WP. 33
ENGLISH ONLY

**UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE**

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Neuchâtel, Switzerland, 5-7 October 2009)

Topic (vi): New and emerging methods

POTENTIAL OPERATIONAL IMPROVEMENTS OF THE HULLIGER'S CRITERION FOR MANAGING OUTLIERS IN LONGITUDINAL SAMPLING SURVEYS

Supporting Paper

Prepared by Roberto Gismondi, ISTAT, Italy

I. OUTLIERS' DETECTION AND TREATMENT IN SAMPLING SURVEYS

In survey sampling theory, the interest usually lies in the estimation of finite population parameters such as the total of a variable of interest y for a given finite population. The observed sample may include *outlier* observations, that can be defined as values falling in the left or right tail of the observed empirical y-distribution. A sensitive role is played by *representative* outliers: they are given by anomalous observations not due to mistakes occurred along the data capturing process, but to some events strictly connected with the statistical unit behaviour. Each representative outlier is the sample evidence of a certain sub-group of units in the population which are characterised by a similar profile and that are *represented* by the observed outlier in the estimation process.

The influence of extreme values in the overall estimation error – especially when outliers data are referred to *large* units – could be quite dangerous without a specific system of detection and treatment (Searls, 1966). As a consequence, the following problems must be faced:

a) how to identify outlier observations;
b) how to treat them after the identification, according to one (or a combination) of these criteria:
    1) outliers are excluded from further calculations (their sampling weight is put equal to zero) or included as self-representative (the sampling weight is put equal to one);
    2) outlier data are re-estimated as they were missing observations or according to some trimming rule;
    3) outlier data are not changed, but their sampling weight is reduced.

As regards point a), a basic quite accepted idea is that the nature of "outlier" for a statistical observation is not just an *intrinsic individual feature*, but an attribute that can arise evaluating the role played by each unit *in the estimation process* in comparison with that played by the other sampling units taken as a whole. Generally speaking, this evaluation is based on the use of proper acceptance thresholds. As regards point b), it is recommended that use of re-weighting is always driven by the evaluation of the trade/off between bias reduction and increase of variance (Welsh and Ronchetti, 1998). Useful outlier detection procedures should be as much as possible time saving – especially when large data-sets are managed (Latouche and Berthelot, 1992) –, contain under a reasonable level the number of sampling units detected as outliers (Gismondi, 2002) and be founded on objective rules for fixing thresholds or applying trimming (Kokic and Bell, 1994). Moreover, in the field of official statistics strategies for dealing with outliers should not be too heterogeneous, in order to guarantee a common theoretical background according to which the outliers' treatment is carried out. In particular, as regards short-term business statistics some late best practices are commented in AA.VV. (2008).

More in details, Chambers (1986) proposed an estimator that reduces to 1 the weights of extreme

observations, while Hulliger (1995, 1999) presented an estimator under a model based approach (Särndal *et al*, 1993) based on weights for outliers that are reduced (but not necessarily equal to 1) with respect to the original ones according to a standardised function, expressing the difference between observed and expected values. Chambers *et al*. (2000) re-analyse the recourse to trimming as an alternative to re-weighting, while Beaumont and Alavi (2004) focus more on the estimation process, evaluating performances of a family of robust generalised regression estimators. In this context, we will deal with the Hulliger's criterion (section II), according to which one can identify and treat outliers at the same time: i) without the need of complex elaborations and ii) applying a model-based alternative to weights' trimming[1] (Elliott and Little, 2000). In particular, we propose some changes that may improve its efficiency: they concern both the choice of the threshold for detecting outliers and the rule for re-weighting (section III). We also show links with other methods for dealing with outliers (section IV) – proposing in sub-section IV.A a post-stratification approach – and present the main outcomes of two empirical attempts based on true turnover data (section V). Perspective conclusions have been drawn in section VI.

The main proposal consists in the choice of the acceptance threshold based on a "calibration" approach, which can be implemented using past *y*-data normally available in the frame of longitudinal surveys. The only constraint – even though fundamental – is the possibility to know (or to estimate) the correspondent past true estimation error. Ren and Chambers (2002) already introduced the principle of robust imputation via reverse calibration; herein we develop an operational strategy not just aimed at modifying observed values or re-estimating missing ones, but at fixing an objective threshold that would have been optimal if applied to past data. Availability of time series of historical data referred – at least in part – to the same subset of units represents an information bulk not always fully exploited in the frame of longitudinal surveys. Even though the discussion is more focused on business surveys data, with simple adaptations the basic criteria can be applied to more general contexts as well.

II. THE HULLIGER'S ROBUSTIFIED RATIO ESTIMATOR

Given a population *P* with size *N*, the target is the estimation of the population total Y_P through a sample *s* with size *n* and on the basis of sampling weights *w*. We suppose the regression superpopulation model *R* defined as: $y_i=\beta x_i+\varepsilon_i$, with $E(\varepsilon_i)=0$, $Var(\varepsilon_i)=\sigma^2 x_i$, $Cov(\varepsilon_i,\varepsilon_j)=0$ for each (*i*) or (*i*≠*j*), where *x* is an auxiliary variable available for each unit in the population with total X_P, with β and σ^2 unknown parameters. The one-step robustified ratio estimator proposed by Hulliger is based on an estimate of the ratio $\hat{\beta}_0 = med(y_i, w)/med(x_i, w)$ through weighted quantiles (ISTAT *et al*., 2007, 58-59) and the standardised absolute residuals $a_i = \left|y_i - \hat{\beta}_0 x_i\right| \Big/ \sqrt{x_i}$. Let the median of the absolute residuals be $\hat{\sigma}_a = med(a_i, w)$. Then *robust weights* are defined as:

$$w_{Hi} = u_i w_i \quad \text{where:} \quad u_i = \begin{cases} 1 & if \quad a_i \le c\,\hat{\sigma}_a \\ c\,\hat{\sigma}_a / a_i & if \quad a_i > c\,\hat{\sigma}_a \end{cases} \quad \text{for each } i \tag{1}$$

where *c* is a parameter to be chosen. The one-step robustified ratio estimator (*RRE*) is:

$$T_H = \left(\sum_s w_i u_i y_i\right)\left(\sum_s w_i u_i x_i\right)^{-1} X_P = \left(\sum_s w_{Hi} y_i\right)\left(\sum_s w_{Hi} x_i\right)^{-1} X_P . \tag{2}$$

