STATISTICA, anno LXXIII, n. 2, 2013

IMPROVING EFFICIENCY OF MODEL BASED ESTIMATION IN LONGITUDINAL SURVEYS THROUGH THE USE OF HISTORICAL DATA

Roberto Gismondi[1]
ISTAT, Istituto Nazionale di Statistica, Roma, Italia

1. USEFULNESS AND RISKS OF MODELS IN SURVEY SAMPLING

In last years, the recourse to a model based approach for sampling estimations has become more and more important, as an additional tool compared to design oriented strategies, to increase quality of sample estimates (Kalton, 2002). Still, the use of model based estimations in the context of official statistics is scarce, in Italy as well as in the European Union context. This largely depends on risks due to incorrect knowledge of model parameters and functions which are necessary to implement reliable estimates.

In this paper we will not focus on comparisons between design and model based sampling estimation: a resume of the theoretical properties of a model based approach is found in several manuals (see Särndal *et al*., 1999), usefulness of models for managing non responses is widely discussed in Särndal and Lundström (2005) and is emphasized in the late works by Ibrahim *et al*. (2008) and by Slud and Bailey (2010). However, most of the imputation techniques may not reduce bias enough to balance the increase of variance due to imputation (Watson and Starick, 2011). The calibration approach (Lundström and Särndal, 1999) is also often applied in the official statistics context, especially in the frame of structural business statistics and multipurpose surveys. However, it is not always possible to find proper auxiliary variables with known population totals and strongly correlated with the main target variables, especially in a short-term surveys context.

Based on these assumptions, from now on we will suppose that the main target of the statistical survey lies in the estimation of finite population parameters such as the mean or the total of a variable of interest y for a given finite population.

According to a super-population approach, the optimal estimation strategy is based on the minimisation of the mean squared error (*Mse*) with respect to the model that is supposed to better explain observed data. The main risk relates to the need to identify the fittest model, taking into account that in a given domain of interest more than one model

[1] The opinions herein expressed must be addressed to the author only, as well as possible errors or omissions. All tables derive from elaborations on ISTAT data. Figures in Section 4 have no direct relation with official data released by ISTAT.

may occur. Hedlin *at al.* (2001) underlined the risk of additional bias due to a model miss-specification even when the asymptotically design unbiased *GREG* estimator is used. In particular, the choice of a model implies two main aspects: 1) its mathematical form; 2) the specification of parameters and/or model functions; efficiency gains can result either from a more appropriate model, of from a better parametrization, or both (Lehtonen *et al*., 2003).

Without loss of generality, we will define as *non respondents* all the units whose data are not available at the estimation stage. The non response bias often depends on a model misspecification, e.g. because respondents and not respondents follow different patterns. A late experience regarding employment data (Copeland and Valiant, 2007) showed that non response bias may be not systematic, but could happen for some survey occasions and/or for some domains only.

Models include individual variances and/or variance functions, which should be correctly specified in order to guarantee optimality of the estimation procedure and to avoid bias. However, it is quite difficult to specify the correct model variance function at the single unit level. In this work, given the model mathematical form (aspect 1 before mentioned), we will focus attention only on the possibility to achieve a more reliable estimation of the model unit variances, which are fundamental to guarantee not too biased estimates (aspect 2). This attempt follows other relevant works on the same field, as Laird and Ware (1982) and Knaub (2004). The estimation is completely driven by the availability and the proper use of historical micro-data for the same variable of interest observed in previous surveys, as it is often the case in many real longitudinal survey contexts.

After a brief resume of the basic known results on optimal prediction of a population mean under a linear model (Section 2), alternative estimation techniques are described in Section 3.1. A further more common criterion to obtain better estimates is post-stratification: in Section 3.2 two simple clustering methods are proposed based on the availability of historical micro-data analogous to those used in Section 3.1. Finally, a detailed empirical attempt is proposed in Section 4, with the purpose of comparing the performances of different alternative estimation strategies. This exercise has been carried out in two ways: a) using real historical data; b) on the basis of historical data, with a simulation based on 1000 random replications of pseudo non responses, with the goal of testing the various strategies according to different potential non respondents' profiles. Perspective conclusions have been drawn in Section 5.

2. CLASSICAL MODEL BASED ESTIMATION

Under a model-based approach, quality evaluations can be carried out on the basis of the mean squared error with respect to the particular model underlying observed data. The choice of the model is crucial and its misspecification can lead to serious bias and lack of efficiency.

The next considerations and proposals address the issue of robustness of predictors of population quantities using a two-fold strategy: a) imposing restrictions to the possible super-population models adopted (only the following model (1) will be taken into account); b) using predictors that adaptively consider the possibility that each one out of a series of alternative models is the correct model, focusing on variance estimation at the single unit level.

We can introduce a quite general and known linear model, often used in practice and

further improved according to more complex assumptions (Park and Fuller, 2008). We indicate as i a generic population unit, s is the observed sample, $\overline{s}$ is the not observed population, while the whole population is $U = s \cup \overline{s}$. The main purpose is the estimation of the population mean $\overline{y}_U$. In a longitudinal framework, estimates often refer to months or quarters; however, in this section a time label is not necessary and simpler formulas can be used. For each population unit we suppose the model:

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

where expected values E, variances V and covariances are referred to the model and not to any sampling design, x is an additional variable strongly correlated with y and to be specified, as well as the function v_i (which will be defined as *model unit variance*) with β and σ^2 given, but generally unknown parameters. The model (1) is analogous to the one used by Copeland and Valliant (2007) in a sampling context for employment data where $v=x$ for each unit i. If $\overline{y}_s$ is the sample mean and $f = n / N$, a general linear predictor of the unknown mean is given by:

$$\hat{\overline{y}} = f\overline{y}_s + (1-f)\sum_s \alpha_i y_i = f\overline{y}_s + (1-f)Z_s = f\overline{y}_s + (1-f)\overline{x}_{\overline{s}}\hat{\beta} \tag{2}$$

where Z_s is a predictor of the not observed mean $\overline{y}_{\overline{s}}$ which should guarantee model unbiasedness: $E(\hat{\overline{y}} - \overline{y}_U) = 0$ and the lowest $Mse(\hat{\overline{y}}) = E(\hat{\overline{y}} - \overline{y}_U)^2$. Possible solutions depend on hypotheses on c_{ij} and v_i. Since *Mse* is evaluated according to the model (1), for each n its level depends on the particular sample selected. If we suppose $c_{ij}=0$ for each $(i \neq j)$, it is well known (Cicchitelli *et al*., 1992, 385-390) that the optimal predictor of the not observed y-mean is $\overline{x}_{\overline{s}}\hat{\beta}^*$, where the *BLUP* of β and the *Mse* of the optimal $\hat{\overline{y}}^*$ (that is equal to its model variance) will be, respectively:

$$\hat{\beta}^* = \left(\sum_s \frac{x_i y_i}{v_i}\right)\left(\sum_s \frac{x_i^2}{v_i}\right)^{-1} \tag{3a}$$

$$Mse(\hat{\overline{y}}^*) = \frac{\sigma^2}{N^2}\left[x_{\overline{s}}^2\left(\sum_s \frac{x_i^2}{v_i}\right)^{-1} + v_{\overline{s}}\right] \tag{3b}$$

where $x_{\overline{s}}$ and $v_{\overline{s}}$ are sums over units not in the sample. The recourse to an estimator given by the sample mean is coherent with the homoschedastic model implied by $x_i=v_i=1$ for each i and, as a consequence, the sample mean will be optimal *if and only if* $x=v=1$. On the other hand, when $v=1$, $v=x$ or $v=x^2$ one gets, respectively, that β is estimated by: 1) a regression through the origin, 2) a ratio between means and 3) a mean of ratios y/x, all

based on sample units only. In symbols, we have:

$$\hat{\beta}_1^* = \left(\sum_s x_i y_i\right) \Big/ \left(\sum_s x_i^2\right) \qquad \hat{\beta}_2^* = \frac{\bar{y}_s}{\bar{x}_s} \qquad \hat{\beta}_3^* = \sum_s \frac{y_i}{n x_i}. \tag{4}$$

The main risks underlying the model based prediction based on (3a) are:

1. even though the model (1) is correctly specified, it is not possible to use good estimates of the unit variances v_i in the prediction process;
2. the observed sample includes some units which follow the model defined by (1), but also some units which follow another model (or other models) to be specified. That is coherent with the risk that non responses derive from an *informative* drop-out mechanism (Little, 1995).

