This module is part of the

# Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

# Theme: Weighting and Estimation – Main Module

## Contents

# General section

## 1. Summary

The present module gives an overview of the methods that can be used to obtain estimates for parameters such as totals, means or ratios, from the observed sample data. It is assumed that data have already been processed to treat potential errors and item non-response (see the modules “Statistical Data Editing – Main Module” and “Imputation – Main Module” for introduction to treatment of errors and item non-response).

Commonly, in official statistics, probability-based sampling designs are carried out, and a design weight can be associated to each sampled unit. This design weight equals the inverse of the inclusion probability. It can be thought as the number of population units each sample unit is representative of. Hence, a simple method to obtain estimates of the target parameters is to use these design weights to inflate the sample observations (see subsection 2.1). Design weights are strictly related to sampling design implemented for the survey (see the module “Sample Selection – Main Module”). Moreover, design weights can be adjusted also to consider non-response (see subsection 2.2), and/or they can be modified to take into account of auxiliary information (Särndal et al., 1992). An example of use of external information is given by the calibration estimator (see the module “Weighting and Estimation – Calibration”) or the GREG estimator (see the module “Weighting and Estimation – Generalised Regression Estimator”), which is a special case of calibration estimator.

The previous estimators are unbiased or approximately unbiased in a randomisation approach (or design-based approach: properties are assessed on the set of all possible samples). Note that even if, in some cases a model is assumed (as for GREG), the properties of the estimators do not depend on the model and the estimators remain design unbiased even in case of model failure. For this reason, this class of methods is robust. However, their efficiency depends strongly on model assumptions and relationships on auxiliary variables affect their variances.

In fact, when the distribution of the target variable in the population is highly skewed, as it often happens in business surveys, representative outliers may occur in the sample. The values of such units are true values and then they do not need to be edited (see the topic “Statistical Data Editing”). Nevertheless, even if estimators remain unbiased, presence of these outlying units has a large impact on variance estimators. The module “Weighting and Estimation – Outlier Treatment” gives an overview of methods that have been suggested in literature for reducing variance of the estimates, while controlling for the presence of bias.

A relevant approach for estimation is given by model-based approach: differently from design-based approach, where, as stated above, properties are assessed on the set of all possible samples, in this framework, the assumption of a model is the basis to obtain estimators that are the best in terms of model Mean Square Error: Best Linear Unbiased Predictor (Royall, 1970, Vaillant et al., 2000). In official statistics, the class of model-based estimator is applied in specific situations, such as when the sample size is not large enough to obtain estimates with sufficient accuracy (small area estimators, see also the module “Weighting and Estimation – Small Area Estimation”). A second important field of application of model-based estimation is given by preliminary estimation, when for short term statistics a provisional estimate is calculated on a sub-sample of the sample units. The auto-selection of units in the preliminary sample may be the most relevant issue for preliminary estimates. Moreover, when the sample is selected with a non-probabilistic mechanism, model-based estimates can be applied for inference, and model-based variance can be evaluated.

The peculiarity of panel surveys is also highlighted. In panel surveys, the same units are observed in several occasions (waves), allowing for reduction of estimators' variance and estimation of longitudinal parameters (e.g., gross change and measure of frequency). Cross-sectional and longitudinal weights have to be determined according to the target parameters (see subsection 2.6).

Finally, the use of administrative data is mentioned in subsection 2.9.

To conclude the review of relevant issues in weighting and estimation, subsection 2.10 underlines some of the most typical matters in applied cases.

## 2. General description

### *2.1 Weighting – basic weighting*

A very important methodology in sampling strategy is provided by the use of weights to obtain estimates of the parameter of interest such as totals (levels), means, differences (or ratios), *etc*. In official statistics, probabilistic sample designs are regularly implemented and a design weight equal to the inverse of the inclusion probability can be associated to each sample unit.

The design weight can be thought as the number of units in the population, a unit in the sample represents.

On this basis, a very simple principle to obtain estimates is to use the design weights. Estimates are produced by summing up the sample data multiplied by their design weights, i.e., the data are inflated with the weights for reproducing the whole population.

Let $y_i$ be the value of the target variable associated to the i-th unit, and let $d_i$ be the weight equal to the inverse of the inclusion probability, an estimation of the total of Y is given by:

$$\hat{Y}_{HT} = \sum_{i \in s} d_i y_i \ . \tag{1}$$

The resulting estimator is called the Horvitz-Thompson estimator.

The Horvitz-Thompson estimation of the mean is

$$\hat{\bar{Y}}_{HT} = \frac{1}{N} \sum_{i \in s} d_i y_i \ .$$

Whereas, in estimating the mean value, if the amount of population is estimated as well, the Hájek estimator is obtained

$$\hat{\bar{Y}}_{H} = \frac{1}{\hat{N}} \sum_{i \in s} d_i y_i = \frac{1}{\sum_{i \in s} d_i} \sum_{i \in s} d_i y_i \ . \tag{2}$$

The use of design weights is relevant in particular whenever the sample design assigns different inclusion probabilities to units in the population, e.g., to account for different size of population units if size is thought to be related with the main target variable.

In the case of stratified simple random sampling design (see the module "Sample Selection – Main Module"), for unit *i* belonging to stratum *h*:

$$d_{ih} = \frac{N_h}{n_h},$$

$\hat{Y}$ reduces to

$$\hat{Y} = \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{i=1}^{n_h} y_i ,$$

whereas

$$\hat{\bar{Y}}_{HT} = \frac{1}{N} \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{i=1}^{n_h} y_i = \sum_{h=1}^{H} \frac{N_h}{N} \left( \frac{1}{n_h} \sum_{i=1}^{n_h} y_i \right) = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_h .$$

More complex indicators can be estimated by replacing true values by their respective HT estimators. For example, estimation of change of variable Y in a given lag-time *l* is given by

$$\hat{Y}_{t+l} - \hat{Y}_t . \qquad (3)$$

Estimation of relative change is given by

$$\frac{\hat{Y}_{t+l} - \hat{Y}_t}{\hat{Y}_t}, \qquad (4)$$

where $\hat{Y}_t$ and $\hat{Y}_{t+l}$ are the estimates of Y at different times *t* and *t+l*, obtained by applying formula (1).

