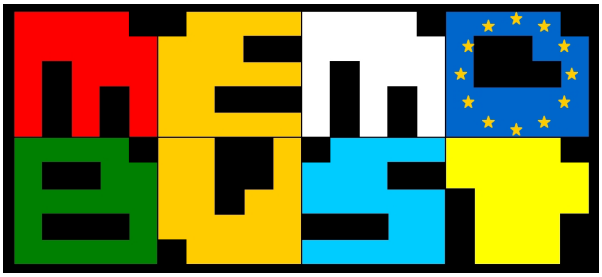

This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Method: Subsampling for Preliminary Estimates

Contents

General section

1. Summary

Among the main components of the quality in official statistics, the timeliness seems to be one of the most relevant both for producers and users of statistical data. In particular timeliness is becoming a pressing target especially for short term statistics (EUROSTAT, 2000). Therefore, in recent years, in many fields of official short term statistics the timeliness is becoming the driving issue, both for the increasing demand of users and the need to fill the gap comparing to data release standards already achieved by USA and other developed countries. The Amendment EU Regulation on Short Term Statistics (introduced in August 2005, EUROSTAT) requests all the statistical institutes of the EU Member States to transmit preliminary short term indicators to EUROSTAT with a reduced delay comparing to the timeliness set in the original 1998 Regulation. Frequently, in the NSIs short term statistics are based on fixed panel surveys of enterprises or rotating panels with a partial overlap from one year to another. Auxiliary variables coming from the previous survey occasions are often available.

A common approach for dealing with *preliminary estimates* focuses essentially on the study and the definition of efficient estimators, exploiting almost exclusively auxiliary information in the estimation phase. In such context sampling has a marginal role. Preliminary estimation merely involves the use of the quick respondent units. In fact, in order to obtain "good" preliminary estimates, standard survey strategy often aims to achieve high quick response rate by means of a well-structured plan of follow-up. In some surveys the "largest" units are carefully supervised. Following this approach, we point out that there is no explicit definition of sampling design for preliminary estimation, but that for the approach trying to observe large units. Hence, the preliminary estimates are usually drawn by a non-probabilistic sample design. A useful documentation on preliminary estimation problems (even though not comprehensive) can be downloaded from the OECD web site[1].

The topic investigates alternative sampling approaches for planning the subsamples for preliminary estimates. These designs try to exploit the auxiliary information in an efficient way according to the estimator used for the preliminary and final estimation. Therefore, an *overall strategy* for the production of preliminary estimates is developed, involving both the sample design and the estimator definition.

2. General description of the method

Given a sampling survey, a preliminary (or provisional) estimate is defined. It means the estimation of a parameter of interest obtained on the basis of a sample of quick respondent units available within a time lag Δ_t' after the reference time point (or end of the reference period) t of the survey, while the correspondent final estimate is based on both quick and late respondents (final sample), observed within a time lag Δ_t $(> \Delta_t')$. The indicators measuring the statistical quality of a preliminary estimation method are based on the differences between the two estimates. These differences are known as *revisions* or *revision errors*. For a detailed description of the indicators for statistical quality

[1] For the issue of the preliminary subsample the link is:
http://www.oecd.org/document/17/0,3746,en_2649_33715_30386193_1_1_1_1,00.html.

of preliminary estimations see, for instance, Di Fonzo (2005). The quick respondents can be observed according to different sampling processes. In particular we can observe the sample of quick respondents

(a) without any sampling and follow-up plans: we denote it as *Unplanned Preliminary Observed Sample* (UPOS);

(b) without any sampling plan but with a follow-up plan for the large final sampled units: we denote it as *Partially Unplanned Preliminary Observed Sample* (PUPOS);

(c) with a planned subsample for preliminary estimates. Then a *Preliminary Theoretical Sample* (PTS) is drawn and an intensive follow-up of the PTS units is planned so that *Planned Preliminary Observed Sample* (PPOS) will be as close to PTS as possible.

Before exposing the topic devoted to the sampling process (c), we make some general remarks:

1. preliminary estimation has two goals: producing accurate estimates of the parameters of interest; producing estimates with small revisions comparing to the final estimates. In some sense, this second goal can be more important for a NSI than the first one because the statistical users can compare the preliminary and final estimates and they can have a concrete perception of the sampling errors. Typical examples are the trend estimates where the preliminary and final estimates could have opposite signs, although the two estimation procedures can produce accurate and unbiased estimates;
2. the preliminary estimation issue arises also in surveys based on administrative data. Many aspects of the topic are suitable for such kind of surveys, but some others not. The topic does not tackle these peculiarities. Baldi et al. (2003) gives many interesting indications illustrating the problem and the possible solutions for the Employment, Wages and Labour Cost Survey conducted by the Italian NSI;
3. few references in literature on sampling design aiming at the preliminary estimation are available (cf. OECD link; D'Alò et al., 2007; Righi and Tuoto, 2007).

