New proposals for linkage error estimation
Tiziana Tuoto (tuoto@istat.it), Niki Stylianidou

Keywords: probabilistic record linkage, linkage quality assessment, mixture models

1 INTRODUCTION

The use of combined data from different sources is an advantage that the National Institutes of Statistics try to take advantage of in more and more occasions.

In a context in which large information are produced by different sources, achieved through integrated and comparison operations (methods and techniques), it becomes increasingly urgent the need to equip the results of data integration and linkage with quantitative assessments of the quality of such combining operations.

As part of the integration of data at a micro level, the currently applied methodologies of record linkage, are widely used within the Italian National Institute of Statistics (Istat). These linkage methodologies generally produce good results, mainly when applied by means of the strong identification matching variables. However evaluation information i.e. quantitative indicators of quality of the output of the matching procedures, is rarely available. In official statistics, evaluation information regarding the output (matching data) of a procedure has a fundamental role since “certify” the accuracy and credibility of the data; especially if the aim is then to use them for further analysis or for inference.

2 DATA INTEGRATION AND QUALITY

The following matrix identifies the possible combinations that a matching methodology may produce compared to the truly linked data. The given matrix is used for building the quality indicators of the record linkage procedure:

Table 1 – Output of the record linkage procedure

<table>
<thead>
<tr>
<th></th>
<th>Linked</th>
<th>No Linked</th>
</tr>
</thead>
<tbody>
<tr>
<td>Matched</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>Unmatched</td>
<td>c</td>
<td>d</td>
</tr>
</tbody>
</table>

The following indicators can be calculated:

- **Rate of false matching:** \[ f = \frac{b}{(a+b)} \]
- **Rate of missing matching:** \[ m = \frac{c}{(c+a)} \]

Usually, the variable Linked / No Linked, that determines the true state of a match, is not observed. The methods for the estimation of the Linked / No Linked variable are mainly two:

a) with very accurate manual revision of the matching outputs by clerks;

b) the use of statistical methods such as latent variable mixture models (Fellegi and Sunter, 1969; Belin and Rubin, 1995).

---

1 Istat – Italian National Statistical Institute
2.1 The Belin and Rubin method

The most defused method for determining the matching procedure failures, mainly of those of false matches, it has been proposed by Belin and Rubin (1995). Their proposal had as a starting point a training set extracted from the records of declared up matches. Once calculated the matching weights, $R^2$ are firstly logarithmic transformed, and thus a second transformation is applied for rescaling them in the range 1-1000. Finally, last a Box-Cox transformation is applied on the log-rescaled R’s, in order to give to the weights a normal distribution shape. At this point, the authors apply a Fitting Transformed Normal Mixtures with Fixed global parameters. The model refers to the latent variable that expresses the state of false matches for records declared matches, ie the cell $b$.

The proposed method of Belin and Rubin (1995) already-known problems that suffers the are:

- the difficult construction of a representative training set for which crucial information are known as the rates of truly linked matching records and thus being able to reproduce the weight distribution of the two populations Linked and No linked as observed in the universe of Matches;

- the fact that the distribution of the transformed weights of matches does not always be adapted to a normal distribution through a Box-Cox transformation.

Here in Istat a practical replication of the Belin and Rubin method has been tempted facing the above exposed problems and thus aiming no results. The attempt to replicate the proposed methodology by Belin and Rubin (1995) has been helpful for focusing on the need to seek a rule that distinguishes the True Linked vs False Linked from the matched records dataset. For this reason, we have tried to apply the methodology of Canonical Discriminant Analysis in the context of record linkage. The seek result seems to be the very similar as the one of Fisher’s when in 1936 introduced this methodology regarding a dataset with two different kind of Iris flowers.

2.2 Discriminant Analysis

For exposing the method Discriminant Analysis, some notation is necessary:

Let’s the vector of the match records be $Y = (Y_1, Y_2)$

Where $Y_1$ are the true linked matched records (from table 1 is $a$)

$Y_2$ is its complement alias the false-linked but matched records (from table 1 is $b$).

The covariance matrix of the matched records is: $X = (X_1, X_2)$

denoting with $\bar{x}_1$ the vector that contains the averages of the matched truly linked covariates

and with $\bar{x}_2$ the vector that contains the averages of the matched false linked covariates.

Respectively are denoted with $S_1$ = the matrix of Variance of the matched true linked and $S_2$ = the Variance matrix of the matched false linked.

$n_1$ number, indicates how many matched true-linked records are and $n_2$ is the corresponding number of matched false linked records.

Calculating the Variance and Covariance matrices:

---

2 for example, the ratio of the likelihoods obtained through a procedure proposed by Fellegi and Sunter, although several other different procedures can also be considered for obtaining the weights
Between clusters

\[ B = \frac{n_1 n_2}{n_1 + n_2} (\bar{x}_1 - \bar{x}_2)(\bar{x}_1 - \bar{x}_2)^t \]

and

within clusters

\[ W = \frac{S_1(n_1 - 1) + S_2(n_2 - 1)}{n_1 + n_2 - 2} \]

that are necessary for construction so called Fishers’ equation of discrimination:

\[ Y = XW^{-1}(\bar{x}_1 - \bar{x}_2) \]

The discrimination point between the two groups (Breaking point) is given by the:

\[ \text{break} = \frac{\bar{y}_1 + \bar{y}_2}{2} \]

and thus the assignment rule for each unit is given by:

\[ \min \triangleq \{ |\bar{y}_1 - y_i|, |\bar{y}_2 - y_i| \} \forall y_i \]

That is each unit of \( y \) will be assigned to the population whose center of gravity \( \bar{y}_i \) is the least distant.

