IMPROVING ROBUST RATIO ESTIMATION IN LONGITUDINAL SURVEYS WITH OUTLIER OBSERVATIONS

Roberto Gismondi

1. 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 $y$ in a given finite population. The observed sample may include outlier observations, which are values falling in the left or right tail of the observed empirical $y$-distribution. The influence of extreme values on the overall estimation error 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.

Good outlier detection procedures should satisfy the following conditions: a) to be as much as possible time saving – especially when large data-sets are managed (Latouche and Berthelot, 1992); to bound under a reasonable level the number of sampling units detected as outliers (Gismondi, 2002); to be founded on objective rules for fixing thresholds or applying trimming (Kokic and Bell, 1994). Moreover, in the field of official statistics – where the use of standard rules as regards the main methodological issues is recommended – strategies for dealing with outliers should not be too heterogeneous, in order to guarantee a common theoretical background for outliers’ treatment. As regards structural and short-term business statistics, some late best practices are described and commented in AA.VV. (2008a; 2008b).

---

1 The opinions herein expressed must be addressed to the author only, as well as possible errors or omissions. All tables derive from elaborations on ISTAT data. A preliminary shorter version of this work is available in Gismondi et al. (2009).
More in details, while winsorisation downweights outliers substituting the anomalous values with proper estimates, re-weighting is aimed at reducing their sample weight according to some criterion: for instance, Chambers (1986) proposed an estimator that reduces to one the weights of extreme observations. Lee (1991) introduced a family of estimators that are robust to outliers under a model based approach. Croux et al. (1994) introduced a new class of regression estimators, called generalised S-estimators, which are focused on the optimal estimation of the slope parameter of a linear model and can have a 50% breakdown point like S-estimators, but attain a much higher efficiency. Hulliger (1995, 1999) analysed in depth an estimator under a model assisted survey framework (Särndal et al., 1993), based on weights for outliers that are reduced (but not necessarily equal to 1) with respect to the original ones. Re-weighting is based on a standardised function, which expresses the difference between observed and expected values. Chambers et al. (2000) re-analysed the recourse to trimming as an alternative to re-weighting, while Beaumont and Alavi (2004) focused more on the estimation process, evaluating performances of a family of robust generalised regression estimators. Late applications of robust methods for outlier detection can be found in Todorov et al. (2009).

In this context, we will deal with the Hulliger's criterion mentioned above (section 2), according to which outliers are identified and managed at the same time: i) without the need of complex elaborations and ii) applying a model-based alternative to weights' trimming\(^2\) (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 3). We also present and discuss the main outcomes of two empirical attempts based on true turnover data (section 4). Perspective conclusions have been drawn in section 5.

The main proposal consists in the choice of the acceptance threshold based on a “calibration” approach, which can be implemented using past data of the target variable that are often available in the frame of longitudinal surveys. The only basic constraint 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 based on calibration (reverse calibration); herein we develop an operational strategy not just aimed at modifying or re-estimating observed values, but at fixing an objective threshold that would have been optimal if applied to past data. Availability of time series of historical data referred 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, the basic criteria can be adapted to more general contexts as well.

\(^2\) However, it is worthwhile to note that reducing weights is equivalent to apply the original weights to trimmed values, and vice-versa.
2. The Robustified Ratio Estimator

Given a population \( P \) with size \( N \), the target is the estimation of the population total \( P_Y \) through a sample \( s \) with size \( n \) and on the basis of sampling weights \( w \).

We suppose the regression super-population model \( R \) defined as:
\[
y_i = \beta x_i + \varepsilon_i,
\]
where \( \varepsilon \) is an auxiliary variable available for each unit in the population with total \( X_p \), with \( \beta \) and \( \sigma^2 \) unknown parameters. The one-step robustified ratio estimator proposed by Hulliger is based on an estimate of the ratio between weighted medians:
\[
\hat{\beta}_0 = \frac{\hat{q}_{0.50}(y_i,w)}{\hat{q}_{0.50}(x_i,w)}
\]
and on the standardised absolute residuals
\[
\hat{a}_i = \left| y_i - \hat{\beta}_0 x_i \right| / \sqrt{x_i}.
\]
Let the median of the absolute residuals be \( \hat{\sigma}_a = \hat{q}_{0.50}(a_i,w) \). Then robust weights are defined as:
\[
w_{1H} = \frac{u_i w_i}{u_i} \text{ where: } u_i = \begin{cases} 
1 & \text{if } a_i \leq c \hat{\sigma}_a \\
\frac{c \hat{\sigma}_a}{a_i} & \text{if } a_i > c \hat{\sigma}_a
\end{cases} \text{ for each } i \in s \quad (1)
\]
where \( c \) is a parameter to be chosen. The one-step robustified ratio estimator (RRE) is:
\[
T_{1H} = \left( \sum_i w_{1H} y_i \right) \left( \sum_i w_{1H} x_i \right)^{-1} X_p = \left( \sum_i w_{1H} y_i \right) \left( \sum_i w_{1H} x_i \right)^{-1} X_p. \quad (2)
\]

The RRE is a linear estimator based on weights given by \( (X_p w_{1H}) / \sum_i w_{1H} x_i \).

