# The Procedure of Sampling Coordination for Business Surveys Implemented at INSEE: Methodology and Practice

Emmanuel Gros[1]

**Abstract**

The public statistical system carries out each year a significant number of businesses and establishments surveys. The objective of the negative coordination of samples is to foster, when selecting a sample, the selection of businesses that have not already been selected in recent surveys, while preserving the unbiasedness of the samples. This coordination contributes to reduce the statistical burden of small businesses – large businesses, from a certain threshold, are systematically surveyed in most surveys. On the other hand, positive coordination aims at maximizing the overlap between coordinated samples, either to obtain a panelised sample or once again with the aim of reducing the statistical burden by reducing the size of questionnaires.

We present here the sampling coordination method operationally used at Insee since the end of 2013. This method – whose theoretical foundations were presented at ICES-IV by Olivier Sautory – belongs to the family of sample coordination procedures based on Permanent Random Numbers (PRN), and is based on the notion of coordination function. These functions, defined for each unit and each new drawing taking into account the past response burden of each unit, transform permanent random numbers so as to meet the objective of negative or positive coordination.

After a brief reminder of the main principles of the method limited to the case of stratified simple random sampling, we first present the results of simulation studies assessing the properties of this coordination method. Then, we focus on how the method allows the coordination of samples relating to surveys based on different kinds of units, for example legal units and local units. Finally, we address two drawbacks of these coordination procedure – feed back bias and incompatibility with systematic sampling – and expose options chosen on both issues.

**KeyWords:** Sampling coordination, permanent random numbers, coordination function.

## 1. The Procedure of Sampling Coordination for Business Surveys Implemented at INSEE

We present here the main principles of the method, detailed in [1], limited to the case of stratified simple random sampling, which is the sampling design the most frequently used at Insee for business surveys. This method was proposed by C. Hesse in 2001 in [2], and studied by P. Ardilly in 2009 in [3].

[1] Emmanuel Gros, Insee, 18 boulevard Adolphe Pinard, 75675 Paris cedex 14, FRANCE, email: emmanuel.gros@insee.fr

## 1.1 Coordination Functions – Samples selection

### *1.1.1 Definition of a coordination function*

The concept of coordination function plays an essential role in the method.

> A coordination function g is a measurable function from [0,1] onto itself, which preserves uniform probability: if P is the uniform probability on [0,1], then the image probability $P^g$ is P. It means that for any interval I = [a, b[ included in [0,1] :
>
> $$P\left[g^{-1}(I)\right] \overset{def}{=} P^g(I) = P(I) = b - a$$

The length of the inverse image of any interval under g equals the length of this interval: a coordination function preserves the length of intervals – or union of intervals – by inverse image.

### *1.1.2 Selection of samples*

Each unit k of the population is given a permanent random number $\omega_k$, drawn according to the uniform distribution on the interval [0,1[. The drawings of the $\omega_k$ are mutually independent.

We consider a sequence of surveys t = 1, 2,…(t refers to the date and the number of the survey), and we denote by $S_t$ the sample corresponding to survey t. Suppose that one has defined for each unit k a "wisely chosen" coordination function (see 2.2) $g_{k,t}$ which changes at each survey t.

The drawing of the sample $S_t$ by stratified simple random sampling is done by selecting, within each stratum (h,t) of size $N_{(h,t)}$, the $n_{(h,t)}$ units associated with the $n_{(h,t)}$ smallest numbers $g_{k,t}(\omega_k),\ \ k = 1...N_{(h,t)}$ .

**<u>Proof</u>**

> The $N_{(h,t)}$ random numbers ($\omega_k$) associated to the $N_{(h,t)}$ units of the stratum have been independently selected according to the uniform probability on [0,1], denoted P. Since we have $P^{g_{k,t}} = P$ for each k, the N numbers $g_{k,t}(\omega_k)$ are also independently selected according to P. Then, using a well-known result, the $n_{(h,t)}$ smallest values $g_{k,t}(\omega_k)$ give a simple random sample of size $n_{(h,t)}$ in the stratum.

