Profiling: a New Way to Increase the Quality of Statistics on Research and Development

Thomas Balcone [1] | *Ministry for Higher Education, Research and Innovation, Paris, France*

Abstract

Currently, statistics on Research and Development (R&D) carried out in the business sector are computed in France on the sole basis of legal units: firstly, a survey is addressed to them to collect the data and then, statistics on R&D are disseminated at legal unit level. Considering the increasing importance of the enterprise group in the French economy, it seems difficult today to go on using only the legal units to calculate business statistics. Indeed, assimilating the legal unit to the enterprise is not relevant anymore for group's affiliates and subsidiaries. Taking into account the European definition of an enterprise will help to disseminate more consistent and relevant R&D statistics on the business sector.

The French business statistical register established by the French National Statistical Institute (INSEE), called SIRUS, contains notably all the legal units and all the enterprises. The main contribution of this register is to make possible the calculation and dissemination of statistics at another level than the legal unit one: the enterprise level.

This article first describes why the data should go on being collected at the legal unit level and not at the enterprise one. Indeed, it seems that such a change in the data collection can be dangerous because it could result in a substantial increase of the response burden. Then, this article presents the process based on SIRUS that leads to the computation of key indicators on R&D at enterprise level. To conclude, it compares these key indicators with the ones calculated at the legal unit level to show the impact of moving to the enterprise level on French R&D statistics.[2]

INTRODUCTION

In the context of globalization, Research and development (R&D) is a major issue for firms and countries to stay competitive by bringing notably innovation. The globalization of corporate R&D has become

---

a considerable phenomenon in the last few years. Therefore, there is a huge need to collect consistent data on R&D to follow properly the evolution of R&D expenditure, notably in the business sector. This is the goal of the R&D survey, which is based on the Frascati Manual, reference work of the OECD which gives all the countries common method and definitions for the conduct of a R&D survey.

In France, this survey led by the Ministry of Higher Education, Research and Innovation since 1963, is based on questioning legal units. With the globalization and the more and more noticeable presence of big groups in the business sector, a new definition of the enterprise appeared in the European regulation,[3] more based on economy than legality: 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. Therefore, computing R&D data at this enterprise level seems to be more relevant to study R&D in the business sector and could become soon mandatory with the implementation of the future framework regulation for the business statistics FRIBS[4] which is currently under negotiation.

In the first part of this article, we will see how we can get data at the enterprise level from the R&D survey. Then, the article presents a method to compute key indicators on R&D at this new level. To conclude, it compares these key indicators with the ones calculated at the legal unit level to show the impact of moving to the enterprise concept on French R&D statistics.

1 HOW TO GET THE DATA AT THE ENTERPRISE LEVEL?
1.1 First approach: collect the data at the enterprise level

Currently, the data collection unit in the French R&D survey is the legal unit. The most natural idea to compute statistics on R&D at the enterprise level with the aim to disseminate more consistent and relevant statistics, is to choose the enterprise as the new data collection unit. However, such a change of data collection unit presents two major risks.

The first risk is a drop of our response rate which is currently over 90%. The legal unit considered as the decision-making unit of the enterprise is not necessarily the one which conducts the R&D activity, and in some large enterprises several legal units can be active in R&D. This can be an issue if we have only one contact in the decision-making unit and may probably increase the response burden for the enterprises. As a consequence, the response rate for large enterprises with several legal units active in R&D, which constitute a major part of the total internal expenditure on R&D, could drop considerably. Moreover, currently, with the legal unit as data collection unit, it's already hard for some contacts to collect data in all the establishments making up the legal unit.

The second risk is the one of no longer being able to do an analysis at the legal unit level. This kind of analysis is relevant to understand how the enterprises organize their R&D activity. For example, it is interesting to know if this activity is located in a sole legal unit (or in a dedicated R&D centre) or divided between several legal units. In the French R&D survey on the 2015 data, nearly 28% of the enterprises likely to carry out R&D activities[5] have more than one legal unit in the survey population, i.e. the population of the legal units likely to carry out R&D activities. Moreover, the legal unit concept is used in French national accounts and there will be no change in the near future. So, it is essential to keep computing consistent and relevant R&D statistics at the legal unit level.

In conclusion, choosing the enterprise as the new data collection unit does not seem the best way to get the best analysis of the R&D activities carried out by the business sector. So, we keep the legal unit as the data collection unit.

