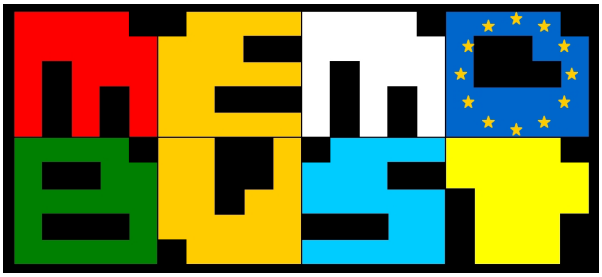

This module is part of the

# Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

# Theme: Small Area Estimation

## Contents

# General section

## 1. Summary

Business surveys carried out by National Statistical Institutes are usually aimed to obtain estimates of target parameters, e.g., the overall amount of industrial turnover for the whole population of business enterprises. Analogous parameters are usually defined with respect to relevant population sub-sets, i.e., sub-populations corresponding to geographical partitions (e.g., administrative areas) or sub-populations associated to economic cross-classification (e.g., enterprise size and sector of activity). An example is given by the estimation of the industrial turnover for each administrative region (e.g., NUTS2 level), or for each sector of activity (e.g., 2-digit NACE). An estimator of the parameter of interest for a given sub-population is said to be a *direct estimator* when it is based only on sample information from the sub-population itself. Unfortunately, for most of surveys the sample size is not large enough to guarantee reliable direct estimates for all the sub-populations. A 'small area' or 'small domain' is any sub-population for which a direct estimator with the required precision is not available. Even though the term 'small domain' may seem to be proper in the business survey context, 'small area' is intended in the literature as a general concept and it is used to indicate a general partition of the population according to geographical criteria or other structural characteristics (socio-demographic variables for household surveys or economic variables for business surveys). In the following we will utilise preferably the term small domain but the term small area will be used too in its wide and meaningful definition.

When direct estimates cannot be disseminated because of unsatisfactory quality, an ad hoc class of methods, called small area estimation (SAE) methods, is available to overcome the problem. These methods are usually referred as *indirect estimators* since they cope with poor information for each domain borrowing strength from the sample information belonging to other domains, resulting in increasing the effective sample size for each small area.

## 2. General description

Sampling designs for business surveys are usually stratified one stage designs, where strata are defined as the cross-classification of structural characteristics of the enterprises as geographical area, economic activity, size in terms of number of workers, etc. Planned domains of interest are usually given by the different sets of marginal strata. In this context small domains are defined as *planned domains* when they are obtained as strata or aggregation of strata. Furthermore small domains are defined as *unplanned domains* when they cut across strata.

This situation is showed in Figure 1, where domains of interest are geographical areas. The example is referred to a one stage stratified sampling design, in which $h$ is the generic stratum ($h = 1, \ldots, H$) and the dots are the sampling units. The figure shows, three different types of small areas that can be potentially encountered:

- the first type, denoted by $d$, is an example of unplanned small area, being the union of complete and incomplete strata. Then the corresponding sample size is a random variable;
- the second kind of small area, denoted by $d'$, is a special case of unplanned small area when no sample units are selected in the target small area;

– finally the third type, denoted by *d''*, is an example of planned small area, being the union of complete strata.

