New proposal for linkage error estimation

Tiziana Tuoto
Istat – Italian National Statistical Institute

Joint work with: Niki Stylianidou

Brussels, 11th March 2015
Outline

1. Motivations: linkage errors and their effects

2. Drawbacks of current solutions for evaluating linkage quality

3. A new proposal

4. Concluding remarks and future works
Motivations

Why we are still studying linkage errors in order to:

✓ provide necessary information for proper statistical analyses on linked data

✓ Increase the level of automatism of integration process
  → Move from black box towards toolbox
Solutions for linkage quality

Approaches based Latent Mixture Model:

✓ **Fellegi and Sunter** (1969)

The procedure is reliable in identifying matches, not so accurate in assessing errors. A precious outcome of the F-S procedure is the posterior probability of being match for couple $(a,b)$

$$
post _ {prob}(a,b \mid \gamma) = \frac{P[\gamma(a,b) \mid M] \times P[M]}{P[\gamma(a,b) \mid M] \times P[M] + P[\gamma(a,b)U] \times P[U]}
$$

✓ **Belin and Rubin** (1995)

After the F-S procedure, their suggestion assumes transformation of the Matches and Unmatches distributions in order to normalize them. They only achieve estimate of false match rate. Actually at the end, normality is not guaranteed.
Scenario at the end of Fellegi-Sunter application...

Increasing values of comparison values - likelihood ratio

\[ w(ab) = \frac{f(y(ab)|M)}{f(y(ab)|U)} \]
A new proposal for evaluating matching errors

**Main idea:** enriching F-S results via auxiliary info in a second step

- The Fellegi and Sunter probabilistic record linkage provides estimates of the probability of being correctly matched – the $post\_prob$ – applying log-linear model on the most informative variables while other variables are not included for modeling issues due to their (relative) small power to distinguish between matches and non-matches (as sex, marital status, ... few categories uniformly distributed)

- Those variables could be exploited in a second step in order enrich the F-S results particularly to measure matching errors

Tuoto "New proposals for linkage error estimation" Brussels, 11 March 2015
The proposal: second step

- From matches of the F-S procedure, sampling a training set for which the true matching status is assigned and modeling such status via the \( post\_prob \) and the comparisons on other variables not included in the F-S model

\[
\text{ts}y_{(a,b)} \sim \logit \left( post\_prob_{(a,b)} + I \left( X_2_{(a,b)} \right) + I \left( X_3_{(a,b)} \right) + \cdots \right)
\]

- The model is applied on the full data

\[
\hat{y}_{(a,b)} \sim ts\beta_1 post\_prob_{(a,b)} + ts\beta_2 I \left( X_2_{(a,b)} \right) + ts\beta_3 I \left( X_3_{(a,b)} \right) + \cdots
\]

- Matches are assigned on the basis of \( \hat{y}_{(a,b)} \)

\[
\text{If } \hat{y}_{(a,b)} > 0.5 \text{ then the pair is matched}
\]

- Errors are estimated on the basis of \( \hat{y}_{(a,b)} \)

\[
\text{false matches } \sum_{\hat{y}_{(a,b)}>0} (1 - \hat{y}_{(a,b)}) \\
\text{missing matches } \sum_{\hat{y}_{(a,b)}<0} \hat{y}_{(a,b)}
\]

Tuoto "New proposals for linkage error estimation" Brussels, 11 March 2015
Some results on fictitious data

First Step:

✓ **Full data**: 1500 units from fictitious population census data and administrative register, the true matching status is known (ESSnet DI, 2011)

✓ Probabilistic record linkage, according Fellegi and Sunter theory, using the most informative variables (name, surname, day and year of birth), other variables remain out of the F-S model (sex, month of birth, postal code...)

Results of the F-S procedure on the basis of \( post\_prob > 0.5 \):

<table>
<thead>
<tr>
<th>True matching errors</th>
<th>Estimated matching errors</th>
</tr>
</thead>
<tbody>
<tr>
<td>0%</td>
<td>4%</td>
</tr>
<tr>
<td>2%</td>
<td>0%</td>
</tr>
</tbody>
</table>

Tuoto "New proposals for linkage error estimation" Brussels, 11 March 2015
Some results on fictitious data

Second Step:

- Training set randomly sampled from the F-S results (size 10%)
- Logistic model selection to predict the true matching status using post\_prob
  plus other variables remain out of the F-S model: sex, month of birth, postal code

Some notes on second step:

1. the best model (in terms of AIC, area under the ROC curve, entropy and 0-1 loss functions) on the TS and the one on full data (if true linkage status were known) could not be the same: this seems to not compromise the prediction power of the selected model

2. from first evidence, the logistic model should include as explanatory variable the post\_prob
Some results

True matching errors

- 17%
- 2%

Estimated matching errors

- 15%
- 2%

On the training set

True matching errors

- 1.9%
- 0.5%

Estimated matching errors

- 1.7%
- 0.3%

Projection on the full data

Declared
False
Missing

Tuoto "New proposals for linkage error estimation" Brussels, 11 March 2015
Concluding remarks

Proposal: two-step procedure in order to assign linkage status and estimate linkage errors as well

► Step 1. Fellegi-Sunter procedure is applied on the most strong variables
► Step 2. on the F-S results, a training set is sampled, the true matching status is modeled on the basis of the post_prob and still not-exploited comparison variables

Advantages: all available comparison variables are exploited
Advantages: Estimates in the same time false and missing matches

Further works:
► 1. analyze the impact of training set selection
► 2. analyze the impact of model selection on the TS and the projection on the full data
References

