STATISTICA, anno LXVII, n. 1, 2007

SCORE FUNCTIONS AND STATISTICAL CRITERIA TO MANAGE *INTENSIVE FOLLOW UP* IN BUSINESS SURVEYS

Roberto Gismondi[1]

1. NON RESPONSE PREVENTION AND OFFICIAL STATISTICS

Among the main components on which the statistical definition of quality for business statistics is founded (ISTAT, 1989; EUROSTAT, 2000), accuracy and timeliness seem to be the most relevant both for producers and users of statistical data. That is particularly true for what concerns *short-term* statistics that by definition must be characterised by a very short delay between date of release and period of reference of data.

However, it is well known the *trade/off* between timeliness and accuracy. When only a part of the theoretical set of respondents is available for estimation, in addition to imputation or re-weighting, one should previously get responses by all those units that can be considered "critical" in order to produce good estimates. That is true both for census, *cut/off* or pure sampling surveys (Cassel *et al.*, 1983).

On the other hand, the number of recontacts that can be carried out is limited, not only because of time constraints, but also for the need to contain response burden on enterprises involved in statistical surveys (Kalton *et al.*, 1989). One must consider that, in the European Union framework, the evaluation of response burden is assuming a more and more relevant strategic role (EUROSTAT, 2005*b*); its limitation is an implicit request by the European Commission and influences operative choices, obliging national statistical institutes to consider with care how many and which non respondent units should be object of follow ups and reminders.

In Italy, business surveys managed in the frame of official statistics are mostly based on a system of reminders. For instance, in the frame of the 2 most important yearly business surveys carried out by ISTAT – referred to enterprises with until 99 and with more than 99 persons employed[2], and aimed at estimating the main structural indicators as turnover, investments, value added and employment

[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] They are identified, respectively, as PMI (Piccole e Medie Imprese) and SCI (Sistema dei Conti delle Imprese).

– normally 2 distinct reminders are used: by post and by telephone for the former survey, only by post for the latter. Moreover, in the main short-term business statistics (monthly industrial production and turnover, monthly retail trade sales, quarterly turnover indicators for the market service activities) a composite remainders' system is used, on the basis of the conjoint recourse to post, fax-server, telephone and e-mail and on a number of follow-ups - for the same reference period - ranging from 1 to 4.

The main reason that justifies the recourse to reminders is the not possibility to apply successfully corrections for non-response when coverage of respondents – in terms of a given auxiliary variable correlated with that of interest – is too low, or when available responses are the result of a *self-selection* process (Royall, 1992; Drudi and Filippucci, 2000). Late experiences on that have been presented and commented in ISTAT (2006).

Even though procedures based on re-weighting are widely known and recommended, their empirical effectiveness strongly depends on the availability of one or more auxiliary variables, in order to estimate the individual response probabilities (Cicchitelli *et al.*, 1992, 422-424), or to implement calibration estimators (Lundström and Särndal, 1999). A crucial point is that auxiliary variables must be measured on all the units in the population, or at least their total referred to the units not belonging to the sample must be known. For instance, for all the main short-term business survey aimed at estimating dynamics of monthly or quarterly turnover, the only auxiliary variables that satisfy this condition are the yearly turnover and the number of persons employed derived from the business register ASIA managed by ISTAT, both referred to the year before that under observation and not always strictly linked to infra-annual output indicators. On the other hand, recourse to imputation – even though recommended in the context of infra-annual surveys repeated along time (Hidiroglou and Berthelot, 1986) – generally leads to poor results when large units data are missing, or a non-response bias occurs (Bolfarine and Zacks, 1992, 128-133).

Finally, generally speaking large enterprises data can not be object of any estimation or re-weighting procedures, even in the case when respondents' coverage is high. It is the case of self-representative units – as those object of a census inside some *cut/off* strata – which data must be obtained before the release of estimates and could be quite poorly estimated using imputations based on "average" dynamics" or donor procedures (ISTAT, 2006).

We can indicate as *FU* a general current *Follow Up* action, and as *IFU* an *Intensive Follow Up* action addressed to a certain subset of non respondent units. The particular problem herein faced concerns the procedure to be used for the identification of this subset. Units belonging to this subset will be indicated as *IFUs*.

One must note that, in a sampling design context, the most natural and recommended way to perform a *FU* process simply consists in choosing units to be recontacted at random among non respondents (Cochran, 1977, 365-367; Droesbeke *et al.*, 1987, 181-182). In particular, actions to be carried out are:

a. determining the lowest number of units to be recontacted. That can be done according to the sampling variance formula and in order to guarantee a given precision level, or on the basis of operative constraints and deadlines for publication of provisional data.
b. Choosing these units at random among non respondents, according to their preliminary inclusion probabilities. In this case, the two fundamental problems concerned with *IFU* – choice of the number of units to be recontacted and identification of these units – are faced in two separate steps, while in an *IFU* context they are generally solved simultaneously, according to the preliminary definition of an individual *score function*, as described in paragraphs 2 and 3.

However, there are several practical situations when recourse to an *IFU* strategy can be an alternative to the previous random selection procedure and, sometimes, even necessary:

- the available sample does not derive from a predefined sampling design, but is a natural sample available, for instance, from administrative sources. In this case, the concept of sampling variance looses its meaning if the randomness taken into account refers only to the *sampling design*.
- The available sample has been selected in a deterministic way, and/or according to a superpopulation model, so that the final *MSE* will be evaluated according to the model and not to any sampling design.
- The survey is a census, or a *cut/off* sampling survey, so that the identification of an *IFU* strategy should be related to the reduction of the possible final undercoverage.
- Finally, even when the sample derives from a specific sampling design, one would still have the possibility to apply reminders for a specific subset of non respondents. In this case, the final inclusion probabilities will *change* and the final *MSE* estimation will be based on a mixture of old and new inclusion probabilities, concerned respectively with units respondent without and as a consequence of an *IFU* action.

Even though this topic is assuming an increasing relevance, especially in the frame of official business statistics carried out by national statistical institutes, until now only a few attentions have been spent on that. Moreover, the main available theoretical proposals often refer to data editing problems rather than to intensive follow up management[3]. With these premises, in this context the main purposes are:

1) to resume into a general methodology theoretical criteria and best current practices, focusing on the definition of a generalised "score function" to be calculated for each non respondent unit.

[3] See, for instance, Granquist and Kovar (1997); Lawrence and McKenzie (2000); De Jong (2003); Hedlin (2003); Philips (2003).

2) To propose a new “score function”, valid both for estimation of level and change.
3) To evaluate and compare some criteria for identifying, according to their score function, critical units that should be considered as *IFUs*.
4) To compare the various criterions in the frame of an empirical attempt carried out using real business data.

Points 1) and 2) have been faced in paragraphs 2 and 3, point 3) in paragraph 4 and point 4) has been developed in paragraph 5; summary conclusions have been drawn in paragraph 6.

2. ESTIMATION OF LEVEL

Let’s suppose that a theoretical sample *s* with size *n* is drawn from a population composed by *N* units, with the aim to estimate a given population parameter. When estimation must be carried out, only an effective sample s_R including n_R respondent units is available. If $s_{\bar{R}}$ is the sub-sample including the $n_{\bar{R}} = (n - n_R)$ non respondents, the main purpose is the identification of a sub-sample $s_{\bar{R}*}$ including the $n_{\bar{R}*}$ *IFU* units (*IFUs*) to be recontacted, with $n_{\bar{R}*} \leq (n - n_R)$.

These units are those fundamental in order to guarantee enough good estimates[4] of the unknown population mean - or the change of this mean between the period of reference *t* and a previous time (*t*-1) - and should be object of *IFU* in case of non-response or *late* response.

For instance, it could be the case of a sampling survey where the theoretical sample has been selected according to a *purposive* scheme and not to a probabilistic design: if a regression super-population model though the origin based on an auxiliary *X* variable has been adopted, it is well known (Cicchitelli *et al.* 1992, 385-387) that the units having the largest *X*-values should have the largest influence on the model *MSE* as well. Further, even when a probabilistic design is used, one could decide to speed up responses of those units leading to the largest gain in precision of *provisional* estimates, even though – as already remarked – that produces a modification of the individual response probabilities and a more complicated estimation of the final *MSE*. In particular, under a *design-based* approach, if *ad hoc* instead of random reminders are used, the sampling design could become complex and the estimate of inclusion probabilities needed in order to implement the Horvitz-Thompson (HT) could be quite difficult. Fattorini (2006) proposed an estimation technique based on a Monte Carlo simulation and *M* independent replications of the sampling scheme: he also evaluated properties of the resulting estimator, showing its convergence towards the ordinary HT estimator for $M \rightarrow \infty$ and proposing the option $M=10^6$ as empirical rule.

With these premises, the problems to be faced concern: a) the definition of a score function based on observed data that expresses the *statistical risk* concerned

[4] For what concerns the concept of “goodness of estimates”, see paragraph 4.

with the not availability of data referred to a certain unit; b) the choice of a statistical criterion able to detect which of these individual scores are particularly high and, as a consequence, can lead to the identification of the *IFUs*.

It is worthwhile to remark that all the following considerations can not be applied to outstanding units who are new to the survey or had been a non respondent for all the previous survey occasions. A cautional option consists in assigning to them the highest *IFU* priority.