## 1.2 Construction of a Coordination Function from the Cumulative Response Burden

### *1.2.1 Response burden and coordination function*

Let $\Omega$ denote the vector of random numbers $\omega_k$ given to the population units k, and $\gamma_{k,t}$ be the response burden of a questioned business k at survey t. The cumulative burden for unit k is a random variable, function of $\Omega$, equal to:

$$\Gamma_{k,t}(\Omega) = \sum_{u \le t} \gamma_{k,u} . 1\!\!1_{k \in S_u}(\Omega) \quad (1)$$

We wish to define, for each unit k, a coordination function $g_{k,t}$ based on $\Gamma_{k,t-1}$, the cumulative burden of unit k until survey t-1. To meet the objective of negative coordination – **to draw as a priority, for a given sample selection, units that have had the lowest response burden during the recent period** – and taking into account the selection scheme of the units – **the higher the probability for the unit to be selected the smaller the number $g_{k,t}(\omega_k)$** –, a desirable property for any coordination function is the following:

$$\Gamma_{k,t-1}(\Omega^{(1)}) < \Gamma_{k,t-1}(\Omega^{(2)}) \Rightarrow g_{k,t}(\omega_k^{(1)}) \le g_{k,t}(\omega_k^{(2)})$$

where $\omega_k^{(i)}$ (i=1,2) denotes the kth component of vector $\Omega^{(i)}$.

This condition is not easy to handle, because the function $\Gamma_{k,t-1}(\Omega)$ is a function of vector $\Omega$: it depends not only on the random number $\omega_k$ given to unit k, but on all the other random numbers $\omega_1 \ldots \omega_N$. We will see on ❸ how we can replace this function by a function $\Gamma'_{k,t-1}(\omega_k)$ which depends only on $\omega_k$. The desirable property for any coordination function $g_{k,t}$ will become :

$$\Gamma'_{k,t-1}(\omega_k^{(1)}) < \Gamma'_{k,t-1}(\omega_k^{(2)}) \Rightarrow g_{k,t}(\omega_k^{(1)}) \le g_{k,t}(\omega_k^{(2)}) \quad (2)$$

### *1.2.2 Construction of a coordination function*

For the sake of simplicity, we omit the subscripts k and t. So $\omega$ is now a simple real number between 0 and 1. We note C the cumulative burden function – supposed to be a bounded measurable function: $\omega \in [0,1] \to C(\omega) \in \mathbb{R}$ – and we wish to associate to it a coordination function g such that:

$$C(\omega^{(1)}) < C(\omega^{(2)}) \Rightarrow g\ (\omega^{(1)}) \le g\ (\omega^{(2)}) \quad (2')$$

Let us define the function $G_C = F_C(C)$, with $F_C$ the cumulative distribution function of C:

$$\forall \omega \in [0,1], G_C(\omega) = P\big(u \big| C(u) < C(\omega)\big)$$

We can show that the range of $G_C$ is included in [0,1], and that $G_C$ satisfies (2'), but is not a coordination function if C has "levels", that is subsets of [0,1] where C is constant ($G_C$ has then the same levels). However, we can construct a bijective coordination function on [0,1] $g_C$ equal to $G_C$ outside the levels and composed of line segments having a slope equal to 1 on the levels of $G_C$, as illustrated in figure 1, where C is a step function, with 4 levels.