A way to reduce risks due to model-based approach is the recourse to a mixed approach, based on both a model and a design driven inference. In particular, under model (1), an alternative robust estimation strategy could be based on the generalised regression estimator (Cicchitelli *et al*., 1992, 399). Given a sampling design with fixed size n and such that all the inclusion probabilities $\pi_i > 0$ for each i, if $\bar{x}_U$ is the x population mean, under model (1) the *GREG* estimator can be written as:

$$\hat{\bar{y}}_{GREG} = \beta^* \bar{x}_U + \sum_s (y_i - \beta^* x_i) / \pi_i \tag{5}$$

where β^* is given by (3a). Let's note that $\hat{\bar{y}}^* = f \hat{\bar{y}}_{GREG} + (1-f)\beta^* \bar{x}_U$. Since β^* is unbiased respect to the model, then $\hat{\bar{y}}_{GREG}$ is unbiased respect to the model as well. This estimator is also asymptotically unbiased respect to the sampling design and its expected sampling variance respect to the model is the lowest in the class of strategies where the predictor is design-unbiased. When a simple random sampling is used, it becomes: $\hat{\bar{y}}_{GREG} = \bar{y}_s + \beta^* (\bar{x}_U - \bar{x}_s)$.

However, it is clear that neither *GREG* can tackle properly the problem due to the lack of knowledge as regards the functional form of the model unit variances v_i. A more problem solving approach could be based on the proper use of individual historical data available in many surveys contexts, which may provide information on the individual "empirical variability" observed along past periods.

The effects due to the not exact knowledge of the function v cannot be known in advance, but of course they are strictly connected to the estimation of the slope parameter β. We can suppose that, instead of the right values v_i, we can use the wrong values z_i in order to implement the formula (3a), in order to estimate the parameter $\hat{\beta}_0$ instead of the right estimate $\hat{\beta}^*$. We can also suppose a certain situation – simplified with respect to the real contexts – where the sample s can be written as: $s = s_1 \cup s_2$, and the relation between the wrong and right unit variances is as follows:

$$z_i = a_1 v_i \quad \text{if} \quad i \in s_1$$
$$z_i = a_2 v_i \quad \text{if} \quad i \in s_2.$$

Of course, if $a_1 = a_2$ then $\hat{\beta}_0 = \hat{\beta}^*$. Otherwise, it can be shown that, putting:

$\omega = \left[\left(\sum_{s_1} \frac{x_i^2}{v_i}\right) + \left(\sum_{s_2} \frac{x_i^2}{v_i}\right)\right]\left[\frac{1}{a_1}\left(\sum_{s_1} \frac{x_i^2}{v_i}\right) + \frac{1}{a_2}\left(\sum_{s_2} \frac{x_i^2}{v_i}\right)\right]^{-1}$ and putting also $\Sigma^{-1} = \left(\sum_{s} \frac{x_i^2}{v_i}\right)^{-1}$, we will have:

$$\hat{\beta}_0 = \hat{\beta}^* + \Sigma^{-1}\left[\left(\frac{\omega}{a_1} - 1\right)\left(\sum_{s_1} \frac{x_i y_i}{v_i}\right) + \left(\frac{\omega}{a_2} - 1\right)\left(\sum_{s_2} \frac{x_i y_i}{v_i}\right)\right] =$$
$$= \hat{\beta}^* + \Sigma^{-1}\left[(\omega_1)(\gamma_1) + (\omega_2)(\gamma_2)\right] = \hat{\beta}^* + \Omega$$

If $a_1 < a_2$ (the relative unit variance estimation error is lower for the units in the sub-sample 1), then $\omega_1 > \omega_2$, with $\omega_1 > 0$ and $\omega_2 < 0$. As a consequence, if $|\omega_1 \gamma_1| > |\omega_2 \gamma_2|$ (as it may happen if the sub-sample 1 includes larger units on average), then $\hat{\beta}_0 > \hat{\beta}^*$, and vice-versa. Only if the combination among the four parameters involved is such that $\omega_1 \gamma_1 \approx -\omega_2 \gamma_2$ we will have $\Omega \approx 0$ and $\hat{\beta}_0 \approx \hat{\beta}^*$, so that the problem due to the need to estimate correctly the model unit variances may be neglected.

It can be easily shown that, if in the estimation process the (wrong) parameter $\hat{\beta}_0$ is used in place of the right parameter $\hat{\beta}^*$, then the absolute value of model bias concerning the correspondent predictor $\hat{\bar{y}}_0$ will be given by:

$$\left|Bias(\hat{\bar{y}}_0)\right| = \left|E(\hat{\bar{y}}_0 - \bar{y})\right| = (1-f)\bar{x}_{\bar{s}} \left|E(\hat{\beta}_0 - \hat{\beta}^*)\right| = (1-f)\bar{x}_{\bar{s}} |\Omega| .$$

3. USE OF HISTORICAL MICRO-DATA

3.1 Model unit variance estimation

In this context, we propose a simple method for estimating each variance component v_i, which uses the real variability of historical data. From the original model (1) we have: $v_i = Var(y_i) / \sigma^2 \approx \hat{Var}(y_i)$, where the last term is an estimate of v_i unless a constant term which in the formula (3a) disappears. A strategy for achieving to the estimate $\hat{Var}(y_i)$ is based on the hypothesis to deal with a sample survey context, for which current estimates refer to a given period *p* of a certain year *T*. A period may be given by a month or a quarter and we can suppose to have *P* periods along a whole year (*P*=12 or *P*=4). Furthermore, a database of historical micro-data, derived from past survey

occasions, is supposed to be available. For sake of simplicity, a first assumption is that the historical database includes all and only the units belonging to the theoretical sample in the year *T*. We also suppose that the database includes micro-data referred to *k* consecutive years before *T*, so that, for each unit, it will contain *k*x*P* observations referred to the *y* variable object of interest. In this framework, for each unit *i* which turns out to be *respondent* as regards the estimation period *p* in the current year *T*, an estimate of the individual model variance will be given by:

$$\hat{Var}(y_{Tpi}) = \sum_{t=T-k}^{T} (y_{tpi} - \hat{\bar{y}}_{Tpi})^2 / (k+1) \quad \text{where:} \quad \hat{\bar{y}}_{Tpi} = \sum_{t=T-k}^{T} y_{tpi} / (k+1). \qquad (6a)$$

