The unit problem: A first assessment of the impact of profiling on sampling.

Ronan Le Gleut[1] Anaïs Levieil[2] Élodie Martal[3] Thomas Merly-Alpa[4]

Abstract

In many countries of the European Union, business statistics are undergoing great changes. In France, for instance, business surveys are currently based on the observation of legal units that have a juridical definition. However, from now on, business statistics will be more and more based on the economic notion of enterprise, which is the smallest combination of legal units that is an organisational unit producing goods or services with a certain degree of autonomy. The use of this statistical unit as reporting unit has become compulsory due to economy globalization. To this end, an important methodological operation of "profiling" – which consists in a manual delineation of enterprises within complex business groups on the one hand and to consider the other groups as one enterprise on the other hand – is ongoing at the French National Institute of Statistics and Economic Studies (Institut National de la Statistique et des Études Économiques, INSEE). We present here a first assessment of the impact of profiling on sampling frames and survey designs. Since the statistical units (enterprises) are now different from the data collection units (legal units), the survey design can be seen as a two-stage cluster sampling. As a cluster, an enterprise is randomly selected, and then all legal units within this enterprise are included in the sample. The aim of this paper is to optimize the survey design in order to have a good precision on estimators at the enterprise level, under the constraint of a limited number of collection units.

Key Words: profiling, cluster sampling, survey design optimisation.

1. Introduction

In France as in many other countries of the European Union, business statistics are undergoing great changes. Until now, business surveys were based on the observation of legal units that have a juridical definition. In order to comply with the Structural Business Statistics (SBS) European regulation [1], business statistics will be more and more based on the economic notion of enterprise.

A legal unit is a legal entity of public or private law. This legal entity can be:

- A legal entity, whose existence is recognized by the law independently of the persons or institutions which possess it or are part of it;
- A natural person, who, as independent, can exercise an economic activity.

It must be declared to the competent administrative bodies in order to exist.

However, from now on, business statistics will be more and more based on the economic notion of enterprise. An enterprise is the smallest combination of legal units that is an organizational unit producing goods and services, enjoying a certain decision-making autonomy, especially for the allocation of its current resources. To this end, an important methodological operation of "profiling" – which consists in appointing legal units within complex business groups – is ongoing at INSEE. The aim of this operation is to analyze the groups in order to identify relevant economic structures which are named "profiled" enterprises, and to get contact information on each of these enterprises to facilitate data collection. There are three "profiling targets" at INSEE:

- The largest groups, for which profiling is tailor-made by specialists in close co-operation with these groups: Target 1

[1]Insee, 18 boulevard Adolphe Pinard 75014 PARIS, ronan.le-gleut@insee.fr
[2]Drees, 11 place des 5 Martyrs du Lycée Buffon, 75014 PARIS, anais.levieil-guillon@sante.gouv.fr
[3]Insee, 36 rue des Trente Six Ponts BP 94217 31054 TOULOUSE Cedex 4 , elodie.martal@insee.fr
[4]Insee, 18 boulevard Adolphe Pinard 75014 PARIS, thomas.merly-alpa@insee.fr

- The smallest groups, for which profiling is automatic: Target 2
- Medium-sized groups, for which profiling is also automatic: Target 3

As of today, the profiling of Target 1 concerns 44 groups. The results of this analysis are used in many business surveys, such as ESA[2] and EAP[3]. These two surveys are part of the ESANE[4] process, which produces structural business statistics in order to answer the SBS European regulation, using both survey and administrative data [2].

ESA's purpose is to observe the different activities of a company by breaking down its turnover into activities (sectoral classification) and thereby to deduce its principal activity. The scope of this survey concerns activities of trade, services, transport and construction. The sample is very large, with almost 130 000 legal units surveyed each year, of which 116 000 legal units are located in Metropolitan France (without French overseas regions).

EAP's scope concerns manufacturing industry. It has two main goals:

- To identify the different activities carried out by an enterprise via the breakdown of its turnover into sub-sectors, and to deduce its main activity;
- To provide elements allowing the production of high-quality data on the production of manufactured goods in order to meet the requirements of the European ProdCom regulation.

The second goal also meets demands from national users, in particular from professional organisations. The sample is composed of about 35 000 units in Metropolitan France. ProdCom, defined by the European Regulation CEE 3924/91 adopted on 09/12/1991, provides statistics on the production of manufactured goods. The term comes from the French "PRODuction COMmunautaire" (Community Production) for mining, quarrying and manufacturing. This regulation defines a list of 3 500 products whose production needs to be quantified by the European countries.

In order to comply with the SBS regulation, INSEE plans to use the results related to the profiling Targets 2 and 3 in the ESANE process, especially during the sampling of these two surveys in 2016. The construction of the groups and therefore the enterprises is based on the LIFI[5] process, which uses administrative data in order to get information on the financial relations between two or more units [8].

Therefore, the drawing of the ESA and EAP samples will be done on a sampling frame composed of enterprises, instead of legal units. However, the data collection unit remains the legal unit. Most groups, especially the smaller ones, are indeed not in the situation of estimating their consolidated turnover, let alone the breaking down of this consolidated turnover into their consolidated activities. Information is therefore collected at the legal unit level, and data on groups are produced through automatic consolidation based on that information. Moreover, some users of the business data (the National Accounts for example) still use the information at the legal unit level. The survey design can be seen as a two-stage cluster sampling. Enterprises are randomly selected, and all legal units within these enterprises are included in the sample. But the cost constraint is still based on the number of legal units surveyed, which is now random. For example, if at the enterprise level, we keep the sampling rates used for a survey design at the legal unit level, this decreases the number of primary units drawn, while increasing the number of legal units to be surveyed. This survey design could also lead to the selection of units within the selected enterprises which do not belong to the scope of the ESA and EAP surveys. On the other hand, there are units which could never be selected, because they belong to enterprises which do not belong to the scope of the surveys; this may lead to bias for a study at the legal unit level.