It is equivalent to the ordinary ratio estimator applied to couples \((x_i,y_i)\) that, when \( a_i > c \hat{\sigma}_a \), modify into the new couples of truncated values \((x_i, \text{tr} y_i)\). It is also different with respect to the ordinary ratio estimator, e.g. the model BLU predictor (Cicchitelli et al., 1992, 385-387).

The RRE form and its performance strictly depend on 3 methodological issues:
1) the rule linking \( w_{1H} \) 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 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

---

3 A weighted median is calculated as follows: 1) order the observations \( y(1) \leq ... \leq y(n) \). 2) Let \( w(y_0) \) be the weight of \( y_0 \). The partial sums of weights of the ordered observations are defined as:
\[
k_j = \sum_{i=1}^{j} w_i / \sum_{i=1}^{n} w_i. \text{ In fact, } k_j \text{ is the estimate of the distribution function of } y \text{ at the point } y_j. \quad 3)
\]
Find the index \( j \) with \( j = \min \{ k_j > 0.5 \} \). 4) The weighted median is \( \hat{q}_{0.50}(y,w) = y(j) \). Note that the weighted median may not be expressed as a simple weighted mean. For more details, see also I-STAT et al., 2007, 58-59.
will be lower than the sum of the original ones, but that should not produce additional bias of estimates, because in the estimator (2) weights operate both at numerator and denominator.

If the corrector $u$ in (1) is quite lower than one, the number of units in the whole population which are represented by the sample outlier observation concerned will be quite lower than the number represented by the original weight $w$. In other terms, the correctors modify the formal connection between the sample and the population density distributions. On the other hand, a subjective choice of a fixed threshold parameter $\varepsilon$ may lead to wrong conclusions, especially in the frame of short-term statistics, where seasonal effects may be better managed using different parameters, depending on the month concerned and/or other stratification criteria.

A limit of the Hulliger’s criterion – as well as of many other outlier detections techniques – is that a second iteration of the procedure (with the same parameters) which uses the truncated $y$ and $x$ values in place of the original outliers ($\varepsilon\hat{\sigma}_a y_i/a_i$ and $\varepsilon\hat{\sigma}_a x_i/a_i$ respectively in place of $y_i$ and $x_i$) might generate these effects: a) these units are still detected as outliers and/or b) new outliers are found.

3. POTENTIAL IMPROVEMENTS

3.1 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 $w_{11b} \to 0$ (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 $w_{11b}$ and $w_i^*$ may be neglected only when $N$ is quite larger than $n$.

3.2 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 $a$ aimed at increasing or decreasing the quickness of the change of the original weights $w$. We still suppose that $u_i = 1$ if $a_i \leq \varepsilon\hat{\sigma}_a$. Moreover, we can put:

$$u_{ai} = (\varepsilon\hat{\sigma}_a/a_i)^a \text{ if } a_i > \varepsilon\hat{\sigma}_a. \quad(3)$$

When $a=1$, then $u_a = u_0$. When $a>1$ ($a<1$), $u_a$ tends more (less) quickly to zero than $u_0$ as well as the corresponding weights $w_a$. Since each weight expresses the number of not observed units in $P$ represented by the corresponding sample unit, the
option $a>1$ 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 $a\leq 1$, and vice-versa.

3.3 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 $(t-1)$ – whose $y$ total is known at the time $t$ 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 $Y_{P(t)}$.

b) If we suppose to apply the same sample weights at times $t$ and $(t-1)$, the RRE calculation at time $(t-1)$ is carried out trying a set of values for $c$. For each $c$ the absolute error of estimates is calculated, according to the formula:

$$AE_{(t-1)} = |T_{H(t-1)} - Y_{P(t-1)}|.$$ That can be also defined as calibration error.

c) We choose the particular optimal $c^*(t-1)$ such that: $AE_{c^*(t-1)} = \min \{AE_{c(t-1)}\}$.

d) The optimal $c^*(t-1)$ is applied 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 $c^*(t-1)$, which can be defined as “pseudo-optimal”.