### 2.2 *Weight adjustment for non-response*

The principle of weighting is also applied to account for unit non-response of sample units. In fact, design weights can be adjusted also to consider non-response in order to reduce the possible bias of resulting estimates, which may arise when there is a different propensity in answering for different groups. For example, the sample can be partitioned into sub-groups of units where the response rates are assumed to be constant, and where it can be assumed that non-respondents behave similarly to respondents. More precisely, the method is based on the assumption that the non-response depends on variables that define the sub-sets, but conditionally on these variables it is independent of the target variable (non-response is missing at random, MAR, see Little and Rubin, 2002). This grouping may differ from the sampling strata and cut across them.

A response rate, possibly weighted by the initial sampling weights, is determined in each class and a new weight is defined as the product of the design weight and the inverse of the response rate. The new weights are used in the weighting process of respondent sample units in order to get the estimates.

Let us assume for simplicity in notation that sample design is stratified and that sub-groups (or post-strata) coincide with design strata, the response rate in stratum *h* is evaluated as:

$$r_h = \frac{n_{rh}}{n_h}.$$

Then the initial weight of unit $i$ in stratum, $d_{hi} = \frac{N_h}{n_h}$, is replaced with the new weight $w_{hi} = \frac{d_{hi}}{r_{hi}} = \frac{N_h}{n_{rh}}$ and the usual HT is given by:

$$\hat{\bar{Y}}_{HT} = \frac{1}{N}\sum_{h=1}^{H}\frac{N_h}{n_{rh}}\sum_{i=1}^{n_{rh}} y_i = \sum_{h=1}^{H}\frac{N_h}{N}\left(\frac{1}{n_{rh}}\sum_{i=1}^{n_{rh}} y_i\right) = \sum_{h=1}^{H}\frac{N_h}{N}\bar{y}_{rh} \ .$$

Occasionally, unit non-response can also be treated by imputation methods (see the module "Imputation – Main Module").

### *2.3 Weight adjustment*

Besides the modification of weights for handling non-response, weights adjustment may also be carried out to take into account of auxiliary information, for example by means of the calibration estimator (see the module "Weighting and Estimation – Calibration") or GREG estimator (see the module "Weighting and Estimation – Generalised Regression Estimator"). The use of auxiliary information can have the aim to insure consistency among estimates of different sample surveys. Indeed, when good covariates are available, some improvement in the precision of estimators may be achieved by exploiting the relationship between target variable and extra information.

Auxiliary data can be used to improve the precision of the estimators as long as the values of the auxiliary variables are collected for all surveyed units and known population totals are available for these variables from another reliable source in case a linear relationship is assumed. Otherwise, totals do not suffice; see comments in the module "Weighting and Estimation – Generalised Regression Estimator".

A general method for exploiting auxiliary information is calibration estimators (see the module "Weighting and Estimation – Calibration"). The weights are adjusted so that applying the estimators on the auxiliary variables, one is able to reproduce the known covariates totals. Calibration includes well-known estimators such as the regression, the ratio and the raking-ratio estimators (Deville and Särndal, 1992).

However, the calibration estimator may introduce high variability in weights and consequently an increase in variance of the estimator which may be relevant whenever the auxiliary variables are not enough correlated with the target variable. In particular, in official statistics, where the same set of weights is used for several target variables, it may happen that the set of covariate used in determining the final weights is not appropriate for reducing the variance of the estimators of a sub-set of the target variables.

Besides the aim of actually improving the accuracy of the sample estimators, calibration is often applied in practice to attain consistency of estimates obtained with different sources. In fact, the estimates calculated with a survey should be consistent with information on known totals obtained, for example, from a larger survey or from reliable administrative sources. Though, problems to achieve consistency can be encountered in practice if weights are also forced to lie within a given range.

An important use of calibration estimators is for further reducing the effect of unit non-response on the estimators and the possible coverage error of the sampling frame (see Lundström and Särndal, 2001).

In fact, calibration estimators may offer some protection against non-response bias when non-response is related with variables used in calibration. Poststratification and regression estimation, both special cases of calibration estimators, are widely used techniques to attempt to reduce non-response bias in sample surveys. Särndal and Lundström (2005) suggest the use of calibration to handle non-response.

Finally, weighting can be applied to combine different samples (sources) and produce estimators that are more accurate than the estimators based on any of the single samples individually (e.g., see Renssen and Nieuwenbroek, 1997, Houbiers et al., 2003).

Once the weights are obtained, estimators of totals, means are easily obtained as described above.

### *2.4 Robust estimation in the presence of outliers*

In business surveys, the statistical distribution of target variables is often highly skewed, hence in observed sample observations that differ substantially from most of the other observations occur. These units, referred as *representative outliers* (see Chambers, 1986), are true values in the finite population and should not be considered as gross errors.

In particular, presence of this kind of outlying values in the sample does not affect the bias of the HT or calibration estimators described in 2.2 and 2.3. However their occurrence has usually a great impact on their variability.

Outlier treatment at estimation stage (robust estimation) aims at reducing the effect on variance of outliers, also controlling the possible bias of the estimator.

The methods for dealing with outliers can be broadly classified as winsorisation, modification of weight, and M estimation, i.e., methods for robust estimation in classical theory properly adapted in the finite population estimation framework.