Once established the effectiveness of this method, using different applications, it may be proposed to improve the data produced by the record linkage using a training set to extract the optimal rule to be applied on the whole of Matched (in order to estimate \( b \)) but also on all the unmatched (in order to estimate \( c \)).

3. Practical Applications

The methods proposed, the Bellin and Rubin and the Discriminant Analysis, have been experimented with the aid of simulated data, for which the real state of linkage, is known through the availability of a univocal identification code.\(^3\)

3.1 The data in use and the linkage scenarios

The data used in this exercise have been created by Paula McLeod, Dick Heasman and Ian Forbes, of the Office of National Statistics (UK), within the European project: ESSnet Date Integration. The two methods have been tested in two different sceneries of linkage: the first one introduces a rate of false linkage slightly lower than 2%; this it will be pointed out as Gold Scenario. In this case it has been drawn out a sample from the total of the available data, around 1500 unities. Highly identifying variables has been used for the matching procedure according to the model of Fellegi-Sunter, as implemented in the open source software for the record linkage RELAIS.

In the first matching exercise, as the dimensions of the two datasets are small, it has been applied the matching cross product of all the records. The identifiable variables used in the matching procedure are: Name, Surname, day and year of birth, allowing individualizing 1300 matches of which 45 false linked matches. The matching threshold of \( p_{\text{post}} \) is 0.6.

A second scenario of linkage has been built considering all the records that compose the ons reference file that is 26000 records. As the number records is high the search space dimension has been reduced using the SimHash function in selected blocks. As matching identifiable variables has been selected the same as in the Gold Scenario. This matching strategy gave as output 21560 matched records of which 1336 are false linked matches.

\(^3\) Thus the main aim of this work is to have a methodology that will guarantee us a high level quality matched records of micro data and thus an estimation of the possible false linked matches that may contain (error).
linked matches ie around 6%. The matching threshold was set at p_post=0.8. This scenario will be indicated as Silver.

A last scenario, named Platinum, takes the outcome of the Gold Scenario that is the 1300 matched records and within them are run with RELAIS other two matching models using different identificative variables ie address and number of cap. Thus the Platinum scenario is a merge of 3 different matching models giving us an output of 1652 pair matches with 327 false linked and 1325 true linked.

### 3.2 The Results of Discriminant Analysis method

As a first approach it has been used the method indicating with Y the dummy variable of true vs false linked record regressed on only one covariable, X, the P_post (post probability, calculated from RELAIS that a record is truly matched).

The first application on the Silver scenario, gave us as a break point b=0.9571004. In fact applying the b on the p-post we have the following results:

<table>
<thead>
<tr>
<th></th>
<th>forcast_true</th>
<th>foresct_false</th>
</tr>
</thead>
<tbody>
<tr>
<td>TRUE</td>
<td>17555</td>
<td>2669</td>
</tr>
<tr>
<td>FALSE</td>
<td>16</td>
<td>1320</td>
</tr>
</tbody>
</table>

P_POST | forcast_false | foresct_true
0.8666  | 706          | 0
0.93581 | 3283         | 0
0.99387 | 0            | 17571

The above results assure us that if we use as indication the above break-point (just one covariate) the estimated failure of this method is 12,5% ((2669+16)/21,560) while the actual false records that are indicated as true have a rate of 0,1% (16/21,560).

Similar results we receive from the second application, still using only one covariance the p_post on the Gold Scenario:

<table>
<thead>
<tr>
<th></th>
<th>forcast_true</th>
<th>foresct_false</th>
</tr>
</thead>
<tbody>
<tr>
<td>TRUE</td>
<td>999</td>
<td>256</td>
</tr>
<tr>
<td>FALSE</td>
<td>0</td>
<td>45</td>
</tr>
</tbody>
</table>

P_POST | forcast_true | foresct_false
0.66549 | 17           | 0
0.79116 | 64           | 0
0.8376  | 47           | 0
0.86571 | 173          | 0
0.99982 | 0            | 999

The estimated error in this case is nearly 20% while the rate of false linked matches considered as true is 0%.

Applying the method on the Platinum scenario and thus having three covariate ie the three p_post of each model elaborated by RELAIS, the failure rate lowers to 8% however as the break point is a combination of three different models is not easy to handle.

### Conclusions
The aim of this exercise is to find a method that can help us on the estimation of the errors in the record linkage. In this particular exercise has been took under consideration two main methods the Bellin Felligi method and the Discriminant Analysis.

The Bellin-Fellegi method gave us no results as the assumptions that must hold in this method are too restrictive and thus not applied in our datasets in hand. The Discriminant Analysis method give us reasonable results that can be considered satisfying.