[2] Annual Sectoral Survey.
[3] Annual Production Survey.
[4] ESANE for the French Élaboration des Statistiques Annuelles d'Entreprise
[5] Financial Links.

The aim of this paper is to assess the impact of the process of profiling on the sampling frame and on the survey design. In particular, it will address the process of building an efficient survey design in order to draw a sample of enterprises for the ESA and EAP surveys in Metropolitan France. Thereby, we try to achieve a good precision on estimators at the enterprise level, under the constraint of a limited number of collection units.

2. Materials and Methods

2.1 The sampling frame

The sampling frame used for ESA and EAP surveys for the year 2015 (see Section 1 for more information on these surveys) was composed of legal units, except for the largest groups in Target 1. We now have to build a new sampling frame on the population of enterprises, as defined in Section 1. We use the existing data on legal units and on the financial relations between these units during 2013[6] in order to identify the independent units, i.e. the ones which constitute an enterprise on their own, and the enterprises built of more than one legal unit.

The treatment of these units will depend on the following criteria:

- The legal units which have been categorised as independent are treated as before, with no modification of the scope of the survey. The units linked to a Target 1 enterprise are treated in the same way.
- For the Target 2 and 3 enterprises, we need to define their main activity, in order to know if they belong to the scope of the survey. The principal activity of an enterprise was computed using data on financial links and economic weights of the units within it. This construction might lead to coverage problems, because legal units within the scope of the survey might belong to enterprises whose principal activity is not concerned by this kind of survey.

We will now focus on the coverage of this sampling frame, and especially on the coverage problem for a study at the legal unit level. If we do not take care of the problems of overcoverage or undercoverage, this may lead to bias [12].

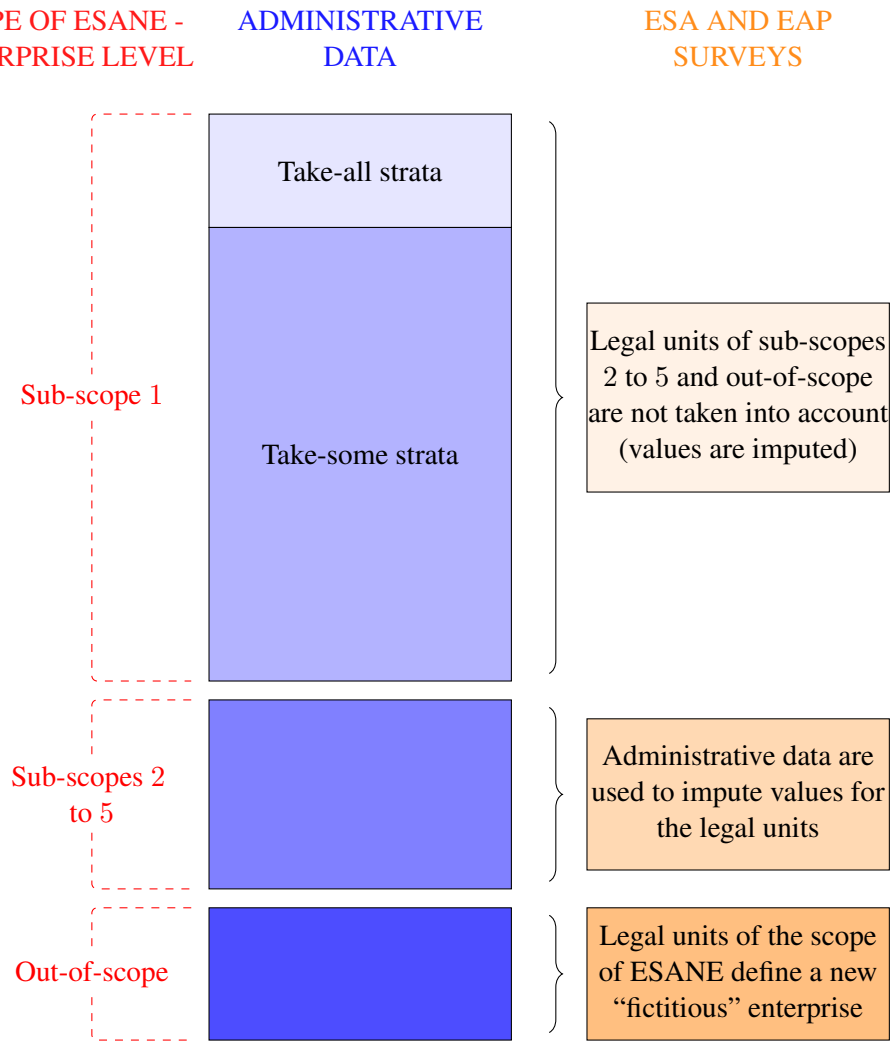
The scope of ESA and EAP surveys (defined with the business activities of the units) is usually called ESANE sub-scope 1. Administrative data give information on the turnover of the legal units that belong to this sub-scope. ESANE has 4 other sub-scopes, where the administrative data can also give information on the turnover of the legal units. These units are also taken into account to respond to the SBS regulation or ProdCom regulation, but the use of administrative data is tolerated to break down the turnover into activities. Finally, some legal units do not belong to the scope of ESANE, and we do not have any administrative information on them at INSEE.

With the new sampling frame, some enterprises from the sub-scope 1 could contain legal units that do not belong to this sub-scope. In order to avoid this overcoverage, we delete the out-of-scope units within an enterprise from sub-scope 1, and ignore their information (number of employees, turnover, etc.) to define the enterprise.

Some enterprises from the sub-scopes 2 to 5 could contain legal units that belong to the sub-scope 1. In this case, we would have an undercoverage problem, because these enterprises have an inclusion probability of zero. In order to solve this problem, we will use the administrative data, and in particular the turnover of each unit. We will use these administrative information to characterise the enterprise.

[6] It was the most recent data available at that period.