In particular, winsorisation consists in modifying the outlying observations so that they have less impact on the estimation. Sample observations whose values lie outside certain pre-set cut-off values are set equal to the cut-off (type I winsorisation) or are transformed as a linear combination of the observed value and the cut-off (type II winsorisation) with coefficients for the observed values equal to the inverse of the sampling weights (see Gross et al., 1986, Kokic and Bell, 1994).

In case of simple random sample (s.r.s.), the winsorised estimator is

$$Y_{WR} = \frac{N}{n}\sum_{i\in s} Y_i^*$$

where

$$Y_i^* = \begin{cases} fy + (1-f)K & \text{y} > \text{K} \\ y & otherwise \end{cases}$$

and $f$ is a coefficient in [0,1]. When $f = 0$, the winsor estimator is said winsor of type I, whereas, when f = $n/N$, a winsor type II estimator is obtained.

An extension for a stratified sampling design is in Gross et al. (1986). Choice of cut-off under superpopulation models are in Kokic and Bell (1994), Chambers and Kokic (1993). See also the module “Weighting and Estimation – Outlier Treatment” for a detailed description of winsorisation estimator and the choice of cut-off for a general sampling design. Once the data are transformed the estimation process consists in applying the chosen estimator (e.g., GREG) to the new set of data.

The cut-off values are chosen to approximately minimise the MSE of the resulting estimator, usually under model assumptions (e.g., see Kokic and Bell, 1994, for optimal cut-off in stratified sampling design), the efficacy of this method is highly dependent on the goodness of cut-off(s) choice.

An alternative class of methods relates to modification of sampling weights, i.e., weights are reduced to decrease the impact of outlying units. Various methods for weight reduction have been proposed (see Hidiroglou and Srinath, 1981, Lee, 1995).

For s.r.s., Hidiroglou and Srinath (1981) suggested

$$\hat{Y}_{HS} = \lambda \sum_{i \in s_2} Y_i + q(\lambda) \sum_{i \in s_1} Y_i ,$$

where $s_1$ is the sub-sample of *inliers* and $s_2$ is the subsample of outliers, $q(\lambda)$ is a function of the downweighting factor $\lambda$ of the outlying units, such that

$$\lambda n_2 + q(\lambda) n_1 = N .$$

Hidiroglou and Srinath (1981) proposed a method to determine $\lambda$ in order to minimise the conditional mean squared error, which is difficult to apply in practice. Chambers (1986) obtained an optimum value that minimises the model-based mean squared error. This method requires estimation of unknown parameters of the two different models underlying the subpopulation of inliers and the subpopulation of outliers.

Finally, the class of M estimators (Huber, 1981) is applied to HT or GREG estimators in the finite population sampling framework (e.g., see Chambers, 1986, Hulliger, 1995, Beaumont and Alavi, 2004).

See Beaumont and Rivest (1999) for a description of the methods and a presentation of practical issues. The module "Weighting and Estimation – Outlier Treatment" in this handbook provides a review of methods for dealing with outliers at estimation stage focusing on winsorisation methods.

### *2.5 Model-based estimators*

The weighting methods described previously rely on inference that is based on the randomisation introduced by the sampling mechanism. This approach is more robust to model failure, i.e., less dependent on model assumptions on super-population[1] relationships between the target variable and auxiliary variables and for this reason commonly applied in official statistics. Though, model-based framework for inference in finite population sampling (see Valliant et al., 2000) is applied in specific fields of application, as it can produce more reliable estimators than those obtained with the traditional design-based (or model-assisted[2]) approach, and it may be preferable in cases where the sample size is very limited. We mention here some circumstances where model-based estimation is applied in official statistics.

[1] A mechanism generating the realised finite population.

[2] The model-assisted approach assumes a super-population relationship, as well. However, on the contrary to model-based approach, the properties of the estimators are still based on the randomisation approach. The calibration estimator described in the previous subsection is an example of model-assisted estimator.

An important field of application of model-based estimators, as we will see in subsection 2.8, is on small area estimation[3]. The issue of small area estimation arises whenever the sample size of a target domain is not large enough, so the direct estimator has too large variability to be published (see EUROSTAT, 2013, for some examples of threshold on reliability of the estimators). A large development in terms of methods and software, as well as real applications in official statistics has been produced in recent years (see Rao, 2003, EURAREA project, and WP2 and WP6 reports of ESSnet SAE).

Model-based estimation has also been proposed for the dissemination of short term statistics where the need for timely estimates conflicts with the need to observe the whole planned sample (Rao et al., 1986). In this case, besides the problem of estimation in presence of few observed data, one has to deal with risk of presence of (auto) selection bias. See the module "Weighting and Estimation – Preliminary Estimates with Model-Based Methods" for model-based methods to tackle the preliminary estimation issue.

Finally, note that whenever the sample is selected without a randomisation mechanism but units are chosen purposely, model-based estimation represents the framework for assessing the obtained estimators. More specifically, the implicit model that is underlying the estimation method can be evaluated in order to give support to, or, on the contrary, to invalidate the strategy used for estimation (see Kalton, 1983). For example, the ratio estimator is commonly associated with cut-off sampling, which is often chosen for convenience and cost consideration. This strategy can be justified under a ratio model. Validity of this model can be verified to assess the whole sampling strategy and measures of variability can be provided following this approach (see Valliant et al., 2000, and the module "Sample Selection – Main Module" for a review of sampling designs). See also Benedetti et al. (2010) where a model-based estimator is proposed for the unobserved subpopulation in a cut-off framework.

Models that are used at micro level to cope, for example, with non-response or to edit units (for these issues, see the topics "Statistical Data Editing" and "Imputation") are not reported within the weighting and estimation topic.

