Weighting and estimation methods: description in the Memobust handbook

Loredana Di Consiglio, Fabrizio Solari,
Department of Integration, Quality, Research and Production Networks Development
Istat - Italian National Institute of Statistics
diconsig@istat.it, solari@istat.it

1.Introduction

The Memobust project aims at identifying best practices and at developing common methodology for designing and conducting statistical business surveys. These objectives are supported by guidelines consisting in an update of an existing handbook on methodology and design of business statistics (Willeboordse, 1998). The need for a new handbook has arisen both to include new issues in statistical methodology for business statistics and to make more flexible the existing handbook. The new handbook is planned as a set of separate and interconnected documents serving as introductory, contextual or background material or describing specific methods.
It is primarily aimed at persons working at statistical institutes in the area of business statistics, in particular survey managers and statisticians involved in the production process. However, the handbook should particularly be useful for methodologists.
First, *Theme modules* give a general description of topics and subtopics. They may be mainly useful for statisticians and survey managers when dealing with the production process, since they are thought for introducing the reader to problems that may occur, moreover suggestions for handling them are given. Then, there are also more technical modules, which refer to the methods used in the context of each topic. This module typology, named *Method module*, can be useful for survey managers who wish to deep the knowledge, and for methodologists that may find an updated state-of-the-art of methods on the issue.
These modules are conceptually connected within each topic and between topics. At the same time, they are conceived as self-contained, so that the reader can access directly to the module describing the argument she/he is interested without the need of a preliminary reading of other sections.
The content of the handbook is roughly structured according to the components of Generic Statistical Business Process Model (GSBPM). A detailed description of the handbook and of the structure of the modules is given in Willenborg *et al.*, (2012). A complete overview of the planned contents of the handbook can be found on the project's website (http://www.cros-portal.eu/content/memobust).
In this paper we give an account of the structure of one of the topic of the handbook: *Weighting and estimation*, underlying the new content of the updated handbook under writing, with respect to the Handbook on the Design and Implementation of Business Surveys (Willeboordse, 1998).

2. Weighting and estimation methods

This section of the handbook is aimed to give an overview of the methods that can be used to provide estimates.
Standard methods such as weighting (HT, GREG or calibration) are described in specific method modules, and special focus is given to central issues such as *robust estimation*, i.e. methods to deal with *representative outliers* a common issue for highly skewed distribution as those encountered in business surveys, *preliminary estimation*, i.e. estimation methods to deal with provisional estimates that have to be disseminated on the basis of sub-sample of the planned sample; *small area estimation*, i.e. methods to be applied when the sample size is not large enough to guarantee the release of direct estimates at the desired level of disaggregation;.

Special attention is also given to the use of administrative data in the estimation process.

2.1. 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, the implementation of probabilistic sample design is very common and a design weight equal to the inverse of the inclusion probability is associated to each unit.
The principle of weighting is also applied to account for unit non-response of sample units. The design weights can be adjusted also to consider non-response in order to reduce the possible bias of resulting estimates. Besides the modification of weights for handling with non-response, weights adjustment may also be carried out to take into account of auxiliary information, for example by means of the calibration estimator (Deville and Särndal, 1992) and GREG estimator (Särndal, *et al*. 1992) or to insure coherence among estimates of different sample surveys. Indeed, when good covariates are available, some improvement in the precision of estimates may be achieved by exploiting the relationship between target variable and extra information.
A method for using auxiliary information is by calibration: the weights are adjusted so that applying the estimators on the auxiliary variables, one is able to reproduce the known totals. Calibration includes well-known estimators such as the regression, the ratio and the raking-ratio estimators (Deville and Särndal, 1992).
In the handbook the main theme module introduces weighting method; moreover two specific method modules on GREG and calibration estimators are presented.

2.2 Robust estimation

In business surveys, 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 to be considered as gross errors. The handbook describes statistical methods to handle these units, as representative outliers affect the variability of the HT or GREG estimators. In particular a module describing the winsorization method is released.
Winsorization consists in modifying the outlying observations so that they have less impact on the estimation. In particular sample observations whose values lie outside certain preset cut-off values are set equal to the cut-off (type I winsorization) or are transformed as a linear combination of the observed value and the cut-off (type II winsorization) with coefficients for the observed values equal to the inverse of the sampling weights.
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 dependant on the goodness of cut-off(s) choice.
One specific module introducing robust estimation and the most classical approaches (such as weight trimming) and describing in details the winsorization method is included in the handbook.