2.1 *The univariate case*

In this case, only one y variable is observed (or is considered as relevant) and object of estimation is the population mean $\overline{y}$. If y_i is the y-value reported by the *i-th* unit, and $\hat{y}_i$ is its estimate got using a whatever imputation technique, the first step consists in defining the following transformation:

$$r_i = \left| z_{1i} y_i - z_{2i} \right|, \tag{1}$$

where the new variables z_1 and z_2 must be determined. In particular, one can put:

a) $z_{1i} = 1$ and $z_{2i} = 0$ (2)

so that (1) reduces to the simple absolute y-value related to the unit;

b) $z_{1i} = 1$ and $z_{2i} = \hat{y}_i$ (3)

so that (1) becomes the absolute difference between the true and estimated y-value;

c) $z_{1i} = \hat{y}_i^{-1}$ and $z_{2i} = 0$ (4)

so that (1) becomes the absolute ratio between the true and estimated y-value;

d) $z_{1i} = y_i^{-1}$ and $z_{2i} = \hat{y}_i y_i^{-1}$ (5)

so that (1) becomes the absolute relative difference between the true and estimated y-value.

According to case b), the function r_i equals that proposed by Mckenzie (2003, 476). Moreover, according to case c), the function r_i becomes similar to that proposed by Latouche and Berthelot (1992, 392), even though in that case z_{2i} was given by the y-value referred to a previous time (t-1). Transformation d) was proposed by Gismondi (2006) and has the advantage, respect to b), to deal with functions independent from measure unit and individual magnitude, so that they can be summed up over different units.

The next step consists in multiplying r_i by the individual sampling weight w_i

and a factor measuring the importance of the unit according to its size, so that a first score function will be given by:

$$\Phi_{1i} = r_i \cdot w_i \cdot (MAX(y_i, z_{1i}, z_{2i}))^U , \tag{6}$$

where *MAX* indicates the highest value. The recourse to the *MAX* function is coherent with options b) and c), while in case a) one could put *U*=0. As also remarked by Hidiroglou and Berthelot (1986), the exponent U $(0 \le U \le 1)$ provides a control on the importance associated with the magnitude of the data. This parameter is not very sensitive and the same value can be used for many variables of the survey.

According to case a), the function (6) is equivalent to that used by Pursey (2003), Chen and Xie (2004) and Succi and Cirianni (2005) when *U*=0, supposing that the *y*-value used to implement (1) can be obtained by a business register, hypothesis which is realistic if *y* is turnover or a general business revenue. In particular, under a simple random sampling design also the weights *w* are constant, so that Φ_{1i} reduces to the original value y_i.

The main difference between a) and b) or c) is that, in these last two cases, a large unit is not necessarily characterised by a high score function.

At the third step, the first score function (6) is transformed into a second score function defined as follows – where $q_{(\Phi_1)0,25}$, $q_{(\Phi_1)0,50}$ and $q_{(\Phi_1)0,75}$ are, respectively, the first quartile, the median and the third quartile of the score function (6):

$$\Phi_{2i} = \left(\frac{\Phi_{1i} - q_{(\Phi_1)0,50}}{q_{(\Phi_1)0,75} - q_{(\Phi_1)0,25}} \right). \tag{7}$$

With this transformation, the final score function will have a distribution more uniform and symmetric than (6), which form is strongly influenced by the original *y* distribution (Gismondi, 2000). This aspect will be considered again in paragraph 4.

A further choice for z_1 and z_2 can be obtained supposing to evaluate the effect of the not availability of a certain unit on the final level estimate. If the *i-th* unit is not respondent, one can decide to estimate its *y*-value ant to carry on the sampling estimation of average level including this estimate in calculations. Then, a score function for the *i-th* unit can be given by the absolute difference between the estimate got using the true y_i value (first round brackets in (8)) and the estimated one (second brackets); if the estimator of the population mean is given by $N^{-1}\sum_{i=1}^{n} y_i w_i$, the score will be given by:

$$N^{-1} \left| \left(\sum_{j \ne i \quad j=1}^{n-1} y_j w_j - y_j w_j \right) - \left(\sum_{j \ne i \quad j=1}^{n-1} y_j w_j - \hat{y}_j \right) \right| = N^{-1} w_j \left| y_j - \hat{y}_j \right| , \tag{8}$$

and the last term is similar to that derived from (1) in case b), where the sampling weight w is already included in the score function before transformation (6).

The situation changes if we suppose that the *i-th* not respondent unit's value is not estimated and is excluded from calculation. That happens, for instance, when in the survey context no imputation procedure has been planned, or it is not planned for particularly large and relevant units[5], whose values must be obtained directly.

If the sampling design is based on inclusion probabilities π_i, the sampling weight used for the estimation of the unknown mean is given by $w_i^{(n)} = (\pi_i^{(n)})^{-1}$, where the upper ($n$) means that the estimation is based on n units. If the sampling design is based on (n-1) units, we can suppose[6] that for whatever n: $\pi_i^{(n)} = \alpha\pi_i^{(n-1)}$. Since we must have $\sum_{i=1}^{N}\pi_i^{(n)} = n$, it follows immediately that $\alpha = n(n-1)^{-1}$, so that this relation will hold:

$$\pi_i^{(n)} = n(n-1)^{-1}\pi_i^{(n-1)} \rightarrow w_i^{(n)} = (n-1)n^{-1}w_i^{(n-1)}. \tag{9}$$

The absolute difference between the estimates based on n and (n-1) units – meaning as (-i) the estimate based on all units except the *i-th* – will be given by:

$$\begin{aligned}
\left|\hat{\bar{y}}^{(n)} - \hat{\bar{y}}_{(-i)}^{(n-1)}\right| &= N^{-1}\left|\sum_{j=1}^{n} y_j w_j^{(n)} - \sum_{j\neq i \quad j=1}^{n-1} y_j w_j^{(n-1)}\right| = \\
&= N^{-1}\left|\left(\sum_{j\neq i \quad j=1}^{n-1} y_j w_j^{(n)} + y_i w_i^{(n)}\right) - \sum_{j\neq i \quad j=1}^{n-1} y_j w_j^{(n-1)}\right| = \\
&= N^{-1}\left|(n-1)n^{-1}\left(\sum_{j\neq i \quad j=1}^{n-1} y_j w_j^{(n-1)} + y_i w_i^{(n-1)}\right) - \sum_{j\neq i \quad j=1}^{n-1} y_j w_j^{(n-1)}\right| = \\
&= n^{-1}\left|(n-1)N^{-1} y_i w_i^{(n-1)} - \hat{\bar{y}}_{(-i)}^{(n-1)}\right| = n^{-1}\left|\hat{\bar{y}}_{(i)}^{(1)} - \hat{\bar{y}}_{(-i)}^{(n-1)}\right|.
\end{aligned} \tag{10}$$

The previous quantity is proportional to the absolute difference between the population mean estimates based, respectively, only on the *i-th* unit and all the remaining (n-1) units different from the *i-th*. Under a simple random sampling design, the previous estimates reduce to sample means based, respectively, on 1 and (n-1) observations. The previous function can be related to (1) putting:

$$\text{e) } z_{1i} = N^{-1}w_i^{(n)} \text{ and } z_{2i} = n^{-1}\hat{\bar{y}}_{(-i)}^{(n-1)}. \tag{11}$$

[5] Not large units can be relevant as well if they belong to very small and heterogeneous strata.

[6] Simple random sampling and *PPS* designs satisfy this rule.

In order to calculate score functions, one must take into account some additional aspects:

- of course, score functions can be evaluated with reference to the actual time *t* only if they are calculated using an auxiliary variable; in a longitudinal survey context, the same *y*-variable referred to one ore more previous occasions (*t*-1) is often used. If a unit is included in the survey for the first time, it could be *a priori* excluded from (or included in) follow up actions, or its score can be estimated according to an auxiliary variable quite correlated with *y*.
- If the survey is a census, sampling weights *w* in (6) disappear. In order to guarantee generality to (6), they can be put all equal to one.
- If stratification is used, score functions (6) should be defined and evaluated separately in each stratum. On the other hand, if all units are considered as a whole, and W_v is the relative weight of the *v-th* stratum, we can evaluate the new function $W_v\Phi_{1vi}$[7].
- A more general way to deal with comparable score functions - rather than transformations c) or d) - consists in dividing transformations (2) or (3) by the true overall mean $\overline{y}$, or its estimate $\hat{\overline{y}}^{(n)}$. Even though this adding factor does not have consequences on the choice of *IFUs*, it can be helpful whenever it is necessary to sum up score functions referred to different domains that could be expressed in *different measure units* (paragraph 2.2).

Finally, given the vectors of observations **y** and score functions **Φ**, the general rule useful to identify the *IFUs* consists in analysing the behaviour of a general transformation *f* such as:

$$f(\Phi_{2i}, y_i)\,. \tag{12}$$

The function *f* can be based only on Φ_i, only on y_i or, more generally, on both of them. Some alternative options are discussed in paragraph 4. Let's note that the identification of *IFUs* is not *necessarily* based on the definition of a threshold for *f* – e.g. some *f** – even though this possibility is explicitly considered in paragraph 4.2.

2.2 *The multivariate case*

The identification of *IFUs* could be based on more than one indicator derived from the survey. Indicators can be given by single variables (as turnover, costs, number of persons employed in the case of business surveys) or by particular functions applied to the same variable. If *k* indicators are taken into account, a simple way to proceed simply consists in considering as *IFUs* all those units that turn out to be *IFUs* for *at least one* indicator *h*, according to the general rule (12).

[7] In a stratified random sampling context, units to be re-contacted could be distributed among strata according to the Neyman allocation rule.