### *2.6 Panel surveys*

Short term surveys make use of repeated surveys (see the module "Repeated Surveys – Repeated Surveys") to produce estimation of monthly or quarterly changes. For this reason, overlapping of samples, instead of renewing the sample at each occasion of the same survey, is applied as it allows reducing the variance of estimation of net changes. In fact, variations over time are measured more accurately with overlapping samples with respect to the case where samples on different occasions do not overlap (see for example Eurostat, 2013 for estimation of variance of changes when samples overlap). Actually, standard errors of the estimate of changes over time are minimised by using complete overlap of samples (Kish, 1965) if the correlation between observations in different periods is positive, as it is usually the case. Estimation of changes is a relevant objective for short term statistics and the use of panel (or rotating panels, see Kish, 1987), where the same set of units is observed each month or quarter of the year(s), is a mean to attain the aim of reducing its variance.

Note that, whereas, in a repeated survey with independent samples at each occasion, net change reflects a combination of changing values and changing population composition, on the contrary in a

[3] Model-based estimators are not the only class of methods applied in this field, even if they have a central role.

panel survey, unless steps are taken to incorporate new entrants at later waves as in rotating panels, net change reflects only changing values but refer to the initial population. See Duncan and Kalton (1987) for a comprehensive review of the design and analysis of longitudinal data. See also the module "Repeated Surveys – Repeated Surveys" for a discussion of possible alternative sampling designs to be applied in repeated surveys. In this module focus is given to panel surveys and in particular way to issues in estimation and in determination of sampling weights.

An important preliminary matter when a panel survey is conducted is the definition of continuity rules in order to establish whether an enterprise represent the same unit over the different sampling occasions (waves). This definition, of course, affects definition of target population and statistical units and have effect on sample definition. The interested reader can refer to the modules "Repeated Surveys – Repeated Surveys" and "Dynamics of the Business Population – Business Demography" where aspects of continuity rules are discussed, here we focus on relevant issues in the determination of sampling weights. In this respect, let us note that in a panel survey, two types of weights can be calculated: cross-sectional weights and longitudinal weights. This distinction depends on the nature target populations and related parameters. Cross-sectional weights refer to a population of a given wave and are used to estimate parameters of the given population. Longitudinal weights are used to estimate parameters referring to the longitudinal population, i.e., the population in different occasions. Examples of the latter are gross change[4] and measures of frequency, timing and duration of events occurring within a given time period.

Definition of cross-sectional weights for each wave of the survey to reproduce the target population proceed similarly to the standard cross-sectional surveys, but may require specific computation when using panel surveys.

Evaluation of cross-sectional weights for the first occasion of the survey follows the standard steps described in subsections 2.1, 2.2 and 2.3: determination of a design weight equal to the inverse of the inclusion probability and subsequent adjustment for non-response and for improving estimators.

It has to be underlined that if no-renewal is done in the panel, the sample is in fact representative only of the initial population. Moreover, even if the population is fixed, after the first wave, determination of weights should take attrition into account. Then, at each subsequent wave, the first operation should consist in adjusting the first wave weights for non-response due to attrition. On the other hand, as population is subject to changes, it is important to modify weights to reflect these changes, as well. If updated totals are available then calibration to new totals can reduce presence of bias (see also subsection 2.10).

If, on the contrary, a refreshment of panel is done to represent the population dynamics, the sample in a given wave is composed of different parts. To obtain estimate of cross-sectional indicators, two different approaches may be applied. One approach is determining weights for each component and then combine the estimates, the second consists in pooling the two samples by assigning a unique weight. This second method may be less straightforward to apply in practice due to complexity in computation of inclusion probability in both samples.

[4] On the contrary, estimation of net change requires use of cross-sectional weights at each wave and proceed as described in sub-section 2.1.

Determination of longitudinal weights requires first definition of the target population, which may be for example the set of units present both at time $t_0$ and at time $t$, or the initial population at time $t_0$ only. In the first case, for example, one assigns weights only to overlapping units in the two different samples and the longitudinal weight is given by the product of cross-sectional weight in $t_0$ and the conditional weights to units in $t$ that belong to $t_0$.

Use of panel survey is much more established in sampling surveys on households where examples of weights definition can be found (e.g., Verma et al., 2006) . An example of panel in a business survey with discussion of different weights usage can be found in Australian Bureau of Statistics (2000, pages 9-20).

It has been mentioned at the beginning of this subsection, that repeated observations on the same unis allows reduction of variance of the estimators of changes. An additional advantage of observing repeatedly the same units consists in the possibility of explicitly exploiting the temporal correlation which arises between observations on the same units at the different occasions of the survey to improve estimators on the basis of models which take into account the autocorrelation between observation on the same units at different time points to estimate cross-sectional measures. Model-based estimators for panel data are for example proposed in Fabrizi et al. (2007).

Before concluding this subsection, it is important to note that overlapping of samples induced with (rotated) panels require also special concern for variance estimation both for estimation of levels when more sampling occasions are involved (e.g., means of quarters in a year) and for estimation of changes. In fact, for example, when estimating a measure of change, the variance estimation of this estimate has to account for the correlation between estimators at different times of the repeated survey. See for example, Nordberg, (2000) for a proposal for coordinated sample with permanent random numbers, Qualité and Tillé (2008) for a proposal of variance estimation of changes in repeated surveys, Berger (2004) and finally, Knottnerus and Van Delden (2012) for variance estimation of changes in rotating panels. See EUROSTAT (2013) for review of variance estimation methods and key references.

### *2.7 Preliminary estimates*

As already mentioned in subsection 2.5, timeliness in disseminating the estimates is a very important aspect of the quality of short term statistics and it is also one of the main peculiarities of them.