Regarding the definition of the preliminary samples we point out that:

4. sampling process (b) is a special case of sampling process (c);
5. as far as sampling processes (b) and (c) are concerned, we highlight that preliminary estimation has a distinctiveness with respect to the standard estimation process (producing the final estimates). The researcher using PUPOS or PPOS will obtain responses also from other sampled units of the final sample not included in the preliminary subsample and/or in the follow-up plan;
6. the sampling process (c) assumes that there is significant difference between early respondents and late respondents Then intensive follow-up of the PTS aims to survey all units that would belong to the two categories in a standard context. The small size of the PTS had to guarantee a small nonresponse rate;
7. using sampling process (c), the researcher can define an overall strategy taking into account the functional form of the parameters of interest (in general totals, indices or ratios) and the preliminary and final estimators. In an extensive vision of the problem,

the sampling design for PTS must be coordinated with the sampling design of the final sample and with the provisional and final estimation process. The ideal situation is to plan all these elements at the same time and, in practice, the two samples are coordinated according to an optimisation problem that takes into account of the trade-off between publishing early (risking a high revision error) and publishing late (which is not attractive) with a smaller risk for a high revision error. Nevertheless, the topic does not treat such huge context, but considers a restricted field typical for many sampling surveys. The provisional estimation goals and consequently the PTS are defined after the final sample and estimator were fixed, while the time lag of the provisional estimates is given by some legislative regulation on official statistics. In this case the possibility of defining different types of sampling strategy is restricted. If the strategy used for final estimation is optimal (within a given family of estimators and according to a design- or model-based approach), there is no particular reason for justifying the use of a quite dissimilar strategy for preliminary estimation. Secondly in order to reduce the revisions the form of the provisional estimator should be similar to the given final estimator.

The basic issue of using the PTS is the intensive follow-up that must be guaranteed for applying the method (point 6). If the PTS is affected by high non response, in general, small revisions cannot be obtained by letting the preliminary estimator resembling the final estimator. In this case it needs to define a specific provisional estimator following the approach typically used when the estimates are based on UPOS. An example of such approach is given by Rao et al. (1989). As far estimation process is concerned, the modules "Weighting and Estimation – Preliminary Estimates with Design-Based Methods" and "Weighting and Estimation – Preliminary Estimates with Model-Based Methods" shows some techniques. Here we pay the attention to the sampling phase.

Strategies for contact and follow-up of sampled units are dealt with in the module "Data Collection – Design of Data Collection Part 2: Contact Strategies".

2.1 The inferential approach

The sampling design for a PTS has to be defined according to the inferential approach. We distinguish two classical alternative inferential paradigms: the design-based/model-assisted and the model-based approaches (see also "Weighting and Estimation – Main Module"). The literature has studied the two approaches for the final estimates from different points of view and currently neither of them is dominant, although in the official statistics the design-based/model-assisted prevails. However, in the preliminary estimation context the reference framework is unlike from the context considering only the final estimate. We refer to the elements described in points 2.2 and 2.3, common in the preliminary estimation and missing in the final estimation process. Such elements can drive to prefer the model-based approach.

2.2 Sampling design for design-based/model-assisted approach

The PTS (selected from the final sample) has been drawn according to a random selection procedure. Standard sampling designs (simple random sampling, stratified simple random sampling, unequal probability sampling design etc.) can be implemented (see "Sample Selection – Main Module").

The choice of a sampling design depends on:

- the explicative power of the auxiliary variable (known at the design phase) for the variables of interest;
- the sampling design for the final estimates.

The second condition is established essentially to define a preliminary estimation process as similar as possible to the final estimation process. Then, if a stratified design is used for the final sample, in general it is better to use the same design for the PTS even though the stratification should be more aggregate because of a smaller sample size. The aim is to limit the revisions.

Advantages

The main advantages of a random PTS are the following:

- the inference of the design-based/model-assisted approach with the PTS does not suffer from bias;
- if the final estimates are design-based/model-assisted, it is better to use the same approach for the preliminary estimates in order to bound the revisions.

The most important condition for achieving these advantages is that the sampling design should be followed by a good follow-up plan for the preliminary and final sample. If the PTS and PPOS are quite dissimilar and the final sample has a high non response rate, the revisions can be very high and systematic, producing an highly undesirable upward or downward revisions in each survey occasion.

Disadvantages

The drawbacks of using a random subsample depend on whether the inferential approach is suitable. In fact, the preliminary estimation has a special parameter as the final estimate. Comparing the two estimates we obtain the revision. In the ideal situation, when the PPOS is the PTS and the final sample is fully observed, the revision represents simply estimate error. The inferential paradigm assures that under the preliminary sampling design the expected revision is zero. Nevertheless, such condition holds rarely and preliminary response and non-response for the units not belonging to the PTS have to be dealt with. However, there is a further complexity due to the variability of the final estimates because of final or late non-response. It means that the final estimates are random variables depending on the unknown non-response mechanism of the final sample. Since the inferential approach based on random sample does not cover the possibility of non-fixed parameters of interest (except for some special model assumptions), we have to use the model-based approach. In this case a non-random sample can be drawn.

