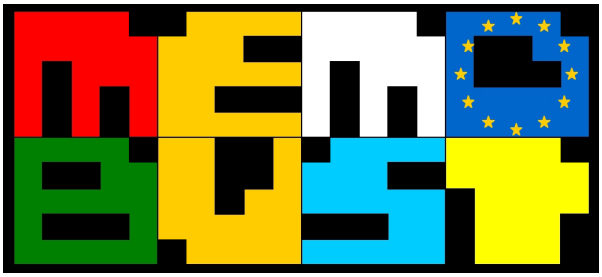

This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Method: Composite Estimators for Small Area Estimation

Contents

General section

1. Summary

In surveys conducted by statistical offices one of the main problems is to have reliable estimates for domains for which the sample size is too small or even equal to zero. It is the consequence of the fact that many institutions need more detailed information not only for the whole country but also for some specific subdomains such as geographic areas or other cross-sections. It also concerns business statistics where increasing demand exists for information for different classification of activities (e.g., trade, manufacturing, transport, construction, etc.) including small, medium and large enterprises and many variables (e.g., revenue, operating costs, taxes, etc.). In such situations direct estimates based only on specific domain sample data are insufficient because of high variability and small precision. The remedy could be the methodology of small area estimation (SAE) which plays an important role in the field of modern information provision, which aims to cut survey costs while lowering the respondent burden.

Thanks to their properties, SAE methods enable reliable estimation at lower level of spatial aggregation and with more specific domains, where direct estimation techniques display too much variance. Another advantage over direct estimators is that small area estimation can be used to handle cases with few or no observations for a given domain in the sample. Therefore it is necessary in many situations to use indirect estimates that borrow strength by taking into account values of the variables of interest from related areas and from that point of view increasing the "effective" sample size.

Generally speaking there are basically two types of indirect estimators: the synthetic and the composite estimators which can be derived under a design-based approach or taking into account the fact that an explicit area level or unit level model exists. In this part of the handbook only design-based composite estimators are described. For details on model-based composite estimators see Rao (2003) or the modules mentioned in section24 below. The main aim of this module is to provide a set of principles for composite estimators. Information about the first group of estimators can be found in the module "Weighting and Estimation – Synthetic Estimators for Small Area Estimation".

2. General description of the method

Composite estimators provide a broad class of indirect estimators and are used in situations when the direct estimator is not taken into account because of its large variance and the synthetic estimators giveunacceptable results because of bias. Composite estimators can be seen as estimators which give a compromise between the large variance of direct estimators and the bias of synthetic estimators and from that point of view they are built for balancing the properties of the direct and the synthetic estimator. When the sample size is quite large the direct estimator is valuable. On the other hand when the sample size is small or even equal to zero synthetic estimators are more valuable. From that point of view a composite estimator can be considered as an estimator that usually takes into account a direct and an indirect estimate and is better in the sense of having smaller bias and variance.

One common type of the composite estimator is a weighted average of two estimators – direct ($\hat{Y}_{dir,d}$) and synthetic ($\hat{Y}_{synth,d}$). Generally speaking, this class of estimators is a very easy solution to the problem of large bias of synthetic estimators and large variance of direct estimators. Composite estimators can be defined as follows:

$$\hat{Y}_{com,d} = \gamma_d \hat{Y}_{dir,d} + (1 - \gamma_d)\hat{Y}_{synth,d} \quad (1)$$

where γ_d is a weight from the interval [0,1] in the small area d. The above expression is a convex combination of the direct and synthetic estimators and, in general, the choice of a proper weight $\gamma_d$depends on the size of the sample in the small area d. If the sample size in the small area is large enough, then the direct estimator should receive a bigger weight. Otherwise if the sample size gets smaller than the synthetic part receives a bigger weight.

