Sample Coordination and Response Burden for Business Surveys: Methodology and Practice of the Procedure Implemented at INSEE

Emmanuel Gros, Ronan Le Gleut

European Establishment Statistics Workshop 2017 - Southampton

August 30, 2017

Summary

Sommaire

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.

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 $\rightarrow$ minimising the overlap between samples;
- **positive coordination**: foster the selection of businesses that have already been selected in some other surveys $\rightarrow$ maximising the overlap between samples.

Sommaire

INSEE's procedure of sample coordination

This procedure is a **Permanent Random Number (PRN)** technique: each unit k of the population U is given a PRN $\omega_k \sim \mathcal{U}[0;1]$.

It rests on the concept of **coordination function**: measurable function $g : [0;1] \to [0;1]$ which preserves uniform probability $\mathbb{P}$.

INSEE's procedure of sample coordination

This procedure is a **Permanent Random Number (PRN)** technique: each unit k of the population U is given a PRN $\omega_k \sim \mathcal{U}[0;1]$.

It rests on the concept of **coordination function**: measurable function $g : [0;1] \to [0;1]$ which preserves uniform probability $\mathbb{P}$.

$\Rightarrow$ if the $\mathbf{\Omega} = (\omega_k)_{k \in U}$ are independent random numbers selected according to $\mathbb{P}$, the transformed numbers $(g(\omega_k))_{k \in U}$ are also independently selected according to $\mathbb{P}$.

INSEE's procedure of sample coordination

This procedure is a **Permanent Random Number (PRN)** technique: each unit k of the population U is given a PRN $\omega_k \sim \mathcal{U}[0;1]$.

It rests on the concept of **coordination function**: measurable function $g : [0;1] \to [0;1]$ which preserves uniform probability $\mathbb{P}$.

$\Rightarrow$ if the $\mathbf{\Omega} = (\omega_k)_{k \in U}$ are independent random numbers selected according to $\mathbb{P}$, the transformed numbers $(g(\omega_k))_{k \in U}$ are also independently selected according to $\mathbb{P}$.

The selection of a sample is done 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 cumulative response burden function

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

$$C_{k,t-1}(\boldsymbol{\Omega}) = \sum_{u \leq t-1} c_{k,u} \cdot \mathbb{1}_{k \in S_u}(\boldsymbol{\Omega}) \ ,$$

with $c_{k,u}$ the response burden of a questioned business k at survey u.

The cumulative response burden function

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

$$C_{k,t-1}(\mathbf{\Omega}) = \sum_{u \leq t-1} c_{k,u} \cdot \mathbb{1}_{k \in s_u}(\mathbf{\Omega}) \ ,$$

with $c_{k,u}$ the response burden of a questioned business k at survey u.

To meet the objective of negative coordination, a desirable property for any coordination function is the following:

$$C_{k,t-1}\left(\mathbf{\Omega}^{(1)}\right) < C_{k,t-1}\left(\mathbf{\Omega}^{(2)}\right) \ \Rightarrow \ g_{k,t}\left(\omega_k^{(1)}\right) < g_{k,t}\left(\omega_k^{(2)}\right)$$

The expected cumulative burden function

Because $\mathbb{1}_{k \in S_u}(\boldsymbol{\Omega})$ depends not only on ω_k but on all $(\omega_k)_{k \in U}$, we approximate the indicator function by its conditional expectation:

$$\mathbb{1}^{a}_{k \in S_u}(\omega) = \mathbb{E}\Big(\mathbb{1}_{k \in S_u}(\boldsymbol{\Omega}) \mid \omega_k = \omega\Big) = \mathbb{P}\Big(k \in S_u \mid \omega_k = \omega\Big)$$

The expected cumulative burden function

Because $\mathbb{1}_{k\in S_u}(\mathbf{\Omega})$ depends not only on ω_k but on all $(\omega_k)_{k\in U}$, we approximate the indicator function by its conditional expectation:

$$\mathbb{1}^a_{k\in S_u}(\omega) = \mathbb{E}\Big(\mathbb{1}_{k\in S_u}(\mathbf{\Omega}) \mid \omega_k = \omega\Big) = \mathbb{P}\Big(k \in S_u \mid \omega_k = \omega\Big)$$

