Sampling Weights Adjustment for Improving Crops Early Estimates' Precision

Roberto Gismondi
ISTAT (Italian National Statistical Institute), Roma, Italy - gismondi@istat.it

Loredana De Gaetano
ISTAT (Italian National Statistical Institute), Roma, Italy - degatean@istat.it

The Robustified Ratio Estimator (*RRE*) technique consists in re-weighting sample units identified as outliers through the calculation of standardized differences between observed and expected values. The original sample weights are reduced on the basis of the magnitude of residuals and are used for implementing robust ratio estimation, which is a time saving approach reducing the risk of anomalous estimates. The *RRE* can be also applied when measurement errors occur because of response errors. In the *RRE* process, a crucial phase concerns the choice of the acceptation threshold beyond which a measurement is detected as outlier. In this context, we propose improvements of the *RRE*, based on: i) the use of weights not lower than one; ii) transformation of the original variables in order to tackle the presence of zero values and guarantee the possibility to calculate residuals; iii) an objective criterion for fixing the acceptation threshold and the re-weighting rule. Results of an empirical attempt referred to early estimates of agricultural crops in Italy show positive effects of estimates precision, on the basis of strategies which combine in different ways the improvements. The non response bias has been also estimated for any estimation strategy implemented, in addition to sampling variance.

Keywords: Agriculture, Bias, Calibration, Crop, Outlier, Robust ratio, Sampling, Weighting.

1. Introduction

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 anomalous observations, which are values falling in the left or right tail of the observed empirical *y*-distribution. These observations may be outliers of wrong data derived from measurement errors. As a consequence, the following problems must be faced:

a) how to identify anomalous observations;
b) how to treat them after the identification, according to one (or a combination) of these criteria:
   - anomalous micro-data are re-estimated as they were missing observations,
   - anomalous micro-data are not changed, but their sampling weight is reduced.

From now on the term *outlier* and *anomalous* will be supposed equivalent. Good outlier detection procedures should satisfy the following conditions: i) to be as much as possible time saving, especially when large data-sets are managed and the time lag between the end of data capturing and the release phases is very short; ii) to avoid the risk of massive re-imputation of micro-data supposed to be wrong, due to the risk to overlap the original data with subjective estimations; iii) to be founded on objective rules for defining acceptation thresholds and for modifying original data and/or sampling weights.

Re-weighting is a process aimed at reducing outliers sample weight according to some criterion. In particular, Hulliger (1995) analysed in depth an estimator under a model assisted survey framework, based on weights for outliers that are reduced with respect to the original ones: re-weighting is based on a standardised function, which expresses the difference between observed and expected values. In this context, we will deal with the Hulliger's criterion (section 2), according to which outliers are identified and managed without the need of complex elaborations. 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 discuss the main outcomes of an empirical attempt concerning early estimates of agriculture crop production in Italy, characterized by many null observations which affect calculability of residuals (section 4). The estimation of bias due to non responses is also provided. Perspective conclusions have been drawn in section 5.

2. The robustified ratio estimator: basic features

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 super-population model $\mathcal{M}$ 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) and ($i\neq j$), 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 is based on the ratio between weighted medians β_0 and on the standardized absolute residual a_i defined as, respectively:

$$\beta_0 = q_{0.50}(y_i, w)/q_{0.50}(x_i, w) \qquad\qquad a_i = \left|y_i - \hat{\beta}_0 x_i\right| \Big/ \sqrt{x_i}\,. \tag{1}$$

Let the median of the absolute residuals be $\hat{\sigma}_a = q_{0.50}(a_i, w)$. Then *robust weights* are defined as:

$$w_{*i} = u_i w_i \quad \text{where:} \quad u_i = \begin{cases} 1 & if \quad a_i \le c\,\hat{\sigma}_a \\ c\,\hat{\sigma}_a / a_i & if \quad a_i > c\,\hat{\sigma}_a \end{cases} \quad \text{for each } i \in s \tag{2}$$

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

$$T_{RRE} = \left(\sum\nolimits_s w_i u_i y_i\right)\left(\sum\nolimits_s w_i u_i x_i\right)^{-1} X_P = \left(\sum\nolimits_s w_{*i}\, y_i\right)\left(\sum\nolimits_s w_{*i}\, x_i\right)^{-1} X_P\,. \tag{3}$$

The *RRE* is a linear estimator based on weights given by $(X_P w_{*i}) \Big/ \sum_s w_{*i} x_i$. 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). The re-weighting system (2) can be viewed as a robust estimation criterion that reduces the outliers' weight according to the standardized distance between the observed and the theoretical y-value. A major advantage due to (2) 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 should not produce additional bias of estimates, because in the estimator (3) weights operate both at numerator and denominator. If the corrector u in (2) 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. The *RRE* method can be used with or without a preliminary micro-data editing process. Let's also note that, in a multivariate context, different sampling weights will be assigned to the same unit.

