Small Area Estimation on linked data: a comparison between frequentist and Bayesian approaches

Brunero Liseo

MEMOTEF Department
Sapienza University of Rome

NTTS 2017 Conference
Bruxelles, 15/03/2017

1joint work with Dario Briscolini, Andrea Tancredi (Sapienza University of Rome), Loredana Di Consiglio (Eurostat), Tiziana Tuoto (Istat)
Outline

1 Motivation

2 Overview on linkage techniques
   Notation

3 Overview on Small Area Estimation

4 Record Linkage in Small Area Model
Linkage Problems

⇒ Linking two or more data sets can be important for two different and complementary reasons:

(i) per sé, i.e. to obtain a larger reference data set or frame
   • Useful for administrative tasks
   • More accurate summary statistics

(ii) to calibrate statistical models via the additional information which could not be extracted from either one of the two single data sets.
   • Linear and logistic regression
   • Population size estimation
   • Small area estimation
Summary

• In this talk we will focus on the methodological aspects of (ii) in the small area setting.

• We present a comparative analysis of several different estimators of the parameters of a unit-level small area model both from a classical and a Bayesian perspective.

• We compare the results on a pseudo population whose values \( Y \) and \( X \) are obtained from the survey on Household Income and Wealth, Bank of Italy and the person identifiers come from the fictitious population census data created for the ESSnet DI, an European project on data integration run from 2009 to 2011.
Linked data: the bias effect

• Assume we observe $Y, V_1, \ldots, V_h$ in a file and $X, V_1, \ldots, V_h$ in the other one.

• Consider a regression of $Y$ on $X$ based on pairs declared matches after a RL analysis based on $(V_1, \ldots, V_h)$


• The presence of false matches reduces the observed level of association between $Y$ and $X$.
  
  • bias effect towards zero when estimating the slope of the regression line.

• Similar biases may appear in any statistical procedure: for example, false matches reduces the final estimate of $N$ when RL methods are used in capture-recapture models for estimating population size.
Standard RL methods

- Records (or transformations thereof) are compared each other
- Some metric is used to measure “distance” between pairs
- A decision (either based on a test or a posterior probability) is taken.
- Output: few matches and a huge number of non matches.
- Curse of dimensionality; difficult to generalize to $k$ files
The ingredients

- Two datasets $A$ and $B$ consisting respectively of $n_A$ and $n_B$ units
- The datasets $A$ and $B$ include variables $V_1, V_2, \ldots, V_k$ called key variables or matching variables.
- The degree of overlapping between the two datasets is generally unknown.
- The role of the key variables makes the linkage process feasible. Age, sex, marital status, etc., are examples of key variables: they are usually categorical.
- The identifying variable (i.e., tax code) is usually missing.
- The key variables are subjected to measurement error.
The Frequentist record linkage (Fellegi-Sunter)

The linkage process between $A$ and $B$ can be viewed as a classification problem in $M$ (matched pairs) and $U$ (unmatched pairs) on the basis of:

$$ r_{ab} = \frac{P(\gamma_{ab}|(a, b) \in M)}{P(\gamma_{ab}|(a, b) \in U)} = \frac{m(\gamma_{ab})}{u(\gamma_{ab})} $$

where $m(\gamma_{ab})$ is the conditional probability of the comparison vector $\gamma$ between $a \in A$ and $b \in B$ given that the pair belongs to set $M$ and $u(\gamma_{ab})$ is the conditional probability of the comparison vector $\gamma$ given that the pair belongs to set $U$.

The $m$ and $u$ can be estimated by assuming the true link status is a latent variable, using, for instance, the EM algorithm.
**The Frequentist record linkage: the errors**

Thresholds on the ratio $r$ are chosen to minimize false link probability as well as false non link probability.

The frequentist record linkage provides as output an evaluation of the probability of being a correct link given that the link is assigned:

$$\lambda_{ij} = \frac{m(\gamma_{ij})P(M^*)}{m(\gamma_{ij})P(M^*) + u(\gamma_{ij})P(U^*)}$$

These quantities $\lambda_{ij}$ will be exploited for adjusting the linkage errors in the small area estimation framework described in the next section.
The Bayesian approach: the C matrix
We call $C$ a $n_A \times n_B$ 1/0 matrix such that for a generic pair $(i \in A, j \in B)$:

$$C_{ij} = \begin{cases} 1 & \text{if the pair } (i, j) \text{ is a match} \\ 0 & \text{otherwise} \end{cases}$$

The aim of the record linkage process is to estimate the $C$ matrix.