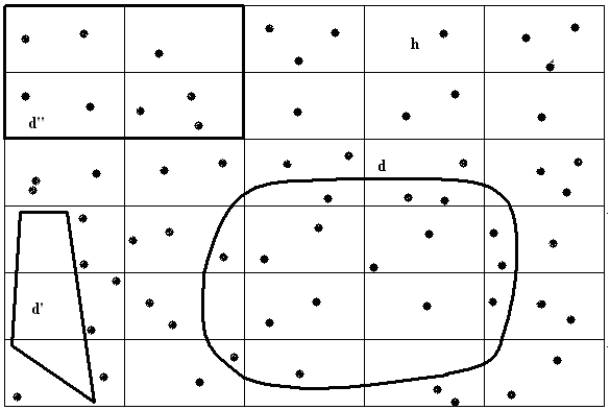


*Figure 1. Different types of small areas.*

Direct estimators, that are obtained within the design-based approach, may produce reliable domain estimates of the target parameters only when the domain sample sizes are sufficiently large. When the realised domain sample sizes are not large enough to guarantee reliable direct domain estimates, indirect methods provide tools to overcome the problem. The main idea underlying these techniques is to increase the effective sample size for each domain by means of the information from the units belonging to other domains considered "similar" (with respect to structural characteristics) to the small domain of interest. The set of all domains from which estimation methods borrow strength will be referred as *broad domain*. For instance in figure 1 the broad domain may be given by the union of all the *H* strata defining the largest rectangle. The more straightforward way to borrow strength is given by the *synthetic estimator*. According to Gonzalez (1973) an estimator is called a synthetic estimator if a reliable direct estimator for a large area (i.e., broad domain), covering several small areas, is used to provide small area estimates under the assumption that all the small areas have the same characteristics as the large area. Synthetic estimators increasing the effective sample size result in smaller variances than direct estimators. On the other hand, bias can seriously affect synthetic estimators, since they make too strong use of information from other areas allowing too little for local variation (overshrinkage). In order to balance the potential bias of a synthetic estimator against the instability of a direct estimator is to take a weighted average of the two estimators. The resulting estimators are known as *composite estimators* (Schaible, 1979). Synthetic and composite estimators are usually referred as *indirect methods*.

It is useful to consider the following classification of direct estimators according the use of auxiliary population information:

– no use of auxiliary population information, corresponding to Horvitz-Thompson (H-T) estimator (Horvitz and Thompson, 1952; Cochran, 1977);

– use of auxiliary population information, these methods improve the efficiency of H-T estimator by means of unit level auxiliary information (observed for each respondent unit) and the corresponding known population totals or means. This class of estimators may be further divided

in methods using: *Domain Specific level* (DS) auxiliary population information and *Aggregated Domain level* (AD) auxiliary population information. The former refers to auxiliary population information available for each small domain, the latter is related to the case of auxiliary population information for aggregations of two or more domains.

Almost all large scale business surveys use direct estimators exploiting auxiliary population information, such as Generalised regression estimator (GREG) or more in general Calibration estimator. Calibration estimator satisfies constraints entailing the equivalence between known auxiliary variables population totals, or means, and the corresponding calibrated estimates. Calibration weights are derived minimising a distance between survey and calibration weights. Deville and Särndal (1992) showed that GREG estimator is a particular case of Calibration estimator under the chi-square distance. For both DS and AD information, it is possible to obtain some well-known special cases of GREG estimator, e.g., Ratio, Post-stratified, Post-stratified Ratio and Ratio-raking estimators, that are broadly used in large scale surveys.

GREG estimator, obtained under the model-assisted framework, allows to define approximately unbiased, and in many cases consistent, direct estimators, exploiting the correlation between the target variable and a set of covariates. A linear fixed model is defined to obtain a reduction of design variance of H-T estimator.

In the case of AD auxiliary population information, Generalised Regression Estimator, $GREG_{AD}$, is approximately unbiased if the overall sample size is large enough, but consistency is obtained only under a large expected domain sample size. Note that, under AD auxiliary information, residuals are different to zero for all units belonging to the sample, then large negative residuals for all the sampled units not belonging to domain $d$ can produce inefficiency.

When DS auxiliary population information is used, the corresponding Generalised Regression Estimator, denoted with $GREG_{DS}$, is approximately unbiased only if the domain sample size is sufficiently large. For $GREG_{DS}$, unlike $GREG_{AD}$, the sample residuals of the units outside domain $d$ are null. Therefore $GREG_{DS}$ can be more efficient than $GREG_{AD}$.

An approximately unbiased direct estimator that may overcome the problems related to the above GREG estimators is known as Modified Direct (MD) estimator. It is equal to $GREG_{DS}$ estimator, but $GREG_{AD}$ regression coefficient vector is used. Then MD estimator borrows strength for estimating the regression coefficients but does not increase the effective sample size as indirect estimators. It is approximately unbiased as the overall sample size increases, also when the domain sample size is small. Note that, like $GREG_{DS}$, residuals are null outside the target domain. Then when the DS and AD regression coefficients are close each other, the MD estimator may results more efficient than $GREG_{DS}$. For more details on the above estimators, see Rao (2003).

SAE methods are characterised by the different ways to borrow strength from information other than the observed values of the target variable in each small domain.

Figure 2, taken from Elazar (2005), synthesises well the different approaches in borrowing strength:

(a) cross-sectional way;

(b) using auxiliary data;

(c) exploiting spatial relationship;

(d) using over time relationship.

