Small Area Estimation for Official Statistics

Fabrizio Solari

Istat

solari@istat.it

Istat
Istituto Nazionale di Statistica

OUTLINE

1. What is Small Area Estimation?
2. When Using SAE
3. Main SAE Methods
4. SAE and Official Statistics
5. SAE and Repeated Surveys
6. Benchmarking and Reconciliation
7. Case Studies

What is a Small Area?

Small Area (or Small Domain, Local Area, Sub-Domain, Small Subgroup, Minor Domain, etc.):

is a population for which inadequate or even no direct reliable information is available for the variable of interest. For instance:

— in intercensal years, direct population counts are often not available for many domains of interest;

— in the census, population counts are frequently not accurate for certain minority groups;

— a sample survey designed for a large population may select a small number of elements or even no element for the small area of interest. Other nonsampling errors such as nonresponse, late response, etc., may further reduce sample size for a particular small area.

Classification of Domains

— **Planned Domains:** these are domain for which separate samples have been planned, designed and selected.

— **Unplanned Domains:** these cut across the sampling design and sampling units; e.g., age, sex, occupation and education classes widely spread across the sampling units. These domains have not been distinguished in the sample selection and tend to concentrate unevenly in primary units.

Classification of Domains

— **Major Domains:** comprising 1/10 of the population or more. Ex: major regions, 10-year age groups, major categorical classes like occupation.

— **Minor Domains:** comprising between 1/10 and 1/100 of the population. Ex: major municipality and province populations, single year of age, two-fold classifications like occupation X education.

— **Mini Domains:** comprising between 1/100 and 1/10,000 of the population. Ex: small province populations, a three-fold classifications like age X occupation X education.

— **Rare types of individuals:** comprising less than 1/10,000 of the population. Ex: populations with specific health problems classified by health service area.

Classification of Domains

— **Geographic:** province, school district, health service area, metropolitan area, census tract.

— **Demographic:** age-sex-education level group within a geographic area.

— **Geographic and demographic:** poverty status of senior citizens in provinces.

Demand for Small Area Statistics

— Considerable demand for small area statistics both from the public and the private sectors.

— Increasing government concern with the issues of distribution, equity and disparity; underprivileged subgroups (geographical or demographic) may need an upliftment; to implement remedial program reliable data needed for such subgroups.

— Many governments use small area statistics for fund allocation and planning.

— Many private businesses need small area statistics on income, population and environmental data to evaluate markets for new products, and to determine areas for the location, expansion and contraction of their activities.

Why Small Area Estimation?

— Most national surveys are planned to produce accurate estimates at the national and large regional level. Clearly, with a sample of 10,000 persons at the national level most municipalities (and therefore small aggregations of municipalities as local labour market areas) will have zero samples.

— Within regions some considerations may be given to produce accurate subclass totals.

— Detailed survey analysis frequently uses the national data to meet demand for data in small subclasses, which were not planned.

— These subclasses may cut across planned areas or domains or may be small areas (municipalities, local labour market areas, health service areas, etc.) or demographic categories (sex, education level, etc.) within a region.

— Many of these subclasses have very small samples (or even no samples).

— Statistical precision of estimates depends mainly on effective sample size.

Goals in Small Area Estimation

Budget and other constraints usually prevent drawing large samples from each of the small areas. Small areas of interest are often specified only after survey has already been carried out. Clients often ask for more than initially planned.

Problem:

— small area sample sizes are too small to warrant the use of a direct estimator (based only on the area-specific sample data).

Goal:

— how to produce reliable estimates of characteristics of interest (means, counts, etc.) for small areas or domains, based on very small samples taken from these areas;

— how to assess the estimation error.

When and How SAE?

When to use SAE methods:

Whenever direct estimators which are based only on sampling units observed for each small area are not reliable (small sample size or sometimes even no observed units), i.e. coefficient of variations (CVs) or other measures of sampling variability of direct estimates are considered to be too high for the target indicator at area level

How to use SAE methods:

Exploiting relationship among variables and areas by means of explicit or implicit modeling (borrowing strength)

Issues in Small Area Estimation

- definition of small areas;
- identification of relevant source of information;
- sampling design issues;
- method of combining information;
- accuracy of the SAE method;
- robust validation;
- computer software;
- presentation of SAE statistics.

Domain Estimates

Sample surveys are cost-effective means for gathering information for the Total Population as well as for many Subpopulations or Domains.

- **Direct Domain Estimates**
- **Undirect Domain Estimates**

Direct Domain Estimates

A domain estimate is a direct estimate if it is based only on domain-specific sample. It may use auxiliary information related to the interest variable. It is typically design-based but can be motivated/justified by models.

A domain is large if the domain-specific sample is large to yield direct estimates of adequate precision. Otherwise, direct estimates present large variances.

When direct estimates are not reliable?

From Statistics Canada (2010). Guide to the Labour Force Survey SAE methods:

Statistics Canada applies the following guidelines on LFS data reliability:

-if CV ≤ 16.5% then direct estimates are disseminated without restrictions (no release)

-if 16.5% < CV ≤ 33% then the estimates should be accompanied by warning (release with caveat)

-if CV > 33% then the estimates are not recommended for use (no release)

These threshold may be good for labour market variables. What if we are dealing with variables with very low incindency?

Undirect Domain Estimates

If a domain sample cannot support direct estimates of adequate precision, often indirect estimate is used to "borrow strength" by using values from related areas and/or time periods to increase effective sample size.

These values are used in the estimation through a model (implicitly or explicitly) that provides a link to related areas and/or time periods through supplementary information such as recent census counts or current administrative records related to the variable of interest.

Undirect Domain Estimates

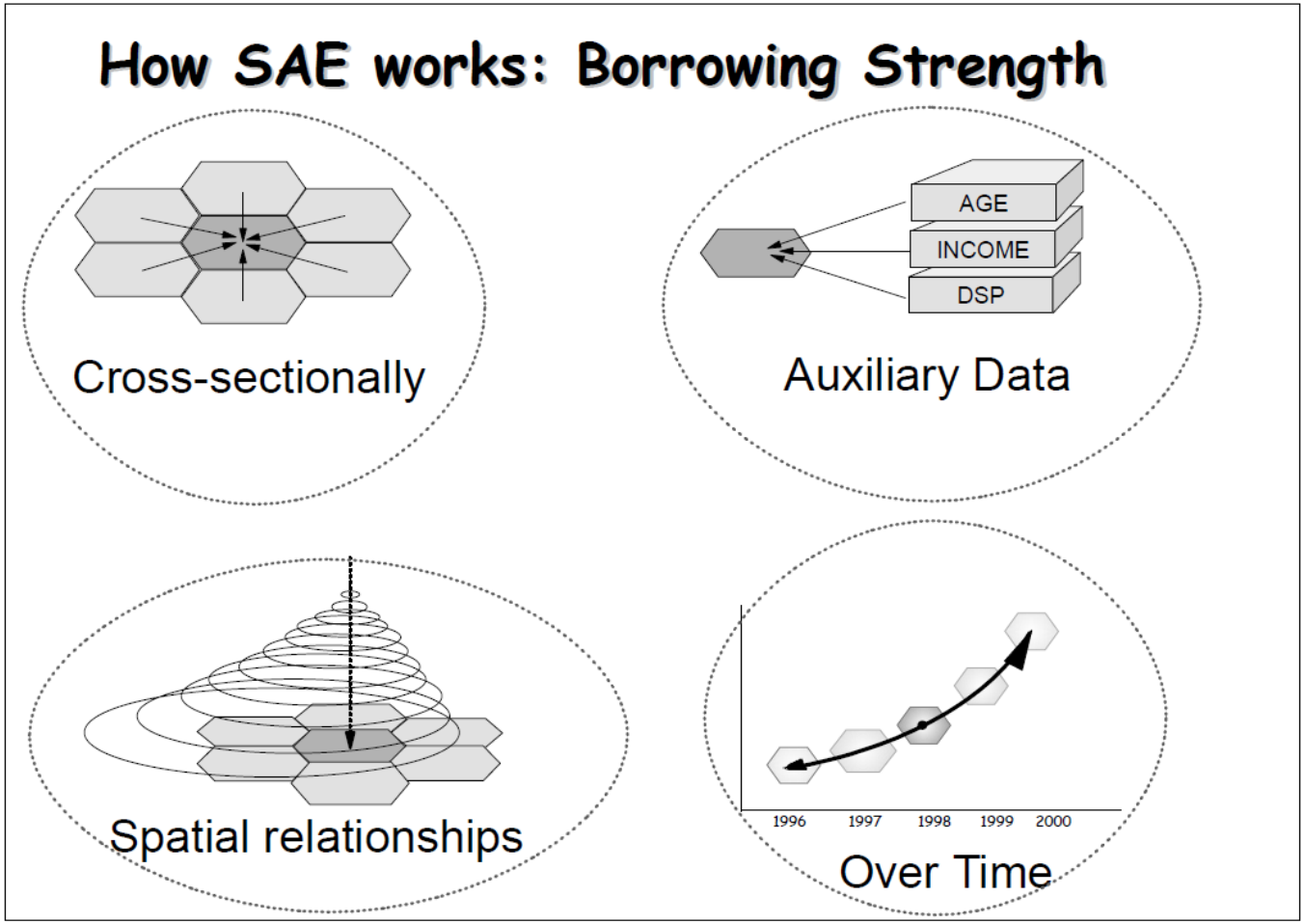
It is possible to divide the indirect estimates into:

— **Domain Indirect:** uses values from another domain but not from another time.