The estimation criterion (6a) consists in the calculation, for each unit in the observed sample, of an empirical longitudinal variance based on historical data, where the second function in (6a) is the empirical longitudinal mean. The underlying assumption is that the mean of y_{tpi} is approximately constant over time. The main advantage derived from the recourse to the observed historical variability of individual data is that it avoids any a priori exact - but dangerous - formulas for modelling unit variances, as it is implicitly supposed on the basis of relations (4). On the other hand, the use of (6a) implies that a reliable estimate of the model variance functions v, which refers to a *certain period p of a given year T*, may be approximated by a *longitudinal* estimate, derived from a synthesis of the individual unit variability along time. Even though the available database is supposed to include *k* historical observations concerning the same period *p*, the number of terms in the sums defined in the formula (6a) is equal to $(k+1)$, since we may add to the sum the last observation as well, picked up through the current sample referred to the period *p* of the year *T*. Of course, if the unit *i* is *non respondent* as regards the estimation period *p* in the current year *T*, an estimate of the individual model variance will be based on *k* historical observations only, and will be given by:

$$\hat{Var}(y_{Tpi} / i \notin s) = \sum_{t=T-k}^{T-1} (y_{tpi} - \hat{\bar{y}}_{(T-1)pi/i\notin s})^2 / k \quad \text{where:}$$
$$\hat{\bar{y}}_{(T-1)pi/i\notin s} = \sum_{t=T-k}^{T-1} y_{tpi} / k. \qquad (6b)$$

The estimation criteria (6a) and (6b) can be used if at least two historical observations of the same unit and the same period *p* are available. They may be also applied using a shorter time series, for instance because: a) the observations referred to one or more years are potential outliers which may generate wrong variance estimates; b) the observations related to the most recent years are more useful for achieving to reliable estimates.

It is worth noting that both criteria are built so as to save seasonality of estimates, since each variance is estimated using only historical data referred to the same period *p* (month or quarter), without mixing together different periods data. This option is fully

justified in many short-term surveys contexts[2], where seasonal effects are influential and summing up data referred to *different months or quarters* would mean mixing *different variables*. Of course, this issue disappears if we refer to yearly surveys.

An implicit assumption is that seasonality concerning each month and unit is quite steady along time. Weights inversely proportional to *y*-size may be used to balance the squared differences in (6a) or (6b) if the *y*-magnitude of the same unit along time is quite different from year to year.

The historical variance criterion above mentioned may be further developed according to some considerations. For simplicity, we shall refer to (6a) only, since the only difference with respect to (6b) is the availability of one more period in the sums.

A first modification of the criterion is based on the idea to use more steady estimates of the model variances *v*, less depending on some anomalous individual historical observations. To this purpose, the new estimation criterion (7) is founded on the product, for each unit *i*, of the average empirical longitudinal coefficient of variation (*Cv*) – calculated on the whole available units – say *m* – and which represents the common factor for the estimation of each unit variance function – by the squared longitudinal mean related to the particular reference unit *i*:

$$\hat{v}_{Tpi} = \left[\frac{1}{m} \sum_{j=1}^{m} \sqrt{\frac{\hat{Var}(y_{Tpj})}{\hat{\bar{y}}^2_{Tpj}}} \right]^2 \hat{\bar{y}}^2_{Tpi} = \left[\frac{1}{m} \sum_{j=1}^{m} Cv(y_{Tpj}) \right]^2 \hat{\bar{y}}^2_{Tpi} . \qquad (7)$$

The basic rationale which justifies (7) is the need to reduce as much as possible the risk to obtain some unrealistic model variance estimates through (6a), because of some anomalous micro-data in the historical database. It is important to note that, on the basis of a slight adaptation of (7), it would be possible to estimate the model variance function *v* even for a unit for which time series includes only one observation (on the basis of which individual variability could not be estimated otherwise). If the only available observation related to the period *p* is referred to the generic year *t*, it is enough to use (7) putting y_{tpi} in place of $\hat{\bar{y}}_{Tpi}$.

A second kind of modification stems from the idea that the calculation of an empirical variance may be more reliable if a larger number of observations is used. A simple method for increasing the number of addenda consists in removing the seasonality constraint which has been implicitly supposed in the definition of (6a) and (6b). Even though this constraint is recommended in short-term surveys contexts, the counterbalance is given by the larger number of observations used for estimating variances. In symbols, we would have:

$$\hat{Var}(y_{Ti}) = \sum_{t=T-k}^{T} \sum_{p=1}^{P} (y_{tpi} - \hat{\bar{y}}_{Ti})^2 / P(k+1) \quad \text{where:} \qquad (8)$$

[2] For instance, the industrial production index, the industrial turnover index, the retail trade index managed by ISTAT. They all are monthly indicators ($P = 12$).

$$\hat{\bar{y}}_{Ti} = \sum_{t=T-k}^{T} \sum_{p=1}^{P} y_{tpi} / P(k+1) .$$

Let's note that if the criterion (8) is used, then the *same* model unit variance estimates for any reference period p will be used. This assumption may be unrealistic if the y-variable follows a clear seasonal pattern. A particular case when the criterion (8) is quite recommended if $k=1$, e.g. when the time series of historical data is very short and contains one complete year only: the original criterion (6a) may still be used, but on the basis of two only observations. On the other hand, a criterion quite similar to (8) is absolutely mandatory if we deal with a completely new survey: the individual variances may be estimated starting from the period $p=2$, mixing in the variance formula the observations related to the periods $p=1$ and $p=2$.

A third alternative criterion starts from the calculation of whatever kind of empirical variance estimates (6a), (7) or (8), say $\hat{Var}(y_{Tpi})$. Since the series of these empirical estimates may contain some outlier data, we could use them for the estimation of the parameters a and b of the model: $\hat{Var}(y_{Tpi}) = a x_{Tpi}^{b}$. After the logarithmic linearization, *OLS* estimates $\hat{a}$ and $\hat{b}$ may be used for the calculation of the new model based variance estimates given by:

$$\hat{\hat{Var}}(y_{Tpi}) = \hat{a} x_{Tpi}^{\hat{b}} \tag{9}$$

Similarly to criterion (7), criterion (9) is driven by the idea to reduce the number of unrealistic model unit variance estimates as much as possible. Another advantage derived by (9) is that it can provide an estimate of the model unit variance even for those units for which criteria (6), (7) or (8) cannot be applied, for instance because no (historical) data are available, except for the only value x_{Tpi} (which may be given by $x_{Tpi} = y_{(T-1)pi}$).

3.2 Post-stratification

According to observed data, within a given reference domain different models may exist, since individual expected values and/or variances could follow different patterns. Post-stratification may be used as a tool to reduce the model misspecification bias.

Broadly speaking, the issue of post-stratification is connected with the non response problem. Following the classical approach (Särndal and Lundström, 2005, 94-96), the design bias of a post-stratified estimator can be strongly reduced if in each post-stratum the y-mean of respondents and non respondents are quite similar and different post-strata are characterised by different average response rates. Furthermore, post-stratification is an important tool for testing the presence of different sub-populations in which model variances may be approximately homoschedastic. Of course, in this case the problem concerning estimation of model variance functions would disappear.

Beyond the several applications concerning the need to find an optimal post-stratification (Block and Segal, 1989; Djerf, 1997; Dorfman and Valliant, 2000), in this context we propose two simple procedures which are fully coherent with the operational

context described above: an adaptation of the well known Dalenius-Hodges method and a new procedure, driven by the empirical variance estimation techniques described in Section 3.1. The estimation techniques resumed in Section 2 and the model unit variance estimation criteria proposed in Section 3.1 can be applied separately in each post-stratum.

The method proposed by Dalenius and Hodges (1959) to stratify a population, in order to minimise the variance of estimates in a stratified random sampling context, is based on the hypothesis to know the values z_{Tpi} assumed on the *i-th* population unit by an auxiliary variable *z* correlated with *y*. For each period *p* in the year *T*, putting as U_T the reference population and as *i'* the place of the *i-th* unit in the decreasing ranking of *z*-values, the rule for identifying boundaries of *r* sub-populations is based on the following equality:

$$\sum_{i' \in U_{Tph}} \sqrt{z_{Tpi'}} \approx \sum_{i' \in U_T} \sqrt{z_{Tpi'}} / r \,. \tag{10}$$

where U_{Tph} is the *h-th* sub-population including N_{Tph} units. Rule (10) means that the sum of the square roots of the *z*-values in the *h-th* sub-population must be almost equal to the same sum calculated in each of the other (*r*-1) sub-populations. The variable *z* may be given by the same variable *x* in model (1): for instance, *z* may be equal to $x_{(T-1)p}$, or to a proper synthesis of past values. However, even though the Dalenius-Hodges method is quite simple, it may not properly take into account individual variability.