If we suppose that the coordination functions are bijective functions:

$$\mathbb{1}^a_{k\in S_u}(\omega) = \mathbb{P}\Big(k \in S_u \mid g_{k,u}(\omega_k) = g_{k,u}(\omega)\Big) = b_{k,u}\Big(g_{k,u}(\omega)\Big)\ ,$$

where $1 - b_k(x)$ is the cumulative distribution function of a beta distribution with parameters n and $N - n$.

The expected cumulative burden function

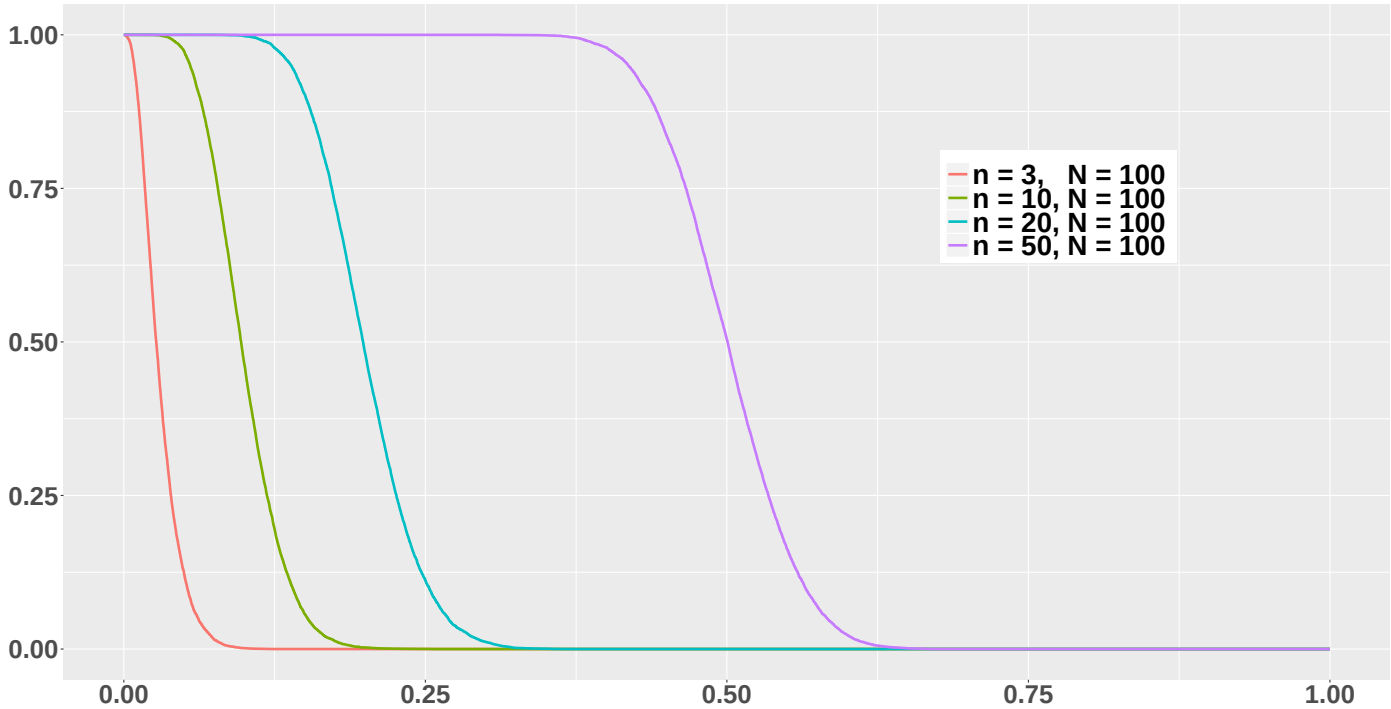


Figure: $b_k(x)$ function for some values of n and N.

