THE IMPACT OF PROFILING ON SAMPLING: HOW TO OPTIMIZE SAMPLE DESIGN WHEN STATISTICAL UNITS DIFFER FROM DATA COLLECTION UNITS

Emmanuel Gros [1] & Ronan Le Gleut [2]

[1] *Insee, 18 boulevard Adolphe Pinard 75014 PARIS,emmanuel.gros@insee.fr*
[2] *Insee, 18 boulevard Adolphe Pinard 75014 PARIS, ronan.le-gleut@insee.fr*

Abstract

In France as in many other countries, business statistics are undergoing great changes. Until now, business surveys were based on the observation of legal units that have a juridical definition. From now on, business statistics will be more and more based on the economical 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. To this end, an important methodological operation of "profiling" – which consists on one hand in a manual delineation of enterprises within complex business groups and on a other hand to consider the other groups as one enterprise – is ongoing at INSEE, the French NSI.

Since the statistical units (enterprises) are now different from the data collection units (legal units), the sample design can be seen as a two-stage cluster sampling. Enterprises are randomly selected, and then all legal units within those enterprises are included in the sample. The main drawback could be a loss of precision due to the similarity of units in a cluster, but:

1. more than 95% of enterprises consist of a single legal unit
2. the legal units of an enterprise may have different activities

A further drawback to cluster sampling is that we cannot completely control the final sample size. If, at the enterprise level, we keep the sampling rates used for a survey design at the legal unit level, this may decrease the number of primary units drawn, while increasing the number of legal units to be surveyed.

This paper present how the sample design of the French structural business survey was optimized, in order to have a good precision on estimators at the enterprise level under a constraint pertaining to the number of surveyed legal units.

Keywords. Profiling, stratified cluster sampling, survey design optimisation.

1 Introduction and context

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" is ongoing at INSEE, which has a major impact on the Esane[1] process, which produces structural business statistics in order to answer the SBS European regulation, using both survey and administrative data (Brion and Gros 2015).

1.1 The French structural business surveys

In the Esane device, two structural business surveys are used to produce structural business statistics:

- the ESA (Annual Sectoral Survey), which scope concerns activities of trade, construction, services and transport. The sample is very large, with almost 116 000 legal units surveyed each year in Metropolitan France.
- the EAP (Annual Production Survey), which scope concerns manufacturing industry. The sample is composed of about 35 000 units in Metropolitan France.

The purpose of these two surveys is to deduce the main activity of a company by breaking down its turnover into activities (sectoral classification). Until reference year 2015, these two surveys were drawn according to a stratified simple random sampling of legal units.

1.2 The SBS European Regulation

The Structural Business Statistics (SBS) regulation (Regulation 1993) cover industry, construction, distributive trades and services. Presented according to the NACE activity classification, they describe the structure, conduct and performance of businesses across the European Union. These statistics can be broken down to a very detailed sectoral level, as well as according to the size of enterprises. In order to comply with this regulation, business statistics produced by ESA and EAP surveys will be based on the economic notion of enterprise instead of the juridical definition of a legal unit.

2 Methods

Since the statistical units (enterprises) are now different from the data collection units (legal units), the sample design can be now seen as a stratified cluster sampling. As a cluster, an enterprise is randomly selected, and then all legal units within this enterprise are included in the sample.

2.1 Definition of the take-all strata

We keep the same historical thresholds on the number of employees and the turnover, but modulate them by a coverage rate of the turnover to reach in each business sector. Enterprises composed of

1. ESANE for the French Élaboration des Statistiques Annuelles d'Entreprise.

more than 20 legal units, with more than 200 employees or with more than 50M€of turnover are automatically included in the take-all strata.

In order to decrease the number of legal units in the take-all strata, we add a cut-off rule within each enterprise. This rule avoids to survey legal units with a very low turnover, for which we can suppose that they only have one activity.

This results in about 70 000 legal units taken exhaustively, 40 000 for ESA and 30 000 for EAP.

2.2 Stratification and domains of interest

The take-some strata are defined by crossing the business sector of the French classification in five positions[2] with the number of employees in each enterprise, in nine classes.

Two domains of interest are considered :

- The business sectors of the French classification in five positions: Activities.
- 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.

2.3 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 (Koubi and Mathern 2009). 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 and Mathern (2009) 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. This classification is a sub-classification of the European classification in four positions.

2.4 Variability of the number of legal units to survey

We introduced in Section 2.3 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:

$$\hat{N}_{LU} = \sum_{h=1}^{H} \sum_{k \in S_h} N_{LU,k}$$

We can rewrite this quantity as the Horvitz-Thompson estimator of the variable z_k:

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

which is an unbiased estimator of N_{LU}.

In the case of a stratified simple random sampling, the variance of this estimator can finally be expressed as:

$$\mathbb{V}_p \left[\hat{N}_{LU} \right] = \sum_{h=1}^{H} n_h (1 - f_h) S^2_{N_{LU},h}, \text{ with } S^2_{N_{LU},h} = \frac{1}{N_h - 1} \sum_{k \in U_h} \left(N_{LU,k} - \bar{N}_{LU,h} \right)^2$$

2.5 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.

