ESANE OU LES MALHEURS DE L'ESTIMATION COMPOSITE : COMMENT GÉRER LES VALEURS NÉGATIVES D'ESTIMATEURS PAR DIFFÉRENCE ?

Emmanuel GROS (*)

(*) Insee, Unité Méthodologie Statistique - Entreprises

Introduction

Le nouveau dispositif de production des statistiques structurelles d'entreprises françaises, Esane (Élaboration des Statistiques ANnuelles d'Entreprises), a été mis en place en 2009. Il s'appuie sur une utilisation intensive de sources administratives, complétées par des données obtenues par une enquête statistique réalisée sur un échantillon d'entreprises. Cette utilisation conjointe de données administratives et de données d'enquête intervient à différentes étapes du processus d'exploitation des données – contrôles de cohérence inter-sources et réconciliation des données, correction de la non réponse, calage, etc. – et en particulier lors de la phase d'estimation via l'utilisation d'estimateurs par différence.

Ce type d'estimateur se révèle particulièrement adapté au contexte du dispositif Esane, puisqu'il permet de mobiliser conjointement les sources administratives et les résultats d'enquête, et présente de nombreux avantages, en particulier en termes de précision. Il ne garantit toutefois pas la positivité des estimations et peut donc conduire à des estimations négatives, et ce même lorsque toutes les données individuelles sont positives ou nulles. Ceci se révèle problématique, d'autant plus lorsque cela concerne des variables pour lesquelles des agrégats négatifs n'ont aucun sens économique. Si ces cas problématiques s'avèrent assez rares tant qu'on raisonne à un niveau suffisamment agrégé, le nombre d'estimations négatives s'accroît sensiblement dès que l'on s'intéresse à des agrégats à un niveau fin de la nomenclature, ou portant sur des variables à occurrences rares. Il est donc nécessaire d'adapter la procédure d'estimation afin de pallier ce problème.

La problématique développée ici est donc celle de la gestion des valeurs négatives problématiques dans le cadre d'estimations par différence. La première partie présente le dispositif Esane, en détaille les particularités et présente les estimateurs par différence initialement envisagés. La seconde partie revient sur les problèmes d'estimations négatives rencontrés et expose la méthode retenue pour gérer ces estimations problématiques.

1. Le contexte du système Esane

1.1. Un dispositif multi-sources

Comme mentionné précédemment, le système Esane[1] repose sur l'utilisation conjointe de différentes données administratives et d'une enquête statistique (figure 1).

Deux sources administratives sont mobilisées dans le cadre du système Esane :

- d'une part, les déclarations annuelles sur les bénéfices[2] adressées par les entreprises à la Direction générale des finances publiques (DGFiP) ;

[1] Cf. [1] pour plus de détails sur le sujet.

[2] Ces déclarations peuvent être utilisées directement à des fins statistiques car les informations comptables demandées par l'administration fiscale française font référence au Plan Comptable Général français, tout comme les variables comptables des enquêtes statistiques auprès des entreprises.

- d'autre part, les déclarations annuelles de données sociales (DADS), établies pour le compte des organismes de protection sociale et contenant des données sur les effectifs employés et les rémunérations.

Le répertoire inter-administratif Sirene sert de cadre de référence à ce dispositif – l'unité légale telle que définie dans ce répertoire constitue en effet l'unité de référence[3] tant du point de vue administratif que statistique – et permet une exploitation conjointe aisée de ces données provenant de sources différentes – le numéro Siren des unités légales de Sirene faisant office dans chaque source d'identifiant unique –. Il vient également compléter les données administratives en fournissant une information sur le classement sectoriel des unités, puisqu'on dispose dans Sirene d'un code d'activité principale (dit code APE) pour chaque unité du répertoire. Ce classement sectoriel ne permet malheureusement pas l'élaboration de statistiques sectorielles satisfaisantes, du fait de son potentiel manque de « fraîcheur » : en effet, le code APE disponible dans le répertoire peut avoir été déterminé depuis bon nombre d'années[4] et ne pas avoir été mis à jour depuis.

