Sampling coordination of business surveys

A new method implemented at INSEE

Gros Emmanuel
Methodologist – Sampling division – Insee
Sample coordination

- The purpose of sample coordination is to take into account the samples of previous surveys when drawing a new sample...

- … in order to reduce the statistical burden of small businesses (large businesses, from a certain threshold, being systematically surveyed in most surveys)…

- … while preserving the unbiasedness of the samples.

- Two kinds of coordination:
  
  ✓ **negative coordination (the most frequently used at Insee):** foster the selection of businesses that have not already been selected in recent surveys → minimising the overlap between samples;

  ✓ **positive coordination:** foster the selection of businesses that have already been selected in some other surveys → maximising the overlap between samples.
The new Insee’s method of coordination

- A Permanent Random Number (PRN) technique…
  - Each unit $k$ of the population is given a permanent random number $\omega_k$, independently drawn according to $U[0;1]$.

- … which rests on the concept of coordination function:
  - coordination function $g = \text{measurable function from } [0;1] \text{ onto itself, which preserves uniform probability:}$

    $$\text{if } P \text{ is the uniform probability on } [0;1], \text{ then we have } P \circ g = P.$$ 

    $$\text{→ if the } (\omega_k)_{k \in U} \text{ are independent random numbers selected according to } P, \text{ the transformed numbers } (g(\omega_k))_{k \in U} \text{ are also independently selected according to } P.$$ 

- Drawing of a sample by stratified simple random sampling: using a “wisely chosen” coordination function $g_k$, select, within each stratum $h$ of size $N_h$, the $n_h$ smallest transformed PRN $g_k(\omega_k)$.
The coordination functions

- For the drawing of a given sample, the coordination function $g_k$ takes into account the cumulative response burden of unit $k$ to meet the objective of negative or positive coordination:

  - for negative (resp. positive) coordination, $g_k$ is defined so that the higher the cumulative response burden of unit $k$, the higher (resp. smaller) the number $g_k(\omega_k)$.

- The coordination functions of previous surveys form the basis for calculating the cumulative response burden...

- … which allows next to determine the coordination function of each unit for the current sample drawing.

$$
\left[
\begin{array}{cc}
  g_{k,1}(\omega) & \gamma_{k,1}\left(g_{k,1}(\omega)\right) \\
  \vdots & \Rightarrow \vdots \\
  g_{k,t-1}(\omega) & \gamma_{k,t-1}\left(g_{k,t-1}(\omega)\right)
\end{array}
\right] \Rightarrow \Gamma_{k,t-1} = \sum_{u \leq t-1} \gamma_{k,u}(g_{k,u}(\omega)) \Rightarrow g_{k,t}(\omega)
$$
A comprehensive and efficient method

A very comprehensive method:

✓ allows both negative and positive coordination;

✓ allows to coordinate a sample with any number of previous surveys…

✓ … while differentiating the response burden assigned to each survey;

✓ allows the coordination of surveys based on different unit types.

First assessments of the method conducted on simulated data highlight its efficiency in terms of response burden allocation over the population units…

… as well as its robustness regarding the surveys parameters: sampling rates, differences of stratification between the surveys, overlapping of the surveys scope, response burden assigned to each survey, etc.
First full-scale test on real data

A simulation study – consisting in the drawing of a sequence of 20 legal unit samples, each sample being negatively coordinated with the whole of previous ones – proved the operational effectiveness of the method, and confirm its efficiency:

<table>
<thead>
<tr>
<th>Cumulative response burden of legal units, except take-all strata</th>
<th>Frequency according to the sampling scheme</th>
<th>Differences between coordinated and independent drawings</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Independent drawings</td>
<td>Coordinated drawings</td>
</tr>
<tr>
<td>0</td>
<td>3 981 423</td>
<td>3 952 718</td>
</tr>
<tr>
<td>1</td>
<td>391 840</td>
<td>445 402</td>
</tr>
<tr>
<td>2</td>
<td>30 494</td>
<td>9 084</td>
</tr>
<tr>
<td>3</td>
<td>3 670</td>
<td>606</td>
</tr>
<tr>
<td>4</td>
<td>374</td>
<td>9</td>
</tr>
<tr>
<td>5</td>
<td>18</td>
<td>0</td>
</tr>
</tbody>
</table>
“Multi-level” coordination

The methods allows the coordination of samples relating to surveys based on different kind of units, for example legal units and local units, according to the following procedure:

1. define a permanent link between the legal unit and one of its local units;
2. assign to this “principal local unit” the same permanent random number as the legal unit;

→ enables coordination between legal unit and principal local unit.

More precisely:

✓ for the drawing of legal units samples, take into account in the cumulative response burden of legal units the response burden of their principal local unit;

✓ reciprocally, for the drawing of local units samples, take into account in the cumulative response burden of principal local units the response burden of their legal unit.
Full-scale test of the “multi-level” coordination

➢ Simulation study:

✓ drawing of a sequence of 20 legal units samples and 8 local units samples;

✓ three different sampling schemes: independent drawings, “level by level” negative coordination and “multi-level” negative coordination.

<table>
<thead>
<tr>
<th>Cumulative response burden of legal units, except take-all strata</th>
<th>Frequency according to the sampling scheme</th>
<th>Differences between drawings:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Independent drawings</td>
<td>&quot;Level by level&quot; coordinated drawings</td>
</tr>
<tr>
<td>0</td>
<td>4 670 676</td>
<td>4 651 954</td>
</tr>
<tr>
<td>1</td>
<td>410 016</td>
<td>439 355</td>
</tr>
<tr>
<td>2</td>
<td>40 095</td>
<td>34 824</td>
</tr>
<tr>
<td>3</td>
<td>8 072</td>
<td>4 679</td>
</tr>
<tr>
<td>4</td>
<td>2 142</td>
<td>813</td>
</tr>
<tr>
<td>5</td>
<td>578</td>
<td>93</td>
</tr>
<tr>
<td>6</td>
<td>121</td>
<td>5</td>
</tr>
<tr>
<td>7</td>
<td>20</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>0</td>
</tr>
</tbody>
</table>
Assessment of two methodological issues

- Theoretically, survey feedback from coordinated samples leads to bias in the estimates. Yet, the largest business survey in France is the annual sectoral survey (ESA), which allows to update the sectoral classification of legal units...

  \[\text{Simulations on past ESA show that the magnitude of the feedback bias was small enough to be negligible.}\]

- The coordination procedure is incompatible with systematic sampling on sorted file (currently used at INSEE).

  \[\text{Transform the implicit stratification on the sorting variable into an explicit stratification with proportional allocation. Simulations show that the increasing of the number of strata does not deteriorate the quality of coordination.}\]
Conclusions

➢ A very comprehensive method:

✓ negative and positive coordination;
✓ differentiation of the response burden assigned to each survey;
✓ coordination of surveys based on different unit types.

➢ Assessments conducted on simulated data as well as full-scale tests highlight both the efficiency and the robustness of the method (and its operational effectiveness…).

➢ Used in production at INSEE since the end of 2013.
Thank you for your attention!

Contact
M. GROS Emmanuel
Tel.: +33 1 41 17 64 91
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