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.

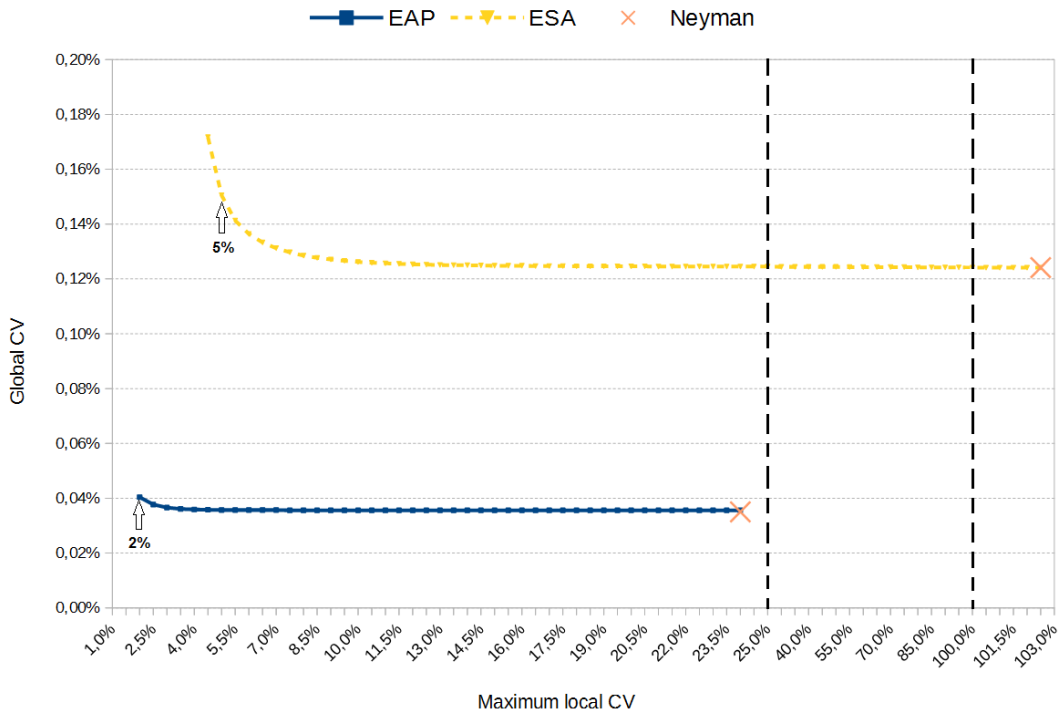

Figure 1: Efficiency boundary for the sectors of the French classification in five positions

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.

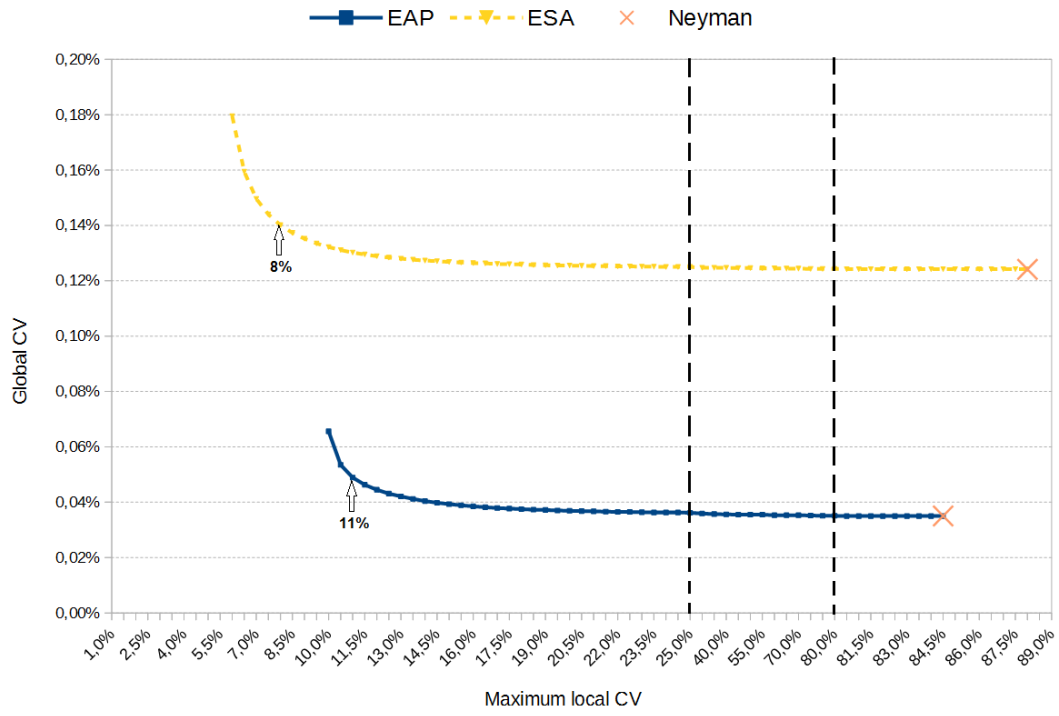

Figure 2: Efficiency boundary for the intersection between the sectors in three positions and the number of employees per enterprise

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 2.4 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]	[115 840 ; 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.