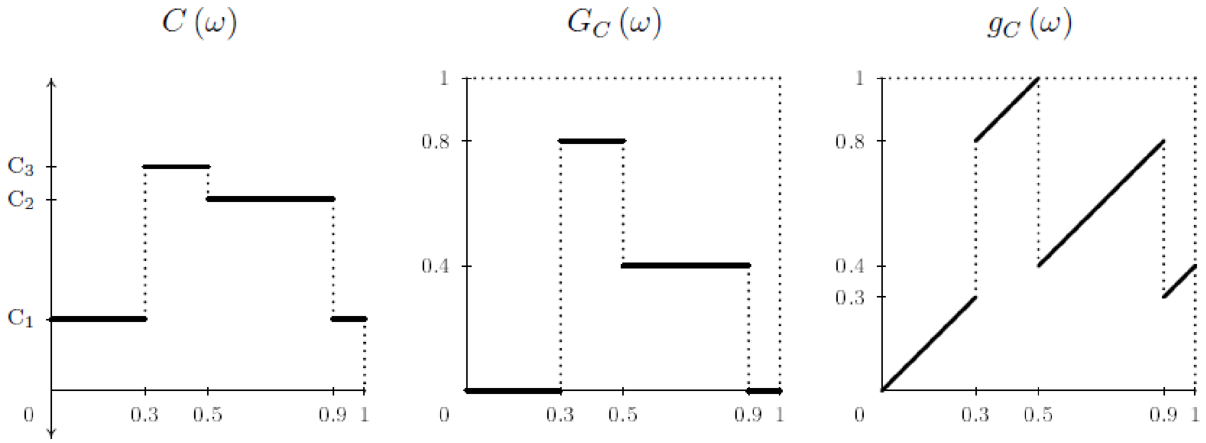


**Figure 1:** Cumulative burden function C, function $G_C$ and coordination function $g_C$

## 1.3 Application to Stratified Simple Random Sampling

With this sampling method, we select a unit k in sample $S_t$ if the random number $g_{k,t}(\omega_k)$ is one of the n lowest numbers $g_{i,t}(\omega_i)$ associated with all the units i of the sampling frame[2]. Then the inclusion of k in $S_t$ depends on the random numbers $\omega_i$ of all the units i. The indicator function $I_{k,t}$, together with the cumulative burden $\Gamma_{k,t}$, are functions of vector $\mathbf{\Omega}$. So there is a need to replace the indicator function $I_{k,t}$ with an approximate indicator function $I'_{k,t}$, which should be a function of $\omega_k$ close to $I_{k,t}$.

### *1.3.1 The approximate indicator function – The expected cumulative burden function*

The best approximation of the indicator function $I_{k,t}(\mathbf{\Omega})$ depending only on $\Omega_k$, in the L2-norm sense, is its conditional expectation given $\Omega_k$:

$$I^a_{k,t}(\omega) = E\left(I_{k,t}(\mathbf{\Omega}) \middle| \Omega_k = \omega\right) = P\left(k \in S_t \middle| \Omega_k = \omega\right)$$

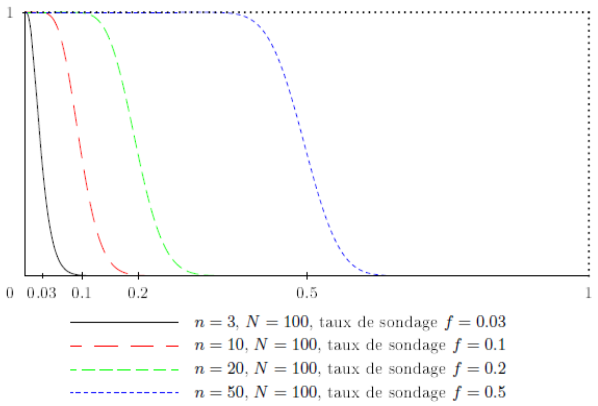
If we suppose that the coordination functions are bijective[3] functions, we can write

$$I^a_{k,t}(\omega) = P\left(k \in S_t \middle| g_{k,t}(\Omega_k) = g_{k,t}(\omega)\right) = b_{k,t}\left(g_{k,t}(\omega)\right)$$

where 1 − $b_{k,t}(x)$ is the cumulative distribution function of a beta distribution with parameters n and N−n. The graph in figure 2 shows the shape of the b(x) function for some values of n and N.

[2] We recall that we omit the stratum index.

[3] This property is satisfied with the method described here, but it is not an intrinsic property of a coordination function.



A b(x) function has the following shape: a first part "almost horizontal" close to 1, corresponding to an "almost certain" selection of the unit in the sample, a third part "almost horizontal" close to 0, corresponding to an "almost certain" non-selection of the unit in the sample. Between them, a decreasing part "with a high negative slope" corresponding to a more or less short interval on the abscissa axis: this interval is nearly centered on the value n/N, equal to the sampling rate. Around this value there is uncertainty about the selection of the unit in the sample.

**Figure 2:** Shape of the b(x) function for some values of n and N.

Due to the substitution of the approximate indicator function for the indicator function, the cumulative burden function itself is replaced, in formula (1) in §1.2.1., by an <u>expected</u> cumulative burden function $\Gamma^{e}_{k,t}$ given $\Omega_k$ :

$$\Gamma^{e}_{k,t}(\omega) = \sum_{u=1}^{t} \gamma_{k,u}\, I^{a}_{k,u}(\omega)$$

To ensure that the algorithm performs well, that is it leads to unbiased samples, it is necessary to use this <u>expected</u> burden instead of the <u>actual</u> burden. The latter is based on the <u>observed</u> inclusions of unit k in the successive samples:

$$\Gamma_{k,t} = \sum_{u=1}^{t} \gamma_{k,u}\ \mathbb{1}(k \in S_u)$$

### *1.3.2 Approximation by step functions*

The approximate indicator functions $I^{a}_{k,t}(\omega)$ and the expected cumulative burden functions are not step functions or functions that can be easily "computed". We will simplify the shape of the approximate indicator functions $I^{a}_{k,t} = b_{k,t}$ as follows:

❶ We divide the interval [0,1[ into L equal subintervals $I_\ell = \left[\frac{\ell - 1}{L}; \frac{\ell}{L}\right[ \quad \ell = 1 \ldots L$[4].