The last case is that some enterprises out-of-scope could contain legal units that belong to the sub-scope 1. In this case, we would also have an undercoverage problem, because some legal units would never be surveyed. In order to solve this problem, we define a "fictitious" enterprise which only consists of the legal units of the scope of ESANE, and calculate a "fictitious" main activity using the information on its legal units. Depending on the new main activity, the "fictitious" enterprise is then included in the sub-scope 1 or in one of the sub-scopes 2 to 5. The treatment of this "fictitious" enterprise is then similar to the one describe above.



2.2 Definition of the take-all strata

In this section, we explain the criteria behind the selection of the enterprises in the take-all strata, i.e. units which are always part of the sample, with a sampling weight of 1. This operation is iterative, and tries to respect a continuity between old and new samples, but with a necessary optimisation due to the new economic nature of the units to be drawn.

The first step was to keep the same criteria for the definition of take-all strata as the ones used today on legal units samples, since these rules were specified in order to take into account disparity between sectors. These rules are the following:

EAP : Units with more than 20 employees or with a turnover exceeding 5M€, as specified by the ProdCom regulation, are in the take-all strata. However, we also need to respect the ProdCom regulation regarding

the list of products. Therefore, a cut-off of the units achieving 85% of the turnover within each sector is made, and these units are also included in the take-all strata.

ESA : The criteria for the take-all strata are different for each business sector (trade, construction, agri-food industries, etc.) regarding the number of employees and the turnover. Enterprises with total assets of more than 75M€ are also exhaustive.

This approach led to the selection of more than 130 000 legal units through take-all strata, which means that almost all the sample is composed of exhaustive units. As the data collection process is constrained with respect to the number of legal units (around 151 000 legal units), this does not represent a viable approach. Therefore, we need to develop new techniques and new criteria to decrease the number of these units.

In order to decrease the number of legal units in the take-all strata, we decided to add a cut-off of the legal units achieving 95% of the turnover within each exhaustive enterprise. With this criteria, the legal units that represent the lowest 5% of the turnover of the enterprise are not surveyed. These legal units with a low turnover often have only one activity, so the breakdown of its turnover is not necessary. Moreover, the administrative data are used to get information on the turnover of these legal units in order to avoid a bias. Using this method allow us to diminish the size of the exhaustive part of the sample to 100 000 legal units. Unfortunately, this is still too much ; our goal is to respect the global structure of the 2015 samples regarding the mix between exhaustive and sampled units. Thus, we need to change some other criteria concerning the take-all strata, keeping this cut-off rule that permits to reduce the number of legal units in the exhaustive part of the sample.

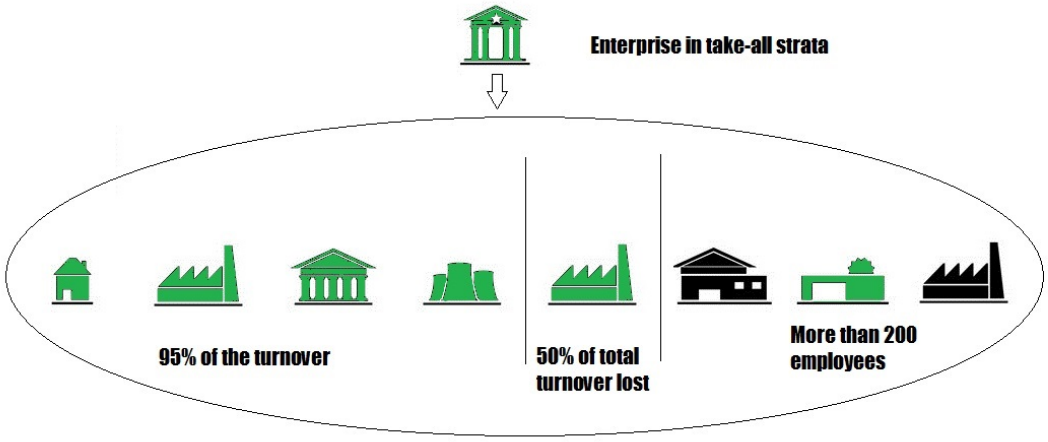
One way to deal with this issue is to keep the economic weight of the exhaustive units in the new sampling process as close as possible to the one in the 2015 sample. We computed the proportion of turnover made by the exhaustive units in each business sector and then used this proportion in order to realise a cut-off, i.e. selecting the biggest enterprises one by one, as long as their combined share of turnover is lower than the 2015 proportion. This approach led to 95 000 legal units in take-all strata, or 70 000 with the cut-off rule within the enterprises. However, this also led to a huge imbalance between sectors, as in some cases this approach implies selecting numerous small units, whereas in other sectors there is only one big unit selected by this cut-off: these discrepancies mean that the take-all strata vary widely in terms of economic weight of legal units between the sectors.

The solution adopted here is to combine these two approaches. That means that we apply the same criteria as before, but we only select the biggest enterprises under a threshold defined by the proportion of economic weight of take-all strata in 2015. This approach limits the number of legal units selected while also prohibiting the accumulation of small enterprises to reach the cut-off threshold. However, this method is only possible on the scope of ESA, because the EAP survey procedure is defined by the ProdCom regulation (see Section 1 for more information), and therefore a 85% cut-off in each sector has to be achieved. This results in about 70 000 legal units taken exhaustively (using the cut-off rule within the enterprises), 40 000 for ESA and 30 000 for EAP.

We add some missing legal units which are relevant to the take-all strata of ESA (defined by a combination of cover rate and actual thresholds of turnover) and EAP (defined by the ProdCom regulation):

ESA : Units within the groups which are not selected by the cut-off rule but have more than 200 employees, or with a very important turnover. Selecting 1 000 legal units among the 20 000 legal units which were put aside initially with the cut-off rule leads to regain half of the turnover.

EAP : Units within the groups which are not selected by the cut-off rule but with more than 20 employees or with a turnover exceeding 5M€ are put into take-all strata. This leads to the selection of 700 more legal units among 3 000 legal units which were not selected initially.



Moreover, the enterprises consisting of more than 20 legal units, with more than 200 employees or with a turnover of more than 50M€ which were not yet included in the take-all strata were added. Indeed, keeping these enterprises in the take-some strata would have implied an excessive variance.