---

[3] The European regulation 696/93.
[4] Framework Regulation Integrating Business Statistics.
[5] An enterprise is likely to carry out R&D activities if at least one of its legal units is in the survey population.

1.2 Second approach: rebuild the enterprises' profiles from legal units

The setting up of the R&D survey population is atypical: contrary to the other surveys conducted in the business sector in France, this population does not come directly from the French business statistical register, called SIRUS, established by INSEE,[6] which contains notably all the legal units and all the enterprises. According to the recommendations of the Frascati Manual 2015[7] (cf. p. 220–221), the target population of the R&D survey at the legal unit level, noted *U(LU)*, is only all legal units likely to perform R&D. This population is built up every year by using the population of the last survey and several other sources linked to R&D or innovation listing legal units (tax credit claimants, innovation survey (CIS),[8] young innovative firms aids…). As the R&D activity carried out by an enterprise can be considered as all the R&D activities performed by its legal units, the target population at the enterprise level, noted *U(EP)*, is naturally all enterprises for which at least one legal unit is likely to perform R&D, i.e. belonging to the target population *U(LU)*.

Following the same logic, the sample at the enterprise level, noted *S(EP)*, is made by all enterprises for which at least one legal unit belongs to the sample *S(LU)*. To compute data on R&D at the enterprise level from the sample *S(EP)*, it is necessary to know the data for all the legal units belonging to an enterprise in the sample and to the target population *U(LU)*. Unfortunately, the data are not available for all these units because first of all, the sample of the R&D survey *S(LU)* does not come from a cluster sampling with the enterprise as the cluster. Indeed, as for most of the other surveys conducted in the business sector in France, the target population at the legal unit level *U(LU)* is divided into an exhaustive stratum and a non-exhaustive one. The exhaustive stratum is made up of the large[9] legal units and of the ones which appear for the first time in the target population. All the other legal units of the target population form the non-exhaustive stratum. In the 2015 survey, the respondent legal units correspond to 1 435 enterprises in *S(EP)* made up of more than one legal unit in the target population *U(LU)* (cf. Table 1). 1 090 of these enterprises are formed by at least one legal unit which is in the target population *U(LU)* but whose R&D

Table 1 The target population at the legal unit level *U(LU)* and the enterprise level *U(EP)* in the 2015 R&D survey

Population		Legal unit (LU) level	Enterprise level			
Target population		U(LU): 25 962	U(EP): 21 466			
Sample population	Respondent	10 552	8 855	*Made up of a sole LU in U(LU)*	7 420	
				Made up of more than one LU in U(LU)	All LU are respondent	345
					There is at least one LU with unknown R&D data	1 090
	Non respondent	1 007	894			
	Total	S(LU): 11 559	S(EP): 9 749			

Source: Ministry for Higher Education, Research and Innovation – 2015 R&D survey

---

[6] The French national statistical institute.

[7] The Frascati Manual is a manual from the OECD which gives the guidelines for collecting and reporting data on reasearch and experimental development.

[8] Community innovation survey.

[9] In the R&D survey, a large legal unit is a unit whose last known internal expenditure on R&D (BERD) is higher than 400 k€.

data are unknown either because it is not in the sample *S(LU)* (2 047 legal units), or because it did not answer to the survey (97 legal units). In the rest of the paper, an enterprise is considered respondent if at least one of its legal units has answered the R&D survey; otherwise, the enterprise is considered non respondent.

The second reason why data are not directly available for all the legal units belonging to an enterprise in *S(EP)* and to *U(LU)* is the existence of "grouped" responses in the survey, i.e. answers which are not about a sole legal unit only. Actually, in some situations, it was decided that some correspondents could answer for several legal units, independently of the concept of enterprise or group, to reduce its response burden.[10] These "grouped" responses, about a hundred or so each year, are matched with a "response outline" composed of several legal units. This kind of answer must be treated in a particular way to get the data for the legal units in the "response outline" and belonging to different enterprises of the sample *S(EP)*. In the 2015 survey, there are exactly 89 "grouped" responses (barely 1% of all the survey responses – cf. Table 2) which relate to 246 legal units and which represent 5.8 billion Euros of internal expenditures on R&D (BERD) (i.e. 18.3% of the total for the business sector). Among these 89 "grouped" responses, 29 mix several enterprises: they relate to 83 legal units and 63 enterprises.