For short term statistics, in fact, the planned sample may occur to be partially observed when the estimates have to be disseminated. Preliminary (provisional or early) estimates are the estimates that are computed using the statistical information available on the basis of the preliminary sample (PS), i.e., the subset of the planned final sample (FS) that is observed at time of first release of the estimates.

The main problem that has to be faced in a short-term preliminary estimation context concerns the possible self-selection of early respondents, since self-selection can lead to biased estimators of the unknown population mean and variances. Early respondents may have systematically different (e.g., lower) values in terms of the target variables from late respondents.

Preliminary estimation methods may be classified in function of the stage on which the preliminary method is applied.

In fact, it is possible to identify different methods according to the stage they are implemented in:

1. the sampling design stage, by selecting a preliminary subsample of the planned sample (see the module "Sample Selection – Subsampling for Preliminary Estimates");
2. the estimation stage, in the following ways:
   a) by means of imputation techniques of missing data, that are applied to non-respondent units in FS but not in PS;
   b) by means of weighting adjustment, i.e., modifying the sampling weights assigned to the units in PS in order to take into account non respondents of the FS;
   c) by applying direct and indirect estimators, using known population totals of auxiliary variables and/or time series of preliminary and final estimates of the variable of interest (see the modules "Weighting and Estimation – Preliminary Estimates with Design-Based Methods" and "Weighting and Estimation – Preliminary Estimates with Model-Based Methods").

The different approaches can be compared in terms of bias and revision error, i.e., the difference between preliminary and final estimates.

See the module "Weighting and Estimation – Preliminary Estimates with Design-Based Methods" for a description of design methods, in particular for a method proposed in Rao et al. (1989) which at time $t$ exploit time $t$ and $t-1$ data aiming at minimising the mean square error of the estimator. Moreover, see the module "Weighting and Estimation – Preliminary Estimates with Model-Based Methods" for a description of a model-based estimator proposed by Rao et al. (1989), which introduces model that use disaggregated auxiliary information coming from survey data at previous times and/or administrative register data. For these, the relationship between the variable of interest and the auxiliary variables is usually formalised through domain level models in which the auxiliary information is expressed in terms of domain known totals or estimates. An estimation technique of the latter class was developed by Rao et al. (1989). In their proposal, preliminary estimates are computed on the basis of a first order autoregressive model for final estimates and revision errors.

### *2.8 Small area estimation*

The aim of small area (domain) estimation methods is to produce reliable estimators for the variable of interest under budget and time constraints. In fact, National Statistical Office surveys are usually planned for large domains. Hence, whenever more detailed information is required, the sample size may be not large enough to guarantee the release of direct estimators at the desired level of disaggregation. In the most extreme cases direct estimator cannot be calculated when no units belonging to the domain occur in the observed sample. For instance, one is interested in the overall amount of industrial turnover for the whole population of business enterprises, and also in estimating analogous parameters 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).

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

This means that their variability does not depend on the sample size of domain *d*, but on sample size of a larger area (see Rao, 2003).

More precisely, the increase in efficiency of SAE is obtained by means of information on units belonging to other areas considered geographically close or similar with respect to structural characteristics to the small area of interest. In practice, an improvement in the efficiency of the estimators can be achieved by assuming, implicitly or explicitly, a relationship which links together sampling units in the small area of interest and sampling units in the small areas which behaves similarly to the small area of interest. Enhanced methods are involved when applying model using complex spatial or temporal information. In particular, the model using temporal information may be useful in case of repeated surveys, i.e., when several survey occasions are available. In fact, in this case it would be possible to use the information from the previous survey occasions or times.

An account of small area estimation is given in the module "Weighting and Estimation – Small Area Estimation". Specific small area methods, both design-based and model-based, are described in the modules "Weighting and Estimation – Synthetic Estimators for Small Area Estimation", "Weighting and Estimation – Composite Estimators for Small Area Estimation", "Weighting and Estimation – EBLUP Area Level for Small Area Estimation (Fay-Herriot)", "Weighting and Estimation – EBLUP Unit Level for Small Area Estimation", and, finally, "Weighting and Estimation – Small Area Estimation Methods for Time Series Data".

Area and unit level EBLUP are both based on linear mixed model assuming a random area (domain) effect to take into account extra variability between areas not accounted for by the linear relationship between target and auxiliary variables. Both estimators are a linear combination of the direct estimator and the synthetic prediction resulting from the model. The area level EBLUP can be applied also when only macrodata referred to domain level are available, in this case variance of the direct estimator has to be (or assumed to be) known. Furthermore, to exploit temporal information a dedicated method module "Weighting and Estimation – Small Area Estimation Methods for Time Series Data" is provided. Some of these methods are based also on linear mixed models, in which time random effect is introduced or alternatively on auto-regressive specifications.

For a review of recent developments on small area estimation, see Pfeffermann (2013).

### *2.9 Integration of administrative sources in the statistical production*

Nowadays there is an increasing interest in using administrative data for production of official statistics. The administrative data are meant not only as a source of auxiliary information or as a tool for building sampling frames, but also as a source of statistical information itself in place of sample surveys and censuses (Wallgren and Wallgren, 2007), in order to reduce costs and statistical burden.

Hence, though, traditionally, administrative records are used to support the survey work, now more and more increasingly, administrative records are given a central role in the statistical process, to completely replace the collection of survey data. Sample surveys are now part of a more complex system where more sources and surveys are combined together. In some cases they represent the supplementary data that may be used to adjust for data quality (see Eltinge, 2011) or to complement administrative data when coverage issues arise.

