SESSIONE I
CAMPIONAMENTO E STIMA

Strategie di benchmark per stimatori per piccole aree EBLUP unit level con effetti casuali spazio-temporali

Michele D’Alo’, Lorenzo Di Biagio, Stefano Falorsi, Andrea Faluso e Fabrizio Solari
Benchmark constraints for space and time unit level EBLUP estimators

Michele D’Alo’  Lorenzo Di Biagio  Stefano Falorsi  Andrea Fasulo  Fabrizio Solari
Istat
dalo@istat.it, lorenzo.dibiagio@istat.it, stfalors@istat.it, fasulo@istat.it, solari@istat.it

Abstract
In this paper benchmark strategies for SAE estimates are compared for the case of space and time unit level EBLUP estimation. In particular the procedure proposed in Di Fonzo and Marini (2011) has been developed to obtain coherent estimates for both temporal and contemporaneous constraints.

Finally two empirical studies are presented. The first study is based on synthetic data generation finalized to show the computational performances of the method. The second one considers the Italian Labour Force Survey (LFS) data.

Key words: Time series, indirect estimator, small area estimation, serial autocorrelation

Introduction
Large scale surveys are planned to produce reliable direct estimates of target parameters for national level and as well as analogous parameters related to relevant population sub-sets, e.g. sub-populations corresponding to geographical partitions or sub-populations associated to structural characteristic of the population. When direct estimates cannot be disseminated because of unsatisfactory quality, an ad hoc class of methods, called small area estimation (SAE) methods, is available to overcome the problem. These methods are usually referred as indirect methods since they cope with the lack of information from each domain borrowing strength from the sample information belonging to other domains, resulting in increasing the effective sample size for each small area.

The most important social surveys on households and individuals and business surveys on enterprises, conducted by Italian National Statistical Institute (Istat) and European and non-European institutes are repeated surveys (see Duncan and Kalton, 1985 and Kish 1987). For this reason, in order to deal with the problem of small area estimation within a wide and general survey framework, in this work we refer to repeated large-scale surveys
aimed to produce estimates of cross-sectional parameters for both the generic survey occasion and more survey occasions.

In this paper, unit level linear mixed models with area and time random effects are considered following the general formulation proposed in D’Alò et al. (2014). The authors propose a block-wise matrix notation in order to derive a useful and new formulation of EBLUP unit level estimators borrowing strength from space and/or time; furthermore, for each random effect (area and time) is introduced a general auto-correlation structure that can be exploited in order to capture the variability of the phenomena under study. When large amount of data need to be processed for estimation under unit level models referred to many survey occasions, this method may result computationally more efficient. Within this context different ways for benchmarking SAE estimates, aimed to guarantee consistency of SAE estimates with higher level direct estimates are considered. Furthermore benchmarking is appealing in order to reduce the potential bias of SAE estimates enforcing the robustness of the results. The two-way linear mixed model under a general formulation and the unit level BLUP are formally introduced in section . Section describe the different benchmarking strategies and includes an empirical analysis based on both synthetic data generation and Italian Labor Force survey data aimed to show the empirical and computational performances of the proposed method.

Two-way linear mixed model

Linear mixed models play an important role in SAE context. Random effects are intended to reduce the extra-variability not explained by fixed effects. Let us suppose that the data obey to the following general linear mixed model

\[ y = X\beta + Zu + e. \] (1)

where \( X \) and \( Z \) are known design matrices of full rank and \( u \) and \( e \) are random vectors independently distributed each other with \( u \sim N(0, G) \) and \( e \sim N(0, R) \). While \( e \) are the sampling errors, \( u \) are the random small area effects, that models the area specific random effects accounting for between area variation not explained by the fixed effects. We assume that covariance matrices \( G = G(\omega) \) and \( R = R(\omega) \) possibly depend on the vector of variance components \( \omega = (\omega_1, \cdots, \omega_q )' \). Hence \( y : N(X\beta, \Sigma) \) is a random vector with covariance matrix \( \Sigma = \Sigma(\omega) = Var(y | X, \beta, \omega) \) given by

\[ \Sigma = [R + ZGZ'\rho^2]. \] (2)