Finding the right value of the weight γ_d constitutes the main problem in the use of composite estimators. This is very important from the point of view of balancing the potential bias of the synthetic estimator against the instability of the direct estimator. The way of selecting this weight is very controversial. One of the most common solution is to take $\gamma_d = n_d/N_d$, where n_d is the sample area size for domain d and N_d is the population area size for domain d. Alternatively γ_d can be obtained by minimising the mean square error (MSE) of the composite estimator, see Rao (2003). In this second approach the weights can be obtained by minimising the MSE of the composite estimator $\hat{Y}_{com,d}$, with respect to γ_d, under the assumption that the covariance between direct and synthetic estimator is small compared to the MSE of $\hat{Y}_{synth,d}$. In this approach it can be shown that the optimal weight is given by the formula:

$$\gamma_d = \frac{MSE(\hat{Y}_{synth,d})}{MSE(\hat{Y}_{dir,d}) + MSE(\hat{Y}_{synth,d})}. \quad (2)$$

Some other ways of finding γ_d are discussed in Ghosh and Rao (1994), Holmoy and Thomsen (1998) and Singh, Gambino and Mantel (1993). Here, our attention will be focused only on the so-called sample size dependent estimator(SSD) which is a special case of the composite estimator with weights γ_d which depend on the domain counts $\widehat{N}_d$ and N_d where $\widehat{N}_d$ is the sum of all design weightsin domain d, i.e.,$\widehat{N}_d = \sum_{i=1}^{n_d} d_i$, and $N_d$is the population size in domain d. In Drew, Singh and Choudhry (1982) the proposition for γ_d is as follows:

$$\gamma_d = \begin{cases} 1 & if\ \widehat{N}_d \geq \alpha N_d, \\ \frac{\widehat{N}_d}{\alpha N_d} & otherwise, \end{cases} \quad (3)$$

where α is subjectively chosen parameter. Generally speaking when the sample size in domain d increases, $\gamma_d$is close to 1 and the composite estimator $\hat{Y}_{com,d}$ is very similar to direct estimator. Otherwise the synthetic estimator has a bigger contribution.

Another proposition can be found in Särndal and Hidiroglou (1989):

$$\gamma_d = \begin{cases} 1 & if\ \widehat{N}_d \geq N_d, \\ \left(\frac{\widehat{N}_d}{N_d}\right)^{h-1} & otherwise, \end{cases} \quad (4)$$

where h is subjectively chosen.When $\alpha = 1$ and $h = 2$, the weight γ_d is the same in the first and the second approach.

A discussion devoted to different types of composite estimators derived under design-based approach can also be found in Rao (2003).

Estimation of the MSE of the composite estimators, even when a weight γ_d is fixed, runs into difficulties similar to those for synthetic estimators. For details, see the module on synthetic estimators

and Rao (2003) where a broad discussion devoted to the problem of MSE estimation of composite estimators can be found.

3. Preparatory phase

4. Examples – not tool specific

In the literature one can find many examples of composite estimators both in real surveys and simulation studies. Eklund (1998) used composite estimators to estimate the net coverage error for the 1997 U.S. Census of Agriculture at the state level. Falorsi, Falorsi and Russo (1994) used the composite estimator of the number of unemployed in Health Service Areas of the Friuli region in Italy. The method was also applied in the Labour Force Survey by Griffiths (1996). An example of the use of the sample size dependent estimator can be found in Farver (2002) where this estimator was used in the estimation of food-animal productivity parameters. A broad discussion devoted to examples of applications of composite estimators can also be found in Rao (2003).

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

Costa, A., Sattora, A., and Ventura, E. (2009), Using composite estimators to improve both domain and total area estimation.http://www.econ.upf.edu/docs/papers/downloads/731.pdf

Drew, D., Singh, M. P., and Choudhry, G. H. (1982), Evaluation of small area estimation techniques for the Canadian Labour Force Survey. *Survey Methodology***8**, 17–47.

Essnet Project on Small Area Estimation (2012a), *Report on Workpackage 3 – Quality Assessment*. Final Version, March 2012.

Essnet Project on Small Area Estimation (2012b), *Report on Workpackage 6 – Guidelines*. Final Version, March 2012.

