Comment calculer la précision des estimateurs composites d'Esane

Thomas Deroyon

Insee - 18 Bd Adolphe Pinard Paris, thomas.deroyon@insee.fr

Résumé. Les statistiques structurelles d'entreprise permettent de décrire l'appareil productif marchand, en analysant par secteur les informations comptables des entreprises et leur production par branche. Depuis 2008, ces statistiques sont réalisées en France par l'Insee grâce au dispositif Esane, qui mobilise à cette fin données administratives et données d'enquête et emploie des estimateurs composites exploitant au mieux la richesse et la diversité de l'information disponible. L'Insee calcule également chaque année la précision de ces estimateurs composites. Cet article a pour but de présenter, sur un cas simple, celui de l'estimation du total du chiffre d'affaires par secteur, les méthodes de calcul de précision utilisées par l'Insee.

Mots-clés. Estimation de variance, Enquêtes entreprises, ...

Introduction

Les statistiques structurelles d'entreprise permettent de décrire le système productif français, en estimant notamment les principaux agrégats comptables des entreprises (chiffre d'affaires, valeur ajoutée, excédent brut d'exploitation) par secteur d'activité, ainsi que la production des branches en France, *i.e.* le chiffre d'affaires de l'ensemble des fragments d'entreprise réalisant une même activité (produisant un même produit, un même service, ayant la même activité commercial). Elles sont produites depuis 2008 grâce au système Esane [1], qui mobilise des sources d'information variées (données administratives, enquêtes). Pour bénéficier au mieux de la diversité de l'information disponible, des estimateurs particuliers, appelés estimateurs composites, sont utilisés pour la production des statistiques structurelles. Le calcul de précision de ces estimateurs est une étape importante du processus Esane. Après avoir résumé l'organisation d'Esane et la forme des estimateurs composites, nous expliquons les méthodes et hypothèses employées pour le calcul de précision de ceux-ci, en nous attardant sur les estimateurs du niveau du chiffre d'affaires sectoriel.

[1] pour Elaboration des Statistiques ANnuelles d'Entreprise

1 Esane, le système de production des statistiques structurelles d'entreprise en France

Esane exploite principalement deux sources de données :

1. les déclarations fiscales des entreprises, dites liasses fiscales ; cette source contient toutes les variables des comptes de résultat et des bilans des entreprises opérant en France (notamment leur chiffre d'affaires). La source fiscale ne suffit cependant pas aux besoins de la statistique d'entreprise, qui doit connaître la décomposition détaillée du chiffre d'affaires de chaque entreprise entre ses différentes activités.

2. des enquêtes (Enquêtes Annuelles de Production – EAP – pour les secteurs industriels et Enquêtes Sectorielles Annuelles – ESA – pour les autres secteurs) ; ces enquêtes permettent principalement de recueillir la ventilation détaillée du chiffre d'affaires par branches d'activité, et ainsi de réévaluer l'activité principale exercée (APE) de chacune des entreprises de l'échantillon. Ces enquêtes ne sont cependant conduites que sur un échantillon de l'ensemble des entreprises du champ d'Esane. Plus précisément, le champ est divisé en deux parties : un ensemble de strates exhaustives, contenant les entreprises les plus importantes de chaque secteur, qui sont toutes intégrées à l'échantillon des enquêtes ; et un ensemble de strates non exhaustives, dont seule une partie des entreprises est interrogée. Cette partie de l'échantillon est sélectionnée aléatoirement sur la base d'un plan de sondages stratifié avec sondage aléatoire simple sans remise dans chaque strate.

De façon à limiter la charge statistique pesant sur les entreprises, seul un nombre limité de variables déjà présentes dans la liasse fiscale sont également demandées dans les enquêtes. Il s'agit du chiffre d'affaires et de sa ventilation agrégée en trois postes (activités commerciales, ventes de biens, ventes de services). Pour ces variables, les valeurs renseignées dans les enquêtes et dans la liasse fiscale sont mises en cohérence, de façon à avoir une valeur unique pour chaque entreprise, appelée « valeur REDI » [2]. Cette valeur est *a priori* plus fiable que les valeurs renseignées dans les liasses fiscales et dans les enquêtes séparément car le processus de réconciliation des données individuelles nous permet de corroborer les déclarations des entreprises en cas de correspondance des résultats ou au contraire de détecter des erreurs en cas de divergence.

Pour chaque entreprise, nous disposons de ce fait :

[2] Le processus de mise en cohérence des données d'enquête et fiscales s'appelle REconciliation des Données Individuelles (Redi).