Let us consider now unit level models, in which design matrices \( X \) and \( Z \) and vector \( e \) are composed by \( N \) rows and variance terms are defined as \( R(\omega) = \sigma^2W^{-1} \) and \( G(\omega) = \sigma^2\Omega \), being \( \sigma^2 \) an unknown variance component, \( W \) a known diagonal matrix whose generic element is \( w_{ii} \), \( \Omega = \Omega(\phi, \rho) \) a known matrix depending on a set of unknown variance components \( \phi = (\phi_1, \phi_2, \cdots, \phi_i) \) and \( \rho = (\rho_1, \rho_2, \cdots, \rho_i) \), being \( \phi = \sigma^2\sigma \), with \( \sigma = (\sigma_1, \sigma_2, \cdots, \sigma_i) \). Then the explicit expression of (2) under unit level model is

\[ \Sigma(\omega) = \sigma^2[W^{-1} + Z\Omega(\phi, \rho)Z'], \] (3)

where \( \omega = (\sigma^2, \phi, \rho) \) is the variance components vector.
Once observed the sample it is useful to partition the model (1) as

\[
\begin{bmatrix}
  y_s \\
  y_r
\end{bmatrix} = \begin{bmatrix}
  X_s \\
  X_r
\end{bmatrix} \beta + \begin{bmatrix}
  Z_s \\
  Z_r
\end{bmatrix} u + \begin{bmatrix}
  e_s \\
  e_r
\end{bmatrix},
\]

(4)

where the subscripts \( s \) and \( r \) refer to sample and non sample population units respectively. Also the covariance matrix of \( y \) is appropriately partitioned as:

\[
\Sigma = \sigma^2 \begin{bmatrix}
  \Sigma_{ss} & \Sigma_{sr} \\
  \Sigma_{rs} & \Sigma_{rr}
\end{bmatrix} = \sigma^2 \begin{bmatrix}
  W_s^{-1} + Z_s \Omega Z_s' & Z_s \Omega Z_r' \\
  Z_r \Omega Z_s' & W_r^{-1} + Z_r \Omega Z_r'
\end{bmatrix}
\]

Standard small area models generally consider only iid area random effects, while more realistic and efficient models should consider further random effects related to meaningful components, such as iid and/or autoregressive time random effects for repeated surveys. Model (4) is a general theoretical framework allowing to add more iid and/or autocorrelated random effects into small area estimation models. In this work additive area and time random effects are considered, to allow spatial and time autocorrelation structures of random effects. The two-way linear mixed model above considered, defines a general structure allowing to take into account different additive spatial and temporal correlation patterns of random effects. The acronym ST-CC is used to denote the general small area model \( M^{ST-CC} \), under spatially correlated area effects and autocorrelated time effects. The same notation is adapted to denote some well known specific models utilized in small area estimation for repeated and cross-sectional large scale surveys. In particular, models \( M^{ST-II} \), \( M^{ST-IC} \) and \( M^{ST-CL} \) denote some popular two-way small area models, respectively with: iid area effects and iid time effects; iid area effects and correlated time effects; spatially correlated area effects and iid area effects. Furthermore one-way linear mixed models \( M^{T-C} \) and \( M^{T-I} \), respectively with correlated time effects and iid time effects, are useful to perform small area estimation borrowing strength from other times (survey occasions) but not from other areas or domains. While one-way models \( M^{S-C} \) and \( M^{S-I} \), respectively with spatially correlated area effects and iid area effects, allow to perform cross-sectional small area estimation borrowing strength from other domains but not from other times. For example, the last model is used to perform standard cross-sectional small area estimation by means of EBLUP (Empirical best Linear Unbiased Predictor) estimation.