This is the *enlarged* criterion, since we should obtain a relatively large number of *IFUs*. On the other hand, a composite average score function can be defined as:

$$\Phi_{2i} = \sum_{h=1}^{k} \Phi_{2i}^{(h)} P^{(h)}, \tag{13}$$

where $P^{(h)}$ is a coefficient related to the *h-th* indicator. The purposes of these coefficients can be: a) to eliminate the effects due to different magnitudes (and/or different measure units) of indicators and guarantee their additivity; b) to assign a specific weight to each indicator. In order to get the first goal, a simple choice consists in putting:

$$P^{(h)} = \overline{\Phi}_2^{(h)}, \qquad \text{where} \qquad \overline{\Phi}_2^{(h)} = n^{-1} \sum_{i=1}^{n} \Phi_{2i}^{(h)}. \tag{14}$$

That is particularly useful when indicators rather than variables are taken into account, as it will be seen in paragraph 5.

If the second purpose is the most relevant, the coefficients *P* can be put equal to some weights *W*, that can be defined with a subjective choice, according to the relative weight of each indicator on the overall variance[8] or on the basis or more particular rules.

For instance, an alternative way to calculate an average score function is based on the following formula, provided that $y_i^{(h)}$ is the value assumed by the *h-th* indicator on the *i-th* unit:

$$\Phi_{2i}^{*} = \sum_{h=1}^{k(i)} \Phi_{2i}^{(h)} W_i^{(h)}, \quad \text{where} \quad W_i^{(h)} = \left(y_i^{(h)} \Big/ \sum_{j=1}^{n^{(h)}} y_j^{(h)} \right) \left[\sum_{h=1}^{k(i)} \left(y_i^{(h)} \Big/ \sum_{j=1}^{n^{(h)}} y_j^{(h)} \right) \right]^{-1}, \tag{15}$$

and $n^{(h)}$ is the number of units on which the *h-th* indicator can be measured, with $k(i) \leq k$. The main difference respect to (13) is that *different* weights are used for each single unit considered. Also in this case, sum of weights for each *i-th* unit is equal to one and weights should be estimated using data referred to a previous time (*t*-1).

Recourse to (15) could be useful when, in the survey, on each *i-th* unit only *k*(*i*) of the *k* indicators can be measured, e.g. $y_i^{(h)} \neq 0$ for $h \in H(i)$, where *H*(*i*) includes *k*(*i*) indicators. A typical example is given by the monthly survey on industrial production, currently carried out in each developed country. In this case, each observation unit (enterprise, local unit or local "Kind of Activity Unit") can produce one or more industrial products. These *k*(*i*) products are generally expressed in different measure units and could vary along time. These considerations justify the recourse to weights as those in the second equality in (15).

[8] In this case indicators should be previously standardized.

One relevant consequence derived from the recourse to (13) or (15) is that one unit could be an *IFU* without being an *IFU* for *any* of the single *k* indicators.

A further criterion can be mentioned: if $H(i)^*$ is the number of indicators for which the *i-th* unit is an *IFU* according to (12), one can calculate:

$$\left(\sum_{h \in H(i)^*} W^{(h)} \Big/ \sum_{h=1}^{k} W^{(h)} \right) \tag{16}$$

and then verify if (16) – according to a preliminary transformation as (7) – is higher than a certain threshold, on the basis of criteria similar to those described in paragraphs 4.2 and 4.3. In this way we consider the *i-th* unit as an *IFU* verifying the relative overall magnitude of variables for which this unit is critical according to the univariate case.

On the other hand, this criterion cannot be always used, because it depends on the possibility to sum up *different* indicators. Moreover, from a logical point of view it implies a double application of score functions: the first to the single *y*-values, the second to weights *W* assigned to indicators for each unit.

Finally, another criterion can be obtained generalising that proposed by Mckenzie (2003, 478). We suppose to have observed *k* variables measured on *n* units along *T* time periods before that under observation. For each variable *h* one can calculate scores $r_i^{(h)}$ (or $\Phi_i^{(h)}$). Then, on the basis of the *n*x(*T*-1) available individual scores (so, excluding data referred to the latest period), one can determine deciles of the empirical score distribution. The same procedure is carried out separately for each variable (and, of course, separately in each stratum derived from the original sampling design). For each unit, one calculates scores referred to the last time *T* and verifies, for each variable, which decile they belong to; finally, a priority *IFU* score correspondent to the maximum decile is assigned, where deciles have been supposed to be numbered from 0 to 9 (0 to 0-10th percentile, 1 to 11th-20th percentile and so on). For instance, if *k*=2 and a unit falls in the second decile for a variable and in the third decile for the other, an *IFU* score equal to 2 will be assigned. Even though this method is relatively simple to be implemented, a certain loss of information must be paid passing from original data to deciles.

3. ESTIMATION OF CHANGE

The main purpose of the most part of short-term business surveys is the estimation of the change $\overline{y}_t / \overline{y}_{(t-1)}$, where *t* is a month or a quarter and (*t*-1) is a generic previous period – for instance, the base year when index numbers are calculated. The individual change will given by $c_{ti} = c_i = y_{ti} / y_{(t-1)i}$.

In this case it is useful to apply a further transformation to the individual change, given by $c_{ti}^* = MAX(y_{ti} / y_{(t-1)i}, y_{(t-1)i} / y_{ti})$. This option derives from the need to assign high priority to units characterised both by a very high *or* a very low change, even though the next transformation (17) can be applied indifferently to *c* of *c**. Of course, final relevance of each unit will be determined according to its *c** value and its magnitude as well, according to function (6).

Even though the logical frame remains similar to that seen in paragraph 2, a relevant difference is that, in this case, the score function – given the individual *y*-magnitude – should increase whenever a unit is characterised by very high or very low rates of change along time. On the other hand, the only univariate case will be considered in details, since all the considerations concerning the multivariate case (paragraph 2.2) remain valid for estimation of change as well.

We also suppose that at times (*t*-1) and *t* the same units are included in the sample with the same inclusion probabilities.

The first step consists in defining a transformation similar to (1), but applied to c_i:

$$r_i = \left| z_{1i} c_i - z_{2i} \right| , \tag{17}$$

where the new variables z_1 and z_2 must be determined. In particular, the most useful options for estimation of change are:

$$\text{a}') \; z_{1i} = 1 \;\text{ and }\; z_{2i} = 0 \tag{18}$$

so that (17) reduces to the simple *c*-value related to the unit;

$$\text{b}') \; z_{1i} = 1 \;\text{ and }\; z_{2i} = \hat{c}_i \tag{19}$$

so that (17) becomes the difference between the true and estimated *c*-values.

All the further steps (6) and (7) seen in paragraph 2 can be applied, with obvious modifications, to the transformation (17), so that, also in this case, a final score function Φ_{2i} can be calculated.

A further choice for z_1 and z_2 can be obtained supposing to evaluate the effect of the not availability of a certain unit on final change estimate. If the *i-th* unit is not respondent and is excluded from calculation, the hypothesis (9) on inclusion probabilities is still valid and symbols introduced in paragraph 2 keep their meaning, one can evaluate the absolute difference between the estimates of change between times *t* and (*t*-1) based on *n* and (*n*-1) units, given by:

$$\begin{aligned} \left| \left(\frac{\hat{y}_t^{(n)}}{\hat{y}_{(t-1)}^{(n)}} \right) - \left(\frac{\hat{y}_{t(-i)}^{(n-1)}}{\hat{y}_{(t-1)(-i)}^{(n-1)}} \right) \right| &= \left| \left(\frac{\sum_{j=1}^{n} y_{tj} w_j^{(n)}}{\sum_{j=1}^{n} y_{(t-1)j} w_j^{(n)}} \right) - \left(\frac{\sum_{j \neq i \;\; j=1}^{n-1} y_{tj} w_j^{(n-1)}}{\sum_{j \neq i \;\; j=1}^{n-1} y_{(t-1)j} w_j^{(n-1)}} \right) \right| = \ldots = \\ &= n^{-1} \left(\frac{\hat{y}_{(t-1)(i)}^{(1)}}{\hat{y}_{(t-1)}^{(n)}} \right) \left| I_{t(i)} - I_{t(-i)} \right| , \end{aligned} \tag{20}$$

where I_t is an index of change between times t and (t-1) and, in particular:

$$I_{t(i)} = \frac{y_{ti}}{y_{(t-1)i}} \qquad I_{t(-i)} = \frac{\sum_{j\neq i \quad j=1}^{n-1} y_{tj} w_j^{(n-1)}}{\sum_{j\neq i \quad j=1}^{n-1} y_{(t-1)j} w_j^{(n-1)}}$$

$$\hat{y}^{(1)}_{(t-1)(i)} = N^{-1} n\, y_{(t-1)i} w_i^{(n-1)} \qquad \hat{y}^{(n)}_{(t-1)} = N^{-1} \sum_{j=1}^{n} y_{(t-1)j} w_j^{(n)} .$$

In this case, the score function is based on the absolute difference between the indexes of change calculated, respectively, on the only *i-th* unit and on the (n-1) units excluded the *i-th*, and plays the same role as (10), obtained for estimation of levels. However, an additional factor - respect to the only individual y-magnitude – that influences the score function is given by the ratio between the level estimates referred to time (t-1) calculated, respectively, only on the *i-th* unit and on all the n units. If (t-1) is the base year of index numbers, this ratio expresses the relative weight of the *i-th* unit on the overall level estimate referred to the base year.