This function is then approximated by a step function to construct an approximate **expected cumulative burden function**:

$$C^e_{k,t-1}(\omega) = \sum_{u \leq t-1} c_{k,u} \cdot \mathbb{1}^a_{k \in S_u}(\omega)$$

Response burden and coordination function

From the expected cumulative burden function C, we define the distribution function G_C of C:

$$G_C(\omega) = \mathbb{P}\Big(u \mid C(u) < C(\omega)\Big)$$

Response burden and coordination function

From the expected cumulative burden function C, we define the distribution function G_C of C:

$$G_C(\omega) = \mathbb{P}\Big(u \mid C(u) < C(\omega)\Big)$$

Then we can construct a bijective coordination function g_C:

$$g_C(\omega) = G_C(\omega) + \sum_i \mathbb{1}_{A_i}(\omega) \cdot \int \mathbb{1}_{A_i \cap [0,\omega]}(u) \mathrm{d}u$$

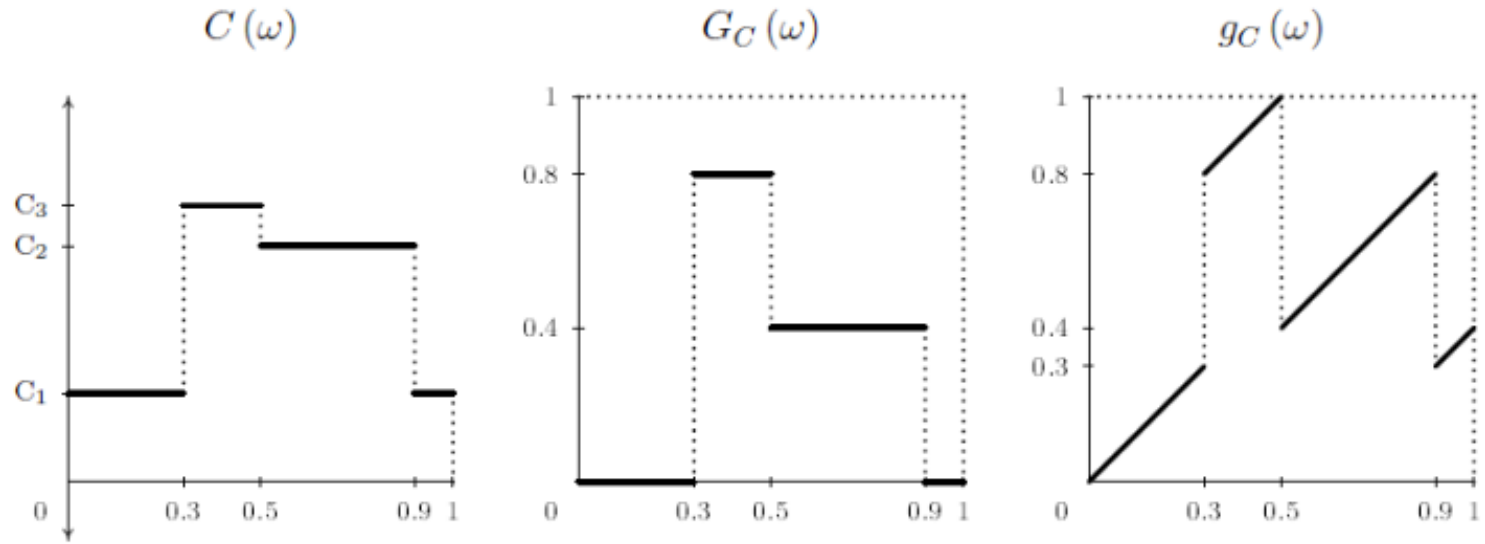


Figure: Burden function C, distribution function G_C and coordination function g_C.