- de deux valeurs de son activité principale exercée : celle, dite APE du répertoire, notée $\mathbb{A}$ renseignée dans le répertoire des entreprises avant l'enquête, et la valeur réévaluée grâce à l'enquête pour les entreprises appartenant à son échantillon, dite APE de diffusion et notée $\mathcal{A}$, égale à l'APE du répertoire pour les entreprises qui ne font pas partie de l'échantillon de l'ESA ou de l'EAP ;

- de deux valeurs des variables de la liasse fiscale : la valeur avant réconciliation des données individuelles (dite valeur IEG, pour Information Économique Générale, nom du processus informatique de production des liasses fiscales, notée $\mathbb{Y}$ par la suite) et la valeur après réconciliation des données individuelles (dite valeur REDI, notée $\mathcal{Y}$). La valeur REDI est égale à la valeur IEG pour toutes les variables non affectées par la réconciliation des données individuelles.

Les estimateurs employés dans Esane, dits estimateurs composites, essaient de tenir compte au mieux de la variété des informations disponibles.

Ainsi, pour estimer le chiffre d'affaires total des entreprises d'un secteur s donné, nous souhaiterions pouvoir calculer la somme des chiffres d'affaires REDI des entreprises dont l'APE correspondant le mieux à leur activité actuelle, i.e. l'APE de diffusion, réévaluée avec des données récentes, les place dans le secteur s. Le chiffre d'affaires REDI et l'APE de diffusion n'étant disponibles que pour les entreprises de l'échantillon de l'enquête (ESA+EAP), l'estimateur naturel est alors l'estimateur classique d'Horvitz-Thompson $\sum_{i\in\mathcal{R}} w_i^C\, \mathcal{Y}_i\, \mathbb{I}(\mathcal{A}(i)=s)$, où $\mathcal{R}$ désigne l'ensemble des répondants à l'enquête et w_i^C leur poids après traitements post-collecte.

Nous préférons cependant employer un estimateur tenant compte de l'information apportée par les liasses fiscales sur l'ensemble des entreprises du champ (voir Brion (2009))

$$Y_s^C = \sum_{i\in\mathcal{R}} w_i^C\, \mathcal{Y}_i\, \mathbb{I}(\mathcal{A}(i)=s)\, \mathbb{I}(i\in\mathcal{U}) + \sum_{i\in\mathbb{U}} \mathbb{Y}_i\, \mathbb{I}(\mathbb{A}(i)=s) - \sum_{i\in\mathcal{R}} w_i^C\, \mathbb{Y}_i\, \mathbb{I}(\mathbb{A}(i)=s) \quad (1)$$

où $\mathbb{U}$ désigne le champ d'Esane au lancement, utilisé comme base de sondage de l'échantillon des enquêtes, et $\mathcal{U}$ désigne le champ en fin de campagne [3].

[3]Le champ d'Esane est défini principalement par des critères d'activité et de catégorie juridique. Il est évalué au lancement de la campagne annuelle, à la fin de l'année sur laquelle porte la campagne, pour la constitution de l'échantillon de l'enquête, mais n'est pas parfaitement connu à cette date : les informations sur les cessations d'activité ou les changements de catégorie juridique ne sont par exemple connus souvent qu'assez tardivement, en fin de campagne (à l'été de l'année $N+2$ pour la campagne portant sur l'année N).

Cet estimateur par différence revient à corriger l'estimateur d'Horvitz-Thompson (égal à la première somme) de l'écart entre la somme des chiffres d'affaires IEG dont l'APE du répertoire appartient au secteur s et son estimateur HT construit sur l'échantillon de l'enquête (respectivement deuxième et troisième somme de l'estimateur). Les variables $\mathbb{Y}_i \, \mathbb{I}(\mathbb{A}(i) = s)$ et $\mathcal{Y}_i \, \mathbb{I}(\mathcal{A}(i) = s) \, \mathbb{I}(i \in \mathcal{U})$ étant en général fortement corrélées, et même très souvent identiques, cet estimateur par différence conduit en général à une amélioration de la qualité des estimations sectorielles par rapport à l'estimateur HT (voir Gros (2012b)).

Les estimateurs par différence posent cependant des problèmes appliqués à des niveaux très fins dans la nomenclature d'activité. En effet, rien ne garantit qu'ils soient toujours positifs. Sur les secteurs envisagés au niveau le plus fin possible, *i.e.* dans les 732 sous-classes de la nomenclature d'activité française, ce problème se pose souvent, si bien que la procédure d'estimation a dû être adaptée (voir Gros (2012a)) :

- dans les secteurs des niveaux agrégés de la nomenclature d'activité, *i.e.* pour les 272 groupes et *supra*, les estimateurs par différences de formule 1 sont employés ;

- dans les secteurs correspondant à des niveaux plus fins de la nomenclature d'activité, des estimateurs obtenus par ventilation des estimateurs par la différence au niveau groupe sont utilisés.