❷ We replace the approximate indicator function $b_{k,t}$ by a piecewise linear function $\tilde{b}_{k,t}$ which takes the same values as $b_{k,t}$ at the endpoints of the intervals $I_\ell$.

❸ We compute the the average value $\beta_{k,t}(\ell)$ of $\tilde{b}_{k,t}$ on each interval $I_\ell$.

❹ We define the function $\beta_{k,t}$ as : $\forall \omega \in I_\ell \quad \beta_{k,t}(\omega) = \beta_{k,t}(\ell)$. $\beta_{k,t}$ is an approximation of the approximate indicator function $I^{a}_{k,t}$ by a piecewise constant function.

[4] L being a "large enough" integer (at least greater than 50).

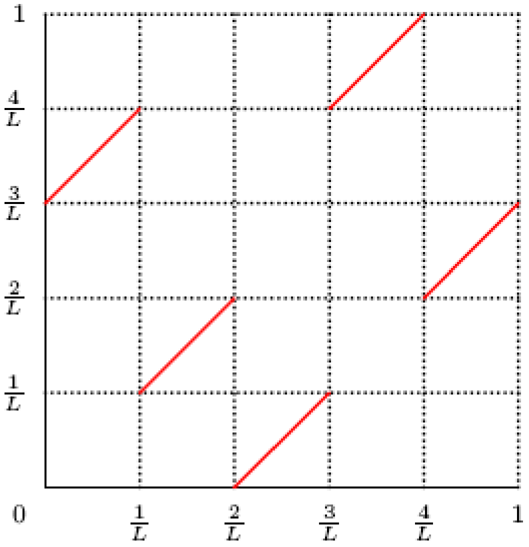
Finally, the cumulative burden function $\Gamma_{k,t}$ is replaced by the approximate expected cumulative burden function $\Gamma^{ea}_{k,t}(\omega)=\sum_{u=1}^{t}\gamma_{k,u}\,\beta_{k,u}\big(g_{k,u}(\omega)\big)$, which is a step function, constant on each $I_\ell$.

### *1.3.3 Construction of a coordination function*

So we are in the same context as in the example presented in §1.2.2: from the function $\Gamma^{ae}_{k,t}$, we construct a "G" function, also constant on each $I_\ell$, and then a coordination function g which looks like in the opposite example wit L=5.

It is entirely defined by a permutation $\sigma$ on {1,2,3...,L}, according to the following formula:

$$\forall\,\omega\in\left[\frac{\ell-1}{L};\frac{\ell}{L}\right[\quad g_\sigma(\omega)=\frac{\sigma(\ell)-1}{L}+(\omega-\frac{\ell-1}{L})$$



The only remaining issue is so the definition of the permutation $\sigma$. To do this, we go back to the fundamental property (2') of the coordination function: the smaller is the criterion (here the cumulative response burden), the smaller is the value of the coordination function $g_\sigma$. Now, on $\left[\frac{l-1}{L};\frac{l}{L}\right[$, the smaller is $\sigma(\ell)$, the smaller is $g_\sigma$. So, we will arrange the values $\sigma(\ell)$ exactly in the same order as the values of the approximate expected cumulative burden function $\Gamma^{ea}_{k,t}(\ell)$:

$$\Gamma^{ea}_{k,t}(\ell_1)\le\Gamma^{ea}_{k,t}(\ell_2)\le\ldots\le\Gamma^{ea}_{k,t}(\ell_L)\Leftrightarrow\sigma(\ell_1)\le\sigma(\ell_2)\le\ldots\le\sigma(\ell_L)$$

where $\ell_i$ is the identifier of the i$^{th}$ standardized interval.

