L'impact du profilage sur les statistiques d'entreprises à l'Insee

Neuvième colloque francophone sur les Sondages

Emmanuel Gros

13 octobre 2016

Institut national de la statistique et des études économiques (INSEE, France)

Sommaire

Aux origines du profilage à l'Insee

Entreprise : vision juridique *versus* concept économique

- L'unité légale (UL) : une unité qui présente de nombreux avantages en pratique...
  - une unité définie de manière directe et non ambiguë ;
  - une inscription obligatoire au répertoire Sirene géré par l'Insee ;
  - Les UL sont tenues légalement d'effectuer de nombreuses déclarations administratives ⇒ informations administratives très riches au niveau UL ;

  ... mais qui ne constitue qu'une vision juridique de l'entreprise.
- Le concept économique d'entreprise : une définition uniforme depuis 1993 en Europe (règlement n°696/93)...

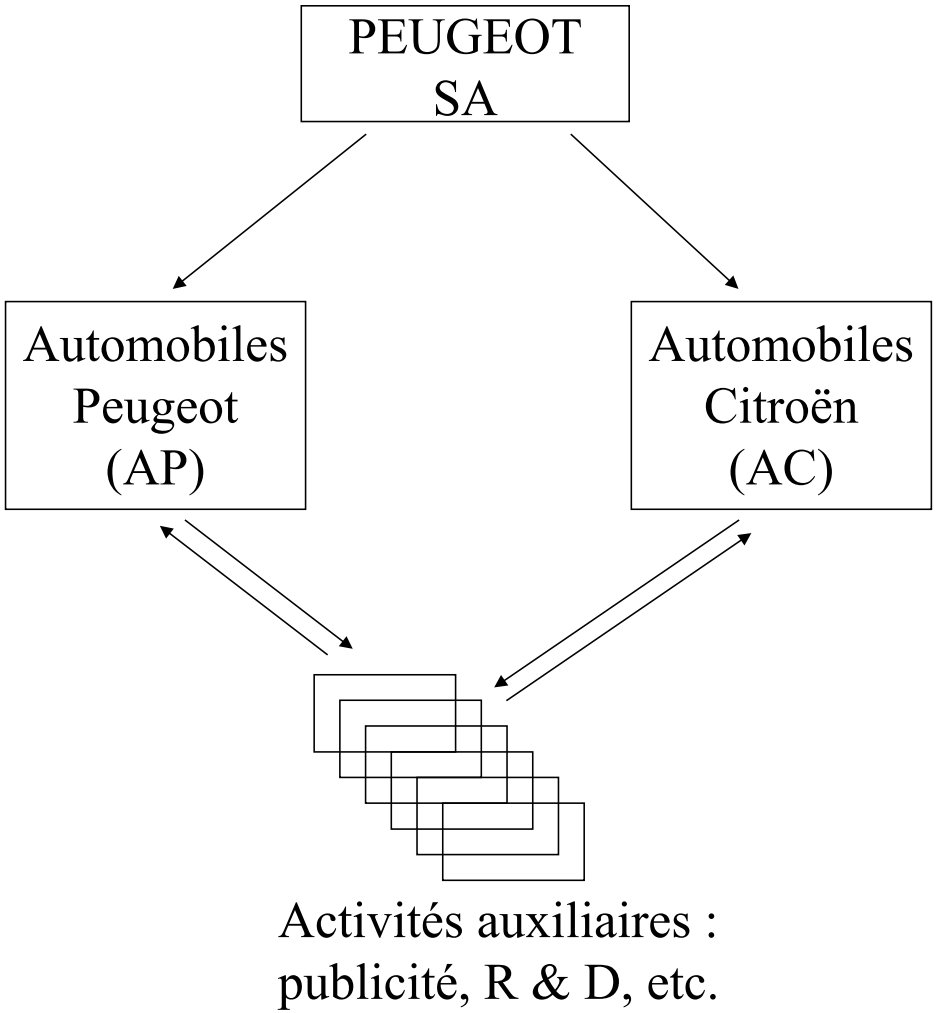
  > *« L'entreprise correspond à la plus petite combinaison d'unités légales qui constitue une unité organisationnelle de production de biens et de services jouissant d'une certaine autonomie de décision, notamment pour l'affectation de ses ressources courantes. »*

  ... mais qui nécessite de construire des unités statistiques non triviales.

Les statistiques d'entreprises jusqu'à la fin des années 90

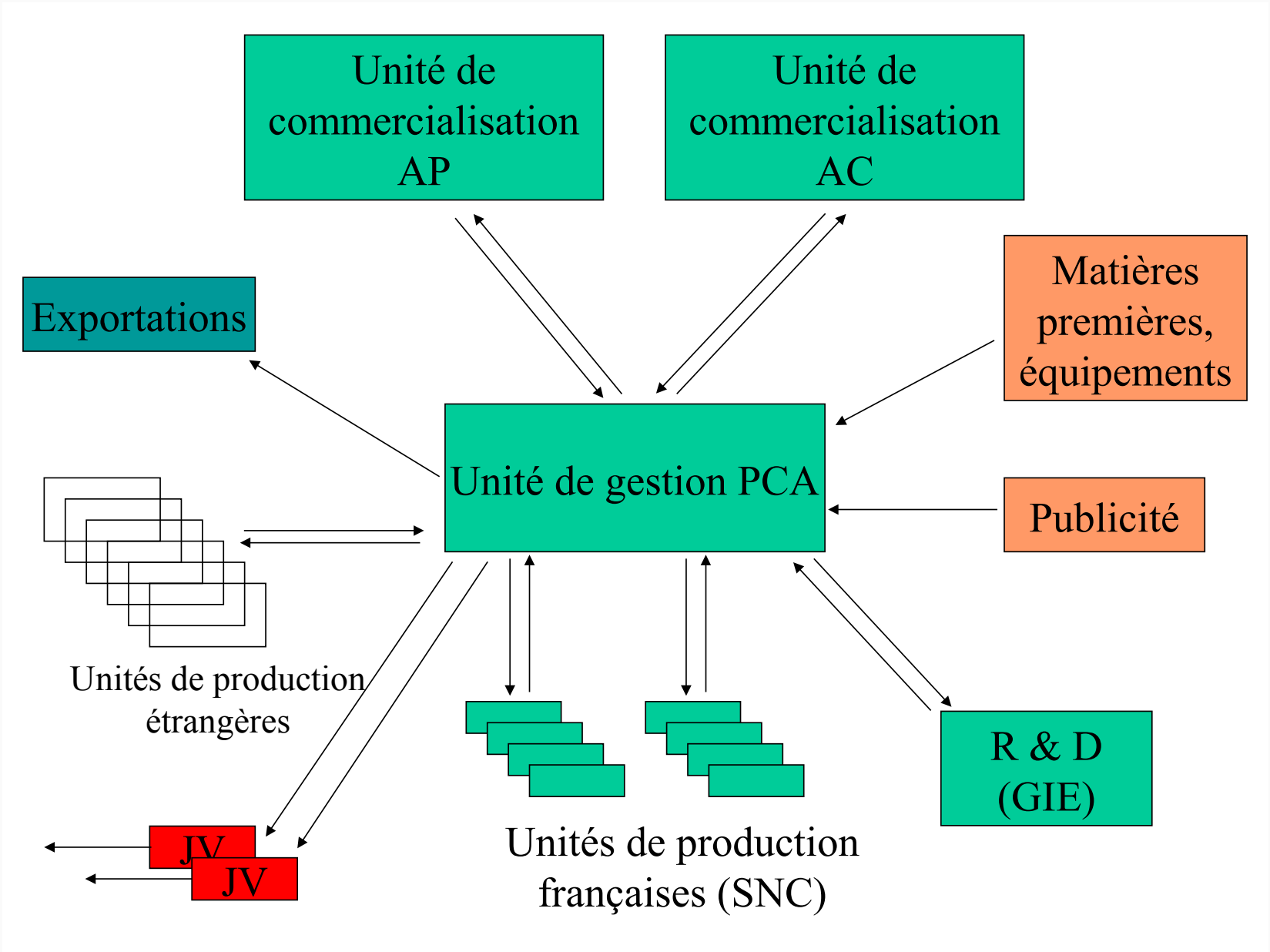
- La statistique d'entreprise s'est longtemps appuyée exclusivement sur la notion d'entreprise au sens juridique d'unité légale :
  - pour d'évidentes raisons de simplicité ;
  - conduisait à un certain nombre de problèmes, liés aux groupes d'entreprises : sous-estimation de la concentration du tissu productif, répartition sectorielle des agrégats perturbée, cohérence temporelle des statistiques en cas de réorganisation d'un groupe...
  - ... mais ces problèmes n'étaient pas rédhibitoires tant que l'importance des groupes dans l'économie et leur complexité restaient limitées.
- À partir des années 1990, l'importance croissante des groupes dans l'économie, leur déploiement multinational et leurs recompositions régulières et de plus en plus fréquentes ont rendu cette approche de moins en moins valide $\Rightarrow$ cf. la restructuration du groupe Peugeot en 1998.

