the latent one, to be predicted by observing the results of the comparisons on observable variables (in models with un-supervised approaches) or by means of the known results in a training set (supervised models). The theory of Fellegi and Sunter [5] to solve record linkage problems can be viewed as un-supervised mixture model approach.

Define the true matching status as a latent variable g (like in [1,6], and others):

$$g_{(a,b)} = \begin{cases} 1 & \text{if } (a,b) \in M \\ 0 & \text{if } (a,b) \in U \end{cases} \quad (1)$$

According to the Fellegi and Sunter theory, this variable distribution (pairs classification criterion) is estimated on the basis of K observable common identifiers (matching variables). For each pair (a, b), a comparison function is applied in order to obtain a comparison

Table 1
Linkage results (in rows) and true linkage status (in columns)

	Matches	Un-Matches
Links	a – true positives	b – false positives or false links
No-links	c – false negatives or missing links	d – true negatives

vector $\boldsymbol{\gamma} = \{\gamma_1, \gamma_2, \ldots, \gamma_K\}$. The ratio

$$r_{(a,b)} = \frac{P(\gamma\,|(a,b) \in M)}{P(\gamma\,|(a,b) \in U)} = \frac{m(\gamma)}{u(\gamma)} \tag{2}$$

is the match weight, given by the ratio between the likelihood that the pair belong to the Matches set – $m(\gamma)$ – and the likelihood that the pair belong to the Un-Matches set – $u(\gamma)$ – given that the result of comparisons across matching variables assumes the pattern in the comparison vector γ.

In practice, once the probabilities m and u are estimated, for instance by means of the EM algorithm, all the pairs can be ranked according to the match weight $r_{(a,b)}$ and the classification criterion is based on two thresholds T_m and T_u ($T_m > T_u$). If the match weight $r_{(a,b)}$ is above the threshold T_m the pair (a, b) is assigned to the set of Link M *, if the weight $r_{(a,b)}$ is below the threshold T_u, the pair is assigned to the No-Link set U *, if the weight assumes a value between the two thresholds no decision on the linkage status for that pair is made. The Fellegi-Sunter rule is based on two assigned tolerable error levels: in fact, a pair composed of different unit could assume a value $r_{(a,b)} > T_m$, the frequency of this error is, attached to decision M*, is

$$\mu = \sum_{\gamma \in \Gamma} u(\gamma) P(M^*|\gamma) = \sum_{\gamma \in \Gamma_{M^*}} u(\gamma)$$
$$\text{where } \Gamma_{M^*} = \{\gamma : T_m \leqslant m(\gamma)/u(\gamma)\} \tag{3}$$

In the same way, a pair, that in fact is constituted by records belonging to the same real word entity, can assume value $r_{(a,b)} < T_u$. The frequency of this error is, attached to decision U*, is

$$\lambda = \sum_{\gamma \in \Gamma} m(\gamma) P(U^*|\gamma) = \sum_{\gamma \in \Gamma_{U^*}} m(\gamma)$$
$$\text{where } \Gamma_{U^*} = \{\gamma : T_u \geqslant m(\gamma)/u(\gamma)\} \tag{4}$$

Fellegi and Sunter suggest selecting the "acceptable" values for the error levels μ and λ, and once they are fixed, the threshold values T_m and T_u can be obtained by solving the above formulas. The error values μ and λ can be estimated by means of the estimated distributions $\hat{m}(\gamma)$ and $\hat{u}(\gamma)$:

$$\hat{\mu} = \sum_{\gamma \in \Gamma_{M^*}} \hat{u}(\gamma) \quad \hat{\lambda} = \sum_{\gamma \in \Gamma_{U^*}} \hat{m}(\gamma) \tag{5}$$

The error estimation as proposed in the Fellegi and Sunter approach is highly dependent on the distributions estimates $\hat{m}(\gamma)$ and $\hat{u}(\gamma)$ precision. Errors in model specification (e.g. the reliability of the conditional independence assumption), lack of information and so on, may cause loss of precision in the distributions estimates and consequently strong bias in the error estimates. It is important to note that the same model specification problems, however, have only a minor effect on the method effectiveness in identifying matches, given that the weight $r_{(a,b)}$ still allows to achieve a correct rank for pairs classification.