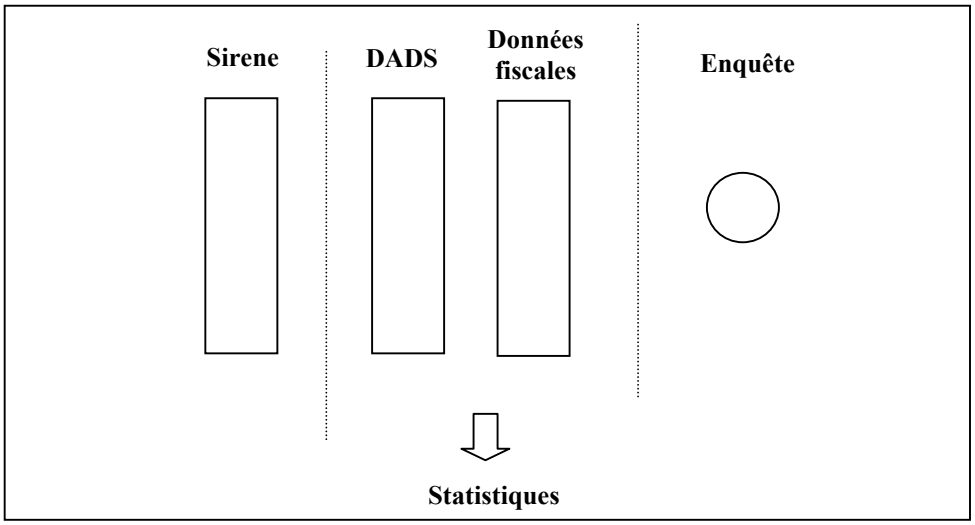
Ainsi, l'utilisation de ces seules sources administratives n'est pas suffisante pour répondre à l'ensemble des besoins exprimés en matière de statistiques structurelles d'entreprises. En particulier, la ventilation du chiffre d'affaires des entreprises selon leurs différentes activités n'est pas disponible dans les données fiscales. Or cette information est indispensable et répond à un double besoin de la statistique d'entreprises :

- d'une part, elle permet aux comptables nationaux d'établir les comptes de branches ;

- d'autre part, cette ventilation du chiffre d'affaires des entreprises permet de réestimer leur code APE selon une approche économique de leurs activités et non plus selon une approche déclarative. C'est ce classement sectoriel, obtenu par application d'un algorithme et basé sur la part relative de chaque activité dans le chiffre d'affaires total, et non celui de Sirene, qui doit être utilisé pour la production des statistiques sectorielles.

Afin de pallier cette incomplétude des données administratives, une enquête statistique[5] – l'ESA, enquête sectorielle annuelle, ou l'EAP, enquête annuelle de production pour le secteur de l'industrie – est réalisée sur un échantillon d'entreprises. Cette enquête permet d'obtenir en premier lieu un tronc commun de variables clefs – telles le chiffre d'affaires et sa ventilation par activité à un niveau fin ou des informations sur les restructurations –, ainsi que des caractéristiques sectorielles spécifiques au secteur de l'entreprise interrogée – superficie des magasins de vente au détail dans le cas du commerce, dépenses en carburant pour les entreprises de transport, etc. – également indisponibles dans les sources administratives.

Figure 1 : structure du dispositif Esane



[3] À l'exception de certaines grandes unités pour lesquelles un profilage, à savoir la définition d'unités de collecte spécifiques, est pratiqué.
[4] Par exemple à partir de la déclaration de l'entreprise, au moment où celle-ci s'est enregistrée dans Sirene.
[5] Cf. [2] pour plus de détails sur le sujet.

1.2. Un processus de réconciliation des données individuelles inter-sources

Une telle organisation ouvre de nouvelles perspectives en matière de contrôle et d'amélioration de la qualité des données individuelles. En effet, les trois sources d'informations mobilisées – données fiscales, données d'emploi et enquête – partagent un certain nombre de variables en commun :

- le chiffre d'affaires, ainsi que sa ventilation agrégée en ventes de marchandises / production de biens / production de services, sont ainsi disponibles à la fois dans les liasses fiscales et via les résultats de l'enquête[6] ;

- les variables « salaires » et « effectif salarié » sont quant à elles présentes à la fois dans les liasses fiscales et dans les DADS.