Plus précisément, l'estimateur composite du total du chiffre d'affaires dans le secteur s, inclus dans le groupe g, est obtenu en ventilant l'estimateur composite du total du chiffre d'affaires dans le groupe g par la répartition du chiffre d'affaires du groupe par sous secteurs observée dans l'enquête, *i.e.* :

$$\hat{Y}_s^C \; = \; \hat{Y}_g^C \, \frac{\sum_{i \in \mathcal{R}} \, w_i^C \, \mathcal{Y}_i \, \mathbb{I}(\mathcal{A}(i) = s) \, \mathbb{I}(i \in \mathcal{U})}{\sum_{i \in \mathcal{R}} \, w_i^C \, \mathcal{Y}_i \, \mathbb{I}(\mathcal{A}(i) = g) \, \mathbb{I}(i \in \mathcal{U})} \tag{2}$$

2 Les traitements post-collecte des ESA ayant un impact sur les calculs de précision

Plusieurs des traitements mis en œuvre dans Esane sont susceptibles d'avoir un effet sur la précision des estimateurs en sortie du dispositif. Nous décrivons ici rapidement ces traitements et leurs effets sur la variance de l'estimateur du total d'une variable Y.

2.1 La variance d'échantillonnage

Les estimateurs d'Esane sont construits à partir des résultats d'une enquête menée sur un échantillon d'entreprises sélectionnées aléatoirement, suivant un plan de sondages stratifié,

l'échantillon étant, hors des strates enquêtées exhaustivement, renouvelé par moitié chaque année (voir Bauer et al. (2009)). La variance d'échantillonnage peut être correctement approximée en supposant que l'échantillon est sélectionné suivant un plan de sondages par strates avec plan de sondage aléatoire simple sans remise dans chaque strate (voir Caron et Fuentes (2002)).

2.2 La correction de la non réponse

L'information attendue n'est pas toujours disponible, dans les liasses fiscales des entreprises ou dans l'enquête. La non-réponse fait l'objet de traitements par imputation (pour la non-réponse partielle et la la non-réponse totale dans les liasses fiscales et les strates exhaustives des enquêtes) et par repondération (pour la partie de l'échantillon tirée dans les strates non exhaustives). Seule l'effet de cette dernière est pris en compte dans les estimations de précision.

La correction de la non réponse (CNR) dans les strates non exhaustives de l'échantillon se fait par groupes de réponse homogène (GRH). Nous supposons ainsi qu'il est possible de partitionner l'échantillon en groupes dans lesquels toutes les entreprises ont la même probabilité de répondre à l'enquête. L'effet de la CNR par GRH sur la variance de l'estimateur d'un total s'obtient en reformulant l'hypothèse de l'existence des GRH en termes de plan de sondage : ainsi, le passage de l'échantillon total à l'échantillon réduit aux seuls répondants peut s'assimiler à une deuxième phase de tirage d'échantillon, suivant un plan de sondage stratifié dont les strates sont égales aux GRH. Sous cette hypothèse, il est possible d'obtenir une estimation de la variance de l'estimateur HT du total d'une variable.

2.3 Les traitements post correction de la non réponse des données de l'ESA et de l'EAP

D'autres traitements sont également mis en œuvre sur les données de l'ESA et de l'EAP : il s'agit de la winsorisation et du calage sur marges appliqué aux unités des strates non exhaustives de l'ESA et de l'EAP.

2.3.1 Winsorisation

La winsorisation est une méthode de statistique robuste visant à diminuer la variance des estimateurs en traitant les unités atypiques non aberrantes. Ces unités sont des entreprises très atypiques qui peuvent faire varier sensiblement la valeur des estimations obtenues sur l'échantillon d'une enquête suivant qu'elles y appartiennent ou pas (voir Guggemos (2010)).

Seules les unités atypiques non aberrantes présentes dans l'échantillon peuvent être traitées par la winsorisation, du fait de l'absence d'information effective sur les autres. La winsorisation consiste à définir, dans chaque strate h non exhaustive de l'enquête et pour une variable donnée Y – pour Esane, le chiffre d'affaires –, un seuil K_h et un coefficient g_h tel que, si Y dépasse, pour un individu, le seuil K_h, elle sera rabotée du coefficient g_h, i.e. l'estimateur d'Horvitz-Thompson du total de Y sur la strate s sera égal à $\sum_{i \in \mathcal{R}_h} w_i \, Y_i^*$ avec $Y_i^* = Y_i$ si $Y_i < K_h$ et $Y_i^* = K_h + g_h \, (Y_i - K_h)$ sinon. L'ajustement effectué sur la variable Y est ensuite appliqué à toutes les autres variables d'une même entreprise.