Construction of a coordination function

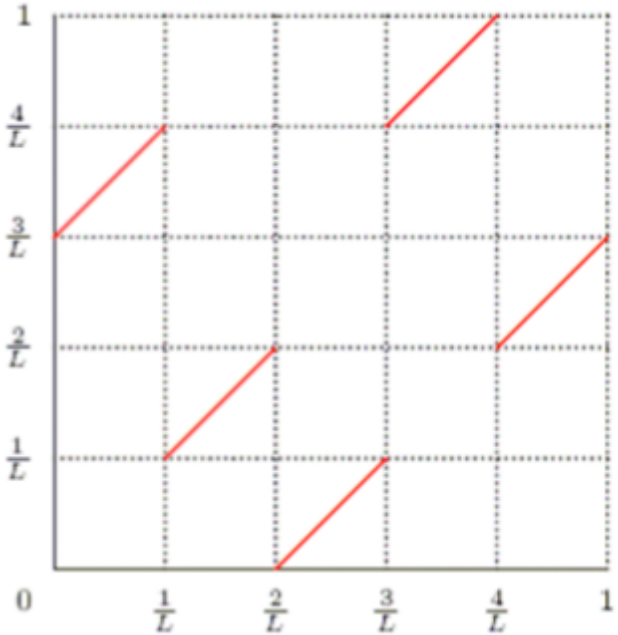


Figure: Example of a coordination function with $L = 5$.

If the approximation by step functions is done by dividing the interval $[0, 1[$ into L equal subintervals $\left[\frac{\ell-1}{L}; \frac{\ell}{L}\right[$, the coordination function g is entirely defined by a permutation σ:

$$g_\sigma(\omega) = \frac{\sigma(\ell) - 1}{L} + (\omega - \frac{\ell - 1}{L})$$

Construction of a coordination function

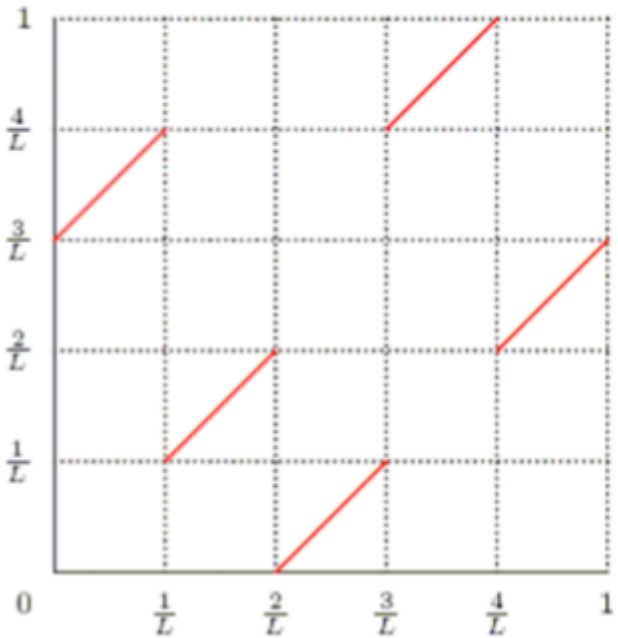


Figure: Example of a coordination function with $L = 5$.

If the approximation by step functions is done by dividing the interval $[0, 1[$ into L equal subintervals $\left[\frac{\ell-1}{L}; \frac{\ell}{L}\right[$, the coordination function g is entirely defined by a permutation σ:

$$g_\sigma(\omega) = \frac{\sigma(\ell) - 1}{L} + (\omega - \frac{\ell - 1}{L})$$

The smaller is $\sigma(\ell)$, the smaller is g_σ. We arrange the values $\sigma(\ell)$ exactly in the same order as the values of the approximate expected cumulative burden function:

$$C^e_{k,t-1}(\ell_1) \leq C^e_{k,t-1}(\ell_2) \leq \ldots \leq C^e_{k,t-1}(\ell_L) \Leftrightarrow \sigma(\ell_1) \leq \sigma(\ell_2) \leq \ldots \leq \sigma(\ell_L)$$

Sommaire

Coordinated drawing of 20 legal units samples

Starting with the 2008 annual sectoral survey (ESA), we perform, in chronological order, the drawings of the 19 other legal units samples:

- respecting the sampling designs used during the actual drawings of these surveys;
- each sample being negatively coordinated with the whole of previous ones.