The *RRE* is a linear estimator based on weights given by $(X_P w_{Hi}) \Big/ \sum_s w_{Hi} x_i$. It is equivalent to the ordinary ratio estimator applied to couples (*x*, *y*) that, when $a_i > c\,\hat{\sigma}_a$, modify into the new couples of *truncated* values (*ux*, *uy*). It is also different with respect to the ordinary ratio estimator, e.g. the model *BLU* predictor (Cicchitelli *et al*., 1992, 385-387).

There are 3 aspects influencing the *RRE* form and its performance: 1) the rule linking w_H to *w*; 2) the definition of correctors *u* in (1); 3) the choice of parameter *c* in (1). The re-weighting system (1) can be

[1] However, it is worthwhile to note that reducing weights is equivalent to apply the original weights to trimmed values, and vice-versa.

viewed as a robust estimation criterion that reduces the outliers' weight according to the standardised distance between the observed and the theoretical *y*-value. A major advantage due to (1) consists in the possibility to detect and treat outliers at the same time; the sum of new weights will be lower than the sum of the original ones, but that does not produce additional bias of estimates because in the estimator (2) weights operate both at numerator and denominator. The correctors *u* in (1) play the role of link between each outlier in the observed sample and the corresponding number of (similar) outlier observations present in the whole population, so that they establish a formal connection between the sample and the population density distributions. On the other hand, a subjective choice of the threshold parameter *c* may lead to wrong conclusions, especially in the frame of short-term statistics, where seasonal effects suggest the recourse to different parameters depending on the month and/or other stratification criteria.

III. POTENTIAL IMPROVEMENTS

A. Weights *w*

We propose the alternative transformation $w_i^* = 1 + u_i(w_i - 1)$, because when $u_i \to 0$ (very anomalous unit) $w_i^* \to 1$ (the *i-th* unit is self-representative). This is a less extreme option with respect to the alternative $w_{Hi} \to 0$ (the unit disappears) and may be preferred especially in case of *representative* outliers. On the other hand, it is still reasonable to reduce as much as possible (even toward zero) the weight of suspicious *non-representative* outliers. The difference between w_{Hi} and w_i^* may be neglected only when *N* is quite larger than *n*.

B. Correctors *u*

Correctors *u* in (1) can be defined on the basis of a lightly different position. The basic idea consists in the introduction of a parameter α aimed at increasing or decreasing the quickness of the change of the original weights *w*. We still suppose that u_i=1 if $a_i \le c\,\hat{\sigma}_a$. Moreover, we can put:

$$u_{\alpha i} = \left(c\,\hat{\sigma}_a / a_i\right)^{\alpha} \quad \text{if} \quad a_i > c\,\hat{\sigma}_a . \tag{3}$$

When α=1, then $u_{\alpha i}$=u_i. When α>1 (α<1), $u_{\alpha i}$ tends more (less) quickly than u_i to zero, as well as the corresponding weights $w_{\alpha i}$. Since each weight expresses the number of not observed units in *P* represented by the corresponding sample unit, the option α>1 implies that the extreme observations (very large or very small) included in the observed sample are considered more rare in the whole population rather than when $\alpha \le 1$, and vice-versa.

C. Selection of *c*

As regards this crucial aspect, a *calibration* approach may improve the *RRE* efficiency, reducing the risk of additional bias due to subjective choices of *c*. In particular, too low levels of *c* may lead to the identification of outliers even when true outliers do not exist. The basic hypothesis consists in the availability of historical data, e.g. the possibility to evaluate the relationship between *y* and *x* using a past sample drawn from a past population – both referred to a time (*t*-1) – which *y* total is known at the rime *t* when current estimates must be released. The procedure follows the steps listed below:

a) at time *t* we observe a sample including *n* units. We suppose to know *y* values of each sample unit referred to time (*t*-1), as well as the total Y_P at time (*t*-1), say $Y_{P(t\text{-}1)}$.
b) Supposing to apply the same sample weights at times *t* and (*t*-1), we carry out the *RRE* calculation at time (*t*-1) trying a set of values for *c*. For each *c* we also calculate the absolute error of estimates, given by $AE_{c(t-1)} = \left|T_{H(t-1)} - Y_{P(t-1)}\right|$, that can be also defined as *calibration error*.
c) We choose that particular *c**(*t*-1) such that: $AE_{c^*(t-1)} = \min_c \left\{AE_{c(t-1)}\right\}$.
d) We apply *c**(*t*-1) for implementing (1) and (2) at time *t*. Let's note that, of course, at time *t* the

optimal (unknown) $c^*(t)$ minimising $AE_{c(t)}$ may be different from $c^*(t\text{-}1)$.

The method – derived by the calibration approach as a tool to reduce bias of sample estimates (Lundström and Särndal, 1999) – is founded on the idea that the optimal c that would have guaranteed a near-calibration of sample estimates with respect to the population total at time (t-1) should work fine at time t as well. There are 2 ways for implementing the procedure. If at times t and (t-1) the variables under study are given by, respectively, $y_{(t)}$ and $y_{(t\text{-}1)}$, then:

a) at time t the auxiliary variable x is given by $y_{(t\text{-}1)}$, while at time (t-1) it is given by $y_{(t\text{-}2)}$;
b) at time t the auxiliary variable x is given by $x_{(t)}$, while at time (t-1) it is given by $x_{(t\text{-}1)}$.