Le groupe Peugeot avant la restructuration

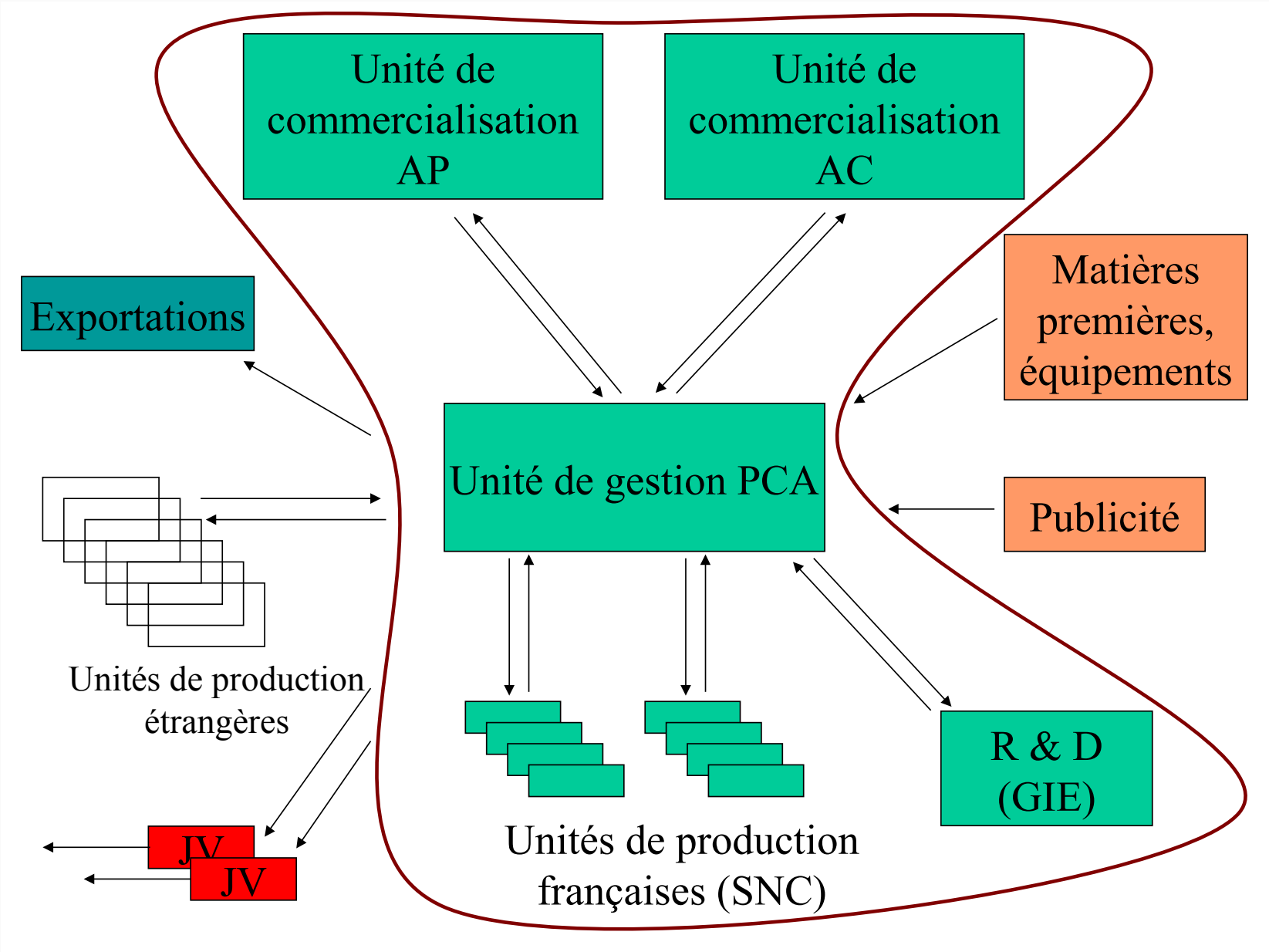


Les sociétés AP et AC produisaient et vendaient des voitures. Elles possédaient leurs propres moyens de production et avaient leurs employés.

La nouvelle organisation du groupe Peugeot



Création d'une entreprises profilée « DAF-PSA »



Vers de réelles statistiques d'entreprises avec le profilage

- Il était donc de plus en plus impératif de produire des statistiques d'entreprises intégrant le concept économique d'entreprise...
- ... et ce d'autant plus qu'en 2008, la Loi de Modernisation de l'Économie (LME) reprend la définition économique de l'entreprise du règlement européen n°696/93, et définit des catégories d'entreprises.

$\Rightarrow$ Création à l'Insee d'une division « Profilage et traitement des grandes unités » chargée de l'opération de « profilage » des entreprises.

Le profilage : principe et mise en œuvre à l'Insee

Objectif & cibles du profilage

- Le profilage consiste, au sein des groupes :
  - à identifier la ou les entreprise(s) pertinente(s) au sens de la LME ;
  - et à reconstituer leurs comptes consolidés.
- La stratégie de profilage sera fonction de la taille des groupes, répartis en 3 « cibles » :
  - cible 1 (~60 groupes) : les plus grands groupes présents en France, qui feront l'objet d'un profilage « sur mesure » *via* des réunions annuelles entre profileurs et interlocuteurs des groupes ;
  - cible 2 (~80 000 groupes) : les groupes petits (moins de 250 salariés) ou simples (composés de moins de 3 UL) seront assimilés à une seule entreprise et feront l'objet d'un algorithme de consolidation automatique ;
  - cible 3 (~5 000 groupes) : les groupes de taille ou de complexité intermédiaire feront à terme l'objet d'un profilage « semi-automatique ».

Le profilage des grands groupes de la cible 1

1. Définition du contour des entreprises profilées :
   - liens entre entreprises et unités légales ;
   - entreprises définies sur le périmètre France.
2. Construction des comptes consolidés des entreprises :
   - méthode *bottom-up* privilégiée : agrégation des comptes des UL à partir des liasses fiscales puis retraitement des flux intra-groupes (fournis par les groupes) ;
   - Méthode mixte : le groupe transmet des comptes aux normes IFRS sur le périmètre France pour les variables non-additives et le profileur les adaptent aux normes françaises ; pour les variables additives, on conserve la somme des liasses fiscales des UL ;
3. Obtention des réponses aux ESA & EAP au niveau des entreprises profilées si possible ;
4. gestion courante des unités profilées, *via* des réunions annuelles avec les groupes.