As the ProdCom regulation concerns mainly legal units, we have to assure that our sample respects the criteria imposed by this regulation. That means that we have to build take-all strata of legal units with the same rules as before: 5 M€ of turnover or more, 20 or more employees and a cut-off rule of 85% of turnover in every business sector.

In a subsequent step, we confront these two sets of take-all units used for EAP, the one on the legal unit level (denoted TA_{PC}) and the one on the enterprise level (denoted TA_{SBS}). 92% of the legal units in TA_{PC} also belong to TA_{SBS}. The other 2 000 legal units are required for TA_{PC} but not originally included for the SBS regulation. However, it is possible to include some of these units to TA_{SBS}:

- Independent legal units are included in TA_{SBS}.
- Enterprise whose legal units all belong to TA_{PC} are also added to TA_{SBS}.

These units are surveyed anyway, so it makes sense to include them in TA_{SBS}, because we do not want to sample them in the following steps of the procedure. This is not the case for enterprises with at least one legal unit that is not in TA_{PC}. Indeed, we have to be able to sample these enterprises in order to avoid undercoverage of the scope of the survey.

Finally, 500 legal units selected through the constraint issued by ProdCom regulation do not belong to the take-all strata defined by the SBS regulation. These units are selected with a sampling weight equal to one. This sampling weight would be used in the case of a study at the legal unit level.

2.3 Stratification and domains of interest

The take-some strata are defined by crossing the business sector of the French classification in five positions[7] with the number of employees in each enterprise. Nine strata are defined for the number of employees : 0 employee / 1-5 employees / 6-9 employees / 10-19 employees / 20-29 employees / 30-49 employees / 50-99

[7]This classification is a sub-classification of the European classification in four positions

employees / 100-199 employees / 200 employees or more.
These strata are the ones already in use, but they could have been chosen in a different way. For example, the strata boundaries of the number of employees could have been determined by an optimal categorisation (more discussion in Section 4).

Two domains of interest are considered :

- The business sectors of the French classification in five positions[8]
- The intersection between the business sectors in three positions and the number of employees aggregating the strata with less than 10 employees, the ones between 10 and 49 employees and the ones between 50 and 199 employees.

2.4 Allocations

The allocations in each strata are calculated using a Neyman allocation on the turnover of the enterprise integrating local constraints on precision on the domains of interest [9]. The advantage of this algorithm in comparison to the classical Neyman allocation is that we can add the constraint of a maximal local CV on the domains of interest.

Since data remain collected on legal units, the survey cost depends on the number of legal units to survey (116 000 units for ESA and 35 000 for EAP). Therefore, we extend the algorithm presented by Koubi *et al.* [9] by introducing costs in the Neyman allocation .

If we denote y_k the turnover of the enterprise k, $\hat{t}_{y\pi}$ the Horvitz Thompson estimator for the total of turnover, $S^2_{y,h}$ the empirical variance of y_k in stratum h, N_{LU} the number of legal units to be drawn in the scope of one survey (ESA or EAP), $N_{LU,k}$ the number of legal units of enterprise k in the same scope, n_h the number of enterprises to survey, N_h the number of enterprises and $f_h = n_h/N_h$ the sampling rate in stratum h, $C_h = \bar{N}_{LU,h} = (1/N_h)\sum_{k\in U_h} N_{LU,k}$ the cost, i.e. the mean number of legal units per enterprise in stratum h, D the whole range of domains of interest and CV_{loc} the local precision we expect, we have to resolve:

$$\begin{cases} \min\limits_{n_1,\dots,n_H} \mathbb{V}_p[\hat{t}_{y\pi}] = \sum\limits_{h=1}^{H} N_h^2 \dfrac{1-f_h}{n_h} S^2_{y,h} \\ u.c. \sum\limits_{h=1}^{H} C_h n_h = N_{LU} \\ u.c.\ n_h \leq N_h \\ u.c. \max\limits_{d\in D} CV_d \leq CV_{loc} \end{cases}$$

As we cannot combine the two different domains of interest in the same Neyman allocation, we calculate both and compare them (see Section 3).

2.5 Variability of the number of legal units to survey

We introduced in Section 2.4 a mean cost per stratum in the Neyman allocation, that leads to the good number of legal units to survey on average N_{LU}. But the number of legal units to be drawn remains random and varies from one sample to another. We will now take an interest in the variability of this number.

[8]The principal activity is determined according to the breakdown of the company's various activities. As the added value of different activity branches is often hard to determine based on statistical surveys, it is the breakdown of turnover or of the workforce according to the branches that is used as a determination criterion.

The Horvitz Thompson estimator of N_{LU} is:

$$\hat{N}_{LU} = \sum_{h=1}^{H} \sum_{k \in S_h} N_{LU,k} = \sum_{h=1}^{H} \sum_{k \in S_h} \frac{z_k}{\pi_k} \qquad \text{with} \quad z_k = \pi_k N_{LU,k} \quad \text{and} \quad \pi_k = \frac{n_h}{N_h}$$

The expectation of this estimator can expressed as:

$$\begin{aligned}
\mathbb{E}_p\left[\hat{N}_{LU}\right] &= \sum_{h=1}^{H} \sum_{k \in U_h} N_{LU,k} \mathbb{E}_p\left[I_k\right] \qquad \text{with} \quad I_k = \mathbb{1}(k \in S_h) \\
&= \sum_{h=1}^{H} \sum_{k \in U_h} N_{LU,k} \mathbb{P}(k \in S_h) = \sum_{h=1}^{H} \sum_{k \in U_h} N_{LU,k} \pi_k \\
&= \sum_{h=1}^{H} \sum_{k \in U_h} N_{LU,k} \frac{n_h}{N_h} = \sum_{h=1}^{H} n_h \bar{N}_{LU,h} = N_{LU}
\end{aligned}$$

In the case of a stratified simple random sampling[9], the variance of this estimator can expressed as:

$$\mathbb{V}_p\left[\hat{N}_{LU}\right] = \sum_{h=1}^{H} N_h^2 \frac{1 - f_h}{n_h} S_{z,h}^2 \,,$$

with $S_{z,h}^2 = \dfrac{1}{N_h - 1} \sum_{k \in U_h} \left(z_k - \mu_{z,h}\right)^2$ the empirical variance of z_k in stratum h.