To link the EBLUP estimator to the correspondent linear mixed model, under which is derived, the same notation above introduced is utilized. Then the general unit level BLUP derived under model \( M^{ST-CC} \) will be denoted as \( \hat{y}_{dt}^{ST-CC} \) and the same criterion is adopted for special estimators.

When \( N_d \) is large, \( f_{dt} \equiv 0 \) and \( \bar{x}_{.,dt} \equiv \bar{x}_{.,d} \), then the general formula of unit level EBLUP with space and time correlation, \( \hat{y}_{dt}^{ST-CC} \), is given

\[
\hat{y}_{dt}^{ST-CC} = \bar{x}_{dt} \beta + \sum_d \sum_i \gamma_{d,i} \bar{y}_{.,d,i}
\]

where \( \gamma_{d,i} = \omega_{d,i} \Gamma_{dt} \left( d',i' \right) \), being \( \Gamma_{dt} \left( d',i' \right) = \left( T_{11,dd'} + T_{12,dd'} + T_{21,dd'} + T_{22,dd'} \right) \). where \( T_{11,dd'} \), \( T_{12,dd'} \), \( T_{21,dd'} \) and \( T_{22,dd'} \) are the generic elements of matrices \( T_{11}, T_{12}, T_{21} \) and \( T_{22} \) being \( T = T^*(\omega) \) is a \((D+T) \) dimensional square matrix given by
\[ T^* = \left[ Z_x W_x Z_x + \Omega^{-1} \right]^{-1} = \begin{bmatrix} T_{11}^* & T_{12}^* \\ T_{21}^* & T_{22}^* \end{bmatrix} \] 

being \( \tilde{\epsilon}_{w,dt} = \tilde{\epsilon}_{w,dt} (\cdot) \) given by

\[ \tilde{\epsilon}_{w,dt} = y_{w,dt} - \bar{x}_{w,dt} \hat{\beta}. \] 

where \( \hat{\omega} \) denotes the REML estimates of \( \omega \) and \( w_{dt} = \sum w_{dt} \).

Through particular settings for \( \Gamma_d \) special cases of estimator (5) are obtained.

**Benchmarking**

National institutes of statistics should often use data reconciliation procedures to ensure coherence between time series. In fact high-frequency (e.g., quarterly) time series need to agree with their low-frequency (e.g., annual) counterparts; moreover for every observed period certain contemporaneous constraints must be fulfilled. Typical examples are national accounts and labour force surveys: in both cases quarterly data should respect annual aggregates; moreover, for each quarter, national accounts must respect accounting totals given by financial constraints, while labour force small areas estimates must comply with (more trustworthy) data from bigger areas.

Since low-frequency series are usually more reliable, we can consider high-frequency series as preliminary and try to correct them imposing the temporal constraints given by the low-frequency series and the contemporaneous constraints.

Di Fonzo and Marini, building on previous works by Denton and Cholette, have developed a new simultaneous reconciliation procedure for high-frequency time series. This procedure addresses, at the same time, both temporal and contemporaneous constraints, in the form of linear combinations of variables, producing benchmarked (or reconciled) time series. Di Fonzo – Marini procedure is a multivariate extension of the modified Denton proportional first-difference benchmarking procedure. It is based on the constrained optimization of an objective function. In order to avoid discontinuities between time periods, the function is chosen so that the proportionate difference between the benchmarked and the preliminary series is as constant as possible through time. In this way the temporal profile of the series are preserved at best, at the expense of (maybe) bigger percentage adjustments between each period of the preliminary and the reconciled series.