Le profilage des groupes des cibles 2 et 3 à court terme

- On considère qu'un groupe = une entreprise. Les données restent collectées au niveau des UL et les données des entreprises s'en déduisent *via* un algorithme de consolidation automatique :

❶ Identification automatique du caractère auxiliaire, commercial ou productif des UL d'un groupe :

  - à partir de l'activité principale exercée par chaque UL...
  - ... et en prenant en compte l'importance, en termes d'emploi dans le groupe, des UL pour déterminer les activités auxiliaires ;

❷ Consolidation automatique du chiffre d'affaires en neutralisant les flux intra-groupes :

  - on retire au chiffre d'affaires du groupe celui des unités légales exerçant une activité auxiliaire...
  - ... ainsi que les flux entre certaines unités légales commerciales et/ou productives.

Le profilage des groupes des cibles 2 et 3 à plus long terme

- À plus long terme, organisation d'une enquête annuelle portant sur les groupes de la cible 3 :
  - afin de collecter les flux intra-groupes pour les groupes les plus importants de la cible 3, et de pouvoir les estimer de façon plus précise pour les autres groupes de la cible 3 ;
  - les résultats de cette enquête serviront également à améliorer l'algorithme de consolidation automatique utilisé sur la cible 2.

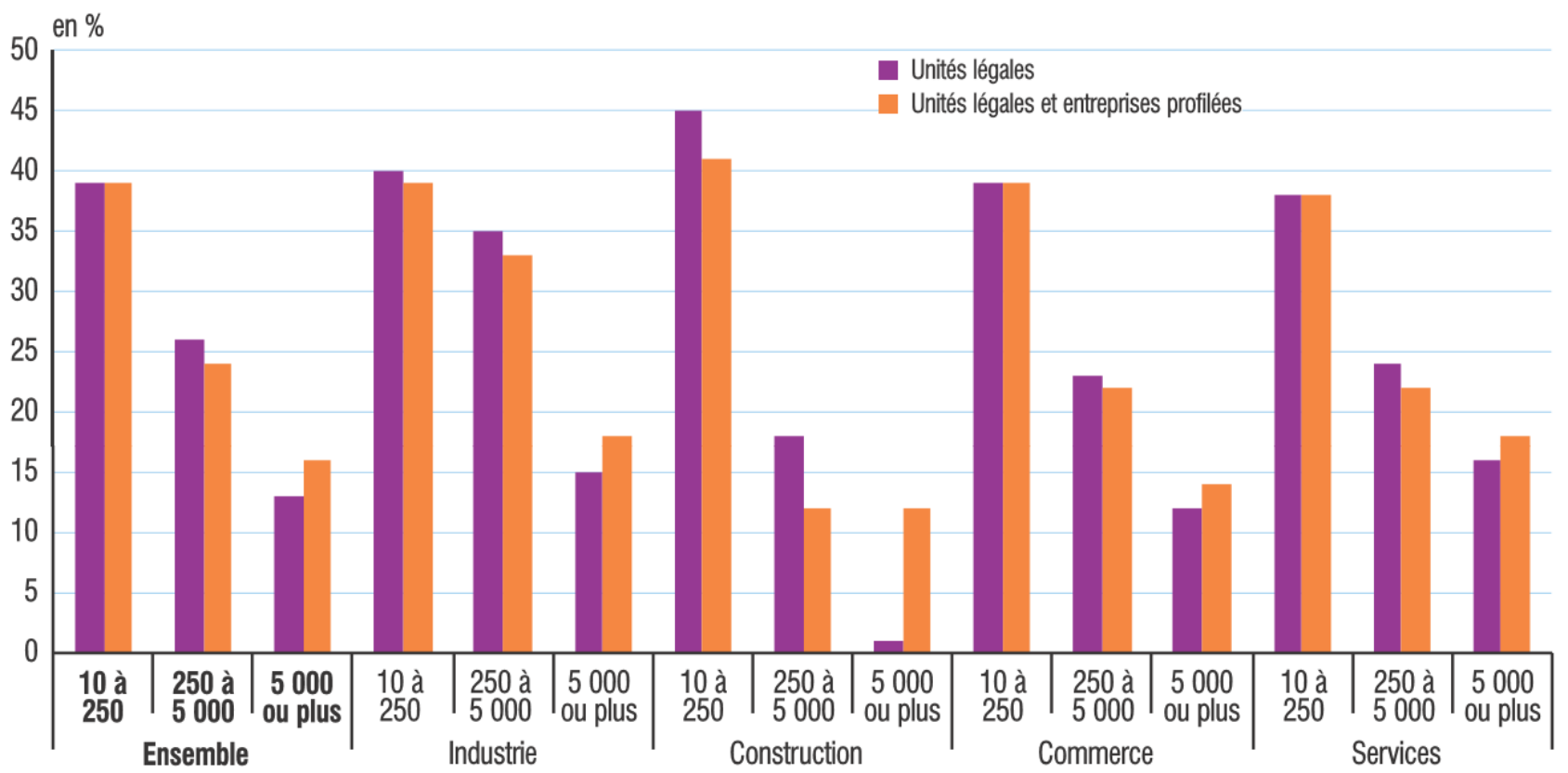
L’impact du profilage sur les statistiques structurelles d’entreprises

L’impact du profilage de la cible 1

- 43 grands groupes traités aujourd’hui à l’Insee, comportant 3 500 unités légales et profilés en 105 entreprises qui représentent :
  - 120 milliards d’euros de valeur ajoutée, soit 12 % de la VA totale ;
  - un million d’emplois salariés, soit 8 % de l’emploi salarié total.
- Trois principaux résultats :
  - une plus forte concentration de l’appareil productif ;
  - une réallocation importante entre secteurs ;
  - Du fait de la consolidation des comptes, une vision plus cohérente de l’appareil productif.

