COMMENT PASSER DU CONCEPT D'UNITÉ LÉGALE À LA DÉFINITION ÉCONOMIQUE DE L'ENTREPRISE DANS LES STATISTIQUES SUR LA R&D ?

Thomas BALCONE (*), Camille SCHWEITZER (*)

(*) SIES, Département des études statistiques de la recherche

thomas.balcone@recherche.gouv.fr
camille.schweitzer@recherche.gouv.fr

Mots-clés : statistiques d'entreprise, unité statistique, collecte, processus de réalisation d'enquêtes, sondages

Résumé

Les moyens financiers et humains consacrés à la recherche et au développement (R&D) par les entreprises en France sont mesurés par une enquête annuelle menée par le service statistique du Ministère de l'enseignement supérieur, de la recherche et de l'innovation (SIES). Les unités interrogées dans le cadre de cette enquête sont les unités légales. Étant donné l'importance croissante des groupes dans l'économie française, en particulier en matière de stratégies de R&D, l'analyse des activités de R&D au seul niveau des unités légales semble incomplète. La prise en compte de la définition européenne de l'entreprise permettrait de diffuser des statistiques sur la R&D dans le secteur privé plus pertinentes.

Une première approche consisterait à interroger non plus les unités légales mais directement les entreprises au sens de la nouvelle définition. Cette piste est difficilement envisageable en raison de la difficulté présumée pour l'entreprise de connaître avec précision les activités de R&D de l'ensemble de ses unités légales. En effet, il est déjà parfois difficile de trouver le bon interlocuteur au sein des unités légales. De plus, l'analyse en termes d'unités légales est importante pour comprendre la façon dont les entreprises organisent leurs activités de R&D. Un traitement post-collecte semble ainsi plus approprié. Il convient alors de reconstruire les contours des entreprises à partir des unités légales constituant la population de l'enquête, à savoir celles qui sont susceptibles de mener des activités de R&D.

A partir de cette nouvelle population, il s'agit alors de construire le meilleur estimateur pour chacune de nos variables d'intérêt que sont la dépense intérieure de recherche et développement (DIRDE), l'effectif de R&D et l'effectif de chercheurs et ingénieurs. La méthode généralisée de partage des poids (MGPP) semble indiquée dans ce genre de situation. Deux versions de la méthode sont testées dans le cadre de cette étude : MGPP avec liens classiques et MGPP avec liens pondérés par la DIRDE. La difficulté dans la mise en application de cette méthode repose sur le traitement des entreprises dont certaines unités légales ne sont pas interrogées. Afin de mesurer la qualité des estimateurs obtenus et de retenir in fine le meilleur, des simulations sont réalisées en tirant à chaque fois un nouvel échantillon dans la strate non exhaustive de la base de sondage.

À partir de l'estimateur retenu, des statistiques peuvent alors être établis au niveau entreprise, permettant notamment une analyse par catégorie d'entreprise. La comparaison de ces résultats à ceux obtenus au niveau unité légale permet d'avoir une meilleure connaissance de l'activité de R&D en 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. Considering the increasing importance of the enterprise group in the French economy and the European definition of an enterprise, it seems important to disseminate more consistent and relevant R&D statistics on the business sector at the enterprise level. This is now possible thanks to the French business statistical register established by the French national statistical institute (Insee), called SIRUS. This article first describes why the data should go on being collected at legal unit level and not at enterprise one. Then, it presents the process based on SIRUS and used to compute key indicators on R&D at enterprise level. To conclude, this paper 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.

Introduction

Dans le contexte de la mondialisation, la Recherche et Développement (R&D) est une problématique importante pour les entreprises et plus largement les nations pour rester compétitif, notamment en permettant l'innovation. Dans le cadre de la stratégie « Europe 2020 », l'Union Européenne s'est fixée un objectif de 3 % de Produit Intérieur Brut (PIB) consacré à la R&D pour chaque pays, afin de promouvoir la croissance et l'emploi. Produire des données de bonne qualité sur la R&D parait donc indispensable afin de suivre l'évolution des dépenses de R&D par rapport à cet objectif, notamment dans le secteur des entreprises. C'est l'objectif de l'enquête R&D, basée sur le Manuel de Frascati, ouvrage réalisé par l'OCDE pour fourir à tous les pays des définitions et méthodes communes pour la réalisation d'enquêtes sur la R&D.

En France, l'enquête sur les moyens consacrés à la R&D dans les entreprises est conduite par le Ministère de l'Enseignement Supérieur, de la Recherche et de l'Innovation depuis 1963, et se base sur les unités légales. Avec la mondialisation et le poids de plus en plus important des grands groupes dans l'économie, une nouvelle définition plus économique que légale de l'entreprise a été introduite dans la législation européenne[1] : une entreprise est la plus petite combinaison d'unités légales qui constitue une unité organisationnelle de production de biens et de services jouissant d'une certaine autonomie de décision, notamment pour l'affectation de ses ressources courantes. Produire des données au niveau des entreprises, selon cette nouvelle définition, semble donc pertinent pour étudier la R&D dans le secteur privé, d'autant plus que cela pourrait devenir obligatoire avec la mise en œuvre du futur cadre de régulation des statistiques d'entreprise FRIBS[2] actuellement en négociation.

Dans la première partie de cet article, nous allons reconstruire une information sur la R&D au niveau entreprise à partir de l'enquête R&D. Nous présenterons ensuite une méthode pour produire des indicateurs clés sur la R&D à ce nouveau niveau. Enfin, la dernière partie de cet article mettra en parallèle ces indicateurs clés avec notre ancien cadre d'analyse au niveau des unités légales pour mettre en évidence l'impact de ce passage au concept d'entreprise dans les statistiques de R&D.