Table 2 "Grouped" responses

Respondent population		Number	Legal unit level	Enterprise level
Non "grouped" responses		10 306	10 306	8 732
"Grouped" responses	A sole enterprise	60	163	60
	Several enterprises	29	83	63
Total		10 395	10 552	8 855

Note: Actually, this number is higher (8 782) because some enterprises have legal units in a "grouped" response and other ones not in "grouped" responses.
Source: Ministry for Higher Education, Research and Innovation – 2015 R&D survey

In conclusion, it is not so easy to compute R&D data for all the enterprises of the sample *S(EP)* from the data collected for the legal units sample *S(LU)*. Post-collection treatments are necessary.

1.3 The estimation of the internal expenditure on R&D (BERD) at the enterprise level

Within the framework of this study, we will consider only the internal expenditure on R&D (BERD) as R&D data. As seen in the previous paragraph, some enterprises in the sample *S(EP)* are made up of legal units in the target population *U(LU)* but whose BERD is unknown. We have to estimate this expenditure at the legal unit level to have a BERD estimator of good quality for such enterprises.

As said before, the "grouped" responses have to be treated in a particular way. As a consequence, firstly, we will concentrate on legal units whose BERD is unknown and which are not in a "grouped" response.

1.3.1 Estimation of the BERD for legal units except the ones in a "grouped" response

All the legal units in the target population *U(LU)* are likely to perform R&D but some of them do not. Unfortunately, this feature is not available for the legal units which are not in the sample S(*LU*) and for

---

[10] This may be due to strong collaboration on R&D activities between several legal units which may be close geographically, or to a unified accounting system (for two legal units in the same group for example).

the non-respondents. Then, a first step is to model the probability of performing R&D for each legal unit of the target population *U(LU)*. This modelling is based on all the respondent legal units except the ones in a "grouped" response (10 306 for the 2015 R&D survey – cf. Table 2), whether they have answered positively or not.[11] We distinguish four respondent sub-populations:

- the large legal units (QG[12] and exhaustive QS[13] – Large LU),
- the new legal units, i.e. the legal units which appear for the first time in the target population *U(LU)* (QS new – New LU),
- the legal units in the non-exhaustive stratum whose BERD in 2014 is unknown,
- the legal units in the non-exhaustive stratum whose BERD in 2014 is known.

For the large legal units, as the non-response and negatively response rates are low (respectively 0.15% and 3.94%), we assume that the probability of performing R&D, noted P(BERD>0) is equal to 1. For the other three respondent sub-populations, we estimate this probability thanks to a logistic regression. For each sub-population, the model is the following:

$$logit\left[P\left(BERD > 0\right)\right] = \beta_0 + \sum_{k=1}^{K} \beta_k X_k + \epsilon, \tag{1}$$

where:

- β_0 is the intercept,
- $\beta_1, ..., \beta_K$ are the coefficients related to the K explanatory variables $X_1, ..., X_K$
- ϵ is the error term.

The explanatory variables for the different sub-populations are mentioned in Table 3. For the sub-population "legal units in the non-exhaustive stratum whose BERD in 2014 is known", there is an additional variable in the model: a dummy variable which equals 1 if the legal unit answered positively in the 2014 survey (BERD>0), 0 otherwise (BERD=0). Then, we estimate the coefficients for each of the three models by using the corresponding respondent legal units. These estimates allow us to estimate the probability of performing R&D for each legal unit whose BERD is unknown and which is not in a "grouped" response. From this estimated probability, we define a "performing R&D" dummy variable, noted IR&D, as follows:

$$I_{R\&D} = \begin{cases} 1 \, if \, P\left(\widehat{BERD > 0}\right) > 0.5. \\ \quad 0 \, otherwise \end{cases} \tag{2}$$

For some of the legal units (629), we cannot compute such a dummy variable because there are missing values for some explanatory variables. In this case, we assume that the "performing R&D" dummy variable equals 0.

Afterwards, a second step is to estimate the internal expenditure on R&D (BERD) for the legal units for which the "performing R&D" dummy variable equals 1 (i.e. $I_{R\&D}$=1). If the legal unit answered positively one of the previous R&D surveys between 2009 and 2014, we assume that the BERD in 2015 equals the last BERD known and corrected by price changes. Otherwise, the BERD is estimated by using a linear regression based on the legal units which answered positively the 2015 R&D survey (8 169[14] legal

---

[11] A legal unit answers positively the 2015 R&D survey if it has performed R&D in 2015 (BERD>0). Otherwise, it answers negatively (BERD=0).