Having administrative data acquired a relevant role in the production of official statistical output, the issue of establishing a framework for assessing, measuring, documenting and reporting on quality of administrative data sources and its statistical potential usability has received a considerable attention. An example of a framework for assessment of quality of administrative data can be found, for example, in Daas et al., 2011, mainly developed within the European project BLUE-ETS (http://www.blue-ets.istat.it/fileadmin/deliverables/Deliverable4.1.pdf; Laitila et al., 2011). In the present handbook, the module "Weighting and Estimation – Estimation with Administrative Data" reports the main aspects to be considered when administrative data are used to replace in part or completely sample surveys.

More in general, the issue of integrating administrative and sample sources has emerged. Chambers et al. (2006) designate the model-based estimation approach as a natural framework for integrating sources in the statistical production. In this context, a proposed solution is fitting a model on the sample and applying it to predict values for units on the unobserved part of the sample using information obtained from administrative data. The ESSnet on Administrative Data (http://www.cros-portal.eu/content/admindata-sga-3) reports various experimental applications of this approach. The module "Weighting and Estimation – Estimation with Administrative Data" describes practical uses of administrative data in business statistics and gives suggestions on which methods is more appropriate according to the informative context (timeliness and coverage) of the administrative source providing data.

### *2.10 Departure from ideal conditions: imperfect frames*

In this subsection some of the most typical departures from the ideal conditions are the basis of sampling and estimation methodologies are highlighted.

The most common case is when sample is selected from a not updated frame. In this context, it may occur that some values of the stratification variables used for sampling differ from those observed in the selected units. For instance, the observed enterprise size, measured as the number of employees, can change from the measure registered in the sample frame. When new totals are available, a post-stratified estimator can be applied in order to take into account of this updated information during the estimation phase. See the module "Weighting and Estimation – Calibration" for more details on how the post-stratification estimator can be applied.

A second important concern arises due to demography of the statistical units (enterprises). Since business population experiences rapid changes, the sampling frame is affected by some degree of overcoverage and undercoverage, i.e., units in the frame are no longer in the target population or vice versa. For units not covered by the frame (undercoverage) there is a zero probability, and this feature may cause biased estimators. See Lundström and Särndal (2001, page 139) for a formalisation of the context and for possible solutions, in particular in Section 11.3 the calibration approach is described when updated known totals can be used at estimation stage.

Finally, an important issue arises when mergers and splits occurs between the construction of the frame and the surveys. In this case sampled units do not correspond to units recorded in the sampling frame. This population dynamics affects the sampling inclusion of the final units and may introduce bias in the final estimates, if it is not properly taken into account. This context can be formalised with indirect sampling (Lavallée, 1995, Deville and Lavallée, 2006). In fact, the sampling step is carried out on a frame not containing the target population units but linked to them. The major difficulty with this approach consists in recognising the links between sampling list and target population produced by mergers or splitting and in determining the correct weights.

The Generalised Weight Share Method (GWSM) has been developed by Lavallée (1995) and Deville and Lavallée (2006). Lavallée and Labelle-Blanchet (2013) present the method for skewed populations. The Swiss Federal Statistical Office applied the weight share method for the estimations of the Quarterly Job Statistics:

http://www.bfs.admin.ch/bfs/portal/fr/index/infothek/erhebungen__quellen/methodenberichte.html?publicationID=3217%20.

More details on imperfect frames are reported in the module "Weighting and Estimation – Design of Estimation – Some Practical Issues".

## 3. Design issues

The choice of estimations methodology is strictly related to the main aspects of quality described in the module "Quality Aspects – Quality of Statistics". The main features are accuracy, coherence, timeliness. Choice of estimation methodology is also highly related to characteristics of the sampling design (e.g., probabilistic or cut-off sampling). Here we give a brief summary of the quality and sampling factors to be considered. First of all, one should determine if administrative data are available, accessible and can be used for direct production of statistical output according to schema in the module "Weighting and Estimation – Estimation with Administrative Data".

Whenever administrative data are not available, and sampling is carried out to achieve the required information, in order to choose the proper estimation methodology, one should take into account of the sampling mechanism. If a non-random mechanism is applied, a model-based estimator can be applied. However, the risk of this sampling strategy is that when the model is not valid bias may be present.

On the other hand, bias can be present, also in case of a probabilistic sampling design, when there are no-respondents and no proper measure are taken, for example, by means of adjustment of weights.

Moreover, when external constraints are given then benchmarking to these external constraints can be obtained with calibration estimators and reweighting. This solution is not always applicable in practice and problems can be encountered in practice to meet too many constraints.

Similarly, to improve accuracy of estimators, in particular to reduce variance one may want to use of auxiliary information correlated with the target (e.g., calibration estimators). However, in practice many target indicators are produced by a survey and a single weight is used for a single survey. Then, the auxiliary variables of the calibration estimator or GREG estimator will likely relates only with one (or few) of the target variables. Another class of methods to (further) improve accuracy when large variability of HT (or calibration) estimators is caused by skewness of the distribution of the target variable is robust estimation.

If calibration estimators still does not satisfy the needed accuracy level (this will often occur with unplanned domain, but also in case of large no-response), small area estimators may represent a possible solution to guarantee the desired degree of information.

Similarly, if there is a need to obtain estimates with incomplete sample observations to meet timeliness, preliminary estimators may be applied.

Many of the aspects recalled in this section, conflict with each other and compromise solutions should be considered between the different competing needs, aiming at guaranteeing the quality of the estimates.

The module "Weighting and Estimation – Design of Estimation – Some Practical Issues" provides more details on practical issues to be considered in designing the estimation methodology.

## 4. Available software tools

There are several software tools to perform estimation using basic weights or calibration estimators together with variance estimation (see the topic "Quality Aspects"). In the following we classify some of them requiring open source R or the commercial software SAS and SPSS.

