L'IMPACT DU PROFILAGE SUR LA REFONTE DU PLAN DE SONDAGE DES ENQUÊTES SECTORIELLES ANNUELLES

Ronan LE GLEUT (*), Thomas MERLY-ALPA (*)

(*) Insee, Direction de la méthodologie et de la coordination statistique et internationale

ronan.le-gleut@insee.fr, thomas.merly-alpa@insee.fr

Mots-clés : Profilage, sondage en grappes, optimisation d'un plan de sondage

Résumé

Dans de nombreux pays de l'Union Européenne, les statistiques d'entreprise sont en grand changement. En effet, afin de répondre au règlement européen *Structural Business Statistics*, les instituts nationaux de statistiques se sont engagés à fournir à Eurostat des agrégats basés sur la notion économique d'entreprise profilée (EP), qui correspond à la plus petite combinaison d'unités légales (UL) qui constitue une unité organisationnelle de production de biens et de services jouissant d'une certaine autonomie de décision. A l'Insee, cela se traduit en particulier par une modification du plan de sondage des Enquêtes Sectorielles Annuelles, dont l'un des objectifs est de déduire l'activité principale d'une entreprise via la ventilation en branches de son chiffre d'affaires (CA). Dans ce nouveau contexte, les unités statistiques (EP) sont différentes des unités de collecte (UL), les réponses étant toujours collectées au niveau des UL pour une diffusion en entreprises. Le nouveau plan de sondage peut ainsi être vu comme un sondage en grappes, où une EP (grappe) est sélectionnée puis toutes les UL qui la composent sont interrogées. Cette méthode de tirage conduit donc à une variabilité du nombre d'unités légales mises en collecte. La refonte du plan de sondage a tout d'abord conduit à revoir les critères d'exhaustivité de l'enquête. Pour cela, les seuils historiques utilisés précédemment (en termes de CA, d'effectifs et de total de bilan) ont été conservés, en les modulant par un taux de couverture du CA à couvrir par activité. Les plus grosses EP en termes de salariés, de CA ou de nombre d'UL sont également forcées dans l'exhaustif. Malgré ces critères, certaines EP ont encore un CA atypique par rapport aux autres unités de leur strate de tirage. Trois méthodes sont alors exploitées afin d'identifier des unités atypiques dans chaque strate, et de les forcer dans l'exhaustif. L'étape suivante de la refonte de plan de sondage a consisté à revoir la méthode de calcul des allocations. La première contrainte de ce calcul portait sur le nombre d'UL à interroger, qui doit être peu variable et rester proche du volume d'unités actuellement mis en collecte. La seconde contrainte portait sur la précision des estimations des totaux de CA sur deux domaines de diffusion : l'APE (activité sur 5 positions) et le croisement groupe (activité sur 3 positions) x tranche d'effectifs. Ce papier présente donc la façon dont le plan de sondage des Enquêtes Sectorielles Annuelles a été réoptimisé afin d'avoir une bonne précision des estimations pour la diffusion en entreprise sous la contrainte d'un nombre fixe d'UL à enquêter.

Abstract

In many countries of the European Union, business statistics are undergoing great changes. In France, for instance, most of the business surveys are currently based on the observation of legal units that have a juridical definition. In order to comply with the Structural Business Statistics European regulation, business statistics will be more and more based on the economic notion of enterprise, which is the smallest combination of legal units that is an organisational unit producing goods or

services with a certain degree of autonomy. However, in most surveys, the data collection units remain the legal units, whereas the statistical units are now the enterprises. Consequently, the sample design of the surveys in this new paradigm can be seen as a two-stage cluster sampling. Enterprises are selected with a probabilistic mechanism, and then all legal units within those enterprises are included in the sample. This paper presents how the sample design of the French structural business survey was optimized, in order to obtain sufficiently precise estimates at the enterprise level under a constraint pertaining to the number of surveyed legal units.

1. Introduction et contexte

En France comme de nombreux autres pays, les statistiques d'entreprise évoluent avec le passage en diffusion de résultats basés sur la notion économique d'entreprise profilée (EP) en remplacement d'agrégats précédemment diffusés au niveau unité légale (UL), ceci afin de pouvoir répondre au règlement européen *Structural Business Statistics* (SBS, EEC 1993). L'objectif de ce règlement est produire des statistiques décrivant la structure, le comportement et les performances des entreprises dans l'ensemble de l'Union Européenne, pouvant être ventilées par secteur jusqu'à un niveau très détaillé, ainsi que par taille d'entreprise.

Cette différence de concept entre une UL, correspondant à une entitié juridique de droit public ou privé, et une EP, correspondant à la plus petite combinaison d'UL qui constitue une unité organisationnelle de production de biens et de services jouissant d'une certaine autonomie de décision, s'accompagne d'un grand changement dans le dispositif d'Élaboration des Statistiques Annuelles d'Entreprise (Esane). Ainsi, une importante opération de « profilage » a été mise en place à l'Institut de la Statistique et des Études Économiques (Insee) afin de pouvoir prendre en considération ce nouveau concept économique. L'objectif de cette opération est de définir des unités statistiques intermédiaires au groupe qui paraissent être les mieux appropriées pour une observation de l'activité économique. Pour cela, les entreprises des plus grands groupes (plus de 10 000 salariés) voient leur contour être défini « à la main » avec un profilage individuel et un entretien annuel avec des « profileurs » ; pour les autres groupes, un profilage automatique avec une consolidation des caractéristiques par un algorithme (Chanteloup, 2018) est effectué, avec l'idée qu'un groupe équivaut à une entreprise.

A l'Insee, deux enquêtes permettent de répondre au règlement SBS :

- L'Enquête Sectorielle Annuelle (ESA) qui couvre le champ des activités de commerce, de construction, de services et de transport ; Environ 116 000 unités légales sont enquêtées chaque année en France métropolitaine.
- L'Enquête Annuelle de Production (EAP) qui couvre le secteur industriel et pour laquelle environ 35 000 unités légales sont enquêtées chaque année.

L'un des principaux objectifs de ces deux enquêtes est de repérer les différentes activités exercées par les entreprises, via la ventilation de leur chiffre d'affaires (CA) en branche, et en déduire leur activité principale (APE). Jusqu'en 2015, les échantillons d'unités légales de ces deux enquêtes étaient sélectionnés selon un sondage aléatoire simple stratifié.