We know that $z_k = \pi_k N_{LU,k}$ and $\pi_k = n_h / N_h$. We can write:

$$\mu_{z,h} = \frac{1}{N_h} \sum_{k \in U_h} z_k = \frac{n_h}{N_h} \frac{1}{N_h} \sum_{k \in U_h} N_{LU,k} = \frac{n_h}{N_h} \bar{N}_{LU,h}$$

In this case, the empirical variance of z_k in stratum h can be rewritten:

$$S_{z,h}^2 = \frac{1}{N_h - 1} \left(\frac{n_h}{N_h}\right)^2 \sum_{k \in U_h} \left(N_{LU,k} - \bar{N}_{LU,h}\right)^2 = \left(\frac{n_h}{N_h}\right)^2 S_{N_{LU},h}^2$$

Finally, the variance of the number of legal units to survey can be expressed as:

$$\mathbb{V}_p\left[\hat{N}_{LU}\right] = \sum_{h=1}^{H} n_h (1 - f_h) S_{N_{LU},h}^2$$

2.6 Efficiency boundaries

In order to find the best local precision on the two domains of interest, we calculate the minimum number of enterprises that should be drawn for different local coefficients of variation (CVs). We also calculate the global CVs that we would obtain for each local CVs.

[9] In order to comply with the sampling coordination method used at INSEE [6], the survey design must be a stratified simple random sampling.

For a given number of enterprises to survey n_{ent}, we call **efficiency boundary** the allocations $(n_1, \ldots, n_H)$ that cannot lead to a better local precision without a deterioration of the global precision. We can represent this boundary in a plot with the maximum (i.e, worst) local CVs on the x-axis and the global CVs on the y-axis.

The plot in Figure 1 represents the efficiency boundary for the first domain of interest: the business sectors of the French classification in five positions. As we could expect, the Neyman allocation without local constraints of precision (represented here with a cross) is a flat optimum. The global precision gets worse if one chooses very strong local precision. We can see that the best local precision without a considerable deterioration of the global precision could be a local CV of 5% for ESA and 2% for EAP.

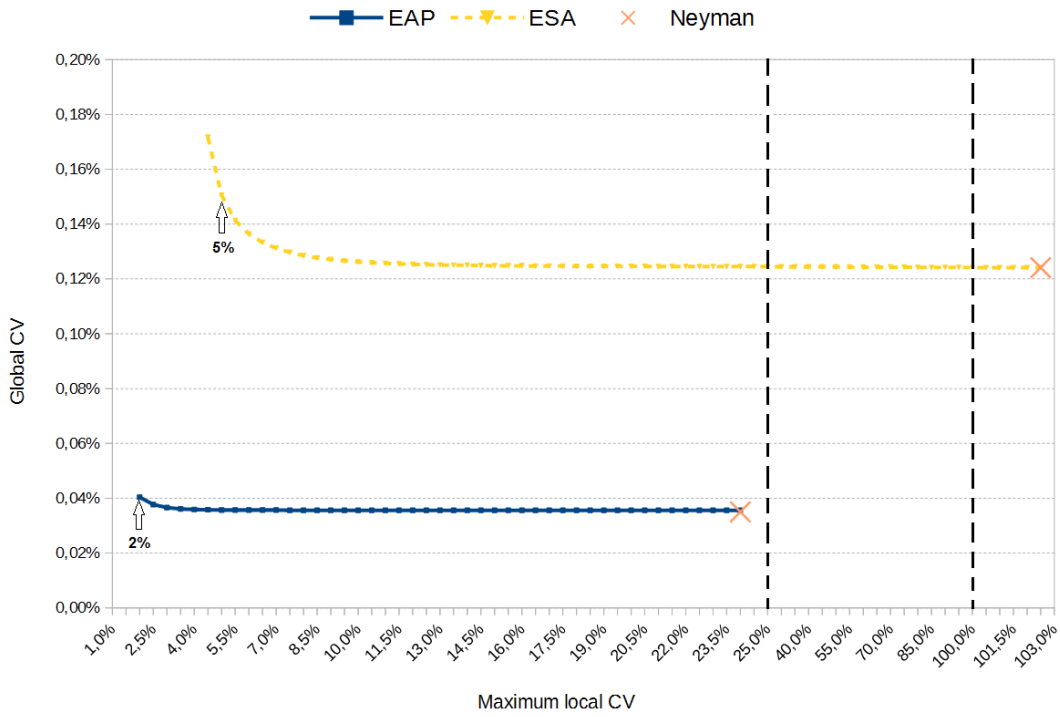


Figure 1: Efficiency boundary for the sectors of the French classification in five positions

The plot in Figure 2 represents the efficiency boundary for the second domain of interest: the intersection between the business sectors in three positions and the number of employees. We can see that the best local precision without a noticeable deterioration of the global precision could be a CV of 8% for ESA and 11% for EAP.