An important outcome of the Fellegi and Sunter procedure is given by the E-step of the EM algorithm applied to estimate probabilities $\hat{m}(\gamma)$ and $\hat{u}(\gamma)$. In fact, it provides the posterior probability that $g_{(a,b)}$ equals 1, in other words the estimates of the probability of being correctly matched given the observed comparison vector γ. Using Bayes theorem,

$$^{FS}\hat{g}_{(a,b)} = E\left(g_{(a,b)}\,|\gamma\,(a,b)\right) = \frac{m\,(\gamma) \times P\,[M]}{m\,(\gamma) \times P\,[M] + u\,(\gamma) \times P\,[U]} \tag{6}$$

where $P\,[M]$ and $P\,[U]$ are the (estimated) probabilities of Matches and Un-Matches sets, respectively.

The scenario at the end of the Fellegi and Sunter procedure is drawn in Fig. 1.

Theoretically, the number of false matches as well as the number of missed matches can be estimated via the $^{FS}\hat{g}_{(a,b)}$

- Estimated number of false matches $= \sum_{^{FS}\hat{g}_{(a,b)} > T_m} (1 - {}^{FS}\hat{g}_{a,b})$
- Estimated number missing matches $= \sum_{^{FS}\hat{g}_{(a,b)} < T_u} ({}^{FS}\hat{g}_{(a,b)})$

Clearly, these estimates are strongly dependent on accuracy of estimates $\hat{m}(\gamma)$ and $\hat{u}(\gamma)$. It is well known that $^{FS}\hat{g}_{(a,b)}$ (as $r_{(a,b)}$) is reliable in ranking the pairs distinguishing in Matches and Un-matches, even if its point estimate is not so accurate in evaluating false matches and missing matches. In the next section, we show how to obtain reliable linkage error estimates enriching the Fellegi and Sunter results with further auxiliary information.

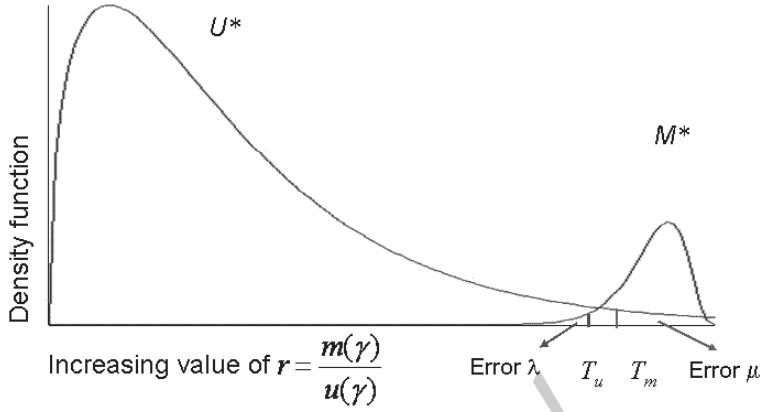

Fig. 1. Estimated probability distributions for Matches and Un-matches.

4. Enhancing Fellegi and Sunter procedure to evaluate linkage errors

The estimates ${}^{FS}\hat{g}_{(a,b)}$ of being correctly matched are dependent on the matching variables introduced in the comparison vector γ. For modeling issues, i.e. for avoiding over-parameterized log-linear models and too sparse tables, only the most informative variables are included in vector γ, while other variables due to their (relative) small power to distinguish between matches and non-matches remains out of the ${}^{FS}\hat{g}_{(a,b)}$estimates. For instance, dealing with linkage of people, obviously variables – such as names, surnames, complete date of birth – will be included in the Fellegi and Sunter model, while other variables with few categories – as sex, marital status, educational level – will remain out of the model. The idea of the proposed method is to exploit the latter variables in a second step in order to enrich the Fellegi and Sunter results particularly to measure matching errors.