2.3 Sampling design for model-based approach

The model-based approach does not require a random sample for making inference. The preliminary sample can be purposive, judgemental or non-random. On the other hand, the preliminary sample has to respect some features depending on the superpopulation model which generates the data for obtaining efficient estimates. In particular, it is important for the preliminary estimation to concentrate on the non-response mechanism. As mentioned earlier, in the real survey context the use of models in preliminary estimate problems is quite common because it is necessary to deal with the preliminary non-response, the preliminary response for the units not belonging to the PTS and the non-response of the final sample. We underline that in the last case, when the researcher has to deal with preliminary

estimation problem, he/she has to model the final non-response before observing it, in order to estimate the expected composition and size of the corresponding final sample.

The researcher usually does not know the non-response mechanisms, thus he/she has to make some assumptions defining the *working models*. Model-based approach makes inference on the working models, assuming that they represent satisfactory approximations of the true non-response mechanisms. Nevertheless, if the working models are seriously incorrect, the estimates can be strongly biased and the revisions can be systematically positive or negative. To avoid these problems, robust sampling strategies can be defined, in the sense that they perform well with the working and alternative models.

As far as the sample selection to protect from bias is concerned, we consider the balanced sampling design for drawing the PTS. Roughly speaking, in the model-based approach a sample is defined as balanced on a set of auxiliary variables if the sample and the known population means of the auxiliary variables are equal (Royall and Herson, 1973; Valliant et al., 2000). According to the considered estimator, different kinds of balanced samples can be used. Therefore, before defining the sampling design, the knowledge of the estimator form is necessary. The use of a balanced sample defines a *bias-robust* strategy.

The example of section 4 suggests how to implement a balanced sample for a real survey.

Advantages

If a model-based approach is used, a random or a purposive sample can be drawn. Nevertheless, there are some theoretical and operative advantages of using a purposive sample. From theoretical point of view:

- a suitable sample according to the working models used in the estimator can be drawn. Assuming that the working models hold, suitable purposive samples produce efficient estimates. When the researcher has no evidence that the working models represent satisfactory approximations, balanced samples produce robust estimates.

Considering the operative aspects, we refer to the short term statistics based on a panel component or longitudinal data. Frequently, these surveys rely on a set of sample units with high quick response rate achieved after a sensitisation work in the previous survey occasions. Then, in the perspective of renovating the preliminary sample with a non-random sample,

- it is easier to include the units that have shown high quick response probability in the preliminary subsample.

Disadvantage

The drawbacks to use a purposive or a non-random sample are linked to the inferential paradigm. If the working models are far from the actual non-response mechanisms, the inferences can be biased. On the other hand, model-based sampling theory suggests that it could be useful to select a random sample with this approach as well. Such samples could preserve the inference from the biasedness.

A second disadvantage can emerge when the final estimation is design-based/model-assisted. However, we remark that even with this approach the use of models is rather widespread.

3. Preparatory phase

4. Examples – not tool specific

4.1 Example: Comparison of different sampling designs for a preliminary subsample based on the Italian Monthly Retail Trade Survey data

The complexity of the preliminary estimation problem allows giving only a few general indications about the steps for defining a preliminary subsample for the provisional estimates. The following example shows how the process can be defined, but the main conclusion we want to highlight is that the preliminary sampling design has to take carefully into account the estimation process. The example is based on the data of the Italian Monthly Retail Trade Survey (MRTS), collected in 2004 (De Sandro and Gismondi, 2004).

4.1.1 Parameters of interest, preliminary and final estimates of the Italian Monthly Retail Trade Survey

The MRTS is based on the monthly measurement of the turnover of a stratified sample of retail enterprises (Division 52 of NACE nomenclature for a population of about 570 thousands) of different types and sizes. The sample is composed by a panel and a non-panel component, drawn every year and observed for 12 months. The survey provides provisional estimates within 30 days after the reference time and final estimates within 54 days according to the EU user needs (Eurostat, 2000). The provisional retail trade indices are referred to the domains: type of product sold (food and non-food retail enterprises) and type of distribution (large and small retail enterprises). Then the parameters of interest at the month *t* are defined as

$$I_d^{t,0} = \left(\sum_{h \in d} I_h^{t-12,0} R_h^t \gamma_h \right) \Big/ \sum_{h \in d} \gamma_h , \qquad \text{with} \qquad R_h^t = \frac{\sum_{i \in U_h^{t,t-12}} Y_i^t}{\sum_{i \in U_h^{t,t-12}} Y_i^{t-12}} ,$$

where *d* is the generic domain of interest; *h* is the generic stratum defined by the cross-classification of the main group of product sold, the class of employed persons and the type of distribution for 120 strata; $I_h^{t-12,0}$ is the retail trade index of the same month *t* of the previous year in the stratum *h* (with *t*=13, 14, …, 24)[2]; γ_h is a stratum weight given by the yearly turnover in 2000, derived from structural business statistics (ASIA archive); Y_i^t and Y_i^{t-12} are the total turnover variables of the unit *i* in month *t* and the same month of the previous year, respectively; $U_h^{t,t-12}$ is the longitudinal population of stratum *h* in time period (*t*, *t*–12). The product term ($I_h^{t-12,0} R_h^t$) represents the elementary index at stratum level.

