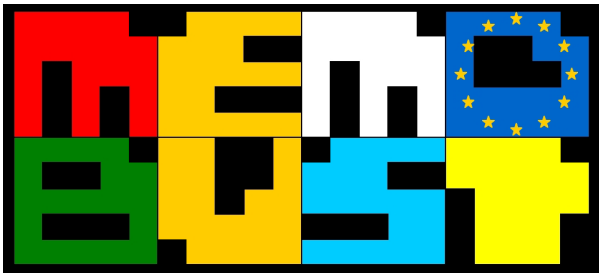

This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Method: EBLUP Unit Level for Small Area Estimation

Contents

General section

1. Summary

The aim of Small Area Estimation (SAE) is to compute a set of reliable estimates for each small area for the target variable(s) of interest, whenever the direct estimates (see "Weighting and Estimation – Main Module" and "Weighting and Estimation – Generalised Regression Estimator") cannot be considered enough reliable, i.e., the correspondent variances (see the module "Quality Aspects – Quality of Statistics") are too high to make those estimates releasable.

Small area methods provide a set of techniques to obtain the estimates of interest in the National Statistical Institutes (NSIs) large scale survey, where more detailed information is required, and the sample size is not large enough to guarantee release of direct estimation. SAE methods which increase the reliability of estimates 'borrowing strength' from a larger area.

The unit level EBLUP estimator, which is described in this module, is a linear combination of the direct information and a regression synthetic prediction of non-sampled units. The fixed part of the model links the target values to some known auxiliary variables, for each units belonging to the larger area to which the small areas of interest belong to. The area specific random effects is instead introduced in order to take into account the correlation among the units with each small area (between area variation).

2. General description of the method

The unit level mixed model can be used when unit-specific auxiliary variables are available in each small area. The area-specific random effect terms are considered in order to take into account the between area variation through the correlation among units within a small area. The basic unit level linear mixed model is the nested error regression model formulated by Battese et al. (1988). It can be expressed as follows:

$$y_{di} = x_{di}^T \boldsymbol{\beta} + u_d + e_{di} \tag{1}$$

where

$$\begin{aligned} &u_d \sim iid\ N\left(0, \sigma_u^2\right) \\ &e_{di} \sim iid\ N\left(0, \sigma_e^2\right) \\ &\forall i = 1, \ldots, N_d \\ &d = 1, \ldots, D \end{aligned}$$

and y_{di} is the variable of interest for the *i*-th population unit in the *d*-th small area. Assuming non informative sampling designs, like simple random sampling, has been used at the sampling stage, the same model assumed for the population values can be applied for the sample units. Therefore, using a matrix notation, the following model can be formalised

$$\mathbf{y}_s = \mathbf{x}_s \boldsymbol{\beta} + \mathbf{z}_s \mathbf{u} + \mathbf{e}_s \tag{2}$$

where $\mathbf{y}_s$ is n-dimensional vector of the observed values for the variable *y*, $\mathbf{x}_s$ is the $(n \times p)$-dimensional matrix of the covariate values observed in the sampling units, $\mathbf{e}_s$ is the *n*-dimensional error vector, $\mathbf{z}_s$ is the $(n \times D)$-dimensional incidence matrix of the sampling units in the small areas, and $\mathbf{u}$ is the D-dimensional vector of area random effects.

In order to obtain the small area estimates based on the above model, either a predictive or a Bayesian approach can be employed (see Rao, 2003, for more details). Following the predictive framework, the Best Linear Unbiased Predictor (BLUP) is obtained by minimising the quadratic loss in the linear unbiased estimator class. The BLUP depends on the variance components and that are usually unknown, so their estimates need to be computed. Both variance components and fixed effects parameters can be estimated in different ways, for example by means of Maximum Likelihood (ML) or Restricted Maximum Likelihood (REML) (Cressie, 1992) methods.

