Adaptation of Winsorization caused by weight share method

Keywords: *Business Statistics, Winsorization, Weight share method, profiling.*

1 Introduction

The French Structural Business Statistics (SBS) production system, known as ESANE, has two main uses:

- production of statistics based on the European SBS regulation;
- estimation of businesses' contributions to GDP for the national accounts.

The system is based on a mix of exhaustive administrative fiscal data and data obtained on a random sample of the business population [1]. ESANE is currently changing to produce estimates based on profiled units or enterprises[1] and no longer on legal units.
Several methodological studies have been conducted to support this change, and the following study concerns the adaptation of the treatment of influential values by winsorization for the dissemination of results at the enterprise level.

1.1 New sampling design

Starting at reference year 2016, the sampling design of SBS surveys selects enterprises [3]. When an enterprise is selected, then all legal units within this enterprise will be surveyed. The entreprise's answer will be built based on the legal units answers.

1.2 Update of the delineation of enterprises

For reference year T, the samples are drawn in November T with links between legal units and enterprises referring to year T-2, the most recent available at this date. Few months later, new links referring to year T-1, so more up-to-date, are available. These new links are used to produce the results at the enterprise level concerning year T thanks to the generalized weight share method (GWSM)[4].

1.3 Influential values treatment

Between the reference years 2008 and 2015, the influential values were treated by winsorization of the variable turnover with Kokic and Bell's thresholds [5]. The Kokic and Bell's thresholds minimize the mean square error of the estimator of the total of the winsorized variable.

[1] An enterprise is defined by law as the smallest combination of legal units that is an organisational unit producing goods or services with a certain degree of autonomy [2]

1.4 Aim of the study

Kokic and Bell's approach is based on the fact that the sampling design is a stratified simple random sampling, which is not the case since the reference year 2016 because of the update of the delineation of enterprises (see 1.2). The aim of this study is so to adapt the Kokic and Bell's method to a weight sharing.

2 Methods

Four scenarios are compared in a simulation study based on the ESANE data 2016.

2.1 Scenario 1

In scenario 1, Winsorization is performed in the drawn sample, then the weight sharing is performed with the winsorized weights instead of the drawing weights. This scenario is theoretically correct, but the winsorization thresholds are determined to optimize the accuracy of the estimator of the turnover before updating the delineations of the enterprises. Some enterprises whose influence would be amplified by weight sharing or non-response correction may therefore not be detected.

2.2 Scenario 2

In Scenario 2, weight sharing is first performed, then winsorization is performed as if the sample after weight sharing was obtained by a stratified simple random sampling. This scenario is not entirely correct theoretically, since the calculation conditions for the Kokic and Bell thresholds are not verified, but it takes into account the impact of weight sharing and non-response correction in the influence of units.

2.3 Scenario 3

In Scenario 3, a winsorization is performed in the sample drawn but not directly on the variable turnover. Indeed, the winsorization is based on a variable taking into account the futur sharing weights, in the sense that the estimator of the total of this variable (variable Z in [4]) with the sampling weights is equal to the estimator of the total of turnover after weight sharing. This scenario is theoretically correct, and it takes into account the sharing of weights in the influence of the units, but it is not guaranteed that the winsorization of the transformed variable Z leads to good results for the original variable Y, which is the one which interests us.

2.4 Scenario 4

Scenario 4 is close to Scenario 3. The difference is that the variable Z is transformed to take into account weight sharing and also nonresponse.

3 Data and Simulation study

A simulation study has been conducted on the ESANE data 2016, based on 50 000 iterations. An iteration consists in a four-steps treatments :

- Draw a sample
- Winsorization.

- Sharing weights.
- Simulation of response behavior and correction of nonresponse by reweighting.

Depending on the scenarios, the steps occur in different orders.
The four scenarios are compared to a basic scenario, number 0, in which no winsorization is performed. The indicator used in the study for a scenario sc is the ratio of the coefficient of variation of the estimator associated to the scenario sc and the coefficient of variation of the estimator associated to the scenario 0, that is without winsorisation.

$$RCV^{sc} = \frac{\sqrt{\frac{1}{rep}\sum_{r=1}^{rep}(\hat{T_y^{sc(r)}} - T_y)^2}}{\sqrt{\frac{1}{rep}\sum_{r=1}^{rep}(\hat{T_y^{0(r)}} - T_y)^2}} \quad (1)$$

With :

- rep : number of iterations (50 000).
- T_y : total of the variable Y computed in the sampling frame.
- sc : the scenario studied (0 = scenario without winsorization).
- $\hat{T_y^{sc(r)}}$: estimation of the total of Y computed with the sample corresponding to the iteration r and with the treatments corresponding to the scenario sc.

