This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Method: EBLUP Area Level for Small Area Estimation (Fay-Herriot)

Contents

General section

1. Summary

Small area (or small domain) estimation methods are a set of techniques allowing the estimation of parameters of interest for domains where the direct estimators (e.g., HT or GREG; see the theme module "Weighting and Estimation – Main Module" and the method module "Weighting and Estimation – Generalised Regression Estimator", respectively) cannot be considered reliable enough, i.e., their variance is too high to be released. National Statistical Office surveys are usually planned at a higher level, hence, whenever more detailed information is required, the sample size may be not large enough to guarantee release of direct estimates and in some cases, smaller domains may happen to be without sample units. Small area methods increase the reliability of estimation by "borrowing strength" from a set of areas in a larger domain for which the direct estimator is reliable. This means that information from other areas is used and/or additional information from different sources is exploited (see the theme module "Weighting and Estimation – Small Area Estimation").

The area level EBLUP, which is described in this module, is a linear combination of the area (domain) direct estimator and a predicted component based on a linear mixed model. The model relates the parameter of interest to known auxiliary variables for each of the domains that constitute the partition of the whole population. An effect to account for (within) domain homogeneity is included in the model.

2. General description of the method

The EBLUP area level is a small area estimation method (see the theme module "Weighting and Estimation – Small Area Estimation"). It is based on a linear mixed model which formulates the relationship between the parameter of interest and auxiliary area level information.

Let θ_d be the parameter to be estimated for each domain d. A linear relationship between θ_d and a set of covariates whose values are known for each domain of interest is assumed. In details

$$\theta_d = \mathbf{X}_d^T \boldsymbol{\beta} + u_d , \qquad (1)$$

where $\mathbf{X}_d$ is the vector of covariates for domain d and the u_d s (d=1,...,D) are domain effects assumed to be distributed with mean zero and variance σ_u^2. The random effects account for the extra variability not explained by the auxiliary variables in the model.

Beside the model on the parameters, let us specify the sampling model. A design unbiased direct estimators $\hat{\theta}_d$ is supposed to be available (but not necessarily for all the domains), that is

$$\hat{\theta}_d = \theta_d + e_d , \qquad (2)$$

where the e_d s are the sampling errors associated with the direct estimators, for which $E(e_d \mid \theta_d) = 0$, i.e., the direct estimator is assumed to be unbiased, and $V(e_d \mid \theta_d) = \varphi_d$, where the variances φ_d are supposed to be known.

Combining equations (1) and (2) a linear mixed model is obtained. The model is formulated as follows:

$$\hat{\theta}_d = \mathbf{X}_d^T\boldsymbol{\beta} + u_d + e_d\,. \tag{3}$$

Normality for e and u is commonly assumed for estimation of the Mean Square Error (MSE), but this assumption is not necessary for estimating the parameter. On the basis of model (3) the empirical best linear unbiased estimator (EBLUP) is

$$\hat{\theta}_d^{\text{EBLUP\_AREA}} = \gamma_d\hat{\theta}_d + (1-\gamma_d)\mathbf{X}_{.d}^T\hat{\boldsymbol{\beta}}\ , \tag{4}$$

where

$$\gamma_d = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \varphi_d}$$

is the weight of the direct estimator and $\hat{\boldsymbol{\beta}}$ is the weighted least square (WLS) estimator of the regression coefficient vector $\boldsymbol{\beta}$, where the weights for estimating $\boldsymbol{\beta}$ are provided by a diagonal matrix whose generic element is given by $\hat{\sigma}_u^2 + \varphi_d$. The estimation for the parameters σ_u^2 and $\boldsymbol{\beta}$ has to be obtained recursively. Moreover, as already mentioned above, in order to avoid identifiability problems for the variance components, the sampling variances $V(e_d \mid \theta_d) = \varphi_d$ (d=1, …, D) are assumed to be known.

Nevertheless, if information at unit level is available, then under the hypothesis of homoscedasticity of the sampling errors, the variance φ_d can be estimated from a unit level model (see the method module “Weighting and Estimation – EBLUP Unit Level for Small Area Estimation”) or a generalised variance function (see Wolter, 2007, or Eurostat, 2013, p. 95). Anyway, this would affect the MSE of the predicted domain values (Bell, 2008).

For more details on model specification, methods for estimation of $\hat{\sigma}_u^2$ see Rao (2003, pp. 115-120). Details on the estimation of the MSE are given in Rao (2003, pp. 103 and 128-130).

For the application of the method the user can use several specific software in SAS or R. A review is available in ESSnet SAE Work Package 4 “Software Tools” downloadable from http://www.cros-portal.eu/content/sae.