[1] Législation européenne 696/93.

[2] Framework regulation integrating business statistics.

1. Comment obtenir des données au niveau entreprise ?

1.1. Première approche : collecter les donnée au niveau entreprise

Actuellement, l'unité de collecte dans l'enquête sur les moyens consacrés à la recherche et au développement dans les entreprises (enquête R&D) est l'unité légale. Afin de produire des statistiques au niveau entreprise au sens économique, plus cohérentes et pertinentes en termes d'analyse économique, l'idée la plus naturelle est de considérer l'entreprise comme unité de collecte de l'enquête et non plus l'unité légale. Cependant, un tel changement d'unité de collecte n'est pas sans risque.

Le premier risque est la chute du taux de réponse, actuellement au dessus de 90 %. En effet, en collectant l'information au niveau entreprise, c'est l'unité légale considérée comme le centre de décision qui serait interrogée, or ce n'est pas forcément cette unité légale qui réalise des activités de R&D au sein de l'entreprise, et plusieurs unités légales d'une même entreprise peuvent réaliser des activités de R&D. Dans le cas où nous aurions un seul contact dans l'unité légale décisionnaire de l'entreprise, sa charge de réponse pourrait être augmentée : il faut repérer les unités légales qui réalisent de la R&D, remonter les informations concernant chaque unité légale. Dans certains cas, ces informations sont déjà consolidées au niveau de l'entreprise ce qui ne poserait pas de problème (certaines réponses se font déjà sur le contour entreprise). En revanche, pour d'autres entreprises, cela reviendrait à multiplier le nombre d'intermédiaires et donc accroître la durée du processus de réponse voire même augmenter la non-réponse. Déjà aujourd'hui, certaines unités légales peinent à remonter les informations de l'ensemble de leurs établissements. Les entreprises composées de plusieurs unités légales représentant une grande part dans le total des dépenses de R&D des entreprises, cela peut donc être dangereux.

Un second risque, en termes de compréhension de l'activité de R&D dans les entreprises, est de ne plus pouvoir analyser la R&D par unité légale. En effet, cette analyse est pertinente pour comprendre la façon dont les entreprises organisent leurs activités de R&D, dans la mesure où près de 28 % des entreprises susceptibles de réaliser des activités de R&D[3] qui sont composées de plusieurs unités légales ont plus d'une unité légale dans la population de l'enquête. Par exemple, l'analyse par unité légale permet de savoir si les entreprises concentrent leurs activités de R&D dans une seule unité légale et si celle-ci est un centre dédié à la R&D, ou si elles divisent leurs activités entre plusieurs unités légales. De plus, la Comptabilité Nationale utilise le concept d'unité légale et continuera à l'utiliser pour le moment. Il est donc important de produire des données consolidées et pertinentes au niveau des unités légales.

Ainsi, choisir l'entreprise comme unité de collecte ne semble pas être la meilleure solution à court terme pour étudier la R&D dans le secteur des entreprises. Nous choisissons donc de conserver l'unité légale comme unité de collecte.

1.2. Seconde approche : reconstruire le contour des entreprises à partir des unités légales

Nous décidons donc de reconstruire le contour des entreprises à partir des unités légales de notre population. La constitution de la population de l'enquête R&D est atypique : contrairement à la plupart des autres enquêtes réalisées auprès des entreprises, elle n'est pas directement issue du répertoire SIRUS. Selon les recommandations du Manuel de Frascati[4] (cf. p.244), la population cible de l'enquête R&D au niveau des unités légales, notée *U(UL)*, est uniquement constituée des unités

[3] Une entreprise est susceptible de réaliser des activités de R&D si au moins une de ses unités légales est dans la population de l'enquête, i.e. est elle-même susceptible de réaliser des activités de R&D.

légales susceptibles de réaliser des activités de R&D. Cette population est redéfinie chaque année grâce à la population de l'enquête précédente ainsi que d'autres sources de données liées à la R&D ou l'innovation (bénéficiaires du Crédit d'Impôt Recherche (CIR), enquête innovation (CIS), jeunes entreprises innovantes,...).

Si l'on considère que l'activité de R&D d'une entreprise correspond à l'ensemble des activités de R&D des unités légales la composant, nous pouvons définir la population cible au niveau entreprise, notée *U(EP)*, comme l'ensemble des entreprises dont au moins une unité légale est susceptible de réaliser des activités de R&D, i.e. appartient à la population cible *U(UL)*. En suivant la même logique, l'échantillon au niveau entreprise, noté *S(EP)*, correspond à l'ensemble des entreprises dont au moins une unité légale appartient à l'échantillon *S(UL)*. Pour produire des statistiques de R&D au niveau entreprise à partir de l'échantillon *S(EP)*, il faudrait connaître l'activité de R&D de toutes les unités légales de la population cible *U(UL)* appartenant aux entreprise de *S(EP)*. Or nous ne disposons pas des données sur l'ensemble de ces unités légales, d'une part car l'échantillon *S(UL)* ne correspond pas à un tirage par grappes de l'ensemble des unités légales d'une entreprise. En effet, comme dans la plupart des enquêtes thématiques sur les entreprises, la population cible des unités légales *U(UL)* est composée d'une strate exhaustive (les unités légales importantes en termes de R&D et les unités légales apparaissant pour la première fois dans la population) et d'une strate non exhaustive, dans laquelle est tiré un échantillon.

