STATISTICA, anno LXVIII, n. 1, 2008

REDUCING REVISIONS IN SHORT-TERM BUSINESS SURVEYS

Roberto Gismondi[1]

1. TIMELINESS OF BUSINESS STATISTICS: CORE OF THE PROBLEM AND LATE DEVELOPMENTS

Many business surveys must strike a balance between timeliness and accuracy in the process of estimates' release. Estimates are generally required to be available soon after the reference period, in order to efficiently drive users and decisions makers. However, speedy delivery can adversely affect survey quality, as non-reporting tends to be higher with shorter collection periods.

Despite the issuance of revised estimates, preliminary estimates are most critical for use and tend to receive the most visibility: in particular, in the field of official statistics the EU Regulation on Short-Term Statistics (EUROSTAT, 2005) requests all the statistical institutes of the EU Member States to collect and transmit to EUROSTAT preliminary short-term indicators with a reduced delay: from 60 to 30 days for retail trade, from 90 to 60 days for the other services activities not including retail trade.

Since deviations between preliminary and revised estimates may be perceived as indicating an inability of the estimation methodology to appropriately correct for non-reporting, the main goal consists in reducing as much as possible the potential for large differences between preliminary and revised estimates. As a matter of fact, many political economy decisions are taken on the basis of short-term preliminary estimates (industrial production, prices, foreign trade), whose degree of precision is fundamental.

More in details, in the frame of a given business survey, we define as "preliminary quick estimate[2]" the estimation of a parameter of interest obtained on the basis of a quick sub-sample available at a time t' before time t correspondent to the "final estimate". This last estimate will be based on a final sample including

[1] The opinions herein expressed must be addressed to the author only, as well as possible errors or omissions. All tables and graphs derive from elaborations on ISTAT data.

[2] The terms "preliminary" and "provisional" are often exchangeable. However, in this context the term "preliminary" is preferred, since it more strictly underlines the need to calculate and release estimates before a *prefixed deadline* for final estimates.

both quick and late respondents. A revision can be calculated as the difference between final and preliminary estimates.

The identification of the optimal *preliminary* estimation and the optimal *final* estimation strategies are always strictly connected problems. Both of them could require the availability of one or more auxiliary variables for *all the units* in the population. Auxiliary variables can be used for the sampling design planning (stratification, evaluation of inclusion or response probabilities), in order to build up and test a super-population model and for carrying out estimates. In Italy, the most part of official short-term statistics are based on fixed panel of enterprises (monthly industrial production and turnover, monthly employment in large firms) or rotating panels with a partial overlap from one year to another (monthly retail trade, quarterly service activities). For all the enterprises in the population, structural variables as the number of persons employed and the yearly turnover – both referred to the last year – are available from the business register ASIA at the *single unit* level.

Generally speaking, if the strategy used for final estimation is optimal (inside a given family of estimators and according to a sampling or a model based approach), there is not a particular reason justifying the use of a different strategy for preliminary estimation. However, in the case of final non response the final sample can not be known in advance and when preliminary estimates are evaluated it could be necessary to estimate its final expected composition. A relevant exception is given by a census survey, when sooner or later all units should respond. If the strategy used for final estimation is not optimal – for instance, because when the survey was planned it was not possible to identify the true model underlying observed data, or variables in this model were not available, or a final non-response bias occurred – the apparent paradox is that the preliminary estimation technique that minimizes the average revision should not be optimal as well.

The non response bias could affect both final and preliminary estimates. The problem due to late respondents is a particular case of the non-response problem: in particular, quick and late respondents could follow different models (in terms of mean and/or variability). Late experiences (ISTAT, 2007) showed that in many empirical contexts referred to business data the non response bias is not systematic, but could happen for some survey occasions and/or for some domains only. The previously mentioned applications pointed out also the relatively poor efficiency of some traditional design based strategies for reducing non response bias. The estimation of individual response probabilities – useful to modify sampling weights of the ordinary Horvitz-Thompson estimator – is quite difficult because of the lack of enough reliable auxiliary variables (Rizzo *et al.*, 1996). The most part of imputation techniques that can be used in a short-term business survey – donor, regression and respondents' mean – normally do not reduce bias enough to balance the increase of variance due to the imputation process, unless it can be based on the propensity to respond (David *et al.*, 1983). These evidences stressed a wide recourse to a model based approach, as remarked in Cassel *et al.* (1983), Särndal *et al.* (1993), Valiant *et al.* (2000), Kalton (2002), Little and Rubin (2002), Särndal and Lundström (2005).

According to a super-population approach, the optimal estimation strategy is based on minimisation of the mean squared error (*MSE*) respect to the model underlying observed data. In a preliminary estimation context, it consists in a re-weighting process applied to respondent units. The main risk is due to the need to identify the right model, taking into account that in a given domain of interest more than one model could occur and when the (final) estimation strategy was established there could have not been enough information for a correct model identification. In a late work Hedlin *at al.* (2001) stressed the risk of additional bias due to a model miss-specification even when the asymptotically design unbiased *GREG* (generalised regression) estimator is used (Särndal *et al.*, 1993). These remarks underline the need to test model rightness.

Finally, a technical constraint is often the shortness of available time series of micro-data, because of the need to contain response burden and to adopt yearly rotations of sample units. That is the main reason why a time series approach – that assumes long time series and regularity along time of the error profile – does not seem the most appropriate for the problem under study, even though useful theoretical suggestions are available in Binder and Dick (1989) and Yansaneh and Fuller (1998). In particular, Rao *el al.* (1989) proposed a preliminary estimation technique that can be based on series of macro-data only.

According to a model based approach, in this context we propose and compare some quick model-based estimation techniques aimed at reducing the average revision in short-term business surveys, with an application to real business data. In details, in section 2 we present the general expression of the optimal preliminary predictor conditioned to the final prediction strategy under a model based approach. Even though the use of models in a provisional estimation context is not new (Copeland and Valliant, 2007; Gismondi, 2007*b*), herein the link between *final* and *provisional* estimation strategies has been analysed more in depth. In particular, on the basis of the general formula (8) we have assessed the formal relationship between *provisional* and *final* predictors; moreover, some particular cases and operational problems have been discussed as well. The presence of a potential non-response bias is faced in section 3, where a particular model-based post-stratification technique is presented, on the basis of a generalisation of a model already proposed in Gismondi (2007*b*). It must be remarked that this technique is not based on the same criteria on which the classical design-based post-stratification is founded (see, for instance, Kalton and Kasprzyk, 1986; Little, 1993). Moreover, that may be seen as an alternative respect to other empirical proposals aimed at evaluating self-selection bias (Billiet *et al.*, 2007), that normally require a huge amount of historical data and re-interviews of reluctant respondents. Finally, section 4 contains the main outcomes of an empirical attempt aimed at the estimation of changes along time of quarterly average turnover. That is based on the use of real data - referred to the period 2003-2006 - picked up in the frame of the quarterly wholesale trade survey currently carried out by ISTAT, with the goal of comparing ten provisional estimation strategies. Some conclusions have been drawn in section 5: among them, we mention the usefulness of the recourse to model based estimation strategies and to a post-stratification of available provisional data.

2. A GENERAL MODEL FOR DERIVING THE OPTIMAL PRELIMINARY PREDICTOR

Given a population U including N units, one supposes that each y-value of the variable y in the population (and in the observed sample) derives from the following general super-population model:

$$y_i = \beta x_i + \varepsilon_i \quad \text{where:} \quad \begin{cases} E(\varepsilon_i) = 0 & \forall i \\ VAR(\varepsilon_i) = \sigma^2 v_i & \forall i \\ COV(\varepsilon_i, \varepsilon_j) = 0 \quad if \quad i \neq j \end{cases} \tag{1}$$

where the x-values concern an auxiliary variable x available for *all* the units in the population and v is a variable determining y variability and to be specified. The parameters β and σ^2 are generally unknown. Even though a specific model (1) could be defined for different reference periods t, at the moment time labels are not necessary. The use of a univariate model through the origin is justified by the particular context under study. In the frame of short-term business surveys, it is not easy to find out more than one auxiliary variable measurable for all the units in the population and significantly correlated with that of interest. Moreover, a simple framework can better remark the link between the provisional and the final estimation strategies. Finally, recent works (Gismondi, 2007*b*) showed that the use of more than one auxiliary variable does not guarantee a significant reduction of revisions.

We suppose that the main purpose of the survey is the estimation of the unknown population mean $\overline{y}$, where $\overline{x}$ is the (known) population x-mean. A predictor T of the population mean is unbiased respect to (1) if $E(T - \overline{y}) = 0$. Let's also indicate as S the *final* sample (including n units), observed at the end of the response process referred to a given period t, and as S_P the *preliminary* sample (including n_P units), on the basis of which *preliminary* estimates are currently calculated and diffused. We can also write $S = S_P \cup S_L$, where S_L is the sub-sample including the $n_L=(n-n_P)$ *late* respondents. If $\overline{S}$ is the part of population not observed in the final sample, $\overline{S}_P$ is the part of population not observed in the preliminary sample, where $\overline{S} = \overline{S}_P \setminus S_L$.

2.1 *The optimal preliminary predictor conditioned to the final prediction strategy*

The general form of a whatever linear predictor used for the calculation of final estimates is:

$$T = \sum_S b_i y_i \tag{2}$$

where b_i are general coefficients applied to each observation belonging to the final sample S. The general expression of a linear predictor that can be used for calculating preliminary estimates is:

$$T_P = \sum_{S_P} a_i y_i \tag{3}$$

where coefficients a_i could be formally different from coefficients b_i. In particular, reasonable conditions that the preliminary predictor (3) should satisfy are the following ones:

$$E(T_P - T) = 0 \tag{4a}$$

$$E(T_P - T)^2 = \underset{\mathbf{a}_P}{Min} \tag{4b}$$

where $\mathbf{a}_P$ is the vector including the n_P coefficients a_i. The logic justifying the conjoint use of (4a) and (4b) is that, on the average, preliminary and final estimates should produce the same results and the variability of differences between preliminary and final estimates – e.g., the average magnitude of revisions – should be the lowest in the class of linear predictors. It is easy to verify that the condition (4a) is simply the request that the preliminary and the final predictors have the same model mean, but it is worthwhile to note that it does not imply that both predictors are unbiased respect to the model (1). Under condition (4a), the expectation in (4b) is equal to the variance of revisions, *given* the form of final predictor. One can write, adding and subtracting $E(T_P)$ and $E(T)$:

$$\begin{aligned} &E(T_P - T)^2 = E[(T_P - E(T_P)) - (T - E(T)) + E(T_P) - E(T)]^2 = \\ &= V(T_P) + V(T) - 2COV(T_P, T) + [E(T_P) - E(T)]^2 . \end{aligned} \tag{5}$$

It is clear from (5) that the condition (4a) for both the predictors implies that the last term in (5) is null. Furthermore, (5) could be minimised by a preliminary predictor that is biased under (1) if the final predictor T is biased itself. Considering condition (4a), taking into account that the last term in (5) does not depend on the choice of $\mathbf{a}_P$ and putting, according to the model (1):

$$V(T_P) = \sigma^2 \sum_{S_P} a_i^2 v_i \quad V(T) = \sigma^2 \sum_{S} b_i^2 v_i \quad COV(T_P, T) = \sigma^2 \sum_{S_P} a_i b_i v_i \tag{6}$$

one can minimise the following Lagrange function:

$$\Phi(\mathbf{a}, \lambda) = \sigma^2 \left(\sum_{S_P} a_i^2 v_i + \sum_{S} b_i^2 v_i - 2 \sum_{S_P} a_i b_i v_i \right) + \beta^2 \left(\sum_{S_P} a_i x_i - \sum_{S} b_i x_i \right)^2 + \lambda \beta \left(\sum_{S_P} a_i x_i - \sum_{S} b_i x_i \right). \tag{7}$$

Putting equal to zero the first derivatives of Φ respect to λ and each a_i, one gets the optimal solution:

$$T_P^* = \sum_{S_P} a_i^* y_i = \sum_{S_P} (b_i + \gamma_{Pi}^*) y_i \quad \text{where:} \quad \gamma_{Pi}^* = \left(\frac{x_i}{v_i} \right) \left(\sum_{S_L} b_i x_i \right) \left(\sum_{S_P} x_i^2 / v_i \right)^{-1} . \quad (8)$$

One must remark that this result can be also seen as a particular case of a more general result due to Valliant *et al.* (2000, 29)[3].

