LE PLAN DE SONDAGE DE L'ESA
(ENQUÊTE SECTORIELLE ANNUELLE DU FUTUR DISPOSITIF DE STATISTIQUES STRUCTURELLES D'ENTREPRISES)

P. Bauer, G. Brilhault, E. Gros (*)

(*)Insee, Direction des statistiques d'entreprises

Introduction

La refonte des statistiques structurelles d'entreprises à l'Insee a conduit dès la fin 2008, à la mise en place d'une nouvelle enquête annuelle sectorielle auprès des entreprises, l'ESA, qui prend le relais de l'EAE (Enquête Annuelle d'Entreprise). Dans le nouveau système Ésane (Élaboration des Statistiques Annuelles d'Entreprises), cette enquête sera utilisée conjointement avec des sources administratives pour produire les statistiques structurelles.

L'Unité de Méthodologie Statistique - Entreprises a été chargée de définir le plan de sondage de cette enquête, en prenant en compte les nouveaux « paramètres » du dispositif :

- un calage sur les données fiscales des données récoltées par l'ESA ;
- un nombre de questionnaires envoyés aux petites entreprises devant diminuer, dans une optique de réduction de la charge statistique ;
- une utilisation des estimateurs « composites » pour certains types de statistiques, en particulier les statistiques sectorielles (voir la communication *«L'utilisation combinée de données d'enquête et de données administratives pour la production des statistiques structurelles d'entreprises »* de P. Brion), afin d'avoir une bonne précision au niveau fin de la nomenclature d'activités et sur les tranches d'effectif.

L'étude présentée ici avait comme objectif de mesurer le gain de précision des estimations apporté par le calage et les estimateurs « composites ». Ce gain s'est révélé suffisamment significatif pour permettre une diminution de moitié de la taille de la partie échantillonnée par rapport à l'EAE.

1. Les estimations dans le cadre d'Ésane

Remarque liminaire : les notations habituelles seront utilisées, sachant que :
U est la population des entreprises du champ de l'ESA, de taille N ;
s est l'échantillon de l'ESA, de taille n ;
i est une entreprise de U ;
Y une variable à estimer.

1.1. Les données dans Ésane

Grâce au module RÉDI (RÉconciliation des Données Individuelles) du processus Ésane, les données administratives, exhaustives sur le champ de l'ESA sont réconciliées avec les résultats de cette enquête. C'est à dire qu'une valeur est déterminée de manière définitive suite à la confrontation des sources disponibles. Ce nouveau module apporte des outils pour améliorer significativement les estimations finales et permet d'assurer la cohérence d'ensemble entre l'ESA et les sources administratives.
Avant Resane, il y avait deux valeurs distinctes pour une variable : celle de la source fiscale, et celle de l'EAE ; maintenant, il y en a trois valeurs :

- IEG : (Information Économique Générale, fichier exhaustif) la valeur issue des sources administratives.
- ESA : la valeur issue des données de l'enquête
- RÉDI : la valeur issue de la confrontation des données IEG, ESA, ou tierce information.

1.2. Le calage dans Ésane

Quand l'enquête est finie, et que la source administrative IEG est disponible, sont réalisés un traitement de la non-réponse totale par repondération et un calage sur les données administratives.
Le chiffre d'affaires étant une donnée importante de l'enquête, et lié à de nombreuses variables, il est acquis que ses totaux dans l'IEG serviront à caler les résultats de l'ESA au niveau groupe de la NAF. D'autres variables seront éventuellement utilisées pour le calage final, mais cela fera l'objet d'études ultérieures sur les données de l'ESA.
Le calage réalisé ici, sur la partie échantillonnée, consiste alors à modifier les poids initiaux des unités de l'échantillon de manière à ce que le chiffre d'affaires de l'échantillon extrapolé d'un secteur X (tel que défini dans le répertoire[1]) permette de retrouver le chiffre d'affaires total du secteur. C'est-à-dire que ces poids doivent respecter la condition suivante :

$$\sum_s wi\, CA^{fiscal}(i)\ 1_{APErep=X}(i) = \sum_U CA^{fiscal}(i) 1_{APErep=X}(i)$$

On obtient alors les poids finaux wi sur lesquels seront basés dans la suite de cette note tous les estimateurs.

1.3. Les estimateurs dans Ésane

Les estimations finales sont établies grâce à différents estimateurs « composites », combinant les données d'enquêtes et les données administratives, et proposés par P. Brion [1].

Plusieurs types d'estimateurs sont à utiliser, selon que la variable est disponible dans l'IEG ou non, ou qu'elle est de type passage secteur branche :

- Pour une statistique s'appuyant sur une variable de l'ESA (absence de données IEG), on utilisera l'estimateur :

$$\sum_s wi Y_i$$

- Pour une statistique sectorielle s'appuyant sur une variable administrative, on utilisera :

$$\sum_U Y^{fiscal}(i) 1_{APErep=X}(i) + \sum_s wi\,(Y^{vrai}(i) 1_{active}(i) 1_{APEenq=X}(i) - Y^{fiscal}(i) 1_{APErep=X}(i))$$

- Pour une statistique correspondant à la ventilation d'une estimation sectorielle, on appliquera une clef de ventilation issue de l'enquête à l'estimateur du total. Par exemple, pour la ventilation du chiffre d'affaires en branches, on utilisera la clef de ventilation ci-dessous, appliquée à l'estimateur par différence du chiffre d'affaires sectoriel :

$$\frac{\sum_s wi\, CA(b,i) 1_{APEenq=X}(i)}{\sum_s wi\, CA(i) 1_{APEenq=X}(i)}$$

$APErep$ étant l'APE dans le répertoire (au moment du lancement de l'enquête ESA)
$APEenq$ étant l'APE résultant de l'enquête ESA (calculé à partir de la ventilation du chiffre d'affaires déclaré).
$Yvrai(i)$ étant la valeur de Y arbitrée REDI.
$Yfiscal(i)$ étant la valeur de Y dans IEG.