Dans l'enquête 2015, les unités légales répondantes représentent 1 435 entreprises de *S(EP)* qui sont composées de plus d'une unité légale dans la population cible *U(UL)* (cf. tableau 1). Parmi elles, 1 090 sont composées d'au moins une unité légale présente dans la population cible *U(UL)* mais dont on ne connaît pas l'activité de R&D, soit parce qu'elle n'est pas dans l'échantillon *S(UL)* (2 047 unités légales), soit parce qu'elle n'a pas répondu (97 unités légales). Dans la suite de ce papier, une entreprise est considérée comme répondante si au moins une de ses unités légales a répondu à l'enquête ; sinon, l'entreprise est considérée comme non répondante.

Tableau 1 – La population cible de l'enquête R&D 2015 au niveau unité légal *U(UL)* et au niveau entreprise *U(EP)*

<table>
<tr><th colspan="2">Population</th><th>Niveau unité légale (UL)</th><th colspan="4">Niveau entreprise</th></tr>
<tr><td colspan="2">Population cible</td><td>U(UL) : 25 962</td><td colspan="4">U(EP) : 21 466</td></tr>
<tr><td rowspan="4">Échantillon</td><td rowspan="3">Répondantes</td><td rowspan="3">10 552</td><td rowspan="3">8 855</td><td>EP composées d'une seule UL dans U(UL)</td><td colspan="2">7 420</td></tr>
<tr><td rowspan="2">EP composées de plus d'une UL dans la population</td><td>Toutes les UL sont répondantes</td><td>345</td></tr>
<tr><td>Au moins une UL dont on n'a pas d'information sur la R&D</td><td>1 090</td></tr>
<tr><td>Non répondantes</td><td>1 007</td><td colspan="4">894</td></tr>
<tr><td colspan="2">Total</td><td>S(UL) : 11 559</td><td colspan="4">S(EP) : 9 749</td></tr>
</table>

Source : MESRI-SIES – enquête R&D 2015

[4] Le Manuel de Frascati, réalisé par l'OCDE, précise la façon dont les données concernant la recherche et le développement doivent être collectées, dans un but d'harmonisation. La dernière édition date de 2015.

D'autre part, l'information n'est pas disponible pour toutes les unités légales de la population *U(UL)* appartenant aux entreprises de *S(EP)* car il existe des réponses « groupées » dans l'enquête, i.e. des réponses qui concernent plusieurs unités légales. Dans certaines situations, il est décidé en accord avec les gestionnaire et le responsable de l'enquête que certains correspondants peuvent répondre pour plusieurs unités légales, indépendamment du concept d'entreprise ou de groupe, afin de réduire la charge de réponse[5]. Ces réponses « groupées », environ une centaine chaque année, sont associées à un « contour de réponse » composé des identifiants des différentes unités légales intégrées à la réponse. Ces réponses doivent être traitées spécifiquement afin d'obtenir *in fine* une information pour chaque unité légale des « contours de réponse » appartenant à différentes entreprises de *S(EP)*.

Dans l'enquête 2015, il y a 89 réponses « groupées » (à peine 1 % de l'ensemble des réponses à l'enquête – cf. tableau 2) qui correspondent à 246 unités légales et 5,8 Md€ de dépenses internes de R&D (DIRD) (i.e. 18,3 % de la DIRD totale du secteur des entreprises). Parmi ces 89 réponses « groupées », 29 mélangent plusieurs entreprises : elles représentent 83 unités légales et 63 entreprises.

Tableau 2 – Réponses « groupées »

Population répondante		**Nombre de réponses**	**Nombre d'unités légales**	**Nombre d'entreprises**
Réponses non « groupées »		10 306	10 306	8 732[6]
Réponses « groupées »	Une seule entreprise	60	163	60
	Plusieurs entreprises	29	83	63
Total		10 395	10 552	8 855

Source : MESRI-SIES – enquête R&D 2015

Il n'est donc pas si simple de produire des statistiques de R&D sur le contour entreprise pour les entreprises présentes dans l'échantillon *S(EP)* à partir des données collectées au niveau de l'échantillon d'unités légales *S(UL)*. Des traitements post-collecte sont donc nécessaires.

1.3. Estimation des dépenses internes de R&D au niveau entreprise

Dans le cadre de cette étude, nous ne considérerons comme données R&D que les dépenses internes de R&D (DIRD). Comme expliqué dans la partie précédente, certaines entreprises de l'échantillon *S(EP)* sont composées d'unités légales de la population cible *U(UL)* dont la DIRD est inconnue. Il est donc nécessaire d'estimer cette DIRD au niveau des unités légales afin d'obtenir un estimateur de la DIRD de bonne qualité pour ces entreprises.

Comme dit plus haut, les réponses « groupées » nécessitent un traitement particulier. Nous allons donc commencer par traiter les unités légales dont on ne connaît pas la DIRD et qui n'appartiennent pas à une réponse « groupée ».

1.3.1. Estimation de la DIRD des unités légales hors des réponses « groupées »

Les unités légales de la population cible *U(UL)* sont toutes susceptibles de réaliser des activités de R&D, mais dans les faits certaines n'en réalisent pas : projets de R&D sur certaines années, changement d'activité, nouvelle unité légale intégrée à la population qui se révèle être hors

[5] Cela peut s'expliquer par une mise en commun des activités de R&D entre plusieurs unités légales pouvant être proches géographiquement, ou a une comptabilité unifiée dans certains groupes par exemple.