[2] For instance January is indicated with t=13 and the same month of the previous year with t-12=1.

The sampling design is stratified simple random sampling, with about 7,500 units. Each year about 30% of the sample is renewed[3].

In the questionnaire of the reference month t both the values of the variables Y^t and Y^{t-12} are collected with some other auxiliary variables.

Starting at the end of 2004, the evaluation of the preliminary estimates is based on an UPOS calculated after Δ'_t=29 days from the end of the reference month. The estimation phase follows a complex procedure. Here the main steps, used in the simulation study, are sketched.

All the non-respondents within Δ'_t are imputed to obtain the provisional estimates. For each domain of interest the provisional estimation process is given by

$$\tilde{I}_d^{t,0} = \left(\sum_{h \in d} \tilde{I}_h^{t-12,0} \, \tilde{R}_h^t \gamma_h \right) \Big/ \sum_{h \in d} \gamma_h \,, \tag{1}$$

with

$$\tilde{R}_h^t = \frac{\sum_{i \in s_{ah(t)}^t} y_i^t + \sum_{i \in \left(\tilde{s}_h^t - s_{ah(t)}^t\right)} \tilde{y}_i^t}{\sum_{i \in s_{ah(t-12)}^t} y_i^{t-12} + \sum_{i \in \left(\tilde{s}_h^{t-12} - s_{ah(t-12)}^t\right)} \tilde{y}_i^{t-12}} \,, \tag{2}$$

where $\tilde{I}_h^{t-12,0}$ is the estimate of $I_h^{t-12,0}$; y_i^t and y_i^{t-12} are the observed values of Y_i^t and Y_i^{t-12}, $\tilde{y}_i^t$ and $\tilde{y}_i^{t-12}$ are the imputed values for the non-respondents; $s_{ah(t)}^t$ and $s_{ah(t-12)}^t$ are the respective sample units giving information for the preliminary estimates about the variables Y^t and Y^{t-12} in stratum h in month t with $s_{ah(t)}^t \subseteq \tilde{s}_h^t$ and $s_{ah(t-12)}^t \subseteq \tilde{s}_h^t$, where $\tilde{s}_h^t$ is the theoretical overall sample for the final estimates in stratum h in month t, while $\left(\tilde{s}_h^t - s_{ah(t)}^t\right)$ and $\left(\tilde{s}_h^t - s_{ah(t-12)}^t\right)$ are the corresponding non-respondent samples after the time lag Δ'_t for the variables Y^t and Y^{t-12}, respectively. We suppose that unit i providing the value of y_i^t gives information on y_i^{t-12} as well; then, $s_{ah(t)}^t$ and $s_{ah(t-12)}^t$ coincide and they are indicated by s_{ah}^t.

The imputation procedure is defined by two steps:

$$\tilde{y}_i^{t-12} = a_i^{t-12} \frac{\sum_{i \in s_{ag}^t} y_i^{t-12}}{\sum_{i \in s_{ag}^t} a_i^{t-12}} \,, \tag{3}$$

$$\tilde{y}_i^t = \frac{\sum_{i \in s_{ag}^t} y_i^t}{\sum_{i \in s_{ag}^t} y_i^{t-12}} \frac{a_i^t}{a_i^{t-12}} \tilde{y}_i^{t-12}, \tag{4}$$

[3] For practical reasons this percentage could be higher. For instance in 2004 data, analysed in the simulation study, about 50% of the sample belongs to the panel component (observed in the 2003 survey) while the other part is a new sample.

where $s_{ag}^t = \bigcup_{h \in g} s_{ah}^t$ represents the sample of the quick respondents of size n_{ag}^t belonging to the imputation cell *g* defined by crossing the type of distribution and the class of employed persons (8 cells, 3 for large and 5 for small retail enterprises); a_i^t and a_i^{t-12} are the numbers of persons employed in the respective months *t* and *t*–12 for the unit *i*, observed in the survey or imputed. Imputation is performed by the following procedure: the missing value of the variable a_i^{t-12} is imputed by the value a_i^t if it is not missing, otherwise it is imputed by the value of the business register; before imputing a_i^t, the outlier values considering the ratio a_i^t / a_i^{t-12} are checked. If the ratio does not belong to the interval (0.1, 10), the value a_i^t is replaced by a_i^{t-12}. If a_i^t is missing, it is imputed by a_i^{t-12}. In the expressions (3) and (4) we ignore this imputation process and always consider these two variables as observed. The final estimation has the same steps as the preliminary one, working with the information of both the quick and late respondents.