In order to consider variability of historical data as well, an alternative procedure may be defined as follows. The main goal is still the use of a cluster analysis algorithm aimed at identifying *r* post-strata. Let's suppose the model (1) as formally suitable for describing observed data, a reference time *T* whose estimates refer to and the availability of historical data for *k* survey occasions before *T*, as supposed along Section 3.1. The sub-populations are supposed to be characterised by different latent levels of β and σ. As a consequence, the data matrix which may be used for clustering contains on the rows the single population units, while column values are the modalities of the two new variables z_{T1} and z_{T2} defined as follows:

$$z_{1\,pi} = \frac{1}{k} \sum_{t=T-k}^{T} \frac{y_{tpi}}{x_{tpi}} \approx \hat{\beta}_{Tp(i)} \tag{11a}$$

$$z_{2\,pi} = \sqrt{\sum_{t=T-k}^{T} (y_{tpi} - \hat{\bar{y}}_{Tpi})^2 \Big/ \sum_{t=T-k}^{T} v_{tpi}} \approx \hat{\bar{\sigma}}_{Tp(i)} \tag{11b}$$

where: $\hat{\bar{y}}_{Tpi} = \sum_{t=T-k}^{T} y_{tpi} / (k+1)$. The first variable is an estimate of the "average slope" which characterises the *i-th* unit along the $(k+1)$ survey occasions. The second variable is an estimate of the "average standard deviation" which characterises the same unit in the same time lag. Its formal structure derives from the variance model formula referred to a generic time t: $\sigma_t^2 = Var(y_{ti}) / v_{ti}$, from which an estimate of the average variance which

characterises the *i-th* unit along the k past periods is: $\hat{\tilde{\sigma}}^2_{Tp(i)} = \hat{Var}(y_{Tpi}) / \hat{\tilde{v}}_{Tpi}$, where $\hat{Var}(y_{Tpi}) = \sum_{t=T-k}^{T} (y_{tpi} - \hat{\tilde{y}}_{Tpi})^2 / k$ and . The basic rationale is that each cluster should contain the units which are more similar to each other in terms of average slope and variability level through the recent past. $\hat{\tilde{v}}_{Tpi} = \sum_{t=T-k}^{T} v_{tpi} / (k+1)$.

From an operational point of view, use of the two only synthetic variables (11a) and (11b) can be preferred to a data matrix containing $(k+1)$ variables – given by k single periods slope estimates y_{ti} / x_{ti} and the "average variance" estimate – especially when single period slopes are quite unsteady and may contain dangerous outliers, and/or domains under study include a low number of units. After the identification of clusters, one or more estimation criteria can be applied separately inside each of them. When research is limited to two clusters, if one of the two clusters include only one unit which is non respondent, then the estimate of its *y*-level can be put equal to that obtained in the frame of the correspondent not post-stratified estimation strategy.

4. APPLICATION

4.1 A database derived from the istat wholesale trade quarterly survey

Since the first quarter of 2001, ISTAT (the Italian National Statistical Institute) elaborates and releases quarterly index numbers on turnover of the "Wholesale trade and commission trade sector" (classification NACE Rev.2 division 46). Provisional estimates are released 60 days from the end of the reference quarter, final indexes are released after 180 days. The actual time series of micro-data cover the period 2001-2012.

The sampling survey is based on a stratified random sample of about 7.500 enterprises, where strata are obtained crossing 9 economic activities and 3 employment classes (1-5; 6-19; >19). On the basis of elementary strata indexes, calculations of higher order indexes – among which the total wholesale trade index – are based on weighted means of lower order indexes, where weights derive from structural business statistics. 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 assumption adopted in the estimation process is that non responses (and the same late responses as well) follow a *missing at random* (*MAR*) pattern that is the theoretical justification of the recourse to the current estimator given by the ordinary sample mean both for provisional and final estimates.

In this context, attention has been addressed to quarterly estimation of turnover means (instead of indexes). To this purpose, a longitudinal database has been built up, including all and only the units belonging to the theoretical sample in each of the 6 years considered (from 2005 to 2010), e.g. *panel* units. The survey sampling design foresees a yearly partial rotation of units (to reduce response burden), so that each yearly sample includes 10% of new units with respect to the previous year sample. Since the main consequence of the rotation process is a number of panel units lower that the average

yearly sample size, in order to deal with more populated strata the 9 original economic activities have been collapsed and reduced to 4: 1) Wholesale on a fee or contract basis; 2) Agriculture raw materials and live animals; 3) Food, beverages, tobacco, household goods; 4) Non agriculture intermediate products, machinery, equipment and supplies, other products. The final stratification of original micro-data led to 12 strata (4 economic sectors by 3 employment classes).

The database included the following variables: identification and stratum codes, quarterly turnover from first quarter 2005 until fourth quarter 2010, quarterly binary variable – concerning the 4 quarters 2010 – equal to 1 if a unit was respondent within 60 days or to 0 otherwise. For further analyses, we considered only not outlier observations[3].

On the basis of model (1), a crucial aspect concerned the choice of the auxiliary x variable. Possible choices could have been given, for each enterprise, by the yearly turnover or the number of persons employed derived from the business register (both variables refer to the year before that under observation), turnover referred to the previous quarter or to the same quarter of the previous year. The empirical evidence[4] showed a stronger correlation between turnover in the quarters p and $(p-4)$, so that $x_p = y_{p-4}$.

TABLE 1

Number of final respondents (average 2010), share of quick respondents and coefficient of variation of turnover for the panel of wholesale trade enterprises included in the comparative exercise for testing efficiency of various estimation strategies

Domain	Average number of respondents in 2010	Average % of quick respondents	Turnover coefficient of variation
Total	4395	79.1	
1: Wholesale on a fee or contract basis	557	74.0	1.96
2: Agriculture raw materials, live animals	345	74.1	1.99
3: Food, beverages and household goods	1957	78.8	2.73
4: Other products	1536	82.4	4.49

Note: figures are averages of 4 quarters.

The comparative exercise herein discussed is grounded on the following rationale: we consider as the main object of estimation the *final* sample mean (based on the N final respondents) and as estimator the prediction $\hat{\bar{y}}$ based on the only n *quick* respondent units. Knowledge of the true *final* sample mean led to the calculation of the real prediction error referred to the estimation strategies resumed in Section 4.2. The y means object of estimation concern the 4 2010 quarters, the auxiliary x variables are turnover data referred to the 4 2009 quarters, while the quarterly turnover data 2005-2010 have been used for implementing the empirical variance estimation criteria described in Section 3.1.

[3]Units with a yearly turnover lower than euro 1000 have been excluded. Two main longitudinal controls were activated at the single unit level: 1) the ratio between turnover referred to quarters p and (p-4) must range between 0.1 and 10; 2) the ratio between the highest turnover between p and (p-4) and the yearly turnover of the previous year must range between 0.05 and 20.

[4] See also Gismondi (2008).

The average number of final respondents in the 4 quarters 2010 was 4395 (Table 1). The number of final respondents ranged from 345 for domain 2 (Agriculture raw materials and live animals) up to 1957 for domain 3 (Food, beverages and household goods).