Once the parameters of the model have been estimated, the Empirical Best Linear Unbiased Predictor (EBLUP) based on unit level linear mixed model is a composite-type estimator. Letting aside the finite population correction factor, it is given by

$$\hat{\boldsymbol{\theta}}_d^{\text{EBLUP\_UNIT}} = \boldsymbol{\gamma}_d \left[\bar{y}_d + \left(\overline{\mathbf{X}}_d^T \hat{\boldsymbol{\beta}} - \bar{\mathbf{x}}_d^T \hat{\boldsymbol{\beta}} \right) \right] + (1 - \boldsymbol{\gamma}_d) \overline{\mathbf{X}}_d^T \hat{\boldsymbol{\beta}} \tag{3}$$

where

$$\boldsymbol{\gamma}_d = \frac{\hat{\boldsymbol{\sigma}}_u^2}{\hat{\boldsymbol{\sigma}}_u^2 + \hat{\boldsymbol{\sigma}}_e^2 / n_d}$$

and $\overline{\mathbf{X}}_d$ is the vector of known population means of the auxiliary variables in the *d*-th area and $\bar{\mathbf{x}}_d$ is the correspondent vector of sample means. Given the model, the fixed effects parameter are estimates using the whole available larger area sample information and of course, when the between area variation is small, the EBLUP estimator tends towards the synthetic estimator (being the variance of random effects small). More weight is instead given to direct information when the variance of random effects is big respect the total variance.

There are several extensions of the above described basic unit level model. Since the basic model does not take into account for sample data collected with a complex sample design, some methodological development have been directed to specify more complex models that take into account the features of the sampling design. For instance, Stukel and Rao (1999) proposed a two-fold nested error regression model sample data for data collected from a stratified two-stage sampling.

The issue is that, when an informative design is used the inclusion probabilities of sampling units depend on the values of the target variable the model which holds for the sample data is different from the model assumed for the population data, so that it would be the cause of severe bias into the prediction. A possible approach with this regard is to explicitly include all the design variables used for the sample selection as covariates or the sampling weights in the specification of the model. These two options can be untenable when too many design variables are involved and when the sample weights are not available for non-sampled areas or non-sampled units. A Pseudo EBLUP estimator was proposed by Prasad and Rao (1999) starting from unit linear mixed model.

Moreover, multivariate nested error regression model has been proposed in order to estimate more than one small area parameters of interest simultaneously. This type of model, applied in Datta et al. (1999), allows to take into account the correlation among the characteristics under study observed in the sample units.

Finally, the linear unit level mixed models should be applicable only for continuous observations, then some enhancement models has been considered in order to deal with categorical dependent variables. In that case, Generalised Linear Mixed Models (GLMM) can be considered. Within this logistic regression models with mixed effects are commonly used for estimating small-area proportions (Malec et al., 1997).

3. Preparatory phase

Model selection is crucial preparatory phase since the objective is to lessen the chances of introducing design-bias into the small area estimates due to poor model specification. Model selection for each target variable was carried out considering diagnostic criteria such as maximum likelihood, AIC, BIC, Conditional AIC (cAIC) , and Cross Validation (CV) such as in Vaida and Blanchard (2005), Boonstra et al. (2008), and Boonstra, Buelens and Smeets (2009). Once one or several models has been selected, it is necessary to assess the fitting quality of the model(s). The study of model residuals by graphical representations, like Histograms, Q-Q plots, box-plots and mapping the residual, allows to check if the model assumptions are fulfilled.

4. Examples – not tool specific

We refer to Battese, Harter, and Fuller (1988) for an example of data for application of EBLUP Unit level model. These data are taken from a sample survey that have been designed to estimate crop areas for large regions. The predictions of the crop area for small areas such as counties has generally not been done for the lacking of available data directed collected from these areas. In order to apply the method, satellite data in association with farm-level survey observations has been used. They considered the estimation of mean hectares of corn and soybeans per segment and the auxiliary variables are the number of pixels classified as corn and soybeans in each county. In the example were considered data for 12 Iowa countries and data obtained from land observatory satellites.

Their example relates to application of SAS macros for computing the predictors under the model.

The same data is used as an example in http://www.cros-portal.eu/content/sae for explaining the use of R function **mixed.unit.sae.R**.

5. Examples – tool specific

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

Battese, G. E., Harter, R. M., and Fuller, W. A. (1988), An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association* **80**, 28–36.

BIAS project website: http://www.bias-project.org.uk/

Boonstra, H. J., Buelens, B., and Smeets, M. (2009), Estimation of unemployment for Dutch municipalities. Small Area Estimation 2009 Conference, Elche, Spain, June 29-July 1.