Dès lors, il est possible d'utiliser cette redondance d'information existant au sein du système Esane pour améliorer la qualité des données individuelles, via une procédure de contrôle et de mise en cohérence des données. Cette opération de réconciliation des données individuelles inter-sources constitue une des principales innovations du dispositif Esane, et fait l'objet du processus[7] de vérification sélective des données spécifique suivant :

- Pour chaque variable commune, une source de référence est définie comme suit :

Tableau 1: Définition de la source de référence pour chaque variable commune

Variable commune	Condition	Source de référence
Chiffre d'affaires	Données fiscales disponibles pour l'entreprise	Données fiscales
	Données fiscales non disponibles pour l'entreprise	Enquête
Ventilation agrégée du CA	Réponse de l'entreprise à l'enquête pour l'année d'intérêt	Enquête
	Pas de réponse de l'entreprise à l'enquête pour l'année d'intérêt	Données fiscales
Salaires	Données fiscales disponibles pour l'entreprise	Données fiscales
	Données fiscales non disponibles pour l'entreprise	DADS
Effectif salarié	Données d'emploi disponibles pour l'entreprise	DADS
	Données d'emploi indisponibles pour l'entreprise	Données fiscales

- Pour chaque variable en commun, un score, mesurant l'importance de l'écart entre les deux sources relatives à cette variable, est calculé selon la formule suivante :

$$\text{score} = \left| \frac{X_{S1} - X_{S2}}{T(X_p)} \right|, \quad \text{avec} \begin{cases} X_{S1} = \text{valeur de la variable X dans la source 1} \\ X_{S2} = \text{valeur de la variable X dans la source 2} \\ T(X_p) = \text{total de la variable X dans la source de référence,} \\ \qquad \text{pour le niveau d'agrégation retenu pour le contrôle.} \end{cases}$$

- Pour les unités dont le score est inférieur à 1 %, la valeur finale de la variable X est celle de la source de référence. Les autres unités font l'objet d'une vérification par un gestionnaire, qui détermine alors la « bonne[8] » valeur de cette variable X et ajuste également en conséquence les valeurs des variables liées à la variable contrôlée. Par exemple, si le gestionnaire modifie suite à contrôle le montant des ventes de marchandises, il doit également vérifier que les achats de marchandises déclarés dans la source fiscale sont toujours cohérents avec cette nouvelle valeur des ventes, ou à défaut ajuster le montant des achats.

[6] La ventilation agrégée du chiffre d'affaires se déduit en effet immédiatement des résultats de l'enquête par agrégation de la ventilation détaillée demandée.
[7] Processus REDI, pour « RÉconciliation des Données Individuelles », décrit plus en détail dans [3].
[8] Il peut s'agir de la valeur de la source de référence, de la valeur de l'autre source, ou d'une tierce valeur, issue du rappel de l'entreprise par le gestionnaire par exemple.

À l'issue de cette phase de réconciliation des données individuelles, on dispose donc, pour chaque variable concernée[9], d'une variable réconciliée[10] contenant la valeur finale de cette variable. Pour les variables « salaire » et « effectif salarié » (ainsi que pour l'ensemble des variables liées à celles-ci), les variables Redi sont disponibles pour l'ensemble des unités. Pour le chiffre d'affaires et sa ventilation agrégée (ainsi que pour l'ensemble des variables liées à celles-ci), ces variables Redi ne sont disponibles que pour les unités de l'échantillon.

1.3. Une procédure d'estimation spécifique pour les statistiques sectorielles.

On s'intéresse ici à l'estimation du total d'une variable administrative Y, pour un secteur X donné, sur l'ensemble des entreprises du champ de l'enquête (noté U). On cherche donc à estimer :

$$\sum_{i \in U} Y_i \, \mathbb{1}_{APE=X}(i) \quad (*)$$

où APE désigne le « vrai » classement sectoriel de l'unité, qui peut bien évidemment différer de celui connu dans le répertoire Sirene. Pour la grande majorité des variables administratives, non concernées par le processus Redi, la meilleure valeur disponible pour la variable Y est celle issue des sources administratives Y_i^{fiscal}. Pour les variables concernées par le processus Redi, la meilleure valeur disponible pour la variable Y est bien évidemment celle issue du processus Redi Y_i^{Redi}. Cependant, pour le chiffre d'affaires et sa ventilation agrégée en ventes de marchandises / production de biens / production de services, cette variable n'est disponible que pour les unités de l'échantillon. Quant au « vrai » classement sectoriel d'une unité, il n'est lui aussi connu que sur l'échantillon, suite aux résultats de l'enquête.