The expression (8) can be correctly evaluated *only* if late respondents are known in advance: that is guaranteed in a census surveys; however, in a sampling context, if final respondents do not correspond to the theoretical sample, one has to estimate which late respondents the final estimate will be based on (section 5). Moreover, the variance structure determines the general form of the optimal preliminary predictor (and, generally, of the final one as well). The solution (8) leads to an expected squared average revision given by:

$$E(T_P^* - T)^2 = \sigma^2 \left[\sum_{S} b_i^2 v_i + \sum_{S_P} (\gamma_{Pi}^{*2} - b_i^2) v_i \right] . \quad (9)$$

Given the form of the final predictor, the previous relation can be seen as the lowest expected revision in the class of predictors defined by (4a) and (4b) and under model (1), useful as benchmark values respect to which revisions currently calculated can be referred and evaluated.

2.2 *General formulas and particular cases*

If the final predictor is given by the sample mean (b_i=1/n), then from (8) it follows that the preliminary predictor minimising the expected revision – that will be model biased as the sample mean – is formally *different*, depending on both x_i and v_i: the optimal preliminary predictor will still be the sample mean (a_i=1/n_P) if x_i=v_i=1 for each i, under the common homoschedastic model.

More generally, under the model (1) we know that (Cicchitelli *et al.*, 1992, 385-387) the optimal linear predictor – e.g. that one minimising *MSE* respect to the model, $E(T - \overline{y})^2$ – is:

$$T^* = f \overline{y}_S + (1 - f) \overline{x}_{\overline{S}} \hat{\beta}^* \quad \text{where:} \quad \hat{\beta}^* = \left(\sum_{S} x_i y_i / v_i \right) \left(\sum_{S} x_i^2 / v_i \right)^{-1} \quad (10)$$

---

[3] According to Valliant *et al.* (2000, 29), if θ is a population parameter given by the linear combination of y's in the population, and $\hat{\theta}$ is a general linear predictor of θ based on a sample S, in symbols we have $\theta = \sum_U d_i y_i$ and $\hat{\theta} = \sum_S c_i y_i$, where d_i and c_i represent the known population coefficients and the sample weights respectively. Under (1), the linear unbiased predictor of θ with the smallest variance is $\hat{\theta}^* = \sum_S c_i^* y_i = \sum_S \omega_i^* d_i y_i$, where: $\omega_i^* = 1 + x_i (d_i v_i)^{-1} (\sum_{\overline{S}} d_i x_i) (\sum_S x_i^2 / v_i)^{-1}$. The optimal provisional unbiased predictor under (1) can be defined simply considering that, in a provisional estimation context, the final sample and the provisional sample play the roles of population and sample respectively. Consequently, in order to obtain (8) we only need to substitute S, S_P, S_L, a_i, b_i, γ_i^* in place of U, S, $\overline{S}$, c_i, d_i, ω_i^*.

with $f=n/N$ and $\overline{y}_S$, $\overline{x}_{\overline{S}}$ equal, respectively, to the sample y-mean and the x-mean referred to the not observed units. The model *MSE* will be equal to:

$$MSE(T^*)=\left[\left(\sum_{\overline{S}} x_i\right)^2 \Big/ \left(\sum_{S}(x_i^2/v_i)\right)+\sum_{\overline{S}} v_i\right]\frac{\sigma^2}{N^2}. \tag{11}$$

As a consequence, when v is a not decreasing function of x the best choice of the sample simply consists, *if it is possible*, in selecting the n units in the universe having the largest x-values. A direct consequence of (8) is that, putting $\overline{y}_{S_P}$ as the y-mean in the *preliminary* sample, the optimal preliminary predictor will be given by:

$$T_P^*=f_P\overline{y}_{S_P}+(1-f_P)\overline{x}_{\overline{S}_P}\hat{\beta}_P^* \quad \text{where:} \quad \hat{\beta}_P^*=\left(\sum_{S_P} x_i y_i/v_i\right)\left(\sum_{S_P} x_i^2/v_i\right)^{-1} \tag{12}$$

with $f_P=n_P/N$, so that it keeps *the same form* of (10), but substituting the final sample S with the preliminary one S_P. In the quite common case when $v=x$, it follows from (10) and (12) that the optimal final and preliminary predictors are given by, respectively, these ratio estimators:

$$T^*_{(v=x)}=\overline{y}_S\frac{\overline{x}}{\overline{x}_S} \quad \text{and} \quad T^*_{P(v=x)}=\overline{y}_{S_P}\frac{\overline{x}}{\overline{x}_{S_P}}. \tag{13}$$

If one considers again the case when, under (1), the final predictor is the sample mean, then the preliminary predictor that is optimal on the basis of (8) is given by:

$$T_P^*=f'_P\overline{y}_{S_P}+(1-f'_P)\overline{x}_{S_L}\hat{\beta}_P^* \quad \text{where:} \quad \hat{\beta}_P^*=\left(\sum_{S_P} x_i y_i/v_i\right)\left(\sum_{S_P} x_i^2/v_i\right)^{-1} \tag{14}$$

with $f'_P=n_P/n$. It has the same model mean than the sample mean and is based on the same estimator of the regression coefficient given by the second formula (12); however, the basic difference respect to the preliminary predictor in (12) is that in this case the role of the universe is played by the *final* sample, because we have n instead of N and S_L instead of $\overline{S}_P$.

The formula (14) concerns the optimal preliminary predictor to be used when the final estimation is based on the (model biased) final sample mean. However, it must be remarked that the different formal structure between the best preliminary predictor and the final predictor is not due to the bias of the latter, but to its lack of optimality under (1). For instance, under (1) a final unbiased predictor is given by:

$$T_\alpha=\alpha[f_P\overline{y}_{S_P}+(1-f_P)x_{\overline{S}_P}\hat{\beta}_P^*]+(1-\alpha)[f_L\overline{y}_{S_L}+(1-f_L)x_{\overline{S}_L}\hat{\beta}_L]=\alpha T_P^*+(1-\alpha)T_L \tag{15}$$

where $f_L=n_L/N$, $\hat{\beta}_L$ is an estimate of β formally similar to $\hat{\beta}_P^*$, but based on the n_L late respondents only, and α is a weight ranging in [0,1]. It is based on a separate use of preliminary and late respondents, for instance because one could assess a larger variance of the late respondents' data (section 5) and wants to contain its effect on efficiency of estimates choosing α near to one (more generally, when one suspects *different* models for preliminary and late respondents; compare next formula (21)). However, on the basis of (8), after some elaborations one gets a less intuitive formula than T_P^* for the optimal preliminary predictor conditioned to the final predictor (15), given by:

$$T_{\alpha P}^* = \frac{\alpha}{N}\sum_{S_P}\left[1+\frac{x_i}{v_i}\left(\sum_{S} x_i^2/v_i \sum_{\bar{S}_P} x_i + \sum_{S_L} x_i^2/v_i \sum_{S_P} x_i^2/v_i\right)\left(\sum_{S_P} x_i^2/v_i\right)^{-2}\right] y_i . \quad (16)$$

There are at least two main reasons that can justify the use of a final predictor that is not optimal under (1):

1) when the estimation strategy was established, it was not possible to verify rightness of the model, so that final weights b_i could actually be different from those which minimise *MSE* under (1). For instance, that is the case when originally no auxiliary variable x was available in order to test model rightness.
2) One could not completely trust rightness of the supposed model (1), so that final estimation could be based on estimators that are optimal according to other criteria. For instance, the well known and widely used *GREG* estimator is not optimal under (1), but it is asymptotically unbiased under whatever sampling design (Cicchitelli *et al.*, 1992, 399). It is a calibration estimator and all calibration estimators converge to the *GREG* (Deville and Särndal, 1992). As well known, its form under a simple random sampling design and a model as (1) is given by:

$$T_{GREG} = \bar{y}_{S_P} + \hat{\beta}^*(\bar{x} - \bar{x}_{S_P}) \quad (17)$$

where $\hat{\beta}^*$ is given by the second relation (10). However, Hedlin *et al.* (2001) showed that even (17) could produce inefficient although design consistent estimates when the model is miss-specified, especially when dealing with the highly variable and outlier prone populations that are the focus of many business surveys.

2.3 *Preliminary estimation when the auxiliary variable is not available for all the units*

The use of a predictor based on a linear combination of two predictors applied separately to 2 sub-samples – as seen for instance in (15) – will be emphasized in section 3, but there is another relevant concrete case when it could be necessary. In many operational contexts it is not always possible to measure the auxiliary x-variable of the model (1) on each observed unit. For instance, in a short-term busi-

ness survey aimed at estimating changes of the average turnover referred to a month m (y-variable), the right x-variable is often given by turnover referred to the month (m-12) (Copeland and Valliant, 2007): however, normally it can be observed only on those units belonging to the sample and *respondent* in the month (m-12).

Both in a preliminary or a final estimation context, a solution consists in using another auxiliary variable z – available for all the units in the population – instead of x. For instance, in a business survey context z often derives from the business register and is given by the yearly turnover or the number of persons employed referred to a previous year. If x can be measured on n_x units belonging to the sub-sample S_x (and can not be measured on the remaining $n_z=n-n_x$), while z can be measured on all the n units, then the final estimation could be carried out using:

1. the only n_x units for which the variable x can be measured;
2. the variable x for the n_x units and the variable z for the remaining ones;
3. all the n units and the only variable z;
4. the variable x for the n_x units and an estimate of x for the remaining ones (based on z).