Boonstra, H. J., van den Brakel, J., Buelens, B., Krieg, S., and Smeets, M. (2008), Towards small area estimation at Statistics Netherlands. *Metron* **LIV**, 21–49.

Brown, J., Chambers, R., Heady, P., and Heasman, D. (2003), Evaluation of small area estimation methods: an application to the unemployment estimates from the UK LFS. *Proceedings of Statistics Canada Symposium 2001*.

Chandra, H. and Chambers, R. (2007), Small area estimation for skewed data. Small Area Estimation Conference, Pisa, Italy.

Cressie, N. (1992), REML Estimation in Empirical Bayes Smoothing of Census Undercount. *Survey Methodology* **18**, 75–94.

D'Alò, M., Di Consiglio, L., Falorsi, S., and Solari, F. (2008), The Use of Sample Design Features in Small Area Estimation. ISI 2009 Conference, Durban (South Africa), 16-22 August.

Datta, G. S., Day, B., and Basawa, I. (1999), Empirical best linear unbiased and empirical Bayes prediction in multivariate small area estimation. *Journal of Statistical Planning and Inference* **75**, 269–279.

Datta, G. S., Ghosh, M., Steorts, R., and Maples, J. (2009), Bayesian Benchmarking with Applications to Small Area Estimation property. Small Area Estimation Conference, Elche, Spain.

Dick, J. P. (1995), Modelling net undercoverage in the 1991 Canadian Census. *Survey Methodology* **21**, 44–55.

EURAREA Consortium (2004), *PROJECT REFERENCE VOLUME*, Vol. 1. http://www.ons.gov.uk/ons/guide-method/method-quality/general-methodology/spatial-analysis-and-modelling/eurarea/index.html.

Ghosh, M. and Rao, J. N. K. (1994), Small area estimation: an appraisal. *Statistical Science* **9**, 55–93.

Malec, D., Sedransk, J., Moriarity, C. L., and LeClere, F. B. (1997), *Small area inference for binary variables in National Health Interview Survey*. *Journal of the American Statistical Association* **92**, 815–826.

Molina, I., Saei, A., and Lombardia, M. J. (2007), Small area estimates of labour force participation under a multinomial logit mixed model. *Journal of the Royal Statistical Society, Series A* **170**, 975–1000.

Montanari, G. E., Ranalli, M. G., and Vicarelli, C. (2009), Estimation of small area counts with the benchmarking property. Small Area Estimation Conference, Elche, Spain.

Pfeffermann, D. (2002), Small area estimation – New developments and directions. *International Statistical Review* **70**, 125–143.

Pfeffermann, D. and Tiller, R. (2006), Small-area estimation with state-space models subject to benchmark constraints. *Journal of the American Statistical Association* **101**, 1387–1397.

Prasad, N. G. N. and Rao, J. N. K. (1999), On robust small area estimation using a simple random effects model. *Survey Methodology* **25**, 67–72.

Pushpal, K, Mukhopadhyay, P. K., and McDowell, A. (2011), *Small Area Estimation for Survey Data Analysis Using SAS Software*. SAS Institute Inc., Cary. http://support.sas.com/resources/papers/proceedings11/336-2011.pdf

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

SAE ESSnet (2012), Deliverables of the project. http://www.cros-portal.eu/content/sae

Saei, A. and Chambers, R. (2003), Small Area Estimation under Linear and Generalized Linear Mixed Models with Time and Area Effects. Methodology Working Paper- M03/15, University of Southampton, United Kingdom.

Stukel, D. M. and Rao, J. N. K. (1999), Small area estimation under two-fold nested errors regression models. *Journal of Statistical Planning and Inference* **78**, 131–147.

Torabi, M., Datta, G. S., and Rao, J. N. K. (2009), Empirical Bayes estimation of small area means under a nested error linear regression model with measurement errors in the covariates. *Scandinavian Journal of Statistics* **38**, 598–608.

Vaida, F. and Blanchard, S. (2005), Conditional Akaike information for mixed-effects models. *Biometrika* **92**, 351–370.