— **Time Indirect:** uses values from another time but not from another domain.

— **Domain and Time Indirect:** uses values both from another domain and time.

SAE: Borrowing Strength from?



Undirect Domain Estimates

We can distinguish the undirect methods into:

— **Synthetic Estimator:** an estimator is called a synthetic estimator if a reliable direct estimator for a large area, covering several small areas, is used to derive an indirect estimator for a small area under the assumption that the small areas have the same characteristics as the large area.

As a consequence, the variance of a synthetic estimator is lower than the corresponding direct estimator but is biased if the model assumptions are not satisfied.

— **Composite Estimator:** it is a linear combination between a direct estimator and a synthetic estimator. For this reason it represents a good compromise in terms of efficiency between the characteristics of the two components.

Undirect Domain Estimates

Further we have:

— **Design Based Approach:** The main concern is unbiasedness. Estimator properties are assessed with respect to the sampling design. This method is traditionally used for small area estimation, mainly because of its simplicity, applicability to general sampling designs and potential of increased accuracy in estimation by borrowing information from similar small areas.

— **Model Based Approach:** The finite population is treated as a random realization from a superpopulation and a suitable model for the interest variable is proposed.

Model Based Approach

Explicit models are used to develop estimates of small area means by ”borrowing strength” from the other small areas. Using the proposed model the predictive distribution of the nonsampling units is obtained.

Predictive distribution, although inherently Bayesian, is also equally accepted by the frequentists. Despite this, model-based approach has a distinct frequentist counterpart since frequentists view the predictive distribution as a conditional distribution.

Quality of the model-based estimates depends on the availability of good auxiliary data (typically from administrative records or other surveys) to develop an adequate model.

To account for between area variation, area-specific random effects are included in a model.

Model Based Approach

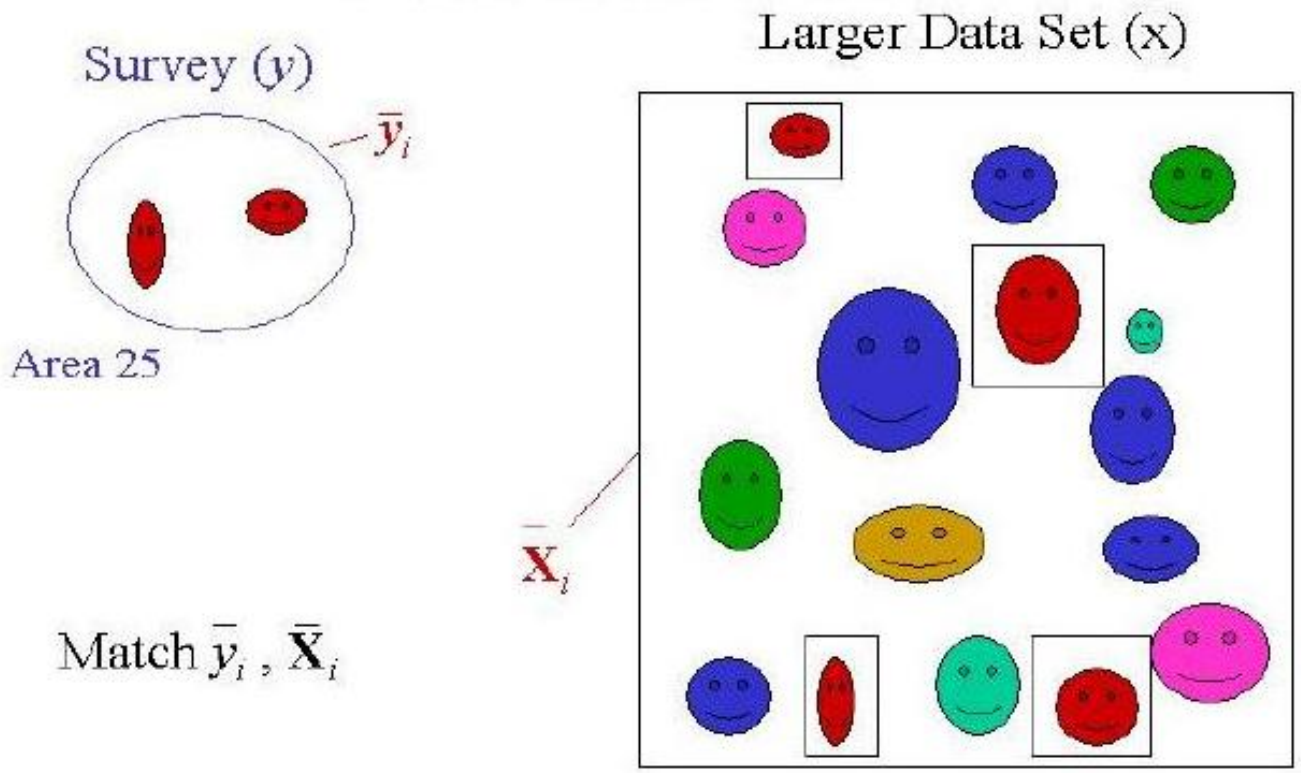
Two types of models depending on the data: area level model and unit level model.

— **Area Level Model:** area level models are appropriate if only area-level summary data available for the auxiliary and/or the response variables. It is possible to take into account the sampling weights into model.

— **Unit Level Model:** unit level models are generally to prefer if unit-specific information is available. They usually ignore the survey weights.