Finally, let us note that, although a probabilistic sample is used and the numerator and denominator of (2) are Horvitz-Thompson estimates with the imputation of the missing values, the sampling weights are annulled. These estimates can be analysed in the model-based context as well.

4.1.2 Definition of the Preliminary Theoretical Sample in the MRTS

The task of planning a subsample integrated with the provisional estimator, defining an overall preliminary sampling strategy needs to give an explicit form of the model of the imputation procedure. In the MRTS the procedure is quite complex. To keep things simple, we consider only the imputation processes defined in (3). In the model-based approach the process is the Best Linear Unbiased provisional estimator with respect to the final estimates if and only if the following superpopulation model generating the data is:

$$\begin{cases} Y_i^{t-12} = \beta_g a_i^{t-12} + \varepsilon_i \qquad (i \in g), \\ E(\varepsilon_i) = 0;\ E(\varepsilon_i, \varepsilon_j) = \begin{cases} \sigma^2 a_i^{t-12} & \text{if } i = j, \\ 0 & \text{else.} \end{cases} \end{cases} \tag{5}$$

Denoting by $\tilde{T}_{Y(d)}^{t-12}$ the provisional estimate of $\hat{T}_{Y(d)}^{t-12}$, the final estimate of the total of the variable Y_d^{t-12} with the final theoretical sample, if (5) is the true model, the expected revision $(\tilde{T}_{Y(d)}^{t-12} - \hat{T}_{Y(d)}^{t-12})$ is zero and it has the smallest variance under (3).

The second imputation step (4) cannot be expressed in a linear superpopulation model. Nevertheless, a reduction of the imputation error in the first step is important for a "good" imputation in the second one.

Under the working model (5), the optimal sampling strategy requires that the quick respondent sample is given by the units whose a^{t-12} values are the largest (Royall and Herson, 1973). However, the (5) is just a working model that likely will be different from the true superpopulation model. When (5) is wrong, selection of the largest units produces quite biased estimates.

In the example we have compared some alternative sampling designs for selecting a preliminary subsample in a simulation. A detailed description of the simulation comparing different preliminary

sampling strategies (preliminary subsamples and preliminary estimators) are given by Righi and Tuoto (2007). In particular, we focused on the selection of balanced sampling, which allows to plan a bias-robust strategy against the model failure (Valliant *et al.*, 2000). We consider two different balanced sampling designs. The first design uses the following balancing equations:

$$\sum_{s_{ag}^t} \frac{\left(a_i^{t-12}\right)^j}{n_{ag}^t} = \sum_{\tilde{s}_g^t} \frac{\left(a_i^{t-12}\right)^j}{\tilde{n}_g^t} \qquad \text{(j=1, 2, …, J),} \tag{6}$$

where $\tilde{n}_g^t$ is the size of $\tilde{s}_g^t \left(= \bigcup_{h \in g} \tilde{s}_h^t\right)$.

The strategy defined by (3) and (6) is called as Simple Balanced (SB) strategy.

The second balanced design tries to satisfy the weighted balancing equation

$$\frac{1}{n_{ag}^t} \sum_{i \in s_{ag}^t} \frac{a_i^{t-12}}{\sqrt{a_i^{t-12}}} = \frac{\sum_{i \in \tilde{s}_g^t} a_i^{t-12}}{\sum_{i \in \tilde{s}_g^t} \sqrt{a_i^{t-12}}}, \tag{7}$$

defining a weighted balanced sample (Royall, 1992). We call this strategy as Weighted Balanced (WB) strategy. Royall (1992) and Valliant et *al.* (2000) give a deeper description of the two strategies and their properties related to the true and working superpopulation model. Here we only highlight that the imputation process (3) becomes more robust with the balanced sampling, even though the variance of the estimator increases with respect to the strategy selecting the largest sampling unit.

4.1.3 Results of the simulation

In order to carry out the simulation study, an artificial sample based on the observed final sample of 2004 has been arranged (Righi and Tuoto, 2007). The main aim of the artificial sample is to make the complete set of data available in terms of target variables and covariates, for all the 7,448 units in the final sample. Starting from the complete data set, denoted as pseudo-sample, the properties of the proposed sampling strategies have been studied in a simulative context. For each strategy 500 PTS, each one with 1,920 units, as recommended by EUROSTAT (2001), have been selected from the pseudo-sample. At each iteration the preliminary estimates are computed for the domains and the revisions are calculated comparing to the final pseudo-sample estimates.

Table 1 shows the UPOS monthly sample size distribution. We observe an average of about 2,340 units, with a maximum value equal to 2,607 and a minimum value equal to 2,068 units.