For instance, in the frame of business surveys, if y is turnover (monthly, quarterly or yearly), option a) can be carried out using as auxiliary variable the correspondent turnover of the previous year, while option b) can be implemented using as auxiliary variable the yearly turnover referred to the year before, derived from a business register. The choice strictly depends on the knowledge of the amount $Y_{P(t\text{-}1)}$: if it is not available, then option ii) might be the only one useful in practice.

The calibration approach should be particularly useful if the relation between the x and y density distributions are quite similar, as well as if the number and the relative magnitude of outlier data are quite the same. Let's also note that it may be used for fixing an objective threshold in the frame of other outliers' detection methods as well, as for instance the Hidiroglou and Berthelot criterion (1986) based on empirical quartiles. It is worthwhile to note that the calibration approach above may lead to the identification of a quite large number of outliers, because that might turn out to be a necessary condition in order to satisfy the calibration constraint. In these circumstances one may guess if all these units are real outliers. The problem could be managed imposing the additional condition that the optimal solution should minimize the calibration error and at the same time guarantee that the relative incidence of outliers' is not larger than a given percent of the whole observed sample (say, 10%). Another strategy could be based on the evaluation – for each c under analysis – of the average difference between original and final weights, and the final choice will fall on the particular c for which this difference rears – whatever is the correspondent calibration error – or under the additional condition that the calibration error is lower of a given threshold. On the other hand, we must underline that a constraint of this approach is the need to try a wide set of potential c in order to find that which minimise the calibration error.

Even though the recourse to different parameters c for different estimation domains is recommended, especially in a short-term survey context one may decide to use a more steady c whatever is the reference month or quarter. The choice of a unique c can be driven by various criteria:

1) minimization of the average calibration error;
2) minimization of the real average estimation error calculated on previous periods;
3) minimization of the variability of c estimates evaluated through a given number of attempts (for instance, different months);
4) the "minimax" approach evaluated on: i) the average calibration error; ii) the number of periods for which a particular c is optimal.

A further, relevant issue concerns the risk due to the use of the optimal c_x in place of the optimal c_y. There is an underlying link between the 2 threshold parameters, which can be assessed through the formalisation of a second model expressing the link between x and z: $x_i=\beta_x z_i+\delta_i$, with $E(\delta_i)=0$, $Var(\delta_i)=\sigma_x^2 z_i$, $Cov(\delta_i,\delta_j)=0$. The threshold rule as regards the x values is given by:

$$\left|x_i-\hat{\beta}_x z_i\right| \Big/ \sigma_x\sqrt{z_i} > c_x \,. \tag{4}$$

Since from the models concerning y (where $\beta=\beta_y$) and x we have that $x_i=E(y_i)/\beta_y$ and $z_i=E(x_i)/\beta_x$, and since it is reasonable to suppose $\beta_y>0$, from relation (4) we have:

$$\frac{\left|x_i-\hat{\beta}_x z_i\right|}{\sigma_x\sqrt{z_i}}=\frac{\left|E(y_i)/\hat{\beta}_y-\hat{\beta}_x z_i\right|}{\sigma_x\sqrt{z_i}}=\frac{\left|E(y_i)-\hat{\beta}_y E(x_i)\right|}{\hat{\beta}_y\sigma_x\sqrt{z_i}}=\left(\frac{\sqrt{\hat{\beta}_x}}{\hat{\beta}_y}\frac{\sigma_y}{\sigma_x}\right)\frac{\left|E(y_i)-\hat{\beta}_y E(x_i)\right|}{\sigma_y\sqrt{E(x_i)}}>c_x$$

so that, finally, we can write:

$$\frac{\left|E(y_i)-\hat{\beta}_y E(x_i)\right|}{\sigma_y\sqrt{E(x_i)}} > c_x\left(\frac{\hat{\beta}_y}{\sqrt{\hat{\beta}_x}}\frac{\sigma_x}{\sigma_y}\right). \tag{5}$$

Since the threshold rule as regards the *y* values is defined through the condition:

$$\left|y_i-\hat{\beta}_y x_i\right| \Big/ \sigma_y\sqrt{x_i} > c_y\,, \tag{6}$$

if $\hat{\sigma}_x$ and $\hat{\sigma}_y$ are estimates of σ_x and σ_y respectively, one may put:

$$c_y \cong \left(\frac{\hat{\beta}_y}{\sqrt{\hat{\beta}_x}}\frac{\hat{\sigma}_x}{\hat{\sigma}_y}\right)c_x\,. \tag{7}$$

From position (6) and (7), if $\hat{\beta}_y = k_\beta\hat{\beta}_x$ and $\hat{\sigma}_y = k_\sigma\hat{\sigma}_x$, finally it follows the proportionality between c_y and c_x, since we have:

$$c_y \cong \left[(k_\beta/k_\sigma)\sqrt{\hat{\beta}_x}\right]c_x = K\,c_x\,. \tag{8}$$

In the table 1 we have reported levels of *K* for some levels of $\hat{\beta}_x$ (rows) and of the ratios k_β/k_σ (columns). Levels selected cover a reasonable range according to many empirical contexts where *y*, *x* and *z* represent the same variable observed at different time points. We can see how *K* is quite always higher than one when $k_\beta > k_\sigma$. Without loss of generality, if we suppose a *y*-trend increasing along time, *K*>1 means that when the model *variability* is increasing less than proportionally with respect to the model *level* we can accept a level for c_y larger than c_x – and, as a consequence, a wider acceptance interval – and vice-versa. This result is fully coherent with the idea that, the other conditions being steady, data with a lower coefficient of variation should include a lower number of outliers, and vice-versa.