Coordinated drawing of 20 legal units samples

Starting with the 2008 annual sectoral survey (ESA), we perform, in chronological order, the drawings of the 19 other legal units samples:

- respecting the sampling designs used during the actual drawings of these surveys;
- each sample being negatively coordinated with the whole of previous ones.

$\Rightarrow$ This simulation study allows to validate the **operational feasibility** of the method.

Coordinated drawing of 20 legal units samples

Starting with the 2008 annual sectoral survey (ESA), we perform, in chronological order, the drawings of the 19 other legal units samples:

- respecting the sampling designs used during the actual drawings of these surveys;
- each sample being negatively coordinated with the whole of previous ones.

$\Rightarrow$ This simulation study allows to validate the **operational feasibility** of the method.

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

Coordinated drawing of 20 legal units samples

Cumulative response burden (except take-all strata)	Frequency according to the sampling scheme		Difference between coordinated and independant drawings
	Ind. drawings	Coord. 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: Allocation of the response burden, except take all-strata, according to the sampling scheme.

We obtain a far better response burden allocation over the population units when the drawings are negatively coordinated.

Sommaire

Conclusion

The coordination procedure implemented at INSEE is a very comprehensive method that allows:

- both negative and positive coordination;
- to coordinate a sample with any number of previous surveys...
- ...while differentiating the response burden assigned to each survey;
- the coordination of surveys based on different unit types (see Discussion).

Conclusion

The coordination procedure implemented at INSEE is a very comprehensive method that allows:

- both negative and positive coordination;
- to coordinate a sample with any number of previous surveys...
- ...while differentiating the response burden assigned to each survey;
- the coordination of surveys based on different unit types (see Discussion).

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).

This method is used in production at INSEE since the end of 2013.

Thank you for your attention!

Sommaire

"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:

- define a **permanent link** between the legal unit and one of its local units;
- assign to this "principal local unit" the same permanent random number as the legal unit.

"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:

- define a **permanent link** between the legal unit and one of its local units;
- assign to this "principal local unit" the same permanent random number as the legal unit.

The simulation study consists in:

- the drawing of a sequence of 20 legal units samples and 8 local units samples;
- with three different sampling schemes: independent drawings, "level by level" negative coordination and "multi-level" negative coordination.

"Multi-level" coordination

Cumulative response burden of legal units (except take-all strata)	Frequency according to the sampling scheme			Difference between coordinated and independant drawings	
	Ind. drawings	"Level by level" coord. drawings	"Multi-level" coord. drawings	Ind. versus "Level by level"	Ind. versus "Multi-level"
0	4 670 676	4 651 954	4 634 250	-18 722	-36 426
1	410 016	439 355	474 286	29 339	64 270
2	40 095	34 824	18 230	-5 271	-21 865
3	8 072	4 679	4 125	-3 393	-3 947
4	2 142	813	737	-1 329	-1 405
5	578	93	92	-485	-486
6	121	5	2	-116	-119
7	20	0	1	-20	-19
8	3	0	0	-3	-3