2.3 Preliminary estimates

Timeliness in disseminating the estimates is a very important aspect of quality of short term statistics and one of the main peculiarities for this type of surveys.
For short term statistics, in fact, it may occur that the planned sample is only partially observed when the estimates have to be disseminated. Preliminary or provisional 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 theoretical sample (TS) that is observed at time of first release of the estimates.
The main problem that has to be faced off in a short-term preliminary estimation context concerns the possible self-selection of early respondents, since self-selection can lead to biased estimates 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. at sampling design stage, by selecting a preliminary subsample of the final theoretical sample, TS; this method is described in the handbook within the topic X. *Sample Selection*
2. at the estimation stage, in the following ways:
    a) by means of imputation techniques of missing data, that are applied to non respondent units in TS 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 in TS;
    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.

The different approaches can be compared in terms of bias and revision error, i.e. the difference between preliminary and final (with the complete theoretical sample) estimates.
The module *Preliminary estimation with design-based methods* focuses on a design based estimator. In particular describes a method proposed in Rao *et al*. (1989) which at time t exploit time t and t-1 data aiming at minimizing the mean square error of the estimate.
Rao *et al*. (1989) propose a basic composite estimator, that is obtained as weighted average of the preliminary estimate and the final estimate of the previous time adjusted for the difference of preliminary estimates of current time and the previous one.
The module *Preliminary estimation with model-based methods* focuses on a model based estimator proposed by Rao *et al*. (1989).These models use disaggregated auxiliary information coming from survey data at previous times and/or administrative register data. For the methods in the latter class, 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.

2.4 Small area estimation

The aim of small area (domain) estimation methods is to produce reliable estimates 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 estimates[1] at the desired level of disaggregation. For instance, one is interested to the overall amount of industrial turnover for the whole population of business enterprises, and also to estimate 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).

[1] 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.

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 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 closed or similar with respect to structural characteristics to the small area of interest. In practice, an improvement in the efficiency of the estimates 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.
In the handbook the topic is introduced in the theme module *Small area estimation*. Four specific method modules describe design based and model based methods.
In particular *synthetic estimators* and *composite estimators*, *EBLUP area level* and *EBLUP unit level* are introduced.
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 macro data 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 on *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.

2.5 Use of administrative data in the estimation process

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.
Sample surveys are then part of a more complex system (where more sources and surveys are combined together) and they in some cases represent the supplementary data to adjust for data quality (see Eltinge, 2011) or a complement of administrative data when coverage issues arises.
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 (Daas *et al*. 2011, Laitila *et al*. 2011).
A module on *Estimation with administrative data* is planned in the handbook showing possible practical use of administrative data in business statistics suggesting alternative methods according to the informative context (timeliness and coverage) of the administrative source.

References

Chambers, R.L. (1986), *Outlier Robust Finite Population Estimation*, Journal of the American Statistical Association 81, 1063-1069

Daas *et al*. (2011) *List of quality groups and indicators identified for administrative data*, Delivarable 4.1, FP7 BLUE-ETS project.

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

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

ESSnet SAE (2012). Report on Workpackage 6 – Guidelines (contributors Istat (Italy), CBS (Netherlands), SSB (Norway), GUS (Poland), INE (Spain), ONS (United Kingdom), http://www.cros-portal.eu/sites/default/files//WP6-Report.pdf

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

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

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

Rao J.N.K., Srinath K.P., Quenneville B. (1989), *Estimation of Level and Change using Current Preliminary Data*, in Panel Surveys, (Kasprzyk, Duncan, Kalton G., Singh eds.), 457-485, John Wiley & Sons, New York

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

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

Willeboordse, A. (ed.) (1998), *Handbook on the Design and Implementation of Business Surveys*. Office for Official Publications of the European Communities, Luxembourg.

Willenborg, L., Scholtus, S., Van Delden, A.,(2012) *Development and Structure of the Memobust Handbook*, Proceedings of Q2012, Athens,