Output editing based on winsorization procedure

*A application in the French system of structural business statistics Esane*

Thomas Deroyon & Emmanuel Gros
Insee – Sampling division
Context / introduction

- Since 2008, SBS aggregates are produced with a new system, called Esane, combining administrative and survey data

- Esane surveys are establishment surveys
  - one-stage stratified sampling with strong differences of sampling rates between strata
  - Highly skewed variables of interest
  - Subject to the problem of (representative) outliers

- Outliers are dealt with at the post-collect treatment step, with specific procedures (winsorization) and at the output editing step, with top-down analysis of contributions

- Use of winsorization at the output editing step
Structure of the system Esane

Business register  Tax data  Employment data  Survey

Statistics
Methodological issue: produce statistical estimates jointly using, in the most efficient way, both administrative data and statistical survey.

Framework

Survey data

Administrative data

Statistical device

⇒ Use of calibration techniques

⇒ Use of specific estimators, mainly difference estimators
Statistical estimates (2)

- Starting point: the weighted estimator

\[ \sum_{i \in R} d_i^{NR, \text{winso}} Y_i \]

- First step: use of calibration techniques:

  \[ \Rightarrow \text{modify weights according to the following calibration equations:} \]

\[ \sum_{i \in R} w_i T^{\text{tax}}(i) \cdot \mathbb{I}_{\text{APE\_rep=X}}(i) = \sum_{i \in U} T^{\text{tax}}(i) \cdot \mathbb{I}_{\text{APE\_rep=X}}(i) \]

\[ \sum_{i \in R} w_i \cdot \mathbb{I}_{\text{APE\_rep=X}}(i) = \sum_{i \in U} \mathbb{I}_{\text{APE\_rep=X}}(i) \]

where APE\_rep is the value of the APE code in the register and \( T^{\text{tax}}(i) \) the value of the turnover of enterprise \( i \) in tax data.

\[ \Rightarrow \text{implemented at the “3-digit” level to limit weights distortion.} \]
Statistical estimates (3)

- Second step: for sector-based estimates, use of a difference estimator, confronting the APE code of the register (APE_rep) and the APE code coming from the survey (APE_enq):

\[
\sum_{i \in R} w_i Y_i^{\text{Tax}} \mathbb{I}_{\text{APE_enq}=X(i)} + \sum_{i \in U} Y_i^{\text{Tax}} \mathbb{I}_{\text{APE_rep}=X(i)} - \sum_{i \in R} w_i Y_i^{\text{Tax}} \mathbb{I}_{\text{APE_rep}=X(i)}
\]

- appropriate to the Esane device, since variables \( Y_i^{\text{Tax}} \mathbb{I}_{\text{APE_enq}=X(i)} \) and \( Y_i^{\text{Tax}} \mathbb{I}_{\text{APE_rep}=X(i)} \) are usually very similar.

NB: for estimations at detailed level, problems of wrongly negative values with these difference estimators \( \Rightarrow \) difference estimators used only for estimations at the “3-digit” level (and upper) of the French nomenclature.
The problem of « representative outliers »

- Sampling plan of the ESA:
  - half of the ESA sample is renewed every year, whereas the other half is maintained in the survey.
  - one-stage stratified sampling design

| strata ≈ economic sectors ⊗ size classes |

leads to appearance of “representative outliers”, that is units:

- which present unusually large values of the variable of interest compared to the other units of their stratum ...
- ... but not due to measurement errors.
The problem of « representative outliers » (2)

Why these representative outliers?

- Because of the time differential between the sampling frame data used to constitute the strata (before the beginning of the survey) and the data collected during the survey.

<table>
<thead>
<tr>
<th>Change in the economic sector</th>
<th>Enterprise classified in a “wrong” stratum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Increase of the number of employees</td>
<td>→ usually, sampling weight too high.</td>
</tr>
</tbody>
</table>
Objective: reduce variance of estimations

→ By decreasing the value of outliers in order to reduce their impact on estimates accuracy.

→ Price to pay: introduce a downward bias in the estimation

⇒ Winsorization procedures seek for the best bias-variance trade-off, by minimizing the mean squared error (MSE) of estimates.
Principle of winsorisation

Reduce the dispersion of measured values by decreasing the value of outliers

\[ y_{hi}^* = \varphi_h(y_{hi}) = \begin{cases} \frac{n_h}{N_h} y_{hi} + \left(1 - \frac{n_h}{N_h}\right) K_h & \text{si } y_{hi} \geq K_h \\ y_{hi} & \text{si } y_{hi} < K_h \end{cases} \]

No winsorisation in take-all strata (weight equal to one)

\[ \hat{Y}_{\text{Winsor}} = \sum_{i \in s} w_i y_i^* = \sum_{h=1}^{H} \left(\frac{N_h}{n_h}\right) \sum_{i \in s_h} y_{hi}^* \]
Kokic & Bell thresholds for winsorization

- Main issue: determine the thresholds $K_h$?
  - Assumption 1: $y_{hi}$ sequence of i.i.d random variables with mean $\mu_h$ and standard deviation $\sigma_h$
  - Assumption 2: thresholds $K_h$ independent from the survey sample
  - Minimization of the MSE of the estimator $\hat{Y}^{\text{Winsor}}$.