3. Preparatory phase

4. Examples – not tool specific

We refer to the example reported in Fuller (2009, section 5.5, table 5.13) dealing with the prediction of wind erosion in Iowa for the year 2002. These data are taken from the U.S. National Resources Inventory. The same data have been used in Mukhopadhyay and McDowell (2011) and ESSnet SAE (2012) to display the use of SAS PROC MIXED and the R function mixed.area.sae respectively when area level model is applied for small area estimation. The data report for the 44 Iowa counties the direct estimates of each county of the cube root of the wind erosion measure, the total number of segments (population size), the sample number of segments (sample size). Auxiliary information is given by the erodibility index. There are 44 counties in Iowa, so all the counties are sampled, but for illustrative purposes 4 additional empty counties are created.

For the computation of area level EBLUP sampling errors of direct estimates are needed. Segments are supposed to be drawn by means of simple random sample. Preliminary analysis supported the hypothesis of a common within area variance. Hence sampling errors can be computed as σ_e^2 / n_d, where σ_e^2 is obtained from the data and n_d is the sample size in county *d*.

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

Bell, W. R. (1999), Accounting for uncertainty about variances in small area estimation. *Bulletin of the International Statistical Institute*. http://www.census.gov/did/www/saipe/publications/files/Bell99.pdf

Bell, W. R. (2008), Examining sensitivity of small area inferences to uncertainty about sampling error variances. U.S. Census Bureau, Small Area Income and Poverty Estimates. http://www.census.gov/did/www/saipe/publications/files/Bell2008asa.pdf

Bell, W. R. (2009), The U.S. Census Bureau's small area income and poverty estimates program: a statistical review. http://cio.umh.es/data2/T1A%20William.R.Bell@census.gov.pdf

Buelens, B., van den Brakel, J., Boonstra, H. J., Smeets, M., and Blaess, V. (2012), Case study, report Statistics Netherlands. *Essnet SAE WP5 report*, 62–81.

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.

Datta, G. S., Ghosh, M., Nangia, N., and Natarajan, K. (1996), Estimation of median income of four person families: A Bayesian approach. In: D. A. Berry, K. M. Chaloner, and J. K. Geweke (eds.), *Bayesian Analysis in Statistics and Econometrics*, Wiley, New York, 129–140.

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.

Ericksen, E. P. and Kadane, J. B. (1985), Estimating the Population in Census Year: 1980 and Beyond (with discussion). *Journal of the American Statistical Association* **80**, 927–943.

ESSnet SAE (2012), *WP4 Final Report Deliverables of the project*. http://www.cros-portal.eu/sites/default/files//WP4report_0.pdf

EURAREA Consortium (2004), *PROJECT REFERENCE VOLUME*, Vol. 1.

http://www.ons.gov.uk/ons/guide-method/method-quality/general-methodology/spatial-analysis-andmodelling /eurarea/index.html

Eurostat (2013), *Handbook on precision requirements and variance estimation for ESS household surveys*. Methodologies and Working papers.

Fay, R. E. and Herriot, R. A. (1979), Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association* **74**, 269–277.

Fuller, W. A. (2009), *Sampling Statistics*. John Wiley & Sons, Hoboken, New Jersey.

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.

Moura, E. A. S. and Holt, D. (1999), Small area estimation using multilevel models. *Survey Methodology* **25**, 73–80.

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

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. (1990), The Estimation of the Mean Squared Error of Small Area Estimation. *Journal of the American Statistical Association* **85**, 163–171.

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

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.

Wang, Y., Fuller, W. A., and Qu, Y. (2008), Small area estimation under restriction. *Survey Methodology* **34**, 29–36.

Wolter, K. M. (2007), *Introduction to Variance Estimation*. Springer, London.

Specific section

8. Purpose of the method

The method is used for small area estimation, which is a specific class of methods used for estimation when sampling size in the domain of interest is too small to attain efficient direct estimation. The method increases the reliability of the estimates by introducing a linear relationship between the direct estimates and known area level auxiliary variables.

9. Recommended use of the method

1. The method can be applied for estimation when few or even no sample data are available for one or more domains of interest.
2. The method can be applied on macrodata referred to domain level.
3. The method is useful to improve direct estimators if a set of covariates with a strong relationship with the variable of interest is available.
4. The variances of the small area direct estimates has to be known. Usually a smoothed model for variance estimation is applied and variances are assumed to be known. This affects the MSE (see Bell, 1999).
5. Covariates are needed only at domain level.

10. Possible disadvantages of the method

1. If the model is not correctly specified the estimator can be affected from bias.
2. When adding up small domains estimates to a larger domain, it is not ensured that direct estimates at larger level are obtained. A simple way to ensure consistency is to ratio adjust the EBLUP area 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 et al., 2008; Pfeffermann and Tiller, 2006; Montanari et al., 2009; Datta et al., 2009).
3. Symmetry of the distribution is required while in business survey skewness may be present. If transformation of variables does not suffice to reduce skewness advanced methods may be employed (Chandra and Chambers, 2007).
4. Assumptions of normality with known variance might be untenable at small sample sizes.
5. Model variance σ_u^2 can be estimated to be zero. This is an undesirable result. Hierarchical Bayesian methods are good alternatives and they always result in strictly positive variances, see, e.g., Bell (1999) and Buelens et al. (2012).