The first case leads to optimal final and preliminary predictors formally similar to what already seen, respectively, according to formulas (10) and (12), with the difference due to the use of n_x units only instead of n. If the model (1) is true, it is helpful to develop option 2, with the aim to evaluate the trade/off between the advantage – respect to option 1 – in the use of n_z additional units, and the additional bias due to the use of z in place of x. This use derives from the implicit idea that a working model alternative to (1) is given by:

$$y_i = \gamma z_i + \delta_i \quad \text{where:} \quad \begin{cases} E(\delta_i) = 0 & \forall i \\ VAR(\delta) = \sigma_u^2 u_i & \forall i \\ COV(\delta_i, \delta_j) = 0 \quad if \quad i \neq j \end{cases} \tag{18}$$

where for simplicity no explicit model bias is introduced, but the hypothesis $VAR(\delta_i) > VAR(\varepsilon_i)$ for each i justifies the higher reliability of model (1). If $f_x=n_x/N$, the final predictors based, respectively, on n_x and n_z units will be given by:

$$T_x^{(2)} = f_x \overline{y}_{S_x} + (1 - f_x)\overline{x}_{\overline{S}_x}\hat{\beta}_x \quad \text{where:} \quad \hat{\beta}_x = \left(\sum_{S_x} x_i y_i / v_i\right)\left(\sum_{S_x} x_i^2 / v_i\right)^{-1} \tag{19}$$

$$T_z^{(2)} = f_z \overline{y}_{S_z} + (1 - f_z)\overline{z}_{\overline{S}_z}\hat{\gamma}_z \quad \text{where:} \quad \hat{\gamma}_z = \left(\sum_{S_z} z_i y_i / u_i\right)\left(\sum_{S_z} z_i^2 / u_i\right)^{-1} \tag{20}$$

where S_z is the sub-sample of units for which x can not be measured. According to model (1), we have: $E(\hat{\gamma}_z) = \beta\left(\sum_{S_z} z_i x_i / u_i\right)\left(\sum_{S_z} z_i^2 / u_i\right)^{-1} = \beta A_z$; as a con-

sequence: $E(T_z^{(2)}) = f_z\beta\overline{x}_{S_z} + (1-f_z)\overline{z}_{\overline{S}_z}\beta A_z = \beta\overline{x} + (1-f_z)\beta(\overline{z}_{\overline{S}_z}A_z - \overline{x}_{\overline{S}_z}) =$ $= E(\overline{y}) + Bias_z$. If $\hat{Bias}_z$ is an estimate of the bias component, it follows that an approximately unbiased final predictor will be given by:

$$T^{(2)} = \alpha T_x^{(2)} + (1-\alpha)(T_z^{(2)} - \hat{Bias}_z) \tag{21}$$

and it is well known that the optimal choice of *a* is:

$$\hat{\alpha} = \hat{VAR}(T_z^{(2)})[\hat{VAR}(T_x^{(2)}) + \hat{VAR}(T_z^{(2)})]^{-1}. \tag{22}$$

Even though under (22) $T^{(2)}$ always improves $T_x^{(2)}$, the crucial point consists in the need to correctly estimate variances in (22) and, in particular, the bias component in (21). That can be done estimating β with $\hat{\beta}_x$ and A_z with the formally equivalent A_x calculated on the units in S_x, while $\overline{z}_{\overline{S}_z}$ is always available if z can be measured on all the units in the population. On the other hand, the estimation of $\overline{x}_{\overline{S}_z}$ (as well as the estimation of $\overline{x}_{\overline{S}_x}$ in (19)) could be problematic: some possibilities consist in modelling x as a function of z using historical data, or when $x=y_{(m-12)}$ using the population y-means estimated one year before. One can also remark that $Bias_z$ is approximately null if the sample S_z is approximately *balanced* respect to x and z (Royall, 1992): for instance, if in (18) $u_i=z_i$ for each i, the bias is negligible if $\overline{x}_{S_z}/\overline{x}_{\overline{S}_z} \approx \overline{z}_{S_z}/\overline{z}_{\overline{S}_z}$, that is the implicit hypothesis justifying the use of (21) without any correction for bias.

As regards preliminary estimation, an approximately optimal solution can be still derived from (8), taking into account that in this case the form of the final predictor is given by (2) plus the estimate of $-(1-\alpha)Bias_z$. The predictor (22) can be written as: $T^{(2)} = \sum_{S_x} b_{xi}y_i + \sum_{S_z} b_{zi}y_i - (1-\hat{\alpha})\hat{Bias}_z$, where:

$$b_{xi} = \frac{\hat{\alpha}}{N}\left[1 + \frac{x_i}{v_i}\sum_{\overline{S}_x} x_i\left(\sum_{S_x} x_i^2/v_i\right)^{-1}\right] \text{ and } b_{zi} = \frac{(1-\hat{\alpha})}{N}\left[1 + \frac{z_i}{u_i}\sum_{\overline{S}_z} z_i\left(\sum_{S_z} z_i^2/u_i\right)^{-1}\right],$$

so that according to (8) the optimal preliminary predictor will be:

$$T_P^{(2)} = \sum_{S_{Px}}\left[b_{xi} + \frac{x_i}{v_i}\sum_{S_{Lx}} b_{xi}x_i\left(\sum_{S_{Px}} x_i^2/v_i\right)^{-1}\right]y_i +$$
$$+\sum_{S_{Pz}}\left[b_{zi} + \frac{z_i}{u_i}\sum_{S_{Lx}} b_{zi}z_i\left(\sum_{S_{Pz}} z_i^2/u_i\right)^{-1}\right]y_i - (1-\hat{\alpha}_P)\hat{Bias}_{Pz} \tag{23}$$

where $\hat{\alpha}_P$ and $\hat{Bias}_{Pz}$ are estimates based on the preliminary sample only.

It is worthwhile to note that option 3 is a particular case of option 2 when $\alpha = 0$, with S_z=S and n=n_z. As regards option 4, generally speaking it leads to an additional random error component due to the imputation process. However, under models (1) and (18) we have $E(y_i) = \beta x_i = \gamma z_i$, so that each missing x-value should be estimated according to $\hat{x}_i = z_i \hat{\gamma} / \hat{\beta}$, e.g. it is supposed to be straightforwardly proportional to the corresponding z-value.

3. A MODEL FOR EVALUATING SELF-SELECTION BIAS

A relevant theoretical problem concerns the possible *self-selection* of quick preliminary respondents. As already remarked in sections 1, it can lead to biased estimates of the unknown population mean and variances. As underlined by Bolfarine and Zacks (1992, 128-133), the question of robustness of predictors of population quantities can be faced using three approaches: 1) imposing restrictions to the possible super-population models adopted; 2) imposing restrictions to the samples to be selected; 3) using Bayes predictors that adaptively consider the possibility that each one out of a series of alternative models is the correct model. The first approach seems the most appropriate for the context under study and one of the most exploited in theory.

In particular, it is possible to model potential structural differences between preliminary and late respondents. We suppose that the population U can be split into 2 separate sub-populations U_A and U_B, including respectively N_A and N_B units, with $U = U_A \cup U_B$ and N=N_A+N_B. For each of the 2 sub-populations (labelled with h, where h=A,B) this model is supposed true:

$$y_{hi} = \beta_h x_i + \varepsilon_{hi} \quad \text{where} \quad \begin{cases} E(\varepsilon_{hi}) = 0 & \forall h, i \\ VAR(\varepsilon_{hi}) = \sigma_h^2 v_i & \forall h, i \\ COV(\varepsilon_i, \varepsilon_j) = 0 \quad \textit{if} & i \neq j \end{cases} \quad \text{for } h=A,B \qquad (24)$$

where all symbols keep the same logical meaning as for model (1). This model is supposed to be valid both for preliminary and late respondents. The basic idea is that these sub-populations do not derive from a preliminary stratification, but depend on some latent factor underlying units under observation. As a consequence, the coexistence of 2 sub-populations *can not* be modelled a priori, e.g. when the estimation strategy is planned, because in that case one could simply carry out separate preliminary and final estimations inside each sub-population using the same criteria seen in the previous section. Even though the 2 sub-populations could differ not only for *different* coefficients β_h and σ_h, but for *different* auxiliary variables x_h and v_h as well[4], in the follow we will suppose x_A=x_B=x and v_A=v_B=v.

[4] For instance, in a short-term business survey frame a classical example is when the auxiliary variable modelling individual turnover in the month m could be given by either turnover in the month (m-1) or turnover in the month (m-12), depending on seasonality and other short-term effects that are peculiar of the context under study.

We suppose that at each estimation stage the split into 2 clusters depends on one or more discriminating variables that could be observed or not, and that information derived from the preliminary sample can be used to correct preliminary estimates for taking into account the effect of this split on the potential bias of preliminary estimates. At each estimation stage (for instance, in 2 following monthly waves of a short-term survey) sub-populations *could change* and must be identified from scratch.

Given model (24), the preliminary and the final samples can be written as, respectively: $S_P = S_{AP} \cup S_{BP}$ and $S = S_A \cup S_B$; they will include, respectively, $n_P = n_{AP} + n_{BP}$ and $n = n_A + n_B$ units. Prospect 1 supplies an overall resuming scheme.

PROSPECT 1

Different patterns for 2 sub-populations A and B

DOMAIN	STRUCTURE			SIZE		
	Population	Sub-population A	Sub-population B	Total	Sub-total A	Sub-total B
Universe	U	U_A	U_B	N	N_A	N_B
Preliminary sample	S_P	S_{AP}	S_{BP}	n_P	n_{AP}	n_{BP}
Late sample	S_L	S_{AL}	S_{BL}	n_L	n_{AL}	n_{BL}
Final sample	S	S_A	S_B	n	n_A	n_B

If $\overline{y}_A$ and $\overline{y}_B$ are the unknown y-means in the sub-populations A and B and $\overline{x}_A$ and $\overline{x}_B$ are the corresponding x-means, the unknown mean to be estimated will be given by:

$$\overline{y} = \overline{y}_A \frac{N_A}{N} + \overline{y}_B \frac{N_B}{N} = \overline{y}_A W_A + \overline{y}_B W_B \quad \text{where: } E(\overline{y}) = \beta_A \overline{x}_A W_A + \beta_B \overline{x}_B W_B \,. \tag{25}$$

3.1 *Optimal final prediction under model (24)*

A general formula for the final predictor of the population mean is given by:

$$T_{(AB)} = T_A \hat{W}_A + T_B \hat{W}_B \tag{26}$$

where T_A and T_B are predictors, respectively, of $\overline{y}_A$ and $\overline{y}_B$ that are based on the units of the final sample belonging to the sub-populations A and B, while $\hat{W}_A$ and $\hat{W}_B$ are estimates of the true population weights W_A and W_B. Variability of these last estimates does not depend on the model (24). Each final predictor can be written as (26) even when the split into 2 sub-populations is not formally introduced, as it happens when the final estimation is carried out mixing together units belonging to the sub-populations A or B - as for predictors (10) or (12), that are biased respect to the model (24). However, without loss of generality we can suppose that at each estimation stage one can know – or at least estimate – the proper sub-population for each unit in the sample, meaning that one can always

calculate two separate predictors for sub-populations A and B and combine them using proper weights. Similarly to (26), it follows that the general form of a preliminary estimator, based on the units belonging to S_P, is given by:

$$T_{(AB)P} = T_{AP}\hat{W}_{AP} + T_{BP}\hat{W}_{BP} \tag{27a}$$

and, in particular, if the optimal predictors are used separately for domains A and B we can put:

$$T^{*}_{(AB)P} = T^{*}_{AP}\hat{W}_{AP} + T^{*}_{BP}\hat{W}_{BP}\,. \tag{27b}$$

If one also supposes that $E(T_{AP}-T_A)=E(T_{BP}-T_B)=0$ – as it happens if preliminary and final predictors are unbiased respect to the model (24) – the condition equivalent to (4a) – e.g. $E(T_P-T)=0$ – is always satisfied if:

$$\hat{W}_{AP} = \hat{W}_A \quad \text{and} \quad \hat{W}_{BP} = \hat{W}_B\,. \tag{28}$$

In particular, when the true W_A and W_B are not known, one could put:

$$\hat{W}_{AP} = \hat{W}_A = n_A / n \quad \text{and} \quad \hat{W}_{BP} = \hat{W}_B = 1-(n_A/n)\,. \tag{29}$$

From (29) it is clear that the only source of model bias for preliminary estimates is due to the fact that when preliminary estimates must be calculated, n_A and n can not be known exactly, being the number of final respondents a random variable itself, with the only but relevant exception of a census survey ($n_A=N_A$). More precisely, that is the only source of bias if the attribution of sample units to the sub-populations A or B is carried out without the risk of a significant miss-classification error. In particular, if one can classify all the units into the 2 sub-populations A and B, then W_A will be known at the stage both of preliminary and final estimates, so that a simple rule that can guarantee a model bias near to zero both for preliminary and final estimates is:

$$\hat{W}_{AP} = \hat{W}_A = W_A \quad \text{and} \quad \hat{W}_{BP} = \hat{W}_B = 1-W_A\,. \tag{30}$$

However, it must be remarked again that bias could not be equal to zero even when (30) is used because: a) at the *preliminary* estimation stage, some units of the preliminary sample are miss-classified in B although their true sub-population is A (and/or vice-versa) – so that $E(T_{AP}-T_A)\neq 0$ and $E(T_{BP}-T_B)\neq 0$ – because some preliminary sample units belonging to B are erroneously used for calculating T_{AP} and vice-versa. The bias could be larger if a miss-classification is paid at the final estimation stage as well; b) at the *final* estimation stage, some units of the final sample are miss-classified in B although their true sub-population is A (and/or vice-versa) – so that $E(T_A-\overline{y}_A)\neq 0$ and $E(T_B-\overline{y}_B)\neq 0$ – because some final sample units belonging to B are erroneously used for calculating T_A and vice-versa (compare formula (34)).