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 $y(t-1)$, then:

1) at time $t$ the auxiliary variable $x$ is given by $y(t)$, while at time $(t-1)$ it is given by $y(t-2)$;

2) at time $t$ the auxiliary variable $x$ is given by $y(t-1)$, while at time $(t-1)$ it is given by $y(t-2)$.

For instance, in the frame of business surveys, if $y$ is turnover (monthly, quarterly or yearly), the option 1) can be carried out using as auxiliary variable the correspondent turnover of the previous year, while the option 2) 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 $Y_{P(t)}$: if it is not available, then the second option might be the only one useful in practice.

The calibration approach should be particularly useful if the number and the relative magnitude of outlier data as regards the $y$ and $x$ empirical distributions are
quite similar. Moreover, this approach may be used for fixing an objective threshold in the frame of other outliers’ detection methods as well.

However, the calibration approach may lead to the identification of a quite large number of outliers, due to the need of satisfying 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 minimizes the calibration error and, at the same time, guarantees that the outliers’ relative incidence is not larger than a given percent of the whole observed sample (say, 10%).

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

- minimization of the average calibration error;
- minimization of the real average estimation error calculated on previous periods;
- minimization of the variability of \( \epsilon \) estimates evaluated through a given number of attempts (for instance, different months);
- the “minimax” approach evaluated on: i) the average calibration error; ii) the number of periods for which a particular \( \epsilon \) is optimal.

One may note that the use of calibration as a robust estimation technique may be carried out without the recourse to the Hulliger’s criterion, but through the ordinary identification of the new calibration weights which minimise the squared difference with respect to the original weights and satisfy the calibration constraint (see section 4 for some empirical efficiency comparisons). If \( D_H \) is the sum of squared differences between new and original weights derived from the Hulliger’s criterion (1) when \( \epsilon \) is determined using calibration, while \( D_C \) is the analogous sum related to ordinary calibration, by definition we have \( D_C \leq D_H \).

Another operational solution derived from (1) may consists in applying the transformation of weights proposed in (1) – but for outlier units only – to all the units, so that \( w_{iH} = \frac{w_i}{w_i} \) for each unit \( i \in S \). We can indicate as \( D_{H'} \) the sum of squared differences between new and original weights derived from this criterion when \( \epsilon \) is still determined using calibration. As a consequence, we must have \( D_{H'} \leq D_H \), that is the main reason justifying this alternative approach. Of course, we still have \( D_C \leq D_{H'} \) by definition.

It is worthwhile to underline that the optimal \( \epsilon \)-level derived from a calibration approach can not be determined on the basis of an explicit formula. If we use a generic calibration variable \( x \) whose known population total is \( X_P \) and the total amount \( Y_P \) is known\(^4\), labelling as \( S_O \) and \( S_G \) the 2 sub-samples including, respectively, the outlier and the non outlier units (“good” units), the calibration approach would imply this equality:

\(^4\) A further relevant issue concerns the risk due to the use of the optimal \( \epsilon \) selected applying calibration to the variable \( x \) in place of the optimal \( \epsilon \) referred to the target variable \( y \).
Improving robust ratio estimation in longitudinal surveys with outlier observations

\[ T_h = \frac{\sum_{S_c} w_i y_i + \sum_{S_o} w_i (\epsilon_{oal} \sigma_x / a_i) y_i}{\sum_{S_c} w_i \chi_i + \sum_{S_o} w_i (\epsilon_{oal} \sigma_x / a_i) \chi_i} X_p = Y_p. \] (4)

As a consequence, the exact calibration \( (4) \) depends at the same time on the sub-sample \( S_o \) and the correspondent \( \epsilon_{oal} \). As a matter of fact, this rule is not operational, because the selection of \( S_o \) is not independent from the value of \( \epsilon_{oal} \) and vice-versa. For instance, for each sub-sample \( S_o \) the particular \( \epsilon_{oal} \) which would satisfy (4) may be such that for some units that do not belong to the sub-sample \( S_o \) we have \( a_i > \epsilon_{oal} \sigma_x \), even though they should be included in \( S_o \) since they are outliers by definition.

4. EMPIRICAL ATTEMPTS

4.1 Application to retail trade turnover data

The monthly 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. A database of monthly turnover data including – on 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 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 (\( N=965, n=729 \)).

The pseudo optimal levels of the parameter \( \epsilon \) used in order to develop the various options of the Hulliger’s criterion have been determined according to the option 1) in section 3.3 (so that \( \varepsilon=\gamma_{a,1} \)). Given the quick sample observed in a given month in 2007, in each domain we have imposed that the final sampling weights to be used for estimation are able to reproduce the final estimate available for the same month 2006. The final sampling weights are the original ones for non outlier units and have been modified according to (1) for outliers. The pseudo optimal \( \epsilon \) has been defined as \( \varepsilon^*(2006) \), while \( \epsilon^*(2007) \) is the real optimal \( \epsilon \), not known at the estimations stage (ex ante), but known ex post.