[1] Répertoire des Entreprises dans lequel est tiré l'échantillon de l'ESA : le futur répertoire statistique

2. Les estimations dans le cadre de l'étude

2.1. Les fichiers utilisés

Afin de mener cette étude, nous avons effectué des simulations à l'aide :

- des résultats de l'EAE passée : l'EAE 2007/2006 en général, sauf pour le Commerce de Gros et les Services aux Entreprises. Pour ces deux secteurs non interrogés lors de cette EAE à cause de la « stratégie petites entreprises » (qui consiste pour la partie non exhaustive de l'échantillon à utiliser une année sur deux les données fiscales à la place des données d'enquête), nous avons utilisé les données de l'EAE 2006/2005,
- et des fichiers administratifs correspondant à l'année 2006.

Par ailleurs, les secteurs de l'Industrie Agro-Alimentaire et de la partie DOM ont été exclus de l'étude puisque pour le premier nous n'avions pas de résultats sur les unités de moins de vingt salariés, et pour le second un plan de sondage spécifique a été réalisé. Néanmoins, les résultats de cette étude leur ont servi indirectement.
Puisque nous travaillions sur les enquêtes passées, nous étions contraints de travailler sur des strates définies en ancienne nomenclature (NAF Rév.1), tandis que le plan de sondage a dû être établi au final en nouvelle nomenclature (NAF Rév.2). Toutefois, au vu de la proximité de ces deux nomenclatures, les résultats en terme de gain de précision associés au calage demeuraient valables en nouvelle nomenclature.

Il faut noter enfin que les simulations ont été réalisées avec la stratification et les seuils d'exhaustivité en vigueur lors des EAE. Ceux-ci pourront être optimisés a posteriori (cf. fin du papier).

2.2. Le choix des variables

De nombreuses variables étaient disponibles pour réaliser l'étude, mais nous nous sommes concentrés sur quatre statistiques-cibles qui nous semblaient pertinentes et très attendues dans les résultats de cette enquête : le chiffre d'affaires sectoriel, l'investissement, le chiffre d'affaires de la branche principale et le chiffre d'affaires d'une branche réalisé par les entreprises d'un secteur.

2.3. Les estimateurs considérés

Dans cette étude nous avons utilisé pour l'investissement et le chiffre d'affaires sectoriel le deuxième estimateur (par différence), et pour le chiffre d'affaires de la branche principale ainsi que le chiffre d'affaires de la branche réalisé par les entreprises d'un secteur le troisième estimateur (par différence avec une clef de ventilation).
Mais n'étant pas encore dans le système Ésane (il n'existe pas de réconciliation des données à l'heure actuelle), nous avons considéré pour tout i : $CA^{fiscal}(i) = CA^{vrai}(i)$.

Nous avons comparé alors la précision entre ces estimateurs « composites » sur l'échantillon total (partie échantillonnée + partie exhaustive) avec :

- D'une part la précision de l'estimateur de l'EAE (estimateur d'Horvitz-Thomson sans calage) ;
- D'autre part la précision de l'estimateur pondéré classique (Horvitz-Thomson) post-calage ;

ceci pour chacune des quatre variables étudiées, et à différents niveaux de nomenclature.

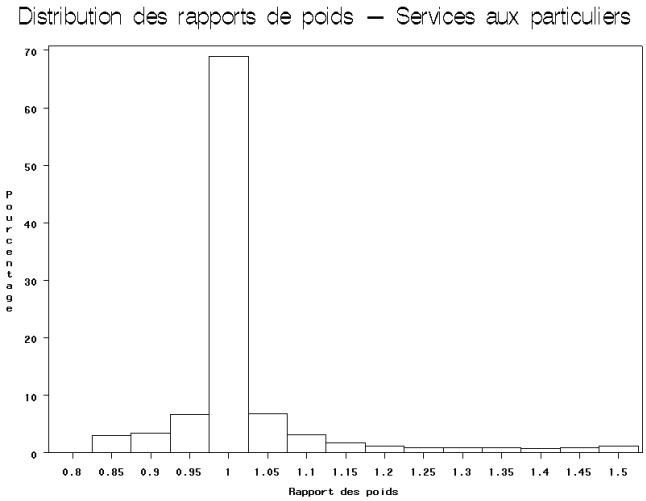
2.4. Le niveau de calage

Les essais de calage au niveau NAF (quatre caractères) n'étaient pas concluants car les poids après calage étaient trop éloignés des poids avant calage, ou parfois même l'algorithme ne convergeait pas. Au niveau trois caractères de la NAF (le groupe) en revanche, les rapports de poids (poids avant calage/poids après calage) obtenus étaient plus concentrés autour de 1.
Le tableau ci-dessous donne des exemples de bornes de rapports de poids qu'il est possible d'atteindre sans déformation trop importante de la distribution des poids.