The following packages R are available from the R-CRAN archives:

*Package survey,* http://cran.r-project.org/web/packages/survey/index.html

*Package sampling,* http://cran.r-project.org/web/packages/sampling/index.html

A full-fledged R system for design-based and model-assisted analysis of complex sample surveys *REGENESEES* is available at http://joinup.ec.europa.eu/software/regenesees/release/all

The following programs allows to calibrate weights and calculate variance estimation:

*BASCULA* http://www.cbs.nl/en-GB/menu/informatie/onderzoekers/blaise-software/blaise-voor-windows/productinformatie/bascula-info.htm

*CALMAR is a SAS macro developed by the French National Statistics Office (INSEE),*

*CLAN* is a system of *SAS* macros developed by Statistics Sweden.

*GENESEES (SAS macro by Italian statistical Institute* http://www.istat.it/it/strumenti/metodi-e-software/software/genesees*)*

*GES, developed in Statistics Canada, is also a system of SAS macros*

*g-Calib (SPSS by Statistics Belgium)*

See Eurostat (2013) for further details on these software tools.

Many tools are available to perform small area estimation methods as well

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. the R functions produced by ESSnet SAE (ESSnet/sae portal http://www.cros-portal.eu/sites/default/files//R_codes_%26_documentations_3.zip
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

A description of functions and software for small area estimation can be found in WP4 final report of ESSnet SAE.

## 5. Decision tree of methods

## 6. Glossary

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

## 7. References


Australian Bureau of Statistics (2000), Business Longitudinal Survey, Confidentialised Unit Record File, 1994-95, 1995-96, 1996-97, 1997-98.

Beaumont, J.-F. and Alavi, A. (2004), Robust generalized regression estimation. *Survey Methodology* **30**, 195–208.

Beaumont, J. F. and Rivest, L. P. (2008), Dealing with Outliers in Survey Data. In: C. R. Rao and D. Pfeffermann (eds.), *Handbook of Statistics, Design, Methods and Applications*, Vol. 29.

Benedetti, R., Bee, M., and Espa, G. (2010), A Framework for Cut-off Sampling in Business Survey Design. *Journal of Official Statistics* **26**, 651–671.

Berger, Y. G. (2004), Variance estimation for measures of change in probability sampling. *The Canadian Journal of Statistics* **32**, 451–467.

Chambers, R. L. (1986), Outlier robust finite population estimation. *Journal of the American Statistical Association* **81**, 1063–1069.

Chambers, R. L. and Kokic, P. (1993), Outlier robust sample survey inference. Invited paper, *Proceedings of the ISI 49th session, Firenze, Italy, August 1993*, 55–72.

Chambers, R., van den Brakel, J. A., Hedlin, D., Lehtonen, R., and Zhang, L.-C. (2006), Future Challenges of Small Area Estimation. *Statistics in Transition* **7**, 759–769.

Daas, P. and Ossen, S. (2011), List of quality groups and indicators identified for administrative data. Delivarable 4.1, FP7 BLUE-ETS project.

Deville, J.-C. and Lavallée, P. (2006), Indirect sampling: The foundations of the generalized weight share method. *Survey Methodology* **32**, 165–176.

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

Duncan, G. and Kalton, G. (1987), Issues of Design and Analysis of Surveys across Time. *International Statistical Review* **55**, 97–117.

Ferrante, M. R. and Pacei, S. (2004), Small Area Estimation for Longitudinal Surveys. *STATISTICAL METHODS & APPLICATIONS* **13**, 327–340.

Gross, W. F., Bode, G., Taylor, J. M., and Lloyd-Smith, C. W. (1986), Some finite population estimators which reduce the contributions of outliers. *Pacific Statistical conference: Proceedings of the congress*, Auckland, New Zealand, 20–24.

Eltinge, J. L. (2011), Two approaches to the use of administrative records to reduce respondent burden and data collection costs. UNECE. http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.42/2011/mtg1/USA_TwoAapproaches.pdf

EUROSTAT (2013), *ESS Handbook on Precision Requirements and Variance Estimation for Household Surveys*. Methodologies and working papers.

Fabrizi, E., Ferrante, M. R., and Pacei, S. (2007) Small area estimation of average household income based on unit level models for panel data. *Survey Methodology* **33**, 187–198.

Hidiroglou, M. A. and Srinath, K. P. (1981), Some estimators of population total from simple random samples containing large units. *Journal of the American Statistical Association* **76**, 690–695.

Houbiers, M., Knottnerus, P., Kroese, A. H., Renssen, R. H., and Snijders, V. (2003), Estimating consistent table sets: position paper on repeated weighting. Discussion paper 03005, Statistics Netherlands, Voorburg / Heerlen. http://www.cbs.nl/NR/rdonlyres/6C31D31C-831F-41E5-8A94-7F321297ADB8/0/discussionpaper03005.pdf

Huber, P. J. (1981), *Robust Statistics*. John Wiley & Sons, New York.

Hulliger, B. (1995), Outlier robust Horvitz-Thompson estimators. *Survey Methodology* **21**, 79–87.

Kalton, G. (1983), Models in the Practice of Survey Sampling. *International Statistical Review* **51**, 175–188.

Kokic, P. N. and Bell, P. A. (1994), Optimal winsorizing cutoffs for a stratified finite population estimator. *Journal of Official Statistics* **10**, 419–435.

Knotterus, P. and van Delden, A. (2012), On variances of changes estimated from rotating panels and dynamic strata. *Survey Methodology* **38**, 43–52.

Laitila, T., Wallgren, A., and Wallgren, B. (2011), *Quality Assessment of Administrative Data*. Research and Development – Methodology reports from Statistics Sweden, 2.