Table 1. Monthly UPOS dimensions

Month	1	2	3	4	5	6	7	8	9	10
Sample size	2,112	2,302	2,275	2,385	2,384	2,482	2,348	2,332	2,607	2,068

The experiment compares the results coming from the proposed sampling strategies, hereinafter the balanced strategies, with both the estimates based on the UPOS and the estimates obtained by the sample of 1,920 units of the largest retail enterprises in terms of turnover or number of employed persons. The largest enterprises samples may be considered as cut-off sampling (cf. “Sample Selection

– Main Module"), which is frequently used in short terms statistics. The last two samples are allocated according to the same technique defining the allocations of the balanced PTS. These three strategies represent the benchmark of the balanced strategies. The balanced samples have been selected by means of the Cube algorithm (Deville and Tillé, 2004).

For evaluating the performances in term of revision, the monthly *Mean Percentage Revision* (MPR) has been computed according to the expression

$$MPR_D^{t,0} = \sum_{d \in D} \left[\left(\frac{\bar{\tilde{I}}_d^{t,0} - \hat{I}_d^{t,0}}{\hat{I}_d^{t,0}} \right) \times 100 \right] \gamma_d \,, \tag{8}$$

where $\hat{I}_d^{t,0}$ is the final estimate for the more disaggregated domain *d* (Large-non-food; Small-non-food; Large-food; Small-food) in month *t*, and $\bar{\tilde{I}}_d^{t,0}$ assumes one of the following values:

- $(1/500)\sum_r \tilde{I}_{d,r}^{t,0}$ with the balanced strategies, $\tilde{I}_{d,r}^{t,0}$ being the provisional estimate on the domain *d* in the *r*-th replication;
- $\bar{\tilde{I}}_d^{t,0} \equiv \tilde{I}_d^{t,0}$ considering the benchmark strategies.

Finally, *D* indicates the generic domain at the more disaggregate level ($d \equiv D$) or at aggregate level (Non-food, Food, Large, Small, Total), and $\gamma_d = \sum_{h \in d} \gamma_h$.

The yearly version of (8) is given by

$$MPR_D = \frac{1}{12} \sum_{t=13}^{24} MPR_D^{t,0} \,. \tag{9}$$

A second type of indicators measures the variability of the estimates by means of the *Mean Absolute Percentage Revision* (MAPR). At monthly level it is defined by

$$MAPR_D^{t,0} = \sum_{d \in D} \left[\left| \frac{\bar{\tilde{I}}_d^{t,0} - \hat{I}_d^{t,0}}{\hat{I}_d^{t,0}} \right| \times 100 \right] \gamma_d \,. \tag{10}$$

We prefer to use the expression (10) for the balanced strategies instead of a more appropriate indicator using the term $(1/500)\sum_r \left| \left(\tilde{I}_{d,r}^{t,0} - \hat{I}_d^{t,0} \right) / \hat{I}_d^{t,0} \right| \times 100$ in the square brackets, since we observed only one preliminary sample for the benchmark strategies. Therefore, in the balanced strategies this alternative indicator catches the variability due to the iterations, not detectable in the benchmark strategies. The yearly MAPR is

$$MAPR_D = \frac{1}{12} \sum_{t=13}^{24} MAPR_D^{t,0} \,. \tag{11}$$

We point out that the monthly MPR and MAPR give rough measures especially for the benchmark strategies, because they are computed for few values and with only one value for the more disaggregate domains. Hence, we show the results of the statistics (9) and (11). The exhaustive description of the simulation results is given in Righi and Tuoto (2007).

Table 2.a shows the values of the statistics (9) for the preliminary domain estimates given by crossing the variables type of sold product (food and non-food retail enterprises) and type of distribution (large and small retail enterprises).

Table 2.a. Yearly Mean Percentage Revision (MPR) by Type of sold product and Type of distribution domains

Method	Large non-food	Small non-food	Large food	Small food
Strategy using UPOS	0.674	0.236	0.505	0.021
Largest Units in terms of Employed Persons (LUEP) strategy	1.606	-0.139	-0.070	-0.592
Largest Units in terms of Turnover (LUT) strategy	0.977	0.126	0.359	-0.309
Simple Balanced (SB) strategy	0.555	0.006	0.733	-0.241
Weighted Balanced (WB) strategy	0.639	-0.077	0.197	-0.303

The table underlines that the balanced approaches using a PTS have, in general, better performances than the benchmark strategies. Especially the WB seems to be the best. The benchmark strategies present a MPR less than the WB strategy only in two cases: for large-food domain the *Largest Units in terms of Employed Persons* (LUEP) strategy has MPR = –0.070, while the WB strategy has MPR = 0.197, and for the small-food domain, where the strategy based on UPOS has MPR = 0.021, a value closer to zero than the value –0.303 of the WB strategy. The SB strategy has good performances except for the large-food domain with MPR = 0.733.