Table 1: *Levels of the K coefficient for some levels of* β_x *and of the ratios* k_β/k_δ

β_x estimate	Ratios k_β/k_δ								
	1,20	**1,15**	**1,10**	**1,05**	**1,00**	**0,95**	**0,90**	**0,85**	**0,80**
1,20	1,31	1,26	1,20	1,15	1,10	1,04	0,99	0,93	0,88
1,15	1,29	1,23	1,18	1,13	1,07	1,02	0,97	0,91	0,86
1,10	1,26	1,21	1,15	1,10	1,05	1,00	0,94	0,89	0,84
1,05	1,23	1,18	1,13	1,08	1,02	0,97	0,92	0,87	0,82
1,00	1,20	1,15	1,10	1,05	1,00	0,95	0,90	0,85	0,80
0,95	1,17	1,12	1,07	1,02	0,97	0,93	0,88	0,83	0,78
0,90	1,14	1,09	1,04	1,00	0,95	0,90	0,85	0,81	0,76
0,85	1,11	1,06	1,01	0,97	0,92	0,88	0,83	0,78	0,74
0,80	1,07	1,03	0,98	0,94	0,89	0,85	0,80	0,76	0,72

IV. LINKS WITH OTHER METHODS

A. A post-stratification approach

Under the model *R* introduced in section II, we know that the optimal predictor of the total is given by:

$$\hat{y}' = \sum_S y_i + \hat{\beta}\sum_{\bar{S}} x_i \quad \text{where the sampling weight of the } i\text{-}th \text{ unit is: } \quad w_i' = \sum_P x_i\left(\sum_S x_i\right)^{-1}. \tag{9}$$

According to definitions already introduced, we can suppose that all the units detected as outliers belong to a sub-sample S_O derived from a sub-stratum of the whole population, say P_O, including all the N_O outlier units. All the remaining (good) units will belong to a second sub-stratum P_G including N_G=N-N_O units. These sub-populations do not derive from any preliminary stratification, but depend on some latent factor underlying units under observation[2]. For each of the 2 sub-populations (labelled with h, where h=O,G) the following model can be supposed: y_{hi}=$\beta_h x_i$+ε_i, with $E(\varepsilon_i)$=0, $Var(\varepsilon_i) = \sigma_h^2 x_i$, $Cov(\varepsilon_i,\varepsilon_j)$=0 for each ($i$) or ($i$≠$j$). Under this model the optimal predictor would be:

$$\hat{y} = \frac{N_G}{N}\hat{y}_G + \frac{N_O}{N}\hat{y}_O \tag{10}$$

where $\hat{y}_G$ and $\hat{y}_O$ keep the formal structure of predictor (9). As a consequence, the (right) sampling weight of the *i-th* unit detected as "good" or "outlier" should be, respectively:

$$w_{Gi*} = \frac{N_G}{N}\sum_{P_G} x_i \left(\sum_{S_G} x_i\right)^{-1} \quad \text{and} \quad w_{Oi*} = \frac{N_O}{N}\sum_{P_O} x_i \left(\sum_{S_O} x_i\right)^{-1} \tag{11}$$

where the asterisk label indicates that the sum of all these new weights do not necessarily reproduce the sum of the original (wrong) weights w' in (9), say N.

The formula (10) can be connected to the Hulliger's criterion, since it can be seen as a more general re-weighting rule to be applied to each of the n_O outliers in S_O detected according to some criterion. We can also note that the new weights (11) may be both lower or larger with respect to the original one. The relation with the Hulliger's re-weighting rule can be directly assessed noting that:

$$w_{Hi} = w_i \frac{c\,\hat{\sigma}_a \sqrt{x_i}}{\left|y_i - \hat{\beta}_0 x_i\right|} \quad \text{and} \quad w_{Oi*} = w_i \frac{N_O}{N}\frac{\sum_{P_O} x_i}{\sum_P x_i}\left(\frac{\sum_{S_O} x_i}{\sum_S x_i}\right)^{-1}. \tag{12}$$

The formal structure of w_{Oi*} is simpler, mainly if the relative outliers' x-weight in the sample and in the whole population post-strata are quite similar, so that the second ratio in the second (12) formula is around one and we simply have: $w_{Oi*} \cong w_i N_O / N$.

The main problem consists in the estimation of the 2 quantities in the second (12) referred to the population P_O (N_O and the x-sum over P_O). In particular, if we suppose a simple random sampling scheme – that is coherent with the position x_i=constant for each i – and $N_O / N \cong n_O / n$, we obtain $w_{Oi*} = N\, n_O / n^2 = w_i\, n_O / n$ and of course we have also $w_{Gi*} = w_i\, n_G / n$. Since we want that the sums of new and old weights are the same, we can finally put:

$$\begin{cases} w_{Gi} = \gamma\, w_{Gi*} = \gamma\, w_i\, n_G / n \\ w_{Oi} = \gamma\, w_{Oi*} = \gamma\, w_i\, n_O / n \end{cases} \quad \text{where:} \quad \gamma = n^2 / \left(n_O^2 + n_G^2\right). \tag{13}$$

The ratio between the "right" weights to be assigned to good and outlier units – that is w_{Gi} / w_{Oi} – is constant with respect to the ratio n/N and decreases as the percent of outliers in the sample n_O/n increases. Since it may happen that the new post-stratified weights (11) – as well as the Hulliger's weights – are lower than one, we can justify this outcome as the result of a process that applies a final weight equal to one (self-representation) to all the outlier observations which original y-values have been properly trimmed (that implies a joint use of both the trimming and re-weighting approaches, as already mentioned in footnote 1).