Specific section

8. Purpose of the method

The method is used for small area estimation, when direct estimates usually applied for official statistics are too unreliable and unit level auxiliary information are available.

9. Recommended use of the method

1. The method can be applied for estimation when auxiliary information/covariates are available for each sample unit. The mean or total population values need to be known at area level.
2. A linear model can be used when the data are continuous and normally distributed. A transformation of the data may be required before modelling to make the data normally distributed.
3. The method is useful to improve direct estimator if a set of covariates with a strong relationship with the target variable is available.
4. If the target variable is not continuous or normally distributed a generalised linear model might be applied. For instance, the variable of interest at unit level is often binary, so that the logistic or probit model should be more appropriate.
5. Both unit and area level auxiliary information can be considered.

10. Possible disadvantages of the method

1. If the model is not correctly specified the estimator can be affected from severe bias.
2. The basic method do not consider the sampling strategy to select the units.
3. When adding up small domains estimates to a larger domain, it is not ensured that direct estimator at larger level is obtained. A simple way to guarantee this type of consistency is by means of ratio adjustment of the EBLUP unit level estimator. Benchmarking can be also set as a constraint to obtain small area estimates. This would produce different methods that will not be reported in the present handbook (Wang, Fuller, and Qu, 2008; Pfeffermann and Tiller, 2006; Montanari, Ranalli, and Vicarelli, 2009; Datta et al., 2009).
4. The model assumes symmetry of the distribution, while in some cases, like in business survey, skewness may be present. If transformation of variables do not suffice to reduce skewness, advanced method may be considered. For instance by employing M-quantile models (Chandra and Chambers, 2007).
5. Standard small area models generally consider only i.i.d. area random effects, whereas more realistic and efficient models might include further structured random effects, such as time for repeated surveys and spatial autocorrelated random effects.

11. Variants of the method

1. Variants of the method are given by the different estimation methods for the variance component of model (3), e.g., Maximum Likelihood ML or Restricted Maximum Likelihood (REML) (see Cressie, 1992), or Method of moments.

2. On the basis of the assumed model, an estimator which uses only the regression component is given by the unit level synthetic estimator:

$$\hat{\boldsymbol{\theta}}_d^{\text{Synth\_unitlevel}} = \mathbf{X}_d^T \hat{\boldsymbol{\beta}}$$

This estimator is always applied for no sampled domain.

3. For repeated sample surveys, extensions aimed to introduce time random effects can be also considered.

4. In order to consider the spatial autocorrelation among areas a unit level model with spatially correlated area effect can be considered. The spatial correlation can be introduced through the variance-covariance matrix of the random effects in function of the distance between areas or by modelling directly the random effects by means of SAR-type model.

5. Multinomial models are considered in Molina et al. (2007).

12. Input data

Input data set can be classified according to the source of information needed to apply the method. The first data set contains sample information whereas the second one contains information provided from auxiliary source at area level. Specific software tools may need various structure for the input to produce estimation. We refer to links in section 27 for software tools that make possible the application of EBLUP unit level.

1. Ds-input1 = a sample data set contains the target variable and auxiliary variables observed for each sampling unit.

2. Ds-input2 = a data set (macrodata) with mean or total values of covariates for each domain, and population size of the domain.

13. Logical preconditions

1. Missing values
    1. EBLUP unit level estimator does not account explicitly for missing values in the sample observations.
2. Erroneous values
    1. Standard small area methods do not take in consideration errors in the target variables and covariates. Possible misspecification of the auxiliary variables or correction in the variables are not taken into account by EBLUP unit level model (see Torabi et al., 2009).
3. Other quality related preconditions
    1.
4. Other types of preconditions
    1. Normality is often assumed for the estimation of the MSE.
    2. Sampling design is usually not considered in the inference.

14. Tuning parameters

1. Parameters for the convergence of the iterative method: number of iterations and/or stopping rule, starting value for the variance components of the models.

15. Recommended use of the individual variants of the method

1. For non-sampled area only synthetic type estimates can be computed.
2. For estimation of random component of the variance, software tools applies ML or REML.

16. Output data

1. Ds-output1 = the target values estimates for each domain and corresponding MSE.

17. Properties of the output data

1. User may check MSE bias diagnostic (see SAE ESSnet site http://www.cros-portal.eu/content/sae) of the resulting estimates.