Ainsi, le matériau dont on dispose peut être représenté sous la forme d'une base de données rectangulaire incomplète, avec d'une part une base de données complète[11] sur l'ensemble du champ pour les variables administratives, et d'autre part des données disponibles uniquement pour les unités de l'échantillon pour les variables d'enquêtes ainsi que certaines variables Redi. Or certaines de ces variables – principalement le code APE, la ventilation du chiffre d'affaires par activités et les variables Redi liées au chiffre d'affaires et à sa ventilation agrégée – constituent des informations cruciales pour l'établissement des statistiques structurelles d'entreprises. Afin de prendre en compte au mieux ces informations centrales disponibles uniquement pour les unités de l'échantillon tout en exploitant au maximum les sources administratives exhaustives, une procédure d'estimation spécifique, combinant calage sur marges et estimateur par différence, a été mise en œuvre.

Tout d'abord, le fait de disposer de données administratives exhaustives permet l'utilisation de techniques de calage, en vue d'améliorer la précision globale des estimations. Les chiffres d'affaires sectoriels et le nombre d'entreprises par secteur constituant deux résultats importants des statistiques structurelles d'entreprises, le calage a été réalisé sur ces deux variables. Plus précisément, l'opération de calage – qui porte uniquement sur la partie échantillonnée de l'enquête et fait suite à une phase de correction de la non-réponse par repondération – consiste en une modification des poids w_i des unités selon les équations de calage suivantes :

$$\begin{cases} \sum_{i \in R} w_i CA^{fiscal}(i) \, \mathbb{1}_{APE\_rep=X}(i) = \sum_{i \in U\text{-exhaustif}} CA^{fiscal}(i) \, \mathbb{1}_{APE\_rep=X}(i) \\ \sum_{i \in R} w_i \, \mathbb{1}_{APE\_rep=X}(i) = \sum_{i \in U\text{-exhaustif}} \mathbb{1}_{APE\_rep=X}(i) \end{cases}$$

[9] i.e. les variables communes contrôlées, ainsi que les variables liées à ces dernières.
[10] Très originalement baptisée « variable Redi »
[11] Au problème des liasses fiscales manquantes près, qui font l'objet d'une procédure d'imputation des données spécifique. Au final, on dispose donc de données fiscales déclarées ou imputées pour l'ensemble des unités actives ou présumées actives du champ, et on considère donc cet ensemble de données comme complet.

où APE_rep est le code APE issu du répertoire Sirene et $CA^{fiscal}(i)$ le chiffre d'affaires de l'entreprise i issu des données fiscales. L'opération de calage conduit ainsi à ajuster les poids w_i de façon à ce que l'échantillon extrapolé permette de retrouver les agrégats sectoriels de chiffre d'affaires et de nombre d'entreprises tels que définis dans le répertoire – et non pas les agrégats sectoriels « réels » au moment de l'enquête, impossibles à délimiter dans le répertoire en raison des changements de secteur non connus de façon exhaustive –.
Dans la pratique, le niveau sectoriel retenu pour ce calage est en général le niveau « groupe » – trois premiers caractères du code APE en NAF Rév.2 –, avec parfois quelques regroupements de groupes, de façon à limiter l'ampleur des modifications de poids liées au calage.