Tableau 1 : bornes possibles des rapports de poids issus du calage au niveau groupe du CA(i), par grand secteur

Secteur	Borne inférieure	Borne supérieure
Construction	0,85	1,50
Commerce de gros	0,75	2,50
Commerce de détail	0,65	6,00
Services aux entreprises	0,40	5,00
Services aux particuliers	0,85	1,50
Transport	0,85	2,00

Graphique 1 : distribution des rapports de poids, pour le secteur des Services aux Particuliers



Remarque : Dans la plupart des cas, même si l'on ne cale qu'au niveau trois caractères de la nomenclature, les estimations sectorielles seront améliorées aussi au niveau plus fin (quatre caractères), grâce à l'utilisation conjointe du code APE du répertoire et du code calculé avec les résultats de l'enquête.

3. Les résultats obtenus en utilisant la taille d'échantillon de l'EAE

3.1. Au niveau des grands secteurs de l'EAE

Le tableau suivant présente les résultats des différentes estimations pour l'ensemble des grands secteurs de l'ESA (la définition de ces grands secteurs est en annexe 1). Le coefficient de variation empirique de chaque estimateur a été calculé. Ainsi, CV_{EAE} désigne le coefficient de variation empirique de l'estimateur actuel de l'EAE, tandis que $CV_{HT\ calé}$ et $CV_{composite}$ désignent respectivement ceux de l'estimateur pondéré classique post-calage et de l'estimateur composite approprié après calage.
Au niveau grand secteur, le chiffre d'affaires d'une branche réalisé par les entreprises du secteur est confondu avec le chiffre d'affaires de la branche principale, c'est pourquoi n'apparaissent que trois variables.

Tableau 2 : précisions obtenues sur les grands secteurs

Variable	Secteur	CV_{EAE}	$CV_{HT\ calé}$	$CV_{composite}$
Chiffre d'affaires sectoriel	Construction	0,518%	0,002%	0,002%
	Commerce de gros	0,636%	0,008%	0,008%
	Commerce de détail	0,299%	0,009%	0,008%
	Services aux entreprises	0,956%	0,007%	0,007%
	Services aux particuliers	1,120%	0,007%	0,007%
	Transport	0,086%	0,001%	0,001%
Investissement	Construction	1,498%	1,408%	0,010%
	Commerce de gros	1,504%	1,362%	0,009%
	Commerce de détail	1,598%	1,339%	0,021%
	Services aux entreprises	1,680%	1,584%	0,005%
	Services aux particuliers	5,168%	5,122%	0,002%
	Transport	0,281%	0,268%	0,001%
Chiffre d'affaires de la branche principale	Construction	0,322%	0,059%	0,013%
	Commerce de gros	0,644%	0,132%	0,014%
	Commerce de détail	0,306%	0,175%	0,015%
	Services aux entreprises	0,985%	0,363%	0,036%
	Services aux particuliers	1,116%	0,081%	0,032%
	Transport	0,075%	0,016%	0,004%

Il faut noter que la précision obtenue est un peu « optimiste », car la variance n'est due qu'aux erreurs d'échantillonnage, et met de côté la variance due à la non-réponse ou aux erreurs de mesure.

On observe que le calage seul améliore déjà la précision au niveau des grands secteurs, surtout sur la variable qui a servi au calage (chiffre d'affaires sectoriel).
En outre, les estimateurs « composites » sont systématiquement meilleurs, et c'est d'autant plus prononcé pour l'investissement et le chiffre d'affaires de la branche principale. Le chiffre d'affaires sectoriel a gagné sa précision sur le calage, et l'estimateur composite ne lui apporte pas beaucoup de précision supplémentaire. Tandis que pour les variables qui n'ont pas participé au calage, la plus-value est moins forte au calage mais plus forte avec l'estimateur « composite ».

3.2. Au niveau des groupes

Il restait à savoir si ces résultats, obtenus au niveau des grands secteurs, demeuraient valides pour des estimations à des niveaux plus fins. Pour ce faire, nous avons reproduit les simulations détaillées ci-dessus au niveau groupe (trois premiers caractères de la NAF), en nous intéressant aux mêmes variables sectorielles, ainsi qu'à la ventilation en branche du chiffre d'affaires sectoriel.
Les résultats obtenus par exemple pour les groupes du secteur des services aux particuliers sont en annexe 2. Notons que pour le chiffre d'affaires sectoriel, puisque nous avons calé au niveau groupe, l'estimateur composite de confond avec l'estimateur classique post-calage.
Les résultats obtenus étaient là encore très positifs : pour la quasi-totalité des estimations, le calage sur données fiscales et l'utilisation d'estimateurs « composites » conduit à une amélioration significative de la précision.

Nous avons aussi été inspecter les résultats des estimations des chiffres d'affaires des branches non principales. Les résultats étaient alors plus mitigés : la mise en œuvre de la méthode d'estimation dans Ésane n'améliore la précision que du quart des estimations, et induit pour 45% une dégradation de la précision supérieure à deux points de coefficient de variation. Mais il faut relativiser ces résultats : en effet, la dégradation importante touche des croisements secteur-branche pour lesquels l'estimateur pondéré pré-calage de l'EAE s'avère déjà très peu précis. Ceci découle en particulier du fait que pour une majorité de ces croisements, seules quelques entreprises contribuent à la valeur de l'estimation. Dans une telle configuration, les estimations initiales s'avèrent peu précises et les modifications de poids liées au calage peuvent conduire à accentuer la variance des estimateurs.

3.3. Au niveau fin de la NAF

Nous avons mené ce même genre d'études à un niveau encore le plus fin de la NAF (quatre caractères), sur le chiffre d'affaires sectoriel et l'investissement. Même à ce niveau là les résultats se sont avérés assez bons.
Les résultats obtenus par exemple sur les NAF dépendant du secteur des services aux particuliers sont présentés en annexe 3.