3. Improvements of the method

The *RRE* method can be improved. Some late proposals (Gismondi, 2010) analyzed the function linking w_* to w, the definition of correctors u in (2) and the choice of c in (2). In this context, the attention focuses on some additional issues and proposes an alternative selection of parameter c.

As first remark, since according to formula (2) the new weights w_* may be lower than one, the additional condition $w_*\geq 1$ can be imposed, because each unit must represent itself at least and weights lower than zero would not have a statistical meaning. This constraint will lead to higher estimates.

Moreover, we observe that the basic conditions on which both formulas (1) are founded: a) $x_i>0$ for any unit i; b) the median of weighted x-values is different than zero, may not be always satisfied in current surveys practice. A broad family of variables which do not satisfy one or both of the previous conditions concern statistical phenomena which may be present or not among the units surveyed. Some examples concern x variables as enterprises' job vacancies or the use of specific drugs, for which both constraints a) and b) may not be satisfied. Furthermore, null values are often observed in the frame of agricultural surveys, as that discussed in section 4. In order to tackle cases when $x=0$, the new variables $y+\theta$ and $x+\theta$ may be used in order to implement the steps (1) and (2), where θ is a small constant defined *a priori* or on the basis of available data. Under the model $\mathcal{M}$, the use of $y'=y+\theta$ and $x'=x+\theta$ in place of y and x implies the new data model $y_i'=\alpha+\beta x_i'+\varepsilon_i$ with $\alpha=\theta(1\text{-}\beta)$. The weighted median β_0 can be calculated using the y' and x' distributions as well. In practice, θ may be put equal to specific percentiles or deciles of the x empirical distribution. However, more driven options can be used in order to tackle more general cases when $q_{0.50}(x_i,w)$=0 and/or $q_{0.50}(y_i,w)$=0. We observe that:

$$\beta_0' = \frac{q_{0.50}(y_i',w)}{q_{0.50}(x_i',w)} = \frac{q_{0.50}(y_i,w)+\theta}{q_{0.50}(x_i,w)+\theta}\,. \qquad (4)$$

As a consequence, if $q_{0.50}(x_i,w)$=0 and $q_{0.50}(y_i,w)$=0 the estimate (4) is equal to one for any θ; moreover, $\beta_0' \to 1$(e.g., the y expected values are supposed to be very similar to the correspondent x values) in the following cases:

a) when $q_{0.50}(x_i,w)\neq 0$ and $q_{0.50}(y_i,w)$=0: if θ is quite larger than $q_{0.50}(x_i,w)$; (5)
b) when $q_{0.50}(x_i,w)$=0 and $q_{0.50}(y_i,w)\neq 0$: if θ is quite larger than $q_{0.50}(y_i,w)$.

Caution options are:

A) as the case a) above, putting θ=$q_{0.50}(x_i,w)$, so that: $\beta_0' = 0.5$; (6)

B) as the case b) above, putting θ=$q_{0.50}(y_i,w)$, so that: $\beta_0' = 2$.

As regards the choice of parameter c in (2), a “pseudo-calibration” approach may improve *RRE* efficiency, reducing the risk of subjective choices. The procedure is based on these steps:

1) We suppose to know y and x values of each sample unit, as well as to know the x total X_P. The x values are also supposed to have been already checked and edited.
2) The same sampling weights may be applied to y and x micro-data. Let’s note that X_P may be itself an estimate derived from the sampling process or from an external source: the basic constraint is that its estimate error is supposed to be quite low. If the estimate is available, we suppose to have obtained the estimate using the ordinary Horvitz-Thompson estimator, without any *RRE* process.
3) The *RRE* process is applied to the available x micro-data: among the various c values attempted, the particular $c_*(x)$ which minimizes the difference between X_P and the estimate $T_{RRE}(x)$ is selected.
4) The pseudo optimal value c to be used for implementing the *RRE* process for estimating Y_P can be put equal to $c_*(x)$, bearing in mind that the “true” optimal $c_*(y)$ is unknown. The true optimal value may be known when the true level Y_P will be available.