As $\sigma$ has to be a permutation, and therefore bijective, we add the following additional constraint: if $\Gamma^{ea}_{k,t}(p)=\Gamma^{ea}_{k,t}(q)$ and $p<q$, then $\sigma(p)<\sigma(q)$. *In fine,* this means imposing strict inequalities in the ranking of the $\sigma(\ell)$, which leads to $\sigma(\ell_i)=i$ and completely defines the permutation $\sigma$.

# 2. Assessment of the Coordination Procedure and Coordination Between Surveys Based on Different Kind of Units

A first empirical assessment of the procedure was conducted in 2012 on simulated data. The results of these simulations, presented in [3], were very satisfactory: the coordination method proved to be both highly efficient[5] – in terms of response burden allocation over the population units – and remarkably robust with regarding the parameters of the different sampling plans – sampling rates, differences of stratification between the surveys, overlapping of the surveys scope, response burden assigned to each survey, etc.

This first study was doubly completed by the work of Kevin Rosamont-Prombo presented in [4]:

- Firstly, additional and more complete tests on simulated data were conducted, and their results confirmed those obtained previously.

- Secondly, a first test on real data was performed, based on the Information and Communication Technologies surveys from 2008 to 2012. The results were once again satisfactory in terms of response burden allocation over the population units. They also showed that the coordination procedure could be used in concert with the method used at Insee for the management of rotating samples in business surveys, and permitted to understand the interaction between these two methods in case of simultaneous use.

Simulations and results presented later in this section enrich and complement earlier work along two axes:

- Full-scale test of the procedure on real data in order to assess its operational feasibility and its performance in a production situation.

- test of a "multilevel" coordination procedure allowing sample coordination between surveys based on different kind of units (legal units and establishments, for example).

## 2.1 Full-scale test on real data: coordinated drawing of 20 legal units samples

To assess the operational feasibility of the coordination procedure, as well as its properties in terms of response burden allocation over the population units in a production situation, we conducted a full-scale simulation study on real data. The simulations consisted in:

- starting with the 2008 annual sectoral survey (ESA), which thus constitutes the survey initiating the sequence of coordinated drawings in our simulations;

- then performing, in chronological order, the drawings of the 19 other legal units samples:

[5] Compared to independent drawings.

- respecting the sampling designs used during the actual drawings of theses surveys: stratification criteria, allocations, positive coordination of the sample of the “retail outlet” survey with the ESA 2009 sample, etc.
- each sample being negatively coordinated with the whole of previous ones.

A sequence of 20 independent drawings was also carried out, in order to assess the efficiency of the coordination process in terms of response burden allocation over the population units.

From an operational standpoint, there is absolutely no problem with the coordination procedure:

- Computation time remains reasonable: about 8 hours for the complete sequence of 20 coordinated drawings;

- The same goes for storage requirements: all the permutations required for defining the coordination functions involved in the drawings take up only 6 GB;

In terms of efficiency of the sampling coordination method, there is, as expected, a far better response burden allocation over the population units when the drawings are negatively coordinated. Table 1 shows the distribution of the population units depending on the number of samples in which they have been selected – i.e. the distribution over the population of the cumulative response burden, defined by the variable "number of selections". As the coordination does not affect the take-all strata, they were excluded from the calculations of cumulative response burden, in order to be able to assess the quality of the procedure on its real scope.

| Cumulative response burden, except take-all strata | Frequency according to the sampling scheme | | Difference between coordinated and independent drawings |
|---|---|---|---|
| | Independent drawings | Coordinated drawings | |
| 0 | 3 981 423 | 3 952 718 | -28 705 |
| 1 | 391 840 | 445 402 | 53 562 |
| 2 | 30 494 | 9 084 | -21 410 |
| 3 | 3 670 | 606 | -3 064 |
| 4 | 374 | 9 | -365 |
| 5 | 18 | 0 | -18 |

**Table 1:** Allocation of the cumulative response burden, except take all-strata, according to the sampling scheme.

As expected, and in line with the results obtained in previous tests on simulated data and on real data based only on ICT surveys, there is a narrowing of the distribution around 1, that is a spreading of the response burden: the number of units selected in more than one sample decreases in significant proportions, as the number of non-sampled units, in favour of a marked increase in the number of units selected in a single sample.