→ Kokic & Bell (1994): asymptotically, when the MSE is minimized:

$$\forall h, (N_h/n_h-1)(K_h-\mu_h) \sim \text{Bias of } \hat{Y}^{\text{Winsor}}(K_1, \ldots K_H)$$

→ system of $H$ equations with $H$ unknowns comes down to one equation involving bias, $F(B)=0$, and then $K_h = -(N_h/n_h-1)^{-1}B + \mu_h$

Definition of thresholds per stratum → 1-dimensional search
ESA is close to the framework of Kokic & Bell method

- Sampling design: one-stage stratified sampling

- Fiscal data available for all members of the population
  ⇒ we have all the information we need to calculate thresholds $K_h$ independent from the sample.

⇒ Since 2008, winsorization is applied to ESA, after non-response treatment and before calibration
Problem:

- We have the information to winsorize each fiscal variable independently
- But these variables have to respect relationships (accounting principles)
- These relationships would not hold if each variable was winsorized independently

Choice:

- We chose a central variable, the turnover in the fiscal source
- After NR adjustment, winsorization is applied to fiscal turnover
- The correction on turnover is applied to all over variables

\[ V_i^W = V_i \frac{T_i^W}{T_i} \]

with $T$ turnover, $V$ another fiscal variable, and exponent $W$ indicating winsorized value
Problem:

- This winsorization is expected to work well for variables strongly related to turnover
- It may work poorly for other variables
- Subject-matter experts report problems with remaining outliers at the output editing step

Output editing procedure:

- Applied to aggregates levels and growth rates
- Top-down procedure to identify spurious observations
- Three sectoral classifications are considered:
  - In 17 sectors (A17)
  - In 38 sectors (A38)
  - The 3-digit level of Nace
Output editing procedure:

- The A17 spurious aggregates growth rates are identified
- A38 aggregates contributing the most to spurious A17 growth rates are identified
- Nace 3-digits aggregates contributing the most to spurious A38 growth rates are identified
- In these Nace 3-digits, units with the biggest contributions are identified are treated
- This procedure is applied to each aggregate of interest, to validate A17, A38 and Nace 3-digits aggregates

Thorough but time consuming procedure

→ Use of winsorization to go faster
Principle:

- Winsorization thresholds are calculated for all variables of interest.
- Units that would be winsorized with these thresholds (and their value after post-collect treatment, i.e. after winsorization on turnover) are identified as potential outliers.
- The effect this winsorization would have on each unit is a measure of its outlierness.

→ Gives a list of outliers with a score of outlierness.

Score of outlierness:

The score has to measure the effect of winsorization on the difference estimates used for SBS aggregates:

\[ \sum_{i \in R} w_i Y_i^{\text{Tax}} \mathbb{I}_{\text{APE_enq}=X(i)} + \sum_{i \in U} Y_i^{\text{Tax}} \mathbb{I}_{\text{APE_rep}=X(i)} - \sum_{i \in R} w_i Y_i^{\text{Tax}} \mathbb{I}_{\text{APE_rep}=X(i)} \]
Score of outlierness:

- The effect of winsorization of variable $Y$ on $Y$ total in sector $X$ is given by

\[ \Delta_i^W = w_i (Y_i^W - Y_i) (I_{APE_{enq}}=X(i) - I_{APE_{rep}}=X(i)) \]

- If $APE_{enq}$ and $APE_{rep}$ are equal, winsorization has no effect on aggregates

- If $APE_{enq} = u$ and $APE_{rep} = v$ are different, winsorization has an effect on two sectoral aggregates
  - $Y$ total in sector $u$ increases by \( w_i (Y_i^W - Y_i) \)
  - $Y$ total in sector $v$ decreases by \( -w_i (Y_i^W - Y_i) \)
Output for subject-matter experts:

- Two lists of units:
  - Potential outliers for 21 variables belonging to the profits and loss accounts
  - Potential outliers for 13 variables describing investments and assets

- In each list, units that would be winsorized and whose winsorization has an effect on difference estimates.

- For each unit of the list, three measures of oulierness are given:
  - Effect of winsorization on A17 sectoral aggregates (with indication of the sectors concerned)
  - Effect on A38 sectoral aggregates (with indication of the sectors concerned)
  - Effect on Nace 3-digit sectoral aggregates (with indication of the sectors concerned)

This procedure will be used for the first time in the actual output editing step in autumn 2015.
A test on 2013 SBS aggregates (1)

- The procedure was applied to SBS 2013 aggregates.
- It identifies a limited number of outliers for some variables of interest in some sectors, but not for all of them.
- The procedure seems to work for sectors and variables secondary to the activity of the sector (for instance, sales in real-estate activities or transportation).