2. Méthodes

Étant donné que l'unité statistique (EP) est maintenant différente de l'unité de collecte (UL), le plan de sondage des Enquêtes Sectorielles Annuelles relatives aux années 2016 et suivantes peut être vu comme un tirage stratifié en grappes, où une EP est sélectionnée de façon aléatoire, puis toutes les UL qui composent cette EP sont inclues dans l'échantillon.

2.1. Définition de l'exhaustif

Dans les enquêtes auprès des entreprises, la part de l'exhaustif dans l'échantillon est souvent assez importante, jusqu'à représenter la moitié de l'échantillon pour certaines enquêtes. Afin de définir l'exhaustif de ce nouveau plan de sondage en EP, nous avons repris les seuils historiques utilisés dans les précédentes Enquêtes Sectorielles Annuelles en termes de CA, d'effectifs et de total de bilan. Les EP de plus de 200 salariés, de plus de 50 000 k€ de CA et celles composées de plus de 20 UL sont également forcées dans l'exhaustif. Ce dernier critère permet notamment de limiter la variabilité du nombre d'UL à interroger, les plus grosses EP étant systématiquement interrogées. Enfin, un cut-off à 95 % du CA au sein de chaque entreprise est effectué, ce qui permet d'éviter d'enquêter les UL avec un très faible CA (pour lesquelles on peut considérer qu'elles sont monoactives).

Cependant, les anciens seuils utilisés dans les précédentes enquêtes conduisaient à un volume d'UL exhaustives trop important, et ce malgré le cut-off à 95 % au sein de chaque EP. Ainsi, afin de limiter la taille de l'exhaustif, ces seuils historiques ont été modulés par un taux de couverture du CA à couvrir par activité, tout en conservant les plus grosses EP (en termes d'effectifs, de CA ou de nombre d'UL) dans l'exhaustif.

Ces premiers critères permettent ainsi de considérer dans l'exhaustif les plus grosses EP de la base de sondage. Cependant, certaines EP ont encore un CA atypique par rapport aux autres unités de leur strate de tirage (voir 2.2. pour la définition des strates). Cela peut être problématique pour le calcul des allocations (voir 2.3.), la dispersion du CA dans une strate pouvant être fortement impactée par une unité influente. Trois méthodes (voir encadré ci-dessous) sont alors exploitées afin d'identifier des unités atypiques dans chaque strate, et de les forcer dans l'exhaustif.

Détection des EP atypiques (avant le calcul des allocations mais après la définition de l'exhaustif)

La contribution à la variance :

Dans chaque strate, on calcule deux indicateurs relatifs à la variance. D'une part, on calcule la contribution d'une unité k à la dispersion du CA de la strate en faisant le rapport suivant :

$$CT_k = \frac{(CA_k - \overline{CA})^2}{\sum (CA - \overline{CA})^2}$$

avec $\overline{CA}$ la moyenne du CA dans la strate. On ne regarde que les unités pour lesquelles le CA est supérieur à la moyenne et ayant une contribution importante à la dispersion. D'autre part, on calcule ce même indicateur en multipliant par la taille de la strate. On obtient ainsi comme indicateur le nombre d'unités que représente une unité atypique en part de variance dans sa strate.

Ces deux approches sont complémentaires. En effet, dans le premier cas, on ne va identifier que les unités ayant une contribution forte (e.g. plus de 75 %) peu importe la taille de la strate. Mais dans certains cas, les CA de ces unités sont très élevés en comparaison aux autres unités, ce qui justifie de les forcer dans l'exhaustif. Dans le second cas, on identifie surtout les unités influentes des strates avec beaucoup d'unités. On identifie ainsi plus facilement des unités influentes ayant une contribution à la variance plus faible (de l'ordre de 20 % par exemple) mais représentant un grand nombre d'unités en part de variance (e.g. plus de 1 000 EP).

La méthode de Kokic et Bell (1994) :

Cette méthode de « winsorization » permet d'identifier des unités atypiques. Elle intervient normalement lors des traitements post-collecte d'une enquête. En plus de la dispersion dans

chaque strate, on prend ici également en compte le taux de sondage de la strate. Celui-ci n'étant pas connu avant le calcul des allocations, on reprend celui de l'année précédente en supposant que ce taux reste assez stable d'une année sur l'autre.

Une unité est alors considérée comme étant influente (de par son CA et selon le taux de sondage de la strate à laquelle elle appartient) lorsque son CA dépasse le seuil en sortie de la méthode, calculé pour l'estimation du total du CA sur l'ensemble de la population (et pas par domaine de diffusion).

L'algorithme des centres mobiles (k-means) :

Une dernière approche consiste à effectuer un k-means sur le CA de l'entreprise pour chaque activité (APE) au niveau EP. Les k-means permettent de déterminer, pour un nombre de classes fixé, une répartition des entreprises qui minimise la variance intra-classe du CA. Afin d'identifier des unités atypiques dans chaque APE, on impose ici de constituer deux classes : une classe contenant un volume « important » d'unités (classe 1), et une classe contenant les unités atypiques (classe 2).

Afin de limiter l'étude aux APE les plus problématiques (l'analyse se faisant « à la main »), on se propose les deux critères suivants :

1. plus l'écart entre les moyennes des deux classes est important, plus les unités atypiques identifiées dans la classe 2 sont éloignées des autres unités ayant la même APE ;
2. la taille de la classe 2 des unités atypiques doit être raisonnable (e.g. pas plus de 5 unités ou un certain pourcentage du nombre d'unités total dans l'APE).

La combinaison des deux premières méthodes permet de détecter des unités influentes dans chaque strate en prenant en compte à la fois le taux de sondage de la strate et le CA des unités. La troisième méthode intervient en validation afin d'être sûr de ne pas avoir manqué quelques cas atypiques au sein de leur activité.

Ces critères conduisent ainsi à un exhaustif d'environ 70 000 UL, 40 000 pour le champ de l'ESA – ce qui laisse envrion 76 000 UL à échantillonner dans la partie sondée – et 30 000 pour l'EAP – soit seulement 5 000 UL restant à échantillonner.

2.2. Stratification et domaines de diffusion

Les strates de tirage sont définies au niveau de l'entreprise en croisant l'APE et la tranche d'effectifs : [0 salarié], [1 – 5], [6 – 9], [10 – 19], [20 – 29], [30 – 49], [50 – 99], [100 – 199], [200 et plus].