Furthermore, it is important to note that this coordination method also allows for positive coordination between surveys: to do this, all you have to do is to assign a negative response burden to the survey(s) with whom you wish to positively coordinate the survey you are drawing. In our simulations, we have also assigned a negative response burden to the sample of ESA 2009 when drawing the sample of the "retail outlet" survey. The results in terms of recovery between the two samples are satisfactory, slightly higher than those observed with the coordination method previously used by INSEE, based on another technique.

Finally, in Table 1, the fact that some units are selected in more than one sample is mainly explained by:

- The positive coordination of the sample of the survey "retail outlet" survey with the sample of the ESA 2009;

- The existence of strata with high sampling rates in some surveys.

Thus, of the 9 084 units present in two samples in the sequence of coordinated drawings, 2 909 are due to the positive coordination mentioned above. For the 6 175 remaining units, 50% of them belong, in one of the two samples in which they are selected, to strata with a sampling rate greater than 50%, and 45% belong to strata with a sampling rate between 20% and 50%.

## 2.2 Sample coordination between surveys based on different kind of units

The methods allows the coordination of samples relating to surveys based on different kind of units, for example legal units and local units. This “multi-level” coordination is performed thanks to the following procedure:

- We first define a permanent link between the legal unit and one of its local units – the head office of the legal unit at the time of its creation – and assign to this “principal local unit” the same permanent random number as the legal unit – the PRN of other local units being drawn according to the uniform distribution on the interval [0,1]. We get so, for each level, a set of permanent random numbers following a uniform distribution on [0,1], with a one-to-one link [legal unit ↔ principal local unit] between these two sets.

- Then, each level is subjected to its own coordination system – which implies in particular the management of coordination functions specific to each level –, the coordination between legal units samples and local units samples taking place exclusively through the [legal unit ↔ principal local unit] link as follows:

  - When drawing a legal units sample, coordination with samples relating to local units is performed by taking into account in the cumulative response burden of legal units the response burden of their principal local unit;
  - Reciprocally, when drawing a local units sample, coordination with legal units samples is performed by taking into account in the cumulative response burden of principal local units the response burden of their legal unit.

We have assessed the efficiency of this multi-level coordination procedure, by incorporating in our simulations 8 local units surveys in addition to the 20 legal units surveys previously mentioned. Three different sampling schemes were used:

- Independent drawings of the 28 samples, respecting the sampling designs used during the actual drawings;
- Coordinated drawings of the 20 legal units samples on the one side, and of the 8 local units samples on the other side, but without multi-level coordination;
- Coordinated drawings of the 28 samples via the multi-level coordination procedure described above.

We then compare the results of these three strategies, in terms of distribution of legal units response burden, the response burden of the principal local units being taken into account in the cumulative response burden of their legal units. The results in table 2 are in line with our expectations and consistent with the previous ones: compared to independent drawings, the strategy of separated coordinated drawings leads to a far better response burden allocation over the population units, and this phenomenon is further strengthened when a multi-level coordination procedure is performed.

| Cumulative response burden of legal units, except take-all strata | Frequency according to the sampling scheme | | | Differences between drawings: | | |
|---|---|---|---|---|---|---|
| | Independent drawings | "Level by level" coordinated drawings | Multi-level coordinated drawings | Independent *versus* "level by level" coordinated | "level by level" *versus* multi-level coordinated | Independent versus multi-level coordinated |
| 0 | 4 670 676 | 4 651 954 | 4 634 250 | -18 722 | -17 704 | -36 426 |
| 1 | 410 016 | 439 355 | 474 286 | 29 339 | 34 931 | 64 270 |
| 2 | 40 095 | 34 824 | 18 230 | -5 271 | -16 594 | -21 865 |
| 3 | 8 072 | 4 679 | 4 125 | -3 393 | -554 | -3 947 |
| 4 | 2 142 | 813 | 737 | -1 329 | -76 | -1 405 |
| 5 | 578 | 93 | 92 | -485 | -1 | -486 |
| 6 | 121 | 5 | 2 | -116 | -3 | -119 |
| 7 | 20 | 0 | 1 | -20 | 1 | -19 |
| 8 | 3 | 0 | 0 | -3 | 0 | -3 |

**Table 2:** Allocation of the cumulative response burden of legal units, except take all-strata, according to the sampling scheme.

## 3. Study of Two Methodological Issues

To conclude this review about the properties of the coordination method, we addressed two methodological issues: the problem of "feedback bias" on the one hand, and the issue of systematic sampling on sorted file on the other hand.