4. Conclusions et choix opérés pour le plan de sondage de l'ESA

4.1. Une précision équivalente à celle de l'EAE, avec une partie échantillonnée deux fois plus petite

Le gain de précision grâce à la combinaison du calage et d'estimateurs composites nous permet de réduire la partie échantillonnée. Les tests montrent qu'une réduction de moitié de la taille de la partie échantillonnée de l'EAE sera possible dans l'ESA, à objectif de précision égal sur une majorité de variables au niveau groupe et supérieur.
En effet, grâce à l'hypothèse simplificatrice qu'une réduction de moitié de l'échantillon augmente la variance de l'estimation par deux, nous obtenons les résultats ci-dessous.

Tableau 3 : précisions obtenues dans l'EAE et pour un demi-échantillon dans le processus Ésane, au niveau des grands secteurs.

	Chiffre d'affaires sectoriel		Investissement		CA branche principale	
Secteur	CV_{EAE}	$CV_{demi\text{-}échantillon}$	CV_{EAE}	$CV_{demi\text{-}échantillon}$	CV_{EAE}	$CV_{demi\text{-}échantillon}$
Construction	0,518%	0,003%	1,498%	0,014%	0,322%	0,018%
Commerce de gros	0,636%	0,011%	1,504%	0,012%	0,644%	0,020%
Commerce de détail	0,299%	0,012%	1,598%	0,029%	0,306%	0,021%
Services aux entreprises	0,956%	0,010%	1,680%	0,007%	0,985%	0,051%
Services aux particuliers	1,120%	0,009%	5,168%	0,003%	1,116%	0,046%
Transport	0,086%	0,001%	0,281%	0,001%	0,075%	0,005%

Au niveau grands secteurs, la précision est même meilleure que dans l'EAE, avec un échantillon réduit de moitié.

Tableau 4 : précisions obtenues dans l'EAE et pour un demi-échantillon dans le processus Ésane, au niveau groupe des services aux particuliers.

	Chiffre d'affaires sectoriel		Investissement		CA branche principale		CA branche du secteur	
Secteur	CV_{EAE}	$CV_{demi\text{-}échantillon}$	CV_{EAE}	$CV_{demi\text{-}échantillon}$	CV_{EAE}	$CV_{demi\text{-}échantillon}$	CV_{EAE}	$CV_{demi\text{-}échantillon}$
551	1,562%	0,098%	5,790%	0,053%	1,753%	0,363%	1,738%	0,997%
552	2,022%	0,156%	23,074%	0,103%	2,174%	0,454%	1,710%	0,801%
553	1,304%	0,096%	3,597%	0,048%	1,270%	0,145%	1,143%	0,477%
554	3,787%	0,586%	8,132%	0,086%	4,201%	1,929%	3,183%	2,893%
555	1,255%	0,053%	5,058%	0,296%	0,928%	0,108%	0,935%	0,264%
633	1,354%	0,016%	3,258%	0,017%	1,289%	0,069%	1,257%	0,414%
701	6,853%	0,045%	22,535%	0,026%	6,465%	0,266%	6,258%	0,630%
702	1,595%	0,105%	7,239%	0,025%	1,698%	0,097%	1,707%	0,231%
703	1,943%	0,113%	11,705%	0,059%	1,908%	0,269%	1,879%	0,580%
921	1,146%	0,058%	4,837%	0,125%	1,121%	0,128%	1,113%	0,383%
922	0,373%	0,025%	0,645%	0,007%	0,351%	0,010%	0,351%	0,148%
924	4,720%	0,605%	3,663%	3,624%	4,555%	0,853%	4,516%	1,007%
930	1,595%	0,044%	5,010%	0,009%	1,571%	0,210%	1,560%	0,416%

La remarque précédente est toujours valable au niveau groupe de la NAF.

4.2. Suppression de la stratégie « petites entreprises »

Cette réduction de la charge statistique pesant sur les entreprises permet de supprimer la stratégie « petites entreprises » appliquée dans l'EAE, qui consistait à n'interroger la partie échantillonnée (soit les petites entreprises) des secteurs du Commerce et des Services qu'une année sur deux. Ceci

permettait de limiter la charge sur les petites entreprises, mais présentait des évolutions heurtées dans les séries. La suppression de cette méthode participe donc à l'amélioration des résultats de l'ESA.

4.3. Un rééquilibrage entre les grands secteurs

Cette étude a aussi permis de comparer la précision des différents grands secteurs de l'EAE. Les différences sont flagrantes, et cela est dû au fait que la taille des échantillons des EAE était basée sur les moyens du service enquêteur. Maintenant que l'ESA est réalisée entièrement à l'Insee, il n'y a plus de raison de favoriser sensiblement un secteur plutôt qu'un autre. Nous avons donc proposé une réallocation des questionnaires afin de rééquilibrer la précision respective des grands secteurs :

Tableau 5 : précisions obtenues dans l'EAE et pour un demi-échantillon dans le processus Ésane, après rééquilibrage entre les secteurs.