In a Bayesian setting one looks for the posterior distribution of the $C$ matrix.

Data are the values observed on the key variables.

Prior for $C$: A uniform prior can be used. Alternatively if $M$ is the number of matches the following structure can be assumed:

$$M \sim p(M = m); \quad P(C|M = m) \sim \text{Uniform.}$$
Unit Level Small Area Estimation

Let $Y$ be the target variable and $X$ auxiliary variables, the population units can be grouped in $D$ domains

$$Y = X\beta + Zu + e$$

where $Z$ is the area design matrix

$$Z = \text{Blockdiag}(Z_d = 1_{N_d}) \quad d = 1 \cdots D$$

The total variance is then $V(Y) = V = \sigma_u^2 ZZ^T + \sigma_e^2 I$ or

$$V = \text{Blockdiag}(V_d; d = 1 \cdots D)$$

with

$$V_d = \sigma_e^2 I_{N_d} + \sigma_u^2 Z_d Z_d^T$$


**EBLUP**

\[
\hat{Y}_{BLUP}^{d} = \frac{1}{N_d} \left( \sum_{i \in s_d} y_{id} + \sum_{i \in s^c_d} \hat{y}_{id}^{BLUP} \right)
\]

where \( \hat{y}_{id}^{BLUP} = X_{id}^T \hat{\beta} + \hat{u}_d \) with

\[
\hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} y
\]

and

\[
\hat{u} = \sigma_u Z^T V (y - X \hat{\beta})
\]

An EBLUP is obtained replacing the estimations \( \hat{\sigma}_u \) and \( \hat{\sigma}_e \) in the previous equations.
The unit linear mixed model under RL

- Let $y^*_{id}$ be the value of the variable matched with the value $X_{id}$.
- Let $Z_2$ be a blocking variable, measured without error on both the Y-register and the X-register, that partitions both registers so that all linkage errors occur within the groups of records defined by the distinct values of this variable.

Given the assumptions:

1. the linkage is complete, i.e. the X-register and Y-register refer to the same population and have no duplicates, so at least the smallest Y-register is contained in the biggest X-register
2. the linkage is one to one between the Y- and X-registers
3. exchangeable linkage errors model
The unit linear mixed model under RL

Under previous assumptions, the observed variable is a permutation of the true one

\[ Y^* = AY \]

where \( A \) is a random permutation matrix such that

\[ E(A|X) = E \]

Being \( Pr(a_{ii} = 1|X) = Pr(\text{correct linkage}) = \lambda \) and
\( Pr(a_{ij} = 1|X) = Pr(\text{uncorrect linkage}) = \mu \) then the expected value can be written as:

\[ E = (\lambda - \mu)I + \mu 11^T \]

This implies that \( \lambda + (N-1)\mu = 1 \) and the first order properties of the probability-linkage mechanism are completely specified by one parameter \( \lambda \).
Ratio type and BLU estimators

Samart and Chambers (2010) proposed a ratio type corrected estimator for $\beta$

$$\tilde{\beta}_R = (X^T V^{-1} E X)^{-1} X^T V^{-1} y^*$$  \hspace{1cm} (2)

Furthermore, by exploiting the relationship between the variable $y^*$ and $X$ a BLUE can be obtained

$$\tilde{\beta}_{BLUE} = (X^T E^T \Sigma E X)^{-1} X^T E^T \Sigma y^*$$  \hspace{1cm} (3)
Variance

The derived variance of the observed $y_d^*$

$$V(Y_d^*) = \Sigma_d^{-1} = \sigma_u^2 K_d + \sigma_e^2 I_d + V_d$$

where

$$V_d \approx \text{diag}((1 - \lambda_d)(\lambda_d(f_i - \bar{f}_d) + \bar{f}_d^{(2)} - \bar{f}_d^2))$$

being $f_i = X_i \hat{\beta}$ and $\bar{f}_d$, $\bar{f}_d^{(2)}$ denote the area $d$ averages if the components of $f_d$ and their squares respectively.