Eklund, B. (1998), Small area estimation of coverage error for the 1997 Census of Agriculture. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 335–338.

Falorsi, P. D., Falorsi, S., and Russo, A. (1994), Empirical Comparison of Small Area Estimation Methods for the Italian Labour Force Survey. *Survey Methodology***20**, 171–176.

Farver, T. B. (2002), Comparison of ratio-synthetic, sample-size dependent and EBLUP estimators as estimators of food-animal productivity parameters. *Preventive Veterinary Medicine***52**, 313–332.

Ghosh, M. andRao, J.N.K. (1994), Small Area Estimation: An Appraisal.*Statistical Science***9**,55–76.

Griffiths, R. (1996), Current Population Survey Small Area Estimations for Congressional Districts.*Proceedings of the Section on Survey Research Method*, American Statistical Association, 314–319.

Holmoy, A.M. K. and Thomsen, I. (1998), Combining Data from Surveys and Administrative Record Systems. The Norwegian Experience.*International Statistical Review***66**, 201–221.

Longford, N.T. (2005), *Missing Data and Small-Area Estimation. Modern Analytical Equipment for the Survey Statistician*. Springer.

MEETS (2011), *Use of Administrative Data for Business Statistics*. Final Report, Poznań (Poland).

Molina, I. andMarhuenda, Y. (2013), *Package SAE*. http://cran.r-project.org/web/packages/sae/sae.pdf

Opsomer, J.D., Botts, C. and Kim, J.Y. (2003), Small area estimation in a watershed erosion assessment survey.*Journal of Agricultural, Biological, and Environmental Statistics***8**, 139–152.

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

Särndal, C.-E. andHidiroglou, M.A. (1989), Small domain estimation: a conditional analysis.*Journal of the American Statistical Association***84**, 266–275.

Singh, M.P., Gambino, J.G., and Mantel, H. (1993), Issues and options in the provision of small area statistics.In:G. Kalton, J. Kordos, and R. Platek (eds.),*Proceedings of the International Scientific Conference on Small Area Statistics and Survey Designs*, vol. 1, Central Statistical Office, Warsaw, 37–75.

Ugarte, M.D., Goicoa, T., Militino, A.F., and Sagaseta-Lopez, M. (2009), Estimating unemployment in very small areas.*SORT***33**, 49–70.

Specific section

8. Purpose of the method

The method is used for small area estimation and involves some variants of combining two estimators into one by taking a weighted average of these estimators. Even though many small area estimators, both design- and model-based, have the basic form of a linear weighted combination of two estimators, the most common approach is to take the direct and synthetic estimator in the formula for the composite estimator. The aim of this intervention is to balance the potential bias of a synthetic estimator and the high variance of a direct one.

9. Recommended use of the method

1. This estimator can be useful in domains in which a direct estimator has a large variance.
2. This estimator can be useful in surveys when analysed domains vary very much in terms of sample size. To avoid the inconvenience related to switching from a direct estimator to a synthetic one, the composite approach can be used, balancing the influence of the used estimators.
3. Because of the simplicity of composite estimators they should be recommended in all surveys when methods of small area estimation are used. They are easy to implement and not difficult to understand by the users. With direct and synthetic estimators they form the so-called triplet of small area estimates and can always be produced using existing data, see Essnet Project on Small Area Estimation (2012b).