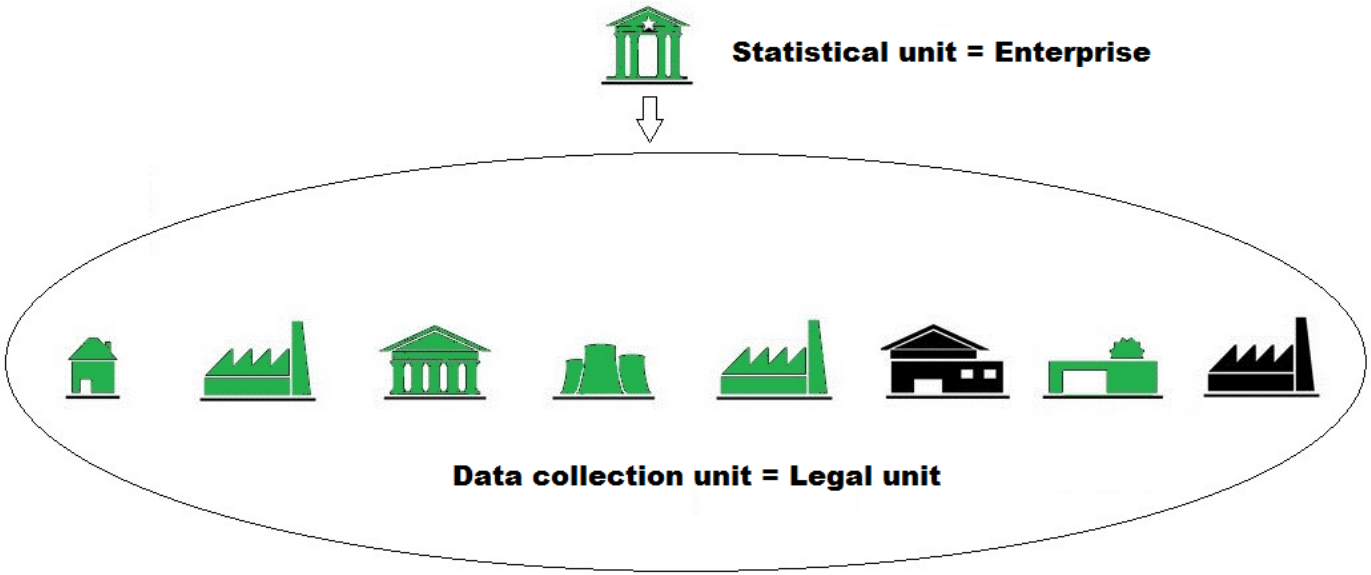
Table: Allocation of the response burden, except take all-strata, according to the sampling scheme.

We obtain a better response burden allocation over the population of legal units when the drawings are negatively coordinated accorded to the "multi-level" coordination procedure.

Coordination with surveys at the enterprise level

In order to comply with the Structural Business Statistics (SBS) European regulation, business statistics will be more and more based on the economic notion of **enterprise**.

Since the statistical units (enterprises) are different from the data collection units (legal units), the sample design can be seen as a **single-stage cluster sampling**.



Coordination with surveys at the enterprise level

In this context:

- we also define a **permanent link** between the enterprise and one of its legal units;
- we assign to this "principal legal unit" the same permanent random number as the enterprise.

Coordination with surveys at the enterprise level

In this context:

- we also define a **permanent link** between the enterprise and one of its legal units;
- we assign to this "principal legal unit" the same permanent random number as the enterprise.

More precisely:

- for the drawing of enterprises samples: take into account the response burden of their principal legal unit only;
- reciprocally, for the drawing of legal units samples: for the principal legal unit only, take into account the response burden of its enterprise.

Coordination with surveys at the enterprise level

In this context:

- we also define a **permanent link** between the enterprise and one of its legal units;
- we assign to this "principal legal unit" the same permanent random number as the enterprise.

More precisely:

- for the drawing of enterprises samples: take into account the response burden of their principal legal unit only;
- reciprocally, for the drawing of legal units samples: for the principal legal unit only, take into account the response burden of its enterprise.

$\Rightarrow$ **How can we account for the response burden of all the legal units of an enterprise?**

Ardilly, P. (2009).

Présentation de la méthode JALES+ conçue par Christian Hesse.

INSEE working paper.

Gros, E. (2016).

The procedure of sampling coordination for business surveys implemented at INSEE: methodology and practice.

ICES-V, Geneva.

Guggemos, F. and Sautory, O. (2012).

Sampling coordination of business surveys conducted by INSEE.

ICES-IV, Montréal.

Hesse, C. (2001).

Généralisation des tirages aléatoires à numéros aléatoires permanents, ou la méthode JALES+.

INSEE working paper E0101.

Rosamont-Prombo, K. (2012).

La coordination des échantillons d'entreprises.

INSEE internship report.