The procedure – inspired by the traditional 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 guarantees a near-calibration of sample estimates with respect to the x population total should work fine as regards the target y variable as well. However, even though this approach should guarantee coherence between x and y totals estimates, since x micro-data have been supposed to be not affected by relevant measurement errors, it may happen that $c_*(x) > c_*(y)$. A further step can be carried out in a multivariate context, where there are k target variables $y_1,\ldots,y_k$. If for at least one of them – say y_1 – the “true” total Y_{1P} is available, then $c_*(y_1)$ is known and we can suppose that:

$$\hat{c}_*(y_r) = c_*(x_r)\left[c_*(y_1)/c_*(x_1)\right] \qquad \text{for } r=2,3,\ldots,k. \qquad (7)$$

A very common empirical context when the pseudo-calibration procedure can be used is when x is given by the y variable observed at a previous time, such as in longitudinal surveys aimed at estimating changes of the target variable, as it happens in the case study commented in section 4.

4. Case study: crops early estimates in Italy

The sampling survey “Crop early estimates” has the goal of producing anticipated estimates as regards the main agriculture crops. The last survey wave has been carried out between November 2014 and January 2015 through the CATI technique and was aimed at interviewing a sample of holders for collecting early estimates regarding area use for agricultural purposes in the agrarian year (*ay*) 2014-15. Since the main survey target is to produce estimates of changes occurred between two following years, information on agricultural land use in the *ay* 2013-14 has been asked as well. The reference population is given by the agricultural holdings which had arable lands according to the 2010 agriculture census (586,722 units). The random sample is composed by 13,575 holdings, stratified within 120 strata obtained crossing 4 geographic area (North-West, North-East, Centre, South-Islands) and surface classes referred to the main 5 crops categories: cereals, legumes, industrial plants, vegetables, others[1]. The final respondents sample size (7,898) derives from the exclusion of: a) not responding holdings; b) units momentary not active; c) units which stopped the activity. In the survey framework, y is given by the *forecasted* size of agricultural land to be used for a specific crops in the *ay* 2014-15, while x is the correspondent *known* size in the previous *ay*. Sample estimates have been calculated inside each stratum; estimates for Italy as a whole have been obtained through weighted arithmetic mean of strata estimates. Original sampling weights have been changed twice: 1) for taking into account non responses; 2) as result of calibration (Lundström and Särndal, 1999). The (known) population totals used as calibration constraints are surfaces for the main crops (durum wheat, common wheat, grain maize, soya) broken down by geographic areas, derived from the agricultural census. The survey outcomes are the percent changes of land use between the two agrarian years.

A crucial aspect consists in the data editing and imputation process carried out before calibration. Its complexity depends on micro-data quality and may take a long time, without being sure that edited data will lead to better estimates The table 1 summarizes the situation observed before running data editing. As regards UAA, only the 15.7% of percent changes 2015/2014 was not larger than 25%, while the 33.7% of changes was larger than 100%. Situation is even more problematic for specific land uses; moreover, the presence of zero values in 2014 (variable x) affects legumes (93.2%), industrial plants (82.7%) and vegetables (90.1%) dramatically. Overall, it is not realistic to foresee that in 2015 about the 75% of agricultural holdings will increase agricultural land use more than 50% (as it would derive according to the column *Total*). Micro-data seem affected by many measurement errors concerning the y variable (foreseen land use in the *ay* 2014-15). They may be due to: i) the questionnaire complexity, since the survey questionnaire asked for 32 specific kinds of crops, then summarized in 5 broad categories; ii) the need to select the sample from the list derived from the last agriculture census, which is old and not updated; iii) the insufficient skill of some interviewers.