The simplest way to borrow strength is using the values assumed by the target variable in all the domains included in the broad domain. This implies assuming that all the domains have a common mean value of the target variable (see case (a) in figure 2). If it is possible to divide the population in sub-groups according to one or more auxiliary information, the following step is to assume common mean values for all the domains within each sub-group. This is a particular case of assuming linear relationship between the target variable and a set of covariates (see case (b) in figure 2). It must be underlined that in case (a) only small domain population counts are needed, while in case (b) users must know small domain population counts for each sub-group when dealing with categorical auxiliary variables or small domains population means when using quantitative variables. In both cases small domains play a symmetric role and have the same importance in the estimation process. Enhanced methods are involved when using spatial or temporal information. Case (c) in figure 2 is related to the case of using spatial information in the estimation process. The main idea is that units belonging to the closest geographical areas should be given more importance in the borrowing strength process. This implies the need of additional information such as distance or neighbourhood matrices among the domains. When small domains are not related to geographical areas, like frequently happen in business surveys, it may be difficult to identify appropriate distance or neighbourhood concepts for domains. The last way to borrow strength from other sources of data may be applied in case of repeated surveys, that is when several survey occasions are available. In this case it would be possible to use the information from the previous survey occasions or times.

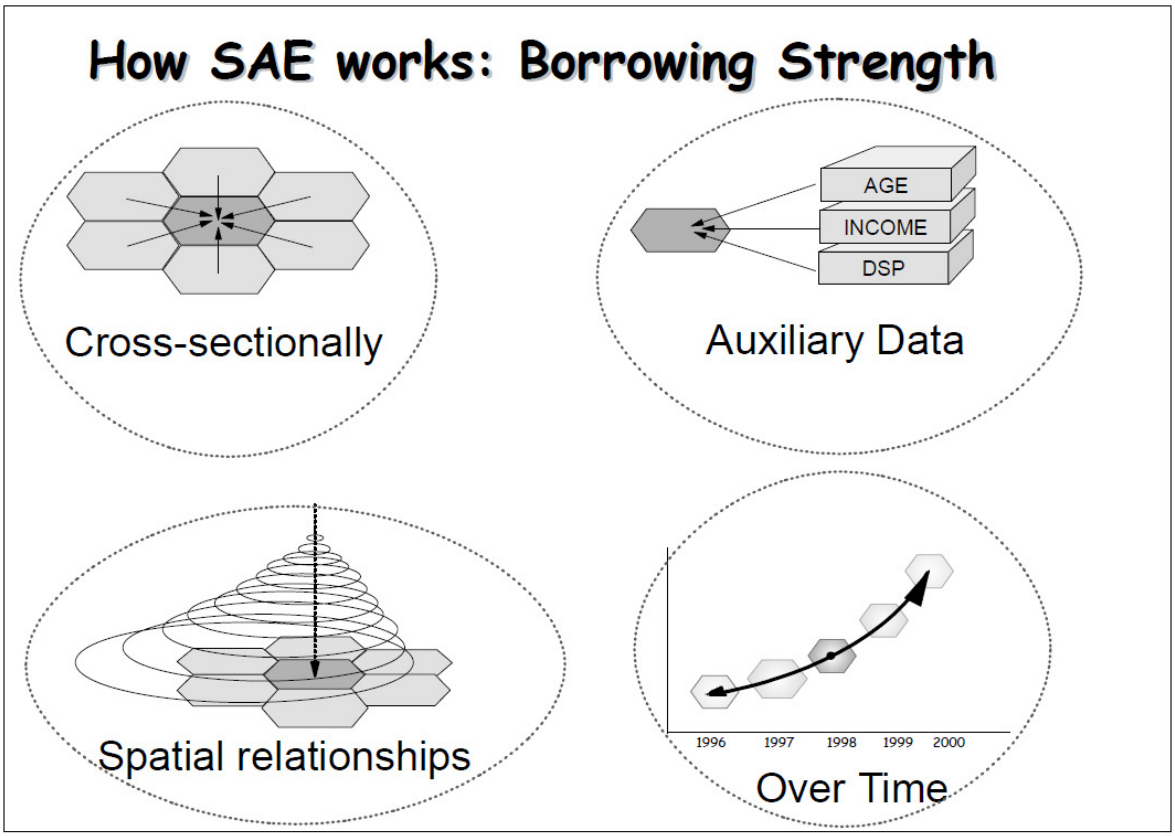


*Figure 2. How to borrow strength: (a) Cross-sectionally, (b) using auxiliary data, (c) exploiting spatial relationship, (d) using over time relationship.*