An obvious generalisation of model (24) and of all the considerations leading to (30) consists in supposing k sub-populations instead of 2. Real data seem to fit better with a modelling based on k=2 or k=3, depending on the algorithm used for identifying sub-populations (section 3.3).

If the *true* model (24) is ignored, and the only model (1) is taken into account, the model bias due to the use of the preliminary predictor (12) optimal under (1) can be evaluated. It can be shown (Appendix 6.1) that this bias is negligible if these conditions hold:

$$\frac{\sum_{S_{AP}} x_i^2 / v_i}{\sum_{S_P} x_i^2 / v_i} \approx \frac{\sum_{S_{AP}} x_i}{\sum_{S_P} x_i} \quad \text{and} \quad \frac{\sum_{S_{BP}} x_i^2 / v_i}{\sum_{S_P} x_i^2 / v_i} \approx \frac{\sum_{S_{BP}} x_i}{\sum_{S_P} x_i}. \tag{31}$$

When x=v, the previous conditions will be approximately satisfied if both S_{AP} and S_{BP} are preliminary samples *balanced* respect to the whole populations U_A and U_B.

3.2 *Some features of the optimal strategy*

Similarities between the strategy based on the predictor (27a) and the post-stratification process often used – under a design-based approach – in order to reduce non response bias can be evaluated. Following Cicchitelli *et al.* (1992, 419-421), the design bias ($Bias_d$) of the post-stratified estimator based on the 2 same post-strata labelled as A and B – that formally is similar to (27a), taking into account that in this context non response is equivalent to a late response – is given by:

$$Bias_d(T_P) = W_A \frac{n_{AP}}{n_A}(\bar{y}_{AP} - \bar{y}_{AL}) + W_B \frac{n_{BP}}{n_B}(\bar{y}_{BP} - \bar{y}_{BL}) \tag{32}$$

where $\bar{y}_{AP}$ is the mean of the sub-population including units in sub-group A which respond as preliminary and all the other population means have an analogous meaning. Since under the model (24) $E(\bar{y}_{AP} - \bar{y}_{AL}) = E(\bar{y}_{BP} - \bar{y}_{BL}) = 0$, it follows $E[Bias_d(T_P)] = 0$ as well, so that even when the design bias (32) is far from zero, its model expectation will always be zero. The 2 post-stratifications have different goals: while the common design-based post-stratification is aimed at finding post-strata A and B such that $(\bar{y}_{AP} - \bar{y}_{AL}) \approx (\bar{y}_{BP} - \bar{y}_{BL}) \approx 0$, the strategy based on model (24) derives from the hypothesis of 2 existing sub-populations A and B, but without imposing the previous quite restrictive constraint that in each sub-population preliminary and late respondents must have the same *realised* y-means.

The model (24) itself is a generalisation of a model proposed by Gismondi (2007*b*). In that case, the sub-populations A and B were given by the sub-

populations of preliminary and late respondents, meaning that the preliminary sample was supposed to include only units of the sub-population *A* (containing all and only the units that could potentially be *quick* respondents) and the late sample was supposed to include only units of the sub-population *B* (containing all and only the units that could potentially be *late* respondents). Under that model the preliminary sample does not contain, by definition, any information on sub-population *B*, while under model (24) both the preliminary and the final samples contain units belonging to the 2 sub-populations and the bias of preliminary estimates can be reduced conditioning to a right identification of sub-populations just at the preliminary estimation stage. In other terms, instead of imposing *a priori* that preliminary and late respondents have different model means or variances, model (24) implies that the preliminary estimation bias can be even eliminated if, for instance, $n_{AP}/n_A \approx n_A/n$, e.g. when the preliminary and the final samples contain approximately the same share of units belonging to each sub-population.

The idea that there is a structural link between the delay of response and the model parameters is not new (Loosveldt and Carton, 2001; Billiet *et al.*, 2007), but it is difficult to be tested, because in the most part of official statistics (referred to business of households) responses are not monitored with a high frequency (daily or weekly), so that only largely approximated evidences could be found. However, if z is the basic variable determining the split into 2 sub-populations (let's note that it could be $z=x$), a simple model coherent with (24) is given by:

$$y_i = \begin{cases} \beta x_i + \varepsilon_i & if \quad z_i \le z_A \\ (\beta + \theta) x_i + \varepsilon_i & if \quad z_i > z_A \end{cases} \quad \text{where:}$$

$$VAR(\varepsilon_i) = \begin{cases} \sigma^2 v_i & if \quad z_i \le z_A \\ (\sigma^2 + \tau) v_i & if \quad z_i > z_A \end{cases} \tag{33}$$

where θ and τ are constant higher or lower than zero. If z is the number of days of delay between date of response and the end of the reference period, model (31) implies that all the units responding with a delay higher than z_A days have a model mean – given the x level – higher by θ respect to the model mean of units responding within z_A days. If the deadline for the calculation of preliminary estimates is z_P, when $z_P > z_A$ the preliminary sample will include units belonging to both the sub-populations, so that all the considerations seen above can be used in order to reduce bias of preliminary estimates. In this case $n_A = n_{PA}$ and $n_B = n_{PB} + n_L$, so that for implementing the preliminary predictor (27a) one must estimate n_L only. A more problematic case occurs if $z_i \le z_A$, because in this case the preliminary sample does not include any information on the sub-population *B*. If one supposes that $U_A = U_P$ and $U_B = U_L$, it can be shown (Gismondi, 2007*b*, 11-14) that, under the model (24) and according to (15) when $\alpha = 0{,}5$, the *final* best linear unbiased predictor will be given by $2T_{0,5}$: the main difference respect to (15) is that in this case $\overline{S}_P$ and $\overline{S}_L$ refer respectively to $(N_P - n_P)$ and $(N_L - n_L)$ not ob-

served units, instead of $(N\text{-}n_P)$ and $(N\text{-}n_L)$ as in (15). However, the implementation of the optimal preliminary predictor requests the estimation of β_L and β_P, that could be carried out using data referred to previous survey waves, but that could also lead to wrong preliminary estimates if non-response bias is unsteady along time in terms of magnitude and/or algebraic sign.

3.3 *The identification of sub-populations*

If the model (24) is true, even when the optimal predictor (27b) is used possible sources of bias could be given by: a) errors in the estimation of the population weight W_A; b) errors in the post-classification of sample units into the 2 sub-populations, as already remarked in section 3.1. If one takes into account these potential sources of error, it can be shown that an estimate of the consequent expected revision is given by (Appendix 6.2):

$$E(T^*_{(AB)P} - T^*_{(AB)}) = (\hat{W}_{AP} - \hat{W}_A)\left[\beta_A \bar{x}_A \left(\frac{n_A^R}{n_A} - \frac{n_B^W}{n_B}\right) - \beta_B \bar{x}_B \left(\frac{n_B^R}{n_B} - \frac{n_A^W}{n_A}\right)\right] \quad (34)$$

where upper labels "*R*" and "*W*" indicate, respectively, units classified into U_A or U_B "Rightly" or "Wrongly". From (34) one deduces that if the sub-populations' weights estimated at the preliminary and final stages are quite the same $(\hat{W}_{AP} \approx \hat{W}_A)$, then the expected revision is zero even when a miss-classification of sample units occurs; otherwise, if no miss-classification occurs the expected revision will reduce to:

$$E(T^*_{(AB)P} - T^*_{(AB)}) = (\hat{W}_{AP} - \hat{W}_A)(\beta_A \bar{x}_A - \beta_B \bar{x}_B)\,. \quad (35)$$

As already underlined, a crucial aspect concerns the technique to be used for detecting the 2 sub-populations *A* and *B*. One must note that it is a common practice to split the sample units between outlier and not outlier observations, because the effect of anomalous data could be fundamental in order to achieve robust estimates. As suggested by Hedlin *et al.* (2001), since outliers do not fit with models as (1) or (24), an operational rule consists in the combination of two estimators, given by an expansion estimator applied to outliers and an estimator applied to not outliers that is optimal according to some criterion (the *GREG* estimator in the mentioned case). It is easy to verify that this approach is a particular case of that based on (26) and (27a) respectively for the final and the preliminary estimation, where T_A and T_{AP} are expansion estimators.

More generally, the basic idea is that structural differences could be tested evaluating different average levels of the *y*-variable and/or the *x*-variable, even when structural differences could concern variability as well $(\sigma_A^2 \neq \sigma_B^2)$: the implicit, but realistic underlying hypothesis justifying this approach is that different average levels imply a different average variability, and vice-versa.

When a preliminary sample is available, a simple procedure can be based on

the method proposed by Cochran (1977, 128-130) for stratifying a given population, in order to minimise the variance of estimates in a stratified random sampling context. If one orders the observed units according to the not decreasing x-values, and (i) is the place occupied by the i-th unit in this ranking, then the rule is based on the following equality:

$$\sum_{i \in S_{AP}} \sqrt{y_{(i)}} \approx \sum_{j \in S_{BP}} \sqrt{y_{(j)}} \tag{36}$$

meaning that the sum of the square roots of the y-values in sub-population A must be equal to the same sum calculated in sub-population B. Since y can not be measured on the not observed units, weights W_A and W_B must be estimated in another way: for instance, supposing that they are equal to the corresponding preliminary sample weights, or applying the rule (36) to all the units in the population using the x-variable.

Even though the Cochran method is quite simple, it does not properly take into account individual variability. Then, a more detailed procedure can be based on the following steps:

a) the preliminary sample is split into 2 sub-samples S_{AP} and S_{BP} on the basis of a clustering algorithm.

b) According to a discriminant analysis or a logistic model, one can identify the most significant variables (available for all the units in the population) which the previous split depends on. One can use the only x-variable, or also additional auxiliary variables available for all the units in the population, but not included in the model (24); let's note that if a *CHAID* technique is used (Kass, 1980), then steps a) and b) are carried out at the same time.

c) Using parameters estimated in the previous step b), one can assign the units in the population not belonging to the preliminary sample to U_A or U_B and estimate W_A and W_B.

Let's note that in step a) one can use the only y-variable, and/or the same auxiliary variables used in step b). However, the first option seems more reasonable, because it corresponds more strictly to the logic of a post-stratification based on the y-values, while in the second case one carries out a stratification that could have been done also before drawing the sample, that is less coherent with the strategy underlying model (24).