For estimation of change, the individual score function depends both on: 1) the contribution given by the unit to the overall level estimate at time (t-1) and 2) the difference between the individual trend and the overall average trend evaluated on the remaining (n-1) units. Relation with (1) can be easily obtained - given that $\hat{y}^{(1)}_{(t-1)(i)} \Big/ \hat{y}^{(n)}_{(t-1)} = g(\mathbf{y}_{(t-1)})$ - putting:

$$\text{c´)} \quad z_{1ti} = n^{-1} g(\mathbf{y}_{(t-1)}) y^{-1}_{i(t-1)} \quad \text{and} \quad z_{2ti} = n^{-1} g(\mathbf{y}_{(t-1)}) I_{t(-i)} . \tag{21}$$

4. IDENTIFICATION OF UNITS TO BE RECONTACTED

The number of recontacts can be determined in different ways. In all cases, one can first choose a score function among those described in paragraphs 2 and 3; then, scores must be ordered in a not decreasing way; finally, *IFUs* will be given by those units occupying the first positions in the ranking. The problem is the definition of a rule to decide how many *first positions* must be considered.

The number of *IFUs* – and the same selective choice of each unit to be recontacted – can be determined:

1) according to operative constraints, such as the maximum number of units that can be effectively followed with particular care, given technical and human resources devoted to the survey.
2) Evaluating the relation existing between reduction of pseudo-bias (paragraph 4.1) and number of follow-ups.
3) On the basis of some other statistical test different from 2) carried out on the individual score functions.

In case 1), having fixed *a priori* the number of recontacts that can be managed given the operative constraints and deadlines for publication, the choice of units can be done according to rules as those described in paragraph 4.2. However, in current practice - especially under non probabilistic sampling designs, or in case of *cut-off* samplings - it is quite common to recontact all and only the units that, added to those already available, guarantee a given coverage level referred to one or more main variables observed in the survey (Pietsch, 1995; Sprent, 1998).

In both cases 2) and 3) the use of a score function is joined to the search of a threshold for choosing units to be re-contacted. Two families of criteria have been resumed in paragraphs 4.1 and 4.2. In the follow, we can suppose that:

- all the available *n* observations are independent each other;
- in each stratum, one can deal with an enough large number of units, so that estimators' distributions can be approximated by a normal density;
- as already remarked, score functions have been preliminarily ordered in a decreasing ranking.

4.1 *Evaluation of the bias ratio and the pseudo bias*

In the frame of case 2) mentioned above, one can consider a "test data set", which could derive from some previous periods of the survey, or could be an early batch of data in the current survey period. This dataset must contain all the units, including those that at the current time are not respondents.

The main idea is to test significance of the difference between the *y*-estimate based on a complete data set of respondents and the data set not including a certain unit (Deming, 1953). A fundamental aspect to be determined is the form assumed by function *f* defined in (12): while score functions as (7) are used in order to create a ranking of units according to their not increasing score level, *y*-values are those *effectively taken into account* to test significance. If $\hat{y}$ is the benchmark reference for assessing precision of the estimate $\hat{y}_{(-i)}$ not including the *i-th* unit, one can evaluate the *bias ratio* of the estimate. Since the global error of this estimate is the sum of squared bias and sampling variance, the bias ratio is defined as the relative incidence of the former error component on the latter - provided that variance under square root depends on the estimator and the sampling design used:

$$BR(\hat{y}_{(-i)}) = \frac{\left|\hat{y} - \hat{y}_{(-i)}\right|}{\sqrt{Var(\hat{y}_{(-i)})}} . \tag{22}$$

On the basis of (22), the selective choice of units to be recontacted can be driven by the evaluation of how much bias one should accept. At each step, starting from the not respondent unit having the highest score, one by one all the non respondent units are supposed to be excluded from calculations and used in order to evaluate $\hat{\bar{y}}_{(-i)}$. If sample estimates approximately follow a normal distribution, the bias ratio is approximately $N(0,1)$.

We can also define the *coverage probability*, that is the probability that the unknown mean is contained in a confidence interval derived from the standardised normal distribution Z. This probability is given by: $\Pr[-z_{1-\alpha/2} - BR(\hat{\bar{y}}_{(-i)}) < Z < z_{1-\alpha/2} - BR(\hat{\bar{y}}_{(-i)})]$ - where $z_{(1-a/2)}$ is the percentile of the standardised normal cumulated distribution leaving on the right a probability equal to $a/2$ - from which it follows that the coverage probability equals the nominal, desired confidence level (1-*a*), only if the bias ratio is zero.

However, according to Cicchitelli *et al.* (1992, 65-66) and Särndal *et al.* (1993, 163-165), we can consider that a bias ratio lower than 10% gives a loss of coverage probability less than 1%, which therefore is entirely negligible compared with other shortcomings of common variance estimation. An operational rule concerned with (22) consists in ordering units according to their not increasing score, identifying as *IFUs* all the units for which, progressively, the bias ratio keeps higher than 10% and stopping as soon as the first unit such that the bias ratio falls under 10% is found.

It is worthwhile to underline how the use of (22) can be strictly connected with a statistical test useful for evaluating the distance between one unit and a group of units. If one consider a generic X variable measured on n units, each X_i-value is compared with the mean $\bar{X}_{(-i)}$ alculated on the remaining units excluded the *i-th*. At the first step, when n units are considered, the test is based on $T_{(n-2)} = (X_i - \bar{X}_{(-i)}) \Big/ \sqrt{S^2_{X(-i)} n/(n-1)}$, where $S^2_{X(-i)}$ is the X-variance calculated on the whole sample excluded the *i-th* unit and $T_{(n\text{-}2)}$ is the Student's *t* with (*n*-2) degrees of freedom. In its original version, the procedure – based on a unilateral test since $X_i > \bar{X}_{(-i)}$ – stops if the unit with the highest score is not detected as critical, otherwise it is carried out again after recalculation both of sample mean and variance.

Given that, it is easy to verify that from (22) – under a simple random sampling without replacement design and putting $Var(\hat{\bar{y}}_{(-i)}) = \hat{\sigma}^2_{(-i)}/(n-1)$, where $\hat{\sigma}^2_{(-i)}$ is an estimate of σ^2 got using all units except the *i-th* – one obtains: $BR(\hat{\bar{y}}_{(-i)}) = T_{(n-2)}/\sqrt{n}$, so that, unless a constant term, the two tests are similar.

Since the 10% threshold could be too conservative, other choices are possible, using for instance the empirical (*pseudo*) bias:

$$EB(\hat{\bar{y}}_{(-i)}) = \frac{\left|\hat{\bar{y}} - \hat{\bar{y}}_{(-i)}\right|}{\hat{\bar{y}}}, \qquad (23)$$

that can be calculated on the basis of late data referred to some previous survey occasions. A similar choice was proposed by Latouche and Berthelot (1992), with the aim to find the lowest number of recontacts for which empirical bias registers a strong decrease. However, different thresholds for evaluating (23) can be used, so that critical values of the empirical bias could be also evaluated according to methods described in paragraphs 4.2 and 4.3.

It is worthwhile to note that test functions based on (22) or (23) could be used also when the number of units that can be recontacted is *given* because of operative constraints (as in the previous case 1)), in order to evaluate how much large could be the bias gap due to the not possibility to recontact all the necessary units.

4.2 *Parametric tests based on thresholds*

When it is not possible to use complete datasets in order to evaluate (22) or (23), or just in order to carry out additional comparative tests, one can consider a series of statistical procedures based on the simple idea to verify if a given unit belongs or not to the same population of the others. Commonly, similar tests are used for identifying outlier observations in sampling surveys frames.

Herein X is a general variable, that could be given by a score function as (7) or the relative gain in pseudo bias reduction (23). In both cases, we suppose that units have been preliminarily ordered according to their not decreasing X-values.

When the form of the X distribution is unknown, a very general and simple tool is given by the Chebyshev inequality. If μ_X and σ_X are mean and standard deviation of X in the population – that can be estimated according to previous surveys or current available observed data – and $Z=(X-\mu_X)/\sigma_X$, one can consider a specific X-value as critical if

$$1-(1/Z^2) > Pr, \tag{24}$$

where Pr is a given probability level. Since the test is bidirectional, when suspected X-values are higher than μ_X the choice Pr=0,10 means that critical values will be all those placed in the higher 5% of the empirical distribution. The main limits of the criterion are: 1) it is much less powerful than others based on the knowledge of X distribution; 2) it does not supply an *exact* probability that the test function is critical.

A second criterion is based on the standardised normal distribution and on the hypothesis that Z-values are approximately normally distributed. In this case, if both μ_X and σ_X are estimated using the whole available sample (*including* potential critical units), one can consider a specific X-value as critical if:

$$Z > z_{(1-a)}\,, \tag{25}$$

where $z_{(1-a)}$ is the percentile of the standardised normal cumulated distribution leaving on the right a probability equal to a, using an unidirectional test. When n<100 a better approximation can be achieved using the Student's t distribution.

Let's note that test (25) – and test (24) as well when mean and variance are estimated using current sample data – should be carried out one unit at a time: when the first unit is detected as an *IFU*, mean and variance should be recalculated, until no more units are critical. The procedure stops immediately if the unit with the highest score is not detected as critical.

Even though test (25) is more precise and more powerful than test (24), its use depends on the assumption of normality for *X*; moreover, Shiffler (1988) remarked that it can lead to wrong conclusions because the maximum limit for *Z* is $(n-1)/\sqrt{n}$, so that it will be easier to identify one unit as critical with a lower number of sample units, just because the highest value that *Z* could reach will be lower.

A further test connected with Student's *t* is the *Extreme Studentized Deviate* test. It was originally proposed by Grubbs (1969) and in this context will be based on:

$$(X_{Max} - \bar{X})/S_X\,, \tag{26}$$

where X_{Max} is the highest *X*-value among the *n* available. The unit characterised by X_{Max} is an *IFU* unit if (26) is higher than a critical value that can be derived from tables originally elaborated by Quesenberry and David (1961). The procedure goes on one unit at a time, excluding at each step from calculations units already identified as *IFUs*.