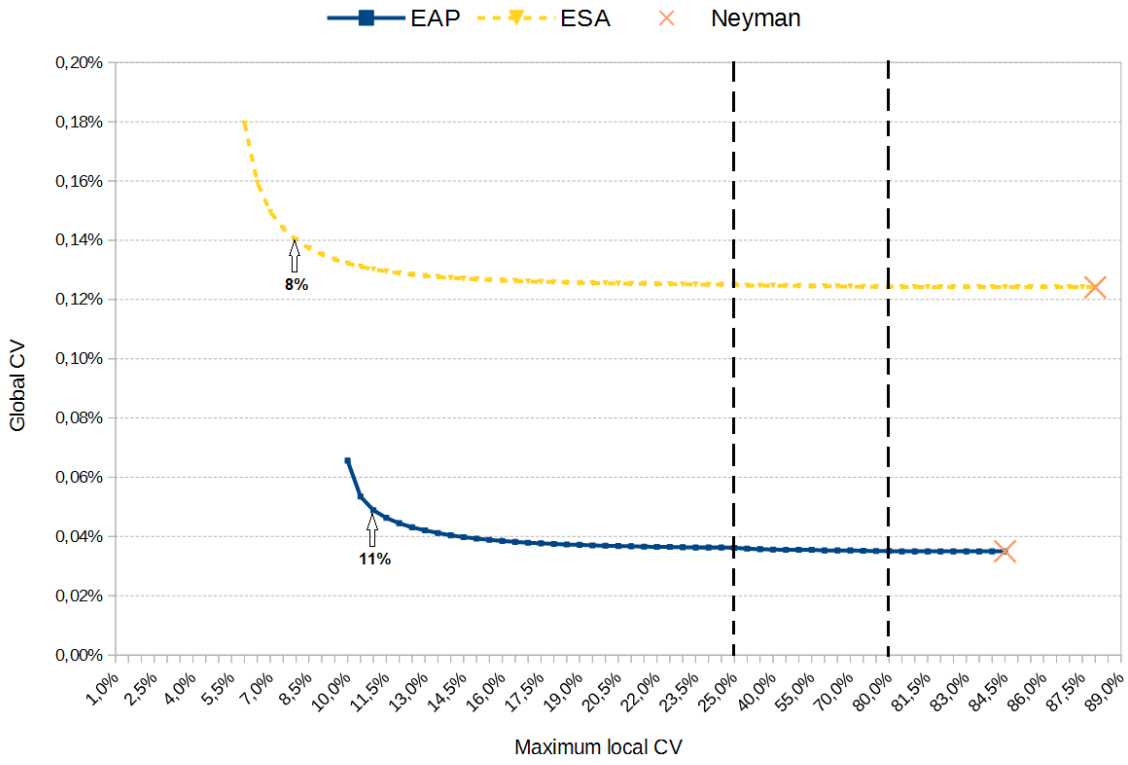


Figure 2: Efficiency boundary for the intersection between the sectors in three positions and the number of employees per enterprise

3. Results

3.1 Number of enterprises to draw

For the first domain of interest (business sectors in five positions), in order to get the best local CVs and the good number of legal units to survey on average, we have to draw $n_{ent,1} = 109\,900$ enterprises, with 27 000 enterprises from EAP's scope and 82 900 enterprises from ESA's scope.

For the second domain of interest (business sectors in three positions crossed with the number of employees), we have to draw $n_{ent,2} = 109\,500$ enterprises, with 27 000 enterprises from EAP's scope and 82 500 enterprises from ESA's scope.

We also calculated the mean between these two allocations, in order to get a "mix" between the best local precision on each domain of interest. We will discuss the results of this approach in Section 3.3. This "mixed" allocation leads to draw $n_{ent,mix} = 109\,700$ enterprises, with 27 000 enterprises from EAP's scope and 82 700 enterprises from ESA's scope.

All these values for n_{ent} ($n_{ent,1}$, $n_{ent,2}$, $n_{ent,mix}$) lead to the selection of approximately 35 000 legal units from EAP's scope and 116 000 legal units from ESA's scope on average.

3.2 Variability of the number of legal units to survey

If we now have a look on the variability of these results, using the formula in Section 2.5 for the variance of the number of legal units to be drawn, we can see in Table 1 that the variability is very low in general and for each survey's scope. The results are approximately the same for all the n_{ent} described above. We present here the results for the "mixed" allocation.

Table 1: Confidence intervals for the number of legal units to survey

	Total	ESA	EAP
$n_{ent,mix}$	109 700	82 700	27 000
$\mathbb{E}_p\left[\hat{N}_{LU}\right]$	151 000	116 000	35 000
$CI_{95\%}\left(N_{LU}\right)$	[150 830 ; 151 170]	[159 840 ; 116 160]	[34 970 ; 35 030]

Another result we have to check was the number of legal units that would be drawn in aggregated sectors. In fact, the legal units drawn in a sample are treated by different teams, depending on the business sector. We have to check whether there are substantial changes in the number of legal units per aggregated sector.

The result is that this survey design at an enterprise level increases the number of legal units to treat in the trade activities and decreases this number in the service activities. This remains true for all values of n_{ent} obtained above. The variability of these results in each aggregated sector is very low.

We also notice that this survey design leads to survey slightly more legal units with 1 to 5 employees, and slightly less legal units with 30 to 49 employees, and that for all three n_{ent}. The variability of these results in each strata of enterprise size is also very low.

We could have expected that the number of legal units with less than 10 employees to survey would be different between the allocations $n_{ent,1}$ and $n_{ent,2}$. Indeed, we could have imagined that it was necessary to survey more enterprises with less than 10 employees in $n_{ent,2}$ than in $n_{ent,1}$ in order to improve the precision in the strata of small enterprises. But in fact, if we compare $n_{ent,1}$ to $n_{ent,2}$ in these strata, the number of enterprises to survey increases (sometimes only marginally) in the ones with bad precision (see Section 3.3) and decreases in the ones with good precision.

3.3 Precision at the enterprise level

As explained in Sections 3.1 and 3.2, all three n_{ent} give approximately the same results for:

- The number of legal units in aggregated sectors
- The number of legal units in each strata of enterprise size
- The variability of the number of legal units to be drawn

In order to find the best allocation between $n_{ent,1}$ (allocation considering the business sector in five position as the domain of interest), $n_{ent,2}$ (allocation considering the intersection between the business sector in three position and the number of employees as the domain of interest), and $n_{ent,mix}$ (mix between $n_{ent,1}$ and $n_{ent,2}$), we calculate in Table 2 the precision at the enterprise level on the domains of interest:

- Business sector in five position with the Neyman allocation $n_{ent,2}$ and $n_{ent,mix}$
- Intersection between the business sector in three position and the number of employees with the Neyman allocation $n_{ent,1}$ and $n_{ent,mix}$

Table 2: Distribution of local CVs of the total of turnover depending on the allocation and the domain of interest (without the take-all strata of units with more than 200 employees for the second domain of interest).