Estimation criteria have been compared in table 1. On average, the sampling rate is equal to 74.3%. According to the general estimator (2), 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 \( a \) (1, 0.5, 2). All figures are averages of 2007 monthly results; levels of \( \epsilon \)
## Comparison among estimation strategies – Average of monthly 2007 estimates for the retail trade turnover

<table>
<thead>
<tr>
<th>Parameter</th>
<th>$c$</th>
<th>MAPE</th>
<th>Number of outliers</th>
</tr>
</thead>
<tbody>
<tr>
<td>$w=N/n$</td>
<td>-</td>
<td>2.26</td>
<td>-</td>
</tr>
<tr>
<td>$w_H$ and $\sigma=1$</td>
<td>206.3</td>
<td>1.86</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>129.3</td>
<td>2.47</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>129.3</td>
<td>1.95</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>129.3</td>
<td>1.42</td>
<td>66</td>
</tr>
<tr>
<td>$w$ and $\sigma=1$</td>
<td>95.3</td>
<td>0.98</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>140.0</td>
<td>2.02</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>140.0</td>
<td>1.07</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>206.3</td>
<td>1.33</td>
<td>42</td>
</tr>
<tr>
<td>$w_H$ and $\sigma=0.5$</td>
<td>185.5</td>
<td>1.92</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>113.6</td>
<td>1.27</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>113.6</td>
<td>0.81</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>187.0</td>
<td>1.35</td>
<td>67</td>
</tr>
<tr>
<td>$w$ and $\sigma=0.5$</td>
<td>76.6</td>
<td>1.60</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>16.0</td>
<td>2.24</td>
<td>115</td>
</tr>
<tr>
<td></td>
<td>16.0</td>
<td>1.41</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td>16.0</td>
<td>1.64</td>
<td>32</td>
</tr>
<tr>
<td>$w_H$ and $\sigma=2$</td>
<td>218.4</td>
<td>1.27</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>151.8</td>
<td>1.84</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>151.8</td>
<td>0.62</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>180.8</td>
<td>1.30</td>
<td>6</td>
</tr>
<tr>
<td>$w$ and $\sigma=2$</td>
<td>68.3</td>
<td>0.94</td>
<td>64</td>
</tr>
<tr>
<td></td>
<td>16.8</td>
<td>1.89</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>16.8</td>
<td>0.98</td>
<td>37</td>
</tr>
<tr>
<td></td>
<td>88.1</td>
<td>1.22</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>151.8</td>
<td>1.38</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>151.8</td>
<td>1.41</td>
<td>66</td>
</tr>
<tr>
<td></td>
<td>118.2</td>
<td>0.87</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>125.9</td>
<td>0.62</td>
<td>2</td>
</tr>
</tbody>
</table>

The $c$ listed are: $c^*(2007)$, $c^*(2006)$ and $\text{avg}[c^*(2006)]$, 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, while we could obtain a lower MAPE using $\text{avg}[c^*(2006)]$ instead of $c^*(2006)$.

All MAPEs in bold identify cases where 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 $w_H$ instead of $w_H/\sigma=1$ is quite useful, because it always leads to lower levels of MAPE, except for D2, using $\text{avg}[c^*(2006)]$ and $c^*(2006)$ coupled with $\sigma=0.5$.
3) When $w_H$ is used, the option $\sigma=0.5$ always improves the standard $\sigma=1$, except for D2 and $\text{avg}[c^*(2006)]$, while the option $\sigma=2$ is not useful, except for D1 with $c^*(2006)$ and for D2 with $\text{avg}[c^*(2006)]$.
4) When $w_H$ is used, the option $\sigma=0.5$ still improves the standard $\sigma=1$ – with a light exception for D2 – while the option $\sigma=2$ is less useful, because it reduces MAPE only for D2 and D4 using $\text{avg}[c^*(2006)]$.

---

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

6 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.
On the whole, for each domain the best strategy (bold figures in boxes) is based on the use of \( w^* \) and \( a=0.5 \) with \( \text{avg}[c^*(2006)] \), since the average MAPE (mean of 4 domains) would be 1.67, against 2.03 got using the ORE.

4.2 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 (\( t-2 \)). In this context, domains of interest have been given by D1: Food products in large enterprises (on average of 2007 months, \( N=121, \) \( n=111 \)), 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 (\( N=1,235, \) \( n=1,055 \)). The average sampling rate is equal to 88.9%, that is quite higher with respect to the retail trade context. The selection of the pseudo optimal \( c \) parameter has been carried out on the basis of a methodology analogous to the retail trade case.