4.3 *Not parametric tests based on thresholds*

If the empirical scores distribution is quite far from normality, one could use not parametric tests. Among the wide set of available methods, we propose two criterions that can be easily adapted to the problem.

A first non parametric test can be based on the *MAD* function (*Mean Absolute Deviation*). When *n X*-values are available, we can define as MAD_X the median of the *n* absolute differences $\left|X - q_{(X)0,50}\right|$, where $q_{(X)0,50}$ is the *X* median; it is an estimator of the population standard deviation less efficient but, generally, more robust than the sample standard deviation S_X. A method defined by Sprent (1998) as simple and reasonably robust is based on the following rule:

$$(X - q_{(X)0,50})/MAD_X > Max\,, \tag{27}$$

where *Max* is a critical threshold to be determined. Running test (27) once at a time, functions $q_{(X)0,50}$ and MAD_X must be recalculated at each step, excluding units already identified as *IFUs*. For the choice of *Max*, one can consider that, in an outlier detection frame, Sprent and Smeeton (2001) suggested to put *Max*=5, since we can consider the empirical relation 5*MAD*=3*S* and that if available data - excluded the unit under observation - follow *approximately* a normal distribution, then anomalous values should be more distant than 3*S* from their mean. In an *IFU* context, a lower choice for *Max* could be acceptable, even though this subjectivity is probably the most relevant limit of the method.

A further non parametric test is based on the outlier detection procedure proposed by Hidiroglou and Berthelot (1986) and revised by Davila (1992). A unit will be an *IFU* if:

$$X > q_{(X)0,50} + \alpha(q_{(X)0,75} - q_{(X)0,50}) \,, \tag{28}$$

where $q_{(X)}$ are X-quantiles already defined in paragraph 2.1 and a is a subjective coefficient. As for (27), also in this case the method could be very sensitive respect to the choice of a, which level could be quite different from that currently used for outlier detection. Normally, it ranges from 2 to 5.

5. A COMPARISON STUDY

Some of the proposed methodologies have been applied to a real case, given by the Italian retail trade monthly survey, carried out by ISTAT.

The survey is aimed at estimating monthly retail trade turnover indexes (Division 52 of the NACE nomenclature); actually, it is the only Italian monthly survey concerning the service sector carried out in the frame of official statistics and represents a fundamental short-term indicator in the whole European Union.

In 2004 the sample was based on 7.500 enterprises - drawn from a population of about 600 thousands - object of a partial yearly rotation concerning about 2.000 enterprises.

The survey design is a stratified random sampling based on 150 strata; in each stratum, the average turnover is estimated using the ordinary sample mean.

Outlier values and real missing values due to non-response are estimated on the basis of a complex procedure involving different methods[9], depending on the availability of data at the single enterprise level, for which we address to ISTAT (1998).

No postal reminders are used, but sensitive firms (about 400, including very big and some other enterprises belonging to small strata) are contacted by telephone after about 25 days from the reference month in order to speed up data collection and guarantee the possibility to calculate and diffuse a provisional retail trade index after 30 days from the reference month. The final definitive release is planned after 52 days, according to requests of the EU Short-term Statistics Regulation (EUROSTAT, 2005*a*).

Delay mostly depends on response burden (on the average, enterprises are requested to remain into the sample for at least 36 months) and some inefficiencies affecting ordinary mail. Because of attrition and wave non-response, the effective monthly sample size is about 4.500 units. Moreover, the need to use for simulations only units for which values y_m, $y_{(m\text{-}12)}$ and $y_{(m\text{-}24)}$ were available – where m is a

[9] Individual estimates of missing and outlier values are not available, so that options b), c) and d) (formulas (3), (4) and (5)) have not been taken into account.

month – reduced size of the monthly sample effectively taken into account to about 2.600 units[10].

The main final indexes are referred to 10 domains, obtained crossing each other 2 groups of product sold ("food" and "non food") and 5 classes of persons employed (1-2, 3-5, 6-9, 10-19, >19). Higher level indexes are obtained on the basis of an arithmetic weighted mean, where the weight of each stratum is given by the yearly turnover referred to the base year 2000, derived from structural business statistics.

The empirical attempts aimed at identifying *IFUs* have been carried out applying these operational simplifications:

1) we have supposed that the main object of estimation is the average turnover and not an index number; in this way, all criteria introduced in paragraphs 2, 3 and 4 can be applied without any further adaptation;
2) separate simulations have been carried out into the 10 main domains on a monthly basis: inside each domain we have also supposed a simple random sampling design (while, in the survey context, a further stratification is applied in each domain);
3) quality indicators have been calculated – unless otherwise indicated – supposing to use for estimation *IFUs* only, even though in current practice estimates are based on non *IFUs* as well.

Under the previous assumptions, in (6) we put U=0,5 so that, adopting option (2), for each unit the function Φ_1 – according to definitions (1) and (6) – is proportional to $y^{1,5}$. This option is quite recommended (Hidiroglou and Berthelot, 1986; Davila, 1992) and widely used in practice (Grandquist, 1990; Hunt *et al.*, 1999), because it is intermediate between the extremes 0 and 1 and choices nearer to 1 could provide a too large influence of individual size on the final score function[11]. The compared criteria for identifying *IFUs* are reported in the following resuming scheme:

Formula	*Definition*
(22)	Bias ratio (coverage probability)
(23)	Empirical (pseudo) bias – 1% and 5% options
(24)	Chebyshev inequality
(25)	Standardized normal distribution – 1% and 5% options
(26)	Grubbs test (extreme studentized deviate)
(27)	Sprent test (Sprent and Smeeton, *MAD* test)
(28)	Hidirogluou-Berthelot – a=3,5 and a=5 options

All criteria have been implemented on the basis of function Φ_2 defined by formula (7), applying options (a) given by (2) and (e) given by (11) and separately for the estimation of level or change. One can note that methods based on bias

[10] As a consequence, figures reported in the following tables can not be compared with those effectively released by ISTAT on a monthly basis.

[11] Results got putting, for instance, U=0, are quite similar (some light differences occurred for methods as empirical bias (23) and the Chebyshev criterion (24)) and have been omitted.

ratio and empirical bias might not identify any *IFU*, because of the intrinsic meaning of the corresponding functions (22) and (23).

Main results reported in the next tables have been obtained as means of the 12 months of year 2004, while results for the domains "food" and "non food" derive from weighted means of the correspondent five employment classes.

Finally, in all the applications the variables used to calculate score functions were $y_{(m\text{-}12)}$ for level (data referred to 2003) and the ratio $y_{(m\text{-}12)}/y_{(m\text{-}24)}$ for change (2003 and 2002), while estimate errors have been evaluated on y_m for level (2004) and on the ratio $y_{m)}/y_{(m\text{-}12)}$ for change.

A first overall result deriving from table 1 is that, generally speaking, the largest differences among the final number of units identified as *IFUs* do not derive from the use of different score functions or the focus on level or changes, but from the particular identification criterion chosen. For instance, for estimation of level the number of *IFUs* obtained using bias ratio (22) is equal, respectively, to 106 when using formula (2) and to 102 using formula (11). Moreover, similar results are also obtained for estimation of change: still 106 units when using formula (2) and 103 units with formula (11). The same considerations hold for all the other identification criteria, with a partial exception for empirical bias at the 1% level; in this case, both for level and change the option (2) leads to a quite lower number of *IFUs* than the option (11): 113 against 95 for level and 134 against 122 for change.

In particular, from table 2 – that reports, for the "Total retail trade", the percent difference between the number of *IFUs* detected using formula (2) or formula (11) derived from table 1 – one deduces that the use of different score functions as (2) or (11) does not particularly influence the final number of *IFUs* for bias ratio (22), the Chebyshev method (24), both the Grubbs tests (26), the Sprent test (27) and the recourse to a standardized normal (25) but only at the 5% level, while in the remaining cases differences could be higher than 10%, with the highest differences due to the empirical bias method (23) for level estimation. Moreover, for the most part of methods lower percent differences occur for change rather than for level.

A clearer evidence derives from the comparison among identification criteria: both for level and change, according to the number of units identified as *IFUs* they can be divided into 4 groups:

1) Sprent test (27) and Hidiroglou-Berthelot (28) with a=3,5, with more than 8% of units identified as *IFUs* (this percentage can be defined as *IFUs ratio*);
2) bias ratio (22), empirical bias (23) at 1% and Hidiroglou-Berthelot (28) with a=5, with a *IFUs* ratio ranging from 4% to 6%;
3) standardized normal curve at 5%, with a *IFUs* ratio around 3%;
4) all the other 5 criteria, with a *IFUs* ratio lower than 2%.