The proposed method for evaluating linkage errors considers a training sample selected from the outcome of the Fellegi and Sunter procedure, before the assignment of the thresholds for classifying pairs, so in the sample both suspected matches and suspected unmatches are included. In the sample selection a crucial role is played by the ${}^{FS}\hat{g}_{(a,b)}$ outcome; since it is able to rank pairs in order to distinguish among Matches and Un-Matches, the training sample has to be selected considering the ${}^{FS}\hat{g}_{(a,b)}$ as a stratification variable.

For the pairs belonging to the training sample the true matching status is assigned, i.e. by means of clerical resolution. Then such true matching status is modeled using as input variables the ${}^{FS}\hat{g}_{(a,b)}$ outcome of the Fellegi and Sunter procedure and the results of comparisons on other variables not included in the Fellegi and Sunter model.

Indicate with ${}_{ts}g_{(a,b)}$ the matching status assigned to the pair (a, b) belonging to the training sample *ts*, let $X_1^{-F.S}$, …, $X_K^{-F.S}$ be K matching variables not included in the Fellegi and Sunter procedure and

$$I\left(X_k^{-F.S}{}_{(a,b)}\right) = \begin{cases} 1 \text{ if } X_k^{-F.S} \text{ assumes the same value in } a \text{ and } b \\ 0 \text{ otherwise} \end{cases} \quad (7)$$

The matching status can be modeled, for instance, according to logistic function as

$$pr({}_{ts}g_{(a,b)}) \sim \frac{\exp({}_{ts}\beta_0\ {}^{FS}\hat{g}_{(a,b)} +_{ts} \beta_1 I(X_1^{-F.S}{}_{(a,b)}) + \cdots}{1 + \exp({}_{ts}\beta_0\ {}^{FS}\hat{g}_{(a,b)} +_{ts}\beta_1 I(X_1^{-F.S}{}_{(a,b)}) + \cdots} \frac{+ \cdots +_{ts} \beta_K I(X_K^{-F.S}{}_{(a,b)}))}{+ \cdots +_{ts} \beta_K I(X_K^{-F.S}{}_{(a,b)}))} \quad (8)$$

The model estimated parameters ${}_{ts}\hat{\beta}_0$, ${}_{ts}\hat{\beta}_1, \ldots, {}_{ts}\hat{\beta}_K$ are then applied on the full data in order to obtain a prediction for the matching status:

$$pr({}^{2}\hat{g}_{(a,b)}) = \frac{\exp({}_{ts}\hat{\beta}_0^{FS}\hat{g}_{(a,b)}) +_{ts} \hat{\beta}_1 I(X_1^{-F.S}{}_{(a,b)}) + \ldots}{1 + \exp({}_{ts}\hat{\beta}_0^{FS}\hat{g}_{(a,b)}) +_{ts} \hat{\beta}_1 I(X_1^{-F.S}{}_{(a,b)}) + \ldots} \frac{+ \ldots +_{ts} \hat{\beta}_K I(X_K^{-F.S}{}_{(a,b)})}{+ \ldots +_{ts} \hat{\beta}_K I(X_K^{-F.S}{}_{(a,b)})} \quad (9)$$

Coherently, links can be assigned on the basis of ${}^{2}\hat{g}_{(a,b)}$, accordingly the usual criterion that the pair is considered a link if ${}^{2}\hat{g}_{(a,b)} > 0.5$. Moreover, the associated errors can be estimated:

Table 2
Summary of true and estimated values for False Match and False Un-Match Rates as resulting from the Fellegi and Sunter procedure