The relative share of quick respondent units on final respondents was 79.1%. The share of quick respondents was also significantly lower in domains 1 and 2 rather than in domains 3 and, in particular, in domain 4. However, for domain 4 the larger quick response rate is counterbalanced by the very high relative variability of turnover, since the correspondent coefficient of variation is equal to 4.49[5]. It is worthwhile to note that lower quick response rates are coupled with lower coefficients of variation.

The observed context seems to be particularly suitable for testing estimation strategies in presence of significant and realistic non response rates, as it often occurs in other real survey frameworks.

4.2 Compared strategies

A wide set of different estimation techniques has been selected and tested. They have been combined with post-stratification too. The main goal consists, for each 2010 quarter and each domain, in using response from quick respondents to estimate the turnover mean which will include late respondents as well.

Two main estimators have been used and compared: the optimal model based predictor (*OMBP*) defined by (2) and (3a) and the *GREG* estimator (5). They have been implemented according to various criteria for estimating model unit variance. Each combination between estimator and criterion for variance estimation will be defined as a *strategy*.

It is worth remarking that a first estimation strategy tested is the simplest one and, for each quarter, is given by the y sample mean calculated on quick respondents. As already mentioned in Section 2, it is equivalent to the model based estimator founded on (3a) when $v_i = x_i = 1$ for each unit i. The efficiency of this strategy was always quite poor, especially in comparison with performances of other strategies. Still, it is an important benchmark for assessing the usefulness of more complex techniques.

The groups of criteria for estimating model unit variances are the following ones:

1. the first group of criteria is derived from the optimal prediction under model (1), still based on (2), with the common positions $v = 1$, $v = x$ and $v = x^2$ for each unit. The correspondent predictors are listed in the formula (4).

2. The second group of criteria is based on the argumentations provided in Section 3. In particular, we have taken into account the following criteria: (6a), labelled as *5 years*, which uses for the estimation of the variances the five past observations from 2005 to 2009 plus the last observation in 2010 ($k = 5$, 6 observations overall for quick respondents; 5 observations overall for late respondents). Of course the formula to be used becomes (6b) for units which are late respondents in the reference 2010 quarter. The same formulas have been implemented using only the most recent past data from 2007 to 2009 ($k = 3$, 4 observations overall for quick respondents; 3 observations overall for late respondents). This criterion has been

[5] The coefficients in Table 1 are means of 4 quarters and 3 employment classes.

labelled as *3 years*. Model (8) has been implemented too, testing the hypothesis of not seasonality of variances (for this reason the label adopted in the following tables is *No season*). Finally, also method (9) has been tested (labelled as *Model*).

3. A third criterion, labelled as *Pseudo best*, has been defined as follows and tested. For each of the strategies listed above it was possible to calculate the mean of absolute per cent errors of estimates (*MAPE*), evaluated only on quick respondent units. This is possible thanks to the knowledge of the true values newly estimated through the specific technique tested. For each domain and quarter, the best strategy was the one with the lowest mean of errors. Of course, that is a "pseudo" best strategy, as, when applied to late respondents, it does not guarantee that the mean of per cent estimate errors is still the lowest among the various criteria. Risks of scarce efficiency will be high when the response patterns of quick and late respondents are quite different. The identification of the pseudo best strategy has been carried out for each of the 4 domains.

TABLE 2
Estimation strategies tested and compared in the empirical attempt

Code	Model unit variances estimation criteria	Estimators applied
(1)	$v=x=1$	Sample mean calculated on quick respondents
(2)	Homoschedasticity ($v=1$)	Optimal model based predictor defined by (2) and (3a) *GREG* estimator (5)
(3)	Model unit variances proportional to size $v=x$	Optimal model based predictor defined by (2) and (3a) *GREG* estimator (5)
(4)	Model unit variances proportional to squared size $v=x^2$	Optimal model based predictor defined by (2) and (3a) *GREG* estimator (5)
(5)	Model unit variances estimated using the whole time series of past data (5 years); use of formulas (6a) and (6b) with $k=5$	Optimal model based predictor defined by (2) and (3a) *GREG* estimator (5)
(6)	Model unit variances estimated using a part of the time series of past data (3 years); use of formulas (6a) and (6b) with $k=3$	Optimal model based predictor defined by (2) and (3a) *GREG* estimator (5)
(7)	Model unit variances estimated through a model based on past data; use of formula (8)	Optimal model based predictor defined by (2) and (3a) *GREG* estimator (5)
(8)	Model unit variances estimated through a model based on past data; use of formula (9)	Optimal model based predictor defined by (2) and (3a) *GREG* estimator (5)
(9)	For each domain the estimation criterion characterised by the lowest *MAPE* calculated on quick respondents is used; pseudo best criterion	Optimal model based predictor defined by (2) and (3a) *GREG* estimator (5)
Code	Additional options combined with all the previous estimation strategies	
(I)	Post-stratification coupled with all the techniques from (1) to (9); method defined by (11a) and (11b)	

Further, both the post-stratification criteria defined by (10) and the couple (11a)-(11b) have been tested with the Ward clustering method, using the whole past data time series ($k=5$) and putting the number of post-strata $r=2$. In this way the number of sub-domains into which basic estimates have been calculated passed from 12 (4 domains by 3

employment classes) to 24. However, the next tables 5, 6, 10 and 11 include only the results achieved with the post-stratification criterion based on (11a) and (11b), which led to better results and is also the most coherent with the argumentations proposed along Section 3. Post-stratification has been coupled with all the basic estimation criteria described in the previous points from 1) to 3). A resume of the estimation strategies compared is provided in Table 2.

Basically, goodness of estimates has been evaluated on the basis of the synthetic error measure given by the mean of absolute per cent errors. For each of the 4 domains D which identify the main economic activity, and each quarter p of 2010, the general estimator of the unknown mean based on both quick and late respondents is: $\hat{\bar{y}}_{2010,p,D} = \sum_{d=1}^{3} \hat{\bar{y}}_{2010,p,D,d} W_{D,d}$, where the label d identifies the employment class and the exogenous weights W are such that: $\sum_{d=1}^{3} W_{D,d} = 1$. The mean of absolute per cent errors will be given by:

$$MAPE_{2010,D} = \sum_{p=1}^{4} \frac{\left| \bar{y}_{2010,p,D} - \hat{\bar{y}}_{2010,p,D} \right|}{4 \hat{\bar{y}}_{2010,p,D}} \cdot 100 \,. \tag{12}$$

The empirical outcomes (12) have been reported in the rows labelled from 1 to 4 in all the tables from 3 to 6 and from 8 to 11. Moreover, for each quarter p of 2010, the general estimator of the wholesale trade sector mean, based on both quick and late respondents, is: $\hat{\bar{y}}_{2010,p} = \sum_{D=1}^{4} \hat{\bar{y}}_{2010,p,D} W_D$ where: $\sum_{D=1}^{4} W_D = 1$. The consequent mean of absolute per cent errors is:

$$MAPE_{2010} = \sum_{p=1}^{4} \frac{\left| \bar{y}_{2010,p} - \hat{\bar{y}}_{2010,p} \right|}{4 \hat{\bar{y}}_{2010,p}} \cdot 100 \,. \tag{13}$$

The empirical outcomes (13) have been shown in the last row in all the tables from 3 to 6 and from 8 to 11.

4.3 Main results from the empirical attempt

The main results derived from the application to real data have been resumed in the tables from 3 to 6. As a matter of fact, the strategy (1), given by the sample mean of quick respondents, led to the worst estimates for all the domains and has been excluded from the

following tables[6].