As it could have been guessed, lower *IFUs* ratios lead, on the average, to higher percent estimate errors, even though the relation between them is not linear. For instance, from table 1 one can note that while the use for estimations of the only 16 *IFUs* identified with empirical bias at 5% leads to an estimate error equal to

TABLE 1

Main results using function Φ_2 with various criteria for estimating level and change

	(22)	(23)	(23)	(24)	(25)	(25)	(26)	(26)	(27)	(28)	(28)
Domain	Bias ratio	Empirical bias 5%	Empirical bias 1%	Chebyshev	Stand. normal 5%	Stand. normal 1%	Grubbs test 5%	Grubbs test 1%	Sprent test	Hidiroglou α=3,5	Hidiroglou α=5
Function Φ_2 and option a), formula (2) – Estimation of level											
Number of IFUs											
Food	67	12	71	18	33	20	13	12	78	73	53
Non food	39	4	42	20	40	23	15	14	142	138	92
Total	106	16	113	38	73	43	28	26	220	211	145
% ratio IFUs/sample											
Food	8,0	1,4	8,5	2,2	4,0	2,4	1,6	1,4	9,4	8,8	6,4
Non food	2,2	0,2	2,4	1,1	2,3	1,3	0,9	0,8	8,1	7,9	5,2
Total	4,1	0,6	4,4	1,5	2,8	1,7	1,1	1,0	8,5	8,2	5,6
% estimate error using IFUs (level)											
Food	2,4	3,0	2,4	4,2	3,0	4,3	6,6	7,7	3,4	2,6	4,4
Non food	5,5	9,4	5,2	6,0	5,0	5,6	6,1	6,4	3,7	3,7	4,4
Total	4,2	6,9	4,1	5,3	4,2	5,1	6,3	6,9	3,6	3,3	4,4
Function Φ_2 and option e), formula (11) – Estimation of level											
Number of IFUs											
Food	64	9	63	19	33	19	14	11	79	78	59
Non food	38	5	32	20	39	22	15	14	149	148	103
Total	102	14	95	39	72	41	29	25	228	226	162
% ratio IFUs/sample											
Food	7,7	1,1	7,6	2,3	4,0	2,3	1,7	1,3	9,5	9,4	7,1
Non food	2,2	0,3	1,8	1,1	2,2	1,3	0,9	0,8	8,5	8,4	5,9
Total	3,9	0,5	3,7	1,5	2,8	1,6	1,1	1,0	8,8	8,7	6,3
% estimate error using IFUs (level)											
Food	2,4	3,1	2,5	4,2	3,0	4,4	6,7	7,5	3,3	2,5	4,2
Non food	5,3	9,5	5,2	6,0	4,9	5,5	6,1	6,4	3,8	3,7	4,4
Total	4,2	7,0	4,2	5,3	4,2	5,1	6,3	6,9	3,6	3,2	4,3
Function Φ_2 and option a), formula (2) – Estimation of change											
Number of IFUs											
Food	64	8	68	18	33	20	13	11	77	74	54
Non food	42	8	66	21	43	25	15	13	146	141	93
Total	106	16	134	39	76	45	28	24	223	215	147
% ratio IFUs/sample											
Food	7,7	1,0	8,2	2,2	4,0	2,4	1,6	1,3	9,2	8,9	6,5
Non food	2,4	0,5	3,8	1,2	2,5	1,4	0,9	0,7	8,3	8,0	5,3
Total	4,1	0,6	5,2	1,5	2,9	1,7	1,1	0,9	8,6	8,3	5,7
% estimate error using IFUs (change)											
Food	2,5	4,5	2,5	4,1	3,1	4,2	6,5	7,3	3,3	2,6	4,2
Non food	8,1	17,0	6,6	8,5	7,6	7,8	8,6	9,1	5,3	5,2	6,2
Total	5,9	12,1	5,0	6,8	5,9	6,4	7,8	8,4	4,5	4,2	5,4
Function Φ_2 and option e), formula (11) – Estimation of change											
Number of IFUs											
Food	63	8	64	19	32	19	13	12	77	73	56
Non food	40	7	58	21	41	22	14	13	148	143	100
Total	103	15	122	40	73	41	27	25	225	216	156
% ratio IFUs/sample											
Food	7,6	1,0	7,7	2,3	3,8	2,3	1,6	1,4	9,2	8,8	6,7
Non food	2,3	0,4	3,3	1,2	2,3	1,3	0,8	0,7	8,4	8,2	5,7
Total	4,0	0,6	4,7	1,5	2,8	1,6	1,0	1,0	8,7	8,3	6,0
% estimate error using IFUs (change)											
Food	2,5	4,5	2,9	4,1	3,1	4,2	6,4	7,4	3,6	2,8	4,4
Non food	8,3	16,6	6,9	8,7	7,7	8,3	8,8	9,1	5,3	5,2	6,0
Total	6,0	11,9	5,3	6,9	5,9	6,7	7,9	8,4	4,6	4,3	5,4

TABLE 2

Percent absolute difference between the number of IFU units detected using formula (2) or formula (11) for the "Total retail trade"

	(22)	(23)	(23)	(24)	(25)	(25)	(26)	(26)	(27)	(28)	(28)
Domain	Bias ratio	Empirical bias 5%	Empirical bias 1%	Chebyshev	Stand. normal 5%	Stand. normal 1%	Grubbs test 5%	Grubbs test 1%	Sprent test	Hidiroglou α=3,5	Hidiroglou α=5
Level	3,9	14,3	18,9	2,6	1,4	4,9	3,4	4,0	3,5	6,6	10,5
Change	2,9	6,7	9,8	2,5	4,1	9,8	3,7	4,0	0,9	0,5	5,8

TABLE 3

Percent incidence of units by number of months for which the same unit has been identified as IFU using function Φ_2 (total IFUs = 100)

	(22)	(23)	(23)	(24)	(25)	(25)	(26)	(26)	(27)	(28)	(28)
Number of months for which the same unit is IFU	Bias ratio	Empirical bias 5%	Empirical bias 1%	Chebyshev	Stand. normal 5%	Stand. normal 1%	Grubbs test 5%	Grubbs test 1%	Sprent test	Hidiroglou α=3,5	Hidiroglou α=5
Function Φ_2 and option a), formula (2) – Estimation of level											
1	54,2	57,0	44,5	59,6	54,8	57,6	59,6	61,7	56,6	55,5	58,3
12	0,0	0,0	1,7	0,0	0,0	0,0	0,0	0,0	0,5	0,5	0,6
≤ 3	68,3	68,8	61,4	73,1	73,1	72,7	73,1	74,5	66,7	67,3	69,2
≤ 6	80,3	86,0	70,1	86,5	81,7	81,8	82,7	87,2	73,1	74,9	75,6
≤ 9	92,3	93,5	79,4	92,3	90,4	89,4	92,3	93,6	81,3	83,4	81,4
Function Φ_2 and option e), formula (11) – Estimation of level											
1	60,9	63,1	51,2	63,3	64,2	62,1	63,3	63,3	48,9	44,6	46,0
12	0,0	0,0	0,6	0,0	0,0	0,0	0,0	0,0	0,6	0,6	0,6
≤ 3	75,9	76,9	67,7	73,5	77,8	75,9	73,5	75,5	69,5	66,4	67,8
≤ 6	88,5	83,1	79,3	87,8	87,7	86,2	87,8	89,8	78,8	77,7	78,2
≤ 9	96,6	93,8	89,0	98,0	96,3	96,6	98,0	98,0	89,4	89,0	90,0

6,9%, the use of the 113 *IFUs* got using empirical bias at 1% (that is, the 606% more) leads to an estimate error equal to 4,1% (that is, the 40,6% less).

Of course, in the context of a survey repeated along time *IFUs* could vary from occasion to occasion. Table 3 shows (for level estimation) the percent incidence of units by number of months for which the same unit has been identified as *IFU* using function Φ_2. If one puts the total number of *IFUs* = 100, on the average quite always more than 50% of them are *IFUs* for one month only, meaning that – given the intrinsic seasonality of the variable under study – methods tested are quite elastic and do not depend too heavily on some relevant large units only. For instance, using bias ratio the percent of *IFUs* for just one month is equal to 54,2% with option (2) and to 60,9% using option (11), while the same percentages referred to units that are *IFUs* for not more than 3 months are equal, respectively, to 68,3% and 75,9%.

Option (2) leads to a higher incidence of units that are *IFUs* in almost all the months respect to option (11): for instance, according to bias ratio (22) the relative incidence of units that are *IFUs* in at least 10 months on 12 is 7,7% with option (2) and only 3,4% using option (11). A similar result also occurs when other criteria are used, with the only exceptions of empirical bias (23) at 1% and 5%, for which the 2 incidences are quite similar.

As a resume, if steadiness of the subset of *IFUs* along time is a relevant feature of the follow up process in a short-term survey, one could prefer option (2); on the other hand, if a lower number of *IFUs* represents a constraint, one can choose option (2) or (11) depending on the criterion adopted. For instance, from table 1 we have – especially for estimation of level – that while criteria based on bias ratio or empirical bias lead to a lower number of *IFUs* using option (11), the reverse is true if one uses the Hidiroglou-Berthelot criterion.

For what concerns choice among methods, while the first 3 (Bias ratio and Empirical bias at the 5% and 1% level) evaluate the incidence of each unit on the overall estimate error, the others are based on a measure of "distance" between each unit and the remaining ones. Generally speaking, one should prefer methods where only a few subjective elements occur (so that the Hidiroglou-Berthelot method could be dangerous, leading to quite unsteady results depending on the parameter *a*), and that lead to a number of *IFUs* that can be managed according to the operational survey constraints. Since the need to evaluate the link between number of follow ups and expected decrease of the estimate error seems fundamental, one choice could fall on the bias ratio (22), that does not seem particularly influenced by the score function used and that leads to results quite similar to those obtained using empirical bias (23) at the 1% level. It is worthwhile to note that analogous results have been got in a previous attempt concerned with the monthly industrial production survey (Gismondi, 2006).