[6] En réalité, ce nombre est supérieur car certaines entreprises (8 782) apparaissent à la fois dans une réponse non « groupée » et dans une réponse « groupée »,

champ. Malheureusement, cette information n'est pas connue pour les unités légales qui ne sont pas dans l'échantillon *S(UL)* ou qui ne répondent pas. Par conséquent, une première étape consiste à modéliser la probabilité de réaliser des activités de R&D, notée P(DIRD>0), pour chaque unité légale de la population cible *U(UL)* non interrogée ou non répondante.
Ce modèle est basé sur l'ensemble des unités légales répondantes en excluant celles des réponses « groupées », soit 10 306 réponses dans l'enquête 2015, à la fois positives et négatives[7]. On distingue quatre sous-populations d'unités légales répondantes :

- les unités légales importantes en termes de R&D (strate exhaustive QG[8] et QS[9] exhaustifs),
- les unités légales nouvellement intégrées à la population cible *U(UL)* (QS new),
- les unités légales de la strate non exhaustive dont la DIRD en 2014 est inconnue,
- les unités légales de la strate non exhaustive dont la DIRD en 2014 est connue.

Pour les unités légales importantes en termes de R&D, comme le taux de non-réponse et le taux de réponse négative sont tous les deux très faibles (respectivement 0,15 % et 3,94 %), nous supposons que la probabilité de réaliser des activités de R&D P(DIRD>0) est égale à 1. Pour les trois autres sous-populations, on estime cette probabilité grâce au modèle de régression logistique suivant :

$$logit\left[P\left(DIRD>0\right)\right] = \beta_0 + \sum_{k=1}^{K} \beta_k X_k + \varepsilon \quad \textbf{(1)}$$

Où :

- β_0 est la constante,
- $\beta_1,...,\beta_K$ sont les coefficients liés aux K variables explicatives $X_1,...,X_K$,
- ε est le terme d'erreur.

Les variables explicatives utilisées dans le modèle pour les différentes sous-populations sont présentées dans le tableau 3 :

[7] Une unité légale a répondu positivement à l'enquête R&D 2015 si elle a déclaré avoir réalisé des travaux de R&D en 2015 (DIRD>0). Sinon, elle a répondu négativement (DIRD=0).
[8] Questionnaire général, envoyé aux unités légales dont la DIRD est supérieure à 2 000 k€, plus détaillé.
[9] Questionnaire simplifié, moins détaillé. Les QS exhaustifs correspondent aux unités légales dont la DIRD est entre 400 k€ et 2 000 k€ et qui sont interrogées de manière exhaustive.

Tableau 3 – Les variables explicatives utilisées dans les modèles de régression logistique

Variables explicatives	**Valeurs**			
Chiffres d'affaires (k€)	[0 ; 200[	[200 ; 1 120[	[1 120 ; 5 700[	[5 700 ; +∞[
Effectif salarié	[0 ; 2]	[3 ; 9]	[10 ; 32]	[33 ; +∞[
Part du chiffre d'affaire à l'export (%)	0	]0 ; 5[	[5 ; 20[	[20 ; 100]
Âge	[0 ; 2[	[2 ; 12[	[12 ; 23[	[23 ; +∞[
Secteur d'activité	Industrie		De haute technologie	
			De moyenne-haute technologie	
			De moyenne-faible technologie	
			De faible technologie	
	Secteur primaire, énergie, BTP			
	Services (hors R&D)			
	R&D (division 72 de la NACE Rev. 2)			
Appartenance à un groupe	Unité légale indépendante			
	Appartenance à		Un groupe français	
			Un groupe étranger	
Région d'activité	Île-de-France		Autres régions	
Demande de Crédit d'Impôt Recherche (CIR)	Oui		Non	

Source : MESRI-SIES

Pour la sous-population des unités légales de la strate non exhaustive dont la DIRD 2014 est connue, on ajoute une variable indicatrice qui vaut 1 si l'unité légale a répondu positivement à l'enquête 2014 (DIRD>0) et 0 sinon (DIRD=0).

Nous estimons ensuite les coefficients pour chacun des trois modèles sur les unités légales correspondant aux trois sous-populations. Ces estimateurs nous permettent d'estimer la probabilité de réaliser des activités de R&D pour toutes les unités légales dont on ne connaît pas la DIRD en 2015 et qui ne sont pas dans des réponses « groupées ». À partir de cette probabilité estimée, on définit une variable indicatrice de la réalisation de travaux de R&D, notée $I_{R\&D}$:

$$I_{R\wedge D} = \begin{Bmatrix} 1\ si\ \hat{P}\left(BERD>0\right) > 0{,}5, \\ 0\ sinon \end{Bmatrix}$$

Pour certaines unités légales (629), l'estimation de la probabilité de réaliser des activités de R&D n'est pas possible en raison de valeurs manquantes pour des variables explicatives. Dans ces cas, nous fixons $I_{R\&D}$ à 0.

Une fois que nous avons estimé quelles unités légales réalisent effectivement des travaux de R&D en 2015, une seconde étape est d'estimer le montant de leurs dépenses internes de R&D (DIRD). Dans le cas où l'unité légale a répondu positivement à l'enquête R&D dans les années précédentes (entre 2009 et 2014), nous faisons l'hypothèse que la DIRD en 2015 est égale à cette DIRD ancienne, corrigée de la variation des prix. Sinon, la DIRD est estimée grâce à un modèle de régression linéaire basé sur les unités légales ayant répondu positivement à l'enquête en 2015 (8 169[10] unités légales, cf. tableau 4). Dans ce modèle (modèle **(2)**), la variable à expliquer est

[10] En réalité, le modèle se base sur 7 609 unités légales lorsque l'on retire les unités avec des valeurs manquantes et les observations aberrantes (5).

log(DIRD) et les variables explicatives sélectionnées par une procédure stepwise (seuil de significativité retenu de 5 %) sont les suivantes :

- le logarithme des variables continues suivantes : chiffre d'affaires, nombre de salariés et âge de l'unité légale ;
- la part du chiffre d'affaires réalisé à l'export ;
- les variables indicatrices suivantes : une par secteur d'activité, localisation de l'activité en Île-de-France, demande de Crédit d'Impôt Recherche (cf. tableau 3) ;
- la variable indicatrice « l'unité légale n'est pas importante en termes de R&D », qui vaut 1 si l'unité légale n'est pas importante en termes de R&D i.e. n'est pas un QG ou un QS exhaustif, 0 sinon.