De plus, l'existence de deux codes APE – celui *ex ante* du répertoire, connu de façon exhaustive, et celui issu de l'enquête statistique, disponible uniquement sur échantillon –, conduit à proposer d'utiliser l'estimateur par différence suivant, pour un secteur donné X :

$$\hat{Y}_{diff}^{X} = \sum_{i \in exh \oplus R} w_i Y_i^{Redi} \, \mathbb{1}_{APE\_enq=X}(i) + \sum_{i \in U} Y_i^{fiscal} \, \mathbb{1}_{APE\_rep=X}(i) - \sum_{i \in exh \oplus R} w_i Y_i^{fiscal} \, \mathbb{1}_{APE\_rep=X}(i)$$

où « exh » désigne la partie exhaustive de l'échantillon – pour laquelle la correction de la non-réponse est effectuée par imputation, qui n'est pas impliquée dans les opérations de calage et dont les unités conservent donc un poids w_i unitaire – et R l'ensemble des répondants de la partie échantillonnée – pour lesquels les poids w_i sont ceux résultant des phases de correction de la non-réponse par repondération et de calage –. Pour les variables non concernées par le processus Redi, on pose par convention $Y_i^{Redi} = Y_i^{fiscal}$.
Les variables $Y_i^{Redi} \, \mathbb{1}_{APE\_enq=X}(i)$ et $Y_i^{fiscal} \, \mathbb{1}_{APE\_rep=X}(i)$ étant en général fortement corrélées, et même très souvent identiques[12], cet estimateur par différence améliore en général la qualité des estimations sectorielles[13] par rapport à l'estimateur usuel de type Horvitz-Thompson post-calage $\sum_{i \in exh \oplus R} w_i Y_i^{Redi} \, \mathbb{1}_{APE\_enq=X}(i)$.

2. Estimateurs par différence et gestion des estimations négatives

2.1. Problématique

L'estimateur par différence post-calage proposé ci-dessus présente de nombreux avantages : il s'agit en effet d'un estimateur linéaire – propriété particulièrement intéressante dans le contexte du dispositif Esane, dont les variables présentent la particularité d'être très fréquemment reliées entre elles par des équations comptables, qu'il convient de respecter lors des estimations – globalement plus efficace que les estimateurs usuels de type Horvitz-Thompson. Il ne garantit cependant pas la positivité des estimations et peut donc conduire à des estimations négatives, alors même que toutes les données individuelles sont positives ou nulles. Ceci se révèle problématique, d'autant plus lorsque cela concerne des variables pour lesquelles des agrégats négatifs n'ont aucun sens économique, comme les ventes de marchandises par exemple[14].

En pratique, ce type d'estimations problématiques relève majoritairement de deux cas de figure bien distincts :

[12] Hors effet Redi, ce cas de figure arrive d'autant plus fréquemment qu'on se place à un niveau sectoriel plus agrégé…
[13] On trouvera dans [4] une estimation *ex ante* du gain de précision attendu, et dans [5] une première évaluation *ex post* de l'impact en termes de précision des estimations de ce nouveau système.
[14] Symétriquement, on observe un problème similaire dans le cas d'estimations positives de quantités censément négatives, telle la variable « Perte comptable de l'exercice ».

- d'une part l'estimation de variables fortement affectées par le processus Redi, et pour lesquelles la valeur finale Y_i^{Redi} diffère fortement de la valeur de la source fiscale Y_i^{fiscal}. Ceci concerne les variables de ventilation agrégée du chiffre d'affaires « ventes de marchandise », « production de biens » et « production de services », ainsi que les variables d'achats et de variations de stocks correspondantes. En effet, pour ces variables, l'enquête constitue la source de référence du processus Redi, et il est donc très fréquent que la ventilation agrégée du chiffre d'affaires post-Redi, ainsi que les achats correspondants, s'éloigne sensiblement des déclarations fiscales ;

- d'autre part, l'estimation de statistiques portant sur des petits domaines. Ce cas de figure se rencontre soit lorsqu'on s'intéresse à l'estimation de quantités à un niveau fin – classement sectoriel au niveau sous-classe, estimation par [groupe ⊗ tranche de taille] ou encore par [groupe ⊗ région], etc. –, soit lorsqu'on travaille sur des variables comptables à occurrences rares, du type « Autres charges et dépenses somptuaires ». Dans ces conditions, les estimations par différence sont peu robustes[15] – taille de la population U d'intérêt de l'ordre de quelques centaines, quelques dizaines d'unités seulement concernées dans l'échantillon – et il suffit qu'une unité de l'échantillon avec un poids élevé change de secteur pour que l'on obtienne parfois une estimation négative.