	False Match Rate		False Un-Match Rate	
	True values	Estimates	True values	Estimates
Min	0	0.019	0.176	0.007
Q1	0.001	0.023	0.192	0.008
Median	0.001	0.023	0.199	0.008
Mean	0.002	0.023	0.200	0.008
Q3	0.003	0.024	0.207	0.009
Max	0.006	0.025	0.225	0.013

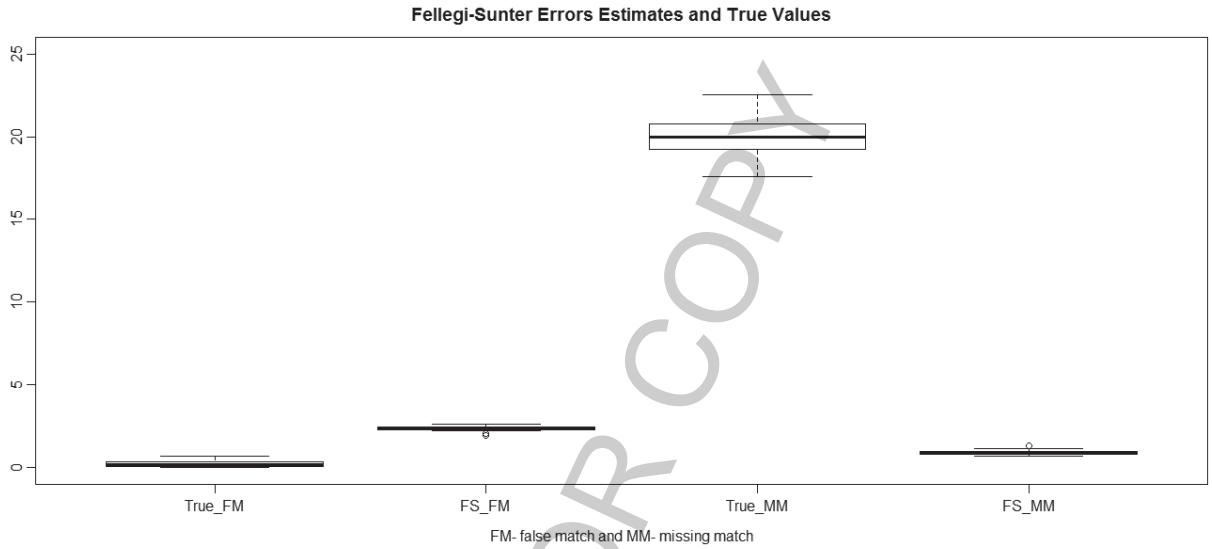

Fig. 2. Boxplot of true and estimated values for False Match and False Un-Match Rates (%) as resulting from the Fellegi and Sunter procedure.

Table 3
Logistic models for prediction in training samples

Model	Explanatory variables
Model 1	${}^{FS}\hat{g}_{(a,b)}$
Model 2	${}^{FS}\hat{g}_{(a,b)}$, *Sex*
Model 3	${}^{FS}\hat{g}_{(a,b)}$, *Postal code of residence*
Model 4	${}^{FS}\hat{g}_{(a,b)}$, *Day of birth*
Model 5	${}^{FS}\hat{g}_{(a,b)}$, *Sex*, *Postal code of residence*.
Model 6	${}^{FS}\hat{g}_{(a,b)}$, *Sex, Day of birth*
Model 7	${}^{FS}\hat{g}_{(a,b)}$, *Postal code of residence*, *Day of birth*
Model 8	${}^{FS}\hat{g}_{(a,b)}$, *Sex*, *Postal code of residence*, *Day of birth*
Model 9	*Sex*, *Postal code of residence*, *Day of birth*

- Estimated number of false matches $= \sum_{{}^{2}\hat{g}_{(a,b)}>0.5} \left(1 - {}^{2}\hat{g}_{(a,b)}\right)$
- Estimated number missing matches $= \sum_{{}^{2}\hat{g}_{(a,b)}<0.5} \left({}^{2}\hat{g}_{(a,b)}\right)$

Error estimates obtained in this way are more accurate and reliable than those based on the ${}^{FS}\hat{g}_{(a,b)}$, thanks to both the exploitation of further information contained in variables $X_1^{-F.S}, \ldots, X_K^{-F.S}$ and to the knowledge of the true matching status on a data subset. Nevertheless the Fellegi and Sunter outcome summarized by the ${}^{FS}\hat{g}_{(a,b)}$ variable is fundamental in training sample selection and model selection.