Table 1: *% changes 2015/2014 calculated on raw micro-data*

Class of change %	UAA	Arable land	Cereals	Legumes	Industrial	Vegetables
0%	5.1	5.0	2.1	0.1	0.6	0.2
0.01% - 25.0%	10.6	10.0	5.7	0.1	0.9	0.3
25.1% - 50.0%	11.3	10.7	6.5	0.1	0.6	0.1
50.1% - 75.0%	14.9	13.1	7.2	0.1	0.6	0.2
75.1% - 100.0%	24.4	24.7	32.0	6.2	13.6	8.6
>100.0%	33.7	33.3	16.9	0.2	1.2	0.4
Surface 2014 = 0	0.0	3.3	29.5	93.2	82.7	90.1
Total cases	100.0	100.0	100.0	100.0	100.0	100.0

Source: elaboration on ISTAT data. Total arable land is the sum of cereals, pulses, industrial plants, vegetables, other crops.

The tables 2 summarizes the main results obtained using different estimation strategies. The first row shows land use estimation in the *ay* 2013-14 (thousand hectares), as obtained applying the ordinary Horvitz-Thompson estimator (equivalent to *REE* when u=1) and weights adjusted for non responses. Macro-estimates of land use in the *ay* 2014-15 based on the *HT* estimator applied to raw micro-data are very anomalous (second row). Afterwards, the *RRE* strategy based on formulas (1), (2) and (3) has been applied combining various options among those discussed in section 3. Beyond the ordinary *RRE* with c=1 (according to which no *RRE* may be applied for legumes, industrial plants and vegetables because of x zero values), the choice of c using the pseudo-calibration approach, the selection of θ

[1] The sum of the 5 categories leads to the total "Arable land", while the "Utilized Agricultural Area" (UAA) is the sum of arable land, permanent crops and pasture.

according to formula (6) and the use of weights not lower than one have been introduced and combined each other. The compared *RRE* techniques – selected among the whole set of combinations attempted – lead to reasonable estimates if compared to 2014 data, even though the ordinary *REE* (c=1) leads to suspect estimates of UAA in 2015; they are not particularly different from those obtained applying calibration after the long and costly data editing process carried out for reducing the effect of wrong micro-data. This outcome implies the possibility to avoid data editing, using *RRE* with the proper combination of suggested improvements, even though it is difficult to choose the best strategy among those compared without introducing any error estimation.

Table 2: *Main results of compared estimation strategies*

Method	Year	UAA	Arable land	Cereals	Legumes	Industrial	Vegetables
2014 micro-data, *HT* estimation (*RRE*: u=1)	2014	9,795	6,082	3,055	117	334	207
Raw micro-data, *HT* estimation (*RRE*: u=1)	2015	>20,000	>20,000	9,922	512	1,328	544
RRE (c=1)	2015	8,405	7,337	3,096	-	-	-
RRE (c=1, θ: rule (6))	2015	10,010	6,372	3,049	117	305	148
RRE (c=1, θ: rule (6), w≥1)	2015	10,215	7,020	3,066	119	306	148
RRE ($c=c_*(x)$, θ: rule (6))	2015	10,857	6,871	3,161	310	416	98
RRE ($c= c_*(x)$, θ: rule (6), w≥1)	2015	11,894	7,901	3,164	310	417	103
Calibration	2015	10,712	6,896	3,146	239	440	188
Agriculture census data	2010	12,856	7,009	3,619	139	343	300

Source: elaboration on ISTAT data. Surfaces in thousand hectares.

In order to achieve to an objective estimate of estimation strategies precision, their Mean Squared Error (*MSE*) has been calculated. The sampling variances have been estimated using the ISTAT software GENESEES[2] (ISTAT, 2004), which runs estimation for calibration estimators and has been adapted in order to estimate sampling variances concerning the *RRE* methods as well. Moreover, bias due to the non response process has been estimated as well (Bethlehem, 2010). Broadly speaking, for each variable y whose total is object of estimates, if h is the stratum label, H is the number of strata, n_{hR} is the number of respondent units for the particular variable concerned in the stratum h, Y is the correspondent variable measured in the last agriculture census (available for both respondent and not respondent units), $\overline{Y}_{hR}$ and $\overline{Y}_{h\bar{R}}$ are, respectively, the estimated population means of respondent and not respondent units in the stratum h, then the stratum and the global non response biases have been estimated with the respective formulas:

$$Bias_h \approx \left(1-\frac{n_{hR}}{n_h}\right)\left(\overline{Y}_{hR}-\overline{Y}_{h\bar{R}}\right) \quad \text{for } h=1,\ldots,H \qquad\qquad Bias \approx \sum_{h=1}^{H}\left(\frac{N_h}{N}\right)Bias_h . \qquad (8)$$

Afterwards, the estimated Mean Squared Error (*MSE*) will be given by the sum of sampling variance and the squared bias in (8). The relative estimation error has been calculated through the Coefficient of variation (*Cv*) given by: $Cv = 100\sqrt{M\hat{S}E}\Big/\hat{T}$, where $\hat{T}$ is the agricultural area estimate.