### 3.1 The "feedback" bias issue

The feedback bias issue is a well known problem which appears in the context of sampling coordination: if we update the sampling frame from a sample A, and then draw in this sampling frame another sample B coordinated with the sample A, this may leads to bias in the results for survey B.

This phenomenon is particularly problematic in the context of a global coordination system for business surveys. Indeed, the sampling frames of the majority of business surveys conducted by Insee are derived from the business register Sirus, which is regularly updated from the results of different surveys. For example, dead units identified thanks to surveys are deleted from the business register. Another example is the sectoral classification of units in Sirus – a classification which constitutes a stratification variable in almost all business surveys –, which is updated each year based on the results of the annual sectoral survey. Therefore, the establishment of a global coordination system for business surveys requires:

- either to prohibit feedback from surveys to the business register, which seems unrealistic because it means to deny oneself the use of all available information;

- either to exclude from the global coordination system the annual sectoral survey (ESA), which is the most important survey used to update the business register. However, insofar as the ESA is the largest – in terms of sample size – business survey, representing a high response burden, this solution would not be completely satisfactory;

- either to ignore the problem of feedback bias, assuming that it is low enough to be negligible compared to the disadvantages of the two alternatives outlined above.

In order to settle the argument between the two last options, we conducted a simulation study, based on data from the SBS production device Esane, to quantify the magnitude of the feedback bias. More specifically, we performed, on the wholesale trade sector, a sequence of 5 000 independent drawings “ESA 2008 → ESA 2009 → ESA 2010 → ESA 2011”, and another sequence of 5 000 drawings with negative coordination. Then, we compute, for each strategy, relative bias for estimators by sectors and by size groups thanks to tax data available for all units. Table 3 shows the distribution of these relative bias for sector based estimates concerning the main variables of the Esane device.

**Independent drawings**

| Variable | Number of entreprises | Turnover | Total purchases | Salary | Value added | Gross operating profit | Accounting result | Total assets | Total liabilities |
|---|---|---|---|---|---|---|---|---|---|
| Mean | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% | -0,1% | 0,0% | 0,0% |
| Maximum | 0,3% | 0,1% | 0,1% | 0,2% | 0,2% | 6,8% | 4,1% | 0,9% | 1,0% |
| P99 | 0,3% | 0,1% | 0,1% | 0,2% | 0,2% | 6,8% | 4,1% | 0,9% | 1,0% |
| P95 | 0,1% | 0,1% | 0,1% | 0,1% | 0,1% | 1,0% | 2,4% | 0,1% | 0,1% |
| P90 | 0,1% | 0,1% | 0,1% | 0,1% | 0,1% | 0,2% | 0,2% | 0,1% | 0,1% |
| P75 | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% | 0,1% | 0,1% | 0,0% | 0,0% |
| Median | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% |
| P25 | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% | -0,1% | -0,1% | 0,0% | 0,0% |
| P10 | -0,1% | -0,1% | -0,1% | -0,1% | -0,1% | -0,2% | -0,6% | -0,1% | -0,1% |
| P5 | -0,1% | -0,1% | -0,2% | -0,1% | -0,1% | -1,1% | -1,9% | -0,3% | -0,2% |
| P1 | -0,2% | -0,6% | -0,6% | -0,4% | -0,4% | -11,8% | -6,7% | -1,0% | -1,1% |
| Minimum | -0,2% | -0,6% | -0,6% | -0,4% | -0,4% | -11,8% | -6,7% | -1,0% | -1,1% |