Estimation criteria have been compared in table 2, that keeps the same formal structure of table 1. In this case, we have the following outcomes:

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Parameter ( c )</th>
<th>MAPE</th>
<th>Number of outliers</th>
</tr>
</thead>
<tbody>
<tr>
<td>( w=N/n )</td>
<td>D1</td>
<td>D2</td>
<td>D3</td>
</tr>
<tr>
<td>( w=N/n ) and ( a=1 )</td>
<td>18.0</td>
<td>52.2</td>
<td>20.8</td>
</tr>
<tr>
<td>( w^* ) and ( a=1 )</td>
<td>15.1</td>
<td>36.4</td>
<td>27.6</td>
</tr>
<tr>
<td>( w^* ) and ( a=0.5 )</td>
<td>14.6</td>
<td>8.8</td>
<td>17.0</td>
</tr>
<tr>
<td>( w^* ) and ( a=0.5 )</td>
<td>7.8</td>
<td>14.0</td>
<td>19.7</td>
</tr>
<tr>
<td>( w^* ) and ( a=0.5 )</td>
<td>7.8</td>
<td>14.0</td>
<td>19.7</td>
</tr>
<tr>
<td>( w^* ) and ( a=0.5 )</td>
<td>16.9</td>
<td>33.9</td>
<td>18.1</td>
</tr>
<tr>
<td>( w^* ) and ( a=0.5 )</td>
<td>17.4</td>
<td>27.3</td>
<td>24.8</td>
</tr>
<tr>
<td>( w^* ) and ( a=0.5 )</td>
<td>17.4</td>
<td>27.3</td>
<td>24.8</td>
</tr>
<tr>
<td>( w^* ) and ( a=0 )</td>
<td>14.2</td>
<td>4.3</td>
<td>16.1</td>
</tr>
<tr>
<td>( w^* ) and ( a=0 )</td>
<td>9.6</td>
<td>17.8</td>
<td>9.8</td>
</tr>
<tr>
<td>( w^* ) and ( a=0 )</td>
<td>9.6</td>
<td>17.8</td>
<td>9.8</td>
</tr>
<tr>
<td>( w^* ) and ( a=2 )</td>
<td>18.8</td>
<td>46.8</td>
<td>18.1</td>
</tr>
<tr>
<td>( w^* ) and ( a=2 )</td>
<td>15.7</td>
<td>30.6</td>
<td>14.5</td>
</tr>
<tr>
<td>( w^* ) and ( a=2 )</td>
<td>15.7</td>
<td>30.6</td>
<td>14.5</td>
</tr>
<tr>
<td>( w^* ) and ( a=2 )</td>
<td>15.6</td>
<td>13.5</td>
<td>18.2</td>
</tr>
<tr>
<td>( w^* ) and ( a=2 )</td>
<td>10.8</td>
<td>21.7</td>
<td>16.0</td>
</tr>
<tr>
<td>( w^* ) and ( a=2 )</td>
<td>10.8</td>
<td>21.7</td>
<td>16.0</td>
</tr>
</tbody>
</table>

The 3 listed are: \( c^*(2007), c^*(2006) \) and \( \text{avg}[c^*(2006)] \). MAPE = Mean of Absolute Percent Errors. Bold: MAPEs lower than MAPE obtained with \( a=N/n \). In box: the lowest MAPE for each domain.
1) the RRE improves the ORE in each domain, but in a lower number of cases with respect to retail trade.

2) The use of \( w^* \) instead of \( w_{31} \) is quite useful, because it always leads to lower levels of MAPE, except for D2, using \( \text{avg}[c^*(2006)] \) coupled with \( a=0.5 \), and D3 using \( a=1 \).

3) In the most part of cases, the use of \( w_{31} \) should be coupled with the standard option \( a=1 \).

4) On the other hand, the recourse to \( w^* \) leads to lower MAPEs with respect to the standard option \( a=1 \) when the alternative option \( a=0.5 \) is used, while the option \( a=2 \) quite always leads to worst results. This result is similar to the one obtained in the retail trade context.

On the whole, as regards wholesale trade a real best strategy does not exist (bold figures in boxes), because one should prefer \( w^* \) for D1 and D4 and \( w_{31} \) for D2 and D3. Three strategies — all based on the new proposal \( w^* \) — might be preferred: \( w^* \) and \( a=2 \) (6 bold figures and 3 boxes), \( w^* \) and \( a=1 \) (5 bold figures and 1 box), \( w^* \) and \( a=0.5 \) (5 bold figures).

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

The overall percent gain due to the use of the best RRE with respect to the ORE \( \gamma \) 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.