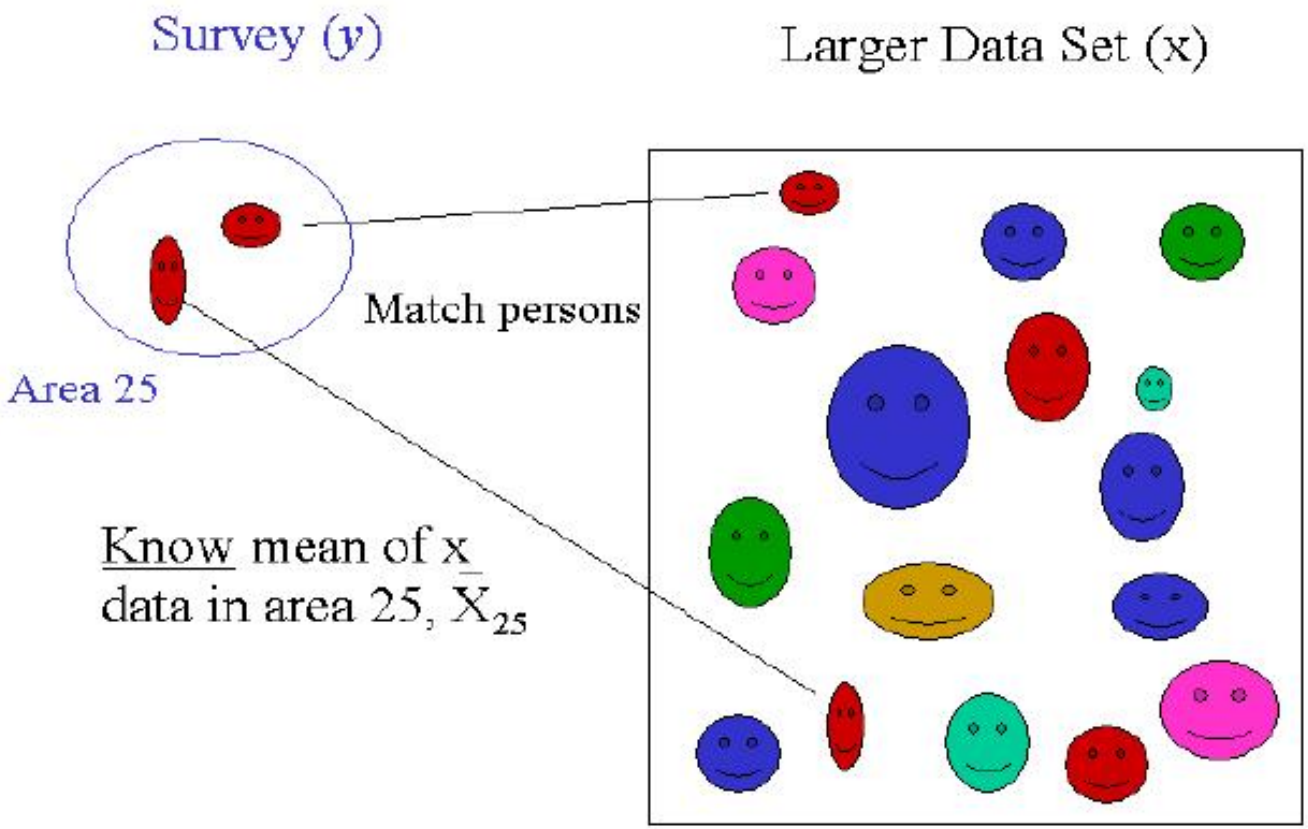
Schematic for Area Level Approach

When only summary data (such as the traditional survey estimator) for the response variable is available at the small area level:



Schematic for Unit Level Approach

When data for both the response variable and auxiliary variables are available at the unit level:



Area Level BLUP

Design based estimators $\hat{\bar{Y}}_d$ of $\bar{Y}_d$ are used in small area modelling.

Fay-Herriot model:

$$\hat{\bar{Y}}_d = \bar{X}_d^T \boldsymbol{\beta} + v_d + e_d,\ d = 1, \ldots, D,$$

v_d, e_d are independent, $v_d \sim N(0, \sigma_v^2),\ e_d \sim N(0, \sigma^2/n_d),\ \sigma^2$ is known.

Area Level BLUP

Fay-Herriot model can be viewed as a hierarchical model:

$$\hat{\bar{Y}}_d = \bar{Y}_d + e_d \qquad \text{(sampling model)}$$

$$\bar{Y}_d = \bar{X}_d^T \boldsymbol{\beta} + v_d \qquad \text{(linking model)}$$

It's a matched linking model; a non-linear function of $\bar{Y}_d$ used in second stage in an unmatched linking model.

Area Level BLUP

The BLUP of $\overline{Y}_d$ is

$$\tilde{\overline{Y}}_d(\sigma_v^2) = \gamma_d \overline{Y}_d + (1-\gamma_d)\overline{x}_d^T \hat{\beta}(\sigma_v^2),$$

that is a linear combination between the direct estimator $\hat{\overline{Y}}_d$ and the synthetic regression estimator $\overline{x}_d^T \hat{\beta}(\sigma_v^2)$, with weight

$$\gamma_d = \sigma_v^2 / (\sigma_v^2 + \sigma^2 / n_d),\ 0 \le \gamma_d \le 1.$$

Unit Level BLUP

Consider the

Nested error regression model:

$$Y_{di} = x_{di}^T \boldsymbol{\beta} + v_d + e_{di},\ d = 1, \ldots, D,\ i = 1, \ldots, N_d$$

v_d, e_{di} are independent $v_d \sim N(0, \sigma_v^2),\ e_{di} \sim N(0, \sigma^2).$

Unit Level BLUP

The resulting BLUP of $\overline{Y}_d$ is given by

$$\tilde{\overline{Y}}_d(\sigma_v^2,\sigma^2) = \gamma_d[\overline{y}_d + (\overline{X}_d - \overline{x}_d)^T \hat{\beta}(\sigma_v^2,\sigma^2)] + (1-\gamma_d)\overline{x}_d^T \hat{\beta}(\sigma_v^2,\sigma^2),$$

which can be viewed as a weighted average between the survey regression estimator $\overline{y}_d + (\overline{X}_d - \overline{x}_d)^T \hat{\beta}(\sigma_v^2,\sigma^2),$ and the synthetic regression estimator $\overline{x}_d{}^T \hat{\beta}(\sigma_v^2,\sigma^2),$ where

$$\gamma_d = \sigma_v^2 / (\sigma_v^2 + \sigma^2 / n_d),\ 0 \le \gamma_d \le 1.$$

Area and Unit Level EBLUP

Fay-Herriot model and nested error regression model are particular cases of the linear mixed model.

Unknown variance components can be estimated by ANOVA method, method of moments, maximum likelihood (ML), restricted maximum likelihood (REML).

The EBLUP is obtained by plugging the variance component estimates in the BLUP.

Empirical Bayes and EBLUP: Under normality and uniform prior for β, the Bayes estimator is also BLUP. Estimating the variance parameter from the marginal distribution of the data we get empirical Bayes estimator of $\overline{Y}_d$. EB and EBLUP are identical under normality

MSE Estimation

Estimation of MSE can be achived by means of:

— **Delta Method**

Prasad and Rao (1990), Datta and Lahiri (2000), Jiang and Rao (2004).

— **Weighted Jackknife Method**

Jiang, Lahiri and Wan (2002), Rao (2003), Chen and Lahiri (2007).

— **Parametric Bootstrap Method**

Butar (1997), Butar and Lahiri (2003), Pfeffermann and Glickman (2004).

Spatial Estimators

Borrowing strength over space

Here, we allow for the spatial autocorrelation of random area effects in the nested error regression model (in the same way for the Fay-Herriot model).

$$y_{id} = x_{id}^T \boldsymbol{\beta} + v_d + e_{id}, \quad \underline{\mathbf{v}} \sim MN(0, \sigma_v^2 \mathbf{A}), \quad \underline{\mathbf{e}} \sim MN(0, \sigma_e^2 \mathbf{I})$$

E.G. the matrix **A** depends on the distances between the areas and on an unknown scale parameter α connected to the spatial correlation between the areas:

$$\mathbf{A} = \{a_{dd'}\} = \left\{ \left[1 + \delta_{dd'} \exp\left(\frac{\mathrm{dist}(d, d')}{\rho} \right) \right]^{-1} \right\}, \ \delta_{dd'} = \begin{cases} 0 & \text{if } d = d' \\ 1 & \text{otherwise} \end{cases}$$

Spatial Estimators

Possible choice for the distance:

- Euclidean distance (take into account the spherical distance if the areas are not closed);
- Street distance;
- Driving time distance.

Spatial Estimators

Other formulations for the spatial correlation:

SAR (Petrucci and Salvati, 2006)

$e_d \sim i.i.d\ N(0,\phi_d)$, ϕ_d known

$$\nu = \rho W \nu + u$$

$$\nu = (I - \rho W)^{-1} u$$

W : proximity matrix between small areas

$u_d \sim i.i.d\ N(0,\sigma^2)$,

σ^2 and ρ unknown

Space and Time and Estimators

Borrowing strength over time

It is possible to add another random component in the nested error regression estimator to take into account extra-variability due to time. The general formulation of unit linear mixed model is:

Mixed models with spatial and temporal random effects:

$$y_{dti} = \mathbf{X}'_{dti}\boldsymbol{\beta}_t + v_{1t} + v_{2d} + e_{dti}$$

$$\mathbf{v}_1 \sim MN(0, \sigma_1^2 \mathbf{A}_1)$$

$$\mathbf{v}_2 \sim MN(0, \sigma_2^2 \mathbf{A}_2)$$

$$\mathbf{e} \sim MN(0, \sigma_e^2 \mathbf{I})$$

$i = 1, \ldots, N_{dt}$, $t = 1, \ldots, T$,; $d = 1, \ldots, D$.

Space and Time and Estimators

— In case of a **first order autoregressive** AR(1) process the matrix $\mathbf{A}_1$ depends on a correlation parameter and on a temporal lag.

— In case of iid time effects $\mathbf{A}_1 = \mathbf{I}_T$.