				Chiffre d'affaires sectoriel		Investissement		CA branche principale	
Secteur	Echantillon EAE initial	Echantillon ESA initial	Echantillon ESA « équilibré »	CV_{EAE}	CV_{ESA} « équilibré »	CV_{EAE}	CV_{ESA} « équilibré »	CV_{EAE}	CV_{ESA} « équilibré »
Construction	19 819	15229	12 934	0,518%	0,004%	1,498%	0,020%	0,322%	0,026%
Commerce de gros	24 160	15667	15 667	0,636%	0,011%	1,504%	0,012%	0,644%	0,020%
Commerce de détail	38 281	23745	18 900	0,299%	0,015%	1,598%	0,036%	0,306%	0,026%
Services aux entreprises	33 523	21976	26 821	0,956%	0,009%	1,680%	0,006%	0,985%	0,043%
Services aux particuliers	34 410	20263	24 342	1,120%	0,008%	5,168%	0,003%	1,116%	0,040%
Transport	13 833	10265	8 481	0,086%	0,002%	0,281%	0,002%	0,075%	0,007%
Total	164 026	107 145	107 145						

Par exemple, les secteurs de la construction et des transports ont chacun été déchargé d'environ 2 000 questionnaires, en faveur des services aux entreprises, mais leurs précisions globales restent les meilleures quasiment partout.

4.4. Mise en place dans l'ESA 2009/2008

Pour le tirage de l'échantillon de l'ESA 2009/2008, nous sommes repartis des poids en vigueur dans l'EAE, et les avons ajustés de façon à obtenir la réallocation ci-dessus. Cela s'est traduit en moyenne par un doublement des poids de l'EAE.
Pour des strates qui avaient un faible nombre d'unités dans leur population, ce nouveau poids amenait à ne tirer aucune entreprise. C'est pourquoi nous avons adapté la pondération de ces cas particuliers.

4.5. Optimisation des limites de strates

L'ensemble de l'étude est basé sur une stratification et des seuils d'exhaustivité hérités de l'EAE, et donc définis il y a bien longtemps.
Maintenant que nous avons fixé une précision cible et une taille d'échantillon maximale, pourquoi ne pas étudier si ces strates et seuils ne pourraient pas être améliorés ? L'idée sous-jacente encore étant de diminuer la taille de l'échantillon.
Il existe un algorithme développé par Lavallée & Hidiroglou [5], amélioré par Rivest, qui propose des limites de strates (par NAF, strates définies par l'effectif) optimisant la précision globale de la variable d'intérêt (selon Neyman). La conclusion du stage réalisé à l'UMS-E par A. Dreyer [6] est que cet algorithme n'atteint qu'un optimum local. L'étude de ce sujet sera prolongé en 2009 par A. Dreyer, par le développement d'un algorithme qui permettra d'atteindre un optimum global, afin de l'appliquer en priorité au plan de sondage de l'ESA.

Les calculs de précision dans l'étude

La précision, ou coefficient de variation, relative à l'estimation du total, est obtenue comme le rapport de la racine carrée de l'estimation de la variance sur l'estimation du total lui-même.
A chaque fois que l'on travaille sur le domaine X, on fait le produit de la variable considérée (par exemple, le chiffre d'affaires) par l'indicatrice d'appartenance au domaine (par exemple, le secteur) :

$$Y = Z.1_{DOMAINEenq=X} \ ,$$

avec
- Z=chiffre d'affaires, investissement,
- DOMAINE =APE, GROUPE ou SECTEUR.

Remarque : la partie exhaustive sert au calcul de l'estimation, mais pas à celui de la variance, car on ne considère ici que la variance d'échantillonnage.

L'estimation du total est simple, tandis que celle de la variance du total nécessite des étapes différentes pour chaque méthode . C'est pourquoi nous détaillons cela dans les lignes qui suivent.

On pose A le total à estimer de la variance d'intérêt Y : $A = \sum_U Y_i$.

Pour la précision dans l'EAE
La variance calculée est la variance empirique d'un sondage aléatoire stratifié. La macro Calker la calcule et nous la fournit par DOMAINE demandé :

$V\left(\hat{A}\right) = \sum_{h=1}^{H} N_h{}^2 \left(1 - \frac{n_h}{N_h}\right) \frac{s_h^2}{n_h}$, avec h=1..H les strates de tirage

Pour la précision issue du calage
Après un calage, pour n grand on a : $V(\hat{A}) \approx V(\sum_s d_i res_i)$, avec d_i les poids avant calage et res_i les résidus issus de la régression linéaire de Y sur les variables de calage : $Y_i = \sum_j b_j . CA^{fiscal}(j).1_{GROUPErep=j}(i) + res_i$, avec j=1..J les groupes du secteur sur lequel on cale.
Nous avons donc obtenu la variance sur les domaines voulus par application de la macro Calker sur les résidus res_i.

Pour la précision des estimateurs « composites »
La démarche est la même que pour le calage simple, sauf qu'au lieu de régresser Yi sur les variables de calage, on régresse :

- pour le CA sectoriel et l'investissement : $T_i = Z(i).(1_{DOMAINEenq=X}(i) - 1_{DOMAINErep=x}(i))$
- pour le CA branche principale et CA branches d'un secteur :

Ces statistiques sont du type passage secteur-branche, donc sont estimées par le produit d'un total et du ratio de deux totaux, que l'on peut écrire ainsi : $\hat{R} = \hat{M}.\frac{\hat{P}}{\hat{Q}}$,
avec M l'estimateur « composite » du chiffre d'affaires sectoriel (cf. page 2), $\hat{P} = \sum_s w_i . CA(b,i).1_{APEenq=X}(i)$ et $\hat{Q} = \sum_s w_i . CA(i).1_{APEenq=X}(i)$.