Results of the post-stratification can be tested verifying the statistical significance of the difference between parameters estimated separately into 2 subsamples, using the same tools available for testing rightness of a general linear model. As regards expected values, according to the model (24) one can test the null hypothesis $\beta_A = \beta_B$ against the alternative $\beta_A \neq \beta_B$ estimating 2 separate regression coefficients on the basis of the second (12) and considering that, in a preliminary estimation context, the random variable:

$$(\hat{\beta}_A - \hat{\beta}_B) \Big/ \sqrt{[(n_{AP} - 2)VAR(\hat{\beta}_A) + (n_{BP} - 2)VAR(\hat{\beta}_B)]/(n_P - 4)} \tag{37}$$

is approximately a Student's t with (n_P–4) degrees of freedom.

A quite similar rationale can be used to test the difference $\sigma_A^2 - \sigma_B^2$ as well. According to (24), one can consider the model:

$$\tilde{s}_{hi}^2 = \sigma_h^2 y_i + \omega_{hi} \qquad \text{for } h=A,B \tag{38}$$

where for the *i-th* unit $\tilde{s}_{hi}^2$ is the *empirical y*-variance and ω_{hi} is a common random error. In this case an additional preliminary step consists in building up a series of n_{hP} couples ($\tilde{s}_{hi}^2$, v_i), where each $\tilde{s}_{hi}^2$ can be calculated using empirical historical data for the same unit: for instance, using data of various years referred to the same period t, or more simply data of the same year referred to different periods t.

4. AN APPLICATION TO REAL WHOLESALE TRADE DATA

4.1 *Revisions in the quarterly wholesale trade survey*

Starting from the first quarter of 2001, ISTAT (the Italian National Statistical Institute) elaborates and releases 8 quarterly index numbers (with base 2000=100) concerning turnover of the "Wholesale trade and commission trade sector" (classification NACE Rev.1, activities from 51.1 to 51.9). Indexes refer to the 7 economic activities plus the total wholesale (in the follow named as "groups"), according to the following scheme: 1) NACE 51.1 - Wholesale on a fee or contract basis; 51.2 - Agriculture raw materials and live animals; 3) 51.3 - Food, beverages and tobacco; 4) 51.4-51.5-51.6 - Household goods; 5) 51.7 - Non agriculture intermediate products; 6) 51.8 - Machinery, equipment and supplies; 7) 51.9 - Other products. Available time series of final indexes (released after 180 days from the end of the reference quarter) and preliminary indexes (after 60 days) are available for the 16 quarters from I-2003 to IV-2006.

The survey is based on a stratified random sampling including, in 2006, about 7.800 units. The stratification considered in this context is based on 21 strata, obtained crossing each other the 7 above mentioned groups and 3 employment classes (1-5; 6-19; >19)[5]. An elementary index is currently calculated in each stratum. Calculations of higher order indexes – among which the total wholesale trade one – are based on weighted means of lower order indexes, where weights are based on yearly turnover referred to 2000.

In order to better manage wave non-response, each turnover index with base 2000=100 is calculated in this way: first the ratio between average turnovers re-

[5] The stratification used in the survey is based on 3 geographic areas as well (81 strata). However, some preliminary analyses based on ANOVA showed that geographic area has a quite poor effect on the turnover variability; moreover, the use of 81 elementary domains could lead to estimates based on a too few number of respondents in each stratum. These reasons – as well as different rules for detecting outliers – explain why the index numbers calculated in this framework differ from those released by ISTAT.

ferred to quarters *t* and (*t*-4) is calculated; secondly, it is multiplied by the index number with base 2000=100 referred to the quarter (*t*-4) and calculated one year before. This option is guaranteed by the presence, in each quarterly questionnaire, of questions concerning turnover of both quarters *t* and (*t*-4), and quality checks are mostly based on the turnover ratio *t*/(*t*-4) control. Non responses are mainly due to deliberate refuses, while late responses depend on delays of the response mechanism and some random factors. Up to now, the implicit hypothesis assumed is that non responses (and the same late responses as well) follow a *missing at random* mechanism; that is the theoretical justification of the recourse to the current estimator given by the ordinary sample mean, used both for preliminary and final estimates (no re-weighting or imputation are carried out).

In all the following analyses we will consider only the units for which, in a given quarter, turnover concerning both quarters *t* and (*t*-4) was available. Moreover, some anomalous observations were detected and excluded from calculations. The problem of robustness and influence of anomalous observations on estimates is well known, as remarked by Hedlin *et al.* (2001). In this context, it was used a set of rules that is simplified respect to the one currently used in the survey, but guaranteeing that a *preliminary* unit detected as outlier will be considered outlier when *final* estimates are calculated as well[6]. For these reasons, the analysis concerned a reduced database of "non outlier" observations, including, on the average, the 98,1% of the real final respondents.

For each quarter and each estimation domain, revisions have been calculated as differences between the final and the preliminary estimated percent rate of change between the quarters *t* and (*t*-4): averaging the absolute value of single revisions one can calculate the *Mean of Absolute Revisions* (*MAR*) for whatever level of details.

The average number of final respondents was 5.650 (table 1), ranging from 5.182 in 2003 to 5.917 in 2004. The average weighted preliminary response rate on the final sample was 74,7%, against the not weighted 80,7%. The 2005 lower levels depend on some changes introduced in the overall survey methodology, concerning the technical tools used for receiving questionnaires and management of remainders. The overall mean of absolute revisions was 1,6%, ranging from 1,3% in 2003 and 2005 and 2,0% in 2006. As a matter of fact, the average linear correlation between the quarterly weighted response rates and revisions is poor (-0,12) and negative only in 2005 (-0,67) and 2006 (-0,65). The consequent outstanding issue is that when preliminary coverage is high (around 80% excluding 2005), there is no further clear negative interaction with the revision level.

In particular (table 2), even though there is a prevalence of positive revisions, there is not the presence of a clear systematic bias along the observed period. Using the actual estimation strategy based on the sample mean both for preliminary and finale estimates, a prevalence of underestimations due to preliminary esti-

[6] Basically, 2 main controls were activated at the single unit level: 1) the ratio between turnover referred to quarters *t* and (*t*-4) must range between 0,1 and 10; 2) the ratio between the highest turnover between *t* and (*t*-4) and the yearly turnover of the previous year derived from the business register must range between 0,05 and 20.

mates occur for the total wholesale trade (11 positive revisions against only 5 negative) and for groups 7 (13 positive revisions), 3 and 6 (12), while groups 2, 4 and 5 are perfectly equilibrated (8 positive revisions). However, the alternation of positive and negative revisions does not follow a regular pattern and it may be verified that there is not a clear negative link between absolute revision and preliminary response rate.

TABLE 1

Number of final respondents by group and year
(data include outliers – weighted response rate based on the business register turnover)

INDICATOR	2003	2004	2005	2006	Mean
Theoretical sample	6.875	7.978	7.622	7.832	7.577
Preliminary respondents	4.435	5.124	3.955	4.718	4.558
Final respondents	5.182	5.917	5.764	5.738	5.650
Preliminary response rate on final sample %	85,6	86,6	68,6	82,2	80,7
Preliminary weighted response rate on final sample %	84,0	79,8	57,5	77,6	74,7
Mean of absolute revisions (MAR) using the sample mean	1,3	1,8	1,3	2,0	1,6
Linear correlation between weighted response rate and revision	0,43	0,49	-0,67	-0,65	-0,12

TABLE 2

Quarterly revisions using the sample mean (actual estimator and excluding outliers) by group and quarter

	2003				2004				2005				2006			
Group	I	II	III	IV	I	II	III	IV	I	II	III	IV	I	II	III	IV
1	10,1	6,1	0,1	2,0	1,0	1,1	-2,3	-1,7	1,5	1,2	-0,8	6,6	-1,8	-1,9	-1,0	0,2
2	1,0	0,2	-0,0	-1,1	-1,9	-0,2	-0,4	2,1	-7,3	-13,6	-3,2	0,5	0,9	1,4	0,6	1,6
3	-0,2	0,8	0,5	15,6	3,0	-1,0	1,7	-2,1	0,1	0,3	0,7	3,3	1,5	6,4	0,2	-2,5
4	0,1	0,6	-0,0	-1,0	0,1	-0,2	1,3	0,1	1,3	-4,8	-3,6	-2,5	1,4	-6,8	7,6	-1,1
5	0,8	-0,0	0,0	-1,6	0,8	3,5	8,6	-0,4	0,3	-0,4	2,2	2,9	-3,9	6,8	1,2	-0,3
6	4,4	-1,7	-5,7	8,5	1,7	-1,2	0,5	-0,1	0,3	4,6	3,9	9,7	0,3	0,6	0,4	1,1
7	-1,1	0,4	-0,1	2,6	2,3	0,2	1,2	0,6	2,1	5,2	0,6	0,9	-2,1	3,0	0,9	1,7
Total	2,2	-0,8	0,7	1,4	1,2	1,1	3,7	-1,3	0,8	-1,1	2,0	1,4	-1,3	3,7	2,1	-1,1

4.2 *Model identification*

A crucial issue concerns the choice of the auxiliary x variable. Possible theoretical options have been given, for each enterprise, by the yearly turnover or the number of persons employed derived from the business register ASIA managed by ISTAT (both referred to the year *before* that under observation), turnover referred to the previous quarter or turnover referred to the same quarter of the previous year. Even though the two last options probably are the most appropriate – since the empirical evidence shows a strong reliability of self-regressive models for quarterly turnover – they could not be applied, because the x variable must be available for *each unit* in the population, or it must be known at least its total referred to units not belonging to the sample, while quarterly data derive from the survey itself.

The final choice was the yearly turnover derived from the business register. Preliminary analyses showed that the use of this auxiliary variable for each unit

led to better results than those got using the other 3 options listed in section 2.3. Let's note how a more general rule is that the auxiliary variable can be given by a power of the business register turnover.

Before using auxiliary information to implement some preliminary estimation strategy, one should always carry out a preliminary empirical validation of model (1), for instance according to some among the tests suggested in White (1980) or Gismondi (2007*a*).

As regards expected values, a simple technique consists in evaluating results of the regression model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{39}$$

verifying its overall significance and, in particular, statistical significance of model parameters. A quite similar model can be used to test model heteroschedasticity as well. It can be written as:

$$\tilde{s}_i^2 = \sigma_0^2 + \sigma_1^2 v_i + \omega_i \tag{40}$$

where for the *i-th* unit $\tilde{s}_i^2$ and ω_i have been already defined according to (38) and one can put $v=x$ or $v=x^2$. The idea to test the hypothesis $Var(y_i)=\sigma^2 v_i$ through ordinary regression is based on the availability of n couples $(Var(y_i), v_i)$: one can elaborate an individual estimate $\tilde{Var}(y_i) = \tilde{s}_i^2$ as described at the end of section 3.3. The inclusion in (40) of a constant term σ_0^2 and comparison with the significance level of σ_1^2 gives the implicit possibility to test the hypothesis $v=1$ as well.

In order to test models (39) and (40), a reduced database was built up for each year, including only units always respondent in all quarters. Results reported in the table 3 refer to 2005, that is the complete year with the largest number of final respondents. For testing model (40) with or without the constant term, the y variable was given by the *total* yearly turnover, got summing up *quarterly* data for each unit.

Testing of model (39) led to satisfactory results, except for group 3, where the correct R^2 is lower than 0,4 and, on a lesser extent, group 4 (about 0,7). It seems quite clear that the right model should not include a constant term, which statistical significance is poor, with a partial exception for group 2; moreover, for groups 4 and 5 the constant term would be negative, that is a nonsense in the context under study.