Table 4
Summary of the Area Under the ROC Curve (AUC) and prediction power according 0–1 loss function of the best model in replications

	AUC	Prediction power
Min	0.946	0.893
Q1	0.977	0.923
Median	0.984	0.942
Mean	0.982	0.941
Q3	0.991	0.959
Max	0.999	0.985

5. Results with fictitious data

In this section, the proposed method effectiveness is shown with respect to fictitious data representing population census and administrative register, where the true matching status is known. Data are characterized by presence of errors in matching variables that mimics reality. These data were created for the ESS-

net DI [4], which was a European project on data integration (Record Linkage, Statistical Matching, Micro integration Processing) run from 2009 to 2011. The data are freely available online at: http://www.cros-portal.eu/content/job-training (accessed 20 July 2015).

For this application, 100 settings were replicated, each one consisting of a population of about 1500 units, selected from census and register on the basis of the *Day of birth* variable. In each setting, the variable *Day of birth* assumes only two values.

Then, probabilistic record linkage was performed, according Fellegi and Sunter theory as implemented in the Relais software [7], using the most informative variables (namely *Name, Surname, Month and Year of Birth*), while other available common variables remain out of the Fellegi and Sunter model (namely S*ex, Day of birth, Postal code of residence*).

The outcomes of the Fellegi and Sunter procedure were assigned to the Links and No-Links sets on the basis of a single threshold corresponding to ${}^{FS}\hat{g}_{(a,b)} = 0.5$.

Results of that procedure are shown in Table 2, where summary values of the 100 replications are reported for the False Match and the False Un-Match Rates, comparing the observed true values and their estimates.

Table 2 clearly shows that the Fellegi and Sunter procedure works well with respect to the false match error, even if the correspondent rate overestimates its level. On the other hand, the Fellegi and Sunter procedure doesn't perform so well with respect to the missing matches, while it strongly underestimates the corresponding error. The situation is clearly represented in Fig. 2.

According to the proposal method explained in previous section, the results of the Fellegi and Sunter procedure were enriched: a training sample of about 150 pairs (about 10% of the original data) was selected from the Fellegi and Sunter results before the thresholds assessment. The outcome ${}^{FS}\hat{g}_{(a,b)}$ was used as stratification variable, pairs were drawn with equal probabilities within strata of equal sample size.

Model selection procedure was applied to predict the true matching status; nine models were considered in each replication, as reported in Table 3: the ${}^{FS}\hat{g}_{(a,b)}$ variable was included as explanatory one plus a combination of the others remained out of the Fellegi and Sunter model: *Sex, Day of birth, Postal code of residence.* The only model not containing ${}^{FS}\hat{g}_{(a,b)}$ as input variable is Model 9, other models without ${}^{FS}\hat{g}_{(a,b)}$ were tested but not reported here due to their poor predicting power.

In each replication, the nine models were compared on the basis of the residuals deviance, the AIC index, the area under the ROC curve, optimality with respect to entropy and 0–1 loss functions. Finally, at each replication the best predictive model was selected on the basis of the highest number of best performances with respect to the previous five indicators. For instance, in Table 4 summary of the area under the ROC curve and prediction power according 0–1 loss function are reported for the best models in the replications.

On the 100 replications, Model 8 was the best 59 times, Model 5 was the best 37 times, finally Model 7 was the best 4 times.