In tables from 3 to 6 and from 8 to 11, for each estimation domain the best strategies (corresponding to the lowest *MAPEs*) have been pointed out as follows:

- a bold figure marks the best strategy for each estimation domain (by rows);
- an underlined figure marks the cases when post-stratification improves the corresponding strategy without post-stratification;
- an asterisk is put before cases when *GREG* improves *OMBP*.

A first outstanding and clear result (Tables 3 and 4) is that the new strategies from (5) to (8) performed better than the classical strategies from (2) to (4), for any estimation domain and for both estimators compared, e.g. *OMBP* and *GREG*, with the only exception for *GREG* and domain 2, where the latter estimator improves the former using the criterion (3) $v=x$ (Tables 3 and 4).

Among the 10 estimation domains taken into account (the 4 domains plus the total wholesale, e.g. 5 for *OMBP* and 5 for *GREG*), the criterion (5) – *5 years* – was the best one in 5 cases, the criterion (6) – *3 years* – was the best one in 3 cases, while the criteria (7) – *No season* – and (3) – $v=x$ – were the best ones in 1 case each (the latter criterion is the only one optimal when coupled with *GREG*, as already underlined).

Overall, the best criterion turned out to be the number (5) – *5 years* – since:

- when *OMBP* is used, it is the optimal one for domain 4 and total wholesale, is the second best for domains 1 and 3 and is the third best for domain 2;
- when *GREG* is used, it is the optimal one for domain 3, 4 and total wholesale, is the second best for domains 1 and is almost the best one for domain 2 as well, if compared with the other new criteria only.

Among the 3 classical criteria, the criterion (3) $v=x$ is clearly the best one when *OMBP* is used, while there is not a clear best classical criterion using *GREG*.

As regards the single estimation domains:

1. Domain 1 is that for which the new criteria led to the largest efficiency gains if compared with the classical ones when *OMBP* is used, as well as – even though at a lower extent – with *GREG*. This outcome implies that the historical micro-data available for enterprises operating in the wholesale on a fee or contract basis sector are able to produce reliable estimates of the model unit variance functions to be specified in the model (1).
2. As regards domain 2, when *OMBP* is used the same considerations as the previous domain 1 hold; on the other hand, using *GREG* no significant efficiency gains have been obtained, and this outcome may imply that at least in this domain the recourse to historical micro-data should be coupled with a full model based estimation procedure.

[6] MAPE for the 4 domains was, respectively: 8.14, 4.10, 9.12, 11.61. MAPE for the total wholesale trade was 9.36.

3. Domain 3 is the one for which the efficiency gains due to the new criteria and *OMBP* are the poorest, when compared with classical criteria as $v=x$ or $v=x^2$. In this case the historical micro-data do not produce unit variance estimates more reliable than those defined through the classical criteria. Larger efficiency gains can be obtained using *5 years* and *GREG*.
4. Finally, for domain 4 the same conclusions as for domain 1 may be drawn, but with an important difference, since no efficiency gains have been obtained with the criterion *3 years*. Both *OMBP* and *GREG* should be coupled with *5 years*, so that for this domain long time series of historical micro-data are recommended.

Broadly speaking, the usefulness of the new criteria is evident in all domains and is not strictly dependent on the ratio between quick and final respondents (compare Table 1). The largest efficiency gains derived from *OMBP* have been obtained in domains 1 and 2, characterized by the lowest relative variability of quarterly turnover.

On average, the recourse to post-stratification confirmed the better performance of the new criteria (Tables 5 and 6). Among the 10 estimation domains taken into account, the criterion (7) – *No season* – was the best in 7 cases, the criterion (5) – *5 years* – was the best one in 3 cases, while the criterion (9) – *Pseudo best* – was the best one in 2 cases and the criteria (4) – $v=x^2$ – and (6) – *3 years* – were the best ones in 1 case each (the former criterion is optimal when coupled with *GREG*). Moreover, post-stratification enforced the better performance of *OMBP* compared with *GREG*, since the former led to the lowest *MAPE* for all the 10 estimation domains.

Among the 13 optimal criteria identified (for 3 domains two criteria have the same lowest *MAPE*), post-stratification led to an optimal strategy (improving the correspondent not post-stratified strategy) in 3 domains both for *OMBP* (3, 4 and total wholesale) and *GREG* (1, 3 and total wholesale). As a consequence, post-stratification has proved not useful only for domain 2 (no efficiency gain); among the domains for which efficiency gains have been obtained the largest *MAPE* concerned domain 1 (*MAPE*=1.12 for *OMBP* and *MAPE*=1.36 for *GREG*).

A tentative explanation of these results may be obtained from Table 7, where for each domain we have reported the empirical coefficients of variation (*cv*) concerning the variables z_1 (average slope, formula (11a)) and z_2 (average standard deviation, formula (11b)) - on which the clustering algorithm described in Section 3.2 is based. Calculations have been done using the whole available historical database. As a matter of fact, domain 2 (no usefulness of post-stratification) is characterized by the largest *cv* of z_2 (6.32) and the second largest *cv* of z_2 (0.69), and this high unit variability for both the key variables used for clustering may explain the poor results of post-stratification. On a lower extent, the largest *cv* concerning z_1 which characterises domain 1 (0.85) may explain its largest *MAPE* even when an optimal strategy (based or not on post-stratification) is used.

It is worth underlining the better performance of the criterion *No season* when post-stratification is used, instead of *5 years*. Since through *No season* model unit variances are estimated using the largest number of observations, we can conclude that the additional support provided by post-stratification is particularly important when the huge longitudinal variability of the *y*-variable may objectively compromise the correct estimation of unit variances based on too short time series.

In order to assess steadiness of results, an additional simulation study has been carried out. Starting from the same database, late responses have been randomized 1000 times, imposing 20% of late responses at each replication. In this way, randomization is carried

out supposing a late response rate quite similar to the average 79.1% which characterizes real 2010 data (Table 1). Selection of late respondents has been carried out at random for each quarter and each sub-domain (4 domains by 3 employment classes), assigning to each unit the same probability (0.8) to be or not to be quick respondent.

TABLE 3
Comparison among MAPEs obtained through the use of the optimal model based predictor and different criteria for model unit variance estimation
Mean of 4 quarters – Real data referred to 2010

Domains	Basic model based prediction			New criteria for estimating individual variance				*Pseudo best*
	$v=1$	$v=x$	$v=x^2$	*5 years*	*3 years*	*No season*	*Model*	
1	1.52	1.15	1.86	0.80	0.90	**0.72**	1.04	1.52
2	1.92	1.76	1.98	1.72	**1.48**	1.54	1.78	1.72
3	2.20	1.15	1.05	1.14	**1.04**	1.20	1.29	1.20
4	1.42	1.97	2.58	**1.22**	2.25	1.29	1.87	**1.22**
Total	1.06	1.07	1.55	**0.68**	1.17	0.70	1.00	0.78

Legenda 1: Wholesale on a fee or contract basis; 2: Agriculture raw materials and live animals; 3: Food, beverages and household goods; 4: Other products; Total: Total wholesale trade. The definition of criteria is available in Table 2.

TABLE 4
Comparison among MAPEs obtained through the use of GREG estimation and different criteria for model unit variance estimation
Mean of 4 quarters – Real data referred to 2010

Domains	Basic model based prediction			New criteria for estimating individual variance				*Pseudo best*
	$v=1$	$v=x$	$v=x^2$	*5 years*	*3 years*	*No season*	*Model*	
1	2.19	2.13	1.51	1.84	**1.49**	1.59	1.86	2.19
2	1.97	***1.45**	2.13	2.04	2.13	2.05	2.02	2.04
3	4.23	2.97	2.73	**2.29**	2.74	2.44	3.16	2.44
4	1.70	1.92	2.13	**1.48**	1.98	1.52	1.88	**1.48**
Total	1.53	1.22	1.26	**1.01**	1.13	1.08	1.32	1.13

See *legenda* of Table 3. The definition of criteria is available in Table 2.