[12] General questionnaire: questionnaire for legal units whose BERD exceeds 2 000 k€.

[13] Simplified questionnaire. Exhaustive QS are intended for legal units whose BERD exceeds 400 k€ but not 2 000 k€.

[14] Actually, the model is based on 7 609 legal units due to missing values and outliers.

Table 3 The explanatory variables[15] in the logistic regression models

Explanatory variable	Values			
Turnover (in k€)	[0;200[	[200;1 120[	[1 120;5 700[	[5 700;+ ∞[
Number of employees (headcount)	[0;2]	[3;9]	[10;32]	[33;+∞[
Share of exports in the turnover (in %)	0	]0;5[	[5;20[	[20;100]
Unit legal age	[0;2[	[2;12[	[12;23[	[23;+∞[
Business sector	High		Technology industry	
	Medium-high			
	Medium-low			
	Low			
	Primary sector, energy, construction industry			
	Services sector (excluding R&D)			
	R&D (division 72 of the NACE Rev.2)			
Part of business group	Independent legal unit			
	Part of		a French group	
			a foreign group	
Region of location	Ile-de-France		Other regions	
Applying for R&D tax credit	Yes		No	

Source: Ministry for Higher Education, Research and Innovation

units – cf. Table 4). In the corresponding model (model (2)), the variable to be explained is log(BERD) and the explanatory variables selected by using a stepwise procedure (the chosen significance threshold is 5%) are the following ones:

– the logarithm of the following continuous variables: turnover, number of employees and unit legal age;
– the share of export activities in the turnover;
– dummy variables corresponding to each value of the business sector, to the location in the region Ile-de-France and to the variable "Applying for R&D tax credit" (cf. Table 3);
– a "not large legal unit" dummy variable which equals 1 if the legal unit is not a large one, 0 otherwise.

---

[15] The continuous variables are turned into categorical ones with four values by using the quartiles.

Table 4 The number of legal units (LU) except the ones in a "grouped" response

Sub-populations		Respondent population		Non-respondent or non-sampled population			
				$I_{R\&D}=0$	$I_{R\&D}=1$		
					BERD estimated by		Estimated BERD (k€)
		DIRD=0	DIRD>0		previous BERD	linear regression	
Large LU		156	3 800	0	5	1	10 053
New LU		685	1 830	86	18	246	34 122
Non-exhaustive stratum	LU with unknown BERD in 2014	770	1 115	6 534	3 537	2 141	975 502
	LU with known BERD in 2014	526	1 424	733	2 069	43	464 011
Total		2 137	8 169	7 353	5 629	2 431	1 483 687

Source: Ministry for Higher Education, Research and Innovation – 2015 R&D survey

1.3.2 Estimation of the BERD for legal units in a "grouped" response

As seen in the paragraph 1.2., 29 "grouped" responses mix several enterprises. As a consequence, it is necessary to estimate the BERD for the 83 legal units involved in these "grouped" response to be able to estimate the BERD of the 63 enterprises related to these units in the end (cf. Table 2).

To estimate the BERD for legal units in a "grouped" response, firstly, we use the same model[16] (model (2)) and the same observations as the ones used in the previous paragraph (§1.3.1.) to get a first estimate. Then, we compute the share related to each legal unit in the estimated total BERD of the "grouped" response. Finally, we get the final estimate of the BERD for each legal unit by multiplying the corresponding share by the collected BERD of the "grouped" response which remains the same before and after estimation.

In conclusion, we have managed to get a value, collected or estimated, for the BERD of each legal unit of the target population *U(LU)*. As a consequence, we can compute a BERD for each enterprise of the sample *S(EP)*. To obtain an estimate of the total of the BERD from these values available at the enterprise level, we have now to determine a weight for each enterprise in the sample *S(EP)*.