In each replication, the best model was applied to predict matching status and linkage errors on the full data. Table 5 reports summary of the False Match and the False Un-Match Rates, comparing the observed true values and their estimates, obtained on the basis of the predicted values ${}^{2}\hat{g}_{(a,b)} = 0.5$.

Results in Table 5 can be compared with results in Table 2. Figures 3a and 3b show the comparison of the true values and the estimates as resulting from the Fellegi and Sunter (FS_) procedure and the proponed method (New_), for the False Match Rate and the False Un-Match Rate respectively.

Figures 3a and 3b clearly show that the proposal method produces estimates of error rates closer to their relative true values than the Fellegi and Sunter procedure. This fact is particularly important in order to evaluate in a proper way the integration procedure itself. Moreover in this way it is possible to achieve more appropriate values for error rates and linkage probabilities to include in subsequent analyses on linked data.

6. Conclusions

The integration procedure quality assessment is a pressing task due to the increasing combined use of data coming from different sources. In this article a method to improve the linkage errors estimates in the Fellegi and Sunter framework is proposed. The proposal enriches the Fellegi and Sunter procedure, applied on the most strong matching variables, via a training sample, where the known true matching status is modeled using as explanatory variables the Fellegi and Sunter outcome plus other still not-exploited comparison variables.

The proposed method allows to estimate in the same time both false matches and missing matches. In the

Table 5
Summary of true and estimated values for false and missing match errors as resulting from the proposed procedure

	False Match Rate		False Un-Match Rate	
	True values	Estimates	True values	Estimates
Min.	0.009	0.000	0	0
1st Qu.	0.019	0.009	0.006	0.005
Median	0.023	0.012	0.007	0.007
Mean	0.023	0.014	0.007	0.007
3rd Qu.	0.027	0.016	0.009	0.010
Max.	0.043	0.053	0.013	0.015

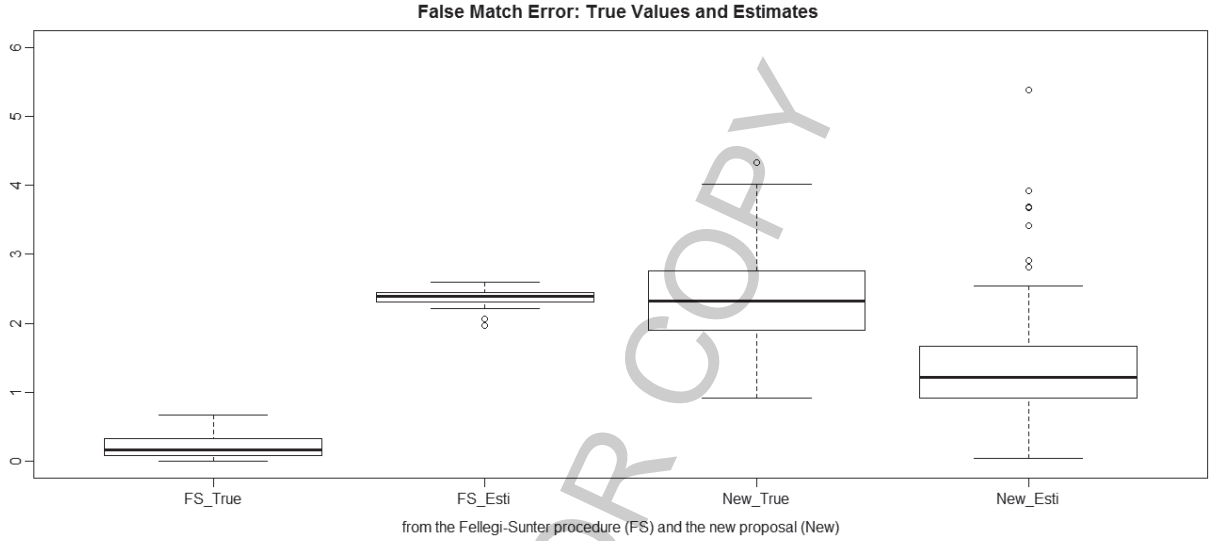

Fig. 3A. Boxplot of true and estimated values for False Match Rates (%) as resulting from the Fellegi and Sunter procedure and the proposed method.