Les domaines de diffusion correspondent à l'APE de l'entreprise et au croisement entre le groupe d'activité et la tranche d'effectifs regroupée (groupe x teff) : [0 – 9], [10 – 49], [50 – 199], [200 et plus].

2.3. Calcul des allocations

2.3.1. Nombre moyen d'UL à enquêter

La première contrainte du calcul des allocations porte sur le fait que le nombre d'UL obtenues au final dans l'échantillon doit être peu variable et rester proche du volume actuellement mis en collecte, et ce pour chacune des deux enquêtes ESA (environ 116 000 UL en Métropole) et EAP (35 000 UL). Des premières simulations ont conduit à la conclusion qu'il était nécessaire de séparer le champ des deux enquêtes au niveau entreprise afin de pouvoir respecter ces contraintes sur le

nombre d'UL à enquêter pour chacune des deux enquêtes. Afin de respecter ces contraintes, nous avons introduit des contraintes de coûts dans le calcul de l'allocation de Neyman :

Calcul de l'allocation de Neyman sous contraintes de coûts pour une enquête donnée (ESA ou EAP) :

Si l'on note y_k le CA de l'entreprise k , $\hat{t}_{y\pi}$ l'estimateur d'Horvitz Thompson du total de CA, $S^2_{y,h}$ la variance empirique de y_k dans la strate h , N_{UL} le nombre d'UL à sélectionner pour une enquête donnée (ESA ou EAP), $N_{UL,k}$ le nombre d'UL de la même enquête de l'entreprise k , n_h le nombre d'entreprises à sélectionner, N_h le nombre d'entreprises dans la strate h et $f_h = n_h/N_h$ le taux de sondage dans la strate h , nous devons résoudre :

$$\begin{cases} \underset{n_1,\dots,n_H}{Min}\ V[\hat{t}_{y\pi}] = \sum_{h=1}^{H} N_h^2 \frac{(1-f_h)}{n_h} S^2_{y,h} \\ s.c. \sum_{h=1}^{H} C_h n_h = N_{UL} \\ s.c.\ n_h \leqslant N_h \end{cases}$$

avec $C_h = Moy_{UL,h} = \frac{1}{N_h} \sum_{k \in U_h} N_{UL,k}$ le coût, i.e. le nombre moyen d'UL par entreprise pour une enquête donnée (ESA ou EAP) dans la strate h .

Les résultats sur le nombre d'entreprises sélectionnées et le nombre d'UL à enquêter sont donnés dans la partie 3.1.

2.3.2. Variabilité du nombre d'UL à enquêter

La formule précédente donne le nombre moyen d'UL à enquêter (le coût introduit dans l'optimisation étant un coût moyen par strate). Cela permet donc de respecter en espérance (i.e. sur tous les échantillons possibles) le bon nombre d'UL à enquêter. Cependant, le nombre exact reste aléatoire et peut varier d'un échantillon à l'autre. Nous allons donc maintenant nous intéresser à la variabilité de ce nombre.

Calcul de la variance du nombre d'UL à enquêter :

L'estimateur d'Horvitz Thompson de N_{UL} s'écrit :

$$\hat{N}_{UL} = \sum_{h=1}^{H} \sum_{k \in S_h} N_{UL,k} = \sum_{h=1}^{H} \sum_{k \in S_h} \frac{z_k}{\pi_k} \quad \text{avec} \quad z_k = \pi_k N_{UL,k} \quad \text{et} \quad \pi_k = \frac{n_h}{N_h}$$

Cet estimateur est sans biais par rapport à N_{UL} . Dans le cas d'un sondage aléatoire simple stratifié, la variance de cet estimateur peut s'écrire :

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

avec $S^2_{N_{UL},h} = \frac{1}{N_h - 1} \sum_{k \in U_h} (N_{UL,k} - Moy_{UL,h})^2$ la variance empirique de $N_{UL,k}$ dans la strate h .

La démonstration est donnée en annexe.

Les résultats sur la variabilité du nombre d'UL à enquêter sont donnés dans la partie 3.2.

2.3.3. Contraintes de précision locales

La variable d'intérêt des enquêtes étant la ventilation en branches du chiffre d'affaires, la seconde contrainte du calcul des allocations porte sur la précision des estimations des totaux de CA sur les deux domaines de diffusion. Pour cela, nous avons utilisé un algorithme proposé dans Koubi et Mathern (2009) permettant de réaliser une optimisation de Neyman (1934) et d'ajouter des contraintes sur un coefficient de variation (CV) maximal à ne pas dépasser sur les domaines de diffusion.

Afin de prendre en considération ces contraintes de précision tout en imposant un nombre moyen d'UL à enquêter fixe (contraintes de coûts), nous avons généralisé l'algorithme proposé par Koubi et Mathern (2009) :

Calcul de l'allocation de Neyman sous contraintes de coûts et de précisions locales :

Si l'on note C_h le coût, i.e. le nombre moyen d'UL par entreprise pour une enquête donnée dans la strate h , nous devons résoudre :

$$\begin{cases} \underset{n_1,\dots,n_H}{Min}\ V[\hat{t}_{y\pi}] = \sum_{h=1}^{H} N_h^2 \frac{(1-f_h)}{n_h} S^2_{y,h} \\ s.c. \sum_{h=1}^{H} C_h n_h = N_{UL} \\ s.c.\ n_h \leqslant N_h \\ s.c.\ \underset{d \in D}{Max}\ CV_d \leqslant CV_{loc} \end{cases}$$

avec D l'ensemble des domaines de diffusion (ensemble des APE ou ensemble des croisements groupe x teff) et CV_{loc} la précision locale maximale attendue.

Cependant, il n'est pas possible de combiner deux domaines de diffusion (APE et groupe x teff) en même temps dans l'algorithme d'optimisation. Ainsi, nous avons calculé les allocations optimisées sur chacun des deux domaines de diffusion, puis nous les avons comparées.

Afin de connaître quelle est la meilleure précision locale que nous pouvons atteindre sans détériorer la précision globale, nous avons tracé ce que Koubi et Mathern (2009) ont appelé la frontière d'efficacité (voir encadré ci-dessous).