2 HOW TO GET AN ESTIMATE OF THE TOTAL BERD FROM THE RESPONDENT ENTERPRISES?
2.1 The respondent enterprises

In this paragraph, we will naturally estimate the total BERD by considering the 8 855 respondent enterprises (cf. Table 1), i.e. the ones whose at least one unit legal answered the 2015 R&D survey. The whole of these enterprises is noted *S(EP)_r*. Thanks to the paragraph 1, a BERD value can be computed for each of them by adding the BERD, collected or estimated (cf. Table 5), of each of their legal units.

---

[16] Actually, the explanatory variables are not exactly the same: the "not large legal unit" dummy variable is substituted by two dummies, namely a "QG" dummy and an "exhaustive QS" one, because only 2 "grouped" responses are not in this sub-population.

Table 5 The legal units of the respondent enterprises

	Number of respondent enterprises	Number of legal units in the respondent enterprises (BERD (k€))		
No estimated BERD for all the legal units in the enterprise	7 688	8 124 (15 303 200)		
Estimated BERD for at least one of the legal unit in the enterprise	1 167	Collected BERD		2 345 (11 176 213)
		Estimated BERD	"Grouped" responses connected to several enterprises	83 (3 459 533)
			Non-sampled or non-respondent	2 143 (367 006)
TOTAL	$S(EP)_r$: 8 855	12 695 (30 305 952)		

Note: The total is not weighted for the BERD.
Source: Ministry for Higher Education, Research and Innovation – 2015 R&D survey

To obtain the total BERD, we just have to determine a weight for each respondent enterprise. We compute this set of weights by using generalized weight share method (GWSM) (Lavallée, 2007).

2.2 Generalized weight share method (GWSM)

In the framework of the generalized weight share method (GWSM), we distinguish two populations: the collection population, here *U(LU)*, and the population of interest, here *U(EP)*. The GWSM allows to obtain an unbiased estimator of the variable of interest, here the total BERD, from the data available at the interest population level and the weights of the units sampled in the collection population, i.e. here the legal units in the sample *S(LU)*. In this study, only the weights of respondent legal units, i.e. the legal units in $S(LU)_r$, are considered because total non-response is handled here through reweighting.

The estimator of the total BERD related to the GWSM, noted $\widehat{BERD\_tot}_{GWSM}$ is expressed as follows:

$$\widehat{BERD\_tot}_{GWSM} = \sum_{EP_i \in S(EP)_r} pond\_EP_i^{GWSM} \times BERD\_EP_i, \tag{3}$$

where:

- $BERD\_EP_i$ is the BERD computed for the enterprise EP_i by adding the BERD of each of its legal units which are in the collection population *U(LU)*:

$$BERD\_EP_i = \sum_{LU_{k,i} \in EP_i \cap U(LU)} BERD\_LU_{k,i}, \tag{4}$$

- $pond\_EP_i^{GWSM}$ is the weight for the enterprise EP_i computed thanks to the GWSM. It is expressed as follows:

$$pond\_EP_i^{GWSM} = \sum_{LU_{k,i} \in EP_i \cap S(LU)_r} \theta_{k,i} \times pond\_LU_{k,i}. \tag{5}$$

Pond_LU$_{k,i}$ is the final weight of the legal unit k of the enterprise i, i.e. the weight got after reweighting to handle total non-response.

The values taken by the coefficients $\theta_{k,i}$ depend on the version of the used GWSM. In this paper, we consider the two following versions:

- the GWSM with classical links: in this version, the weight of an enterprise is based on the number of its legal units which are in the collection population $U(LU)$:

$$\forall LU_{k,i} \in EP_i \cap S\left(LU\right)_r, \quad \theta_{k,i}^{classical} = \theta_i^{classical} = \frac{1}{\sum_{LU \in EP_i \cap U(LU)} 1}, \tag{6}$$

- the GWSM with links weighted by the BERD: in this version, the BERD of legal units is introduced as a weight in the calculation of the coefficients $\theta_{k,i}$:

$$\forall LU_{k,i} \in EP_i \cap S\left(LU\right)_r, \quad \theta_{k,i}^{weighted\_BERD} = \frac{BERD\_LU_{k,i}}{BERD\_EP_i} \tag{7}$$

Then, we have built two new estimators of the total BERD, $\widehat{BERD\_tot}_{GWSM}^{classical}$ and $\widehat{BERD\_tot}_{GWSM}^{weighted\_BERD}$. In the next paragraph, we will compare them with the current estimate of the total BERD, noted $\widehat{BERD\_tot}_{LU}$, got from the BERD collected at the legal unit level and which can be written as:

$$\widehat{BERD\_tot}_{LU} = \sum_{k \in S(LU)_r} pond\_LU_k BERD\_LU_k, \tag{8}$$

and we will try to identify the best of them.