As regards the outcome of the variance structure model (40), we reported results using the hypothesis $v=x^2$, that according to the correct R^2 index works better than the hypothesis $v=x$ for groups 1, 4, 5, 6 and 7. On the other hand, poor results have been got for group 3[7]. As a matter of fact, on the whole turnover data concerning enterprises classified in group 3 can not be explained in a satisfactory way through a model defined by (1) nor as regards expected values or

[7] As regards group 2, the hypothesis $v=x$ should be preferred to $v=x^2$.

variability. Under the more realistic hypothesis $v=x^2$, the presence of a variance component not dependent on v (constant term) is quite always refused, with partial exceptions for groups 2 and 4; on the whole, such a variance model turned out to be particularly realistic for groups 6 and 7.

TABLE 3

Linear model test (39) and variance structure test (40) with $v=x^2$ by group – Year 2005 (outliers are not included)

			Linear model test (39)				Variance structure test (40) with $v=x^2$			
					Parameters' significance				Parameters' significance	
Group	Constant in the model	Degrees of freedom	Correct R^2	Fisher's F (1)	Constant	Regression coefficient	Correct R^2	Fisher's F (2)	Constant	Regression coefficient
Group 1	Yes	865	0,921	7.702	0,122	0,000	0,745	1.941	0,306	0,000
	No	866	0,922	7.970		0,000	0,746	1.963		0,000
Group 2	Yes	398	0,975	12.832	0,042	0,000	0,445	264	0,006	0,000
	No	399	0,974	13.844		0,000	0,449	270		0,000
Group 3	Yes	682	0,366	333	0,987	0,000	0,034	21	0,323	0,000
	No	683	0,390	370		0,000	0,035	22		0,000
Group 4	Yes	1.642	0,692	3.071	0,056	0,000	0,481	1.267	0,057	0,000
	No	1.643	0,707	3.312		0,000	0,480	1.264		0,000
Group 5	Yes	1.048	0,996	219.395	0,715	0,000	0,754	2.772	0,474	0,000
	No	1.049	0,995	223.154		0,000	0,754	2.784		0,000
Group 6	Yes	682	0,976	22.532	0,396	0,000	0,951	10.962	0,249	0,000
	No	683	0,975	23.716		0,000	0,950	11.034		0,000
Group 7	Yes	361	0,833	1.484	0,122	0,000	0,930	3.942	0,149	0,000
	No	362	0,834	1.537		0,000	0,927	3.958		0,000

(1) Fisher's F significance level is always 0,00000. (2) Fisher's F significance level is always 0,00000, except for group 3 (0,00001).

4.3 *Comparison among estimation strategies and main results*

In order to implement some of the possible preliminary estimators introduced, except the sample mean, it is necessary to know - just when the early estimate is carried out - the number of final respondents (that is, of late respondents) and the correspondent x-total, while obviously this information can not be available at the moment of quick estimate, since late respondents can vary from a quarter to another. As a consequence, these figures must be estimated. The technique herein used is based on the following steps:

1) for each unit in the theoretical sample, a final response probability has been estimated. This probability is equal to 1 if the unit is a quick respondent, otherwise it is estimated according to a *logit* model, based on these auxiliary variables: yearly turnover derived from the business register, percent of occasions (on the 4 quarters before that under study) for which the unit was a final respondent.
2) For each of the 27 strata, the *expected number of final respondents* (*enfr*) for quarter t is estimated as the average between the effective number of final respondents at times (t-1) – short-term effect – and (t-4) – seasonal effect.

3) In each stratum, units in the theoretical sample have been ordered according to their not increasing final response probabilities estimated at the step 1).
4) The units that are estimated to be final respondents are those that, in the rank referred to step 3), are placed in the first *enfr* positions.

Estimators used and compared have been listed in prospect 2. In all cases, it was possible to estimate the year-to-year change $t/(t\text{-}4)$ by the ratio between average turnover of the two periods, using the same auxiliary x variable and putting as y, respectively, turnover referred to quarters t and $(t\text{-}4)$. In order to get estimates of index numbers with base 2000=100, these changes have been multiplied by the correspondent final indexes with base 2000=100 already calculated and referred to the same quarter of the previous year.

PROSPECT 2

List of preliminary estimation strategies compared in the empirical attempt

Code	Definition	Notes	Final estimator	
			The same estimator	Sample mean
I	Sample mean	It is the preliminary estimator actually used. It corresponds to predictor (10) when $v=x=1$ for each unit. When a rate of change is estimated, it is equivalent to the ratio estimator (predictor (10) with $v=x$).	Yes	Yes
II	Predictor (10) with $v=1$	Optimal preliminary estimator under the model defined by (1) and when the homoschedastic hypothesis $v=1$ is assumed.	Yes	Yes
III	Predictor (10) with $v=x$	Optimal preliminary estimator under the model defined by (1) and when the hypothesis $v=x$ is assumed (standard deviations are less than proportional respect to x).	Yes	Yes
IV	Predictor (10) with $v=x^2$	Optimal preliminary estimator under the model defined by (1) and when the hypothesis $v=x^2$ is assumed (standard deviations are proportional respect to x) .	Yes	Yes
V	Estimator (17) with $v=1$	*GREG* estimator, model unbiased under (1) and asymptotically design unbiased as well, when $v=1$.	Yes	Yes
VI	Estimator (17) with $v=x$	*GREG* estimator when $v=x$.	Yes	Yes
VII	Estimator (17) with $v=x^2$	*GREG* estimator when $v=x^2$.	Yes	Yes
VIII	Predictor (27b) with $v=1$	Under the model (24) and the homoschedastic hypothesis $v=1$, optimal preliminary estimator (10) applied separately in 2 sub-groups A and B (formula 27b) identified using (36) on the main variable of interest (quarterly turnover). The final predictor is given by (26) and weights W are based on the preliminary units both for the preliminary and the final estimates ($\hat{W}_{AP}=\hat{W}_{A}=n_{AP}/n_{P}$ and $\hat{W}_{BP}=\hat{W}_{B}=n_{BP}/n_{P}$).	Yes	No
IX	Predictor (27b) with $v=x$	Under the model (24) and the hypothesis $v=x$, the same strategy described in VIII.	Yes	No
X	Predictor (27b) with $v=x^2$	Under the model (24) and the hypothesis $v=x^2$, the same strategy described in VIII.	Yes	No

All methods have been applied using a final estimation strategy formally equivalent to the preliminary one; moreover, the actual final estimation strategy based on the sample mean was coupled with each of the preliminary estimation methods labelled from I to VII as well. Let's note that, by construction, it does not have a real statistical sense to apply as final estimator the sample mean when the preliminary estimation strategies VIII, IX and X are used. What is expected be-

fore calculations is a lower average revision when both preliminary and final estimates are based on the same strategy.

As already remarked, precision of preliminary indexes (60 days) with respect to final estimates released after 180 days have been evaluated according to one of the most important and used quality indicators given by *MAR*[8] (Czajka and Hinkins, 1993; Cantwell *et al.*, 1995).

In the table 4 – referred to the case when all the final estimates are based on the sample mean – all results have been calculated as arithmetic means of the 16 quarters' outcomes. The final column *Average* is the arithmetic mean of the 7 groups plus the total wholesale. The best average result (calculated as arithmetic mean of the 8 absolute average revisions for the 7 groups plus the total wholesale trade) is got using strategy VII (*GREG* when $v=x^2$), with a *MAR* equal to 2,14%, lower than strategy III (optimal predictor under (1) when $v=x$, *MAR*=2,16%) and the actual method I based on the sample mean (*MAR*=2,17%). Differences among average revisions are low, but it is important to underline that the actual preliminary estimation strategy I is not the best for any group and can be always improved by other strategies: strategy VII in 4 cases (groups 1, 2, 5 and the total), strategy V in 2 cases (groups 3 and 6), strategies II and IV in 1 case (groups 7 and 4 respectively). Moreover, among the best strategies the most cautious is III, because it is the second best in 5 cases and does not produce particularly wrong preliminary estimates in any occasion; on the other hand, the only strategy that should be avoided is II ($v=1$).

When the final estimator varies according to the particular preliminary estimation strategy used (table 5), the main evidence is the quite good performance of the strategy X, based on the combination of optimal preliminary predictions in 2 sub-strata when $v=x^2$: *MAR* is 1,64% only, and this strategy produces 5 best performances (groups 2, 5, 6, 7 and the total) and 1 second best (group 4), with just a wrong performance for group 1. The methodology proposed along section 3 improves results also when $v=1$ (strategy VIII), while worsens them when $v=x$ (strategy IX). A second, fundamental outcome is that, even if one prefers not to use the above methodology, when final estimates are not given by the sample mean the strategy IV leads to a quite lower *MAR* (1,97%) than when sticking to the sample mean (2,22%), that improves the strategy I performance (*MAR*=2,17%) with 1 best and 1 second best outcomes. On the other hand, even though the use of a different final estimator for each preliminary estimation strategy normally improves quality of preliminary estimates, that is less true for the *GREG* estimator, turning out to be quite more efficient when the final estimator is the sample mean (levels of *MAR* are 2,14% from table 4 and 2,46% from table 5).

[8] Mean of Absolute Revisions. It is conceptually analogous to the most common *MAE* (Mean of Absolute Errors).

TABLE 4

MAR for some preliminary estimation strategies – Average 2003-2006
(x = yearly turnover – final estimates based on the sample mean)

Strategy	GROUPS							Total	Average
	1	2	3	4	5	6	7		
I	2,47	2,26	2,49	2,03	2,11	2,79	1,56	1,62	2,17
II	2,65	2,66	2,94	2,63	3,13	3,00	**1,43**	2,15	2,57
III	2,48	2,26	2,49	2,09	2,09	2,79	1,51	1,58	2,16
IV	2,44	2,67	2,62	**1,52**	2,37	2,90	1,55	1,67	2,22
V	2,54	2,38	**2,42**	2,22	2,63	**2,51**	1,59	1,92	2,28
VI	2,47	2,26	2,50	2,13	2,13	2,80	1,56	1,55	2,18
VII	**2,40**	**2,13**	2,60	2,02	**2,08**	2,92	1,65	**1,34**	**2,14**

Note: in bold the "best" estimator, underlined the "second best".

TABLE 5

MAR for some preliminary estimation strategies – Average 2003-2006
(x = yearly turnover – different final estimates for each strategy)

Strategy	GROUPS							Total	Average
	1	2	3	4	5	6	7		
I	2,47	2,26	2,49	2,03	2,11	2,79	1,56	1,62	2,17
II	2,68	2,76	2,63	2,54	3,40	4,16	1,46	2,61	2,78
III	2,49	2,26	2,39	2,04	2,09	2,80	1,51	1,57	2,14
IV	2,62	2,41	2,37	**1,39**	2,08	2,09	1,44	1,40	1,97
V	3,07	2,13	2,75	1,87	2,15	1,98	1,61	1,44	2,13
VI	2,48	2,26	2,40	1,98	2,13	2,80	1,57	1,63	2,15
VII	2,59	2,30	2,68	2,04	2,76	3,53	1,79	2,02	2,46
VIII	**2,43**	2,94	**1,91**	2,21	2,60	4,33	1,43	2,04	2,48
IX	2,87	2,56	1,97	1,87	2,36	3,47	1,49	1,87	2,31
X	2,84	**1,44**	2,08	1,42	**1,07**	**1,94**	**1,40**	**0,94**	**1,64**

Note: in bold the "best" estimator, underlined the "second best".