	Domains of interest					
	Business sectors in five positions			Sectors in three positions × number of employees		
Levels	$n_{ent,1}$	$n_{ent,2}$	$n_{ent,mix}$	$n_{ent,1}$	$n_{ent,2}$	$n_{ent,mix}$
100% Max	5%	74,4%	23,1%	89,3%	11%	43,1%
90%	5%	9%	6,3%	20,8%	11%	12,5%
75% Q3	5%	4,9%	4,4%	9,2%	8%	8,9%
50% Median	2%	2%	2%	4,2%	4,6%	4,2%
25% Q1	0,9%	0,8%	0,8%	0,1%	0,2%	0,2%
10%	0,2%	0,1%	0,2%	0%	0%	0%
0% Min	0%	0%	0%	0%	0%	0%
Column number	*1*	*2*	*3*	*4*	*5*	*6*

As we could have expected, the "mixed" allocation seems to do better on both domains of interest at the same time (columns 3 and 6), in comparison to the precision we obtain if the domain of interest used to calculate the Neyman allocation *ex ante* is different from the domain of interest *ex post* (columns 2 and 4).

On the other hand, the "mixed" allocation degrades the precision if the two domains of interest are the same (columns 1 and 5). Indeed, the maximum local CVs is equal to 5% for the business sectors in five positions and 11% for the intersection between the business sectors in three positions and the number of employees, as we could expect from the precision seen in Section 2.6. However, this degradation concerns only the domains of interest with the highest 10% local CVs.

Moreover, the differences of precision between these three allocations concern only the domains of interest of the last quartile of the distribution of the local CVs. For all three n_{ent}, the value of the third quartile is close to 5% for the business sectors in five positions and 8-9% for the intersection between the business sectors in three positions and the number of employees.

3.4 Precision at the legal unit level

Some users of the business data (the National Accounts for example) still use the information at the legal unit level. In this case, the structure of the enterprise is not taken into account and the weight of a legal unit corresponds to the weight of the enterprise it belongs to. In this context, some legal units with similar characteristics have different weights, which would lead to a higher weight dispersion. We compare in Table 3 the precision at:

- The enterprise level using the "mixed" allocation ($n_{ent,mix}$)
- The legal unit level with the new survey design using the "mixed" allocation ($n_{LU,mix}$)
- The legal unit level using the allocation of the 2015 ESA and EAP survey designs ($n_{LU,2015}$)

If we denote y_k the turnover of the legal unit k, $\hat{t}_{y\pi}$ the Horvitz Thompson estimator for the total of turnover at the legal unit level, n_h the number of enterprises to survey, N_h the number of enterprises and $f_h = n_h/N_h$ the sampling rate in stratum h, $Y_g = \sum_{k \in g} y_k$ the sum of the turnover of the legal units of an enterprise g,

$\bar{Y}_h = (1/N_h) \sum_{g \in U_h} Y_g$ the empirical mean of Y_g in stratum h, the variance of $\hat{t}_{y\pi}$ with the two-stage cluster sampling is obtained using the following formula:

$$\mathbb{V}_p[\hat{t}_{y\pi}] = \sum_{h=1}^{H} N_h^2 \frac{1 - f_h}{n_h} S_{Y,h}^2 \qquad \text{with} \quad S_{Y,h}^2 = \frac{1}{N_h - 1} \sum_{g \in U_h} \left(Y_g - \bar{Y}_h\right)^2$$

Table 3: Distribution of local CVs for the total of turnover at the legal unit level depending on the survey design and the domain of interest (without the take-all strata of units with more than 200 employees for the second domain of interest).

	Domains of interest			
	Business sectors in five positions		Sectors in three positions × number of employees	
Levels	$n_{LU,mix}$	$n_{LU,2015}$	$n_{LU,mix}$	$n_{LU,2015}$
100% Max	14,9%	47,4%	38,3%	48,5%
90%	5,9%	7,5%	10,6%	12,8%
75% Q3	3,9%	3,8%	7,3%	5,4%
50% Median	2,1%	1,8%	3,3%	1,3%
25% Q1	0,9%	0,6%	0,6%	0%
10%	0,2%	0,1%	0%	0%
0% Min	0%	0%	0%	0%
Column number	*1*	*2*	*3*	*4*

The distribution of the local CVs for the total of turnover at the legal unit level with the "mixed" allocation (columns 1 and 3) is similar to the precision we currently have with the survey design at the legal unit level (columns 2 and 4), and that for both domain of interest. Indeed, the precision at the legal unit level with the new survey design is better in 50% of the cases than the precision we currently have with the actual survey design, and is worst also in 50% of the cases (see Figure 3). However, the two-stage cluster sampling leads to a better precision (i.e. lower local CVs) for the highest values of the local CVs with the current survey design (e.g. more than 40%).