Cette procédure introduit un biais à la baisse dans l'estimation, puisque toutes les valeurs extrêmes sont rabotées, mais entraîne une diminution de la variance. Les seuils K_h et les coefficients g_h sont déterminés de manière à garantir que la baisse de variance soit assez importante pour générer au final une baisse de l'erreur quadratique moyenne et donc une amélioration de l'estimation. Ils sont en pratique calculés à partir des données de l'enquête et conduisent ainsi à des valeurs des variables dépendantes de l'échantillon, donc affectées par un aléa de sondage.
Nous négligeons cependant cet aléa dans les calculs de variance : le seul effet de la winsorisation sur les calculs de variance est qu'il conduit à remplacer, dans toutes les formules, les valeurs d'une variable Y par celles de sa winsorisée (nous n'en tiendrons d'ailleurs pas compte dans les notations par la suite, pour alléger les formules, mais toutes les sommes calculées sur l'échantillon de l'enquête font intervenir dorénavant des variables winsorisées).

2.3.2 Calage sur marges

Le calage (voir Deville et Särndal (1992)) est une opération ayant pour but d'améliorer la précision des estimations obtenues à partir d'une enquête en exploitant l'existence d'une information extérieure à l'enquête, en général l'existence d'informations sur les totaux de variables disponibles pour toutes les unités de l'échantillon. Le calage modifie, le moins possible, les poids des unités de l'échantillon de façon que les estimateurs d'Horvitz-Thompson calculés sur l'échantillon des totaux des variables de calage soient égaux à leur valeur connue par une source externe.

Dans l'ESA et l'EAP, un calage va être appliqué aux unités tirées dans les strates non exhaustives de l'échantillon pour garantir que celui-ci permette d'estimer parfaitement les totaux du chiffre d'affaires et du nombre d'entreprises par groupe ou par division constitués avec l'APE du répertoire.

Le calage a une incidence sur la précision de l'estimateur du total de Y. Dans la mesure où les poids après calage permettent de construire des estimateurs exacts (sans variance) du total des variables de calage, le calage permet de réduire la variance des estimateurs

des totaux de toutes les variables Y corrélées aux variables de calage. Plus précisément, $\hat{\mathbb{V}}(\sum_{i\in\mathcal{R}} w_i^C Y_i) \simeq \hat{\mathbb{V}}(\sum_{i\in\mathcal{R}} w_i \epsilon_i)$, avec w_i^c les poids après calage, w_i les poids avant calage et ϵ les résidus de la régression de Y sur les variables de calage. Le calage permet ainsi de « purger » la variance de l'estimateur du total de Y de sa partie expliquée par les variables de calage.

3 La précision des estimateurs du niveau du chiffre d'affaires sectoriel

3.1 les secteurs au niveau groupe et *supra*

Le calage sur marges du chiffre d'affaires fiscal ne peut pas systématiquement être mis en œuvre au niveau des groupes de la nomenclature d'activité française. Aussi nous nous placerons dans le cas général où $\sum_{i\in\mathbb{U}} \mathbb{Y}_i \mathbb{I}(\mathbb{A}(i) = s) \neq \sum_{i\in\mathcal{R}} w_i^C \mathbb{Y}_i \mathbb{I}(\mathbb{A}(i) = s)$. Dans ce cas,

$$\begin{aligned} \hat{\mathbb{V}}(\hat{Y}_s^C) &= \hat{\mathbb{V}}\{\sum_{i\in\mathcal{R}} w_i^C \left[\mathcal{Y}_i \mathbb{I}(\mathcal{A}(i) = s)\, \mathbb{I}(i \in \mathcal{U}) - \mathbb{Y}_i \mathbb{I}(\mathbb{A}(i) = s)\right]\} \\ &= \hat{\mathbb{V}}[\sum_{i\in\mathcal{R}} w_i \epsilon_i] \end{aligned} \tag{3}$$

avec ϵ_i le résidu de la régression de $\mathcal{Y}_i \mathbb{I}(\mathcal{A}(i) = s)\, \mathbb{I}(i \in \mathcal{U}) - \mathbb{Y}_i \mathbb{I}(\mathbb{A}(i) = s)$ sur les variables de calage estimée sur la population. Ces résidus ne sont pas calculables, les APE de diffusion et la valeur Redi du chiffre d'affaires n'étant disponibles que sur l'échantillon. La variance de l'estimateur du total du chiffre d'affaires sectoriel pour les secteurs de niveau groupe ou *supra* est donc estimée en remplaçant, dans la formule 3, les résidus ϵ_i par les résidus estimés sur l'échantillon de la régression linéaire pondérée par les poids w_i avant calage de $\mathcal{Y}_i \mathbb{I}(\mathcal{A}(i) = s)\, \mathbb{I}(i \in \mathcal{U}) - \mathbb{Y}_i \mathbb{I}(\mathbb{A}(i) = s)$ sur les variables de calage.