**coordinated drawings**

| Variable | Number of entreprises | Turnover | Total purchases | Salary | Value added | Gross operating profit | Accounting result | Total assets | Total liabilities |
|---|---|---|---|---|---|---|---|---|---|
| Mean | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% | 0,1% | 0,0% | 0,0% | 0,0% |
| Maximum | 0,1% | 0,2% | 0,2% | 0,3% | 0,3% | 2,3% | 2,7% | 0,4% | 0,8% |
| P99 | 0,1% | 0,2% | 0,2% | 0,3% | 0,3% | 2,3% | 2,7% | 0,4% | 0,8% |
| P95 | 0,1% | 0,2% | 0,2% | 0,2% | 0,2% | 1,2% | 1,2% | 0,2% | 0,2% |
| P90 | 0,1% | 0,1% | 0,1% | 0,1% | 0,1% | 0,3% | 0,6% | 0,1% | 0,2% |
| P75 | 0,0% | 0,1% | 0,1% | 0,0% | 0,0% | 0,1% | 0,1% | 0,0% | 0,1% |
| Median | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% |
| P25 | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% | -0,1% | -0,1% | 0,0% | 0,0% |
| P10 | -0,1% | -0,1% | -0,1% | 0,0% | -0,1% | -0,2% | -0,3% | -0,1% | -0,1% |
| P5 | -0,2% | -0,1% | -0,1% | -0,1% | -0,1% | -0,4% | -0,7% | -0,1% | -0,1% |
| P1 | -0,2% | -0,1% | -0,2% | -0,1% | -0,1% | -2,4% | -3,0% | -0,5% | -0,6% |
| Minimum | -0,2% | -0,1% | -0,2% | -0,1% | -0,1% | -2,4% | -3,0% | -0,5% | -0,6% |

**Table 3:** Mean and distribution of relative bias for sector based estimates concerning the main variables of the Esane device in 2011, according to the sampling scheme.

As we can see, carrying out coordinated drawings did not appear to induce significant and systematic bias in the estimates compared with a strategy of independent drawings, and the magnitude of the feedback bias seems to be small enough to be negligible.

### 3.2 Systematic sampling and coordination

Samples of business surveys are almost always drawn according to stratified sampling designs, with equal probabilities within each stratum. Moreover, the drawing of the units within each stratum is frequently done by systematic sampling after sorting units within each stratum according to a given criterion. This drawing procedure – which provides, within each stratum, a distribution of sampled units close to that observed in the sampling frame for the sort criterion – is unfortunately totally incompatible with a coordination procedure based on permanent random number. However, as systematic sampling on sorted file is equivalent to an implicit stratified sampling with proportional allocation, it is possible to take into account the criterion previously "controlled" by the systematic sampling in coordinated drawings as follows:

❶ We first redefine the sorting variable as an additional stratification variable in order to define drawing strata;

❷ We then apply the sampling rates computed on the initial stratification to the drawing strata in order to define drawing allocations;

❸ Finally, we merge, if needed, some of the drawing strata in order to avoid strata with an allocation equal to zero.

This procedure leads to an increase of the number of drawing strata, which could affect the quality of the coordination. In order to assess the impact of this “over-stratification” procedure, we performed a simulation study, comparing three sampling schemes: systematic sampling, coordinated drawings without over-stratification and “systematic coordinated drawings”, that is coordinated drawings with over-stratification. Results in table 4 show that the increasing of the number of strata does not deteriorate the quality of coordination.

| Cumulative response burden, except take-all strata | Frequence according to the sampling scheme | | | Difference between independent systematic drawings and "simple" coordinated drawings | Difference between "simple" and "systematic" coordinated drawings |
|---|---|---|---|---|---|
| | Independent systematic drawings | "Simple" coordinated drawings | "Systematic" coordinated drawings | | |
| 0 | 630 452 | 627 016 | 626 896 | -3 436 | -120 |
| 1 | 37 029 | 43 703 | 43 784 | 6 674 | 81 |
| 2 | 3 258 | 213 | 251 | -3 045 | 38 |
| 3 | 188 | 1 | 2 | -187 | 1 |
| 4 | 6 | 0 | 0 | -6 | 0 |

**Table 4:** Allocation of the cumulative response burden, except take all-strata, according to the sampling scheme.

## 4. Conclusion

The sampling coordination method presented in this paper proves, via many simulations studies conducted on simulated as well as real data, to be very efficient – providing significant gains in terms of response burden allocation over the population units – as well as outstandingly robust vis-à-vis sampling design parameters. It is used operationally at Insee since the end of 2013.

## References

[1] F. Guggemos and O. Sautory, *Sampling Coordination of Business Surveys Conducted by Insee*, Proceedings of the Fourth International Conference of Establishment Surveys, June 11-14, 2012, Montréal, Canada.

[2] C. Hesse, *Généralisation des tirages aléatoires à numéros aléatoires permanents, ou la méthode JALES+*, Insee working paper E0101 (2001).

[3] P. Ardilly, *Présentation de la méthode JALES+ conçue par Christian Hesse*, internal Insee working paper (2009).

[4] Kevin Rosamont-Prombo, *La coordination des échantillons d'entreprises*, internship report, Insee, 2012.