Tableau 4 – Nombre d'unités légales (UL) hors réponses « groupées »

Sous-populations		**Population des répondants**		**Population non répondante ou non échantillonnée**			
		DIRD=0	DIRD>0	$I_{R\&D} = 0$	$I_{R\&D} = 1$		
					DIRD estimée par...		DIRD estimée (k€)
					DIRD antérieure	Régression linéaire	
UL importantes en termes de R&D		156	3 800	0	5	1	10 053
Nouvelles UL		685	1 830	86	18	246	34 122
Strate non exhaustive	UL dont la DIRD 2014 est inconnue	770	1 115	6 534	3 537	2 141	975 502
	UL dont la DIRD 2014 est connue	526	1 424	733	2 069	43	464 011
Total		2 137	8 169	7 353	5 629	2 431	1 483 687

Source : MESRI-SIES – enquête R&D 2015

1.3.2. Estimation de la DIRD des unités légales dans les réponses « groupées »

Comme expliqué dans la partie 1.2., 29 réponses « groupées » mélangent plusieurs entreprises différentes. Il est donc nécessaire d'estimer la DIRD des 83 unités légales concernées par ces réponses afin d'estimer la DIRD des 63 entreprises correspondantes. Pour cela, nous utilisons le même modèle (modèle **(2)**) sur les mêmes observations que celui utilisé pour estimer la DIRD des unités légales hors réponses « groupées » (cf. §1.3.1.). Une fois la DIRD estimée pour les unités légales, nous calculons une part pour chacune d'entre elles dans la DIRD estimée pour la réponse « groupée ». Nous appliquons cette part à la DIRD déclarée dans la réponse « groupée » à l'enquête, que nous considérons exacte, afin d'obtenir in fine une DIRD pour chaque unité légale dont la somme est égale à la DIRD déclarée pour la réponse « groupée ».

Par exemple, si l'on considère une réponse « groupée » composée de deux unités légales (UL_1 et UL_2) qui déclareraient une DIRD globale de 100 k€, le tableau 5 présente alors notre méthode d'estimation :

Tableau 5 – Exemple d'estimation de la DIRD des unités légales d'une réponse « groupée »

	DIRD estimée par régression (k€)	**Part dans la DIRD estimée totale**	**DIRD finale (k€)**
UL_1	85	68 %	68
UL_2	40	32 %	32
Total réponse « groupée »	125	100 %	100 (total déclaré)

En conclusion, cette première partie nous a permis d'obtenir une DIRD pour toutes les unités légales de notre population cible *U(UL)*, collectée ou estimée. Il est ainsi possible de reconstruire une DIRD pour chaque entreprise de l'échantillon *S(EP)*. Pour obtenir un estimateur de la DIRD totale de l'ensemble des entreprises de *U(EP)* à partir de notre échantillon *S(EP)*, il s'agit à présent de déterminer un poids pour chacune des entreprises de cet échantillon.

2. Comment obtenir un estimateur de la DIRD totale à partir des entreprises répondantes ?

2.1. Les entreprises répondantes

Dans cette partie, nous allons estimer la DIRD totale des entreprises à partir des 8 855 entreprises répondantes (cf. tableau 6), i.e. les entreprises dont au moins une unité légale a répondu à l'enquête R&D 2015. L'ensemble de ces entreprises est noté $S(EP)_r$. Grâce au travail réalisé en partie 1, nous avons pu attribuer une DIRD à chaque entreprise de $S(EP)_r$, en additionnant la DIRD – collectée ou estimée – de chacune de ses unités légales présentes dans la population *U(UL)*.

Tableau 6 – Les unités légales des entreprises répondantes

	Nombre d'entreprises répondantes	**Nombre d'unités légales (UL) dans les entreprises répondantes (DIRD (k€)[11])**		
DIRD estimée pour aucune des UL de l'entreprise	7 688	8 124 (15 303 200)		
DIRD estimée pour au moins une UL de l'entreprise	1 167	DIRD collectée		2 345 (11 176 213)
		DIRD estimée	Réponses « groupées » concernant plusieurs entreprises	83 (3 459 533)
			UL non interrogées ou non échantillonnées	2 143 (367 006)
TOTAL	$S(EP)_r$: 8 855	12 695 (30 305 952)		

Source : MESRI-SIES – enquête R&D 2015

Afin d'obtenir un estimateur de la DIRD totale, il faut déterminer un poids pour chaque entreprise répondante. Ce jeu de poids est obtenu grâce à la méthode généralisée de partage des poids (Lavallée (2007)).

[11] La DIRD indiquée est non pondérée.

2.2. Méthode généralisée de partage des poids (MGPP)

Dans le cadre de la méthode généralisée de partage des poids (MGPP), nous distinguons deux populations : la population interrogée, ici *U(UL)*, et la population d'intérêt, *U(EP)*. La MGPP nous permet d'obtenir un estimateur non biaisé de la variable d'intérêt, ici la DIRD totale, à partir des données au niveau de la population d'intérêt et des poids des unités échantillonnées de la population interrogée, ici les unités légales de l'échantillon *S(UL)*. Dans cette étude, seuls les poids des unités légales répondantes, c'est-à-dire appartenant à $S(UL)_r$, sont pris en compte car la non-réponses totale est traitée par repondération.