- In trade, all firms make sales ⇒ it is unlikely that a limited number of outliers have a major effect on the aggregate.
- In other sectors for which sales are a secondary part of turnover, if is more likely that strong outliers remain and have an effect on aggregates.
- Winsorization also worked poorly for variables strongly related to turnover (such as salaries or value added).
- Two examples: effect of winsorization in 2013 on total of sales and salaries in A17-sectors.
## A test on 2013 SBS aggregates (2)

<table>
<thead>
<tr>
<th>Sector</th>
<th>Aggregate negative effect</th>
<th>Number of units having a negative effect</th>
<th>Aggregate positive effect</th>
<th>Number of units having a positive effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>AZ Agriculture</td>
<td>0</td>
<td>0</td>
<td>9.29</td>
<td>3</td>
</tr>
<tr>
<td>C1 Manufacture of food products, beverage and tobacco</td>
<td>0</td>
<td>0</td>
<td>0.67</td>
<td>10</td>
</tr>
<tr>
<td>C3 Manufacture of electric equipement, machinery and computer</td>
<td>0</td>
<td>1</td>
<td>0.54</td>
<td>34</td>
</tr>
<tr>
<td>C4 Manufacture of transport equipment</td>
<td>0</td>
<td>3</td>
<td>0.04</td>
<td>11</td>
</tr>
<tr>
<td>C5 Manufacture of other products</td>
<td>-0.03</td>
<td>13</td>
<td>0.46</td>
<td>83</td>
</tr>
<tr>
<td>DE Mining and quarrying</td>
<td>0</td>
<td>0</td>
<td>9.62</td>
<td>18</td>
</tr>
<tr>
<td>FZ Construction</td>
<td>0</td>
<td>0</td>
<td>11.98</td>
<td>9</td>
</tr>
<tr>
<td>GZ Trade</td>
<td>-0.06</td>
<td>81</td>
<td>4.25</td>
<td>5</td>
</tr>
<tr>
<td>HZ Transportation and storage</td>
<td>0</td>
<td>0</td>
<td>0.75</td>
<td>5</td>
</tr>
<tr>
<td>IZ Accomodation and food service activities</td>
<td>0</td>
<td>0</td>
<td>0.4</td>
<td>9</td>
</tr>
<tr>
<td>JZ Information and communication</td>
<td>-0.02</td>
<td>1</td>
<td>6.4</td>
<td>21</td>
</tr>
<tr>
<td>KZ Financial and insurance activities</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>LZ Real estate activities</td>
<td>0</td>
<td>0</td>
<td>17.07</td>
<td>3</td>
</tr>
<tr>
<td>MN Professional, scientific, technical, administrative and support service activities</td>
<td>0</td>
<td>1</td>
<td>6.4</td>
<td>21</td>
</tr>
<tr>
<td>RU Other service activities</td>
<td>0</td>
<td>1</td>
<td>0.07</td>
<td>2</td>
</tr>
</tbody>
</table>
A test on 2013 SBS aggregates (3)

<table>
<thead>
<tr>
<th>Sector</th>
<th>Aggregate negative effect</th>
<th>Number of units having a negative effect</th>
<th>Aggregate positive effect</th>
<th>Number of units having a positive effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>AZ Agriculture</td>
<td>0</td>
<td>0</td>
<td>0.09</td>
<td>1</td>
</tr>
<tr>
<td>C1 Manufacture of food products, beverage and tobacco</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>C3 Manufacture of electric equipement, machinery and computer</td>
<td>-0.01</td>
<td>15</td>
<td>0.01</td>
<td>4</td>
</tr>
<tr>
<td>C4 Manufacture of transport equipment</td>
<td>0</td>
<td>4</td>
<td>0.01</td>
<td>4</td>
</tr>
<tr>
<td>C5 Manufacture of other products</td>
<td>-0.01</td>
<td>36</td>
<td>0.01</td>
<td>12</td>
</tr>
<tr>
<td>DE Mining and quarrying</td>
<td>-0.02</td>
<td>10</td>
<td>0.01</td>
<td>4</td>
</tr>
<tr>
<td>FZ Construction</td>
<td>-0.09</td>
<td>9</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>GZ Trade</td>
<td>-0.01</td>
<td>8</td>
<td>0.09</td>
<td>6</td>
</tr>
<tr>
<td>HZ Transportation and storage</td>
<td>-0.01</td>
<td>4</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>IZ Accomodation and food service activities</td>
<td>-0.03</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>JZ Information and communication</td>
<td>-0.01</td>
<td>5</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>KZ Financial and insurance activities</td>
<td>-0.4</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>LZ Real estate activities</td>
<td>0</td>
<td>0</td>
<td>0.15</td>
<td>4</td>
</tr>
<tr>
<td>MN Professional, scientific, technical, administrative and support service activities</td>
<td>-0.12</td>
<td>13</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>RU Other service activities</td>
<td>-0.07</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Thank you for your attention!

Contact

M. DEROYON Thomas
Tél. : +33 1 41 17 56 87
Mail : thomas.deroyon@insee.fr

M. GROS Emmanuel
Tél. : +33 1 41 17 56 87
Mail : emmanuel.gros@insee.fr

Insee
18 bd Adolphe-Pinard
75675 Paris Cedex 14
www.insee.fr
Informations statistiques :
www.insee.fr / Contacter l’Insee
09 72 72 4000
(coût d’un appel local)
du lundi au vendredi de 9h00 à 17h00