— In case of **spatial autocorrelation structure** the matrix $\mathbf{A}_2$ depends on the distances between the areas and on an unknown scale parameter α connected to the spatial correlation.

— In case of iid spatial effects $\mathbf{A}_2 = \mathbf{I}_D$.

Small Area and Official Statistics: Italy

From 2001: Project "Territorial statistical information and Sector for Structural Policies" responds to multiple information needs expressed by the Ministry of Finance for the proper use of Community resources.

One of the most innovative activities of the project is the realization of estimates of socio - economic parameters , with application programming purposes economic, to a finer spatial detail of the usual administrative units (provinces and regions).

The territorial dimension of choice is the Local System Labor (SLL), as defined by ISTAT on the basis of daily commuting flows.

Small Area and Official Statistics: USA

Small Area Income and Poverty Estimates (SAIPE) improve upon the American Community Survey (ACS) 1-year survey estimates by incorporating information from administrative records, intercensal population estimates, and decennial census data.

SAIPE estimates are broadly consistent with the direct ACS survey estimates, but with the help from other data sources, SAIPE estimates are more precise than the ACS 1-year and 5- year survey estimates alone for most counties and school districts.

ACS 1-year estimates are not available for most of these smaller geographic areas (approx. only 800 counties with a population of 65,000 or more are included in the ACS 1-year estimates).

European Projects on SAE

2001 – 2004: EURAREA (Enhancing Small Area Estimation Techniques to Meet European Needs)

2008 – 2011: AMELI (Advanced Methodology for European Laeken Indicators)

2008 – 2011: S.A.M.P.L.E. (Small Area Methods for Poverty and Living Condition Estimates)

2009 – 2012: ESSnet SAE (ESSnet on Small Area Estimation)

Experimental framework for comparing the performances of the methods

Use of diagnostics to choose the most appropriate estimation method

- External analysis: it is based on simulation studies based on a large number of samples drawn from the population.
- Internal analysis: the properties of the methods are evaluated by means of indicators computed using only one sample.

External anlysis from real population

- Suppose that we know for all the population the values of the interest and auxiliary variables.

- It is possible to assess the properties of the estimators carrying out Monte Carlo simulation studies, drawing a large number of samples using the real sampling design.

External anlysis from pseudo-population

- When the values of the interest and of the auxiliary variables are not known for all the units in the population, it is possible to create a pseudo-population from which select the samples necessary to carry out the simulation study.

- From one or more real samples we can create the pseudo-population replicating each record a number of times proportional to the sampling weight (bootstrap).

External anlysis from pseudo-population

Problem: intraclass correlation coefficient;

It is not easy to assign to replicated units to

- Strata;
- Higher order units (ex. households to municipalities, individuals to household, etc.)

Area Level Evaluation Criteria

% Relative Bias (RB)

$$\mathrm{RB} = \frac{1}{R}\left[\sum_{r=1}^{R} \frac{\tilde{\theta}_d^r - \theta_d}{\theta_d}\right] 100$$

% Relative Root Mean Squared Error (RRMSE)

$$\mathrm{RRMSE} = \sqrt{\frac{1}{R}\left(\sum_{r=1}^{R}\left[\frac{\tilde{\theta}_d^r - \theta_d}{\theta_d}\right]^2\right)}\, 100$$

Overall Evaluation Criteria

% Average Absolute Relative Bias (AARB): $\frac{1}{D}\sum_{d=1}^{D}\left|RB_d\right|100$

% Average Relative Root Mean Squared Error (ARRMSE): $\frac{1}{D}\sum_{d=1}^{D}RRMSE_d\,100$

% Maximum Absolute Relative Bias (AARB): $\max\left(\left|RB_d\right|\right)100$

% Maximum Relative Root Mean Squared Error (MRRMSE): $\max\left(RRMSE_d\right)100$

Internal anlysis

The performances of the estimation methods can be evaluated by means of one real sample and the known totals of the variable of interest both refferred to the same time (generally the census time).

$$\mathrm{ARE} = \frac{1}{D}\sum_{d}\frac{\left|\hat{\theta}_{d,2001} - \theta_{d,2001}\right|}{\theta_{d,2001}} \qquad \mathrm{ASE} = \frac{1}{D}\sum_{d}\left(\hat{\theta}_{d,2001} - \theta_{d,2001}\right)^2$$

Benchmarking

- The problem occurs when the small area estimates are aggregated at larger domain levels. In this case the corresponding estimate differs from the direct estimate (which is reliable) available for that domain.
- Two possible approaches to obtain small area estimates that matches the direct benchmark:
  - ex-ante: introducing constraints directly in the model specification: Ugarte, Goicoa & Militino (2009), Montanari, Ranalli & Vicarelli (2010), Nandram & Sayit (2011)
  - ex-post: post adjustment : Wang, Fuller & Qu (2008), Bell, Datta & Ghosh (2012), Datta, Ghosh, Steorts & Maples (2011), Bell, Datta & Ghosh (2013)

Data reconciliation

- National institutes of statistics often use data reconciliation procedures to ensure coherence between time series.
- In fact, highfrequency (e.g., quarterly) time series need to agree with their lowfrequency (e.g., annual) counterparts.
- 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.

Data reconciliation and benchmarking

— Simultaneous reconciliation: procedure for high-frequency time series, addressing, at the same time, both temporal and geographical constraints, in the form of linear combinations of variables, producing benchmarked (or reconciled) time series (Di Fonzo and Marini, 2012).

Empirical study – Italian LF Survey

Benchmark Issues:

— Contemporaneous constraints

- For every quarter and for each target variables 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 not in labor force must correspond to the provincial population.

— Temporal constraints

- the high-frequency series (quarterly) should be in line with the low-frequency component series (annual), so for each province and for each year the total number of employed (unemployed, inactive) is consistent with the sum of the employed (unemployed, inactive) of the four quarters.

Choosing the macro-area

- The macro-area is defined as the set of small areas for which a common model is specified (and then fitted).
- It is connected to the statistical concept of model group.
- Often NSIs choose as macro-areas administrative regions (e.g., regions, states, etc,), or macro regions meaningful in the geographical sense (e.g. north, center, south).
- This could be a practical but not an optimal solution.

Choosing the macro-area

- We need to find small areas having the same relationship between the target and the auxiliary variables.
- We need to check the residuals distribution of the target variables with respect to the auxiliary variables are similar
- We can consider as similar, and then to be included in the same macro-area, small areas with similar mean value of residuals, in case we fitted a unit level regression model.
- Alternatively, we can compute the residuals from an area level model and consider in the same macro-area the small areas having similar residuals
- An empirical threshold could be $e_d \pm c\,\sigma_d, c < 1$.
- For each small area an ad hoc macro-area is defined.

Space and TIME modelling

- A unit-level model specification is given in Saei and Chambers (2003). Alternatively, only area-level model specifications are described in the literature. Rao and Yu (1992, 1994) proposed an extension of the basic Fay-Herriot model (Fay and Herriot, 1979) to handle time series and cross-sectional data. Datta et al. (2002) and You (1999) use the Rao-Yu model but replace the AR(1) model specification by a random walk model. Pfeffermann and Burck (1990) propose a general model involving area-by-time specific random effects.

- A natural way to overcome the computational problems connected to large sets of data is to use area level models. Unfortunately, they do not take into account the correlation between the individuals.

Space and TIME modelling

- Unit level linear mixed models with area and time random effects are able to better capture the real variability of the phenomena under study, taking into account, also, the correlation existing for each individual.
- When this type of models are considered, large amount of data need to be processed, and computational problems may occur. For instance, the overall data size processed for the data coming from the Labour Fource Survey (LFS) quarterly samples from 2004 to 2014 is more than 7,000,000 records.
- A natural way to overcome the computational problems connected to large sets of data is to use area level models. Unfortunately, they do not take into account the correlation between the individuals.

Space and TIME modelling

- To overcome the computational problems arising when using unit level LMMs, a computationally more efficient model estimation is proposed (D'Alò, Falorsi & Solari, 2015).
- Matrix algebra is used for reducing dimensional matrix, needed for model estimation, from the number of survey records to the number of areas or the number of survey cycles.
- SAS and R cannot handle the size of individual data coming from LFS survey cycles from 2004 to 2014 (SAS can handle them but it is not possible to introduce spatial and time correlation for random effects).
- The proposed algorithm allows to process in five minutes the 7,0 million size data.