Parameters \( c \) are quite unsteady depending on the option used and the domain concerned. Moreover, they are extremely heterogeneous comparing retail trade with wholesale: the only partial exception regards domain D3, for which all the \( c \) parameters range from 5.1 up to 27.6. That confirms how it may not be convenient to use a fixed level of \( c \) whatever is the month and/or the reference domain.

4.3 Non responses’ randomisation and comparison with respect to other criteria

In order to better evaluate efficiency of the various ratio estimators compared, for both retail trade and wholesale trade a lower response rate has been simulated according to a random selection of 1,000 samples in each domain. Each new simple random selection produced a new sample containing 50% of units, used as it were the real (quick) sample available for further calculations.

Moreover, two additional criteria for outlier detection have been evaluated, using both the real samples analysed in section 4.1 and 4.2 and the simulated ones mentioned above. These criteria are:

\( \gamma \) 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.
1) ordinary calibration: outlier data are not identified, while re-weighting is applied to all the sample units, imposing the same constraint used for finding the pseudo optimal \( \hat{c} \) parameters in the Hulliger’s criterion context;
2) the bias ratio criterion exposed in the Appendix, using a 10% threshold and excluding outliers found from further calculations.

We have put in direct comparisons the main outcomes derived from tables 1 and 2 and the average MAPEs derived from sample randomisation, adding results obtained using calibration and bias ratio (tables 3 and 4).

As regards retail trade (table 3), if the response rate is lower than in the real context (50%), the RRE enforces its usefulness, because: a) it always improves the ORE for domains D1, D2 and D3; b) it improves the ORE in D4 as well, using \( \omega_1 \) with \( a=1 \) or \( a=2 \). Moreover, the relative efficiency gain is larger with respect to the real context: for instance, as regards domain D1 the average MAPE decreases from 2.49 (ORE) to 0.64 (RRE), while using the real quick response rates we pass from 2.26 to 1.60. A similar result occurs for D3: MAPE lessens from 2.13 to 1.07, much more than when the real quick responses are used (from 1.95 to 1.41).

Broadly speaking, taking into account results obtained with both real and randomised samples, we can conclude that \( w^* \) should be preferred to \( w_H \) and that it is more realistic to suppose \( a=1 \), even though in the randomised context the option \( a=2 \) should be preferred.

As regards calibration and bias ratio, the former leads to better results in the real contexts, while the latter should be preferred in the randomized frame. In particular, bias ratio is the only criterion able to lessen significantly MAPE as regards D2, since it passes from the original 4.21 referred to the ORE to 1.45. In all the other cases, both calibration and bias ratio can be improved by the RRE coupled with a proper choice of \( \hat{c} \).

**TABLE 3**

Comparison between results obtained using the real response rates and simulated 50% response rates (1,000 random replications\(^8\)); results derived from calibration and bias ratio criteria (retail trade: average of 2007 estimates)

<table>
<thead>
<tr>
<th>Criterion</th>
<th>MAPE - Real response rates</th>
<th>MAPE - 1,000 random replications</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>D1</td>
<td>D2</td>
</tr>
<tr>
<td>( \omega=N/n )</td>
<td>2.26</td>
<td>2.47</td>
</tr>
<tr>
<td>( w_H ) and ( \omega=1 ) (1)</td>
<td>4.53</td>
<td>2.50</td>
</tr>
<tr>
<td>( w^* ) and ( \omega=1 )</td>
<td>1.86</td>
<td>2.20</td>
</tr>
<tr>
<td>( w_H ) and ( \omega=0.5 )</td>
<td>4.38</td>
<td>2.18</td>
</tr>
<tr>
<td>( w^* ) and ( \omega=0.5 )</td>
<td>1.60</td>
<td>2.24</td>
</tr>
<tr>
<td>( w_H ) and ( \omega=2 )</td>
<td>4.38</td>
<td>2.63</td>
</tr>
<tr>
<td>( w^* ) and ( \omega=2 )</td>
<td>1.91</td>
<td>2.13</td>
</tr>
<tr>
<td>Calibration (2)</td>
<td>2.22</td>
<td>2.35</td>
</tr>
<tr>
<td>Bias ratio (2)</td>
<td>2.89</td>
<td>3.55</td>
</tr>
</tbody>
</table>

(1) Results of the Hulliger’s criterion are based on \( \hat{c}^*(2006) \). MAPE = Mean of Absolute Percent Errors.
(2) As exposed in the appendix. Without imputation: outliers are excluded from calculations.

\( \hat{c} \): Bold: MAPEs lower than MAPE obtained with \( \omega=N/n \). In box: the lowest MAPE for each domain.