Even though the score function Φ_2 – by definition – emphasises more largest units, the compared identification methods are not directly driven by a specific *coverage* criterion, meaning as coverage the relative weight of *IFUs* on the total in terms of the variable used to calculate score functions. By the way, graphs 1 and 2 show, for level and change, the link between estimate errors and coverage levels for each criterion. In these graphs coverage evaluated in terms of number of *IFUs* and their turnover have been put on the vertical axis (using 2 different scales), while on the horizontal axis we have put percent error levels (first row) and labels identifying the various criteria (second row).

On the average, both for level and change the increase of *IFUs* coverage – in terms of number of units or turnover – leads to a decrease of the error level: linear correlation between error and coverage is -0,86 for level, while for change it is -0,78 (coverage based on the number of units) and -0,66 (coverage based on turnover).

Considering level (figure 1), one notes that when coverage is about 1% in terms of number and not higher than 25% in terms of turnover, error ranges between 6,9% and 5,1%. On the average, an estimate error lower than 5% is guaranteed for a coverage equal at least to 3% in terms of number and to 30% in terms of turnover (even though criterion (23) is an exception, leading to a higher error level), while an error lower than 4% is associated to a coverage of about 4% in terms of number and to 35% in terms of turnover. The highest coverage (around 8% in number and near to 45% in turnover) reduces error around 3,5%.

Considering change (figure 2), on the average an estimate error lower than 5% is guaranteed for a coverage equal at least to 5% in terms of number and to 35%

in terms of turnover, while the highest coverage (around 8% in number and near to 45% in turnover) contributes to lower error around 4%. All these results depend on flatness of the variance curve when coverage has reached a relatively large level.

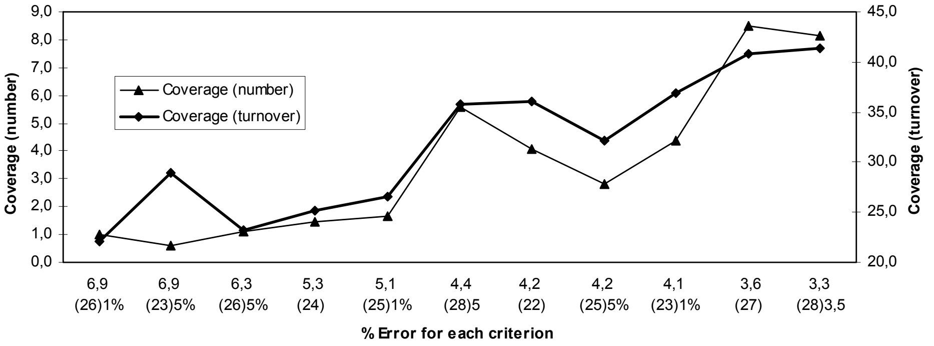

Figure 1 – Estimate errors and coverage using Φ_2 and option a), formula (2) – *Level.*

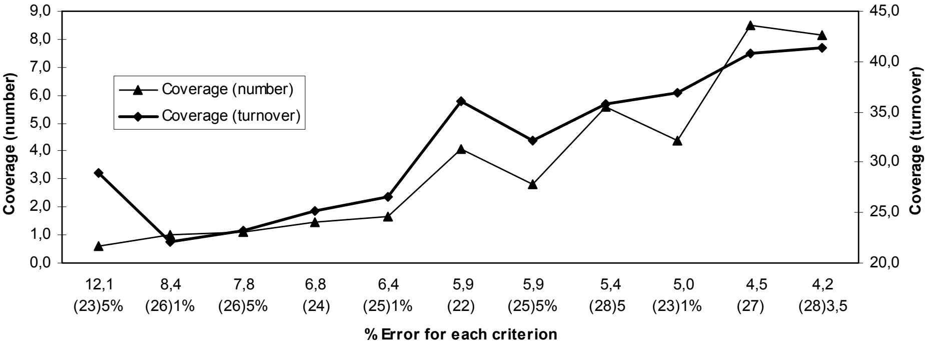

Figure 2 – Estimate errors and coverage using Φ_2 and option a), formula (2) – *Change.*

In the actual retail trade monthly survey, provisional estimates are released after 30 days from the end of the reference month using "quick" respondents. On the average, in 2004 the spontaneous anticipated respondents were 1.315[12] (table 4), that is the 50,8% of the total effective final sample, guaranteeing a turnover coverage equal to 55,4%.

If we consider the total retail trade, the error levels obtained using all these quick respondents are always lower than the correspondent errors based on *IFUs* reported in table 1 (they are equal to 2,7% for level and to 3,3% for change), but

[12] We refer only to the subset of data considered in the simulation: in reality, the actual quick respondents are more than 2.500.

it is worthwhile to note that the same estimate error got for level of non food products (3,7%) can be also reached using *IFUs* identified with option (2) and criteria as the Sprent test and the Hidiroglou-Berthelot method with a=3,5, or option (11) and still the Hidiroglou-Berthelot method. So, a driven choice of influent units could reduce size of a sub-sample guaranteeing a given error level.

A further evidence of the potentially controversial relation linking coverage and precision of estimates derives from table 5. If *IFUs* are stratified according to the 10 domains used for simulations, then a negative correlation between coverage (in terms of turnover) and error level characterises the Hidiroglou-Berthelot method with a=5 (-0,55), but it is not true for the non food sector, where a higher coverage does not guarantee lower error levels. Moreover, bias ratio and the same estimates based on actual quick respondents would even lead to an overall positive correlation (respectively, 0,16 and 0,23), because of the very high error levels concerning the 2 largest employment classes (more than 9 persons employed).

TABLE 4

Main results using actual quick respondents – Average 2004

Domain	Number	Coverage (number)	Coverage (turnover)	% error (level)	% error (change)
Food	479	57,4	58,9	1,3	1,3
Non food	836	47,7	51,8	3,7	4,6
Total	1.315	50,8	55,4	2,7	3,3

A further application concerns methods described in paragraph 2.2 for detecting *IFUs* in the multivariate case. Herein we consider k=2 indicators, given by "level" and "change" referred to average turnover, and the purpose consists in detecting critical units using a score function Φ_2 as that defined by (13) and (14). The main difference respect to results showed in table 1 is that, in this case, *IFUs* for level and change are detected *simultaneously*.

According to table 6, while the number of *IFUs* is fundamentally similar to those detected evaluating *separately* level and change (the *enlarged* criterion mentioned in paragraph 2.2), the consequent error levels are a bit higher. That is because, in this context, errors are based on the use of a unique subset of *IFUs*, that is the same both for level and change, while results of table 1 were got using subsets of *IFUs* for level and change that could be quite different each other, even though their sizes are similar.

For instance, from table 7 one can note that, using bias ratio criterion and formula (2), on the average in 2004 the separate procedures for level and change already analysed – even though leading to 106 *IFUs* both for level and change – would generate a subset of 174 final distinct *IFUs*, since *IFUs* identified for level could be different from those identified for change. A significant saving in the number of distinct *IFUs* to be considered is guaranteed by the conjoint analysis, leading to 105 distinct *IFUs* only, with a quite low loss of estimates' precision for level and change.

TABLE 5

Main results using function Φ_2 and option a), formula (11), for 10 strata, using methods (22), (28) and the actual quick respondents – Estimation of level

	Bias ratio (22)			Hidiroglou α=5 (28)			Actual quick respondents		
Domain	Number of IFUs	Error %	Coverage (turnover)	Number of IFUs	Error %	Coverage (turnover)	Number of IFUs	Error %	Coverage (turnover)
Food 1-2	6	4,8	31,1	10	3,6	37,5	89	2,8	55,8
Food 3-5	10	3,5	41,2	5	7,0	26,6	46	2,4	50,6
Food 6-9	22	2,8	29,3	2	17,2	6,4	59	1,6	43,2
Food 10-19	21	2,3	32,1	4	6,6	12,8	68	1,6	57,9
Food >19	8	1,5	56,2	32	0,4	82,4	217	0,6	87,1
Non food 1-2	4	3,8	25,4	29	2,8	39,4	230	3,4	39,2
Non food 3-5	14	4,0	22,1	10	4,4	18,5	143	1,8	49,8
Non food 6-9	15	3,5	27,2	5	5,7	13,2	90	1,3	53,8
Non food 10-19	3	4,9	52,4	8	4,7	56,3	124	3,9	39,9
Non food >19	3	16,0	43,6	40	8,0	64,8	249	11,0	76,4
Total	106	2,7	36,1	145	4,4	35,8	1.315	2,7	55,4
				Correlations between error and coverage					
Food		-0,58			-0,80			-0,68	
Non food		0,48			0,38			0,76	
Total		0,16			-0,55			0,23	

TABLE 6

Main results using function Φ_2 and option a), formula (2), with options (13) and (14) for a conjoint IFUs identification (level and change)

	(22)	(23)	(23)	(24)	(25)	(25)	(26)	(26)	(27)	(28)	(28)
		Empiri-	Empiri-		Stand.	Stand.				Hidi-	Hidi-
	Bias	cal bias	cal bias	Cheby-	normal	normal	Grubbs	Grubbs	Sprent	roglou	roglou
Domain	ratio	5%	1%	shev	5%	1%	test 5%	test 1%	test	α=3,5	α=5
					Number of IFUs						
Food	64	12	73	18	32	21	13	11	77	73	53
Non food	41	7	61	20	40	23	14	13	142	139	91
Total	105	19	134	38	72	44	27	24	219	212	144
					% estimate error using IFUs (level)						
Food	2,5	4,0	2,4	4,5	3,1	4,6	7,0	7,8	3,7	2,7	4,6
Non food	5,8	11,2	5,3	6,4	5,4	6,2	6,5	6,9	4,1	4,0	4,6
Total	4,5	8,4	4,1	5,6	4,5	5,6	6,7	7,3	3,9	3,5	4,6
					% estimate error using IFUs (change)						
Food	2,5	4,1	2,4	4,5	3,1	4,6	6,9	7,8	3,6	2,6	4,5
Non food	7,0	12,2	6,2	7,3	6,6	7,1	7,5	7,8	4,6	4,6	5,2
Total	5,2	9,0	4,7	6,2	5,2	6,1	7,2	7,8	4,2	3,8	5,0

TABLE 7

Number of IFUs identified using bias ratio criterion (22) with the conjoint and the separate procedures for level and change – Months of 2004 and average

Method	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Average
Separate	165	183	204	206	170	163	157	182	168	157	143	186	174
Conjoint	90	128	136	143	110	105	99	112	90	72	80	94	105

6. CONCLUSIONS

In the wide context of non–response treatment, the aspect concerned with non-response prevention is sometimes under-evaluated, or treated according to criteria not always fully appropriate.