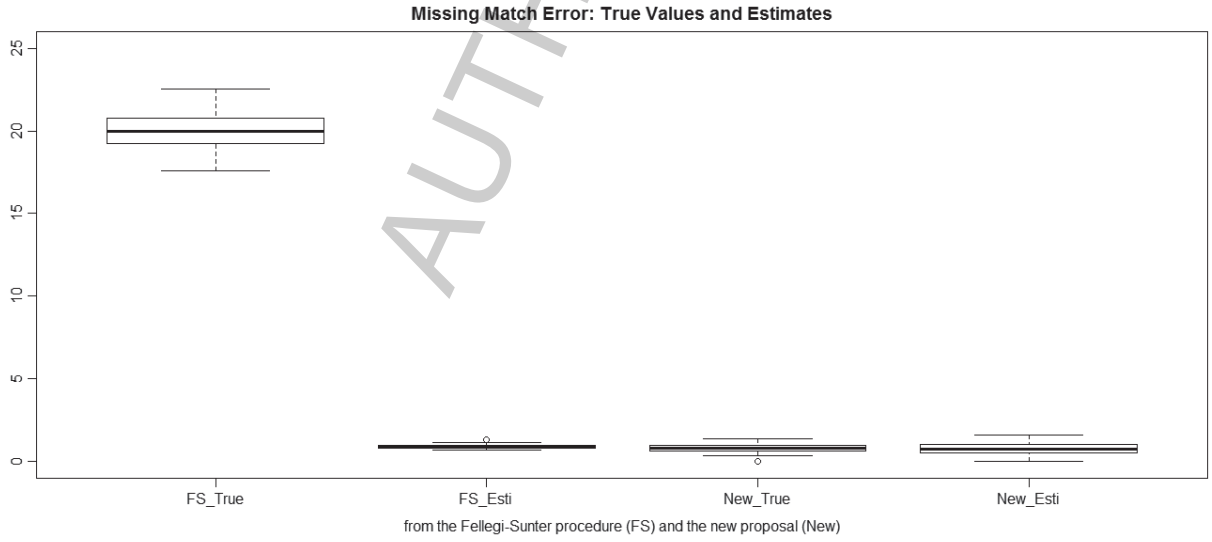

Fig. 3B. Boxplot of true and estimated values for False Un-Match Rates (%) as resulting from the Fellegi and Sunter procedure and the proposed method.

simulation context, the method seems very promising due to its flexibility: differently from other methods based on mixture model [2], it doesn't rely on the distribution assumption. Moreover the method seems robust with respect to training sample selection as well as the model selection on the training sample. The best model (in terms of AIC, area under the ROC curve, entropy and 0–1 loss functions) on the training sample could not be the same of the one on full data. This seems to not compromise the prediction power of the selected model.

Finally, the method can be incorporated in a modular way at the end of the Fellegi and Sunter procedure in order to complete and assess the quality of the procedures itself.

References

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

[2] T.R. Belin and D.B. Rubin, A Method for Calibrating False-Match Rates in Record Linkage, *Journal of American Statistical Association* **90** (1995), 81–94.

[3] R. Chambers, Regression Analysis Of Probability-Linked Data. Official Statistics Research Series 4, 2009.

[4] Essnet DI – McLeod, Heasman and Forbes, (2011) Simulated data for the on the job training, http://www.cros-portal.eu/content/job-training.

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

[6] M.A. Jaro, Advances in record linkage methodology as applied to matching the 1985 Census of Tampa, Florida, *Journal of the American Statistical Association* **89** (1989), 414–420.

[7] RELAIS, (2011). User's guide version 2.2, available at http://joinup.ec.europa.eu/software/relais/release/22.