Again, randomization confirms that the new strategies from (5) to (8) performed better than the classical strategies from (2) to (4), for any estimation domain and for both estimators compared, with the only exception for *GREG* and total wholesale, where the lowest *MAPE* has been obtained using the criterion (3) $v=1$ (Tables 8 and 9).

Among the 10 estimation domains, the criterion (5) – *5 years* – was the best one in 4 domains – always when *OMBP* is used – while the criterion (7) – *No season* – proved the best in 5 cases – of which 4 when coupled with *GREG*. The former result is quite coherent with the outcomes already analysed when the true late response rates are used (Tables 3 and 4), even though from the randomized approach the role of *OMBP* predictor is enforced. On the other hand, the latter outcome is slightly different, since the optimality of *No season* is outstanding (in place of 3 years). In short, a basic final conclusion is absolutely in favour of the strategy *OMBP* and *5 years*, which led to the lowest *MAPE* for

any domain, with the partial exception of domain 1 (for which this strategy is second best after *OMBP* and *No season*). If *GREG* is used, for any domain the best strategy is based on *No season* (estimation of model unit variances using a larger number of individual micro-data than *5 years*), but it is always less efficient than the best strategy obtained with *OMBP*.

TABLE 5
Comparison among MAPEs obtained through the use of the optimal model based predictor and different criteria for model unit variance estimation
Mean of 4 quarters – Real data referred to 2010 – Estimates with post-stratification

Domains	Basic model based prediction			New criteria for estimating individual variance				*Pseudo best*
	$v=1$	$v=x$	$v=x^2$	*5 years*	*3 years*	*No season*	*Model*	
1	1.72	1.31	1.54	**1.12**	1.21	1.15	1.75	1.15
2	1.96	1.84	2.02	1.71	1.55	**1.52**	1.94	1.71
3	1.91	1.12	1.11	1.11	1.43	**0.88**	1.22	**0.88**
4	1.36	1.56	2.12	1.13	2.11	**0.92**	1.36	**0.92**
Total	0.81	0.82	1.41	0.64	1.10	**0.58**	0.89	0.63

See *legenda* of Table 3. The definition of criteria is available in Table 2.

TABLE 6
Comparison among MAPEs obtained through the use of the GREG estimator and different criteria for model unit variance estimation
Mean of 4 quarters – Real data referred to 2010 – Estimates with post-stratification

Domains	Basic model based prediction			New criteria for estimating individual variance				*Pseudo best*
	$v=1$	$v=x$	$v=x^2$	*5 years*	*3 years*	*No season*	*Model*	
1	1.96	1.75	1.62	**1.36**	1.44	**1.36**	1.88	1.44
2	2.12	1.97	**1.67**	1.86	1.75	1.88	2.01	1.86
3	1.56	1.31	1.28	1.15	**1.12**	1.16	1.16	1.16
4	1.32	1.65	1.98	1.13	1.79	**1.04**	1.43	1.32
Total	0.81	0.99	1.14	**0.71**	0.96	**0.71**	0.90	0.76

See *legenda* of Table 3. The definition of criteria is available in Table 2.

On average, the recourse to post-stratification confirmed the better performance of the new criteria as well (Tables 10 and 11). Among the 10 estimation domains taken into account, the criterion (5) – *5 years* – was the best in 5 cases, the criteria (6) – *3 years* – and (8) – *Model* – were the best ones in 2 case each, while the criterion (7) – *No season* – was the best one in 1 case. As for the comparison between *OMBP* and *GREG*, post-stratification fully confirmed the better performance of the former, according to the outcomes derived from Tables 5 and 6. The optimal strategies were always based on the use of *OMBP* for any domain, with the only exception of domain 2 – for which the compared *MAPEs* are in any case quite similar: 1.11 against 1.24.

Post-stratification based on the longitudinal means and standard deviations (11a) and (11b) confirmed to be a useful tool for reducing *MAPE*, as already seen as regards results obtained using real late response rates. When the lowest *MAPEs* concerning post-stratified and not post-stratified estimators are compared, a perfect balance is obtained, since for 5 estimation domains on 10 post-stratification improves the optimal results obtained

without post-stratification. That happens twice when *OMBP* is used (domain 3: *MAPE*=0.61 when post-stratification is used and *MAPE*=0.65 when not; total wholesale: *MAPE* is equal to 0.27 and 0.52 respectively), and three times when *GREG* is used (domain 3: *MAPE*=0.87 when post-stratification is used and *MAPE*=0.95 when not; domain 4: *MAPE* is equal to 0.69 and 0.88 respectively; total wholesale: *MAPE* is equal to 0.68 and 0.73 respectively).

TABLE 7
Average slope (z_1) and average standard deviation (z_2), standard deviation (z_1 and z_2) and coefficient of variation (z_1 and z_2) in each domain
The variables z_1 and z_2 have been defined through formulas (11a) and (11b) – Mean of 4 quarters – Real data referred to 2010

Domain	Average in the group*		Standard deviation in the group		Coefficient of variation in the group	
	z_1 (11a)	z_2 (11b)	z_1 (11a)	z_2 (11b)	z_1 (11a)	z_2 (11b)
1: Wholesale on a fee or contract basis	1.22	241	1.03	1156	0.85	4.79
2: Agriculture raw materials, live animals	1.13	1308	0.77	8267	0.69	6.32
3: Food, beverages and household goods	1.07	1314	0.46	8302	0.43	5.22
4: Other products	1.17	401	0.36	1909	0.31	4.76

* The standard deviations z_2 are in thousand euro.

TABLE 8
Comparison among MAPEs obtained through the use of the optimal model based predictor and different criteria for model unit variance estimation
Mean of 4 quarters – 1000 random replications referred to 2010

Domains	Basic model based prediction			New criteria for estimating individual variance				*Pseudo best*
	$v=1$	$v=x$	$v=x^2$	*5 years*	*3 years*	*No season*	*Model*	
1	2.24	1.71	2.85	1.43	1.85	**1.32**	1.75	1.43
2	2.11	1.60	1.83	**0.86**	1.06	0.87	1.72	1.06
3	3.25	1.20	1.15	**0.65**	0.81	0.71	1.44	0.71
4	1.98	1.08	3.13	**0.34**	1.50	0.37	0.69	**0.34**
Total	1.25	0.94	1.16	**0.52**	1.07	0.65	1.09	0.65

Legenda 1: Wholesale on a fee or contract basis; 2: Agriculture raw materials and live animals; 3: Food, beverages and household goods; 4: Other products; Total: Total wholesale trade. The definition of criteria is available in Table 2.

A joint reading key of both applications (true data and randomization) may be resumed in three main conclusions: a) the new criteria for estimating model unit variances can be preferred to the classical ones; in particular, a larger number of historical micro-data (*5 years* or *No season*) should be used; b) *OMBP* performs better than *GREG*; c) post-stratification based on average longitudinal slope (11a) and average longitudinal standard deviation (11b) is a useful tool for increasing efficiency of estimates.

TABLE 9
Comparison among MAPEs obtained through the use of GREG estimation and different criteria for model unit variance estimation
Mean of 4 quarters – 1000 random replications referred to 2010

Domains	Basic model based prediction			New criteria for estimating individual variance				*Pseudo best*
	$v=1$	$v=x$	$v=x^2$	*5 years*	*3 years*	*No season*	*Model*	
1	2.72	2.10	2.70	1.65	2.47	**1.65**	1.70	**1.65**
2	1.42	1.62	1.77	1.22	1.33	**1.02**	1.46	1.46
3	2,27	1.20	1.39	1.01	1.04	**0.95**	1.41	1.04
4	1.80	1.37	2.91	0.98	1.55	**0.88**	1.02	0.98
Total	0.73	1.03	1.59	0.76	1.11	0.83	1.05	0.83

See *legenda* of Table 8. The definition of criteria is available in Table 2.