It is worthwhile to underline that the four approaches described above can be combined together defining in this way the complete set of SAE methods. In fact methods belonging to (a) or (d) can be used also in combination with (b) and/or (c). For example methods involving spatial correlation between the areas can also use auxiliary information, or methods to be applied when repeated survey data are available can also exploit spatial correlation and the information coming from auxiliary variables. Note that SAE methods in (a), (b) and (c) increase the effective domain sample size exploiting all the sampling information coming from the units belonging to the broad domain. SAE methods related to case (d) increase the domain sample size using the sampling information coming from the units observed from previous survey occasions, within the target domain. The joint use of cross-sectional and temporal information is possible too, e.g., using SAE methods related to cases (a) and (d). These techniques lead to a further increase of the effective sample size.

On the basis of the above description, it is useful to propose a classification of SAE methods. The small area estimators are divided into three groups according to the way they use the sampling and population information:

(1) methods involving *spatial smoothing*, using data of all the small domains for only one survey time;

(2) methods involving *temporal smoothing*, using data for only the small domain of interest for several survey occasions;

(3) methods involving *spatial and temporal smoothing*, using data collected for all the domains at different survey times.

The three classes of methods can be further divided according to the inferential approach: *design-based* (d) or *model-based* (m) approach. In the first approach the target parameters are considered as unknown but fixed quantities while in the second one they are dealt as random variables and inference is based on the definition of an explicit model. The model formalises the relationship between data from several small domains within a broad domain, and/or the link between different survey occasions. Model specification involves extra auxiliary information correlated with the target variable, from census or administrative registers. In order to take into account simultaneously the previous classifications, the notation (d) and (m) will be combined with the indexes (1), (2), (3) denoting three different classes of smoothing, e.g., (d.1) will be denote design-based methods involving spatial smoothing, (d.2) design-based methods with temporal smoothing, and finally (m.3) will indicate model-based methods using spatial and temporal smoothing.

As far as the case (d) is concerned we have:

- (d.1) the so called traditional methods, that is synthetic and composite estimators (see respectively Gonzalez, 1973 and Schaible, 1979). Particular cases of design-based composite estimators are the Sample-size dependent estimator (Drew et al., 1982), the James-Stein estimator (see Rao, 2003).
- (d.2) the methods for which it is possible to assume some time dependent correlation among direct estimators. For repeated surveys based on rotated samples, direct estimators can be suitably combined with a gross change estimator based on the common units in two consecutive samples. This provides additional information allowing to improve the efficiency of the estimator at each time. The original idea by Jessen (1942) and Patterson (1950), was improved using a multivariate framework by Gurney and Daly (1965). They introduced the concept of elementary estimator

related to each rotation group. The elementary estimators have been utilised for linear models, which make use of the correlation structure among the estimators to produce Minimum Variance Linear Unbiased Estimators (MVLUE). In practice, the specification and the inversion of the error correlation matrix may result in unstable estimates. One possible way to overcome this problem was suggested by Gurney and Daly (1965), who defined the class of composite estimators which combine the results of two consecutive samples in order to obtain actual estimates.

- (d.3) this class includes either the Gurney and Daly estimator for the case with more than one small area and the estimator proposed by Purcell and Kish (1980), known as SPREE (Structure Preserving Relation Estimator), for categorical data. This is based on the definition of two structures of data. The first is given by the complete population data related to a previous time. This is used to draw for each small area the associative structure information about the link between the target variable and a set of auxiliary variables (complete contingency table). The second source of information is the allocative structure, that is a set of current estimates for some marginal tables. Estimates preserve the observed relationships in the original associative structure except those specified in the allocative structure.