\( \omega \): Each MAPE is the mean of 12,000 estimates for retail trade (1,000x12 monthly MAPEs) and of 4,000 estimates for wholesale trade (1,000x4 quarterly MAPEs).
As regards wholesale trade (table 4), conclusions similar to retail trade can be drawn. The RRE still improves the ORE for domains D1 and D2, performs better than for retail trade in domain D4 (it does not improve the ORE only using \( w^* \) and \( \alpha=0.5 \)) and improves the ORE in D3 as well, using \( w^* \) with \( \alpha=1 \) or \( \alpha=0.5 \). With respect to retail trade, the relative efficiency gain is lower, even though it is quite large in D3 (MAPE passes from 1.67 to 0.38). Moreover, on average \( w^* \) should be preferred to \( w_H \) at least when \( \alpha=1 \) or \( \alpha=0.5 \), while when \( \alpha=2 \) both options improve the ORE and, in particular, \( w_H \) leads to the lowest MAPE for D1 (0.78). As regards calibration and bias ratio, the former leads to the lowest MAPE for D4 (in particular, in the randomized context MAPE is only 0.09, against the original 3.21 obtained using the ORE), while the latter is optimal in D1 using real response rates (MAPE=0.21).

According to these outcomes, a caution strategy may be based on the joint use of different criteria for detecting and treating outliers, depending on the particular domain under study. However, empirical attempts confirm that the recourse to a RRE strategy would often be the best solution, even though, in some particular cases, other criteria may improve its efficiency.

A more synthetic resume of results has been reported in table 5, which contains, for each criterion, the overall average MAPE – calculated as mean of monthly or quarterly MAPEs in each domain and for each sample replication in the randomised context – separately for retail trade and wholesale trade.

The best strategy is the one based on the RRE with \( w^* \) and \( \alpha=0.5 \): as a matter of fact, this strategy leads to the lowest average MAPE for retail trade using real response rates (MAPE=1.72) and for wholesale trade in the randomised context (MAPE=1.61). Moreover, that improves the ORE also for retail trade in the randomised context (MAPE=2.47 against 2.76 obtained with the ORE) and for wholesale trade using real response rates (MAPE=1.31 against 1.36 obtained with the ORE). However, in the last 2 cases the best criterion is the bias ratio, whose only bad performance concerns retail trade using real response rates (its MAPE is equal to 2.95, quite larger with respect to 2.03 obtained with the ORE). Finally, the RRE with \( w^* \) and \( \alpha=1 \) is a good strategy as well, because it leads to 3 second best performances and it always improves the ORE.

<table>
<thead>
<tr>
<th>Criterion</th>
<th>MAPE – Real response rates</th>
<th>MAPE – 1,000 random replications</th>
</tr>
</thead>
<tbody>
<tr>
<td>( w=N/n )</td>
<td>D1</td>
<td>D2</td>
</tr>
<tr>
<td>( w ) and ( \alpha=1 )</td>
<td>0.75</td>
<td>0.91</td>
</tr>
<tr>
<td>( w^* ) and ( \alpha=0.5 )</td>
<td>0.75</td>
<td>0.79</td>
</tr>
<tr>
<td>Calibration</td>
<td>0.76</td>
<td>1.42</td>
</tr>
</tbody>
</table>

(1) Results of the Hulliger’s criterion are based on \( c^*(2006) \). MAPE = Mean of Absolute Percent Errors.

(2) As exposed in the appendix. Without imputation: outliers are excluded from calculations.

Bold: MAPEs lower than MAPE obtained with \( w=N/n \). In box: the lowest MAPE for each domain.
Table 5
Final comparison among criteria based on the average MAPE in 4 domains: D1, D2, D3, D4 (real response rates and simulated 50% response rates for retail trade and wholesale trade: average of 2007 estimates)

<table>
<thead>
<tr>
<th>Criterion</th>
<th>MAPE – Real response rates</th>
<th>MAPE – 1,000 random replications</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Retail trade</td>
<td>Wholesale trade</td>
</tr>
<tr>
<td>$w = N/n$</td>
<td>2.03</td>
<td>1.36</td>
</tr>
<tr>
<td>$w^*$ and $a=1$ (1)</td>
<td>3.18</td>
<td>1.80</td>
</tr>
<tr>
<td>$w^*$ and $a=0.5$</td>
<td>2.78</td>
<td>1.76</td>
</tr>
<tr>
<td>$w^*$ and $a=0.5$</td>
<td>1.72</td>
<td>1.31</td>
</tr>
<tr>
<td>$w^*$ and $a=2$</td>
<td>3.55</td>
<td>1.97</td>
</tr>
<tr>
<td>$w^*$ and $a=2$</td>
<td>1.85</td>
<td>1.39</td>
</tr>
<tr>
<td>Calibration</td>
<td>2.26</td>
<td>1.46</td>
</tr>
<tr>
<td>Bias ratio (2)</td>
<td>2.95</td>
<td>1.24</td>
</tr>
</tbody>
</table>

(1) Results of the Hulliger’s criterion are based on $\mu(2006)$. MAPE = Mean of Absolute Percent Errors.
(2) As exposed in the appendix. Without imputation: outliers are excluded from calculations.
Bold: the lowest (best) MAPE. Underlined: the second best MAPE.