Given \( M \) preliminary high-frequency time series with length \( n \), \( P_{j,t} \), \((1 \leq j \leq M; 1 \leq t \leq n)\), Di Fonzo – Marini procedure aims at producing the corresponding reconciled time series \( R_{j,t} \) by finding the minimum of the function

\[ F = \sum_{j=1}^{M} \sum_{t=2}^{n} \left( \frac{R_{j,t} - P_{j,t}}{P_{j,t}} - \frac{R_{j,t-1} - P_{j,t-1}}{P_{j,t-1}} \right)^2, \]

given the constraints. Actually, using Lagrange multipliers, the procedure boils down to solve a sparse, symmetric, singular linear system. If the matrix of the system is not too large then linear algebra packages can be applied and a direct solution can be found. Since we will be dealing with short time series, in the following we will also consider a simpler objective function.
$$S = \sum_{j=1}^{M} \sum_{t=1}^{a} \left( \frac{R_{j,t} - P_{j,t}}{P_{j,t}} \right)^2.$$ 

**Empirical results**

Two empirical applications are finally considered in the paper. The first study, based on a synthetic population, has the aim to evaluate the best objective function, and the best routine to solve the linear system. It has also been possible to assess the bias and variance of the small area estimates and the reconciled series using the known totals.

In the second study we consider the quarterly samples from the 2011 Italian Labour Force Survey (LFS) data. Considering the provincial quarterly series for the variables “number of employed”, “number of unemployed” and “number of inactives”, there are two types of constraints that they must satisfy, the contemporaneous and the temporal constraints.

For the contemporaneous constraints for every quarter and for each target variable, the sum of the estimates produced at province level must be equal to direct regional estimates currently produced. At the provincial level, the sum of employed, unemployed and inactives must correspond to the provincial population.

With regard to the temporal constraints, the high-frequency series (i.e., quarterly series) should be in line with the low-frequency component series (i.e. annual series), so for each province and each year the total number of employed (unemployed, inactive) is consistent with the sum of the employed (unemployed, inactive) of the four quarters that make up the year.

The reconciliation procedure takes into account only four quarters at a time, and can be applied separately for each region.

**Simulation results**

The population in the simulation study is based on artificial data, trying to recreate a real situation, with $N = 60,000,000$ individuals scattered among $d = 110$ provinces. $T = 4$ quarters have been considered. The target variable has been created taking into account spatial-temporal correlations between the different areas in the 4 quarters. In order to simulate the labour force characteristics, the target variable was divided into 3 levels. The first level was associated with the 40% of the population (employed) a second level was associated with 5% of the population (unemployed) and the remaining 55% of population was associated with the last third level (not in the labour force). $R = 2,000$ samples have been selected, where each sample was composed of $n = 1,200,000$ individuals. For every sample a design based composite type estimator has been adopted, using an EBLUP based on a unit level linear mixed model estimator.

For the reconciliation procedure, two objective functions have been tested: Di Fonzo–Marini’s function $F$, based on proportional first-differences, and the simpler function $S$. To solve the linear system we have used two different R functions: the function solve in the base package, and the function nlss in the limSolve package. The latter finds only non-negative solutions. In tables 1, 2 we present our results about the benchmarking procedure. The performance of the procedure has been assessed using the following evaluation criteria:
• % Relative Bias \( RB_j = \frac{1}{R} \left[ \sum_{r=1}^{R} \frac{\hat{y}_{jr} - y_j}{y_j} \right] \times 100 \).

• Average Absolute RB: \( AARB = \frac{1}{d} \sum_{j=1}^{d} |RB_j| \).

• Maximum Absolute RB: \( MARB = \max_{j} |RB_j| \).

• % Relative Root Mean Squared Error: \( RRMSE_j = \left( \frac{1}{R} \left[ \sum_{r=1}^{R} \left( \frac{\hat{y}_{jr} - y_j}{y_j} \right)^2 \right] \right)^{\frac{1}{2}} \times 100 \).

• Average RRMSE: \( ARRMSE = \frac{1}{d} \sum_{j=1}^{d} RRMSE_j \).

• Maximum RRMSE: \( MRRMSE = \max_{j} RRMSE_j \).

where \( \hat{y}_{jr} \) is the temporal mean of the EBLUP/benchmarked series of area \( j \) and replicate \( r \), while \( y_j \) is the temporal mean of artificial data for area \( j \).