Choosing macro-area for time series

- It is not important that the time series for the small area estimates are similar. It is important that the regression function between the target and the auxiliary variables is similar over the time.
- Similarly to the cross-sectional case we must check if the time series of the residuals are similar.
- Dissimilarity between time series; metric proposed by Piccolo (1990), compresion based methods (Keogh et al., 2007, and Cilibrasi & Vitànji, 2005), correlation based methods (Golay et al., 1998), and, in frequential framework, multi-resolution wavelet based methods (D'Urso and Maharaj, 2012), and methods based and spectral density (Kakizawa et al., 1988).
- Dissimilarity (or similarity) between small area estimates time series are sought for each small area.
- Small area having small level of dissimilarity are included in the same macro-area.
- For each small area is possible an ad hoc macro-area is used.

SAE and Registers

- Traditionally registers have been used to provide populations counts or totals to be utilized in calibration processes or in computing predicted values of the parameters of interest.
- Now registers are becoming more and more important and can provide individual information for many variables.
- Often variables collected in registers can be considered as proxy of the target variables, or can be highly correlated with them.
- Therefore, we can introduce the register variables in the model specification, and, after fitting the model, compute predicted values for all the units in the register (that is all the population units if the registers is not affected by undercoverage).
- This approach is followed for instance by Kim & Rao (2012), who proposed a design based estimator called projection estimator.
- Model based counterparts of the projection estimator are also possible.

Italian LF Survey Description

- Labour Force (LF) survey is a quarterly two stage survey with partial overlap of final sample units according to a rotation scheme of type (2-2-2).
- In each province the municipalities are classified as Self-Representing Areas and the Non Self-Representing Areas.
- From each Self-Representing Areas a sample of households is selected.
- In Non Self-Representing Areas the sample is based on a stratified two stage sampling design. The municipalities are the primary sampling units, while the households are the secondary sampling units.
- For each quarterly sample about 1350 municipalities and 200,000 individuals are involved.

Small Area Estimation on LF Survey

Since 2000, ISTAT disseminates LF yearly estimates of employed and unemployed counts related to the 684 Local Labour Market Areas (LLMAs).

LLMAs are unplanned domains obtained as clusters of municipalities cutting across provinces (NUTS 3), that are the LF survey finest unplanned domains.

The direct estimates are unstable due to very small LLMA sample sizes (more than 100 have zero sample size). SAE methods are necessary.

Until 2003 a design based composite type estimator was adopted.

Starting from 2004, after the redesign of LF survey sampling strategy, a unit level model based estimator with spatially autocorrelated random area effects has been introduced.

Small area estimation on LF Survey

The choice of unit level EBLUP with spatial correlation of area random effects (EBLUP S) arises from an empirical study carried out by means of a Monte Carlo simulation on 500 LF survey samples drawn from the population census, considering different sets of covariates for the estimation of the unemployment rate.

In the study, the design based composite estimator, previously used by ISTAT, has been compared to standard model based estimators studied within the EURAREA project and the enhanced estimator EBLUP S.

EBLUP S results to outperform the other estimators in terms of Average Absolute Relative Bias (AARB) and Average Relative Root MSE (ARRMSE) .

The EBLUP S showed robustness with respect to the different sets of covariates tested in the model, while the other estimators performed differently in the two cases

Small area estimation on LF Survey

Here there are the main results of the simulation study on LFS for the choice of unemployment rate estimates at LLMA level.

The main aspects of the study were:

— 500 two-stage LFS sample have been drawn from 1991 census data according to the LF survey sample design.

— Comparison of standard small estimators, EBLUP S and design based composite estimator using two different sets of auxiliary variables:

- — Case A - *LFS real covariates* – 28 classes of sex by age and unemployment rate at the previous census;
- — Case B – *EURAREA covariates* - sex, age as continuous variable, education level (only census available) and unemployment rate at previous census.

Some results of LF survey empirical study

***Case* A -** *Average Absolute Relative Bias, Average Relative Root MSE, Maximum Absolute Relative Bias and Maximum Relative Root MSE of the estimators*

Estimator	AARB	ARRMSE	MARB	MRRMSE
COMPOSITE	4.53	29.23	46.45	59.40
GREG	12.41	14.66	82.52	82.83
Unit level Synthetic	10.81	11.84	49.45	52.85
Area level Synthetic	10.85	27.64	50.58	61.66
Unit level EBLUP	10.07	11.90	46.64	47.53
Area level EBLUP	10.81	27.63	50.44	61.52
Unit level Spatial EBLUP	9.60	12.33	49.52	49.89

Some results of LF survey empirical study

***Case* B -** *Average Absolute Relative Bias, Average Relative Root MSE, Maximum Absolute Relative Bias and Maximum Relative Root MSE of the estimators*

Estimator	AARB	ARRMSE	MARB	MRRMSE
GREG	6.68	30.68	36.61	53.87
Unit level Synthetic	12.09	13.35	58.52	58.71
Area level Synthetic	10.14	13.27	39.04	39.29
Unit level EBLUP	11.23	13.37	52.22	53.05
Area level EBLUP	9.56	13.28	34.51	36.07
Unit level Spatial EBLUP	9.36	12.47	43.79	45.10

Some results of LF survey empirical study

— the GREG estimator has the best performance in terms of bias;

— the model based estimators are all almost the same when we consider the average over areas RRMSE criterion, but when we consider the maximum RRMSE the area level EBLUP outperforms the others (the area level synthetic estimator is very similar);

— the introduction of the spatial component in the unit level model improves the estimation in terms of bias over the corresponding model without spatial correlation but it performs worse in terms of mean square error likely because of the higher number of parameters to be estimated;

— the unemployment rate seems to be modelled slightly better in case B than in case A with the exception of the unit level spatial model. In this case the number of coefficients to be estimated may be too high and introduce an extra variability in the estimation.

SAE on Consumer Expenditure Survey

A different context for the application of the SAE methods is in the **estimation of poverty rate** at **province level** based on data of Consumer Expenditure survey (CE).

The sampling design of the Consumer Expenditure survey is a two stage stratified sampling design. The primary sampling units are the municipalities stratified according to their demographic size and they are selected with inclusion probability proportional to their size. The secondary sampling units are the households selected by a systematic sampling.

ISTAT disseminates the estimation of poverty rate at national, three macro geographical areas (North, Centre and South of Italy) and, starting from year 2002, regional level (20 regions).

The sample is planned at regional level, hence provinces are unplanned domains (maybe zero sample size).

Empirical Study on CE Survey

Two sets of covariates have been used:

— Case A - The first set of variables consists of the counts of household members in 8 classes of sex by age and it is the standard set of variables used in ISTAT.

— Case B - The second set consists of a classification obtained by clustering the households in homogenous groups with respect to poverty by means of socio-economic variables. The most important are: household size, household employment rate, age of the household head, region, higher education level of the household members. CART (Classification and Regression Trees) has been applied to define the clusters on CES 1997-2001 data. For each year a model has been derived and being the models similar the model derived 2001 data has been chosen. Estimation of the population totals has been obtained using LF data (LF samples are larger than CE samples).

Empirical Study on CE Survey

A simulation study has been carried out drawing 1000 samples from a pseudo-population. The samples have been drawn according to CE sampling design.

The pseudo-population has been generated using 1997-2001 CE data according to the following steps:

— step1: each record of the data sets has been replicated a number of time equal to its sampling weight divided by five;

— step2: the municipality population is created drawing a number of records equal to the municipality population size from the records in step 1 in the same stratum the municipality belongs. (Municipality size given by administrative records on municipalities - year 2001).