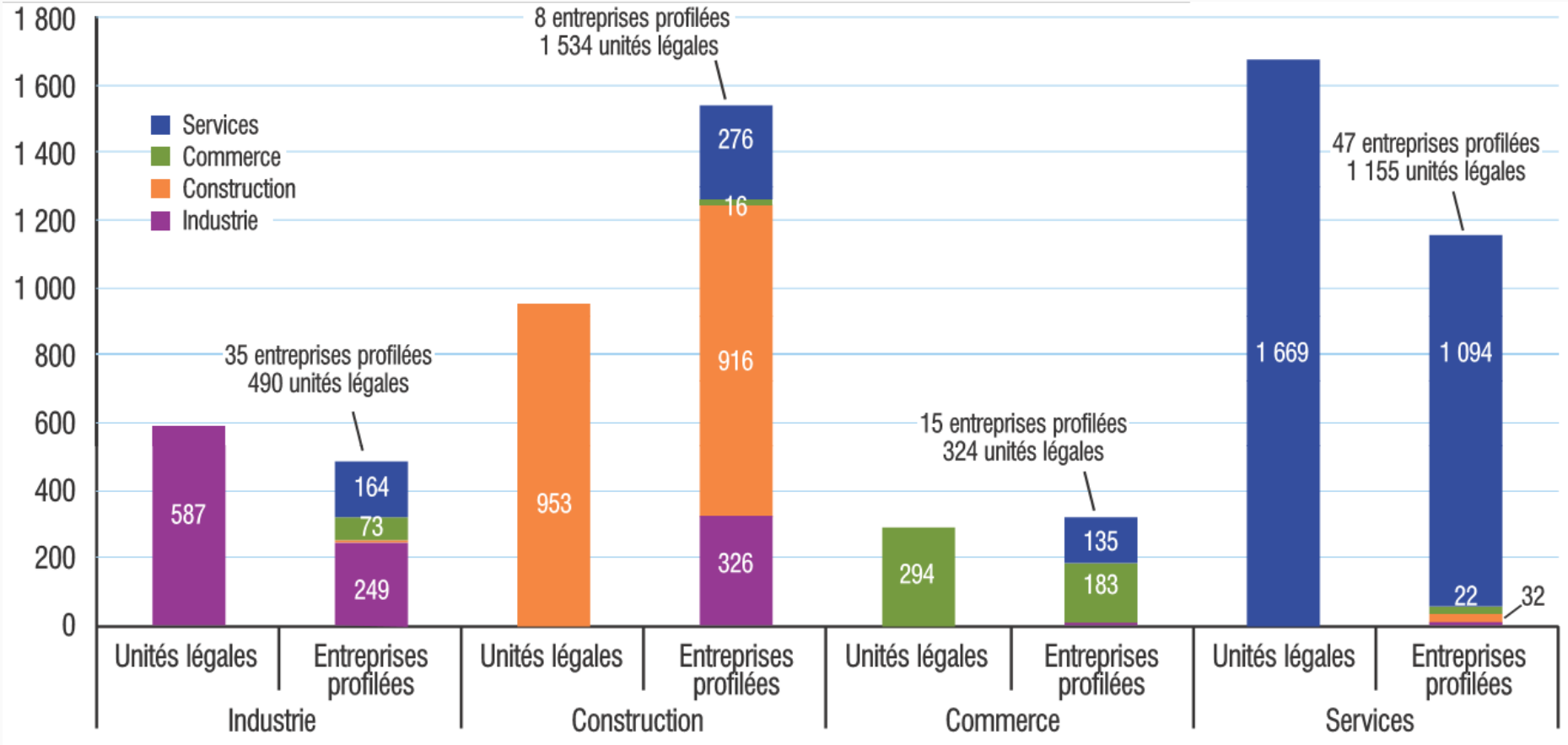
Profilage de la cible 1 : l'effet de concentration

- En termes d'effectifs, sur le million d'emplois salariés des 43 groupes profilés :
  - la moitié des effectifs étaient dans des UL de 5000 salariés ou plus…
  - … tandis que 94 % sont dans des EP de 5 000 salariés ou plus.



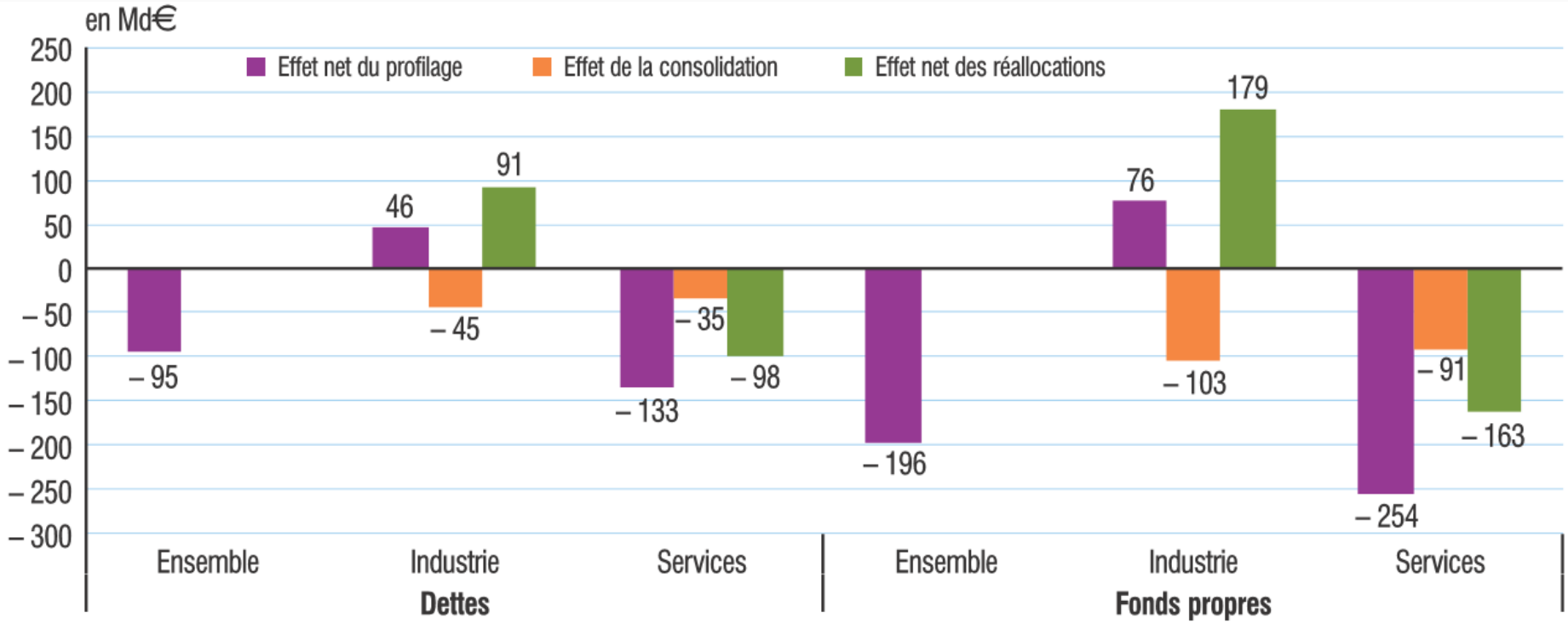
Profilage de la cible 1 : les réallocations sectorielles

- 30 % des unités légales des 43 groupes profilés « changent de secteur »



