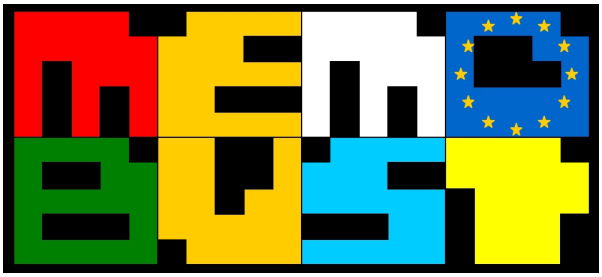

This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Method: Small Area Estimation Methods for Time Series Data

Contents

General section

1. Summary

The aim of small area estimation (SAE) is to produce reliable estimates for each small area for the target variables of interest, whenever the direct estimates cannot be considered enough reliable, i.e., the correspondent variances are too high.

SAE estimators borrow strength from neighbouring areas and auxiliary information deriving from administrative data. Another relevant source of information derives from data measured on previous occasions. In this case specific models can be defined in order to take into account the augmented amount of information with respect to cross-sectional data. Furthermore it is possible to exploit potential correlations between data from the same area on different times. In fact, most repeated survey samples usually include only partial replacement of sample units therefore gain in efficiency can be achieved by borrowing strength from other areas and other time occasions.

Two alternative model specifications are described in the literature. The former is based on linear mixed models in which an additional time depending random effect is added both in unit and area level framework, while the latter refers to state space models specifications.

2. General description of the method

This section describes small area estimation methods using time information. Section 2.1 describes methods based on area level models while section 2.2 illustrates techniques involving unit level model specifications. All the sections are devoted to the description of small area models involving time series data. For the sake of simplicity expressions of predictors are not given but users can easily find them in the references. Once the model parameters are estimated and population means or totals for the auxiliary variables are available, predicted values for each area and time can be straightforwardly computed computing standard expressions.

2.1 Time series area level models

Linear mixed models (LMMs) are one of the more common tools used for model-based small area estimation. LMMs are used for both area and unit models (see the modules "Weighting and Estimation – EBLUP Area Level for Small Area Estimation (Fay-Herriot)" and "Weighting and Estimation – EBLUP Unit Level for Small Area Estimation", respectively). In the first case records refer to units while in case of area level models each record is related to each small area. Basic LMM specifications formulate the relationship between the variable of interest and a set of auxiliary information. Furthermore in order to take into account extra-variability an additional random term is added in the model. In detail a random intercept term is added for each small area. When data for different times are available, this class of models can be straightforward improved introducing an additional random term related to time.

In case of area level models 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. They define a time random component nested in the area random component. In details the following combination of sampling and linking models is proposed:

$$\hat{\theta}_{dt} = \theta_{dt} + e_{dt}, \quad \theta_{dt} = \mathbf{X}_{dt}^T \boldsymbol{\beta} + u_d + v_{dt}, \quad d = 1,\ldots, D, \quad t = 1,\ldots, T, \tag{1}$$

where $\mathbf{X}_{dt}$ is the vector of covariates for the small area *d* at time *t*, the e_{dt}-s are the sampling errors related to the direct estimates $\hat{\theta}_{dt}$, they are uncorrelated over area and time and the variances φ_{dt} are supposed to be known, the u_d-s are domain effects assumed to be distributed with mean zero and common variance, and the v_{dt}-s are time random effects nested into the area effects u_d-s.

Rao and Yu (1992, 1994) suggest a first order autoregressive AR(1) specification to model the time random component v of the model. Hence, the model they propose depends on both area-specific effects and area-by-time specific effects which are correlated across time. More complex ARMA modelling for the time random effect is possible, although it is not clear if such complex modelling will result in efficiency gains (for more details see Rao, 2003). Datta et al. (2002) and You (1999) use the Rao-Yu model but replace the AR(1) model specification by a random walk model.