The RCV^{sc} is computed for several tax variables[2], available for each unit in the data, in particular out of the true ESANE 2016 sample.

4 Results

To evaluate the accuracy of the estimators belonging to each scenario, we compute the RCV indicator for each activity (NACE, 3 positions). We report the distribution of the results in a boxplot (Figure 1). The scenario 4 perform the best, probably because it takes into account the true sampling design and the nonresponse. At a more aggregated level (NACE, A10), scenario 4 does not perfom the best anymore, probably because the biases introduced by winsorization become too high for an aggregated level, let's remember that the winsorization is applyied at the NACE 3 positions level. In the simulations, the number of winsorized units is between 415 and 581 for scenario 4 and between 64 and 245 for the other scenarios.

5 Conclusions

Each scenario improved the accuracy of the estimator, but none of them seems consistently better than the others. Scenario 4 performs the best at the NACE 3 positions level but the worse at the NACE A10 level. Scenario 3 may represent an interesting compromise, giving good results (but not the best) at both NACE A10 and NACE 3 positions level. The choice of the scenario used in production is not made yet, and may take into account practical considerations.

[2]To limit the length of this abstract, the figures reported in this paper only concern the turnover.

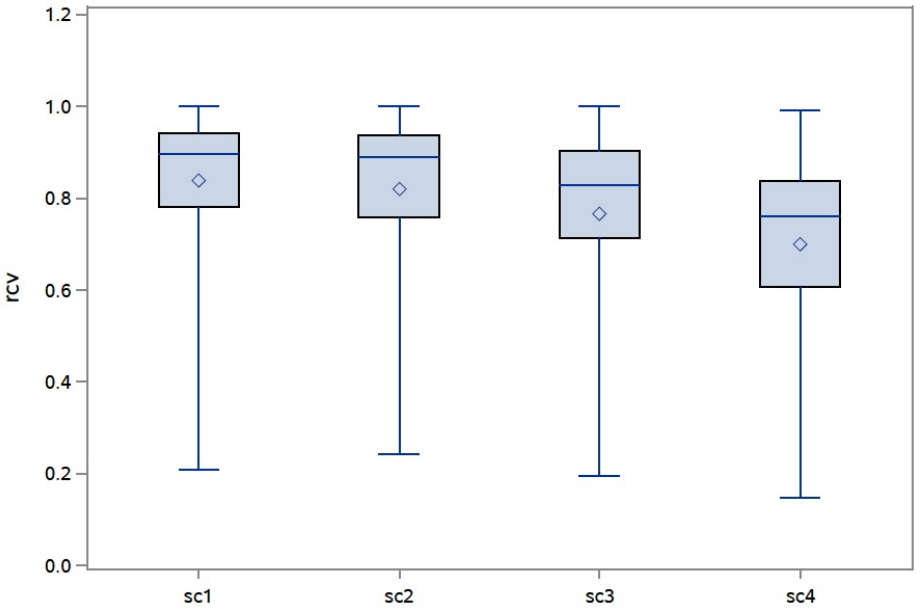

Figure 1: Distribution of *RCV* by NACE, 3 positions

References

[1] P. Brion & E. Gros. Statistical estimators using jointly administrative and survey data to produce french structural business statistics. *Journal of Official Statistics*, 31(4):589–609, 2015.

[2] Council regulation (eec) 696/93 of 15 march 1993 on the statistical units for the observation and analysis of the production system in the community. URL http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31993R0696:EN:HTML.

[3] E. Gros & R. Le Gleut. The impact of profiling on sampling. *presentation at the European Establishment Statistics Workshop*, 2017.

[4] P. Lavallée. *Indirect sampling*. Springer Series in Statistics, 2007.

[5] P. N. Kokic & P. A. Bell. Optimal winsorizing cut-offs for a stratified finite population estimator. *Journal of Official Statistics*, 10(4):419–435, 1994.