The table 3 shows that estimated precision of estimates obtained using the improved *RRE* strategy is comparable with the estimated precision concerning the ordinary calibration approach; all the *REE* strategies pay a precision gap with respect to calibration, which however is relatively small. Moreover, the last strategy – based on the options $c=c_*(x)$, θ selected using rule (6) and sampling weights ≥ 1) – has the same estimated precision than calibration. The parallel gain in timeliness due to elaborations and data manipulations before estimations suggests to reply this comparative exercise next year and as regards other sampling survey contexts, even outside the agriculture statistics framework. We must

[2] The software *GENESEES* (GENEralised software for Sampling Estimates and Errors in Surveys) has been used for calculating calibrated estimates as well.

remind that the above estimated error is *undervalued* with respect to the true unknown error, because other error components cannot be evaluated. Beyond the *list error*, the *response error* – as it is normally defined the overall effect of unit measurement errors – should be evaluated as well, but it requires knowledge of true micro-data, which actually are not known yet. The overall *MSE* of estimates may be obtained adding to the ordinary sampling variance and to the bias estimated through the formula (8): a) the response variance; b) the covariance between sampling error and response error; c) the squared expected difference between the estimator and the true unknown population total.

Table 3: *Cv estimation of 2015 crops estimates using different strategies*

Method	**UAA**	**Arable land**	**Cereals**	**Legumes**	**Industrial**	**Vegetables**	**Avg1**	**Avg2**
Raw micro-data (*RRE*: *u*=1)	8.6	45.8	>50	>50	>50	>50	>50	39.5
RRE (*c*=1)	5.2	8.9	13.6	-	-	-	-	9.2
RRE (*c*=1, θ: rule (6))	5.4	9.3	11.6	25.4	11.5	21.7	14.2	8.8
RRE (*c*=1, θ: rule (6), w≥1)	4.3	8.2	15.7	25.4	11.5	21.7	14.5	9.4
RRE ($c=c_*(x)$, θ: rule (6))	3.7	8.3	11.2	9.5	8.4	32.6	12.3	7.8
RRE ($c= c_*(x)$, θ: rule (6), w≥1)	3.8	7.8	11.2	9.5	8.4	31.2	12.0	7.6
Calibration	3,5	8.3	10.7	9.1	7.9	32.4	12.0	7.5

Source: elaboration on ISTAT data. Avg1: mean of all 6 estimates; Avg2: mean of the first 3 estimates.

5. Conclusions

One of the most relevant features of the *RRE* process is the improvement of the ordinary ratio estimation, obtained through manipulation of sampling weights on the basis of the difference between observed and expected values. The *REE* should also preserve from the risk of biased estimates due to inconsistency between the logic underlying treatment of anomalous micro-data and the final estimation process. Finally, the process is time saving with respect to ordinary data editing process often carried out before running the final estimation phase.

In this framework, we introduced potential improvements of the *RRE*, concerning the extension of its usability to variables for which null values may happen and the choice of the threshold beyond which a unit is detected as anomalous. In particular, selection of the threshold can be driven by a *pseudo-calibration* approach. The empirical attempt concerned the crops' early estimates survey and showed how the use of the improved *REE*, *without any preliminary data editing process*, would lead to estimates characterized by mean squared errors very near to that obtained through the calibration approach currently used in the survey framework, after a complex and long data editing process. Future work should concern:

a) the combination between the improved *RRE* process and a selective editing approach, which limits micro-data corrections to the most influent sample units;

b) the replication of simulation studies to other survey contexts, in order to assess robustness of results achieved as regards the crops' early estimates survey.

References

Bethlehem, J. (2010). Selection Bias in Web Surveys. *International Statistical Review*, 78, 2, 161–188.

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

Gismondi, R. (2010). Improving Robust Ratio Estimation in Longitudinal Surveys with Outlier Observations. *Statistica*, LXX, 1, 23-39.

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

ISTAT (2005). *GENESEES V. 3.0 – Funzione stime ed errori*. Istat, Roma.

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