3.2 Pour les secteurs au niveau inférieur au groupe

Une technique de linéarisation (voir Deville (1999) et Dell et al. (2002)) doit être utilisée pour calculer la variance de l'estimateur donné par la formule 2. Cette méthode permet d'estimer la variance d'un estimateur complexe en l'approchant par la variance du total d'une variable, appelée linéarisée. Ainsi, il est possible de montrer que la variance de l'estimateur du chiffre d'affaires du secteur s inclus dans le groupe g est approximative-

ment égal à $\mathbb{V}(\sum_{i\in\mathcal{R}} w_i^C\,\hat{z}_i)$ avec

$$\begin{aligned}\hat{z}_i \;=\; & \frac{\hat{Y}_s^{HT}}{\hat{Y}_g^{HT}}\,\{\,\mathcal{Y}_i\,\mathbb{I}(\mathcal{A}(i)=g)\,\mathbb{I}(i\in\mathcal{U})\;-\;\mathbb{Y}_i\,\mathbb{I}(\mathbb{A}(i)=g)\,\}\\ & +\frac{\hat{Y}_g^{C}}{\hat{Y}_g^{HT}}\,\{\,\mathcal{Y}_i\,\mathbb{I}(\mathcal{A}(i)=s)\,\mathbb{I}(i\in\mathcal{U})\;-\;\frac{\hat{Y}_s^{HT}}{\hat{Y}_g^{HT}}\,\mathcal{Y}_i\,\mathbb{I}(\mathcal{A}(i)=g)\,\mathbb{I}(i\in\mathcal{U})\,\}\end{aligned}$$

avec $\hat{Y}_s^{HT} = \sum_{i\in\mathcal{R}} w_i^C\,\mathcal{Y}_i\,\mathbb{I}(\mathcal{A}(i)=s)\,\mathbb{I}(i\in\mathcal{U})$.

Au final, la variance de l'estimateur du chiffre d'affaires sectoriel d'un secteur s au niveau infragroupe de la nomenclature d'activité est estimée par $\hat{\mathbb{V}}[\sum_{i\in\mathcal{R}} w_i\,\hat{\eta}_i]$, avec $\hat{\eta}_i$ le résidu de la régression linéaire pondérée par les poids avant calage w_i et estimée sur l'échantillon $\mathcal{R}$ de la variable linéarisée $\hat{z}_i$ sur les variables de calage.

Bibliographie

[1] P. Bauer, G. Brilhault, E. Gros (2009), *Le plan de sondage de l'ESA*, Journées de Méthodologie Statistique, France.

[2] P. Brion (2009), *L'utilisation combinée de données d'enquêtes et de données administratives pour la production des statistiques structurelles d'entreprise*, Journées de Méthodologie Statistique, France.

[3] N. Caron, B. Fuentes (2002),*Estimation de précision des estimateurs de l'EAE-Commerce*, Document de travail de l'Insee, France.

[4] F. Dell, X. d'Haultfoeuille, P. Février, E. Massé, *Mise en œuvre du calcul de variance par linéarisation*, Journées de Méthodologie Statistique, France.

[5] J.C. Deville, C.E. Särndal (1992), *Calibration estimators in survey sampling*, Journal of the American Statistical Association, USA.

[6] J.C. Deville, *Estimation de variance pour des statistiques et des estimateurs complexes : linérisation et technique des résidus*, Techniques d'Enquête, Canada.

[7] E. Gros (2012a), *Esane ou les malheurs de l'estimation composite : comment gérer les valeurs négatives d'estimateurs par différence ?*, Journées de Méthodologie Statistique, France.

[8] E. Gros (2012b), *Le dispositif Esane, ou comment l'utilisation combinée de données administratives et de données d'enquête permet d'améliorer la qualité des données individuelles et des statistiques*, Septième colloque francophone sur les sondages - Rennes, France.

[9] F. Guggemos (2010), *Winsorisation dans les enquêtes annuelles auprès des entreprises françaises*, Sixième collogue francophone sur les sondages - Tanger, France.