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 2.5. 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 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%).

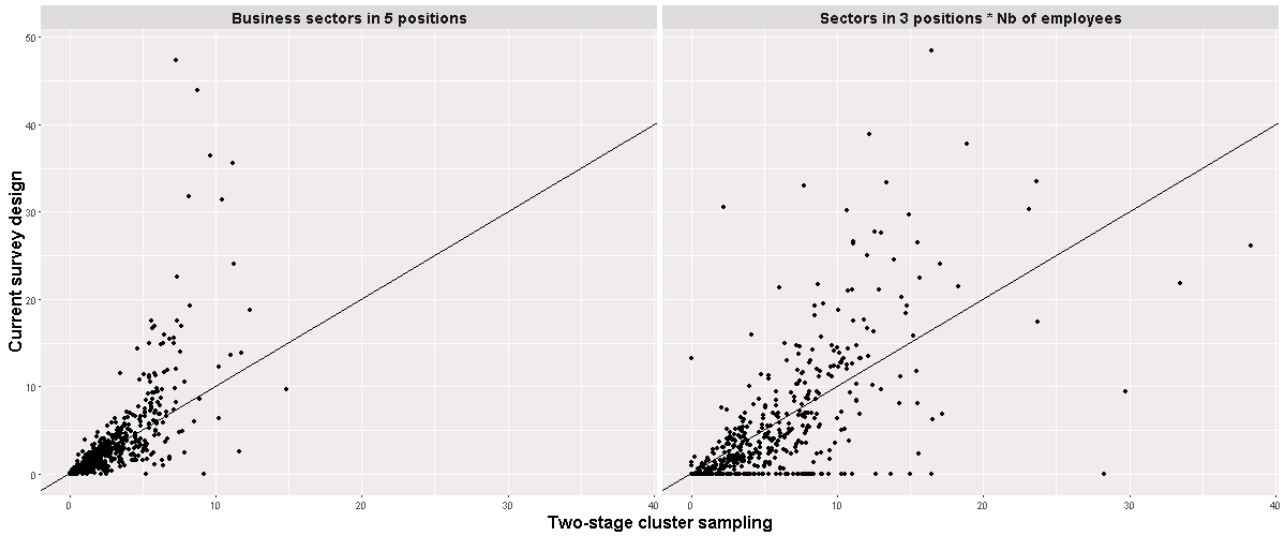

Figure 3: Local CVs for the total of turnover at the legal unit level by Business sectors in five positions (first plot on the left) and for the intersection between the sectors in three positions and the number of employees (second plot on the right) depending on the survey design.

4 Conclusion and future works

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 2.2), one could define an optimal categorization of the number of employees per enterprises using the Dalenius method (Dalenius and Hodges Jr 1959), the geometric method proposed by Gunning and Horgan (2004), or the Lavallée-Hidiroglou method (Lavallée and Hidiroglou 1988). 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 and Rebecq (2016).

Between the selection of the sample and the dissemination of the results, the "perimeter" of an enterprise (i.e. the legal units that belong to this enterprise) can change. For example, a legal unit of an enterprise A can belong to another enterprise B one year later or can become an independent legal unit. This problem can be seen as a particular case of indirect sampling: the sample is drawn in a population of enterprises (with a certain "perimeter") which differs from the population of interest (for the dissemination of the results), but which is linked to this one *via* its legal units. In this context, the generalised weight share method proposed by Deville and Lavallée (2006) would allow us to handle this problem.

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 (Brion and Gros 2015), calibration and winsorization of outliers (Deroyon 2015). 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.

References

Brion, P., and E. Gros. 2015. "Statistical Estimators Using Jointly Administrative and Survey Data to Produce French Structural Business Statistics." *Journal of Official Statistics* 31 (4): 589–609.

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

Deroyon, T. 2015. "Traitement des valeurs atypiques d'une enquête par winsorization - application aux enquêtes sectorielles annuelles." *12èmes Journées de Méthodologie Statistique, Paris.*

Deville, J.C., and P. Lavallée. 2006. "Indirect sampling: The foundations of the generalized weight share method." *Survey Methodology* 32 (2): 165.

Gunning, P., and J.M. Horgan. 2004. "A new algorithm for the construction of stratum boundaries in skewed populations." *Survey Methodology* 30 (2): 177–185.

Koubi, M., and S. Mathern. 2009. "Résolution d'une des limites de l'allocation de Neyman." *10èmes Journées de Méthodologie Statistique, Paris.*

Lavallée, P., and M.A. Hidiroglou. 1988. "On the stratification of skewed populations." *Survey Methodology* 14:35–45.

Merly-Alpa, T, and A. Rebecq. 2016. "Optimisation d'une allocation mixte." *9ème colloque francophone sur les Sondages, Gatineau.*

Regulation, EEC. 1993. "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.