Lavallée, P. (1995), Cross-sectional weighting of longitudinal surveys of individuals and households using weight share method. *Survey Methodology* **21**, 25–32.

Lavallée, P. and Labelle-Blanchet, S. (2013), Indirect sampling applied to skewed populations. *Survey Methodology* **39**, 183–215.

Lee, H. (1995), Outliers in business surveys. In: B. G. Cox, D. A. Binder, B. N. Chinnappa, A. Christianson, M. J. Colledge, and P. S. Kott (eds.), *Business Survey Methods*, John Wiley & Sons, New York.

Little, R. J. A. and Rubin, D. B. (2002), *Statistical Analysis with Missing Data*, 2nd edition. Wiley, New York.

Lundström, S. and Särndal, C.-E. (2001), Estimation in the presence of Nonresponse and Frame Imperfections. Statistics Sweden.

Nordberg, L. (2000), On Variance Estimation for Measures of Changes When Samples are Coordinated by Use of Permanent Random Numbers. *Journal of Official Statistics* **16**, 363–378.

Pfeffermann, D. (2013), New Important Developments in Small Area Estimation. *Statistical Science* **28**, 40–68.

Qualité, L. and Tillé, Y. (2008), Variance estimation of changes in repeated surveys and its application to the Swiss survey of value added. *Survey Methodology* **34**, 173–181.

Rao, J. N. K., Srinath, K. P., and Quenneville, B. (1986), Estimation of Level and Change using Current Preliminary Data. In: Kasprzyk, Duncan, Kalton, and Singh (eds.), *Panel Surveys*, John Wiley & Sons, New York, 457–485.

Rao, J. N. K. (2003), *Small Area Estimation*. John Wiley and Sons, New York.

Renssen, R. H. and Nieuwenbroek, N. J. (1997), Aligning Estimates for Common Variables in Two or More Sample Surveys. *Journal of the American Statistical Association* **92**, 368–374.

Royall, R. M. (1970), On finite population sampling theory under certain linear regression models. *Biometrika* **57**, 377–387.

SAE ESSnet (2012), Deliverables of the project. http://cros-portal.eu/projectdetail/1392.

Särndal., C.-E. and Lundström, S. (2005), *Estimation in Surveys with Nonresponse*. John Wiley and Sons, New York.

Särndal, C.-E., Swensson, B., and Wretman, J. H. (1992), *Model Assisted Survey Sampling*. Springer Series in Statistics, Springer-Verlag, New York.

Vaillant, R., Dorfman, A. H., and Royall, R. M. (2000), *Finite Population Sampling and Inference, a Prediction Approach*. Wiley, New York.

Verma, V., Betti, G., and Ghellini, G. (2006), Cross-sectional and longitudinal weighting in a rotational household panel: applications to EU-SILC. Working paper, 67, University of Siena, available at http://www.econ-pol.unisi.it/dmq/pdf/DMQ_WP_67.pdf.

Wallgren, A. and Wallgren, B. (2007), *Register-based Statistics – Administrative Data for Statistical Purposes*. John Wiley & Sons, Chichester, England.

## Interconnections with other modules

### 8. Related themes described in other modules

1. User Needs – Specification of User Needs for Business Statistics
2. Repeated Surveys – Repeated Surveys
3. Dynamics of the Business Population – Business Demography
4. Sample Selection – Main Module
5. Statistical Data Editing – Main Module
6. Imputation – Main Module
7. Weighting and Estimation – Design of Estimation – Some Practical Issues
8. Weighting and Estimation – Small Area Estimation
9. Weighting and Estimation – Estimation with Administrative Data
10. Quality Aspects – Quality of Statistics

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

1. Sample Selection – Subsampling for Preliminary Estimates
2. Weighting and Estimation – Calibration
3. Weighting and Estimation – Generalised Regression Estimator
4. Weighting and Estimation – Outlier Treatment
5. Weighting and Estimation – Preliminary Estimates with Design-Based Methods
6. Weighting and Estimation – Preliminary Estimates with Model-Based Methods
7. Weighting and Estimation – Synthetic Estimators for Small Area Estimation
8. Weighting and Estimation – Composite Estimators for Small Area Estimation
9. Weighting and Estimation – EBLUP Area Level for Small Area Estimation (Fay-Herriot)
10. Weighting and Estimation – EBLUP Unit Level for Small Area Estimation
11. 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. 5.5 Calculate weights
2. 5.6 Calculate aggregates

## 12. Tools explicitly referred to in this module

1. Software tools for estimation, calibration of weights, variance estimation, application of small area methods

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

1. Sampling, Estimation and Evaluation of Accuracy

# Administrative section

## 14. Module code

Weighting and Estimation-T-Main Module

## 15. Version history

| Version | Date | Description of changes | Author | Institute |
|---|---|---|---|---|
| 0.1 | 28-03-2012 | first version | Loredana Di Consiglio | ISTAT |
| 0.2 | 05-07-2012 | after first review | Loredana Di Consiglio | ISTAT |
| 0.2.1 | 03-09-2012 | new version to take into account the modified workplan | Loredana Di Consiglio | ISTAT |
| 0.3 | 14-05-2013 | revised version | Loredana Di Consiglio | ISTAT |
| 0.4 | 08-10-2013 | after review | Loredana Di Consiglio | ISTAT |
| 0.5 | 10-01-2014 | after CH-HU-SE reviews | Loredana Di Consiglio | ISTAT |
| 0.6 | 24-02-2014 | after HU-SE-EB reviews | Loredana Di Consiglio | ISTAT |
| 0.6.1 | 28-02-2014 | minor revisions | Loredana Di Consiglio | ISTAT |
| 0.6.2 | 11-03-2014 | preliminary release | | |
| 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 13:31 |