18. Unit of input data suitable for the method

Sample unit level information for target variable and covariates to fit the model and to estimates the model parameters included the area random effects. Population area level means or totals for each domain to compute the estimator.

19. User interaction - not tool specific

1. Model selection, the choice of which auxiliary variables to include in the model, e.g., by means of AIC and BIC, cAIC
2. Satisfy the model hypotheses, like symmetry and homogeneity. A transformation of the variable may be needed.
3. Specification of starting value for the variance of the random effects and tuning parameters for convergence
4. Choice of method for variance component estimation
5. The use of the quality method should provide some evidence regarding spatial bias/autocorrelation at different level of aggregation of estimates. Finally MSE for assessing reliability of estimates has to monitored (see guidelines on http://www.cros-portal.eu/content/sae).

20. Logging indicators

1. Run time of the application
2. Number of iteration to attain convergence in the estimation process
3. Out of the range estimation of the target parameter can be attained when linear mixed model is assumed, in this case the normal assumption should be relaxed.

4. Underestimate of MSE can be possible under normality assumption and predictive approach to inference. Generalised linear mixed models and Hierarchical Bayes approach to inference may alternatively be applied.
5. Characteristics of the input data, for instance problems size.

21. Quality indicators of the output data

1. MSE
2. Model Bias diagnostic
3. Benchmark
4. Model selection diagnostic: AIC, BIC, cAIC
5. Analysis of the residual
6. Spatial distribution of area level residual (Maps)

22. Actual use of the method

The method is applied in:

1. Netherlands, for the production of the yearly estimates of unemployment fractions for all Dutch municipalities.
2. Spain, to produce reliable quarterly estimates of consumption expenditures of household and for the survey of the information Society-Families.
3. United Kingdom, to produce 2007/08 middle layer super output area MSOA-level estimates of the proportion of households in poverty for England and Wales, calculated based on equivalised household income after housing costs and produced using the same methodology that was used to produce mean income estimates.
4. Brazil, to generate estimates of poverty and inequality for 5500 Brazilian municipalities.

Interconnections with other modules

23. Themes that refer explicitly to this module

1. Weighting and Estimation – Main Module
2. Weighting and Estimation – Small Area Estimation
3. Quality Aspects – Quality of Statistics

24. Related methods described in other modules

1. Weighting and Estimation – Generalised Regression Estimator
2. Weighting and Estimation – Synthetic Estimators for Small Area Estimation
3. Weighting and Estimation – Composite Estimators for Small Area Estimation
4. Weighting and Estimation – EBLUP Area Level for Small Area Estimation (Fay-Herriot)

5. Weighting and Estimation – Small Area Estimation Methods for Time Series Data

25. Mathematical techniques used by the method described in this module

1. ML or REML by means of Newton-Raphson algorithm

26. GSBPM phases where the method described in this module is used

1. 5.6 Calculate aggregates

27. Tools that implement the method described in this module

1. Eurarea SAS macro and function (http://www.ons.gov.uk/ons/guide-method/method-quality/general-methodology/spatial-analysis-and-modelling/eurarea/index.html)
2. R functions produced by ESSnet SAE (http://www.cros-portal.eu/content/sae)
3. R package sae2 (BIAS project website: http://www.bias-project.org.uk/)
4. SAMPLE project codes in http://www.sample-project.eu/it/the-project/deliverables-docs.html

28. Process step performed by the method

Estimation of parameters in disaggregated domains

Administrative section

29. Module code

Weighting and Estimation-M-EBLUP Unit Level for SAE

30. Version history

Version	Date	Description of changes	Author	Institute
0.1	30-12-2011	first version	Michele D'Alò, Andrea Fasulo	ISTAT
0.2	08-03-2012	second version	Michele D'Alò, Andrea Fasulo	ISTAT
0.2.1	26-03-2012	second version	Michele D'Alò, Andrea Fasulo	ISTAT
0.3	10-09-2013	third version	Michele D'Alò, Andrea Fasulo	ISTAT
0.3.1	12-09-2013	preliminary release		
1.0	26-03-2014	final version within the Memobust project		

31. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	26-3-2014 13:35