Analysis of the results

***Case* A -** *Average Absolute Relative Bias, Average Relative Root MSE, Maximum Absolute Relative Bias and Maximum Relative Root MSE of the estimators*

Estimator	AARB	ARRMSE	MARB	MRRMSE
Direct	1.24	33.22	7.44	126.90
GREG	1.29	33.33	7.58	128.97
SYNTH A	89.29	89.42	498.75	499.00
SYNTH B	37.69	39.85	214.82	216.65
EBLUP A	10.63	28.85	67.42	94.09
EBLUP B	15.48	25.74	151.21	162.70
EBLUP SP	6.92	21.03	46.97	66.68

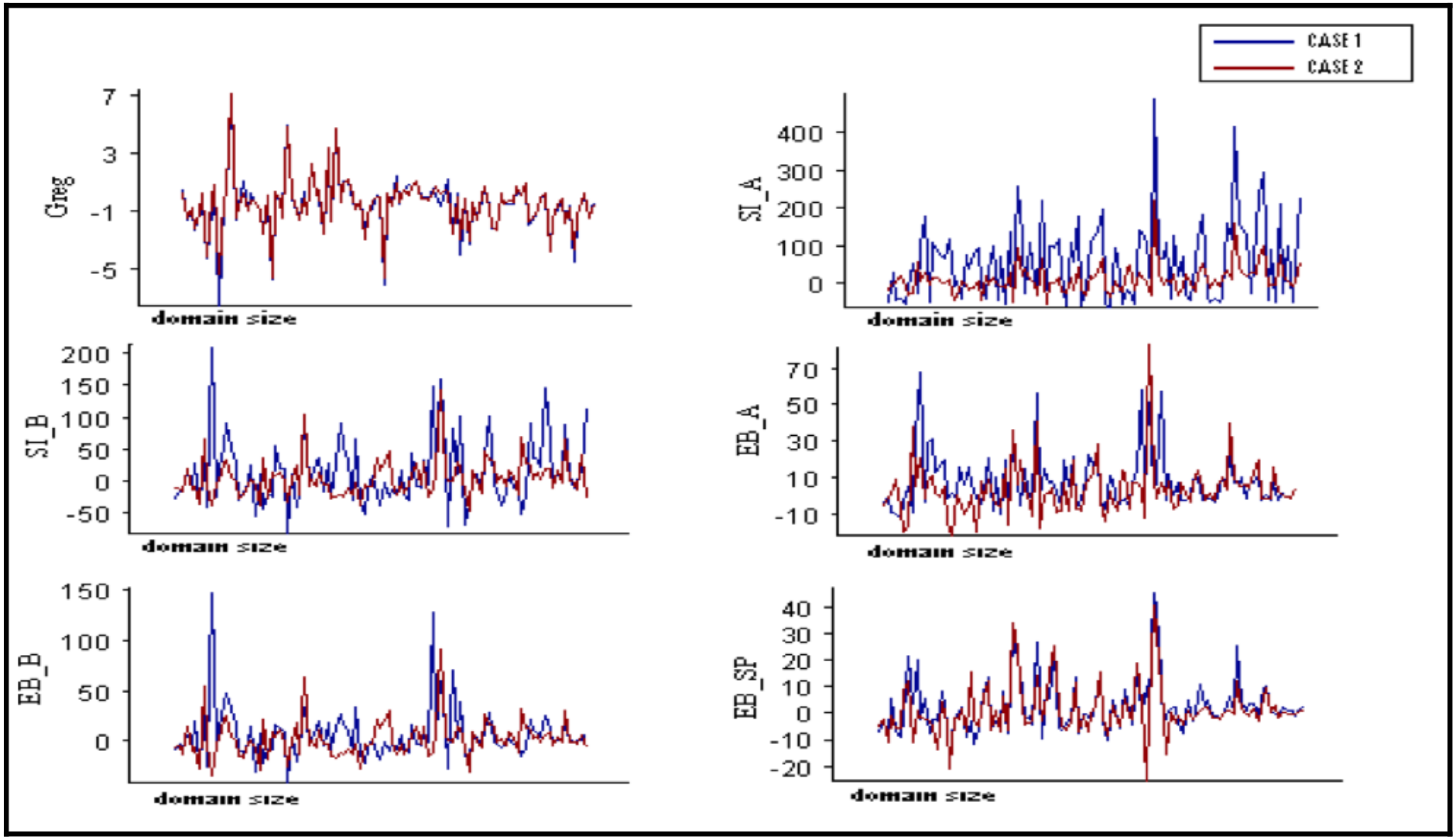
Analysis of the results

***Case* B -** *Average Absolute Relative Bias, Average Relative Root MSE, Maximum Absolute Relative Bias and Maximum Relative Root MSE of the estimators*

Estimator	AARB	ARRMSE	MARB	MRRMSE
Direct	1.24	33.22	7.44	126.90
GREG	1.22	32.65	7.34	124.86
SYNTH A	25.79	26.66	217.55	218.01
SYNTH B	20.14	24.82	144.40	146.99
EBLUP A	9.39	19.71	84.80	96.00
EBLUP B	12.07	20.40	91.06	98.01
EBLUP SP	6.16	24.41	40.34	67.82