10. Possible disadvantagesof the method

1. How to establish the value of the weight γ_d is a matter of discussion.
2. Another problem is how to provide measures of error for a given small area – for example, for bias. It should be mentioned that the bias, even if smaller than for synthetic estimators, is also present for composite estimators.
3. Composite estimators are sometimes called shrinkage estimators because of the fact that all the direct estimates are pulled towards the corresponding synthetic estimate of a broader area. As a consequence composite estimators generally display less between-area variation then they should. In the literature this inconvenience is known as the over-shrinkage problem. For details, see Essnet Project on Small Area Estimation (2012b).
4. For some composite estimators, the estimates $\hat{Y}_d$ for small areas do not add up to the direct large area estimate $\hat{Y}$. In such cases adjustment is needed in order to ensure coherence of estimates at different levels. Potential solution is to use following formula:

$$\hat{Y}_{d,adj} = \frac{\hat{Y}_d}{\sum_d \hat{Y}_d}\hat{Y}. \quad (5)$$

11. Variants of the method

1. Variants of the method depend on which estimators are taken into account in the formula of the composite estimator. In the basic approach, the composite estimator is a weighted average

of a direct and a synthetic estimator. However the expression of composite estimators can be considered as a convex combination of two different estimators than a direct and a synthetic estimator. In the literature devoted to small area estimation many estimators, both design and model-based, have the composite form. Rao (2003) provides many composite estimators including the sample size dependent estimator and the James-Stein method and many examples of their applications.

2. Variants of the method depend also on the way how the weight γ_d is established.

12. Input data

1. The input data set depends on which estimators are taken into account in the formula for composite estimators and the source of information. The input data set can contain individual information for all units in the sample. In this situation the direct and synthetic estimator can be calculated and, as a consequence, the composite estimator is directly established as a weighted sum of these two estimators. The input data set can also contain information coming from auxiliary sources. Specific software may be based on different structures of the input data set in the procedure of estimation using the composite approach.

13. Logical preconditions

1. Missing values
    1. When an area contains no data in the sample, synthetic estimators may be used. In this situation the composite estimator reduces to the synthetic one, i.e., $\gamma_d = 0$.
2. Erroneous values
    1. Standard small area methods do not take into consideration errors in auxiliary variables. A possible misspecification of the area level variables or correction in the variables is not taken into account.
3. Other quality related preconditions
    1.
4. Other types of preconditions
    1.

14. Tuning parameters

1. Because of the fact that a composite estimator consists of a direct and a synthetic estimator, parameters for the convergence of the iterative method may be the same as for the model-based synthetic estimator: the maximum number of iterations, and the convergence criterion. One of the tuning parameter could also be the weight γ_d.

15. Recommended use of the individual variants of the method

1. In some situations where small areas vary strongly in terms of sample size a direct estimator can be good for areas with the largest sample sizes. On the other hand, a direct estimator is very poor when the representation in the sample is very small or equal to zero. In this case a

synthetic estimator may be more effective. Switching from one estimator to the other is inconvenient. The problem can be solved by using composite estimation, which balances inconveniences of these two estimators, see Longford (2005).

2. Because of the fact that composite estimators are easy to implement compared to explicit model-based estimators, they are recommended to use as basic smoothing approach in all surveys when small area estimation methods are taken into account.

3. When the composite weights depend only on the sub-sample sizes, it is possible to derive composite estimates for a large number of target variables at the same time. For comparison at the same time a model applies only to one or very few variables so it is impractical to build models for all variables in the sample. It is usually impractical to build models for all the statistical variables that are collected in the sample, neither at the national level nor at the small-area level, see Essnet Project on Small Area Estimation (2012b). Summing up, composite estimators (especially SSD) are useful when dealing with many variables comparisonwith fitting appropriate models for different variables.

4. Some recommendations devoted to how establish some parameters in composite estimators can be found in the literature. For example, it is recommended, with regard to sample size dependent estimators, that in formula (4) h should be equal to 2, see Särndal and Hidiroglou (1989). For the weight γ_d in formula (3) it is recommended that $\alpha = 1$. However in the Canadian Labour Force Survey α is equal to 2/3.

16. Output data

1. An output dataset usually contains a table with estimates for all small areas. The following measures may also be included in an output data set: MSE, variance, confidence intervals or bias especially in simulation studies when the true value of parameters are known and it is very easy to calculate them.

17. Properties of the output data

1. The user should check the quality of estimates based on their knowledge of the investigated phenomenon and MSE, variance, bias of estimates or confidence intervals if possible, see Essnet Project on Small Area Estimation (2012a).