L'estimateur de la DIRD totale issu de la MGPP, noté $\hat{DIRD\_tot}_{MGPP}$, s'exprime comme suit :

$$\hat{DIRD\_tot}_{MGPP} = \sum_{EP_i \in S(EP)_r} pond\_EP_i^{MGPP} \times DIRD\_EP_i$$

Où :

- $DIRD\_EP_i$ est la DIRD de l'entreprise EP_i, calculée comme la somme de la DIRD de chacune de ses unités légales présentes dans la population *U(UL)* :

$$DIRD\_EP_i = \sum_{UL_{k,i} \in EP_i \cap U(UL)} DIRD\_UL_{k,i}$$

- $pond\_EP_i^{MGPP}$ est le poids de l'entreprise EP_i calculé grâce à la MGPP :

$$pond\_EP_i^{MGPP} = \sum_{UL_{k,i} \in EP_i \cap S(UL)_r} \theta_{k,i} \times pond\_UL_{k,i}$$

  $Pond\_UL_{k,i}$ est le poids final de l'unité légale k de l'entreprise i, c'est-à-dire le poids après repondération pour correction de la non-réponse totale.
  La valeur prise par le coefficient $\theta_{k,i}$ dépend de la version de la MGPP utilisée. Dans ce papier, nous considérons les deux versions suivantes :

  → la MGPP avec liens classiques : dans cette version, le poids de l'entreprise est basé sur le nombre de ses unités légales présentes dans la population interrogée *U(UL)* :

$$\forall UL_{k,i} \in EP_i \cap S(UL)_r,\ \theta_{k,i}^{classique} = \theta_i^{classique} = \frac{1}{\sum_{UL \in EP_i \cap U(UL)} 1}$$

  → la MGPP avec liens pondérés par la DIRD : dans cette version, on introduit la DIRD des unités légales comme un poids dans le calcul des coefficients $\theta_{k,i}$:

$$\forall UL_{k,i} \in EP_i \cap S(UL)_r,\ \theta_{k,i}^{pondéré\_DIRD} = \frac{DIRD\_UL_{k,i}}{DIRD\_EP_i}$$

Grâce à ces deux jeux de poids ainsi construits, nous disposons maintenant de deux nouveaux estimateurs de la DIRD totale, $\hat{DIRD\_tot}_{MGPP}^{classique}$ et $\hat{DIRD\_tot}_{MGPP}^{pondéré\_DIRD}$. Dans la partie suivante, nous allons comparer ces nouveaux estimateurs à l'estimateur actuel de la DIRD totale, noté $\hat{DIRD\_tot}_{UL}$, obtenu à partir des données au niveau des unités légales, et qui s'écrit :

$$\widehat{DIRD\_tot}_{UL} = \sum_{k \in S(UL)_r} pond\_UL_k \; DIRD\_UL_k$$

Il s'agira alors de déterminer lequel de nos estimateurs est le meilleur.

3. Quel estimateur au niveau entreprise retenir ?

3.1. Comparaison des deux estimateurs MGPP

Afin d'étudier la qualité de nos estimateurs, nous souhaitons dans un premier temps les comparer à l'estimateur actuel $\widehat{DIRD\_tot}_{UL}$. Or, la construction de notre estimateur par la MGPP avec liens pondérés par la DIRD redonne mécaniquement cet estimateur :

$$\begin{aligned}\widehat{DIRD\_tot}_{MGPP}^{pondéré\_DIRD} &= \sum_{EP_i \in S(EP)_r} pond\_EP_{i\ pondéré\_DIRD}^{MGPP} \; DIRD\_EP_i \\ &= \sum_{EP_i \in S(EP)_r} \sum_{UL_{k,i} \in EP_i \cap S(UL)_r} \frac{DIRD\_UL_{k,i}}{DIRD\_EP_i} \times pond\_UL_{k,i} \times DIRD\_EP_i \\ &= \sum_{l \in S(UL)_r} DIRD\_UL_l \times pond\_UL_l\end{aligned}$$

Une autre idée est de comparer nos estimateurs à des totaux connus sur la population *U(EP)*. Afin d'obtenir de tels totaux, nous pouvons utiliser des variables extérieures à l'enquête, comme le chiffre d'affaires (CA) ou l'effectif salarié, disponibles dans SIRUS. Nous pouvons également considérer le total de la DIRD calculé sur l'ensemble de la population *U(EP)* grâce aux estimations des DIRD réalisées dans la partie 1 pour les unités légales non répondantes ou non enquêtées. Une façon de juger de la qualité de nos estimateurs est alors de calculer un écart relatif à ces totaux connus sur la population *U(EP)* (cf. tableau 7).

Tableau 7 – Estimations MGPP et écarts relatifs aux vrais totaux

	U(EP)	***S(EP)r*, MGPP avec liens classiques**		***S(EP)r*, MGPP avec liens pondérés par la DIRD**	
	Vrai total	Total	Écart relatif au vrai total	Total	Écart relatif au vrai total
Nombre d'entreprises	21 466	19 106	- 11 %	19 254	- 10 %
Nombre d'unités légales	25 965	23 255	- 10 %	23 838	- 8 %
Chiffre d'affaires (Md€)	2 038	1 976	- 3 %	2 066	+ 1 %
Nombre de salariés (milliers)	5 740	5 416	- 6 %	5 509	- 4 %
DIRD (M€)	31 423	30 640	- 2 %	31 756	+ 1 %