La frontière d'efficacité permet d'évaluer la perte en précision globale lorsque l'on tient à fixer une précision minimale dans des domaines de diffusion. Nous rappelons que la précision locale est

mesurée par le plus mauvais coefficient de variation des estimateurs des totaux de CA dans les domaines de diffusion.

Frontière d'efficacité :

On appelle frontière d'efficacité, l'ensemble des allocations $n_1, \ldots, n_H$ à taille d'échantillon n fixée telles qu'on ne peut augmenter la précision locale dans les domaines de diffusion, sans détériorer la précision globale d'un estimateur (et inversement). Cette frontière d'efficacité est représentée dans le repère (CV global, CV local maximal).

En d'autres termes, la frontière d'efficacité représente, pour une contrainte de précision locale donnée, la précision globale maximale qui peut être obtenue avec les allocations vérifiant la contrainte de précision locale.

La Figure 1 ci-dessous représente la frontière d'efficacité en considérant l'APE comme domaine de diffusion. Comme nous pouvions nous y attendre, l'allocation de Neyman sans contrainte de précision locale (représentée ici par une croix) est un optimum plat. La précision globale ne se détériore que si l'on essaie d'imposer des contraintes de précision locales fortes ; d'où le choix d'imposer des contraintes de précision locale juste avant le coude d'augmentation des CV globaux. Ainsi, on peut voir que la meilleure précision locale atteignable sans trop détériorer la précision globale correspond à un CV dans chaque APE de 5 % pour l'ESA et de 2 % pour l'EAP.

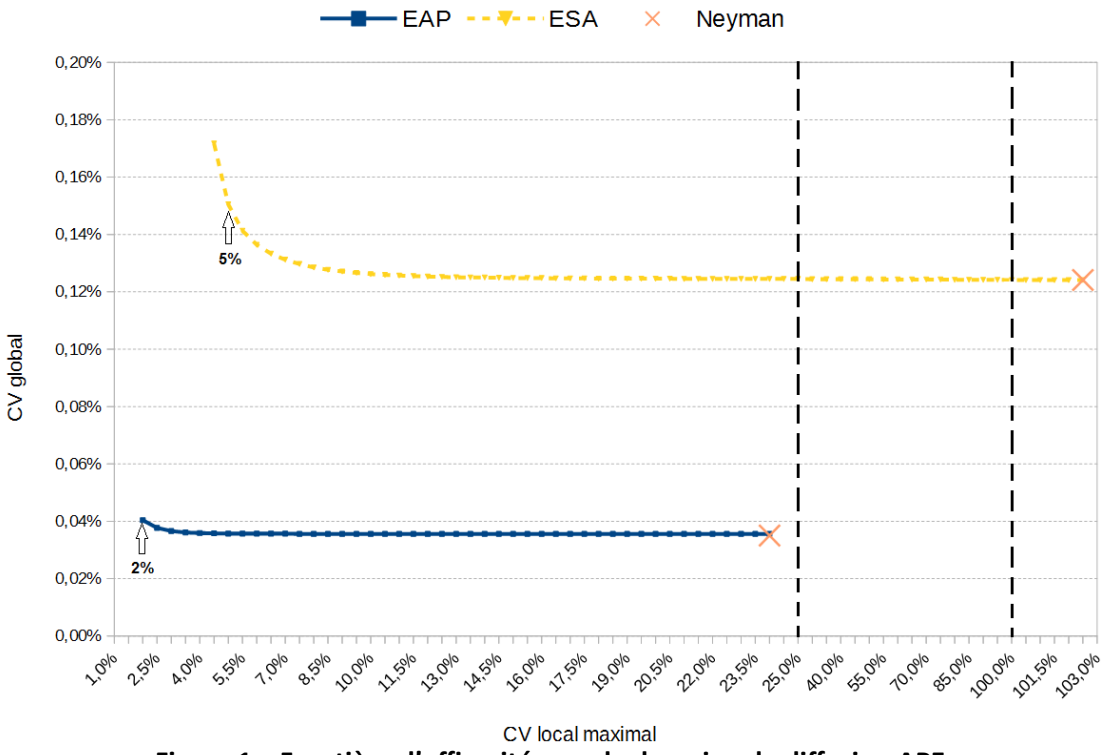

Figure 1 : Frontière d'efficacité pour le domaine de diffusion APE

La Figure 2 ci-dessous représente la frontière d'efficacité en considérant le croisement groupe x tranche d'effectifs comme domaine de diffusion. On peut voir que la meilleure précision locale atteignable sans trop détériorer la précision globale correspond à un CV dans chaque croisement groupe x teff de 8 % pour l'ESA et de 11 % pour l'EAP.