In the model-based approach models are explicitly defined and inference is drawn not anymore from the sample space but from a model on the population values (super-population model). Depending on the level at which the information is specified, area level or unit level models can be specified. In the former the link between target and auxiliary variables is defined for each area, while in the latter the relationship is specified for each unit. The more common methods are Empirical Best Linear Unbiased Predictor (EBLUP), Empirical Bayesian (EB) predictor and Hierarchical Bayesian (HB) predictor. Almost all these methods are based on multilevel models in which one level of the hierarchy is specified at area level. In details these models reduce to linear and generalised linear mixed models in the frequentist approach, and to hierarchical models in the Bayesian framework. Bayesian modelling implies the specification of priors distributions for all the parameters in the model. A regression function with respect to a set of auxiliary variables is introduced. This is usually referred in the frequentist framework as the fixed part of the model and the regression coefficients are indicated as the fixed effects. Moreover to consider the extra-variability not explained by the fixed part of the model, random effects related to each domain are added to the model. On the contrary if area random effects are not included into the model, synthetic model-based estimates are obtained instead of composite model-based estimates. For an extensive overview readers can refer to Rao (2003).

Three classes of model-based SAE methods can be defined:

- (m.1) methods using spatial smoothing. Seminar papers in this context are Fay and Herriot (1979) for area level models, Battese et al. (1988) for unit level models, Morris (1983) for the EB approach, and Datta and Ghosh (1991) and Ghosh (1992) for the HB modelling. Model specifications taking into account spatial correlation of the area random effects are proposed by Cressie (1991), Saei and Chambers (2003), and Pratesi and Salvati (2008);

- (m.2) methods utilising temporal smoothing. Not considering the pioneering works by Scott (1974) and Smith and Jones (1980), worth mentioning are the works by Bell and Hillmer (1987) and Binder and Dick (1989). The proposed methods are based on time series analysis. Milestones of this approach are: (i) considering the observed data over time as a finite subset of a realisation

of a stochastic process; (ii) the definition of state space models (and the application of the Kalman filter to obtain parameter estimates and the correspondent standard errors);

- (m.3) methods using both spatial and temporal smoothing. Some methods are proposed by Pfeffermann and Burck (1990) on the basis of state space modelling. Important results are presented in Singh, Mantel and Thomas (1994), where four generalisations of the Fay – Herriot predictor are proposed. Saei and Chambers (2003) describe models and algorithms to obtain SAE estimates when linear mixed models involving both area and time random effects are defined.

## 3. Design issues

## 4. Available software tools

In the last decade the availability of software tools for SAE is increased significantly. Several routines for multilevel model estimation released by developer teams of R, SAS, SPSS, STATA, MLwiN and WinBUGS or OpenBUGS can be used for small area estimation. Besides, ad hoc software for SAE has been developed by some international projects on the small area estimation topic. It is worth mentioning the SAS macros produced by the EURAREA project, the R functions or libraries released by the projects BIAS, SAMPLE, AMELI and ESSnet-SAE.

Furthermore an extensive review of the SAE software tools is provided in the WP4 of the ESSnet-SAE project.

## 5. Decision tree of methods

In this section we report the step by step process of the activities related to the production of small area estimates as defined in the WP6 of the ESSnet-SAE project. This process is displayed in figure 3, where three separate stages are defined:

(I) clarification: identification of needs and purposes of small area estimation (e.g., estimation of key parameters or ranks for funding allocation);

(II) basic smoothing: direct, and design-based synthetic and composite estimates (triplet) are computed. No change of the inferential framework is needed compared to the direct estimates produced for the survey;

(III) enhancement: it is needed if the basic design-based smoothing is not effective. Quality assessment of the triplet of design-based small area estimates should identify weaknesses in order to work properly for improvements.

Therefore, according to point (III) when design-based methods cannot guarantee the requested precision of small area estimates, enhanced methods based on explicit modelling should be used. After computing model-based estimates, users must verify the validity of the hypotheses underlying the models. Furthermore should also check for the possible bias introduced for misspecification of the model by means a set of bias diagnostics. These diagnostics are based on the comparison between the whole set of direct estimates and model-based estimates. (see Brown et al., 2001). For an overview of model diagnostics for SAE see also the report on WP6 of the ESSnet SAE.