A similar approach consists in supposing to post-stratify the original population in (n_O+1) sub-populations (and sub-models), one for the "good" units and one for each of the n_O outliers: in this way we suppose that each outlier labelled as (i) derives from a specific sub-stratum $P_{O(i)}$ including $N_{O(i)}$ units. Even though this option could often be more realistic from a theoretical point of view, one should

[2] A wider discussion on this topic in relation with the non-response problem is available in Gismondi (2008).

carefully evaluate the trade/off between the right model choice and the possibility to estimate all the additional $2(n_O-1)$ population parameters that are necessary in order to implement estimates, since from (10) and (11) it follows straightforwardly that:

$$\hat{y} = \frac{N_G}{N}\hat{y}_G + \sum_{h=1}^{n_O} \frac{N_{O(i)}}{N}\hat{y}_{O(i)} \qquad \text{where:} \qquad w_{Oi*} = w_i \frac{N_{O(i)}}{N} \frac{\sum_{P_{O(i)}} x_i}{\sum_P x_i} \left(\frac{\sum_{S_{O(i)}} x_i}{\sum_S x_i} \right)^{-1} . \tag{14}$$

B. The bias-ratio criterion

The *bias ratio* criterion derives from an adaptation of the classical theory of confidence intervals. In this sub-section we show how there is an implicit link between the adoption of this criterion for outliers' detection and the choice of a threshold in the Hulliger's approach. Let's suppose that the total Y_P is the benchmark reference for assessing precision of the estimates. If $\hat{y}$ is an estimator of the total based on all the n units of the sample, we can suppose to exclude from estimations a given sub-sample S_A composed by n_A units, so that $S = S_A \cup S_{-A}$, where S_{-A} is the sub-sample used for estimation. If $\hat{y}_{-A}$ is the estimator based on S_{-A}, then the *bias ratio* of this estimate as:

$$BR(\hat{y}_{-A}) = \left| Y_P - \hat{y}_{-A} \right| \left[Var(\hat{y}_{-A}) \right]^{-0,5} . \tag{15}$$

Since the global error of the estimate is the sum of squared bias and sample variance, the bias ratio is given by the squared root of the incidence of the former error component on the latter. 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 within a confidence interval derived from the standardised normal distribution Z. This probability is: by: $\Pr\left[-z_{1-\alpha/2} - BR(\hat{y}_{-A}) < Z < z_{1-\alpha/2} - BR(\hat{y}_{-A}) \right]$ – where $z_{(1-\alpha/2)}$ is the percentile of the standardised normal cumulated distribution leaving on the right a probability equal to $\alpha/2$. Thus the coverage probability equals the nominal, desired confidence level, $(1-\alpha)$, only if the bias ratio is equal to 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% results into a loss of coverage probability lower than 1%, which is therefore negligible if compared with other shortcomings of common variance estimates.

The underlying idea related to the use of (15) in relation with the outliers problem consists in testing the significance of the difference between the y-estimate based on a complete data set and the data set not including a certain sub-set of units. On the basis of a slight adaptation of (15), the *selective* choice of units detected as outliers can be driven by the evaluation of how much bias one should accept at each step. If the estimate $\hat{y}$ substitutes the original (unknown) parameter Y_P, the operational rule is based on these steps:

a) for each unit $i \in S$ we evaluate the *approximate* bias ratio *br*:

$$br(\hat{y}_{-i}) = \left| \hat{y} - \hat{y}_{-i} \right| \left[Var(\hat{y}_{-i}) \right]^{-0,5} \tag{16}$$

and we label with [1] the unit with the largest *br*, while $\hat{y}_{-[1]}$ is the estimate based on the sub-sample excluding the unit [1]. If $br(\hat{y}_{-[1]}) \le \lambda$ – where λ may be equal to 0,10 or to another threshold – no unit is identified as outlier and the procedure stops, otherwise the unit labelled with [1] is detected as outlier and the procedure skips to the step b).

b) If we indicate with the label [2] the unit with the second largest bias ratio after unit [1], we evaluate:

$$br(\hat{y}_{-[1,2]}) = \left| \hat{y} - \hat{y}_{-[1,2]} \right| \left[Var(\hat{y}_{-[1,2]}) \right]^{-0,5} \tag{17}$$

where $\hat{y}_{-[1,2]}$ is the estimate based on the sub-sample excluding *both* units [1] and [2]. If $br(\hat{y}_{-[1,2]}) \le \lambda$, the unit [2] is not detected as outlier and the procedure stops, otherwise the unit labelled with [2] is detected as outlier and the procedures skips to step c).

c) The procedure goes on in the same way as in the step b), until we find the unit labelled as $[n_O]$ that is the last unit such that $br(\hat{y}_{-[1,2,\ldots,n_O]}) > \lambda$ – meaning that $br(\hat{y}_{-[1,2,\ldots,n_O+1]}) \leq \lambda$ – so that the procedure stops with n_O outliers.

It is worthwhile to note that, as for the Hulliger's criterion, the choice of the threshold λ may be based on a calibration approach similar to the one described in section III.

There is a strict link between the Hulliger's and the bias ratio criteria. According to the step a) of the above procedure and taking account optimality of ratio estimators under model (*R*) we can write:

$$\hat{y} = \frac{y_S}{x_S} X_P, \quad \hat{y}_{-[1]} = \frac{y_{-[1]}}{x_{-[1]}} X_P \quad \text{from which we have:} \quad \left|\hat{y} - \hat{y}_{-[1]}\right| = X_P \frac{\left|y_{[1]} \sum_{S_{-[1]}} x_i - x_{[1]} \sum_{S_{-[1]}} y_i\right|}{\sum_{S_{-[1]}} x_i \sum_S x_i}. \quad (18)$$

According to the model *R*, we have also:

$$Var(\hat{y}_{-[1]}) = \sigma X_P \Big/ \sqrt{\sum_{S_{-[1]}} x_i}, \quad \text{so that:} \quad br(\hat{y}_{-[1]}) = X_P \frac{\left|y_{[1]} \sum_{S_{-[1]}} x_i - x_{[1]} \sum_{S_{-[1]}} y_i\right|}{\sigma \sqrt{\sum_{S_{-[1]}} x_i} \sum_S x_i}. \quad (19)$$

The link with the original Hulliger's criterion can be assessed if we suppose to estimate β putting: $\hat{\beta} = \sum_{S_{-[1]}} y_i \Big/ \sum_{S_{-[1]}} x_i$. We obtain: $br(\hat{y}_{-[1]}) \cong a_{[1]} \sqrt{x_{[1]} \sum_{S_{-[1]}} x_i} \left(\sigma \sum_S x_i\right)^{-1}$, from which we have that, approximately:

$$br(\hat{y}_{-[1]}) > \lambda \qquad \leftrightarrow \qquad a_{[1]} > \lambda \sigma \sum_S x_i \left(x_{[1]} \sum_{S_{-[1]}} x_i\right)^{-0,5}. \quad (20)$$