Source : MESRI-SIES – enquête R&D 2015 ; Insee – SIRUS

Les nombres d'entreprises et d'unités légales sont sous-estimés, aussi bien avec les poids obtenus par la MGPP avec liens classiques qu'avec ceux correspondant à la MGPP avec liens pondérés par la DIRD. En revanche, les agrégats économiques, liés ou non à la R&D (DIRD et CA notamment), sont relativement bien estimés, particulièrement par la MGPP avec liens pondérés par la DIRD. Sur les

cinq variables retenues, c'est effectivement l'estimateur issu de la MGPP avec liens pondérés par la DIRD qui semble le meilleur selon le critère de l'écart relatif au vrai total. Cependant, une étude plus poussée sur la qualité de cet estimateur serait nécessaire pour confirmer ce premier résultat. Par exemple, nous pourrions réaliser un certain nombre de simulations du tirage de l'échantillon *S(UL)*, ce qui modifierait l'échantillon d'entreprises *S(EP)* et donc les jeux de pondération associés, afin de pouvoir calculer un biais pour nos différents estimateurs. Nous choisissons néanmoins de retenir dans ce papier l'estimateur MGPP avec liens pondérés par la DIRD comme le meilleur estimateur au niveau entreprise, en se basant uniquement sur le critère de l'écart relatif.

Un autre moyen de juger de la qualité de ce nouvel estimateur est d'étudier sa pertinence en termes d'analyse économique, en comparaison avec l'analyse actuellement réalisée au niveau des unités légales.

3.2. Comparaison avec les résultats obtenus au niveau unité légale : quel est le meilleur niveau pour l'analyse par catégories d'entreprise ?

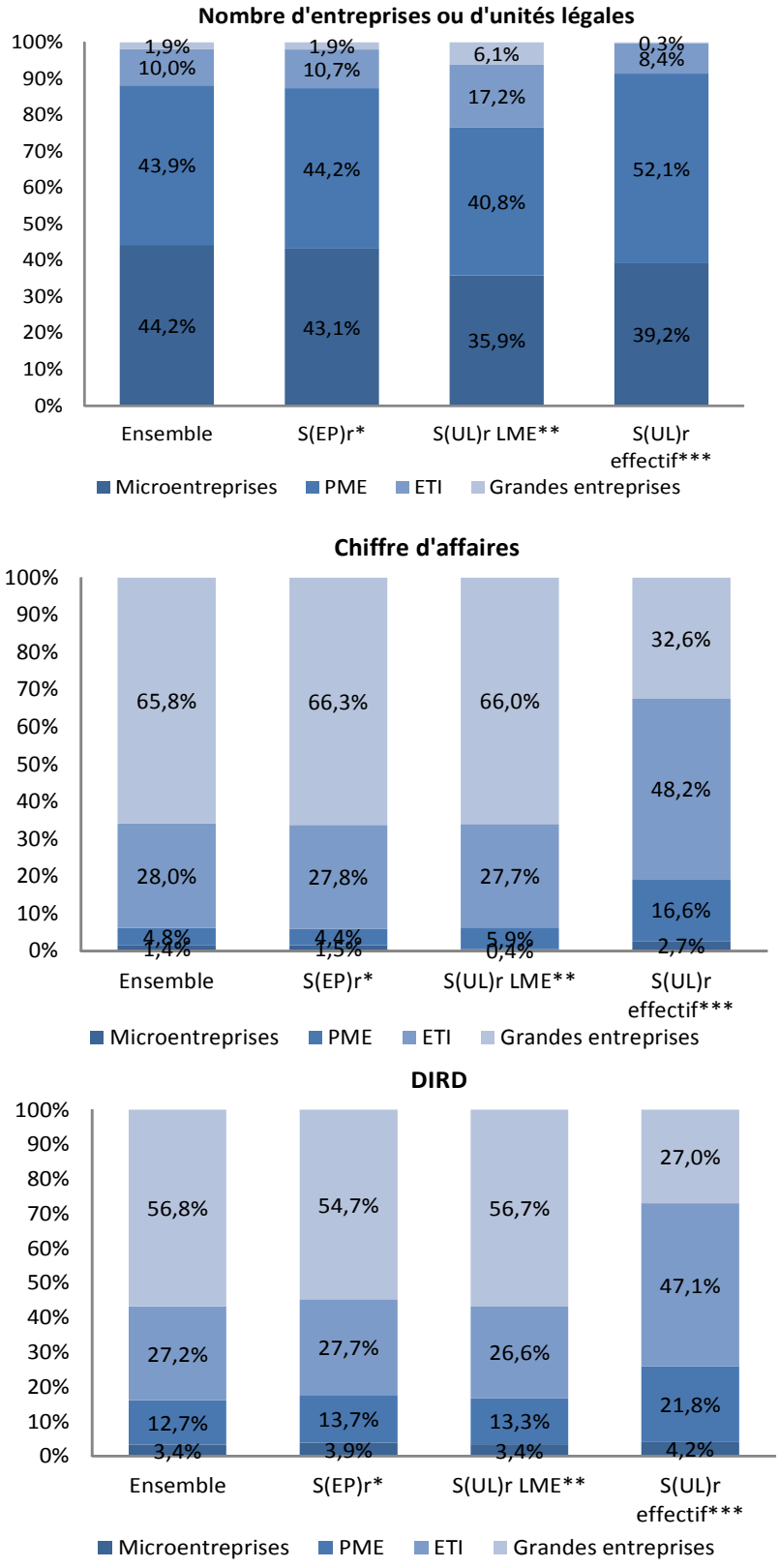
Comme expliqué dans la partie précédente, si on considère la DIRD totale, l'estimateur MGPP avec liens pondérés par la DIRD est mécaniquement identique à l'estimateur actuel obtenu au niveau unités légales. Cependant, une analyse intéressante et d'actualité permise par notre méthode d'imputation au niveau entreprise est l'analyse de la DIRD par catégorie d'entreprise, au sens de la loi de modernisation de l'économie (LME)[12]. Selon l'effectif, le chiffre d'affaire (CA) et le total du bilan de l'entreprise, on distingue quatre catégories :

- les microentreprises : moins de 10 salariés et CA ou total du bilan n'excédant pas 2 M€,
- les petites et moyennes entreprises (PME) : moins de 250 salariés et CA n'excédant pas 50 M€ ou total du bilan n'excédant pas 43 M€,
- les entreprises de taille intermédiaires (ETI) : moins de 5 000 salariés et CA n'excédant pas 1 500 M€ ou total du bilan n'excédant pas 2 000 M€,
- les grandes entreprises : entreprises non classées dans les catégories précédentes.