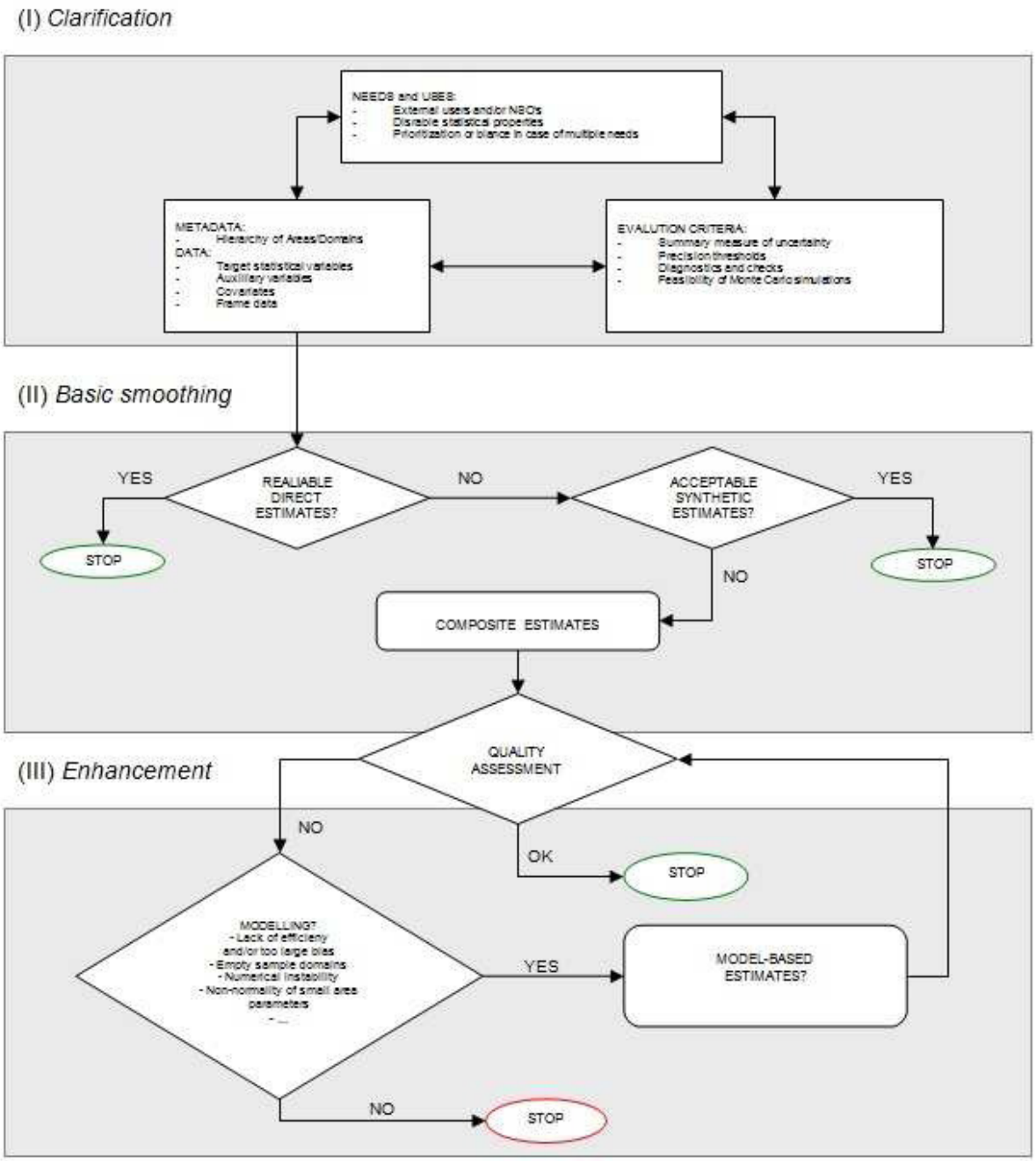


*Figure 3. Flow chart of a SAE process.*

Figure 4 shows the options available when dealing with model-based SAE. For each target variables the model selection phase should include issues as the choice of the more proper set of auxiliary variables and the definition of the small areas to be included in the broad domain. The following step concerns the choice between fixed and mixed effects models. As stated in the previous section the relationship between fixed effects (regression models) and random effects (mixed models) is analogous to that between synthetic and composite estimators. Users are expected to answer a couple of questions: (i) is the regression model good enough? (ii) does the extra computational effort of the mixed model pay off?