5. CONCLUSIONS

In this framework, robust alternatives to the ratio estimator in order to deal with outliers under a model assisted approach have been evaluated. We first defined the robustified ratio estimator. Then we introduced some potential improvements, 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.

One of the most relevant features of the RRE technique is that it guarantees a direct link between the preliminary treatment of micro-data and the original weighting system derived from the sampling design and/or the model. This implicit property should preserve from the risk of very biased estimates due to inconsistency between the logic underlying treatment of outlier data and the final estimation process.

The empirical attempts based on real data confirmed that, in the most part of case studies, the new technical proposals guarantee: i) low levels of MAPE and, in particular, a generalised reduction of MAPE with respect to the ordinary ratio estimator; ii) better performances with respect to the use of ordinary calibration applied to all the available data (without any preliminary treatment of outliers); iii) on average, performances better or at most lightly worst with respect to those obtained using the bias ratio criterion. The original robustified ratio estimator can be improved both in cases when response rates are large enough to contain the effect of extreme observations on the estimation error (as in the two empirical attempts based on real response rates) and when lower response rates occur (as in the randomised simulation).
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 estimation 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 char-
acterised by low response rates – e.g. not larger than 50%;
d) The choice of different thresholds in a multivariate context.

6. APPENDIX: THE BIAS RATIO CRITERION FOR OUTLIERS’ DETECTION

The bias ratio criterion derives from an adaptation of the classical theory of
confidence intervals. If \( \hat{y} \) is an estimator of the population total \( Y_p \) 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 (br) of this estimate is:

\[
br(\hat{y}_{-A}) = \frac{|Y_p - \hat{y}_{-A}|}{\sqrt{\text{Var}(\hat{y}_{-A})}}.
\]

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:

\[
\Pr[-z_{1-\alpha/2} - br(\hat{y}_{-A}) < Z < z_{1-\alpha/2} - br(\hat{y}_{-A})],
\]

where \( z_{(1-\alpha/2)} \) is the percentile of the standardised normal cumulated distribution leaving on the right tail a prob-
ability equal to \( \alpha/2 \). The coverage probability equals the nominal, desired confi-
dence 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 (5) as regards the outliers problem
consists in testing the significance of the difference between the \( Y_p \)-estimates ba-
sed on the complete data set and the data set which does not include a certain
sub-set of units. On the basis of a slight adaptation of (5), 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 the following algorithm:
a) for each unit \( i \in S \) we evaluate the approximate bias ratio \( br \):

\[
br(\hat{y}_{-i}) = \left| \hat{y} - \hat{y}_{-i} \right| \left( \sqrt{\text{Var}(\hat{y}_{-i})} \right)^{-0.5}
\]
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 $br(\hat{y}_{[-1]}) \leq \lambda$ – where $\lambda$ 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:

$$br(\hat{y}_{[-1,2]}) = \left| Y_p - \hat{y}_{[-1,2]} \right| \left[ \text{Var}(\hat{y}_{[-1,2]}) \right]^{-0.5}$$

(7)

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

c) The procedure goes on as in the step b), until we find the unit labelled as $[n_O]$ which 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 $\lambda$ may be based on a calibration approach similar to the one described in section 3.3.

ISTAT, Italian National Statistical Institute

ROBERTO GISMONDI

REFERENCES


SUMMARY

Improving robust ratio estimation in longitudinal surveys with outlier observations

The Hulliger's robust estimation technique consists in the re-weighting of units identified as outliers through a Robustified Ratio Estimator (RRE), according to which outliers...
contribute to the final estimate with a sample weight reduced with respect to the original one. Outlier observations are identified through a standardised function founded on the difference between observed and expected values. A crucial aspect concerns the choice of the acceptance threshold, which plays a role in the re-weighting process as well. In this context, we propose some potential improvements of the $RRE$, concerning the use of an objective criterion for fixing the threshold and the re-weighting rules. Results of two empirical attempts based on real data derived from longitudinal surveys show that, in the most part of case studies, the proposed changes contribute to improve efficiency of estimates with respect to the ordinary ratio estimator.