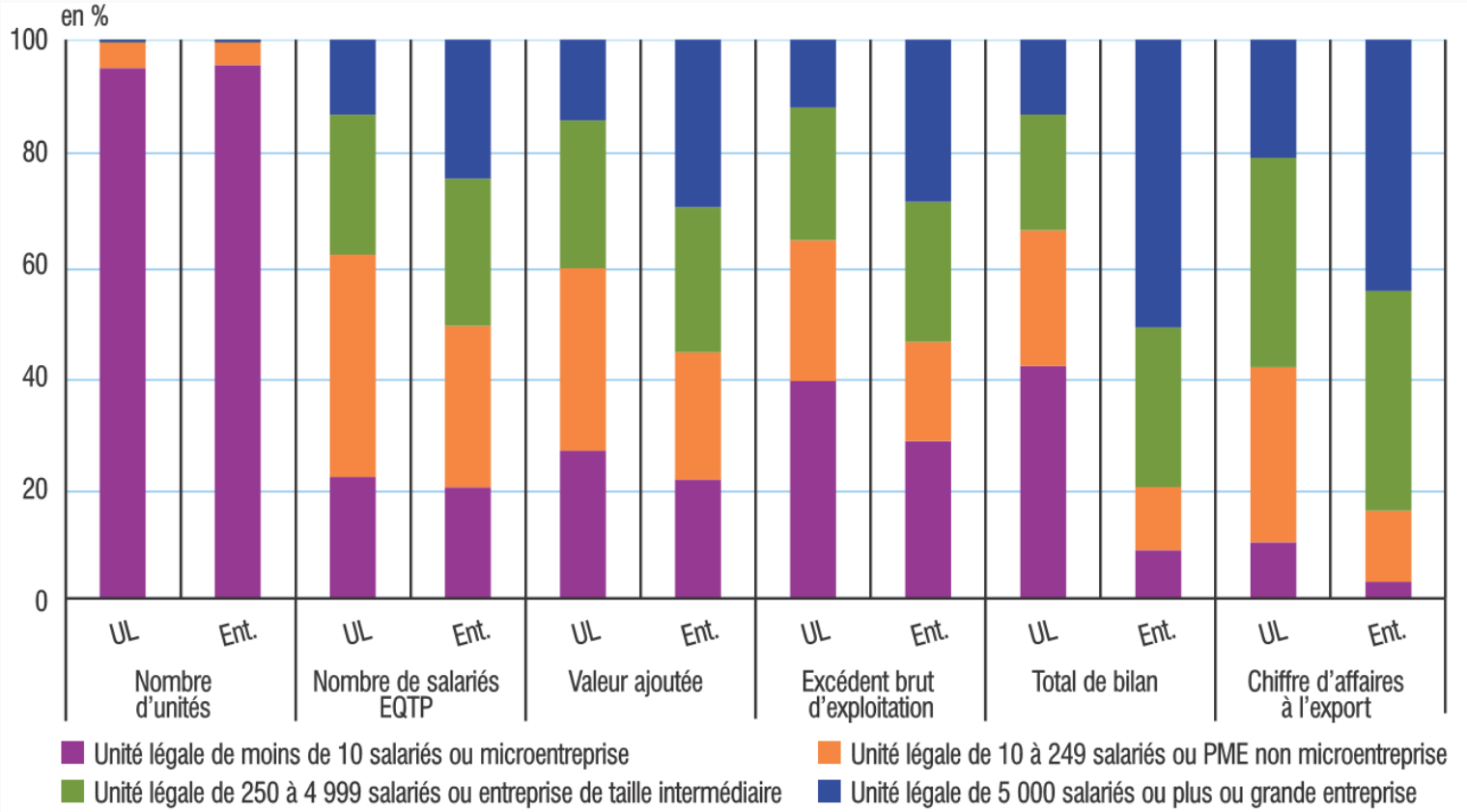
Profilage de la cible 1 : une vision plus cohérente de l'économie

- Après consolidation le CA est réduit de 56 Md€, soit 13 % du CA total des UL des groupes profilés, et 1,5 % du CA total du champ.
- Une vision plus cohérente de l'appareil productif : par exemple, transfert des dettes et fonds propres des services vers l'industrie.



Une première évaluation de l'impact global du profilage

- Une amplification des résultats observés sur la cible 1 :
  - un CA consolidé qui diminue de 220 Md€, soit ~6 % du CA du champ ;
  - l'intégration des sociétés tertiaires renforce le poids de l'industrie et de la construction ;
  - Un tissu productif nettement plus concentré qu'il n'y paraît :



L’impact du profilage sur le plan de sondage des enquêtes structurelles d’entreprises

Le dispositif Esane

- Processus de production des statistiques structurelles d'entreprises :
  - basé sur une exploitation intensive de données administratives (déclarations annuelles sur les bénéfices adressées par les entreprises à la Direction générale des Finances publiques, déclarations annuelles de données sociales)...
  - ...complétées par des enquêtes permettant d'obtenir des informations cruciales non disponibles dans les données administratives.

- Jusqu'à présent, les statistiques structurelles d'entreprises étaient élaborées en UL, SAUF pour les EP de la cible 1 qui étaient intégrées au dispositif, tant au niveau de la collecte que de la production des résultats.

L'impact du profilage sur le dispositif Esane

- À compter des résultats 2016, le système Esane va prendre en compte l'intégralité des entreprises profilées et produire des résultats en EP :
  - pour les grandes entreprises de la cible 1, la collecte s'effectuera toujours directement au niveau de l'EP ;
  - pour les entreprises des cibles 2 et 3, unité de collecte et unité statistique vont en revanche différer : la collecte s'effectuera toujours au niveau UL, avant une consolidation automatique des données par EP et une diffusion en EP ;
  - les UL indépendantes constituent des entreprises à part entière et ne posent aucun problème.

⇒ Nécessité d'adapter le plan de sondage des enquêtes structurelles d'entreprise à ce nouveau paradigme.

Les enquêtes structurelles d'entreprise ESA et EAP

- Il s'agit des principales enquêtes auprès des entreprises réalisées par l'Insee :
  - ESA, Enquête Sectorielle Annuelle :

    Champ : Industrie agroalimentaire, construction, commerce, services et transport

    # 116 000 unités échantillonnées en France métropolitaine

  EAP, Enquête Annuelle de Production :

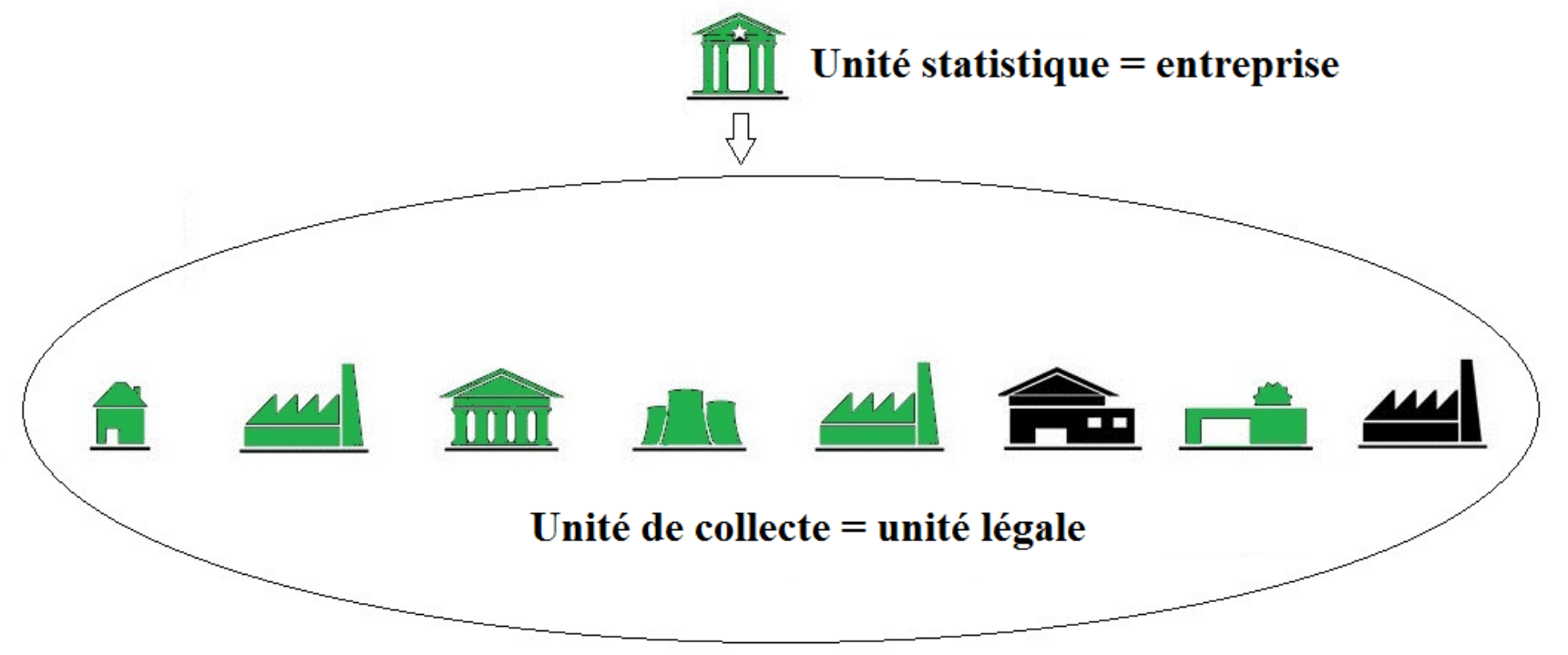
    Champ : industries extractives, industrie manufacturière, énergie

    # 35 000 unités échantillonnées en France métropolitaine

- Objectif principal : obtenir la ventilation du chiffre d'affaires des entreprises selon leurs différentes activités, qui permet ensuite de déterminer leur Activité Principale Exercée (APE).
- Échantillons d'UL sélectionnés par SAS stratifié.