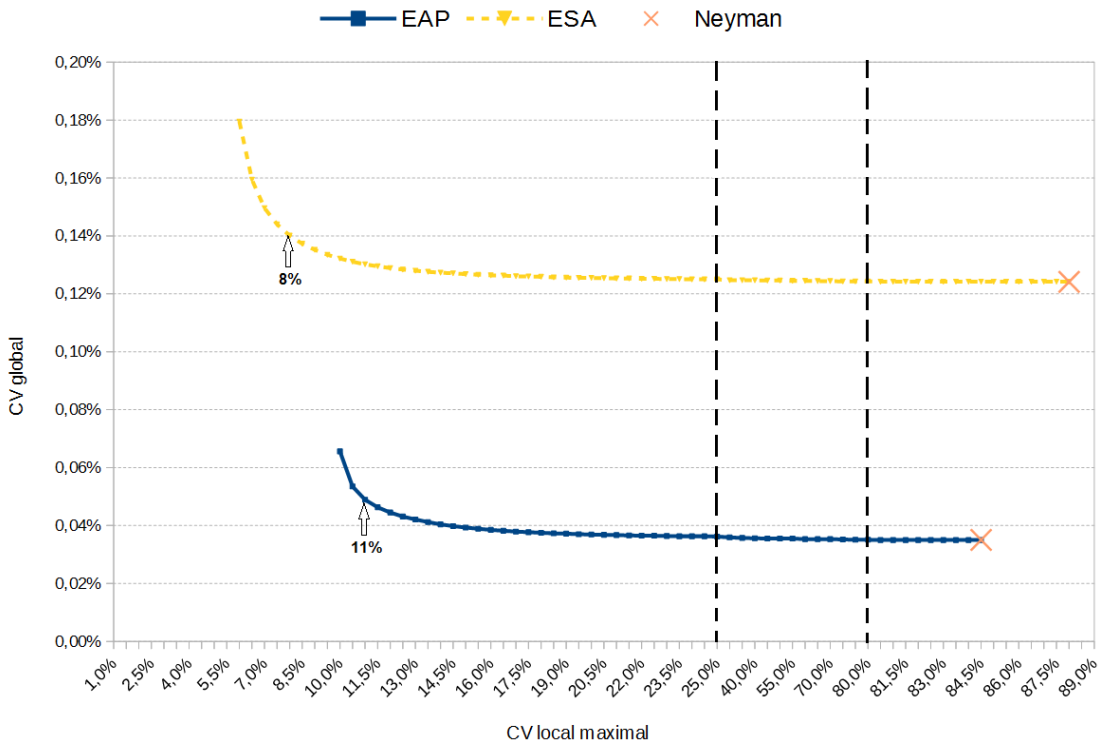

Figure 2 : Frontière d'efficacité pour le domaine de diffusion groupe x teff

3. Résultats

3.1. Nombre d'entreprises à enquêter

Sur le premier domaine de diffusion (APE), l'optimisation sous contraintes de précision locales au niveau EP et sous contrainte d'un nombre fixe d'UL à enquêter conduit à sélectionner 109 900 EP ($n_{et,1}$), 27 000 dans le champ de l'EAP et 82 900 dans le champ de l'ESA. Pour rappel, les paramètres de cette optimisation, en termes de taille d'échantillon, portent sur le nombre d'UL à enquêter pour chaque enquête ESA et EAP, alors que le résultat de l'optimisation donne un nombre d'EP à sélectionner. Concernant le deuxième domaine de diffusion (groupe x teff), l'optimisation conduit à sélectionner 109 500 EP ($n_{et,2}$), 27 000 dans le champ de l'EAP et 82 500 dans le champ de l'ESA.

Afin de garantir une bonne précision sur les deux domaines de diffusion simultanément, nous avons calculé la moyenne entre les deux jeux d'allocations (nombres d'entreprises) obtenus suite à ces optimisations effectuées séparément sur les deux domaines de diffusion. Nous discuterons cette approche et ce choix dans la partie 3.3. Cette allocation « mixte » conduit à sélectionner 109 700 EP ($n_{et,mix}$), 27 000 dans le champ de l'EAP et 82 700 dans le champ de l'ESA.

Toutes ces allocations n_{et} calculées au niveau entreprise ($n_{et,1}$, $n_{et,2}$, $n_{et,mix}$) conduisent bien, via les contraintes de coûts ajoutées au programme d'optimisation, à un échantillon d'environ 35 000 UL pour l'EAP et 116 000 UL pour l'ESA.

3.2. Variabilité du nombre d'UL à enquêter

Les contraintes de coûts introduites dans le programme d'optimisation permettent de respecter en espérance le nombre d'UL à enquêter, le coût correspondant à un nombre moyen d'UL par EP dans une strate donnée (voir Section 2.3.1.). Le nombre exact reste donc variable et dépend de l'échantillon sélectionné.

Ainsi, si l'on s'intéresse maintenant à la variabilité de ce volume, calculée à partir de la formule développée dans la Section 2.3.2., on peut voir dans le tableau 1 que les variations du nombre d'UL à enquêter d'un échantillon à l'autre sont très faibles en général et pour chaque enquête ESA et EAP. Les résultats étant approximativement les mêmes pour chaque jeu d'allocations n_{at} décrit ci-dessus, nous ne présentons ici que les résulats pour l'allocation « mixte ». Cette faible variabilité s'explique en grande partie par le fait que l'exhaustif occupe une part importante de l'échantillon, et que toutes les EP composées de plus de 20 UL sont inclues dans ces strates exhaustives.

Table 1 : Nombre d'EP à sélectionner et intervalles de confiance à 95 % associés au nombre d'UL à enquêter

	EAP	ESA	Total
$n_{at,mix}$	27 000	82 700	109 700
$E[\hat{N_{UL}}] = N_{UL}$	35 000	116 000	151 000
$IC_{95\%}(N_{UL})$	[34 970 ; 35 030]	[115 840 ; 116 160]	[150 830 ; 151 170]

Au cours de la collecte, les UL sélectionnées dans un échantillon sont traitées par différentes équipes selon leur secteur d'activité. Nous devons alors également nous assurer que ce volume d'UL à traiter ne varie pas trop d'une année sur l'autre dans chaque grand secteur d'activité (construction, commerce, industrie, transport, services). Le calcul de la variance de $\hat{N_{UL}}$ dans chaque grand secteur à partir de la même formule de la Section 2.3.2. conduit également à une faible variabilité de ce volume d'UL à traiter par chaque équipe de gestionnaires.

Cependant, la réoptimisation du plan de sondage et le passage d'un tirage stratifié d'UL à un sondage en grappes stratifié sur des caractéristiques d'EP ont conduit à des réallocations sur le volume d'UL à traiter entre les différentes équipes. Le nombre d'UL a ainsi augmenté dans les activités du commerce et diminué dans les services. Également, le nombre de d'UL ayant entre 1 et 5 salariés a augmenté par rapport à l'ancien plan de sondage, au détriment de la tranche d'effectifs [30 – 49 salariés] qui a vu son volume d'UL diminuer.

Ces résultats restent valables quelque soit le jeu d'allocations n_{at} retenu.

3.3. Précision des estimations au niveau EP

Comme expliqué dans les Sections 3.1. et 3.2., les trois jeux d'allocations n_{at} donnent approximativement les mêmes réusltats sur le nombre d'UL à enquêter par grands secteurs et sur la variabilité de ce nombre. Ainsi, afin de départager ces trois jeux d'allocations et de choisir le meilleur, nous avons calculé les précisions des estimateurs des totaux de CA au niveau EP associées à $n_{at,1}$ (allocations optimisées sur l'APE), $n_{at,2}$ (allocations optimisées sur groupe x teff) et $n_{at,mix}$ (moyenne entre $n_{at,1}$ et $n_{at,2}$) sur les deux domaines de diffusion :

- l'APE de l'EP ;
- le croisement groupe d'activité x teff de l'EP.