Alternative model specification to (1) is given in EURAREA Consortium (2004) and Saei and Chambers (2003). More specifically additional independent area and time random effects, and time depending regression coefficients are assumed, that is

$$\hat{\theta}_{dt} = \theta_{dt} + e_{dt}, \quad \theta_{dt} = \mathbf{X}_{dt}^T \boldsymbol{\beta}_t + u_d + v_t, \quad d = 1,\ldots, D, \quad t = 1,\ldots, T. \tag{2}$$

For the time random effects v_t both uncorrelation and first order autoregressive assumptions are made. Furthermore, similarly to model (1), a model with time varying area effects is also specified:

$$\hat{\theta}_{dt} = \theta_{dt} + e_{dt}, \quad \theta_{dt} = \mathbf{X}_{dt}^T \boldsymbol{\beta}_t + v_{dt}, \quad d = 1,\ldots, D, \quad t = 1,\ldots, T. \tag{3}$$

Here the v_{dt}-s are random effects following independent AR(1) processes for $d = 1, 2, \ldots, D$.

The algorithms to obtain the Best Linear Unbiased Estimators (BLUE) of the regression coefficients $\boldsymbol{\beta}$, Best Linear Unbiased Predictor (BLUP) of the small area parameters, and the correspondent Empirical Best Linear Unbiased Predictor (EBLUP), when the variance components are unknown, are described in details in Saei and Chambers (2003). Two estimation methods for variance components are given: Maximum Likelihood (ML) and Residual Maximum Likelihood (REML).

Pfeffermann and Burck (1990) propose a general model involving area-by-time specific random effects. Their model can be written as

$$\hat{\theta}_{dt} = \theta_{dt} + e_{dt}, \quad \theta_{dt} = \mathbf{X}_{dt}^T \boldsymbol{\beta}_{dt}, \quad d = 1,\ldots, D, \quad t = 1,\ldots, T, \tag{4}$$

where the coefficients $\boldsymbol{\beta}_{dt}$ are allowed to vary cross-sectionally and over time. The variation of $\boldsymbol{\beta}_{dt}$ over time is modelled by a state-space model.

2.2 Time series unit level models

The unit level mixed model can be used when unit-specific auxiliary variables are available in each small area. Linear mixed model plays an important role in SAE context. Random effects are intended to reduce the extra-variability not explained by fixed effects. Standard small area models generally consider only i.i.d. area random effects. As reported in section 1 more realistic and efficient models

should take into account additional random effects related to meaningful components, such as time in case of repeated surveys.

Analogously to what described for area level specifications, LMMs are the basic tool to perform small area estimation using time series data. Saei and Chambers (2003) and EURAREA Consortium (2004) adapt the model specifications (2) and (3) to the unit level model framework, obtaining respectively:

$$y_{dti} = x_{dti}^T \boldsymbol{\beta} + u_d + v_t + e_{dti}, \quad d = 1,..., D, \quad t = 1,..., T, \quad i = 1, ..., N_{dt} \tag{5}$$

and

$$y_{dti} = x_{dti}^T \boldsymbol{\beta} + v_{dt} + e_{dti}, \quad d = 1,..., D, \quad t = 1,..., T, \quad i = 1, ..., N_{dt}. \tag{6}$$

As before several assumptions can be made for the random effects v_{dt}. For instance independent area and time random effects can be defined. Saei and Chambers (2003) and EURAREA Consortium (2004) specify models for which the time random effects are modelled according to a first order autoregressive AR(1) process.

3. Preparatory phase

4. Examples – not tool specific

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

Datta, G. S., Lahiri, P., and Maiti, T. (2002), Empirical Bayes estimation of median income of four-person families by state using time series and cross-sectional data. *Journal of Statistical Planning and Inference* **102**, 83–97.

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/downloads/index.html.

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.

Ghosh, M., Nangia, N., and Kim, D. (1996), Estimation of median income of four person families: a Bayesian time series approach. *Journal of the American Statistical Association* **91**, 1423–1431.

Pfefferman, D. and Burck, L. (1990), Robust small area estimation combining time series and cross-sectional data. *Survey Methodology* **16**, 217–237.

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.

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

Rao, J. N. K. and Yu, M. (1992), Small area estimation by combining time series and cross-sectional data. *Proceedings of the Survey Research Section*, American Statistical Association, 1–9.

Rao, J. N. K. and Yu, M. (1994), Small area estimation by combining time series and cross-sectional data. *Canadian Journal of Statistics* **22**, 511–528.

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.

You, Y. (1999), *Hierarchical Bayes and Related Methods for Model-Based Small Area Estimation*. Unpublished Ph.D. dissertation, School of Mathematics and Statistics, Carleton University, Canada.