According to the original Hulliger's criterion, finally we can put:

$$c = c_{[1]} \cong \lambda \frac{\sigma}{\sigma_a} \left(\frac{\sum_S x_i}{\sqrt{x_{[1]} \sum_{S_{-[1]}} x_i}} \right) \qquad \text{with the reasonable position } \lambda\text{=0,10.} \quad (21)$$

The choice of the *c* level given by (21) depends on the particular unit labelled as [1] taken into account. The term in squared brackets is always larger than 1, because it is equal to the double of the ratio between the arithmetic mean and the geometric mean of the 2 quantities given by $x_{[1]}$ and $\sum_{S_{-[1]}} x_i$.

More generally, the estimation of the parameter *c* in the Hulliger's criterion context based on (21) implies different choices of *c* depending on the particular step of the procedure, e.g. on the particular unit concerned. Labelling with [*r*] the *r-th* step, it is easy to verify that the bias ratio threshold would be given by:

$$c_{[r]} \cong \lambda \frac{\sigma}{\sigma_a} \left(\frac{\sum_{S_{-[1,..,r-1]}} x_i}{\sqrt{x_{[r]} \sum_{S_{-[1,2,..,r]}} x_i}} \right). \quad (22)$$

The estimation of the parameter σ is also necessary in order to implement (22).

V. EMPIRICAL ATTEMPTS

A. Application to retail trade turnover data

The retail trade sample survey is carried out by ISTAT, is based on a stratified random design and is aimed at estimating monthly turnover indexes. In this context, we have supposed to focus on the

estimation of total turnover, considering the *preliminary quick sample* – available after 30 days from the end of the reference month – as the observed sample (size *n*), and the final sample observed after 52 days as the population (size *N*). This approach is justified by the random nature of quick respondents and the possibility to know the value of the true parameter, e.g. the total turnover of the final sample (option i)). A database of monthly turnover data including – on a monthly average – 1.507 enterprises has been built up, on the basis of the units always respondent in the same month of the years 2007 (*t*), 2006 (*t*-1) and 2005 (*t*-2). Domains of interest have been given by D1: *Modern food distribution* (on the average of 2007 months, *N*=326 and *n*=240), D2: *Modern non food distribution* (*N*=37, *n*=28), D3: *Small and medium food shops* (*N*=179, *n*=122), D4: *Small and medium non food shops* (*N*=965, *n*=729).

Estimation criteria have been compared in table 2. On the average, the sampling rate is equal to 74,3%. The option $w=N/n$ corresponds to the ordinary ratio estimator (*ORE*), that is the simplest tool for reducing outliers' effect (Gwet and Rivest, 1992; Gwet and Lee, 2000). Six versions of the *RRE* derive from combinations between options for weights (w_H and w^*) and α (1, 0,5, 2). All figures are averages of 2007 monthly results; levels of *c* are: $c^*(2007)$, $c^*(2006)$ and $avg[c^*(2006)]$[3], where the last option (average of 12 $c^*(2006)$) implies the use of a not seasonal steady *c* in each month of 2007. Let's note that, by definition, MAPE got using $c^*(2007)$ is not larger than MAPE obtained using the other two options[4], while we could obtain a lower MAPE using $avg[c^*(2006)]$ instead of $c^*(2006)$.

All MAPEs in bold identify case when the *RRE* improves the correspondent *ORE*. In particular: 1) that happens for all domains and several options, with the partial exception of D4; 2) the use of w^* instead of w_H is quite useful, because it always leads to lower levels of MAPE, except for D2, using $avg[c^*(2006)]$ and $c^*(2006)$ coupled with α=0,5; 3) using w_H, the option α=0,5 always improves the standard α=1, except for D2 and $avg[c^*(2006)]$, while the option α=2 is not useful, except for D1 with $c^*(2006)$ and for D2 with $avg[c^*(2006)]$; 4) using w^*, the option α=0,5 still improves the standard α=1 – with a light exception for D2 – while the option α=2 is less useful, because it reduces MAPE only for D2 and D4 using $avg[c^*(2006)]$.

On the whole, the best strategy (bold figures in boxes) is based on the use of w^* and α=0,5 with $avg[c^*(2006)]$, since the average MAPE (mean of 4 domains) would be 1,67, against the 2,03 got using the ORE.

Table 2: *Comparison among estimation strategies – Average of monthly 2007 estimates for the retail trade turnover*

Criterion	Parameter *c*				MAPE				Number of outliers			
	D1	D2	D3	D4	D1	D2	D3	D4	D1	D2	D3	D4
w=N/n	-	-	-	-	2,26	2,47	1,95	1,42	-	-	-	-
w_H and α=1	206,3	6,9	16,8	191,4	1,86	1,01	0,70	1,33	9	8	5	2
	129,3	9,5	15,1	122,0	4,53	2,50	2,29	3,39	23	6	21	66
	129,3	9,5	15,1	122,0	5,61	**2,00**	**1,73**	6,56	1	2	2	1
w^* and α=1	95,3	4,8	11,1	134,5	0,98	2,02	1,07	1,23	21	11	25	26
	14,0	3,9	5,5	27,1	**1,86**	**2,20**	**1,43**	1,68	60	12	42	198
	14,0	3,9	5,5	27,1	**1,75**	**2,22**	**1,47**	1,78	15	6	10	6
w_H and α=0,5	185,5	5,0	13,5	187,0	1,92	1,27	0,81	1,35	25	14	15	3
	113,6	7,0	12,1	115,0	4,38	**2,18**	1,97	2,60	57	9	31	67
	113,6	7,0	12,1	115,0	4,63	**2,18**	**1,46**	4,03	2	3	3	1
w^* and α=0,5	76,6	3,5	8,7	116,6	1,08	2,14	1,17	1,23	30	15	32	30
	16,0	3,1	5,1	77,0	**1,60**	**2,24**	**1,41**	1,64	62	16	57	127
	16,0	3,1	5,1	77,0	**1,57**	**2,26**	**1,43**	**1,41**	13	7	12	2
w_H and α=2	218,4	7,5	18,0	180,8	1,84	0,87	0,62	1,30	8	6	6	16
	151,8	11,1	18,2	125,9	4,38	2,63	2,49	4,68	20	5	13	66
	151,8	11,1	18,2	125,9	6,27	**1,98**	2,16	9,33	1	2	2	1
w^* and α=2	68,3	5,8	14,5	88,1	0,94	1,89	0,98	1,22	64	8	14	37
	16,8	4,4	10,6	65,4	**1,91**	**2,13**	**1,54**	1,81	31	10	19	84
	16,8	4,4	10,6	65,4	**1,87**	**2,18**	**1,48**	1,67	12	5	3	2