Unité de collecte *versus* unité statistique

- Pour les cibles 2 et 3, les unités statistiques (entreprises) seront donc dorénavant différentes des unités de collecte (unités légales) : le nouveau plan de sondage peut être vu comme un plan de sondage en grappes stratifié.



Contraintes sur le nouveau plan de sondage

- Le nouveau plan de sondage pour une diffusion en entreprises avec une collecte en unités légales a été conçu afin de respecter les contraintes suivantes :
  - le nombre d'unités légales retenues dans chaque échantillon doit rester sensiblement inchangé ;
  - les échantillons doivent être optimisés pour une diffusion en entreprise, afin d'estimer le plus précisément possible le chiffre d'affaires total ;
  - pour cette optimisation, on s'intéresse non seulement à la précision globale, mais également à la précision des estimations par secteur fin et par secteur $\otimes$ tranche de taille ;
  - idéalement, on souhaite pouvoir exploiter ces enquêtes pour produire des résultats en unités légales avec une qualité acceptable.

Les données utilisées – base de sondage

- La base de sondage pour la sélection de ce nouvel échantillon a été constituée en s'appuyant :
  - d'une part sur les bases de sondages en unités légales utilisées pour le tirage des enquêtes ESA et EAP 2015 ;
  - d'autre part sur des données administratives relatives aux liens financiers entre unités légales en 2013 (LIFI 2013), afin d'identifier les unités légales indépendantes et les entreprises composées de plus d'une unité légale (cibles 2 et 3).

Les données utilisées – caractéristiques des entreprises

- Certaines caractéristiques sont issues de LIFI :
  - l'activité principale de l'entreprise : fonction du poids économique et de l'activité principale de chaque unité légale composant l'entreprise ;
  - la localisation géographique : celle de la tête de groupe.
- D'autres sont obtenues en agrégeant les caractéristiques des unités légales de l'entreprise qui sont dans le champ des enquêtes :
  - le chiffre d'affaires ;
  - l'effectif salarié ;
  - le total de bilan.

Objectif de la partie exhaustive

- La constitution de la partie exhaustive de l'échantillon répond à un triple objectif :
  - respecter la structure de l'exhaustif actuel : seuils absolus de chiffre d'affaires ou d'effectifs, ou taux de couverture en termes de chiffre d'affaires ;
  - Conserver une taille d'exhaustif en unités légales et une répartition entre ESA et EAP sensiblement identiques à celles observées actuellement ;
  - limiter la variabilité du nombre d'unités légales interrogées au total.

Définition de l'exhaustif (1)

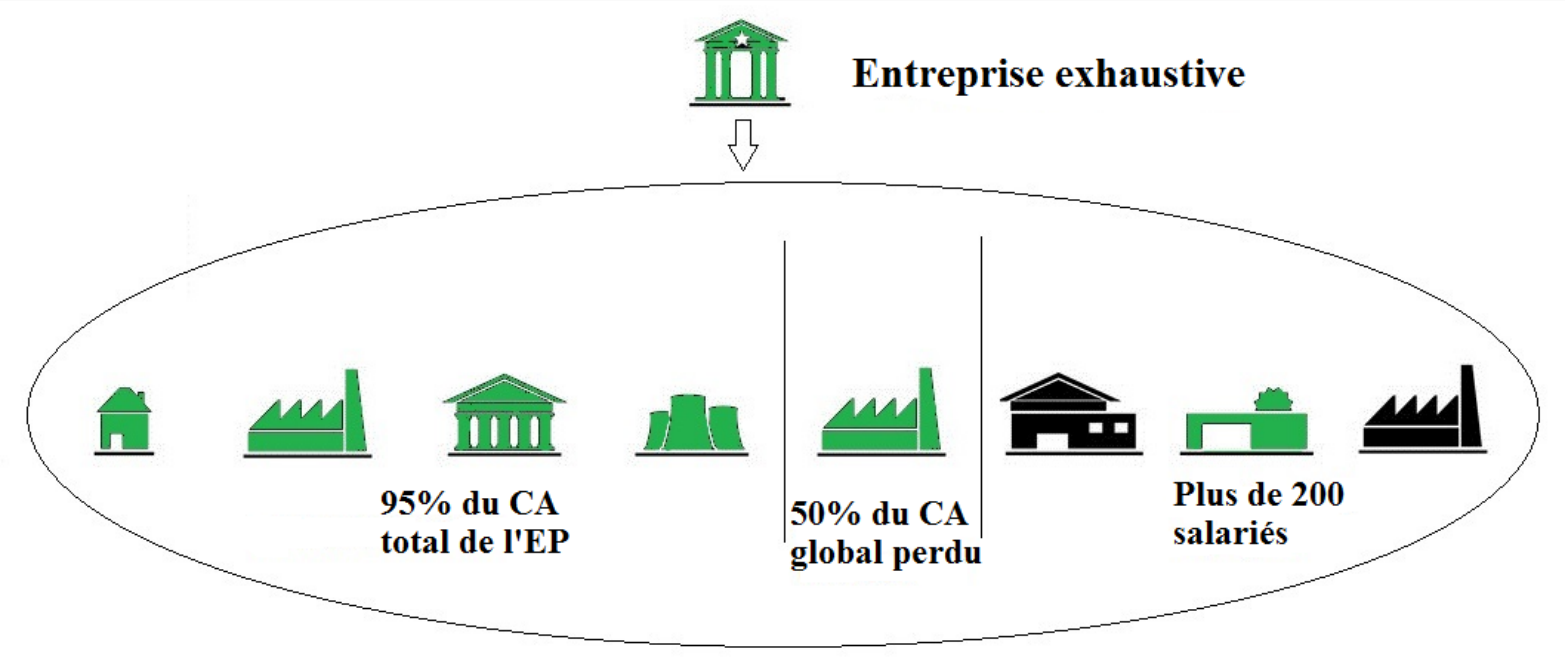
- Pour ce faire, l'exhaustif est dans un premier temps défini selon les règles suivantes :
  - les seuils d'effectifs salariés, de chiffre d'affaires et de total de bilan actuels sont appliqués, puis modulés en fonction du taux de couverture en termes de CA observé actuellement dans chaque secteur d'activité ;
  - les entreprises composées de plus de 20 UL, employant plus de 200 salariés ou réalisant plus de 50 M€ de chiffre d'affaires sont enquêtées exhaustivement.

Définition de l'exhaustif (2)

- Afin de réduire la taille, trop importante, de l'exhaustif ainsi obtenu :
  - au sein de chaque entreprise, on réalise un « cut-off à 95% du chiffre d'affaires » : les plus petites unités légales représentant, en cumulé, moins de 5% du chiffre d'affaires de l'entreprise sont exclues de l'échantillon ; elles ne seront pas enquêtées mais traitées par imputation ;
  - les unités les plus importantes ainsi exclues sont cependant réintégrées à l'exhaustif.

Exemple d'une entreprise exhaustive et de ses unités légales

- Les unités légales en vert sont enquêtées parce que :
  - elles appartiennent au cut-off à 95% du chiffre d'affaires de l'entreprise ;
  - elle pèsent, en cumulé, moins de 5% du chiffre d'affaires de l'entreprise mais emploient plus de 200 salariés ou réalisent un CA important.