Ces résultats en termes de distribution des CV sur les deux domaines de diffusion selon le jeu d'allocations retenu sont donnés dans le tableau 2.

Comme on pouvait s'y attendre, l'allocation « mixte » permet d'améliorer les précisions sur les deux domaines de diffusion simultanément (colonnes 3 et 6), en comparaison avec les précisions que l'on obtient si le domaine de diffusion utilisé pour l'optimisation de l'allocation de Neyman *ex ante* est différent de celui utilisé pour calculer les précisions *ex post* (colonnes 2 et 4). En revanche, l'allocation « mixte » détériore la précision si les domaines de diffusion utilisés pour calculer les allocations et évaluer les précisions sont identiques (colonnes 1 et 5). En effet, le CV local maximal est dans ce cas de 5 % pour l'APE et de 11 % pour le croisement groupe x teff, comme vu dans la Section 2.3.3. Cependant, cette détérioration de la précision ne concerne que les domaines de diffusion dont les CV locaux sont parmi les 10 % les plus élevés.

De plus, les différences en termes de précision entre les trois jeux d'allocations ne concernent que le dernier quartile de la distribution des CV locaux. En effet, pour chaque jeu d'allocations n_{et} , la valeur du troisième quartile est proche de 5 % pour l'APE et de 8-9 % pour le croisement groupe x teff.

Table 2 : Distributions des CV locaux pour l'estimation du total de CA au niveau EP

	Domaines de diffusion					
	APE (Domaine 1)			**Groupe x teff (Domaine 2)**		
Niveaux	$n_{et,1}$	$n_{et,2}$	$n_{et,mix}$	$n_{et,1}$	$n_{et,2}$	$n_{et,mix}$
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 %	0.0 %
0 % Min	0.0 %	0.0 %	0.0 %	0.0 %%	0.0 %	0.0 %

Note : Pour le second domaine de diffusion, la distribution des CV locaux est calculée sans prendre en considération la strate exhaustive des unités ayant plus de 200 salariés.

3.4. Précision des estimations au niveau UL

Certains utilisateurs des données des Enquêtes Sectorielles Annuelles (les comptes nationaux par exemple) utilisent et diffusent toujours de l'information au niveau UL. Dans ce cas, la structure de l'EP n'est pas prise en compte et le poids de l'UL correspond à celui de l'EP à laquelle elle appartient, ce qui se traduit par une hausse de la dispersion des poids dans les domaines de diffusion au niveau UL. En effet, dans les strates de tirage de l'ancien plan de sondage en UL, la dispersion des poids en UL était nulle et ne l'est désormais plus systématiquement dans le plan de sondage en EP, puisque les probabilités de sélection des UL ne dépendent plus de leur secteur ou de leur effectif, mais des caractéristiques de leur EP. A priori, cette dispersion accrue des poids devrait se traduire dans la plupart des cas par une baisse de précision. Il n'est cependant pas certain que cette dispersion des poids dans les strates se traduise forcément par une dispersion beaucoup plus forte dans des domaines de diffusion plus agrégés.

Dans le tableau 3 nous comparons donc les précisions des estimateurs de totaux de CA au niveau UL :

- avec le nouveau plan de sondage en EP et l'allocation « mixte » $n_{UL,mix}$;
- avec l'ancien plan de sondage en UL et l'allocation ayant servi au tirage de l'ESA et et de l'EAP 2015 $n_{UL,2015}$.

Si l'on note y_k le CA de l'UL k , $\hat{t}_{y\pi}$ l'estimateur d'Horvitz Thompson du total de CA au niveau UL, n_h le nombre d'entreprises à interroger, N_h le nombre d'entreprises dans la strate h et $f_h = n_h/N_h$ le taux de sondage dans la strate h , $Y_g = \sum_{k \in g} y_k$ la somme des CA des UL de l'entreprise g et $\overline{Y_h} = \frac{1}{N_h} \sum_{g \in U_h} Y_g$ la moyenne empirique de Y_g dans la strate h , la variance de $\hat{t}_{y\pi}$ est donnée par :

$$V[\hat{t}_{y\pi}] = \sum_{h=1}^{H} N_h^2 \frac{(1-f_h)}{n_h} S_{Y_g,h}^2 \quad \text{avec} \quad S_{Y_g,h}^2 = \frac{1}{N_h - 1} \sum_{g \in U_h} (Y_g - \overline{Y_h})^2$$

Table 3 : Distributions des CV locaux pour l'estimation du total de CA au niveau UL

	Domaines de diffusion			
	APE (Domaine 1)		**Groupe x teff (Domaine 2)**	
Niveaux	$n_{UL,mix}$	$n_{UL,2015}$	$n_{UL,mix}$	$n_{UL,2015}$
100 % Max	14.9 %	47.4 %	38.3 %	48.5 %
90 %	5.9 %	7.5 %	10.6 %	12.8 %
75 % Q3	3.9 %	3.8 %	7.3 %	5.4 %
50 % Médiane	2.1 %	1.8 %	3.3 %	1.3 %
25 % Q1	0.9 %	0.6 %	0.6 %	0.0 %
10 %	0.2 %	0.1 %	0.0 %	0.0 %
0 % Min	0.0 %	0.0 %	0.0 %	0.0 %

Note : Pour le second domaine de diffusion, la distribution des CV locaux est calculée sans prendre en considération la strate exhaustive des unités ayant plus de 200 salariés.

Les distributions des CV locaux pour l'estimation du total de CA au niveau UL sont proches entre l'allocation « mixte » du nouveau plan de sondage en EP (colonnes 1 et 3) et l'allocation de l'ancien plan de sondage ayant servi au tirage en 2015 (colonnes 2 et 4), et ce pour chacun des deux domaines de diffusion.