A better model specification can be obtained if one uses the *square root* of the yearly turnover derived from the business register instead of the original yearly turnover (table 6). The 6 strategies compared to the sample mean improve their *MAR* in 4 cases: in particular, the strongest gain concerns strategy IV, whose *MAR* drops from 2,22% to 2,07% and, in this way, improves the sample mean itself (2,17%); moreover, in the only 2 cases when there is not any improvement (strategies III and VI), *MAR* remains substantially steady. If one takes into account the single domains, the best overall preliminary estimation strategy could be based on strategy IV for all the groups except 2 and 3, for which a better option is based on strategy VII.

When the final estimator varies according to the preliminary one, the square root criterion does not improve methods VIII, IX and X based on separate optimal predictions in 2 sub-strata, but improves 4 of the 6 remaining strategies different from the sample mean (table 7). In particular, the largest improvement concerns strategy VII (*MAR* passes from 2,46% in table 5 to 2,21%). However, the best and the second best strategies still remain, respectively, X and IV.

Even though the main goal of the survey is the estimation of quarterly changes, in table 8 precision of levels' estimation (average turnover per enterprise) is evaluated as well. In each group the sample mean is always improved by all the strategies selected from table 5. Gains are large; in particular, also for level estimation the best and the second best strategies are, respectively, X and IV: they re-

duce *MAPR* (mean of the absolute percent revisions) around 4%, respect to the original 14,15% got using the sample mean.

Table 9 synthesises the further gain in precision of preliminary estimates due to additional (or alternative) options introduced. For instance, if outliers as defined in section 5.1 are not excluded from calculations, the sample mean would lead to a *MAR* equal to 2,73% instead of the actual 2,17%. Taking into account strategies that, on the average, are the best (IV, VII and X), the largest precision gain is always got excluding outliers. For strategy IV, we have already seen how the lowest *MAR* (1,92%) is got excluding outliers, using the square root option for the auxiliary variable and not using the sample mean as final estimate.

TABLE 6

MAR for some preliminary estimation strategies – Average 2003-2006
(x = square root of yearly turnover – final estimates based on the sample mean)

Strategy	GROUPS							Total	Average
	1	2	3	4	5	6	7		
I	2,47	2,26	2,49	2,03	2,11	2,79	1,56	1,62	2,17
II	2,68	2,37	2,69	2,48	2,98	2,84	**1,47**	2,11	2,45
III	2,47	2,26	2,49	2,07	2,11	2,79	1,54	1,62	2,17
IV	**2,25**	2,33	2,51	**1,78**	**1,89**	**2,76**	1,54	**1,49**	**2,07**
V	2,50	2,30	2,50	2,12	2,26	2,78	1,56	1,63	2,21
VI	2,47	2,26	2,50	2,07	2,12	2,80	1,56	1,60	2,17
VII	2,45	**2,22**	**2,48**	2,04	2,06	2,79	1,57	1,59	2,15

Note: in bold the "best" estimator, underlined the "second best".

TABLE 7

MAR for some preliminary estimation strategies – Average 2003-2006
(x = square root of yearly turnover – different final estimates for each strategy)

Strategy	GROUPS							Total	Average
	1	2	3	4	5	6	7		
I	2,47	2,26	2,49	2,03	2,11	2,79	1,56	1,62	2,17
II	2,67	2,47	2,61	2,52	3,17	3,65	1,49	2,43	2,63
III	2,48	2,26	2,46	2,04	2,10	2,80	1,54	1,61	2,16
IV	**2,32**	2,19	2,29	1,67	1,81	2,25	**1,47**	1,37	1,92
V	2,44	2,23	2,49	1,96	2,02	2,22	1,55	1,49	2,05
VI	2,48	2,26	2,46	2,00	2,11	2,80	1,56	1,62	2,16
VII	2,50	2,26	2,51	2,02	2,16	3,02	1,57	1,67	2,21
VIII	2,39	3,10	**1,94**	2,25	2,65	4,25	1,62	2,09	2,53
IX	2,92	2,46	2,23	1,95	2,36	3,21	1,65	1,89	2,33
X	2,49	**1,61**	2,39	**1,61**	**1,38**	**2,04**	1,52	**1,16**	**1,78**

Note: in bold the "best" estimator, underlined the "second best".

TABLE 8

MAPR on levels for some preliminary estimation strategies – Average 2003-2006
(x = yearly turnover – different final estimates for each strategy)

Strategy	GROUPS							Total	Average
	1	2	3	4	5	6	7		
I	12,04	9,01	12,67	2,69	20,21	23,59	19,29	13,70	14,15
II	4,28	3,05	16,48	2,41	3,47	6,86	8,70	5,12	6,30
III	3,64	2,44	14,21	1,96	2,34	4,15	6,61	3,70	4,88
IV	**3,21**	**1,90**	10,42	**1,88**	3,00	3,71	5,41	2,72	4,03
VIII	3,40	2,42	13,10	2,03	3,23	6,99	7,45	4,05	5,33
IX	5,11	7,46	**6,76**	2,17	2,99	5,09	**3,78**	**2,69**	4,50
X	3,39	2,78	9,32	1,98	**1,04**	**3,13**	7,02	2,89	**3,94**

Note: in bold the "best" estimator, underlined the "second best".

TABLE 9

MAR reduction for some preliminary estimation strategies according to some options – Average 2003-2006

OPTIONS	GROUPS							Total	Average
	1	2	3	4	5	6	7		
					Strategy I				
With outliers	2,87	3,14	3,03	2,41	3,21	3,95	1,58	1,66	2,73
Without outliers	2,47	2,26	2,49	2,03	2,11	2,79	1,56	1,62	2,17
					Strategy IV				
With outliers	4,19	3,88	3,51	2,06	3,10	3,96	1,62	1,41	2,97
Without outliers	2,44	2,67	2,62	1,52	2,37	2,90	1,55	1,67	2,22
Without outliers – final estimator	2,62	2,41	2,37	1,39	2,08	**2,09**	1,44	1,40	1,97
Without outliers – square root	**2,25**	2,33	2,51	1,78	1,89	2,76	1,54	1,49	2,07
Without outliers – square root – final estimator	2,32	2,19	2,29	1,67	1,81	2,25	1,47	1,37	1,92
					Strategy VII				
With outliers	7,36	2,82	3,61	2,39	3,34	3,86	1,69	2,48	3,44
Without outliers	**2,40**	**2,13**	2,60	**2,02**	2,08	2,92	1,65	**1,34**	**2,14**
Without outliers – final estimator	2,59	2,30	2,68	2,04	2,76	3,53	1,79	2,02	2,46
Without outliers – square root	2,45	2,22	**2,48**	2,04	2,06	**2,79**	**1,57**	1,59	2,15
Without outliers – square root – final estimator	2,50	2,26	2,51	**2,02**	2,16	3,02	**1,57**	1,67	2,21
					Strategy X				
With outliers – final estimator	3,55	1,80	**2,01**	3,27	1,54	2,76	1,44	1,36	2,22
Without outliers – final estimator	2,84	**1,44**	2,08	**1,42**	1,07	1,94	1,40	**0,94**	**1,64**
Without outliers – square root – final estimator	**2,49**	1,61	2,39	1,61	1,38	2,04	1,52	1,16	1,78

Note: in bold the "best" estimator, underlined the "second best".

5. MAIN CONCLUSIONS

A critical goal for many short-term business surveys is to limit the size of revisions, since the credibility of the survey as useful policy input is strongly affected by precision of early estimates.

The main issue underlying the problem of the average revision reduction in a preliminary estimation context depends on the possibility that late respondent could have different characteristics than reporters used for releasing the quick preliminary estimates.

The need to use a preliminary estimation strategy *different* respect to that used for the final estimation depends on a potential non random *self-selection* of quick respondents and on a consequent non-response bias due to the co-existence of different super-population models for preliminary and late respondents. Moreover, in short-term sampling surveys it is not possible to know the exact composition of the final respondents' sample when the preliminary estimates must be released, because of wave non-response.

Efficiency of the preliminary estimation strategy should always be linked to the final estimation one, even though in some empirical contexts the final estimation is not carried out using the best estimator under a given design or model: it could be due to the not availability of any auxiliary variable useful to implement alternative strategies, or to a too poor or unsteady correlation with the variable under study.

As regards main results, some clear issues can be emphasized:

- theoretical results showed that, under a model based approach, the formal definition of the optimal preliminary predictor is similar to that of the optimal final predictor, unless a bias component due to late responses is introduced. However, precision of preliminary estimates also depends on the not exact knowledge of the true model and the need to know in advance which late respondents will occur.
- Even though the relative shortness of time series does not allow for definitive conclusions on robustness of empirical results, a revision reduction can be obtained using model based estimation strategies. In particular, the recourse to a post-stratification technique (section 3) can lead to a reduction of the average revision from the actual 2,17% down to 1,64%, meaning an efficiency gain of about 25%.

6. APPENDIX[9]

6.1 *Proof of formula (31)*

According to the definition (12) and to symbols introduced in section 3.1, we can write:

$$T_P^* = f_P \overline{y}_{S_P} + (1-f_P)\overline{x}_{\overline{S}_P}\hat{\beta}_P^* = f_P\left(\overline{y}_{S_{AP}}\frac{n_{AP}}{n_P} + \overline{y}_{S_{BP}}\frac{n_{BP}}{n_P}\right) +$$

$$+(1-f_P)\left[\hat{\beta}_{S_{AP}}^*\left(\sum_{S_{AP}} x_i^2/v_i\right)\left(\sum_{S_P} x_i^2/v_i\right)^{-1} + \hat{\beta}_{S_{BP}}^*\left(\sum_{S_{BP}} x_i^2/v_i\right)\left(\sum_{S_P} x_i^2/v_i\right)^{-1}\right]\overline{x}_{\overline{S}_P} = \ldots$$

$$\ldots = \left\{W_A\left[\left(\frac{n_{AP}}{N_P}\right)\overline{y}_{S_{AP}} + \left(1-\frac{n_{AP}}{N_P}\right)\hat{\beta}_{S_{AP}}^*\overline{x}_{\overline{S}_{AP}}\right] + W_B\left[\left(\frac{n_{BP}}{N_P}\right)\overline{y}_{S_{BP}} +\right.\right.$$

$$\left.\left.+\left(1-\frac{n_{BP}}{N_P}\right)\hat{\beta}_{S_{BP}}^*\overline{x}_{\overline{S}_{BP}}\right]\right\} + (1-f_P)\left\{\overline{x}_{\overline{S}_P}\left(\sum_{S_P} x_i^2/v_i\right)^{-1}\left[\hat{\beta}_{S_{AP}}^*\left(\sum_{S_{AP}} x_i^2/v_i\right) +\right.\right.$$

$$\left.\left.+\hat{\beta}_{S_{BP}}^*\left(\sum_{S_{BP}} x_i^2/v_i\right)\right] - \left[\overline{x}_{\overline{S}_{AP}}\hat{\beta}_{S_{AP}}^*\frac{(N_A - n_{AP})}{(N-n_P)} + \overline{x}_{\overline{S}_{BP}}\hat{\beta}_{S_{BP}}^*\frac{(N_B - n_{BP})}{(N-n_P)}\right]\right\}. \qquad (41)$$

The expression in the first graph brackets of (41) is simply equal to the optimal preliminary predictor (27b) when the true model is (24) and $\hat{W}_A = W_A$, $\hat{W}_B = W_B$. Since it is model unbiased, under (24) we have:

[9] This appendix contains original proves.

$$E(T_P^*) = E(\overline{y}) + (1-f_P)\left\{\overline{x}_{\overline{S}_P}\left[\frac{\beta_A\left(\sum_{S_{AP}} x_i^2/v_i\right) + \beta_B\left(\sum_{S_{BP}} x_i^2/v_i\right)}{\left(\sum_{S_P} x_i^2/v_i\right)}\right] - \right.$$

$$\left. -\left[\overline{x}_{\overline{S}_{AP}}\beta_A\frac{(N_A - n_{AP})}{(N-n_P)} + \overline{x}_{\overline{S}_{BP}}\beta_B\frac{(N_B - n_{BP})}{(N-n_P)}\right]\right\}.$$

As a consequence, dividing and multiplying all the terms in the graph brackets by $\overline{x}_{\overline{S}_P}$, we get that the model bias will be approximately null if:

$$\beta_A\left(\frac{\sum_{S_{AP}} x_i^2/v_i}{\sum_{S_P} x_i^2/v_i}\right) + \beta_B\left(\frac{\sum_{S_{BP}} x_i^2/v_i}{\sum_{S_P} x_i^2/v_i}\right) \approx \beta_A\left[\frac{\overline{x}_{\overline{S}_{AP}}}{\overline{x}_{\overline{S}_P}}\frac{(N_A - n_{AP})}{(N-n_P)}\right] + \beta_B\left[\frac{\overline{x}_{\overline{S}_{BP}}}{\overline{x}_{\overline{S}_P}}\frac{(N_B - n_{BP})}{(N-n_P)}\right] \tag{42}$$

and it follows straightforwardly that the relation (42) will be satisfied if conditions (31) hold.