3 WHICH ESTIMATOR AT THE ENTERPRISE LEVEL TO RETAIN?
3.1 Comparison of the two GWSM estimators

To assess the quality of our estimators, we would like to compare them to the current estimator $\widehat{BERD\_tot}_{LU}$. But the computation of the GWSM with BERD-weighted links estimator of the BERD mechanically gives the same result as the current one. Another idea is to compare our two GWSM estimators to the real totals, i.e. the totals on the population $U(EP)$, for the BERD[17] and two variables available in the French business register SIRUS (turnover and headcount). In this paper, we will assess the quality of our two GWSM estimators by computing the relative deviation from the real total for these three variables (cf. Table 6).

Both number of enterprises and legal units are underestimated by the two GWSM estimators, but the estimations of the totals for the two non-R&D variables (turnover and headcount) and for the BERD are quite close to the real ones, especially in the case of the GWSM with BERD-weighted links estimator. As a consequence, the latter seems to be the best one with the "relative deviation" criterion. However, further studies would be necessary to consolidate this first result. For example, we could carry out several simulations to compute several values for both GWSM estimators and, then, to deduce the bias for each estimator. In this paper, we focus only on the "relative deviation" criterion and, so, we decide to choose the GWSM BERD-weighted links estimator as the best one to compute R&D statistics at the enterprise level.

---

[17] Actually, about the BERD, we can only compute a pseudo-real total on the population $U(EP)$ thanks to estimations conducted in the first paragraph.

Table 6 GWSM estimations based on the enterprise population and corresponding relative deviations from the real totals

	U(EP)	S(EP)ₛ GWSM with classical links		S(EP)ₛ GWSM with BERD-weighted links	
	Real total	Estimated total	Relative deviation from the real total	Estimated total	Relative deviation from the real total
Number of enterprises	21 466	19 106	−11%	19 254	−10%
Number of legal units	25 965	23 255	−10%	23 838	−8%
Turnover (B€)	2 038	1 976	−3%	2 066	1%
Headcount (thousands)	5 740	5 416	−6%	5 509	−4%
BERD (M€)	31 423	30 640	−2%	31 756	1%

Source: Ministry for Higher Education, Research and Innovation – 2015 R&D survey; INSEE – SIRUS

Another way to judge the quality of this new estimator is to understand how relevant it is in terms of economic analysis in comparison to an analysis at the legal unit level.

3.2 Comparison with the legal unit level: which level is better for an analysis by business category?

As stated in the previous paragraph, the GWSM BERD-weighted links estimator gives the same total BERD as our current estimator computed at the legal unit level. However, another interesting and topical issue is the analysis by business category. Indeed, the law (LME)[18] specifies now four categories of business for the purposes of statistical and economic analysis, using headcount, turnover and total balance sheet of the enterprise: microenterprises, small and medium-sized enterprises (SME), intermediate-sized enterprises (ISE) and large enterprises. Before the new French definition of enterprise (LME), legal units were only categorized by their size in terms of headcount, with the same threshold than in the new definition. Moreover, data have already been published on LME business categories from the legal unit level: for example, the BERD of SMEs is the sum of the BERD of all legal units belonging to a SME. With this method, we obtain the number of legal units in each enterprise category but not the number of enterprises. So, it's interesting to compare the results got from our new estimator at the enterprise level (the GWSM BERD-weighted links estimator) to those got at the legal unit level, in terms of breakdown by business category (cf. Figure 1).

With the new French definition of enterprise, we completely change what we say about the breakdown between business categories: with the headcount definition on legal units we underestimated the share of large enterprises in the total BERD, and with the LME definition on legal units the shares in the total BERD were well estimated but we could not say anything in terms of number of enterprises. Thanks to our new methodology, we can now compute estimators directly based on the enterprises in our population, for example the number of enterprises carrying out R&D activities.

---

[18] Article 51 of the French law on the modernisation of the economy (LME), 2008.