TABLE 10
Comparison among MAPEs obtained through the use of the optimal model based predictor and different criteria for model unit variance estimation
Mean of 4 quarters – 1000 random replications referred to 2010 – Estimates with post-stratification

Domains	Basic model based prediction			New criteria for estimating individual variance				*Pseudo best*
	$v=1$	$v=x$	$v=x^2$	*5 years*	*3 years*	*No season*	*Model*	
1	2.55	2.15	3.12	2.21	2.65	2.25	**1.89**	2.25
2	2.20	2.33	2.25	1.32	**1.24**	1.47	1.58	1.58
3	2.23	1.10	0.83	**0.61**	0.77	0.78	1.31	0.78
4	0.71	1.24	3.22	**0.44**	1.47	0.47	0.87	0.71
Total	0.87	0.79	1.18	**0.27**	0.88	0.33	0.94	0.33

See *legenda* of Table 8. The definition of criteria is available in Table 2.

TABLE 11
Comparison among MAPEs obtained through the use of the GREG estimator and different criteria for model unit variance estimation
Mean of 4 quarters – 1000 random replications referred to 2010 – Estimates with post-stratification

Domains	Basic model based prediction			New criteria for estimating individual variance				*Pseudo best*
	$v=1$	$v=x$	$v=x^2$	*5 years*	*3 years*	*No season*	*Model*	
1	2.85	2.22	2.65	2.15	2.85	2.34	**1.97**	2.34
2	1.87	1.61	2.10	1.41	1.29	**1.11**	1.45	1.41
3	1.85	1.12	1.16	0.91	**0.87**	0.92	1.22	0.91
4	1.05	1.46	2.31	**0.69**	1.45	0.79	0.93	0.79
Total	0.93	0.86	1.34	**0.68**	1.01	0.78	0.97	0.97

See *legenda* of Table 8. The definition of criteria is available in Table 2.

5. CONCLUSIONS

A critical issue for the proper application of model based estimation in current surveys is

the possibility to obtain reliable estimates of model parameters and functions. One of the limits of many methodological proposals is that only the actual sample information is taken into account for estimation. Indeed, the availability of historical micro-data for the same target survey – or databases which may be helpful in investigating unit variability along time – suggests simple techniques aimed at obtaining estimates of the model unit variance functions, which must be always specified in each non homoschedastic model. These techniques are founded on the idea to approximate each model unit variance function with the empirical longitudinal variance of the same unit calculated on the basis of the historical longitudinal database.

In this context, different model unit variance estimation techniques have been proposed, depending on the number of observations in calculations, on the importance given to seasonal effects and on the opportunity to use an estimation based both on the empirical variance criterion and on a log-linear modelization. An empirical attempt, referred to the quarterly sample survey on wholesale trade carried out by ISTAT, confirmed the usefulness of this family of techniques as opposed to other common *a priori* hypotheses on model unit variance functions.

However, further research is needed to face two main issues, which may lead to additional developments and improvements, and to a more robust assessment of reliability of solutions proposed:

1. other comparative applications are needed. They should consider other longitudinal surveys and, as a consequence, other response rates. Moreover, efficiency of the techniques compared should be investigated unless the presence of the non response (or late response) problem, even though in these cases the true values of the population mean should be known in order to calculate precision of sampling estimates.
2. The various techniques should be tested using other y variables. Turnover represents a not easy task, since other variables – e.g. the number of persons employed – may be more steady. However, discontinuity along time may be larger for investments or changes of stocks, while additional difficulties may rise if the target variable is a binary variable (e.g. the number of job vacancies at the end of the reference period).

REFERENCES

D. A. BLOCK, M. R. SEGAL (1989). *Empirical Comparison of Approaches to Forming Strata Using Classification Trees to Adjust for Covariates*, Journal of the American Statistical Association, Vol.84, 408, pp.897-905.

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

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.

T. DALENIUS, J. L. HODGES (1959). *Minimum Variance Stratification*, Journal of the American Statistical Association, 54, pp.88-101.

K. DJERF (1997). *Effects of Post-stratification on the Estimates of the Finnish Labour Force Survey*, Journal of Official Statistics, 13, pp.29-39.

A. H. DORFMAN, R. VALLIANT (2000). *Stratification by Size Revised*, Journal of Official Statistics, Vol.16, 2, pp.139-154.

R. GISMONDI (2008). *Reducing Revisions in Short-term Business Surveys*, Statistica, LXVIII, 1, 85-115.

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

J. G. IBRAHIM, H. ZHU, N. TANG (2008). *Model Selection Criteria for Missing-Data Problems Using the EM Algorith,* Journal of the American Statistical Association, Vol.103, 484, pp.1648-1658.

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

J. R. KNAUB JR. (2004). *Modeling Superpopulation Variance: Its Relationship to Total Survey Error*, InterStat, http://interstat.statjournals.net/YEAR/2004/articles/0408001.pdf.

N. M. LAIRD, J H. WARE (1982). *Random-Effects Models for Longitudinal Data*, Biometrics, 38, pp.963-974.

R. LEHTONEN, C. E. SÄRNDAL, A. VEIJANEN (2003). *The Effect of Model Choice in Estimation for Domains, Including Small Domains*, Survey Methodology, 29, 1, pp.33-44.

R. J. A. LITTLE (1995). *Modelling the Drop Out Mechanism in Repeated-Measures Studies*, Journal of the American Statistical Association, 90, pp.1112-1121.

S. LUNDSTRÖM, C.E. SÄRNDAL (1999). *Calibration as a Standard Method for Treatment of Nonresponse*, Journal of Official Statistics, Vol.15, 2, pp.305-327.

M. PARK, W. FULLER (2008). *The Mixed Model for Survey Regression Estimation*, Journal of Statistical Planning and Inference.

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

C. E. SÄRNDAL, B. SWENSSON, J. WRETMAN (1999). *Model Assisted Survey Sampling*, Springer.

E. V. SLUD, L. BAILEY (2010). *Evaluation and Selection of Models for Attrition Nonresponse Adjustment*, Journal of Official Statistics, Vol.26, 1, pp.127-143.

N. WATSON, R. STARICK (2011). *Evaluation of Alternative Income Imputation Methods for a*

Longitudinal Survey, Journal of Official Statistics, Vol.27, 4, pp.693-715.

SUMMARY

Improving Efficiency of Model Based Estimation in Longitudinal Surveys Through the Use of Historical Data

In this context, supposing a sampling survey framework and a *model-based* approach, the attention has been focused on the main features of the optimal prediction strategy for a population mean, which implies knowledge of some model parameters and functions, normally unknown. In particular, a wrong specification of the model individual variances may lead to a serious loss of efficiency of estimates. For this reason, we have proposed some techniques for the estimation of model variances, which instead of being put equal to given *a priori* functions, can be estimated through historical data concerning past survey occasions. A time series of past observations is almost always available, especially in a longitudinal survey context. Usefulness of the technique proposed has been tested through an empirical attempt, concerning the quarterly wholesale trade survey carried out by ISTAT (Italian National Statistical Institute) in the period 2005-2010. In this framework, the problem consists in minimising magnitude of *revisions*, given by the differences between preliminary estimates (based on the sub-sample of *quick* respondents) and final estimates (which take into account *late* respondents as well). Main results show that model variances estimation through historical data lead to efficiency gains which cannot be neglected. This outcome was confirmed by a further exercise, based on 1000 random replications of late responses.