Or la macro SAS Calker de calcul de précision ne sait pas calculer directement les variances d'un tel estimateur. C'est pourquoi on crée la variable T_i, qui est la linéarisée de R_i (les détails de cette linéarisation sont explicités en annexe 5) :

Les calculs de précision dans l'étude (suite)

$$T_i = lin(C\hat{A}_{ventilé}(b,X))_i = \frac{C\hat{A}^{enq}_{HT,enq}(b,X)}{C\hat{A}^{enq}_{HT,enq}(X)} CA^{fiscal}(i)(1_{APEenq=X}(i) - 1_{APErep=X}(i))$$

$$+ \frac{C\hat{A}_{diff}(X)}{C\hat{A}^{enq}_{HT,enq}(X)} \left[CA^{enq}(b,i) - \frac{C\hat{A}^{enq}_{HT,enq}(b,X)}{C\hat{A}^{enq}_{HT,enq}(X)} CA^{enq}(i) \right] 1_{APEenq=X}(i)$$

et l'on approxime ainsi $V(\hat{R}) \approx V\left(\sum_s w_i T_i \right)$. Les poids w_i sont ceux issus du calage sur la source fiscale ; d'où la deuxième approximation $V(\sum_s w_i T_i) \approx V(\sum_s d_i res_i)$, avec d_i les poids initiaux et res_i les résidus issus de la régression linéaire de T_i sur les variables de calage.

Annexe 1 : Explicitation des secteurs étudiés, en NAF Rév.1

Entreprises françaises actives, exploitantes et marchandes, des secteurs :

Construction
Division 45
Commerce de Gros
Division 51
Commerce de détail
Divisions 50, 52, et classes 151F, 158B, 158C, 158D
Services aux entreprises
Divisions 64, 67, 71, 72, 74, 90
Services aux particuliers
Groupes 551, 552, 553, 554, 555, 633, 701, 702, 703, 921, 922, 924, 930, sauf classes 552F, 554C, 703E.
Transport
Divisions 60, 61, 62, 63 sauf 633Z, 623Z et 631D

Annexe 2 : Précisions obtenues au niveau groupe des trois estimateurs considérés, sur le secteur des services aux particuliers.

Variable	Groupe	CV_{EAE}	$CV_{HT\ calé}$	$CV_{composite}$
Chiffre d'affaires sectoriel	551	1,562%	0,069%	0,069%
	552	2,022%	0,111%	0,111%
	553	1,304%	0,068%	0,068%
	554	3,787%	0,414%	0,414%
	555	1,255%	0,037%	0,037%
	633	1,354%	0,011%	0,011%
	701	6,853%	0,032%	0,032%
	702	1,595%	0,074%	0,074%
	703	1,943%	0,080%	0,080%
	921	1,146%	0,041%	0,041%
	922	0,373%	0,018%	0,018%
	924	4,720%	0,428%	0,428%
	930	1,595%	0,031%	0,031%
Investissement	551	5,790%	4,801%	0,037%
	552	23,074%	23,736%	0,073%
	553	3,597%	3,320%	0,034%
	554	8,132%	8,474%	0,061%
	555	5,058%	4,940%	0,210%
	633	3,258%	3,181%	0,012%
	701	22,535%	21,474%	0,019%
	702	7,239%	7,251%	0,018%
	703	11,705%	9,689%	0,042%
	921	4,837%	4,515%	0,088%
	922	0,645%	0,590%	0,005%
	924	3,663%	3,476%	2,563%
	930	5,010%	4,567%	0,007%
Chiffre d'affaires de la branche principale	551	1,753%	0,737%	0,257%
	552	2,174%	0,579%	0,321%
	553	1,270%	0,359%	0,103%
	554	4,201%	2,604%	1,364%
	555	0,928%	0,107%	0,076%
	633	1,289%	0,306%	0,049%
	701	6,465%	0,490%	0,188%
	702	1,698%	0,121%	0,068%
	703	1,908%	0,437%	0,190%
	921	1,121%	0,267%	0,091%
	922	0,351%	0,039%	0,007%
	924	4,555%	0,714%	0,603%
	930	1,571%	0,304%	0,148%
Chiffre d'affaires de la branche, réalisé par les entreprises du secteur	551	1,738%	0,730%	0,705%
	552	1,710%	0,580%	0,566%
	553	1,143%	0,360%	0,337%
	554	3,183%	2,086%	2,046%
	555	0,935%	0,175%	0,187%
	633	1,257%	0,299%	0,292%
	701	6,258%	0,515%	0,446%
	702	1,707%	0,153%	0,163%
	703	1,879%	0,438%	0,410%
	921	1,113%	0,270%	0,271%
	922	0,351%	0,052%	0,105%
	924	4,516%	0,708%	0,712%
	930	1,560%	0,302%	0,294%

Annexe 3 : Précisions obtenues pour le chiffre d'affaires sectoriel et l'investissement, au niveau NAF, des trois estimateurs considérés, sur le secteur des services aux particuliers.