Le changement du plan de sondage a donc un effet ambigu sur les précisions des estimateurs sur les UL. En effet, les estimateurs du CA total par APE et par croisement groupe x teff au niveau UL sont plus précis avec le nouveau plan de sondage dans la moitié des cas et moins précis dans l'autre moitié (voir Figure 3). Ces résultats s'expliquent notamment par la variation du nombre d'UL interrogées par secteur ; la précision est par exemple meilleure dans les APE du commerce, où le nombre d'UL échantillonnées augmente avec le nouveau plan de sondage. Celui-ci permet par ailleurs de traiter les APE dans lesquelles la variance dans l'ancien plan de sondage était particulièrement élevée. L'optimisation du plan de sondage en EP permet ainsi de diviser par trois le

CV maximal de l'estimateur du CA total parmi l'ensemble des APE, et d'éviter des CV supérieurs à 40 % comme c'était le cas auparavant. Concernant le domaine de diffusion groupe x teff, la précision est notamment déteriorée dans la tranche d'effectifs [50 – 199 salariés] du fait qu'auparavant, toutes les UL de plus de 100 salariés étaient exhaustives ce qui n'est plus le cas avec le nouveau plan de sondage. Cela a d'ailleurs conduit à revoir les critères d'exhaustivité de l'enquête.

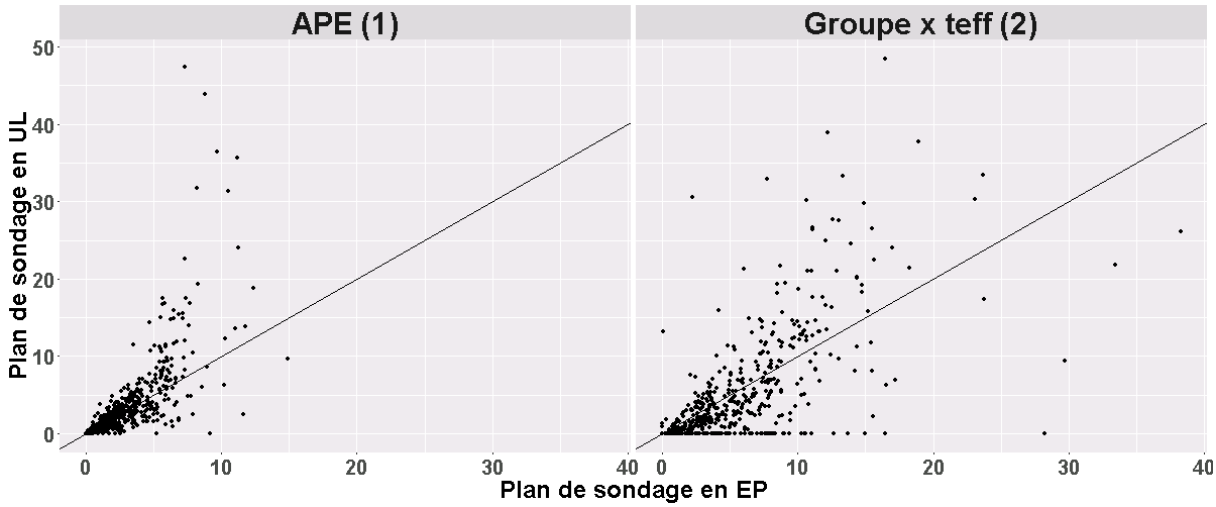

Figure 3 : CV locaux pour l'estimation du total de CA au niveau niveau UL par APE (gauche) et par croisement groupe x teff (droite) selon le plan de sondage

4. Conclusion et perspectives

Dans cette étude, nous avons évalué l'impact sur l'inférence statistique d'un plan de sondage où l'unité de collecte diffère de l'unité statistique. L'objectif était en effet de réoptimiser le plan de sondage des Enquêtes Sectorielles Annuelles afin d'avoir une bonne précision des estimations pour la diffusion en EP sous la contrainte d'un nombre fixe d'UL à enquêter.

La redéfinition de l'exhaustif de l'enquête nous a conduit à considérer environ le même volume d'UL exhaustives qu'en 2015. Cependant, ces critères ont depuis été revus et simplifiés afin de mieux prendre en compte la nouvelle structure de la population d'entreprises (par exemple en forçant dans l'exhaustif les EP et les UL de plus de 100 salariés). Également, un troisième domaine de diffusion (le groupe d'activité) a été intégré dans le calcul d'allocation en sus des deux précédents, et les pondérations allouées à chaque jeu d'allocations ont été modifiées. La variabilité sur le nombre d'UL à enquêter reste très faible, quelque soit l'allocation retenue au niveau EP. L'allocation « mixte » calculée à partir des jeux d'allocations optimisés de façon séparée sur chaque domaine de diffusion permet d'avoir un bon compromis sur la précision au niveau EP sur l'ensemble des domaines à la fois. Au niveau UL, les variances des estimateurs de CA total sur les domaines de diffusion est proche entre le nouveau plan de sondage en EP et l'ancien plan de sondage en UL.

Afin d'améliorer la stratification du plan de sondage (voir Section 2.2.), il aurait pu être envisagé de définir un découpage optimal du nombre de salariés par EP à l'aide de la méthode de Dalenius-Hodges (Dalenius et Hodges Jr, 1959), de la méthode géométique proposée par Gunning et Horgan (2004) ou encore à partir de la méthode Lavallée-Hidiroglou (1988). Cette dernière méthode pourrait également permettre de trouver un seuil optimal de CA par activité au-dessus duquel toutes les toutes les EP seraient exhaustives.

Lors du calcul de l'allocation « mixte » (voir Section 3.1.), en remplacement des pondérations égales allouées à chaque jeu d'allocations, il aurait été possible de calculer des facteurs optimaux afin de mieux répartir l'échantillon selon un certain critère (voir par exemple Merly-Alpa et Rebecq, 2016). L'utilisation d'autres algorithmes de répartition de l'échantillon autorisant l'optimisation sur plusieurs domaines de diffusion à la fois aurait également pu être envisagé, voir par exemple les algorithmes de répartion multivariée de Bethel (1989) ou de Falorsi et Righi (2015).

Entre le tirage de l'échantillon et la diffusion des résultats (intervenant un an plus tard), le périmètre d'une EP (i.e. les UL qui appartiennent à l'EP) peut changer via la mise à jour des données administratives délimitant les contours des groupes ou à partir des remontées des entretiens effectués par les « profileurs » dans les plus grands groupes. Par exemple, une UL d'une EP A peut appartenir à une EP B un an plus tard, ou peut devenir une UL indépendante. Ce problème peut ainsi être vu comme un cas particulier de sondage indirect, où l'échantillon est sélectionné dans une population d'EP (avec un certain périmètre) qui diffère de la population d'intérêt (i.e. au moment de la production des estimations), mais qui est liée à celle-ci via ses UL. Dans ce cas, il est possible d'utiliser la méthode généralisée du partage des poids proposée par Deville et Lavallée (2006) afin de résoudre ce problème (voir Fizzala, 2018b).