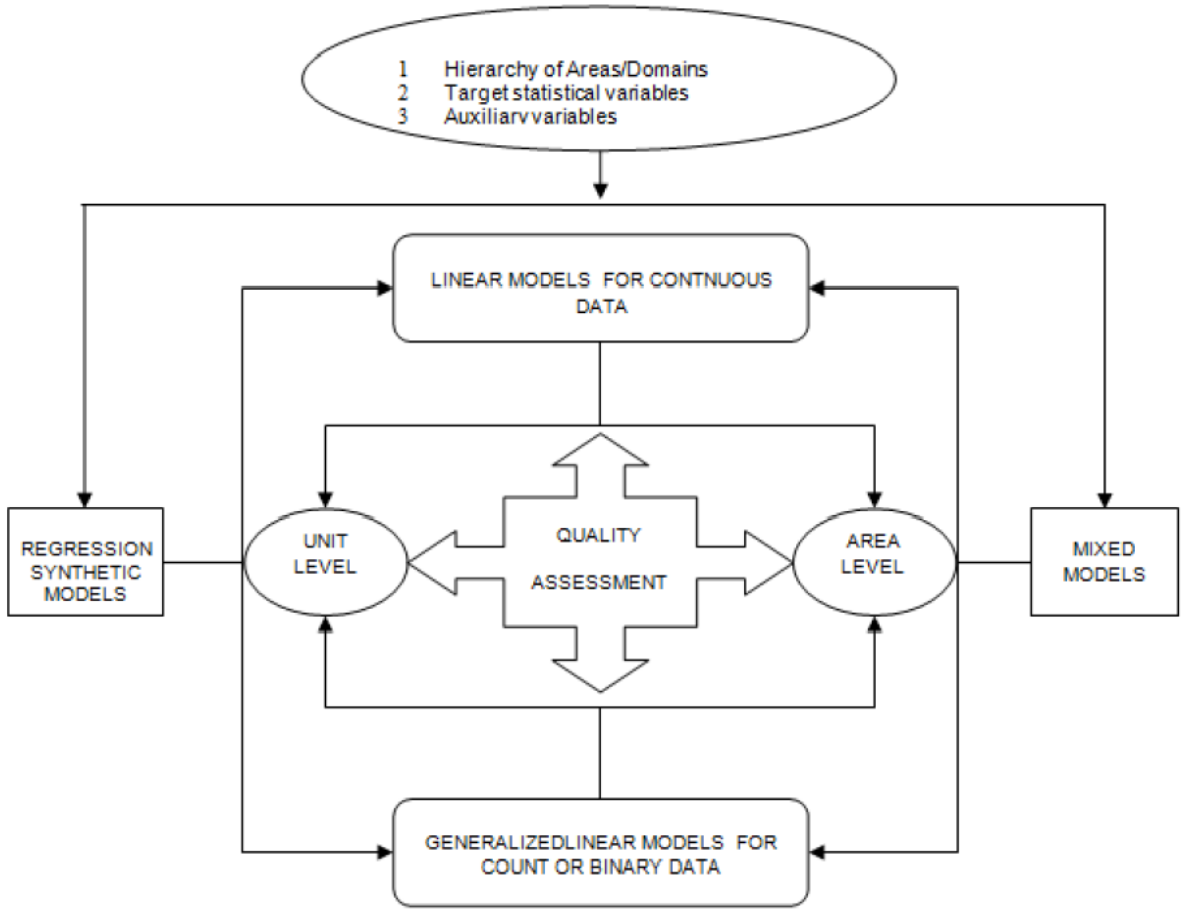


*Figure 4. Flow chart of model-based SAE.*

The following choice depends on the nature and the availability of the data at the two levels. The sampling design also plays a role in the choice. There may design features having a strong impact on the final estimates, e.g., stratification, multistage sample selection and/or clustering. Area level models take into account straightforwardly sampling weights since direct estimates are involved. If unit level models are used, design effects need to be considered as non-informative, given the auxiliary information.

Next step concerns the choice between linear and generalised linear (or nonlinear) models. From theoretical point of view, generalised linear models should be preferred for categorical data. In practice, however, linear models are computationally much easier, and often yield similar results.

## 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* **83**, 28–36.

Brown, G., Chambers, R., Heady, P., and Heasman, D. (2001), Evaluation of small area estimation methods: an application to unemployment estimates from the UK LFS. *Proceedings of Statistics Canada Symposium Achieving Data Quality in a Statistical Agency: A Methodological perspective*, Statistics Canada.