The 3 *c* listed are: $c^*(2007)$, $c^*(2006)$ and $avg[c^*(2006)]$. MAPE = Mean of Absolute Percent Errors.

[3] In the table $c^*(2006)=avg[c^*(2006)]$, since the reported $c^*(2006)$ are means of 12 monthly parameters.

[4] The evaluation of the MAPE got applying $c^*(2007)$ – even though not useful in practice – is helpful in order to assess the lowest limit of MAPE under a given strategy coupled with the *RRE*.

B. Application to wholesale trade turnover data

The quarterly wholesale trade sample survey carried out by ISTAT is characterised by a methodological background quite similar to the retail trade survey's one. Also in this case, we have supposed to focus on the estimation of total turnover, considering the *preliminary quick sample* – available after 60 days from the end of the reference quarter – as the observed sample (size n), and the final sample observed after 180 days as the population (size N). A database of quarterly turnover data including – on a quarterly average – 5.020 enterprises has been built up, on the basis of the units always respondent in the same quarter of the years 2007 (t), 2006 (t-1) and 2005 (t-2). Domains of interest have been given by D1: *Food products in large enterprises* (on the average of 2007 months, N=121 and n=111), D2: *Non food products in large enterprises* (N=3.070, n=2.805), D3: *Food products in small and medium enterprises* (N=594, n=492), D4: *Non food products in small and medium enterprises* (N=1.235, n=1.055). On the average, the sampling rate is equal to 88,9%, a quite higher level with respect to the retail trade context.

Estimation criteria have been compared in table 3, that keeps the same formal structure of table 2. In this case, we have the following outcomes: 1) the *RRE* can improve the *ORE* in each domain, but in a lower number of cases with respect to retail trade; 2) the use of w^* instead of w_H is quite useful, because it always leads to lower levels of MAPE, except for D2, using $avg[c^*(2006)]$ coupled with α=0,5, and D3 using α=1; 3) in the most part of cases, the use of w_H should be coupled with the standard option α=1; 4) on the other hand, the recourse to w^* leads to lower MAPEs with respect to the standard option α=1 when the alternative option α=0,5 is used, while the option α=2 quite always leads to worst results. This result is similar to that obtained in the retail trade context.

Table 3: *Comparison among estimation strategies – Average of quarterly 2007 estimates for the wholesale trade turnover*

Criterion	Parameter *c*				MAPE				Number of outliers			
	D1	D2	D3	D4	D1	D2	D3	D4	D1	D2	D3	D4
W=N/n	-	-	-	-	0,75	0,91	1,01	2,77	-	-	-	-
***w_H* and *α*=1**	18,0	52,2	20,8	52,8	0,12	0,00	0,73	2,51	3	5	30	5
	15,1	36,4	27,6	20,1	0,87	1,56	**0,92**	3,83	6	337	6	58
	15,1	36,4	27,6	20,1	1,13	**0,89**	1,05	3,96	2	6	2	18
***W** and *α*=1**	14,6	8,8	17,0	28,0	0,64	0,72	0,94	2,61	8	368	53	14
	7,8	14,0	19,7	16,9	0,75	**0,79**	**0,98**	**2,68**	15	278	7	41
	7,8	14,0	19,7	16,9	**0,74**	**0,76**	1,08	3,74	6	46	3	11
***w_H* and *α*=0,5**	16,9	33,9	18,1	47,3	0,14	0,05	0,78	2,54	4	190	53	5
	17,4	27,3	24,8	18,6	1,00	1,21	**0,97**	3,84	6	344	7	152
	17,4	27,3	24,8	18,6	0,90	**0,67**	1,07	3,47	2	10	2	9
***W** and *α*=0,5**	14,2	4,3	16,1	23,8	0,67	0,76	0,93	2,65	14	376	96	19
	9,6	17,8	9,8	13,0	0,76	**0,81**	**0,95**	**2,71**	13	395	94	178
	9,6	17,8	9,8	13,0	0,75	**0,81**	1,07	**2,70**	5	29	11	17
***w_H* and *α*=2**	18,8	46,8	18,1	57,6	0,10	0,01	0,80	2,46	2	135	52	5
	15,7	30,6	14,5	21,2	1,16	1,73	1,14	3,86	5	125	50	48
	15,7	30,6	14,5	21,2	1,31	1,17	1,71	4,47	2	8	5	6
***W** and *α*=2**	15,6	13,5	18,2	32,5	0,61	0,69	0,93	2,58	4	82	30	8
	10,8	21,7	16,0	20,0	**0,74**	**0,80**	**1,00**	**2,67**	29	219	64	28
	10,8	21,7	16,0	20,0	0,76	**0,73**	1,12	**2,67**	4	20	6	18

The 3 c listed are: $c^*(2007)$, $c^*(2006)$ and $avg[c^*(2006)]$. MAPE = Mean of Absolute Percent Errors.