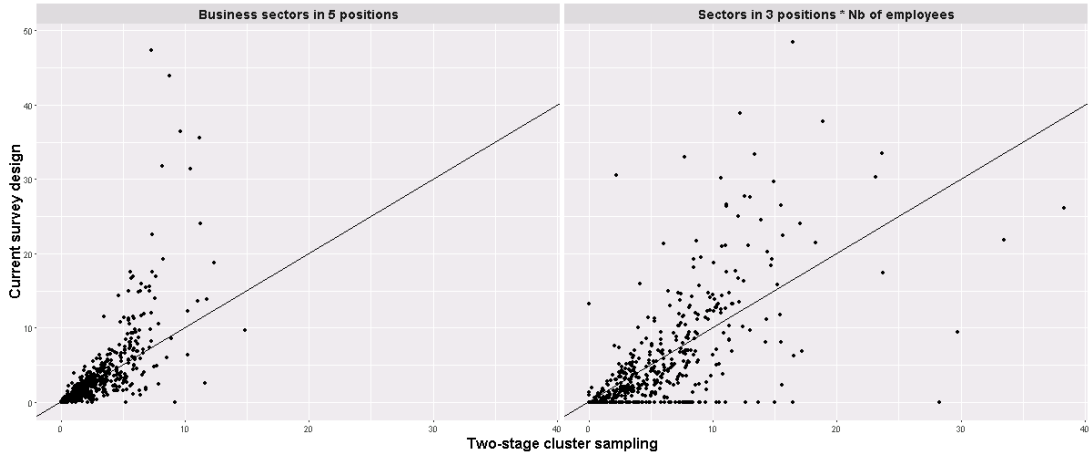


Figure 3: Local CVs for the total of turnover at the legal unit level by Business sectors in five positions (first plot on the left) and for the intersection between the sectors in three positions and the number of employees (second plot on the right) depending on the survey design.

4. Discussion

In this study, we assessed the impact of changes in business surveys that are now based on the sampling of enterprises instead of the sampling of legal units on statistical inference. Our aim is to optimize the survey designs in the resulting two-stage cluster sampling in order to have a good precision of estimators while respecting the constraint of a limited number of legal units selected.

The definition of the take-all strata leads to consider approximately the same amount of legal units in the exhaustive part of the sample as in 2015. The variability of the number of legal units to be drawn in the take-some strata is small for all allocations considered. However, the different allocations yield a different precision at the enterprise level. The variance of the estimator resulting of this optimised survey design was similar to the current one.

To improve the stratification of the survey design (see 2.3), one could define an optimal categorization of the number of employees per enterprises using the Dalenius method [3], the geometric method proposed by Gunning and al. (2004) [7], or the Lavallée-Hidiroglou method [10]. The latter method could also be applied to find an optimal threshold of the turnover in each activity for the definition of the take-all strata.

Instead of using the weights $(1/2; 1/2)$ for the calculation of the "mixed" allocation (see 3.1 and 3.2), it would be advisable to find optimal factors $(\alpha, 1-\alpha)$, as discussed in Merly-Alpa et al. (2016) [11].

The stratification defined in 2.3 leads to strata with only one unit, or with more than one unit but with an empirical variance of the turnover equal to zero. In these cases, the strata are considered as take-all strata. Another idea could be to aggregate these strata, for example via a hierarchical cluster analysis, in order to minimize the variance of the turnover within the aggregated strata and maximize the variance between the aggregated strata.

Each year, the French public statistical system carries out a considerable number of businesses surveys. To reduce the administrative burden of small businesses, a renewed sampling coordination method [6] was introduced in the late 2013 at INSEE. The aim of this coordination of samples is to favor the selection of businesses which have not been surveyed recently, while guaranteeing the unbiasedness of the samples. This method can handle the sample coordination between surveys based on different unit levels (legal units, local units). However, this method does not account for the nested nature of legal units within enterprises, and in particular the fact that the selection of an enterprise leads to the interrogation of all its legal units.

When we are interested in the evolution of a parameter and when the surveys are repeated every year, we usually do not redraw the whole sample for every edition of the survey, but define instead a rotating survey design [4]. With the new survey design at the enterprise level, we will have to decide whether a units belongs to the retained part or to the renewed part of the sampling frame (and of the sample) when the composition of the enterprise changes from one year to another. Indeed, some legal units could change their status between two years, they could for example move to another group or become independent.

Finally, this paper is mainly focusing on the survey design. This is the first step of the production, but ensuring the quality of the surveys requires a lot of post-treatments, such as non-response weight adjustment [2], calibration and winsorization of outliers [5]. All these methods are widely known and discussed, but their application to this survey, while using the economic structure of enterprises, needs to be studied. An issue of particular interest, which will have to be treated in the future, is the question of the correlation between non-response of legal units within enterprises.

Acknowledgement

The authors would like to thank Thomas Deroyon, Sébastien Faivre and Emmanuel Gros for their useful comments and suggestions that helped improving the quality of the paper significantly.

References

[1] 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. *Official Journal*, 46:1–11, 2003.

[2] Philippe Brion and Emmanuel Gros. Statistical estimators using jointly administrative and survey data to produce French structural business statistics. *Journal of Official Statistics*, 31(4):589–609, 2015.

[3] Tore Dalenius and Joseph L Hodges Jr. Minimum variance stratification. *Journal of the American Statistical Association*, 54(285):88–101, 1959.

[4] Elvire Demoly, Arnaud Fizzala, and Emmanuel Gros. Méthodes et pratiques des enquêtes entreprises à l'insee. *Journal de la Société Française de Statistique*, 155(4):134–159, 2014.

[5] Thomas Deroyon. Traitement des valeurs atypiques d'une enquête par winsorization-application aux enquêtes sectorielles annuelles. *JMS, Paris*, 2015:1, 2015.

[6] Emmanuel Gros. The procedure of sampling coordination for business surveys implemented at INSEE: methodology and practice. *ICES-V, Geneva*, 2016:1, 2016.

[7] Patricia Gunning and Jane M Horgan. A new algorithm for the construction of stratum boundaries in skewed populations. *Survey Methodology*, 30(2):159–166, 2004.

[8] Olivier Haag. The French business registers system: How to improve the quality of the statistics by combining different statistical units. *ICES-V, Geneva*, 2016:1, 2016.

[9] Malik Koubi and Sandrine Mathern. Résolution d'une des limites de l'allocation de Neyman. *JMS, Paris*, 2009:1, 2009.

[10] Pierre Lavallée and Michael A Hidiroglou. On the stratification of skewed populations. *Survey methodology*, 14(1):33–43, 1988.

[11] Thomas Merly-Alpa and Antoine Rebecq. Optimisation d'une allocation mixte. *preprint*, 2016.

[12] Olivier Sautory. Les enjeux méthodologiques liés à l'usage de bases de sondage imparfaites. In *Recueil du Symposium 2013 de Statistique Canada*.