	Chiffre d'affaires sectoriel			Investissement		
	CV_{EAE}	$CV_{HT\ calé}$	$CV_{demi\text{-}échantillon}$	CV_{EAE}	$CV_{HT\ calé}$	$CV_{demi\text{-}échantillon}$
551A	1,795%	0,598%	0,148%	6,899%	6,214%	0,141%
551C	4,224%	2,549%	0,039%	14,592%	11,340%	0,015%
551E	11,240%	9,180%	2,211%	31,457%	29,105%	5,015%
552A	55,911%	3,537%	0,064%	0,229%	0,212%	0,000%
552C	3,508%	1,432%	0,000%	18,877%	17,370%	0,000%
552E	1,768%	0,977%	0,094%	49,604%	45,900%	0,001%
553A	1,670%	0,314%	0,125%	5,333%	4,666%	0,048%
553B	1,931%	1,296%	0,210%	6,721%	5,956%	0,194%
554A	4,607%	2,230%	0,073%	14,429%	9,554%	0,000%
554B	5,772%	2,348%	0,273%	12,476%	10,595%	0,143%
555A	0,594%	0,396%	0,001%	5,247%	4,834%	0,000%
555C	0,333%	0,227%	0,000%	0,587%	0,526%	0,000%
555D	6,290%	0,718%	0,219%	16,419%	14,896%	0,686%
633Z	1,354%	0,014%	0,014%	4,198%	3,938%	0,015%
701A	3,339%	2,430%	0,113%	71,748%	66,346%	0,074%
701B	15,332%	12,873%	0,029%	40,128%	37,119%	0,000%
701C	3,372%	2,787%	0,306%	20,842%	19,279%	0,051%
701D	23,091%	17,811%	0,018%	33,962%	31,428%	0,001%
701F	18,129%	2,222%	0,079%	20,411%	18,756%	0,080%
702A	0,797%	0,417%	0,034%	2,129%	1,942%	0,024%
702B	15,829%	8,693%	1,175%	14,083%	13,029%	0,049%
702C	5,188%	0,976%	0,268%	25,656%	23,489%	0,063%
703A	2,657%	0,875%	0,163%	11,037%	10,105%	0,316%
703C	3,381%	1,813%	0,326%	4,636%	4,134%	0,324%
703D	5,014%	2,282%	0,010%	48,250%	44,560%	0,002%
921A	3,926%	1,922%	0,278%	18,636%	17,102%	0,994%
921B	4,452%	2,703%	0,164%	6,931%	6,009%	0,351%
921C	4,215%	2,499%	0,563%	22,993%	21,197%	1,036%
921D	3,298%	2,388%	0,262%	4,862%	4,443%	0,179%
921F	0,971%	0,774%	0,168%	14,002%	12,893%	0,912%
921G	0,955%	0,651%	0,012%	7,640%	7,059%	0,050%
921J	1,958%	1,486%	0,019%	10,548%	9,759%	0,005%
922A	1,061%	0,773%	0,000%	2,137%	1,955%	0,000%
922B	3,800%	1,126%	0,051%	4,988%	4,338%	0,053%
922E	1,501%	0,940%	0,215%	0,673%	0,602%	0,000%
922F	0,024%	0,022%	0,000%	0,000%		
924Z	4,720%	0,464%	0,464%	4,666%	4,139%	3,084%
930A	5,417%	4,682%	0,635%	9,449%	8,715%	0,003%
930B	3,750%	3,079%	0,416%	9,637%	8,904%	0,027%
930D	2,707%	1,369%	0,000%	10,673%	9,749%	0,000%
930E	3,117%	2,606%	0,207%	10,081%	9,310%	0,043%
930G	10,162%	9,097%		8,895%	8,225%	0,000%
930H	2,018%	1,346%	0,012%	4,573%	4,129%	0,001%
930K	1,796%	1,261%	0,013%	3,819%	3,488%	0,011%
930L	6,836%	5,507%	0,058%	12,699%	11,476%	0,001%
930N	7,508%	5,389%	0,167%	9,607%	8,868%	0,221%

Annexe 4 : Le fonctionnement de la macro CALKER

Calker est une macro développée sous SAS de CALKul d'ERreur (écrite par T. Mainaud et S. Hallépée) ; elle permet le calcul de précision :

- pour un total, une moyenne ou un ratio ;
- pour des enquêtes assimilables à un sondage stratifié à un seul degré ;
- en tenant compte de la correction de la non-réponde totale utilisée (deux cas possibles : hot-deck ou repondération) (la non-réponse partielle n'est pas prise en compte).

Les paramètres de la macro sont :

- une table de données individuelles contenant pour chaque observation, son identifiant, son poids final, sa strate de tirage, ainsi que trois indicatrices (réponse, champ, domaine) et les variables dont on veut calculer la précision;
- une table de données sur les strates : nom de la strate h, N_h, n_h et r_h ;
- deux paramètres généraux : type d'estimateur considéré et type de correction de la non-réponse totale.

En sortie, pour chaque domaine de diffusion, sont données les estimations du total, ou de la moyenne ou du ratio, de sa variance, du coefficient de variation et de l'intervalle de confiance à 95%.

Annexe 5 : Linéarisation de l'estimateur du chiffre d'affaires total de la branche b du secteur X

L'objectif est ici de déterminer la variance de l'estimateur du chiffre d'affaires total de la branche b pour les entreprises du secteur X proposé.