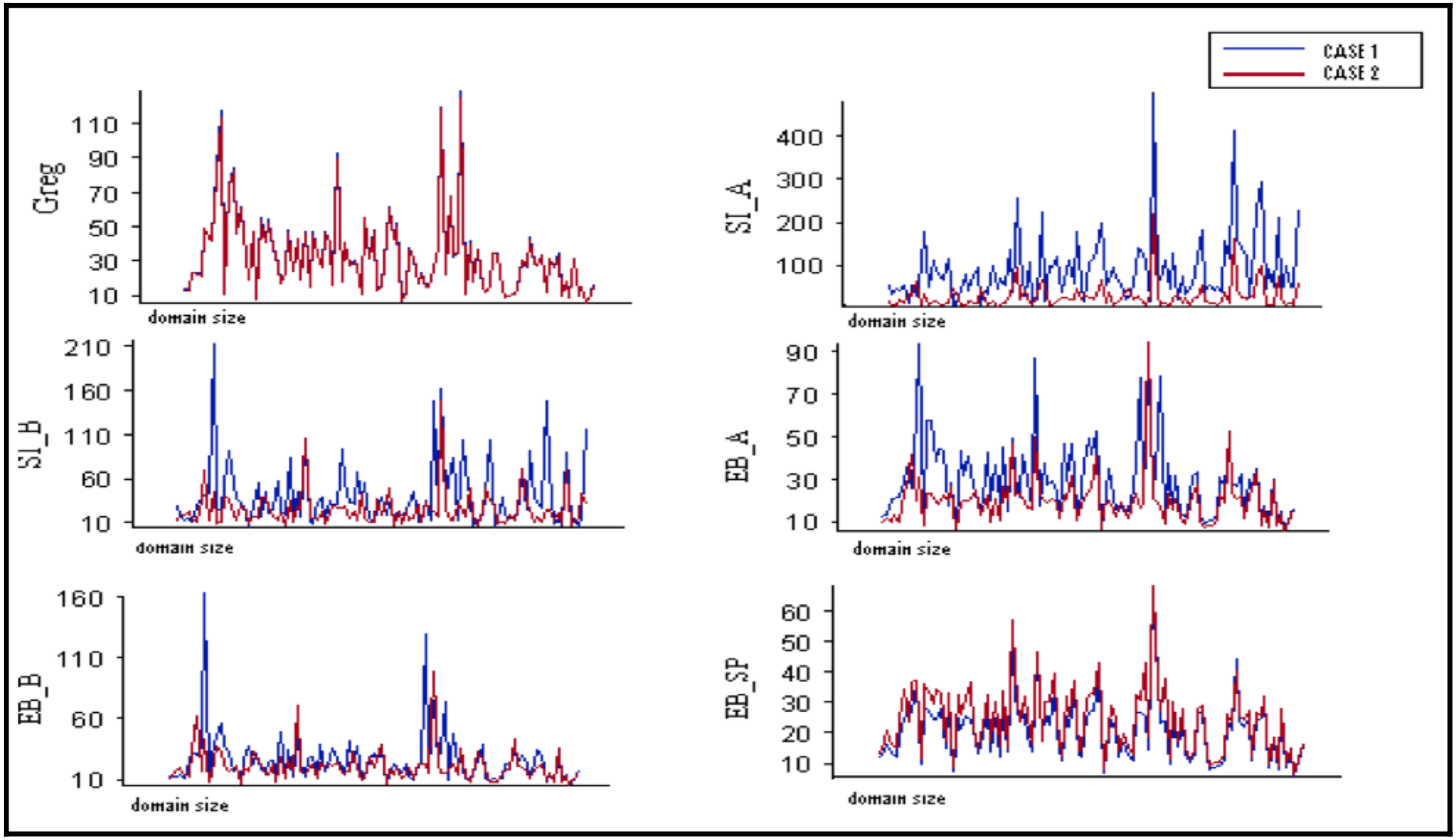
Analysis of the results

Relative Bias of the estimators in Case A and Case B



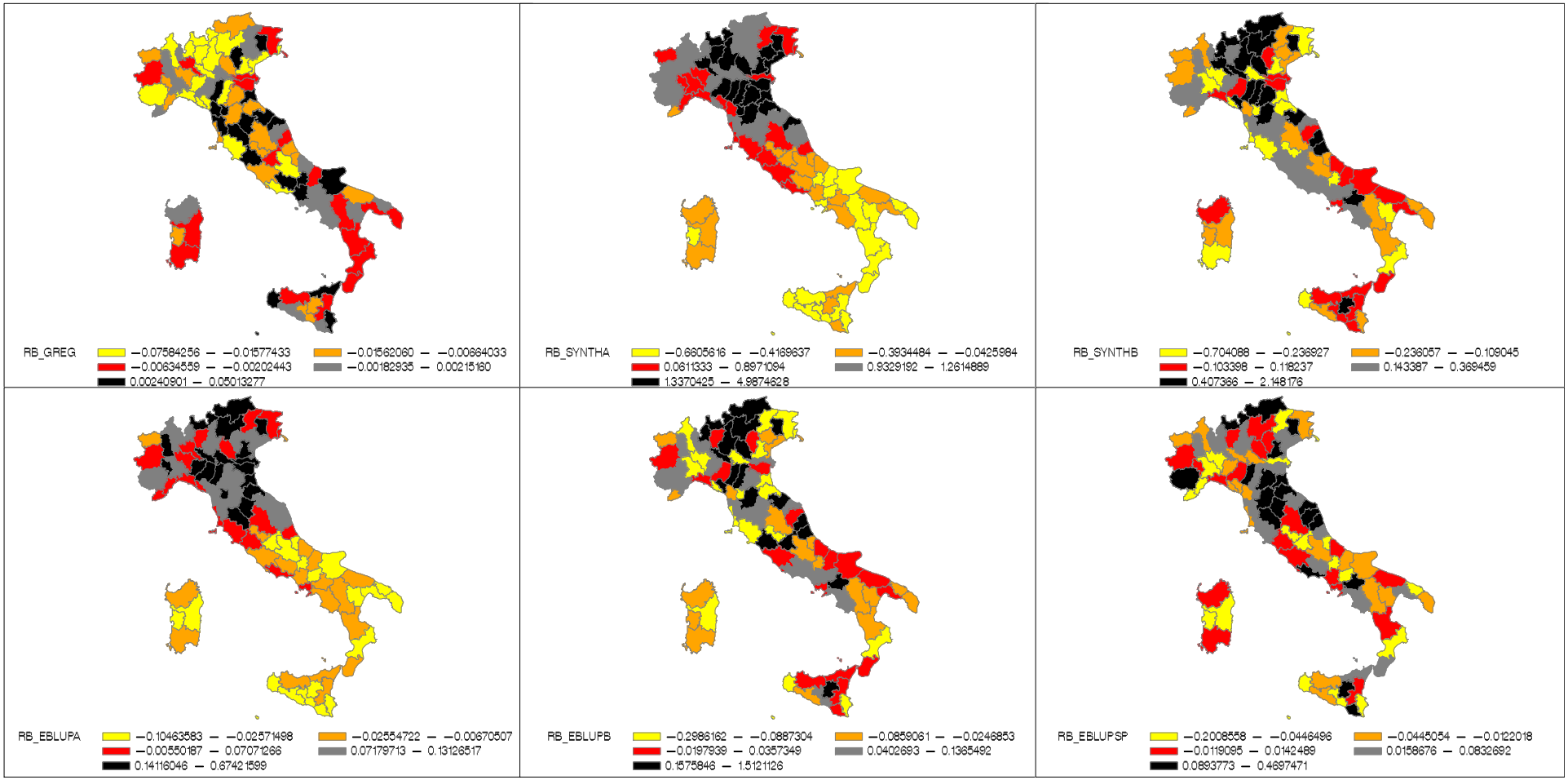
Analysis of the results

Relative Root MSE of the estimators in Case A and Case B



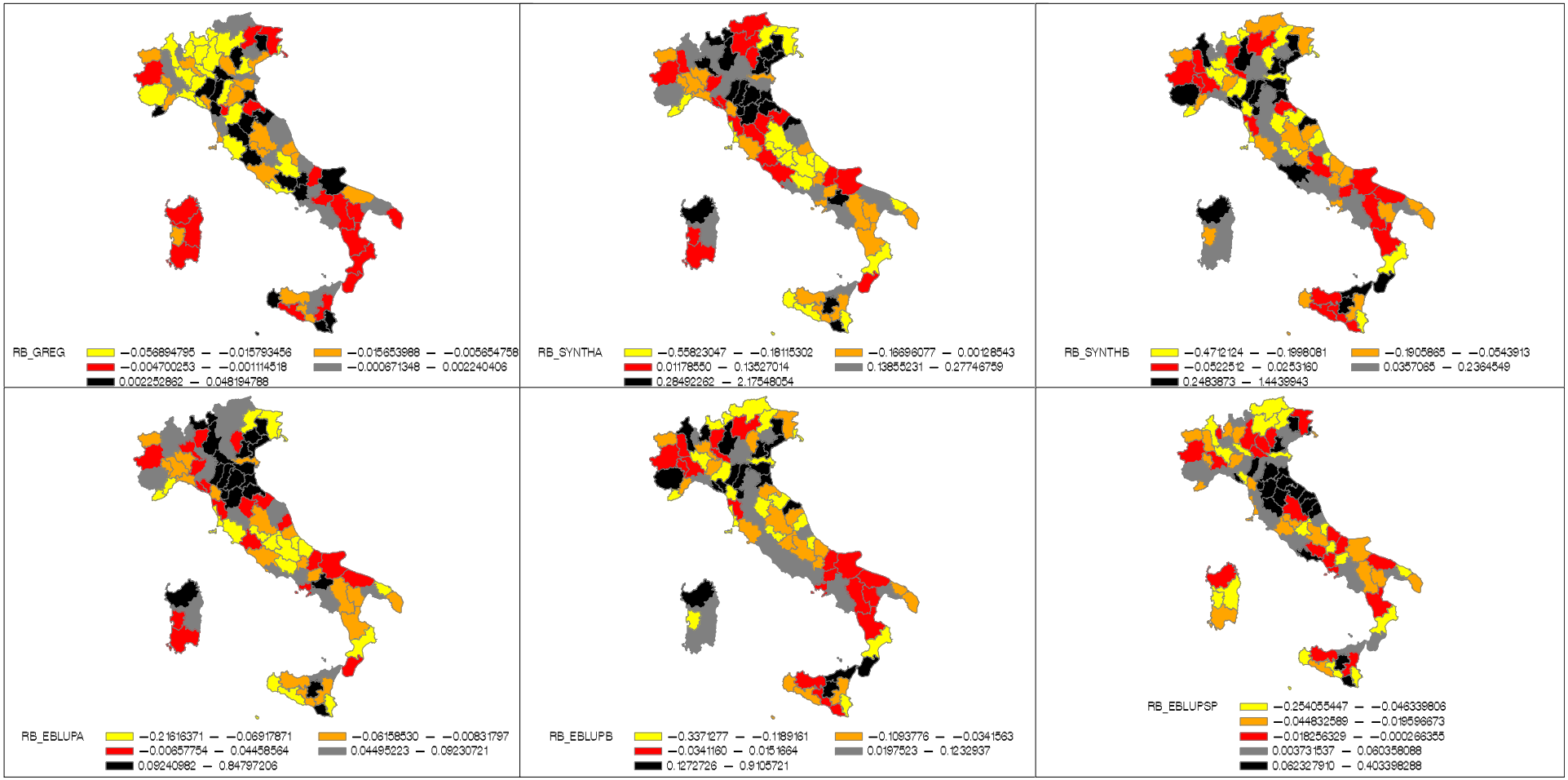
Analysis of the results

Relative Bias geographical distribution in Case A



Analysis of the results

Relative Bias geographical distribution in Case B



Analysis of the results

— The Spatial estimator has the best performance in terms of bias in both cases (apart from the Direct and GREG estimators).

— In terms of MSE the Spatial estimator outperforms the other estimators in Case A.

— In Case B, the unit level EBLUP has the best performance in terms of Average Relative Root MSE, though the spatial estimator displays better in terms of Maximum Relative Root MSE.

— The spatial estimator gives the best results and it shows robustness with respect to the different sets of covariates used in the model.