Table 2.b shows the MPR results for the aggregate domains. The findings must be analysed with caution because of the opposite signs of the MPR at the more disaggregate levels. "Good" results could actually hide an unstable strategy in term of unbiasedness, and this aspect must be taken into account in the conclusive evaluations. The WB strategy is the best method based on PTS, except for the case of small type of distribution with MPR = –0.109, while the SB strategy has MPR = –0.029. In the large domain, LUEP strategy is slightly better. Finally, for the total domain the strategy observing the LUEP has the best MPR with a value equal to –0.021. The WB strategy has MPR = 0.043.

Table 2.b. Yearly Mean Percentage Revision (MPR) by Type of sold product, Type of distribution and Total domains

Method	Type of sold product		Type of distribution		Total
	Non-food	Food	Large	Small	
Strategy using UPOS	0.293	0.396	0.540	0.205	0.334
Largest Units in terms of Employed Persons (LUEP) strategy	0.087	-0.188	0.272	-0.205	-0.021
Largest Units in terms of Turnover (LUT) strategy	0.236	0.209	0.486	0.063	0.225
Simple Balanced (SB) strategy	0.078	0.514	0.697	-0.029	0.250
Weighted Balanced (WB) strategy	0.016	0.084	0.287	-0.109	0.043

Table 3.a gives some findings about the variability of the compared strategies, computed by (11). The methods based on balanced PTS seem to have better performances, especially the WB strategy. The SB has the best MAPR for the large-non-food domain (0.998), but it has a high value for the large-food domain (0.840) with respect to some benchmark strategies. The *Largest Units in terms of Turnover* (LUT) sample strategies have the best results for small-non-food domain with MAPR = 1.001 and LUEP has MAPR = 1.143. The last strategy has the best performance also for the large-food domain (0.317). We note that when the WB has worse results than the largest units strategies, the values are close each other. On the other hand, when WB is the best strategy the MAPR is quite better than the benchmark strategies. Strategy based on UPOS, despite the greatest mean overall sample size, does not operate very well at least with respect to the balanced PTS strategies. Just for the small-non-food domain MAPR = 1.336, while the SB has MAPR = 1.369.

Table 3.a. Yearly Mean Absolute Percentage Revision (MAPR) by Type of sold product and Type of distribution domains

Method	Large non-food	Small non-food	Large food	Small food
Strategy using UPOS	1.493	1.336	1.093	2.091
Largest Units in terms of Employed Persons (LUEP) strategy	2.263	1.143	0.317	2.949
Largest Units in terms of Turnover (LUT) strategy	2.461	1.001	0.587	2.344
Simple Balanced (SB) strategy	0.998	1.369	0.840	2.038
Weighted Balanced (WB) strategy	1.118	1.322	0.392	2.040

For the aggregate domains (Table 3.b) the use of balanced PTS still leads to the best MAPR values. In a few cases the largest units strategies achieve lower values: for the non-food domain, where the LUT strategy is better than the SB strategy (1.191 vs. 1.195) and for the large domain, where the LUEP strategy (0.715) is better than the SB approaches with MAPR values greater than 0.77.

The results in this simulation study show that the WB even though it is not always the best strategy it does appear to be the strategy with the best overall performance.

Table 3.b. Yearly Mean Absolute Percentage Revision (MAPR) by Type of sold product, Type of distribution and Total domains

Method	Type of sold product		Type of distribution		Total
	Non food	Food	Large	Small	
Strategy using UPOS	1.356	1.317	1.175	1.444	1.341
Largest Units in terms of Employed Persons (LUEP) strategy	1.288	0.909	0.715	1.403	1.139
Largest Units in terms of Turnover (LUT) strategy	1.191	0.982	0.970	1.195	1.109
Simple Balanced (SB) strategy	1.195	0.685	0.771	0.881	0.618
Weighted Balanced (WB) strategy	1.168	0.659	0.467	0.840	0.506

5. Examples – tool specific

6. Glossary

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

7. References

Baldi, C., Ceccato, F., Congia, M. C., Cimino, E., Pacini, S., Rapiti, F., and Tuzi, D. (2003), Use of Administrative Data for Short Term Statistics on Employment, Wages and Labour Cost. *Proceedings of the "17th Roundtable on Business Survey frames"*, Rome, 26-31 October 2003. http://www.oecd.org/dataoecd/15/62/36232440.pdf

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984), *Classification and Regression Trees*. Chapman and Hall, New York.

D'Alò, M., De Vitiis, C., Falorsi, S., Righi, P., and Gismondi, R. (2007), Sampling Strategies for Preliminary Estimates Production in Short-Term Business Surveys. In: *Proceedings of the 2007 intermediate conference Risk and prediction*, Società Italiana di Statistica.

De Sandro, L. and Gismondi, R. (2004), Provisional Estimation of the Italian Monthly Retail Trade Index. *Contributi-Istat*, 24/2004.

Deville J.-C. and Tillé, Y. (2004), Efficient Balanced Sampling: the Cube Method. *Biometrika* **91**, 893–912.