Cet estimateur est construit de la façon suivante :

- dans un premier temps, le chiffre d'affaires total du secteur X est estimé grâce à l'estimateur en différence suivant :

$$\hat{CA}_{diff}(X) = \underbrace{\sum_{U} CA^{fiscal}(i)\,1_{APErep=X}(i)}_{\text{indépendant de l'échantillon}} + \sum_{s} w_i\, CA^{fiscal}(i)(1_{APEenq=X}(i) - 1_{APErep=X}(i))$$

- le chiffre d'affaires total de la branche b du secteur X est alors obtenu par ventilation de cette estimation à l'aide de la clef de ventilation suivante issue de l'enquête :

$$\frac{\sum_{s} w_i\, CA^{enq}(b,i)1_{APE=X}(i)}{\sum_{s} w_i\, CA^{enq}(i)1_{APE=X}(i)} = \frac{\hat{CA}_{HT}(b,X)}{\hat{CA}_{HT}(X)}$$

Au final, le chiffre d'affaires total de la branche b du secteur X est estimé à l'aide de l'estimateur « ventilé » suivant :

$$\hat{CA}_{ventilé}(b,X) = \hat{CA}_{diff}(X)\frac{\hat{CA}_{HT}(b,X)}{\hat{CA}_{HT}(X)}$$

Afin de calculer la variance de cet estimateur, nous allons procéder par linéarisation. Si l'on développe l'expression de l'estimateur ventilé, on constate que ce dernier est une fonction de quatre estimations de totaux :

$$\hat{CA}_{ventilé}(b,X) = f(\hat{CA}^{fiscal}_{HT,enq}(X), \hat{CA}^{fiscal}_{HT,rep}(X), \hat{CA}^{enq}_{HT,enq}(b,X), \hat{CA}^{enq}_{HT,enq}(X))$$

avec $f(x,y,z,w) = (C^{te} + x - y)\frac{z}{w}$, et où :

- $\hat{CA}^{fiscal}_{HT,enq}(X) = \sum_{s} w_i\, CA^{fiscal}(i)1_{APEenq=X}(i)$ est l'estimateur du chiffre d'affaires total du secteur X utilisant les APE issues de l'enquête et le chiffre d'affaires fiscal ;
- $\hat{CA}^{fiscal}_{HT,rep}(X) = \sum_{s} w_i\, CA^{fiscal}(i)1_{APErep=X}(i)$ est l'estimateur du chiffre d'affaires total du secteur X utilisant les APE issues de répertoire administratif et le chiffre d'affaires fiscal ;
- $\hat{CA}^{enq}_{HT,enq}(b,X) = \sum_{s} w_i\, CA^{enq}(b,i)1_{APEenq=X}(i)$ est l'estimateur du chiffre d'affaires total de la branche b du secteur X utilisant les APE et chiffres d'affaires issus de l'enquête.

- $\hat{CA}^{enq}_{HT,enq}(X) = \sum_s w_i \, CA^{enq}(i) 1_{APEenq=X}(i)$ est l'estimateur d'Horvitz-Thompson du chiffre d'affaires total du secteur X utilisant les APE et chiffres d'affaires issus de l'enquête.

On a :

$$\frac{\partial f}{\partial x}(x,y,z,w) = \frac{z}{w}, \; \frac{\partial f}{\partial y}(x,y,z,w) = -\frac{z}{w},$$
$$\frac{\partial f}{\partial z}(x,y,z,w) = \frac{C^{te}+x-y}{w}, \; \frac{\partial f}{\partial w}(x,y,z,w) = -\frac{(C^{te}+x-y)z}{w^2},$$

et la linéarisée de l'estimateur ventilé est égale à

$$lin(\hat{CA}_{ventilé}(b,X))_i = \frac{\hat{CA}^{enq}_{HT,enq}(b,X)}{\hat{CA}^{enq}_{HT,enq}(X)} CA(i)^{fiscal} 1_{APEenq=X}(i) - \frac{\hat{CA}^{enq}_{HT,enq}(b,X)}{\hat{CA}^{enq}_{HT,enq}(X)} CA(i)^{fiscal} 1_{APErep=X}(i)$$
$$+ \frac{\hat{CA}_{diff}(X)}{\hat{CA}^{enq}_{HT,enq}(X)} CA^{enq}(b,i) 1_{APEenq=X}(i) - \frac{\hat{CA}_{diff}(X)}{\hat{CA}^{enq}_{HT,enq}(X)} \frac{\hat{CA}^{enq}_{HT,enq}(b,X)}{\hat{CA}^{enq}_{HT,enq}(X)} CA^{enq}(i) 1_{APEenq=X}(i)$$

Soit après factorisation :

$$lin(\hat{CA}_{ventilé}(b,X))_i = \frac{\hat{CA}^{enq}_{HT,enq}(b,X)}{\hat{CA}^{enq}_{HT,enq}(X)} CA^{fiscal}(i)(1_{APEenq=X}(i) - 1_{APErep=X}(i))$$
$$+ \frac{\hat{CA}_{diff}(X)}{\hat{CA}^{enq}_{HT,enq}(X)} \left[CA^{enq}(b,i) - \frac{\hat{CA}^{enq}_{HT,enq}(b,X)}{\hat{CA}^{enq}_{HT,enq}(X)} CA^{enq}(i) \right] 1_{APEenq=X}(i)$$

Bibliographie

[1] Brion P., « L'utilisation combinée de données d'enquête et de données administratives pour la production des statistiques structurelles d'entreprises », *Communication aux JMS2009*, Insee
[2] La macro SAS Calmar, disponible sur le site de l'Insee
[3] Les macros SAS Précalker et Calker, disponibles sur demande à l'UMS-E
[4] P. Ardilly, « Les Techniques de Sondage », éd. Technip
[5] Lavallée P. et Hidiroglou M. A., « Sur la stratification de populations asymétriques », 1988, *Techniques d'enquête*, 14, p. 33-43
[6] Dreyer A. « Optimisation de la stratification sur une variable discrète pour une population finie : Application à l'enquête annuelle d'entreprise », 2009, *rapport de stage Ensae,* Insee.