18. Unit of input data suitable for the method

For the purpose of computations using composite estimators both unit level data and domain level variables can be used.

19. User interaction - not tool specific

1. Select estimators as components of the composite estimator.
2. Establish the weight γ_d as a weighting factor in the formula for the composite estimator.
3. Choose auxiliary variables to be included into the synthetic part of the composite estimator.
4. Establish the level of aggregation.
5. Establish tuning parameters (convergence criteria, starting point, stopping point) if necessary.

6. After the use of the composite estimator quality indicators, if possible, should be checked and verified in order to evaluate the final results (MSE, confidence interval).

20. Logging indicators

1. The logging indicators generally speaking depend on the two estimators taken into account in the formula for the composite estimators and may cover: run time of the application, number of iterations to reach convergence in the estimation process, characteristics of the input data, see also the item "logging indicators" in the module "Weighting and Estimation – Synthetic Estimators for Small Area Estimation".

21. Quality indicators of the output data

1. Compare with quality indicators of the output data for synthetic estimators mentioned in the module "Weighting and Estimation – Synthetic Estimators for Small Area Estimation".

22. Actual use of the method

1. Applications of composite estimators can be found in different areas of statistics. Composite estimators are in use in environmental statistics in a survey conducted in the Rathbun Lake Watershed in Iowa, see Opsomer, Botts and Kim (2003). Other examples of using composite estimators can be found in Costa, Sattora, Ventura(2009). In their article, which was based on a cooperation between The Institute of Statistics of Catalonia(IDESCAT) and the UniversitatPompeuFabra, composite estimators and their application to several areas of interest are described. Sample size dependent estimators are in use in surveys devoted to the labour market. For example, The Canadian Labour Force Survey, uses a sample size dependent estimator to produce Census Division level estimates. Another application of sample size dependent estimators in labour market statistics can be found in Ugarte et al.(2009). Some actual applications of composite estimators in business surveys can be found in documentation of the MEETS project, where composite estimators were implemented to estimate some characteristics (revenue, number of employees, wages) according to short-term and annual statistics of medium-sized and large enterprises. For details, see MEETS (2011). See and compare it with the information devoted to the actual use of the method in the module "Weighting and Estimation – Synthetic Estimators for Small Area Estimation".

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 – Synthetic Estimators for Small Area Estimation
2. Weighting and Estimation – EBLUP Area Level for Small Area Estimation (Fay-Herriot)
3. Weighting and Estimation – EBLUP Unit Level for Small Area Estimation
4. Weighting and Estimation – Small Area Estimation Methods for Time Series Data

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

1. Basic knowledge of linear algebra is needed. When composite estimators are built using the model-based approach the knowledge ofiterative methods is required.

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

1. 5.6 Calculate weights
2. 5.7 Calculate aggregates

27. Tools that implement the method described in this module

1. In many cases own codes are required to implement the above mentioned composite estimators. However there are some functions in R which help to obtain composite estimates. For example, in the SAE package written by Isabel Molina and Yolanda Marhuenda one can find the ssd function which calculates sample size dependent estimators as a composition of direct and synthetic estimators. For details, see Molina and Marhuenda (2013).

28. Process step performed by the method

Estimation of parameters in disaggregated domains.

Administrative section

29. Module code

Weighting and Estimation-M-Composite Estimators for SAE

30. Version history

Version	Date	Description of changes	Author	Institute
0.1	10-02-2012	first version	Marcin Szymkowiak, Tomasz Józefowski	GUS (Poland)
0.2	14-01-2013	second version	Marcin Szymkowiak, Tomasz Józefowski	GUS (Poland)
0.3	31-01-2014	third version	Marcin Szymkowiak, Tomasz Józefowski	GUS (Poland)
0.4	14-03-2014	fourth version	Marcin Szymkowiak, Tomasz Józefowski	GUS (Poland)
0.4.1	17-03-2014	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