Generally speaking, given the estimation strategy, the identification of critical units that should be object of a priority follow up system in case of non response depends on: 1) the particular individual score function adopted, evaluating the risk due to the not availability of a unit for estimates; 2) the criterion used for detecting critical values of the score function and for identifying *IFUs*.

Both the previous aspects have been re-analysed according to the actual mostly used procedures and some new proposals, with an operative application referred to a real short-term estimation process, for which timeliness of decisions is fundamental.

The underlying idea is that statistical relevance and degree of coverage are related concepts, that however should be kept *separate*. The *IFU* feature is an *intrinsic* character of a statistical unit and techniques to identify critical units should not be only based on coverage indicators: according to a procedure driven by coverage only, critical units are automatically detected through a simple decreasing ranking and the selection of all those first units in the rank guaranteeing a certain coverage level. On the other hand, a criterion could lead to a fewer number of *IFUs* than others – and, consequently, to a relatively low *IFU* coverage – but it should be preferred if it is based on a rationale strictly connected with the estimate error evaluation.

Empirical results show that – at least for what concerns the techniques compared – the most sensitive aspect to be carefully evaluated is the choice of the criterion for detecting critical units rather than the possibility to build up the individual score functions in different ways. On the other hand, the score function (2) should lead to a lower monthly variability of the number of units detected as *IFUs* respect to function (11). Different criteria can lead to a very different number of *IFUs* and, on the average, the decrease of estimate error is quite less than proportional respect to the increase of *IFUs*. The other conditions being steady, the final choice should be probably in favour of criteria strictly linked to features of estimator and sampling design, as the bias ratio and the empirical pseudo bias defined in paragraph 4.1.

ISTAT, Italian National Statistical Institute ROBERTO GISMONDI

REFERENCES

H. BOLFARINE, S. ZACKS (1992), *Prediction theory for finite populations*, Springer-Verlag, Berlin.

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.

S. CHEN, H. XIE (2004), *Collection follow up score function and response bias*, "Proceedings of the SSC Annual Meeting – Survey Methods Section", pp. 69-76, Statistics Canada.

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.

H.E. DAVILA (1992), *The Hidiroglou-Berthelot method*, "Statistical data editing methods and techniques", United Nations.

A. DE JONG (2003), *Impect: recent developments in harmonised processing and selective editing*, available on www.oecd.org /dataoecd.

W.E. DEMING (1953), *On a probability mechanism to attain an economic balance between the resulant error of non-response and the bias of non-response*, "Journal of the American Statistical Association", 48, pp. 743-772.

J.J. DROESBEKE, B. FICHET, P. TASSI (1987), *Les sondages*, Economica, Paris.

I. DRUDI, C. FILIPPUCCI (2000), *Inferenza da campioni longitudinali affetti da selezione non casuale*, in C. FILIPPUCCI (ed.), *Tecnologie informatiche e fonti amministrative nella produzione di dati*, pp. 415-432, Franco Angeli, Milano.

EUROSTAT (2000), *Short-term statistics manual*, Eurostat, Luxembourg.

EUROSTAT (2005*a*), *Council regulation N° 1165/98 amended by the regulation N° 1158/2005 of the European Parliament and of the Council – Unofficial consolidated version*, documento non pubblicato, Eurostat, Lussemburgo.

EUROSTAT (2005*b*), *The burden on enterprises resulting from the STS regulation*, technical document discussed in the *Short-term statistics working party*, 21-22 June, Eurostat, Luxembourg.

L. FATTORINI (2006), *Applying the Horvitz-Thompson criterion in complex designs: a computer-intensive perspective for estimating inclusion probabilities*, "Biometrika", Vol. 93, 2, pp. 269-278.

R. GISMONDI (2000), *Metodi per il trattamento dei dati anomali nelle indagini longitudinali finalizzate alla stima di variazioni*, "Contributi", 8, Istat, Roma.

R. GISMONDI (2006), *L'individuazione delle unità statistiche "influenti" nell'indagine mensile sulla produzione industriale*, presentazione nel seminario: *La rilevazione mensile della produzione industriale: aggiornamento metodologico e disegno del nuovo sistema informativo*, 14 marzo 2006, Istat, Roma.

L. GRANQUIST (1990), *A review of some macro-editing methods for rationalizing the editing process*, *Proceedings of Statistics Canada Symposium,* 90.

L. GRANQUIST, J.G. KOVAR (1997), *Editing of survey data: how much is enough?*, in *Survey Measurement and Process Quality*, pp. 415-435, John Wiley & Sons, New York.

F. GRUBBS (1969), *Procedures for detecting outlying observations in samples*, "Technometrics"*,* Vol. 11, 1, pp. 1-21.

D. HEDLIN (2003), *Score functions to reduce business survey editing at the U.K. Office for National Statistics*, "Journal of Official Statistics", Vol. 19, 2, pp. 177-199.

M.A. HIDIROGLOU, J.M. BERTHELOT (1986), *Statistical editing and imputation for periodic business surveys*, "Survey Methodology", 12, pp. 73-84.

J.W. HUNT, J.S. JOHNSON, C.S. KING (1999), *Detecting outliers in the monthly retail trade survey using the Hidiroglou-Berthelot method*, http://www.amstat.org/sections/srms/proceedings/papers/1999_093.pdf.

ISTAT (1989), *Manuali di tecniche d'indagine vol. 4-5*, Istat, Roma.

ISTAT (1998), *La nuova indagine sulle vendite al dettaglio: aspetti metodologici e contenuti innovativi,* Metodi e norme, 3, Istat, Roma.

ISTAT (2006), Report finale del progetto "Sperimentazione di stime anticipate per specifici indicatori congiunturali, finalizzata al rilascio in produzione delle relative metodologie" (a cura di S. FALORSI e R. GISMONDI), Istat, Roma.

G. KALTON, D. KASPRZYK, D. MCMILLEN (1989), *Non-sampling errors in panel swurveys*, in D. KASPRZYK, G. DUNCAN, G. KALTON, M.P. SINGH (eds.), *Panel Surveys*, pp. 249-270, John Wiley & Sons, New York.

M. LATOUCHE, J.M. BERTHELOT (1992), *Use of a score function to prioritise and limit recontacts in business surveys*, "Journal of Official Statistics", 8, pp. 389-400.

D. LAWRENCE, R. MCKENZIE (2000), *The general application of significance editing*, "Journal of Official Statistics", 16, pp. 243-253.

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.

R. MCKENZIE (2003), *A framework for priority contact of non respondents*, available on www.oecd.org /dataoecd.

L. PIETSCH (1995), *Profiling large businesses to define frame units*, in B. COX, D. BINDER, N. CHINNAPPA, A. CHRISTIANSON, M. COLLEDGE, P. KOTT (eds.), *Business Survey Methods*, pp. 101-114, John Wiley & Sons, New York.

R. PHILIPS (2003), *The theory and application of the score function to prioritize and limit recontacts in editing business surveys*, "Proceedings of the SSC Annual Meeting – Survey Methods Section", pp. 121-126, Statistics Canada.

S. PURSEY (2003), *Use of the score function to optimize data collection resources in the unified enterprise*, "Proceedings of the SSC Annual Meeting – Survey Methods Section", pp. 117-120, Statistics Canada.

C.P. QUESENBERRY, H.A. DAVID (1961), *Some tests for outliers*, "Biometrika", 48, pp. 379-390.

L. RIZZO, G. KALTON, M.J. BRICK (1996), *A comparison of some weighting adjustment methods for panel non-response*, "Survey Methodology", 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, New York.

R.E. SHIFFLER (1988), *Maximum Z-scores and outliers*, "American Statistician", 42, pp. 79-80.

P. SPRENT (1998), *Data driven statistical methods*, Chapman and Hall.

P. SPRENT, N.C. SMEETON (2001), *Applied nonparametric statistical methods*, 3d ed., Chapman and Hall.

R. SUCCI, A. CIRIANNI (2005), *La produzione di stime anticipate di fatturato nel settore degli altri servizi*, documento interno, ISTAT, Roma.

SUMMARY

Score functions and statistical criteria to manage intensive follow up in business surveys

In the frame of a statistical survey, the identification of non respondent units that should be object with priority of a reminder action (*Intensive Follow Up - IFU*), with the aim to produce enough good estimates, represents a relevant, but quite not deeply analysed methodological aspect. In this context, we propose and compare some score functions - that can be all reconnected to a generalised function – evaluating how much is dangerous the exclusion from calculations of each unit. Moreover, we evaluate and compare some criteria aimed at identifying *IFU* units by means of suitable statistical tests or thresholds derived by parametric or non parametric methods. A comparative empirical application on a panel of Italian retail trade businesses has been carried out and commented.