Cochran, W. G. (1977), *Sampling Techniques*. John Wiley & Sons, Hoboken, New Jersey.

Cressie, N. A. (1991), *Statistics for spatial data*. John Wiley & Sons, Hoboken, New Jersey.

Datta, G. S. and Ghosh, M. (1991), Bayesian prediction in linear models: applications to small area estimation. *The Annals of Statistics* **19**, 1748–1770.

Deville, J. C. and Särndal, C.-E. (1992), Calibration estimators in survey sampling. *Journal of the American Statistical Association* **87**, 376–382.

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

ESSnet SAE (2012), Report on Workpackage 6 – Guidelines (contributors Istat (Italy), CBS (Netherlands), SSB (Norway), GUS (Poland), INE (Spain), ONS (United Kingdom). http://www.essnet-portal.eu/sae-2

Fay, R. E. and Herriot, R. A. (1979), Estimates of income for small place: an application of James-Stein procedures to Census Data. *Journal of the American Statistical Association* **74**, 398–409.

Ghosh, M. (1992), Constrained Bayes estimation with applications. *Journal of the American Statistical Association* **87**, 533–540.

Gonzalez, M. E. (1973), Use and evaluation of synthetic estimates. *Proceedings of the Social Statistics Section*, American Statistical Association, 33–36.

Gurney, M. and Daly, J. F. (1965), A multivariate approach to estimation in periodic sample surveys. *Proceedings of The Social Statistics Section*, American Statistical Association, 242–257.

Jessen, R. J. (1942), Statistical investigation of a farm survey for obtaining farm facts. *Iowa Agriculture Station Results Bulletin* **304**.

Horvitz, D. G. and Thompson, D. J. (1952), A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* **47**, 663–685.

Morris, C. N. (1983), Parametric empirical Bayes inference: theory and applications. *Journal of the American Statistical Association* **78**, 47–59.

Patterson, H. D. (1950), Sampling on successive occasions with partial replacement of units. *Journal of the Royal Statistical Society, Series B* **12**, 241–255.

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

Pratesi, M. and Salvati, N. (2008), Small area estimation: the EBLUP estimator based on spatially correlated random area effects. *Statistical Methods and Applications* **17**, 113–141.

Purcell, N. J. and Kish, L. (1980). Postcensual estimates for local areas (or domains). *International Statistical Review* **48**, 3–18.

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

Saei, A. and Chambers, R. (2003), Small Area Estimation Under Linear and Generalized Linear Mixed Models With Time and Area Effects. *S3RI Methodology Working Papers*, Southampton Statistical Sciences Research Institute, M03/15.

Schaible, W. L. (1979), A composite estimator for small areas. *National Institute on Drug Abuse, Research monograph,* U.S. Government Printing Office, 24.

Singh, A. C., Mantel, H. J., and Thomas, B. W. (1994), Time series EBLUPs for small areas using survey data. *Survey Methodology* **20**, 33–43.

## Interconnections with other modules

### 8. Related themes described in other modules

1. Sample Selection – Main Module

### 9. Methods explicitly referred to in this module

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

### 10. Mathematical techniques explicitly referred to in this module

1.

### 11. GSBPM phases explicitly referred to in this module

1.

### 12. Tools explicitly referred to in this module

1.

### 13. Process steps explicitly referred to in this module

1.

# Administrative section

## 14. Module code

Weighting and Estimation-T-Small Area Estimation

## 15. Version history

| Version | Date | Description of changes | Author | Institute |
|---|---|---|---|---|
| 0.1 | 05-04-2012 | first version | Stefano Falorsi, Fabrizio Solari | ISTAT |
| 0.2 | 14-08-2012 | changes after review | Stefano Falorsi, Fabrizio Solari | ISTAT |
| 0.2.1 | 25-02-2013 | changes after review | Stefano Falorsi, Fabrizio Solari | ISTAT |
| 0.2.2 | 10-09-2013 | preliminary release | | |
| 0.3 | 24-10-2013 | changes after EB review | Stefano Falorsi, Fabrizio Solari | ISTAT |
| 1.0 | 26-03-2014 | final version within the Memobust project | | |
| | | | | |
| | | | | |

## 16. Template version and print date

| | |
|---|---|
| Template version used | 1.0 p 4 d.d. 22-11-2012 |
| Print date | 26-3-2014 13:34 |