You, Y., Rao, J. N. K., and Dick, P. (2004), Benchmarking hierarchical Bayes small area estimators in the Canadian Census undercoverage estimation. *Statistics in Transition* **6**, 631–640.

Specific section

8. Purpose of the method

This collection of methods can be used for small area estimation, when time series data are available, i.e., when survey data are collected for several survey occasions. Quality of the estimates is improved by introducing linear relationship between target and auxiliary variables, and explicitly introducing in the models time dependent parameters.

9. Recommended use of the method

1. The methods can be applied when auxiliary information is available either for each sample unit (unit level modelling) or for each small area (area level modelling). Mean or total population values need to be known at area level.
2. Normality assumption is requested. A transformation of the data may be required before applying the methods.
3. The methods can be applied even if no sample data is available for one or more areas.

10. Possible disadvantages of the method

1. If the model is not correctly specified bias can seriously affect small area predicted vales.
2. Sampling strategy is indirectly taken into account only when applying area level models.
3. When summing up small area estimates over a larger domain, benchmarking with direct estimator is not guaranteed. Benchmarking can be obtained with a posteriori adjustment of small area predicted values. Elsewhere benchmarking constrains can be included in the model specification (see Pfeffermann and Tiller, 2006, for the frequentist approach, and You et al., 2002, for the Bayesian framework).

11. Variants of the method

1. Bayesian approach; see for instance Ghosh et al. (1996), You (1999), and Rao (2003, pp. 258-262).

12. Input data

1. Ds-input1 = data set containing sample information for each time. Information could refer to unit level data or area level data.
2. Ds-input2 = data set with population size and covariate mean values or totals for each time and for each area.

13. Logical preconditions

1. Missing values
    1. Sampling information in one or more small area can be missing. The methods previously described do not account for missing values in the sample observations.
2. Erroneous values

1. Errors in the target variables are not taken into account.

3. Other quality related preconditions

    1.

4. Other types of preconditions

    1. When applying area level models, some model specifications require sampling variance of the direct estimator to be known (or estimated outside the area level model).

14. Tuning parameters

1.

15. Recommended use of the individual variants of the method

1.

16. Output data

1. Data set output1 = a dataset with predicted small area values for each domain and for each time, error evaluation.
2. Data set output2 = model parameter estimates.

17. Properties of the output data

1. Users should check MSE of the resulting estimates and model bias diagnostics.

18. Unit of input data suitable for the method

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

19. User interaction - not tool specific

1. Selection of the model, auxiliary variables to be included in the model, e.g., by means of AIC, BIC in the frequentist framework, or DIC in the Bayesian context.
2. Transformation of variable may be needed to satisfy model assumptions (symmetry and homogeneity).
3. Tuning parameters for convergence and specification of the starting values for the model parameters in the frequentist approach, choice of the starting values for the parameters in the model and the number of chains in case of Bayesian modelling.

20. Logging indicators

1. Number of iterations needed to attain convergence in the estimation process.
2. Diagnostics criteria to evaluate convergence of MCMC and evaluation of mixing in case of multiple chains.

21. Quality indicators of the output data

1. MSE or Posterior variance.
2. Model Bias diagnostics.
3. Benchmarking.
4. Model selection diagnostic: AIC, BIC, or DIC.

22. Actual use of the method

1.

Interconnections with other modules

23. Themes that refer explicitly to this module

1. Weighting and Estimation – Small Area Estimation

24. Related methods described in other modules

1. Weighting and Estimation – EBLUP Area Level for Small Area Estimation (Fay-Herriot)
2. Weighting and Estimation – EBLUP Unit Level for Small Area Estimation

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

1. ML or REML by means of Newton-Raphson or scoring algorithms.
2. MCMC algorithms.

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).

28. Process step performed by the method

Prediction of totals or mean values for disaggregated domains.

Administrative section

29. Module code

Weighting and Estimation-M-SAE Time Series Data

30. Version history

Version	Date	Description of changes	Author	Institute
0.1	06-04-2012	first version	Michele D'Alò, Fabrizio Solari	ISTAT
0.2	10-05-2012	second version	Michele D'Alò, Fabrizio Solari	ISTAT
0.3	14-11-2013	third version	Michele D'Alò, Fabrizio Solari	ISTAT
0.3.1	18-11-2013	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 13:36