$K$ is a matrix where the element $k_{ij}$ is determined depending on whether $i$ and $j$ are in the same area. $K$ is a function of the number of areas within a block, the number of population units in area and block and $\lambda$s.
Estimation of variance components

We restrict to the ML, assuming multivariate normal distribution. Samart and Chambers (2010) use the method of scoring as an algorithm to obtain the estimators. In the standard case where the variables are recorded on the sample, we have $y \sim N(X\beta; V)$.

For the record linkage case, instead, recall that $y^* \sim N(Ef; \Sigma)$. The scoring algorithm can be applied on the derivatives of the previous likelihood.
The adjusted EBLUP estimator

Additional assumptions:

- The record linkage process is independent of the sampling process. Then the matrices $E$, $V$ and $\Sigma$ depend only on the blocking variables and linkage errors,

- Small area coincides with blocks.

$$\hat{Y}_d^* = \hat{Y}_d$$

Use the distribution of $Y^*$ to obtain the EBLUP estimator:

$$\hat{Y}_d^{*BLUP} = \frac{1}{N_d} \left( \sum_{i \in s_d} y_{id}^* + \sum_{i \in s_c} \hat{y}_{id}^{BLUP} \right)$$

where $\hat{y}_{id}^{BLUP} = E X \tilde{\beta}_{BLUE} + \hat{u}_d$ and $\hat{u} = \sigma_u Z^T \Sigma^{-1} (y^* - E X \tilde{\beta}_{BLUE})$. 
Basic Bayesian Small Area Model

Unit Level Model (Rao(2003))

- \( y_{ij} | \beta, u_i, \sigma^2_e \overset{\text{ind}}{\sim} N(z_{ij}'\beta + u_i, \sigma^2_e), \ i = 1, 2, ..., m; \ j = 1, 2, ...n_i \)
- \( u_i | \sigma^2_u \overset{\text{iid}}{\sim} N(0, \sigma^2_u), \ i = 1, 2, ..., m \)
- \( \pi(\beta) \propto 1 \)
- \( \sigma^2_u \sim IG(a_u, b_u) \)
- \( \sigma^2_e \sim IG(a_e, b_e) \)
- \( \pi(\beta, \sigma^2_e, \sigma^2_u) = \pi(\beta)\pi(\sigma^2_e)\pi(\sigma^2_u) \)

The goals of inference are:

- To obtain a MCMC approximation of the posterior distribution of the parameters of the model \( \pi(\beta, u, \sigma^2_e, \sigma^2_u | y) \).
- To produce a posterior sample from the predictive distribution of the totals at area level.
Standard Gibbs Sampling: Full Conditionals

- $\beta|\mathbf{u}, \sigma_e^2, \sigma_u^2, \mathbf{y} \sim N_K \left( \left( \sum_{i,j} z_{ij} z_{ij}' \right)^{-1} \sum_{i,j} z_{ij} (y_{ij} - u_i), \sigma_e^2 \left( \sum_{i,j} z_{ij} z_{ij}' \right)^{-1} \right)$
- for $i = 1, 2, \ldots, m$:
  
  $u_i|\beta, \sigma_e^2, \sigma_u^2, \mathbf{y} \sim N \left( \gamma_i (\bar{y}_i - \bar{z}_i' \beta), \gamma_i \sigma_e^2 / n_i \right)$

- $\sigma_e^2|\beta, \mathbf{u}, \sigma_u^2, \mathbf{y} \sim IG \left( \frac{n}{2} + a_e, \frac{1}{2} \sum_i \sum_j (y_{ij} - z_{ij}' \beta - u_i)^2 + b_e \right)$

- $\sigma_u^2|\beta, \mathbf{u}, \sigma_e^2, \mathbf{y} \sim IG \left( \frac{m}{2} + a_u, \frac{1}{2} \sum_i u_i^2 + b_u \right)$

where $n = \sum_i n_i$, $\mathbf{u} = (u_1, u_2, \ldots, u_m)'$, $\gamma_i = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2 / n_i}$, $\bar{y}_i = \sum_j y_{ij} / n_i$ and $\bar{z}_i = \sum_j z_{ij} / n_i$
Strategies

Assumption: As before, $C$ is block diagonal

Target: $\pi(C, \beta, u, \sigma_u^2, \sigma_e^2 | X^A, X^B, Y_A, Z_B)$

Two possible strategies of simulations:

- Metropolis within Gibbs Sampling (standard)
- Metropolis within Gibbs Sampling with a unique block for $(\beta, u, \sigma_u^2, \sigma_e^2)$ obtaining a more direct approach through integration.
Bayesian estimates