Stratification, domaines de publication et allocations

- Pour la partie sondée, les strates résultent du croisement entre :
  - le secteur d'activité des entreprises au niveau fin (5 positions) de la Nomenclature d'Activité Française (NAF) ;
  - la taille des entreprises mesurée par leur effectif salarié (9 tranches de tailles)
- Deux domaines de publication sont pris en compte :
  - secteurs d'activité (au niveau fin de la NAF) ;
  - secteurs $\otimes$ tranches de taille (croisement entre le secteur d'activité en 3 positions et la tranche d'effectif en 4 positions)

Allocation de Neyman

- On cherche donc un plan de sondage qui permette d'estimer le plus précisément possible le chiffre d'affaires total des entreprises, sous une contrainte de nombre d'unités légales à enquêter fixé.

  $\Rightarrow$ Ceci correspond à l'allocation de Neyman suivante :

$$\begin{cases} \min\limits_{n_1,\dots,n_H} \mathbb{V}_p[\hat{CA}_\pi] \\ s.c. \sum\limits_{h=1}^{H} C_h n_h = N_{UL} \\ s.c.\ n_h \leq N_h \end{cases}$$

  où les coûts $C_h = \bar{N}_{UL,h}$ correspondent au nombre moyen d'unités légales par entreprise au sein de la strate h.

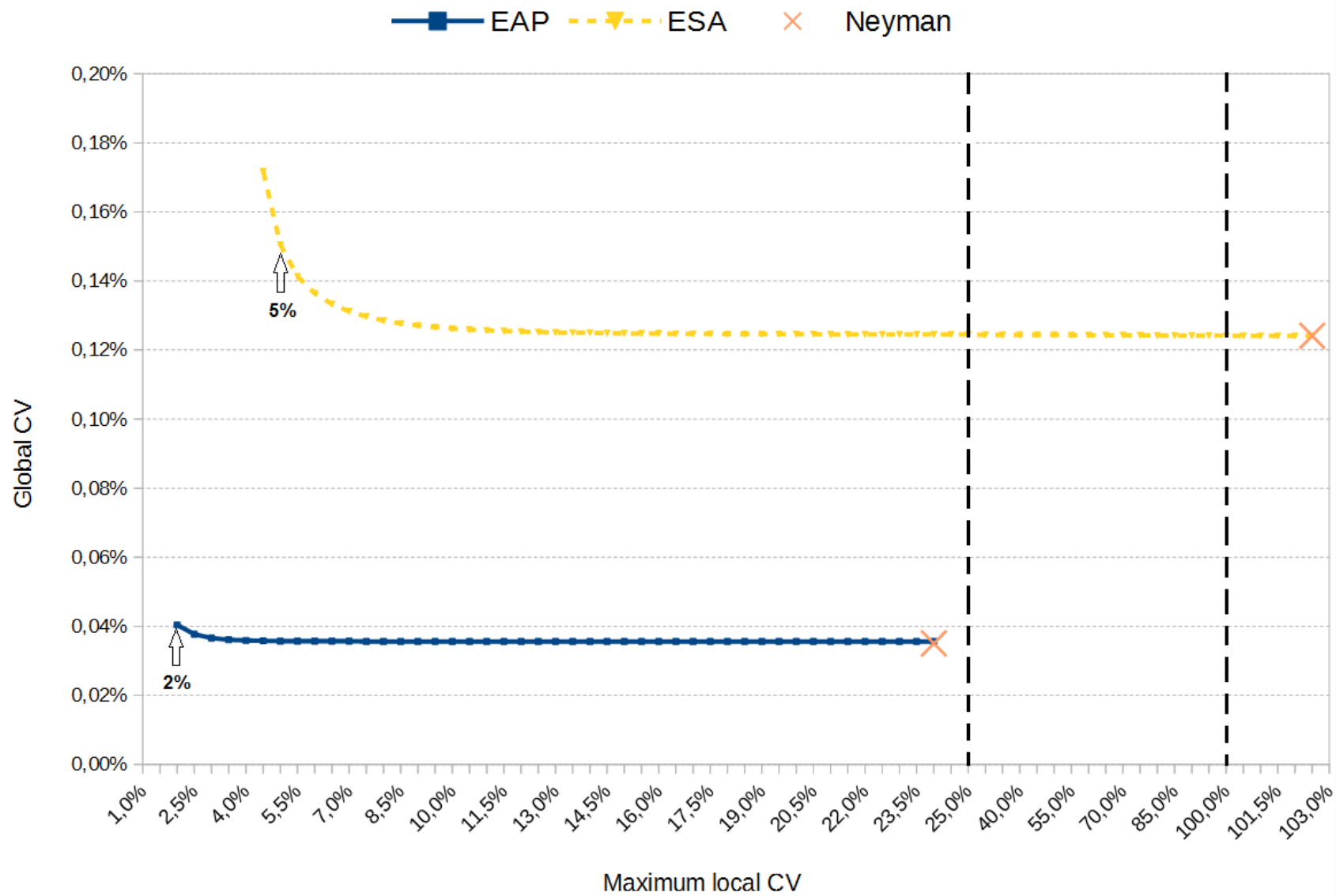
Allocation de Neyman sous contraintes de précision locale

- On souhaite en outre contrôler la précision locale des estimations sur les deux domaines de publication. Pour ce faire, l'algorithme de Koubi & Mathern (2009) a été adapté afin de prendre en compte des coûts par strate :

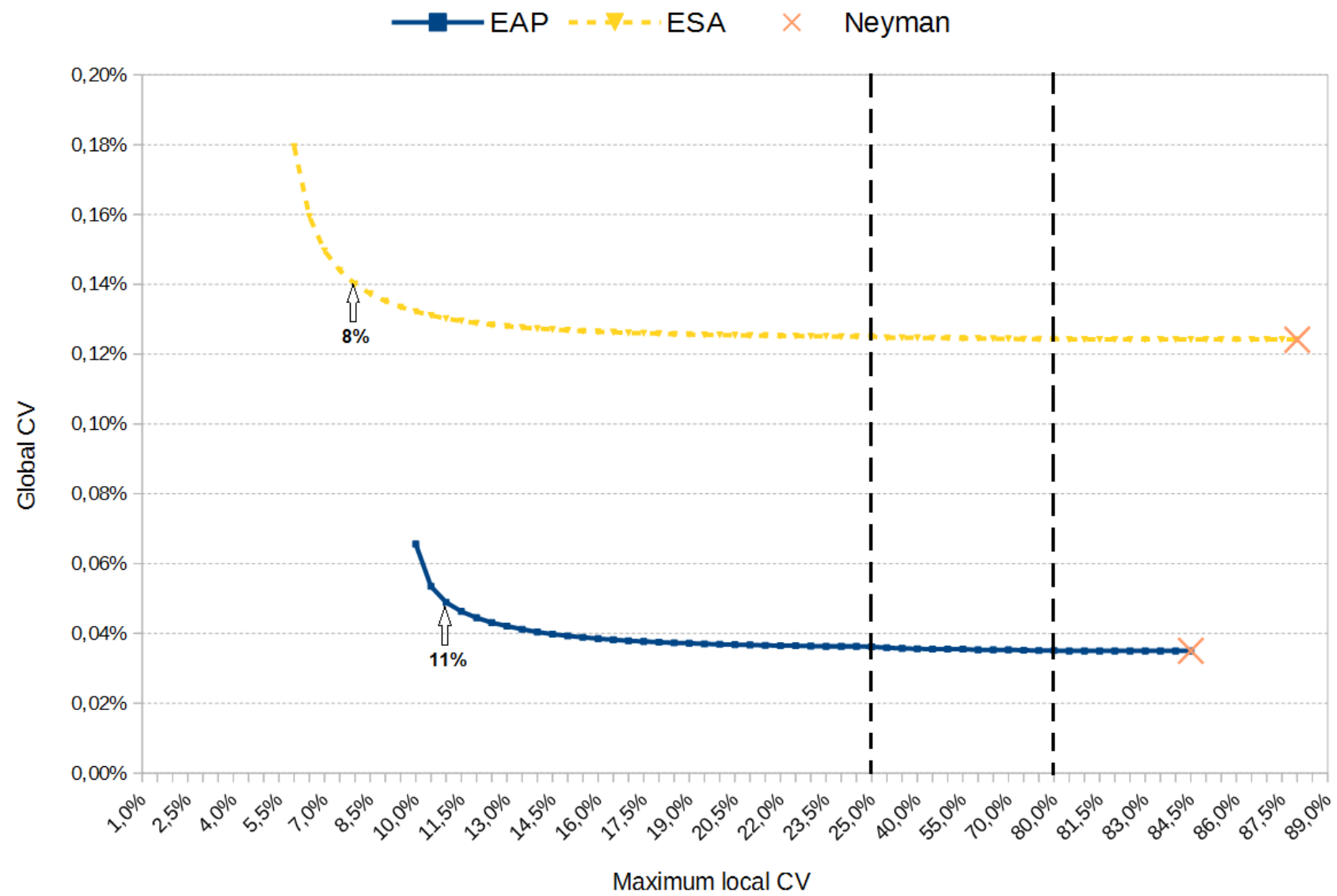
$$\begin{cases} \min\limits_{n_1,\ldots,n_H} \mathbb{V}_p[\hat{CA}_\pi] \\ s.c. \sum\limits_{h=1}^{H} C_h n_h = N_{UL},\ n_h \leq N_h \\ s.c. \max\limits_{d \in D} CV_d \leq CV_{loc} \end{cases}$$