Historiquement, les statistiques de R&D portent uniquement sur les unités légales classées selon leur taille en termes d'effectifs, selon des seuils identiques à la définition économique de l'entreprise sur cette variable. Mais depuis quelques années, la volonté d'analyser la R&D en termes de catégorie d'entreprise selon la nouvelle définition a émergé. Le raisonnement ne se fait pas sur l'entreprise, mais sur l'unité légale appartenant à une entreprise classée dans l'une des quatre catégories énoncées plus haut. Par exemple, la DIRD de la catégorie PME correspond à la somme des DIRD de l'ensemble des unités légales qui appartiennent à une entreprise classée comme PME. Deux analyses par unités légales en termes de catégorie d'entreprise, la première selon la définition LME et la seconde selon une définition basée uniquement sur la taille en termes d'effectif, ont ainsi déjà été produites. Il est donc pertinent de les comparer à notre nouvelle analyse au niveau entreprise rendue possible par la MGPP avec liens pondérés par la DIRD (cf. graphique 1).

[12] Décret d'application (n°2008-1354) de l'article 51 de la loi de modernisation de l'économie (LME), « relatif aux critères permettant de déterminer la catégorie d'appartenance d'une entreprise pour les besoins de l'analyse statistique et économique ».

Graphique 1 – Répartition du nombre d'entreprises, du CA et de la DIRD selon la catégorie d'entreprise, analyse par entreprise ou par unité légale



Source : MESRI-SIES – enquête R&D 2015 ; Insee – SIRUS

*estimateur obtenu par MGPP avec liens pondérés par la DIRD sur les entreprises répondantes S(EP)r

**estimateur au niveau unités légales, catégorie de l'entreprise définie selon la LME

***estimateur au niveau unités légales, catégorie de l'unité légale définie uniquement par l'effectif

Avec une meilleure prise en compte de la nouvelle définition de l'entreprise, l'analyse change considérablement. En effet, l'analyse en termes d'unités légales classées uniquement selon leur effectif sous-estime le poids des grandes entreprises car peu d'unités légales emploient plus de 5 000 salariés (0,3 %), alors que de petites unités légales peuvent être considérées comme grandes entreprises selon la définition LME car elles appartiennent à une grande entreprise (6,1 % d'unités légales appartiennent à une grande entreprise). Dans notre population cible d'entreprises *U(EP)*, les grandes entreprises représentent une part de 1,9 % : elles sont généralement composées de plusieurs unités légales dont certaines peuvent être interrogées dans l'enquête, elles sont donc moins nombreuses au niveau entreprise qu'au niveau unité légale. Cependant, la répartition du CA et de la DIRD entre les différentes catégories d'entreprise au sens LME au niveau unité légale ($S(UL)_r$ LME) est assez proche de celle observée dans l'analyse au niveau entreprise ($S(EP)_r$ et « Ensemble »). Cependant, seul l'estimateur par MGPP avec liens pondérés par la DIRD semble nous permettre d'estimer correctement le nombre d'entreprises (cf. graphique 1 - $S(EP)_r$ et « Ensemble » pour le nombre d'entreprises). Cet estimateur paraît ainsi être le plus pertinent pour mener une analyse économique au niveau entreprise.

Conclusion

Si l'analyse de la R&D au niveau entreprise semble être pertinente, elle n'est pas si simple à mettre en place. En effet, la collecte au niveau unité légale présente un réel intérêt et l'existence de réponses « groupées » et d'unités légales non répondantes appartenant à des entreprises répondantes nous empêchent de reconstruire simplement les données au niveau entreprise. Pour surmonter ces difficultés, nous avons mis en place différents modèles afin de pouvoir reconstruire une DIRD pour chaque entreprise. Cela nous a permis de construire des estimateurs plus pertinents au niveau entreprise, notamment en termes de catégories d'entreprise, grâce à la méthode généralisée de partage des poids (MGPP).

Afin de consolider ces premiers résultats, il faudrait étudier de manière plus approfondie la qualité de l'estimateur retenu, à savoir celui obtenu par la MGPP avec liens pondérés par la DIRD. Une piste consisterait par exemple à réaliser des simulations permettant d'obtenir différents échantillons dans le but de calculer in fine un biais pour l'estimateur retenu. On pourrait également envisager d'utiliser l'estimateur retenu pour estimer d'autres variables de R&D que la DIRD, comme l'effectif de R&D ou le nombre de chercheurs, et d'étudier la qualité des estimations obtenues.

Bibliographie

[1] OCDE. (2016), *Manuel de Frascati 2015: Lignes directrices pour le recueil et la communication des données sur la recherche et le développement expérimental*, Mesurer les activités scientifiques, technologiques et d'innovation, OECD Publishing, Paris.

[2] Lavallée, P. (2007), *Indirect sampling*, Springer.

[3] Haag, O. (2016), « Profiling: a new and better way to apprehend the globalization », European conference on quality in official statistics 2016 (Q2016), Madrid.

[4]Haag, O. (2016), « 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.