Di Fonzo, T. (2005), The OECD project on revisions analysis: First elements for discussion. Paper presented at OECD STESEG meeting, Paris 27-28 June 2005. http://www.oecd.org/dataoecd/55/17/35010765.pdf

EUROSTAT (2000), *Short-term Statistics Manual*. Eurostat, Luxembourg.

EUROSTAT (2001), Conclusion of the First Meeting of the Export Group Contro-Stratified European Sample for Retail Trade, Final Report, July 2001. Eurostat, Luxemburg.

EUROSTAT (2005), Council Regulation No 1165/98 Amended by the Regulation No 1158/2005 of the European Parliament and of the Council – Unofficial Consolidated Version. Eurostat, Luxemburg.

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

Righi, P. and Tuoto, T. (2007), The planning of Preliminary Sample: methodological aspects and an application to the Italian Monthly Retail Trade Survey. In: *Rivista di Statistica Ufficiale* N. 2-3/2007.

Royall, R. and Herson, J. (1973), Robust Estimation in Finite Population. *Journal of the American Statistical Association* **68**, 880–889.

Royall, R. M. (1992), Robustness and Optimal Design Under Prediction Models for Finite Populations. *Survey Methodology* **18**, 179–185.

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

Specific section

8. Purpose of the method

Selection of a preliminary subsample from the sample is used for producing preliminary or provisional estimates. Different sampling designs are suggested according to the survey context. As far as short term statistics are concerned, where the timeliness is a pressing target and the estimation is based on panel or rotating panel, the *balanced sampling design* (see the module "Sample Selection – Balanced Sampling for Multi-Way Stratification") according to the model-based inferential framework is suggested for defining the preliminary sampling design.

9. Recommended use of the method

1. In general the method requires an intensive follow-up of the units belong to the preliminary subsample, so that the rate of quick non-response is low.
2. The method using the balanced sampling design to define the preliminary subsample exploits the time series data of the units in the panel sample.

10. Possible disadvantages of the method

1. The method has no particular theoretical disadvantage.
2. There can be operative disadvantages due to the implementation of the intensive follow-up for the preliminary subsampled units.

11. Variants of the method

1. n/a

12. Input data

1. Data including auxiliary variables for defining the sampling design.
2. In case of a balanced sampling design it is useful to collect data of the previous survey occasions for the panel units.

13. Logical preconditions

1. Missing values
    1. In practice, the method can be adapted to deal with missing values. If the non-response rate is too high it is needed to act in the estimation process.
2. Erroneous values
    1. In practice, the auxiliary variables will inevitably contain some errors. This is not ideal, but the method might still be useful in this case.
3. Other quality related preconditions
    1.
4. Other types of preconditions

1.

14. Tuning parameters

1. The method needs a careful tuning phase of the parameters. The study of the time series of the sampling data (observed in the previous survey occasions) allows to define the suitable sampling design.

15. Recommended use of the individual variants of the method

1. n/a

16. Output data

1. Sample membership indicator variable for the preliminary estimates is added in the input data set.

17. Properties of the output data

1.

18. Unit of input data suitable for the method

Processing full data set.

19. User interaction - not tool specific

1.

20. Logging indicators

1. No specific indicator.

21. Quality indicators of the output data

1. The main quality indicator is the revision, that is the difference between the final and the preliminary estimates.
2. In some survey occasion a simulation study such as the one described in Section 4 might also be used to obtain quality indicators.

22. Actual use of the method

1. Many NSIs base the preliminary estimates on a preliminary subsample. Because of the nature of the problem it means that an intensive follow-up is done for a subsample of the final sample.
2. Istat has used balanced sampling for the MRTS.
3. Balanced sampling that takes into account the estimation process such as in the topic has not been used yet.

Interconnections with other modules

23. Themes that refer explicitly to this module

1. Sample Selection – Main Module
2. Data Collection – Design of Data Collection Part 2: Contact Strategies
3. Weighting and Estimation – Main Module

24. Related methods described in other modules

1. Sample Selection – Balanced Sampling for Multi-Way Stratification
2. Weighting and Estimation – Preliminary Estimates with Design-Based Methods
3. Weighting and Estimation – Preliminary Estimates with Model-Based Methods

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

1. Regression

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

1. 4.1 Select sample

27. Tools that implement the method described in this module

1. Sampling package R
2. SAS Macro downloadable Insee site

28. Process step performed by the method

Sample planning and selection

Administrative section

29. Module code

Sample Selection-M-Subsampling

30. Version history

Version	Date	Description of changes	Author	Institute
0.1	19-03-2012	first version	Paolo Righi	ISTAT
0.2	24-04-2012	second version	Paolo Righi	ISTAT
0.3	18-05-2012	third version	Paolo Righi	ISTAT
0.4	15-04-2013	fourth version	Paolo Righi	ISTAT
0.4.1	06-09-2013	preliminary release		
0.4.2	09-09-2013	page numbering adjusted		
1.0	26-03-2014	final version within the Memobust project		

31. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 17:42