— The Synthentic estimators performs very badly, showing this set of variables is very poor in explaining poverty.

— The above results on the performance of small area estimators on the poverty rate estimation confirms the results on the application of these methods for the estimation of the Unemployment Rate at Local Labour Market Area level.

Emprical study on LF survey

Description of the empirical study:

— LFS sampling data from 1997 to 2001 (5 years by 4 quarters) have been taken into account to assess the empirical performances of spatio-temporal EBLUP estimators.

— The target parameter is the unemployment rate for each LLMA in each quarter.

— The small areas of interest are the 184 LLMA of the (macro) geographical area Centre (Toscana, Umbria, Lazio, Marche).

— The covariates used to fit the models are: cross-classification of sex and 14 classes of age and unemployment rate at 1991 (previous) population census.

Empirical study on LF survey

Four spatio-temporal estimators have been considered:

- correlated spatial and correlated time random effects (EBLUP_cs_ct);
- correlated spatial and uncorrelated time random effects (EBLUP_cs_ut);
- uncorrelated spatial and correlated time random effects (EBLUP_us_ct);
- uncorrelated spatial and uncorrelated time random effects (EBLUP_us_ut);

The autocorrelated spatio and time random effects are expressed as follows:

$$\mathbf{u}_2 \sim MN(0, \sigma_2^2 \mathbf{A}_2)$$

$$\mathbf{A}_2 = [a_{dd'}] = \left\{ \left[1 + \delta_{dd'} \exp\left(\frac{dist(d, d')}{\rho_2} \right) \right]^{-1} \right\}$$

$$\delta_{dd'} = \begin{cases} 0 & \text{if } d = d' \\ 1 & \text{otherwise} \end{cases}$$

$$\mathbf{u}_1 \sim MN(0, \sigma_1^2 \mathbf{A}_1)$$

$$A_1 \sim \frac{1}{1-\rho_1^2} \begin{bmatrix} 1 & \rho_1 & \cdots & \rho_1^{T-1} \\ \rho_1 & 1 & \cdots & \rho_1^{T-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_1^{T-1} & \rho_1^{T-2} & \cdots & 1 \end{bmatrix}$$

Analysis of Results

Average Relative Error, Average Squared Error

ESTIMATOR	ARE (%)	ASE (*10^5)
GREG	36,43	12,99
SYNTH A (unit level)	19,37	2,80
SYNTH B (area level)	37,79	9,65
EBLUP A (unit level)	19,37	3,24
EBLUP B (area level)	36,21	9,42
EBLUP S (unit level)	20,53	2,91
EBLUP_US_UT	17,23	2,46
EBLUP_CS_UT	17,06	2,43
EBLUP_US_CT	17,23	2,46
EBLUP_CS_CT	17,06	2,43

Analysis of Results

All the spatio-temporal estimators performs better with respect to all the cross-sectional estimators (standard and EBLUP_S estimators).

– As far as the cross-sectional estimators it is concerned, the ARE and ASE results reported in the table agree with the ARRMSE computed in LFS simulation study-

– Because of the estimate of time correlation parameter is very small, the use of a AR (1) time correlation structure seems to have no effect-

– Instead, the introduction of a spatial correlation structure enhance the performances of the correspondent spatio-temporal EBLUP with uncorrelated random area effects.

– There is no substantial difference between ARE and ASE in the analysis of the results (same ranking of methods).

Analysis of Results

The introduction of a random effect component depending on time leads to an improvement of cross-sectional methods.

– Further investigation on the temporal covariance structure are needed, expecially to consider seasonal effects.

– Considering appropriate aggregations of survey times for the choice of the model group.

– Covariance structure in the spatial random effect is a function of the euclidean distances among areas, more suitable for physical than socio-economic phenomena.

– Enhancing spatial covariance structure by means of neighboring matrix.

LF survey empirical study

The simulation study on LFS has been carried out to estimate the unemployment rate at LLMA level.

- ✓ 500 two-stage LFS sample have been drawn from 2001 census data set.
- ✓ The performances of the methods have been evaluated for the estimation of the unemployment rate in the 127 LLMAs belonging to the geographical area "Center of Italy ".
- ✓ GREG, Synthetic, EBLUP small area estimators have been applied considering two different sets of auxiliary variables

  Case A - *LFS real covariates* = sex by 14 age classes + employment indicator at previous census;

  Case B – *LFS real covariates* + geographic coordinates (latitude and longitude of the municipality the sampling unit belongs to).

Enhanced Small area estimators

- Spatial EBLUP: A spatial correlation in the variance matrix of the random effects has been considered (EBLUP SP) + Case A covariates

- Nonparametric EBLUP: two semiparametric representations based on penalized splines have been applied (fitted as additional random effects):
  - ✓ geographical coordinates of the municipality (EBLUP-SPLINE SP): this allows for a finer representation of the spatial component vs EBLUP SP (at municipality level instead of LLMA).
  - ✓ age (EBLUP-SPLINE AGE & EBLUP SP-SPLINE AGE)

Nonparametric EBLUP (Opsomer et al., 2008)

$$y_{id} = \mathbf{x}_{id}^T \boldsymbol{\beta} + f(z_{1id}) + f(z_{2id}, z_{3id}) + u_d + e_{id}$$

$$u_d \sim iid\ N(0, \sigma_u^2), \quad e_{id} \sim iid\ N(0, \sigma_e^2)$$

In the literature there are many nonparametric regression methods (kernel, local polynomial, wavelets…) BUT difficult to incorporate in a Small area model

Methods based on *penalized splines* (Eilers e Marx, 1996; Ruppert et al., 2003) can be estimated by means of mixed models -> promising candidate for SAE methods

- ✓ Great Flexibility in definition of model
- ✓ Estimable with existing software using REML
- ✓ Hard to estimate efficiency and test for terms significance (via bootstrap?)

Results – A: LFS covariates; B = A + geog. coord. mun.

ESTIMATOR	AARB	ARRMSE	MARB	MRRMSE
DIRECT	2.9	51.7	20.4	90.7
GREG A	7.2	40.2	83.3	93.8
GREG B	6.9	40.0	71.5	82.8
SYNTH A	14.0	15.8	93.0	93.5
SYNTH B	12.4	16.4	79.7	81.0
EBLUP A	13.2	16.2	92.5	93.1
EBLUP B	11.9	16.7	79.5	80.7
EBLUP SP	12.7	16.3	90.9	91.6
MBD	8.8	35.3	86.3	92.6
EBLUP-SPLINE SP	12.1	16.5	91.1	92.2
EBLUP-SPLINE AGE	13.2	16.5	89.8	90.5
EBLUP SP-SPLINE AGE	12.2	17.3	90.3	90.9

Analysis of results

- The results of GREG, SYNTH and EBLUB in case B, when geographical information is considered in the fixed term, display better performances in terms of bias.

- In terms of MSE standard estimators in case A outperform standard estimators in case B if the ARRMSE is considered as overall evaluation criteria, while better results are obtained in case B if MRRMSE is considered

- Area level estimators (not shown here) perform a little better in terms of Bias but much worse in terms of MSE.

Analysis of results

- EBLUP SP can be compared with the unit level EBLUP with geographical information included as covariates and the EBLUP-SPLINE SP.
  - EBLUP SP show better performances in terms of MSE, while the unit level EBLUP outperform the other estimators in terms of bias.
  - The EBLUP-SPLINE SP displays performances in between the other estimators.
- EBLUP-SPLINE AGE performs similarly to the unit level EBLUP in Case A
  - The use of the age in a nonparametric way is an alternative use of auxiliary information. With respect to case A the model is more parsimonious.
- As it was expected MBDE shows better results in term of bias and performs poorly in term of MSE than other SAE methods
- The use of autocorrelation structure together with the spline on the variable age doesn't improve the performances

Reference

Battese, G.E., Harter, R.M., and Fuller, W.A. (1988) *An error-components model for prediction of county crops using survey and satellite data*. Journal American Statistical Association 83 28-36.

Fay, R.E. and Herriot, R.A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. J. Amer. Statist. Assoc. 74 269-277.

Ghosh, M. and Rao, J.N.K. (1994), *Small Area Estimation: An Appraisal*, Statistical Science, Vol. 9, No. 1, pp.55-93.

Rao J.N.K, (2003) *Small area estimation*, John Wiley & Sons, Hoboken, New Jersey.