Enfin, ce papier porte essentiellement sur la redéfinition du plan de sondage des Enquêtes Sectorielles Annuelles. Cependant, de nombreux traitements post-collecte interviennent suite à la sélection de l'échantillon tels que la correction de la non-réponse, le calage sur marges et la « winsorisation » des valeurs influentes (Fizzala, 2018a). Toutes ces méthodes sont très connues et utilisées, mais leur application à cette enquête tout en tenant compte de la notion économique d'entrprise a nécessité un traitement particulier.

Bibliographie

[1] Bethel J., « Sample allocation in multivariate surveys », *Survey Methodology*, vol 15, n° 1, pp 47-57, 1989.

[2] Chanteloup G., « Consolider les réponses des unités légales pour une statistique d'entreprise plus cohérente », *Acte des Journées de Méthodologie Statistique de l'Insee*, Paris, 2018.

[3] Council Regulation (EEC) 696 / 93, « The statistical units for the observation and analysis of the production system in the Community », 1993.

[4] Dalenius T., Hodges Jr J.L., « Minimum variance stratification », *Journal of the American Statistical Association*, vol 54, n° 285, pp 88-101, 1959.

[5] Deville J.C., Lavallée P., « Indirect sampling : The foundations of the generalized weight share method », *Survey Methodology*, vol 32, n° 2, pp 165-176, 2006.

[6] Falorsi P.D., Righi P., « Generalized framework for defining the optimal inclusion probabilities of one-stage sampling designs for multivariate and multi-domain surveys », *Survey Methodology*, vol 41, n° 1, pp 215-236, 2015.

[7] Fizzala A., « Comment redresser un échantillon d'unités légales tirées via leurs entreprises ? », *Acte des Journées de Méthodologie Statistique de l'Insee*, Paris, 2018.

[8] Fizzala A., « La gestion par partage des poids des changements de contour des entreprises dans l'enquête sectorielle annuelle », *Acte des Journées de Méthodologie Statistique de l'Insee*, Paris, 2018.

[9] Gunning P., Horgan J.M., « A new algorithm for the construction of stratum boundaries in skewed populations », *Survey Methodology*, vol 30, n° 2, pp 159-166, 2004.

[10] Gros E., Le Gleut R., « Sample coordination and response burden for business surveys: methodology and practice of the procedure implemented at insee », *Cambridge Scholars Publishing*, accepté, 2018.

[11] Kokic P.N., Bell P.A., « Optimal Winsorizing cutoffs for a stratified finite population estimator », *Journal of Official Statistics*, vol 10, n° 4, pp 419-435, 1994.

[12] Koubi M., Mathern S., « Résolution d'une des limites de l'allocation de Neyman », *Acte des Journées de Méthodologie Statistique de l'Insee*, Paris, 2009.
[13] Lavallée P., Hidiroglou M.A., « On the stratification of skewed populations », *Survey Methodology*, vol 14, n° 1, pp 33-43, 1988.
[14] Merly-Alpa T., Rebecq A., « Optimisation d'une allocation mixte », *9ème colloque francophone sur les Sondages*, Gatineau, 2016.
[15] Neyman J., « On the two different aspects of the representative method : the method of stratifed sampling and the method of purposive selection », *Journal of the Royal Statistical Society*, vol 97, n° 4, pp 558-625, 1934.

Annexe : Démontration de la formule de variance pour l'estimateur du nombre d'UL à enquêter

L'estimateur d'Horvitz Thompson de N_{UL} s'écrit :

$$\hat{N_{UL}} = \sum_{h=1}^{H} \sum_{k \in S_h} N_{UL,k} = \sum_{h=1}^{H} \sum_{k \in S_h} \frac{z_k}{\pi_k} \quad \text{avec} \quad z_k = \pi_k N_{LU,k} \quad \text{et} \quad \pi_k = \frac{n_h}{N_h}$$

L'espérance de cet estimateur peut s'écrire :

$$E[\hat{N_{UL}}] = \sum_{h=1}^{H} \sum_{k \in U_h} N_{UL,k} E[I_k] \quad \text{avec} \quad I_k = 1 \quad \text{si} \quad k \in S_h$$
$$\Leftrightarrow E[\hat{N_{UL}}] = \sum_{h=1}^{H} \sum_{k \in U_h} N_{UL,k} P(k \in S_h) = \sum_{h=1}^{H} \sum_{k \in U_h} N_{UL,k} \pi_k$$
$$\Leftrightarrow E[\hat{N_{UL}}] = \sum_{h=1}^{H} \sum_{k \in U_h} N_{UL,k} \frac{n_h}{N_h} = \sum_{h=1}^{H} n_h Moy_{UL,h} = N_{UL}$$

avec $Moy_{UL,h} = \frac{1}{N_h} \sum_{k \in U_h} N_{UL,k}$ le nombre moyen d'UL par entreprise dans la strate h .

On se place dans le cas d'un sondage aléatoire simple stratifié. La variance de cet estimateur peut s'écrire :

$$V[\hat{N_{UL}}] = \sum_{h=1}^{H} N_h^2 \frac{(1 - f_h)}{n_h} S_{z,h}^2$$

avec $S_{z,h}^2 = \frac{1}{N_h - 1} \sum_{k \in U_h} (z_k - \overline{z}_h)^2$ la variance empirique de z_k dans la strate h .

Sachant que $z_k = \pi_k N_{LU,k}$ et $\pi_k = n_h / N_h$, on peut réécrire :

$$\overline{z}_h = \frac{1}{N_h} \sum_{k \in U_h} z_k = \frac{n_h}{N_h} \frac{1}{N_h} \sum_{k \in U_h} N_{UL,k} = \frac{n_h}{N_h} Moy_{UL,h}$$

Ainsi, la variance empirique de z_k dans la strate h peut se réécrire :

$$S^2_{z,h} = \frac{1}{N_h - 1}\left(\frac{n_h}{N_h}\right)^2 \sum_{k \in U_h} (N_{UL,k} - Moy_{UL,h})^2 = \left(\frac{n_h}{N_h}\right)^2 S^2_{N_{UL},h}$$

Finalement, la variance de $\hat{N}_{UL}$ peut s'écrire :

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