In this framework, two different estimation strategies - corresponding to two different models - can be conceived

1. Feedback strategy: joint posterior distribution of both the parameters of RL part and the SAE together. The posterior distribution of $C$ will depend on $X$ and $Y$ too (and viceversa)

2. Non-feedback strategy: The record linkage part affects the small area part but the converse is false: we perform a Gibbs sampling for $(\beta, \sigma_u, \sigma_e, u)$ for each single $C$ generated by the algorithm and then we retain the last value of the chain.
An application

⇒ Micro-data from the Survey on Household Income and Wealth, Bank of Italy.
⇒ $X$ is the family income, $Y$ is the consumption.
⇒ Small domains are defined as aggregation of the Italian administrative regions, assigning 18 areas.
⇒ Person identifiers coming from the fictitious population census data created for the ESSnet DI (2011).
⇒ In this data set, the linking variables are distorted with missing values and typos.
Data description

⇒ ESSnet DI population size is over 26K records;
⇒ the consumption/income data contains about 13K records; survey data have been expanded in order to reach the ESSnet data size;
⇒ 100 replications of the following experiment: a sample of size 1K randomly selected w/out replacement from the new population.
⇒ consumption (Y) was assigned to the sample; income (X), is stored on the population register;
⇒ On each replication, the sample containing Y was linked to the register reporting X using the key variables Day, Year of Birth and Gender.
Comparisons

We have considered 6 different estimators of the small area totals $\hat{Y}_i$'s

A. EBLUP with $X, Y$ observed on the same data set, (no linkage step) This is our gold standard for comparisons

B. EBLUP restricted on the subset of linked records. Reduction of sample size due to missed links; no linkage errors;

C. naïve EBLUP restricted on the linked records. $X$ and $Y$ observed on different data sets (no linkage error adjustment)

D. Adjusted EBLUP estimator

E. Posterior mean of the $\hat{Y}_i$'s using a Bayesian approach for the linkage step based on the key variables $V_1, \ldots, V_h$ only.

F. Posterior mean of the $\hat{Y}_i$'s using a Bayesian approach with both the key variables and the regression variables $X$ and $Y$. This introduces a feedback effect which makes the estimation of $C$ also depending on $X$ and $Y$. 
Results

**Table:** Estimates based on the entire population (first line) and average values of the estimates of \((\beta_0, \beta_1)\) based on repeated sampling of size 1000

<table>
<thead>
<tr>
<th>Estimates</th>
<th>Intercept</th>
<th>Slope</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population</td>
<td>3.576</td>
<td>0.538</td>
</tr>
<tr>
<td>A</td>
<td>3.057</td>
<td>0.565</td>
</tr>
<tr>
<td>B</td>
<td>3.030</td>
<td>0.567</td>
</tr>
<tr>
<td>C</td>
<td>5.224</td>
<td>0.450</td>
</tr>
<tr>
<td>D</td>
<td>3.008</td>
<td>0.567</td>
</tr>
<tr>
<td>E</td>
<td>3.527</td>
<td>0.541</td>
</tr>
<tr>
<td>F</td>
<td>2.476</td>
<td>0.607</td>
</tr>
</tbody>
</table>
Results

**Table:** Absolute relative bias of the six different estimation strategies for the area totals $\hat{Y}_i$'s

<table>
<thead>
<tr>
<th>Estimates</th>
<th>ARB</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.033</td>
</tr>
<tr>
<td>B</td>
<td>0.033</td>
</tr>
<tr>
<td>C</td>
<td>0.043</td>
</tr>
<tr>
<td>D</td>
<td>0.035</td>
</tr>
<tr>
<td>E</td>
<td>0.033</td>
</tr>
<tr>
<td>F</td>
<td>0.029</td>
</tr>
</tbody>
</table>
Concluding remarks

• The frequentist proposal produces a slight improvement, when the magnitude of linkage errors is relatively low (in the worst scenario, the average in areas and replications is less than 15 %). One can expect a more sensitive improvement with higher linkage error levels.

• On Bayesian side when the assumed model is correct, the variables used in the small area model may help to detect the correct links. If the model is not correct this advantage may turn in a bias, as in our example. Notice however, from Table 2 that the last estimator performs better in terms of absolute relative efficiency: this may be due to a more accurate estimation of the random errors.
Some references