Figure 1 Breakdown of some variables by business category at the enterprise level and at the legal unit level
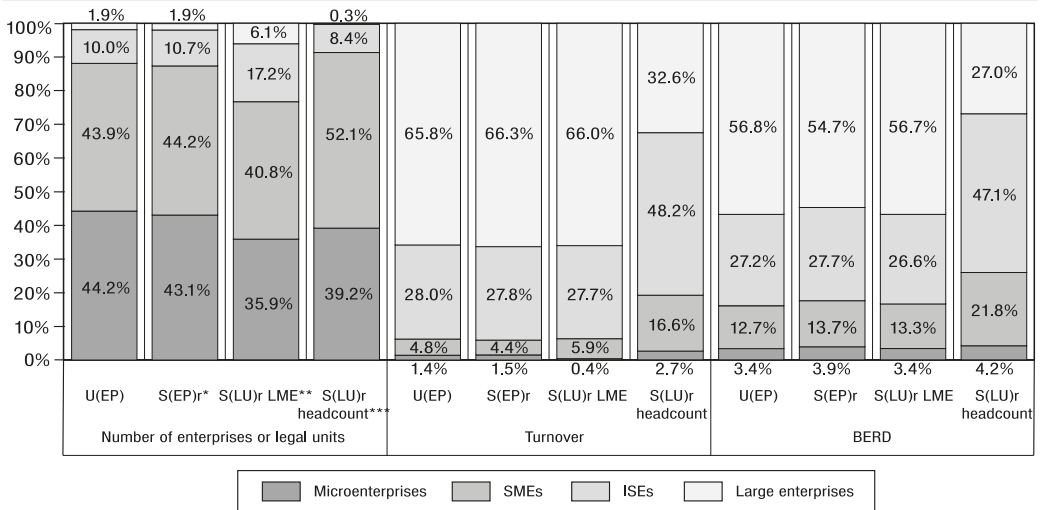
Notes: * Estimators computed with the GWSM with BERD-weighted links on the respondent enterprises *S(EP)*.
    ** Estimators at the legal unit level, LME definition of business categories.
    *** Estimators at the legal unit level, business categories defined only by headcount thresholds.
Reading note: There are 44.2% microenterprises in *U(EP)*, whereas we estimate there are 35.9% legal units belonging to a microenterprise (*S(LU)*, LME).
Source: Ministry for Higher Education, Research and Innovation – 2015 R&D survey; INSEE – SIRUS

CONCLUSION

The data collection at the legal unit level is still interesting and convenient, and the "grouped" responses and non-respondent legal units belonging to a respondent enterprise prevent us to easily rebuild the enterprise data. Different models were therefore used to rebuild a total BERD for each enterprise. This allows us to compute more relevant estimators at this enterprise level, especially in terms of business categories, thanks to the generalized weight share method (GWSM) which generates new sets of weights.

To consolidate our results, it would be necessary to assess more thoroughly the quality of our GWSM BERD-weighted links estimator, for example by carrying out simulations based on different samples to compute different expected values of our total BERD estimator and its bias. Then, we could use our new set of weights at the enterprise level to estimate the totals of other variables, such as R&D personal or researchers, and check the quality of those estimators.

References

OECD. *Frascati Manual 2015: Guidelines for Collecting and Reporting Data on Research and Experimental Development, The Measurement of Scientific, Technological and Innovation Activities.* Paris: OECD Publishing, 2015.
LAVALLÉE, P. *Indirect sampling.* Springer, 2007.
HAAG, O. Profiling: a new and better way to apprehend the globalization. *Q2016 Conference*, Madrid, 2016.
HAAG, O. The French business registers system: How to improve the quality of the statistics by combining different statistical units. *The Fifth International Conference on Establishment Surveys* (ICES-V), Geneva, 2016.
DUNNING, J. H. AND LUNDAN, S. M. The Internationalization of Corporate R&D: a Review of the Evidence and Some Policy Implications for Home Countries. *Review of Policy Research*, 2009, No. 26, pp. 13–33.
FLORIDA, F. The globalization of R&D: Results of a survey of foreign-affiliated R&D laboratories in the USA. *Research Policy*, 1997, 26, pp. 85–103.
MONCADA-PATERNO-CASTELLO, P., VIVARELLI, M., VOIGT, P. Drivers and Impacts in the Globalization of Corporate R&D: the European Experience. *Institute for Prospective Technological Studies Working Paper Series on Corporate R&D and Innovation*, Luxembourg: Publications Office of the European Union, 2011, No. 3.