On the whole, as regards wholesale trade a real best strategy (bold figures in boxes) does not exist, because one should prefer w^* for D1 and D4 and w_H for D2 and D3. Three strategies – all based on the new proposal w^* – might be preferred: w^* and α=2 (6 bold figures and 3 boxes), w^* and α=1 (5 bold figures and 1 box), w^* and α=0,5 (5 bold figures).

Finally, optimal levels of c are more steady with respect to the retail trade case (it may depend on the higher response rate), while both for retail and wholesale trade the lowest number of outliers is obtained using $avg[c^*(2006)]$.

The overall percent gain due to the use of the best *RRE* with respect to the *ORE*[5] is equal to 10,1% for wholesale, while for retail trade it is 19,7%. Since the corresponding sampling rates are, respectively, 88,9% and 74,3%, one may conclude that 14,6 percent points less in response rate correspond to a 9,6% larger gain, e.g. that 1,5 percent points less in response rate correspond to a 1% larger gain due to *RRE*.

VI. CONCLUSIONS

In this paper, we considered robust alternatives to the ordinary ratio estimator under a model assisted approach. We first defined the robustified ratio estimator, originally proposed by Hulliger in order to deal with outliers. Then we introduced some potential improvements of this estimation technique, concerning both the rule linking the original and the robust weights and the choice of the threshold beyond which a unit is detected as outlier – with the consequent reduction of its sampling weight. In particular, choice of the threshold could be driven by a *calibration* approach, that may reduce the risk of additional bias due to a too subjective choice. This approach is particularly useful when a longitudinal database of micro-data is available, as it is common in short-term business surveys as those taken into account in the empirical attempts. The two simulation studies confirmed that the new technical proposals guarantee low levels of MAPE and that the original robustified ratio estimator can be improved even in cases when response rates are enough large to contain the effect of extreme observations on the estimation error.
Future work should concern: a) the search for a quick operational algorithm able to find the optimal level of the threshold avoiding a huge number of iterations; b) the estimations of the mean squared error of the robustified ratio estimator – given the sampling design and the model – under the methodological changes herein introduced and discussed; c) the replication of simulation studies to other real populations in contexts characterised by low response rates – e.g. not larger than 50%.

REFERENCES

AA.VV. (2008), "Seminario: strategie e metodi per il controllo e la correzione dei dati nelle indagini congiunturali sulle imprese: alcune esperienze nel settore delle statistiche congiunturali", *Contributi Istat*, 13/2008, Istat, Roma.

Beaumont J.F., Alavi A. (2004), "Robust Generalized Regression Estimation", *Survey Methodology*, Vol.30, 2, 195-208.-

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

Chambers R., Kokic P., Smigh P., Cruddas M. (2000), "Winsorization for Identifying and Treating Outliers in Business Surveys", *Proceedings of the Second International Conference on Establishment Surveys*, 717-726, Alexandria, Virginia: American Statistical Association.

Cicchitelli G., Herzel A., Montanari G.E. (1992), *Il campionamento statistico*. Il Mulino, Bologna.

Elliott M.R., Little R.J.A. (2000), "Model-Based Alternatives to Trimming Survey Weights", *Journal of Official Statistics*, 16, 191-209.

Gismondi R. (2002), "Confronti tra metodi per l'identificazione di osservazioni anomale in indagini longitudinali: proposte teoriche e verifiche empiriche*", Rivista di Statistica Ufficiale*, 1, 25-60, Franco Angeli, Milano.

Gismondi R. (2008), "Reducing Revisions in Short-term Business Surveys*",* appearing on *Statistica*, CLUEB, Bologna.

Gwet J.P., Lee H. (2000), "An Evaluation of Outlier-Resistant Procedures in Establishment Surveys", *Proceedings of the Second International Conference on Establishment Surveys*, 707- 716. Alexandria, Virginia: American Statistical Association.

Gwet J.P., Rivest L.P (1992), "Outlier Resistant Alternatives to the Ratio Estimator", *Journal of the American Statistical Association*, Vol.87, 420, 1174-1182.

Hidiroglou M.A., Berthelot J.M. (1986), "Statistical Editing and Imputation for Periodic Business Surveys", *Survey Methodology*,12, 73-83

[5] It is given by 100 minus the average of the percent ratios between the bold MAPE in box and the MAPE obtained with the *ORE* for each of the 4 domains.

Hulliger B. (1995), "Outlier Robust Horvitz-Thompson Estimators", *Survey Methodology*, Vol.21, 1, 79-87.

Hulliger B. (1999), "Simple and Robust Estimators for Sampling", *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 54-63.

ISTAT, CBS, SFSO, EUROSTAT (2007), *Recommended Practices for Editing and Imputation in Cross-Sectional Business Surveys*, available on: www.edimbus.istat.it.

Kokic P.N., Bell P.A. (1994), "Optimal Winsorizing Cutoffs for a Stratified Finite Population Estimator", *Journal of Official Statistics*, 10, 419-435.

Latouche M., Berthelot J.M. (1992), "Use of a Score Function to Prioritize and Limit Recontacts in Editing Business Surveys", *Journal of Official Statistics*, 8, 389-400.

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

Ren R., Chambers R. (2002), "Outlier Robust Imputation of Survey Data via Reverse Calibration", *Southampton Statistical Sciences Research Institute Working Paper M03/19*, available on http://eprints.soton.ac.uk/8169/01/s3ri-workingpaper-m03-19.pdf.

Särndal C.E., Swensson B., Wretman J. (1993), *Model Assisted Survey Sampling*, Springer Verlag.

Searls D.T. (1966), "An Estimator for a Population Mean Which Reduces the Effect of Large True Observations", *Journal of the American Statistical Association*, 61 1200-1204.

Welsh A.H., Ronchetti E. (1998), "Bias-Calibrated Estimation from Sample Surveys Containing Outliers", *Journal of the Royal Statistical Society, Series B*, 60, 413-428.