6.2 *Proof of formula (34)*

We can add to symbols already introduced in section 3.3 the following ones:

- T_{AP}^{*R}, that indicates the optimal preliminary predictor (12) applied on the n_{AP}^{R} units correctly classified in domain A, with $E(T_{AP}^{*R}) = \overline{x}_A\beta_A$.
- T_{AP}^{*W}, that indicates the optimal predictor (12) applied on the n_{AP}^{W} units wrongly classified in domain A; since they belong to domain B, we will have $E(T_{AP}^{*W}) = \overline{x}_B\beta_B$.
- T_{BP}^{*R}, that indicates the optimal preliminary predictor (12) applied on the n_{BP}^{R} units correctly classified in domain B, with $E(T_{BP}^{*R}) = \overline{x}_B\beta_B$.
- T_{BP}^{*W}, that indicates the optimal predictor (12) applied on the n_{BP}^{W} units wrongly classified in domain B; since they belong to domain A, we will have $E(T_{BP}^{*W}) = \overline{x}_A\beta_A$.

As a consequence, the (pseudo-optimal) preliminary predictor (27b) can be written as:

$$T_{(AB)P}^{*} = T_{AP}^{*}\hat{W}_{AP} + T_{BP}^{*}\hat{W}_{BP} =$$

$$= \left(\frac{n_{AP}^{R}}{n_{AP}}T_{AP}^{*R} + \frac{n_{AP}^{W}}{n_{AP}}T_{AP}^{*W}\right)\hat{W}_{AP} + \left(\frac{n_{BP}^{R}}{n_{BP}}T_{BP}^{*R} + \frac{n_{BP}^{W}}{n_{BP}}T_{BP}^{*W}\right)\hat{W}_{BP}.$$

It follows that:

$$E(T^{*}_{(AB)P}) = \left(\frac{n^{R}_{AP}}{n_{AP}} \overline{x}_{A} \beta_{A} + \frac{n^{W}_{AP}}{n_{AP}} \overline{x}_{B} \beta_{B} \right) \hat{W}_{AP} + \left(\frac{n^{R}_{BP}}{n_{BP}} \overline{x}_{B} \beta_{B} + \frac{n^{W}_{BP}}{n_{BP}} \overline{x}_{A} \beta_{A} \right) \hat{W}_{BP} .$$

Using analogous symbols, we can also write for the (pseudo-optimal) final predictor (26):

$$E(T^{*}_{(AB)}) = \left(\frac{n^{R}_{A}}{n_{A}} \overline{x}_{A} \beta_{A} + \frac{n^{W}_{A}}{n_{A}} \overline{x}_{B} \beta_{B} \right) \hat{W}_{A} + \left(\frac{n^{R}_{B}}{n_{B}} \overline{x}_{B} \beta_{B} + \frac{n^{W}_{B}}{n_{B}} \overline{x}_{A} \beta_{A} \right) \hat{W}_{B} .$$

If one supposes that:

$$\frac{n^{R}_{AP}}{n_{AP}} \approx \frac{n^{R}_{A}}{n_{A}} \quad \text{and} \quad \frac{n^{R}_{BP}}{n_{BP}} \approx \frac{n^{R}_{B}}{n_{B}} \tag{43}$$

and one takes into account that: $\hat{W}_{BP} = 1 - \hat{W}_{AP}$, $\hat{W}_{B} = 1 - \hat{W}_{A}$, we get easily formula (34).

ISTAT, Italian National Statistical Institute ROBERTO GISMONDI

REFERENCES

J. BILLIET, M. PHILIPPENS, R. FITZGERALD, I. STOOP (2007), *Estimation of Non-response Bias in the European Social Survey: Using Information from Reluctant Respondents*, "Journal of Official Statistics", Vol. 23, 2, pp. 135-162.

D.A. BINDER, J.P. DICK (1989), *Modelling and Estimation for Repeated Surveys*, "Survey Methodology", Vol. 15, 1, pp. 29-45.

H. BOLFARINE, S. ZACKS (1992), *Prediction Theory for Finite Populations*, Springer-Verlag, Berlin.

P.J. CANTWELL, C.V. CALDWELL, H. HOGAN, C.A. KONSCHNIK (1995), *Examining the Revisions in Monthly Trade Surveys Under a Rotating Panel Design*, "Proceedings of the Section on Survey Research Methods", American Statistical Association, pp. 567-572.

C. CASSEL, C.E. SÄRNDAL, J. WRETMAN (1983), *Some Uses of Statistical Models in Connection with the Nonresponse Problem*, in W.G. MADOW, I. OLKIN, D. RUBIN (eds.), *Incomplete Data in Sample Surveys*, vol. 3, pp. 143-160, Academic press, New York.

G. CICCHITELLI, A. HERZEL, G.E. MONTANARI (1992), *Il campionamento statistico*, Il Mulino, Bologna.

W.G. COCHRAN, (1977), *Sampling Techniques*, J.Wiley & Sons, New York.

K.R. COPELAND, R. VALLIANT (2007), *Imputing for Late Reporting in the U.S. Current Employment Statistics Survey*, "Journal of Official Statistics", Vol. 23, 1, pp. 69-90.

J. CZAJKA, S. HINKINS (1993), *Comparing Advance and Final Estimates: 1990 SOI Corporate Sample*, "Proceedings of the Section on Survey Research Methods", American Statistical Association, pp. 592-596.

M.H. DAVID, R. LITTLE, M. SAMUEL, R. TRIEST (1983), *Imputation Models Based on the Propensity to Respond*, "Proceedings of the Section on Business and Economic Statistics", American Statistical Association, pp.168-173.

J.C. DEVILLE, C.E. SÄRNDAL (1992), *Calibration Estimators in Survey Sampling*, "Journal of the American Statistical Association", 87, pp. 376-382.

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*, Internal document, Eurostat, Luxembourg.

W. FULLER (1990), *Analysis of Repeated Surveys*, "Survey Methodology", 16, pp. 167-180.

R. GISMONDI (2007*a*), *Quick Estimation of Tourist Nights Spent in Italy*, "Statistical methods & Applications" Vol. 16, 1, pp. 141-168, Springer & Verlag.

R. GISMONDI (2007*b*), *More Rapid Tourism Statistics Using Auxiliary Variables*, "Statistica Applicata" Vol. 18, 3, pp. 1-38, Rocco Curto editore, Napoli.

D. HEDLIN, H. FALVEY, R. CHAMBERS, P. KOCIC (2001), *Does the Model Matter for GREG Estimation? A Business Survey Example*, "Journal of Official Statistics", Vol. 17, 4, pp. 527-544.

G. KALTON (2002), *Models in the Practice of Survey Sampling (Revisited)*, "Journal of Official Statistics", 18, pp. 129-154.

G. KALTON, D. KASPRZYK (1986), *The Treatment of Missing Data*, "Survey Methodology", 12, pp. 1-16.

G. V. KASS (1980), *An Exploratory Technique for Investigating Large Quantities of Categorical Data*, "Journal of Applied Statistics", Vol. 29, 2, pp. 119-127.

ISTAT (2007), *Testing Different Methodologies to Produce Early Estimates of Short-Term Business Indicators*, Final project report, Eurostat, Luxembourg.

R.J.A. LITTLE (1993), *Post-Stratification: a Modeler's Perspective*, "Journal of the American Statistical Association", 88, pp. 1001-1012.

R.J.A. LITTLE, D.B. RUBIN (2002), *Statistical Analysis with Missing Data*, 2nd edition, J.Wiley & Sons, New York.

G. LOOSVELDT, A. CARTON (2001), *An Empirical Test of a Limited Model for Panel Refusals*, "International Journal of Public Opinion Research", 13, pp. 173-185.

J.N.K. RAO, K.P SRINATH, B. QUENNEVILLE B. (1989), *Estimation of Level and Change Using Current Preliminary Data*, In D. KASPRZYK, G. DUNCAN, G. KALTON, M.P. SINGH (eds.), *Panel Surveys*, pp. 457-485, J.Wiley & Sons, New York.

L. RIZZO, G. KALTON, M.J. BRICK (1996), *A Comparison of some Weighting Adjustment Methods for Panel Non-response*, "Survey Methodology", Vol. 22, 1, pp. 43-53.

R.M. ROYALL (1992), *Robustness and Optimal Design Under Prediction Models for Finite Populations*, "Survey Methodology" 18, pp. 179-185.

C.E. SÄRNDAL, B. SWENSSON, J. WRETMAN (1993), *Model Assisted Survey Sampling*, Springer Verlag.

C.E. SÄRNDAL, S. LUNDSTRÖM (2005), *Estimation in Surveys with Nonresponse*, J.Wiley & Sons, New York.

R. VALLIANT, A.H. DORFMAN, R.M. ROYALL (2000), *Finite Population Sampling and Inference – A Prediction Approach*, J.Wiley & Sons, New York.

H. WHITE (1980), *A Heteroschedasticity-Consistent Covariance Matrix Estimator and a Direct test for Heteroschedasticity*, "Econometrica", Vol. 48, pp. 817-838.

I.S. YANSANEH, W.A. FULLER (1998), *Optimal Recursive Estimation for Repeated Surveys*, "Survey Methodology", Vol. 24, 1, pp. 31-40.

SUMMARY

Reducing revisions in short-term business surveys

Timeliness is a driving feature of national economic statistics, especially in a short-term frame. In a survey sampling context, the current practice normally consists in a data release process based on a first preliminary estimate available for users within a short-time, followed by a final estimate, available when the data capturing process is considered completed. The number of preliminary estimates can be higher than one: for each of them the magnitude of *revisions* can be evaluated, on the basis of the difference respect to the final estimate. In this context, according to a model based approach, we propose and compare some preliminary estimation techniques aimed at reducing the average revision. After the definition of the optimal preliminary estimation strategy when the potential non-response bias is ignored, the case when potential differences between *preliminary* and *late* respondents can not be neglected is considered as well, with the proposal of a particular post-stratification procedure. Further, an empirical comparison among various provisional estimation strategies has been carried out on the basis of the quarterly wholesale trade survey carried out by ISTAT (Italian National Statistical Institute) for the period 2003-2006, aimed at estimating quarterly changes of the average turnover. Results show that a proper model specification leads to preliminary estimation techniques characterised by an average revision lower than that got using the actual respondents' sample mean.