### Table 1 – AARB indicator (MARB)

<table>
<thead>
<tr>
<th>Variables</th>
<th>EBLUP</th>
<th>baseF</th>
<th>nnlsF</th>
<th>baseS</th>
<th>nnlsS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unemployed</td>
<td>45.5 (647)</td>
<td>5.3 (35.4)</td>
<td>5.8 (34.2)</td>
<td>9.7 (156)</td>
<td>13.9 (183)</td>
</tr>
<tr>
<td>Employed</td>
<td>16.7 (117)</td>
<td>1.9 (10.2)</td>
<td>2.4 (11.9)</td>
<td>2.4 (16.7)</td>
<td>4.5 (29.6)</td>
</tr>
<tr>
<td>Inactive</td>
<td>13.2 (77.4)</td>
<td>51.1 (13.1)</td>
<td>1.4 (18)</td>
<td>1.6 (24.1)</td>
<td>3.4 (60)</td>
</tr>
</tbody>
</table>

### Table 2 – ARRMSE (MRRMSE)

<table>
<thead>
<tr>
<th>Variables</th>
<th>EBLUP</th>
<th>baseF</th>
<th>nnlsF</th>
<th>baseS</th>
<th>nnlsS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unemployed</td>
<td>54.5 (670)</td>
<td>19.2 (45.8)</td>
<td>22.8 (60.2)</td>
<td>21.8 (167)</td>
<td>28 (191)</td>
</tr>
<tr>
<td>Employed</td>
<td>19.1 (118)</td>
<td>6.6 (14.7)</td>
<td>8.7 (25.5)</td>
<td>7 (18.9)</td>
<td>10.3 (33.1)</td>
</tr>
<tr>
<td>Inactive</td>
<td>14.6 (77.9)</td>
<td>3.6 (15.1)</td>
<td>5.1 (19.8)</td>
<td>4.1 (26.4)</td>
<td>6.7 (63.5)</td>
</tr>
</tbody>
</table>

Italian Labour Force Survey (LFS) data results

In this exercise we assess the benchmarking procedure against data from the Italian Labour Force Survey 2011 for people that are from 15 to 24 years old. We have considered all the four quarters, and 107 provinces. The original small-area data have been produced using STEBLUP estimates, and the benchmarking procedure is evaluated by calculating the Mean-Squared Percentage Adjustment (MSPA) for each series and the Mean-Squared Adjustment (MSA) for each series. In table 3 a summary statistics for MSPA and MSA are reported. We recall that, for each preliminary series \( P_{j,t} \) and benchmarked series \( R_{j,t} \)

\[
MSPA(R_{j,t}, P_{j,t}) = 100 \sqrt{\frac{1}{n} \sum_{t=1}^{n} \left( \frac{R_{j,t} - P_{j,t}}{P_{j,t}} \right)^2}
\]

and
\[ MSA(R_j, P_j) = 100 \sqrt{\frac{1}{n-1} \sum_{t=2}^{n} \left( \frac{R_{j,t} - R_{j,t-1}}{R_{j,t-1}} - \frac{P_{j,t} - P_{j,t-1}}{P_{j,t-1}} \right)^2} \]

Table 3 – MSPA and MSA indicators

<table>
<thead>
<tr>
<th>Variables</th>
<th>MSPA</th>
<th>MSA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>baseF</td>
<td>nnlsF</td>
</tr>
<tr>
<td>Median</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>Max</td>
<td>84.8</td>
<td>84.7</td>
</tr>
<tr>
<td>Min</td>
<td>0.3</td>
<td>0.4</td>
</tr>
</tbody>
</table>

The tables 3 reports both MSPA and MSA results only for solution found with solve because the function nnls does not fulfilled all constraints, showing a reconciliation problem.

The two series shows very similar performances in terms of median, maximum and minimum of the MSPA index. With respect to the MSA indicator, the baseF series have an index value lower, but very close to, those for baseS.

Bibliografia