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 (or Residual) Maximum Likelihood (REML) (Cressie, 1992), or the method of moments.

2. On the basis of model (3), an estimator making use of only the regression component is given by the area level synthetic estimator:

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

   This estimator uses only the relationship with the covariates and does not exploit the information on the variable of interest in the direct estimator. This estimator can be applied when a domain has no sample data.

12. Input data

Input data sets can be classified according to the source of information needed to apply the method. A first data set contains information calculated on sample data whereas a second one contains information provided from auxiliary sources. Specific software tools may need various structures for the input to produce estimation. We refer to the links in Section 27 below for software tools that make possible the application of the EBLUP area level.

1. Data set input 1 = a data set (macrodata) with direct estimates of the indicators for each domain and their variances.
2. Data set input2 = a data set (macrodata) containing population size and covariates for each domain.

13. Logical preconditions

1. Missing values
   1. Direct estimates in one or more domain can be missing. The EBLUP area level estimator does not account explicitly for missing values in the sample observations.
2. Erroneous values
   1. Standard small area methods do not take into consideration errors in the target variables. Possible misspecification of the area level auxiliary variables or correction in the variables are not taken into account by the EBLUP area level (but 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 variance of the direct estimator has to be known or estimated aside from the area level model.

14. Tuning parameters

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

15. Recommended use of the individual variants of the method

1. Synthetic area level estimator is needed whenever no sample occurs in a specific domain.

2. For the estimation of the random component of the variance, software tools apply ML or REML. The method of moments is more robust with respect to non-normality.

16. Output data

1. Data set output1 = a dataset with predicted (macrodata) values for each domain and possibly MSE.

17. Properties of the output data

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

18. Unit of input data suitable for the method

1. Processing domain level variables for the fitting of the model and the computations of the estimator.
2. Processing unit level data to compute variance estimation of the direct estimator (input for the method).

19. User interaction - not tool specific

1. Select the model, auxiliary variables to be included in the model, e.g., by means of AIC, BIC and cAIC.
2. Determine the aggregate level to which the model is defined, i.e., different models can be assumed for different large domains (aggregation of small domains).
3. Transformation of variable may be needed to satisfy model assumptions (symmetry and homogeneity).
4. Tuning parameters for convergence and specification of starting value for the variance of the random effects.
5. Choice of the method to be used for the estimation of the variance component.
6. After using the method, the quality indicators and logging should be inspected to assess possible presence of bias or inconsistency at different level of aggregation of estimates. Finally MSE for assessing reliability of estimates has to monitored (see guidelines on the ESSnet/sae site http://www.cros-portal.eu/content/sae).

20. Logging indicators

1. Run time of the application.
2. Number of iterations needed to attain convergence in the estimation process.
3. When estimating the variance of the random effects zero or negative values can be attained. This may suggest problems in the variance estimation of the direct estimator. Otherwise hierarchical Bayes to fit model (3) may be applied (Datta et al., 1996).
4. Features of the input data set, e.g., size as it may affect computer time. Anyway problem size does not usually occur with EBLUP area method.

21. Quality indicators of the output data

1. MSE
2. Model Bias diagnostic
3. Benchmarking
4. Model selection diagnostic: AIC, BIC, cAIC

22. Actual use of the method

1. The method is applied by U.S. Census for poverty estimation since 1993, and by Statistics Canada for census undercount estimation.
2. Fay and Herriot (1979)
3. Bell (2009)
4. Dick (1995)

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

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 Unit Level for Small Area Estimation
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. The collection of SAS macros included in the zip file The EURAREA 'Standard' estimators and performance criteria of the EURAREA project (http://www.ons.gov.uk/ons/guide-method/method-quality/general-methodology/spatial-analysis-and-modelling/eurarea/index.html)

2. mixed.area.sae an R function produced by ESSnet SAE (ESSnet/sae site, 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 Area Level for SAE

30. Version history

Version	Date	Description of changes	Author	Institute
0.1	03-06-2011	first version	Loredana Di Consiglio, Fabrizio Solari	ISTAT
0.2	25-11-2011	second version	Loredana Di Consiglio, Fabrizio Solari	ISTAT
0.3	09-01-2012	third version	Loredana Di Consiglio, Fabrizio Solari	ISTAT
0.3.1	17-10-2013		Loredana Di Consiglio, Fabrizio Solari	ISTAT
0.3.2	21-10-2013	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