avec D l'ensemble des domaines de publication et CV_{loc} la contrainte de précision locale.

Frontières d’efficacité – estimations par secteur fin



Frontières d'efficacité – estimations par secteur ⊗ taille



Estimateur du nombre d'unités légales à enquêter

- Cette allocation de Neyman intègre des coûts moyens par strate, ce qui conduit en moyenne au nombre d'unités légales voulu N_{UL}.
- Mais le nombre d'unités légales qui seront réellement enquêtées est aléatoire et dépend de l'échantillon sélectionné :

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

avec $N_{UL,k}$ le nombre d'unités légales de l'entreprise k.

- Cette quantité peut se ré-écrire comme l'estimateur d'Horvitz-Thompson de la variable $z_k = \pi_k N_{UL,k}$:

$$\hat{N}_{UL} = \sum_{h=1}^{H} \sum_{k \in S_h} N_{UL,k} = \sum_{h=1}^{H} \sum_{k \in S_h} \frac{z_k}{\pi_k} = \hat{z}_\pi$$

Variance de cet estimateur et variabilité des résultats

- En conséquence, la variance de cet estimateur est :

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

avec $S^2_{N_{UL},h} = \frac{1}{N_h-1}\sum_{k\in U_h}\left(N_{UL,k} - \bar{N}_{UL,h}\right)^2$ la dispersion de $N_{UL,k}$ dans la strate h.

	Total	*Exhaustif*	*Partie échantillonnée*
N_{Ent}	109 700	39 500	70 200
$\mathbb{E}_p\left[\hat{N}_{UL}\right]$	151 000	76 000	75 000
$CI_{95\%}\left(N_{UL}\right)$	[150 830 ; 151 170]	.	[74 830 ; 75 170]

Table 1: Résultats relatifs au nombre d'entreprises à tirer (N_{Ent}) et au nombre d'unités légales à enquêter (N_{UL}).

Précision pour une diffusion de niveau entreprise

	Domaine d'estimation					
	Secteur fin (1)			Secteur ⊗ taille (2)		
Distribution	$n_{ent,(1)}$	$n_{ent,(2)}$	$n_{ent,mix}$	$n_{ent,(1)}$	$n_{ent,(2)}$	$n_{ent,mix}$
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% Médiane	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%

Table 2: Distribution des CVs des estimations de CA de niveau entreprise selon l'allocation et le domaine considérés (sans la strate exhaustive des plus de 200 salariés pour le domaine secteur ⊗ taille).

Comparaison avec le dispositif en unités légales

	Domaine d'estimation			
	Secteur fin		Secteur ⊗ taille	
Distribution	$n_{UL,2015}$	$n_{ent,mix}$	$n_{UL,2015}$	$n_{ent,mix}$
100% Max	47,4%	**23,1%**	48,5%	**43,1%**
90%	7,5%	**6,3%**	12,8%	**12,5%**
75% Q3	3,8%	4,4%	5,4%	8,9%
50% Médiane	1,8%	2%	1,3%	4,2%
25% Q1	0,6%	0,8%	0%	0,2%
10%	0,1%	0,2%	0%	0%
0% Min	0%	0%	0%	0%

Table 3: Distribution des CVs des estimations de CA selon le niveau (UL/EP) du dispositif et le domaine considérés (sans la strate exhaustive des plus de 200 salariés pour le domaine secteur ⊗ taille).

Précision pour une diffusion de niveau unité légale

	Domaine d'estimation			
	Secteur fin		Secteur $\otimes$ taille	
Distribution	$n_{UL,2015}$	$n_{ent,mix}$	$n_{UL,2015}$	$n_{ent,mix}$
100% Max	47,4%	**14,9%**	48,5%	**38,3%**
90%	7,5%	**5,9%**	12,8%	**10,6%**
75% Q3	3,8%	3,9%	5,4%	7,3%
50% Médiane	1,8%	2,1%	1,3%	3,4%
25% Q1	0,6%	0,9%	0%	0,6%
10%	0,1%	0,2%	0%	0%
0% Min	0%	0%	0%	0%

Table 4: Distribution des CVs des estimations de CA de niveau unité légale selon le plan de sondage (UL/EP) et le domaine considérés (sans la strate exhaustive des plus de 200 salariés pour le domaine secteur $\otimes$ taille).

Conclusions et futurs chantiers méthodologiques

Conclusions

- La prise en compte du concept économique d'entreprise dans les statistiques structurelles constitue un changement de paradigme majeur...
- ... qui implique une refonte du plan de sondage des enquêtes structurelles annuelle...
  - unités de collecte différentes des unités statistiques ;
  - optimisation du plan de sondage pour une diffusion de niveau entreprise sous des contraintes pesant sur les unités de collecte ;
  - un échantillon permettant une exploitation au niveau unité légale avec une qualité satisfaisante.
- ... et qui va permettre une meilleure vision du tissu productif français.

Travaux méthodologiques à venir

- Des évolutions à venir sur le plan de sondage :
  - mise en œuvre d'un plan de sondage rotatif, avec renouvellement de l'échantillon par moitié ;
  - optimisation de l'allocation mixte, *via* des facteurs $(\alpha, 1-\alpha)$ « optimaux » en lieu et place des $(1/2, 1/2)$ actuels.
- Des défis méthodologiques à venir concernant les traitements post-collecte :
  - Problème relatif à la gestion des changements de contour des EP entre base de sondage et base de diffusion.
  - *Quid* de la correction de la non-réponse ? de la gestion des unités influentes ? du calage sur marges ?
  - Quels traitements post-collecte pour la production des résultats en UL à l'attention des comptes nationaux ? et *quid* de la cohérence globale entre résultats de niveau entreprise et résultats de niveau UL ?

Merci de votre attention !