De fait, tant qu'on raisonne à un niveau suffisamment agrégé – en pratique, niveau groupe (trois premiers caractères de la NAF) ou supérieur –, ces estimations problématiques s'avèrent assez rares[16], et une procédure de corrections individuelles (par mise à zéro ponctuelle des agrégats incriminés par exemple) peut être envisagée. En revanche, dès qu'on s'intéresse à des agrégats à un niveau plus fin, le nombre d'estimations négatives – relevant cette fois-ci majoritairement du second cas de figure – s'accroît sensiblement et un traitement au cas par cas garantissant la cohérence des estimations entre les différents niveaux n'est plus envisageable.

2.2. La procédure d'estimation retenue *in fine*

2.2.1. Grands principes

Il a donc été nécessaire de repenser la procédure d'estimation, afin de gérer ce problème d'estimations négatives (respectivement positives) de variables censément positives (respectivement négatives). Cette tâche a été rendue d'autant plus difficile par la richesse et la complexité du système Esane : les statistiques produites par ce système sont en effet soumises à de nombreuses contraintes de cohérence, tant « verticales » – cohérence des estimations portant sur différents niveaux de nomenclature hiérarchiquement imbriqués – que « horizontales » – cohérence lors de l'estimation de variables liées entre elles par des relations comptables –, que la méthode d'estimation se doit de respecter.

Cette multiplicité de contraintes à respecter, ainsi que le nombre important d'estimations problématiques observées dans les premiers tests pour les statistiques de niveau infra-groupe, nous ont rapidement conduit à abandonner l'idée d'utiliser une unique méthode d'estimation basée sur l'estimation par différence pour tous les niveaux, au profit d'une nouvelle procédure d'estimation impliquant des traitements différenciés selon le niveau de détail des statistiques produites :

- pour les agrégats de niveau groupe, il a été décidé de conserver la méthode d'estimation par différence initialement envisagée, et de gérer les estimations problématiques au niveau des agrégats selon des procédures qui seront détaillées en 2.2.2 ;

[15] Tout comme les estimateurs de type Horvitz-Thompson post-calage d'ailleurs, même si sur ces derniers ce problème de robustesse ne se manifeste pas de manière aussi flagrante…

[16] L'ensemble des estimations problématiques niveau groupe – qui ne concerne que les variables de ventilation agrégée du chiffre d'affaires et les variables d'achats et de variations de stocks liées (1er cas de figure), ainsi que quelques variables à occurrences rares (2nd cas de figure) – représente moins de 0,1 % du nombre total d'estimations niveau groupe.

- les estimations de niveaux supérieurs au groupe se déduisent immédiatement des estimations de niveau groupe par simple agrégation ;

- enfin, pour les estimations de niveau infra-groupe, la méthode retenue, détaillée en 2.2.3, consiste à ventiler les estimations de niveau groupe selon une clef de répartition issue de l'enquête seule.

Cette nouvelle mécanique globale d'estimation assure par construction une cohérence parfaite des estimations entre les différents niveaux de nomenclature, et garantit un signe valide aux agrégats obtenus. Elle implique en revanche le recours à des méthodes d'estimations non linéaires, que ce soit pour la gestion des estimations problématiques au niveau groupe ou pour le calcul des agrégats de niveau infra-groupe. Or comme évoqué précédemment, les variables du système Esane présentent la particularité d'être très fréquemment reliées entre elles par des équations comptables, qu'il convient de respecter lors des estimations. Aussi, dès lors que l'on envisage d'utiliser des méthodes d'estimations non linéaires, il n'est plus possible de traiter les variables indépendamment les unes des autres lors des estimations.

Par conséquent, il a été décidé de procéder à des estimations « directes »[17] uniquement pour les variables « élémentaires »[18], et d'en déduire les estimations pour les variables non élémentaires via les égalités comptables.

2.2.2. Méthode d'estimation pour les niveaux groupe et supérieurs

La procédure d'estimation retenue pour les niveaux groupe et supérieurs est la suivante :

❶ Calcul, pour les variables élémentaires, de l'estimateur par différence $\hat{Y}^{G}_{diff}$ au niveau groupe.

$$\hat{Y}^{G}_{diff} = \sum_{i \in exh \oplus R} w_i Y_i^{Redi} \, \mathbb{1}_{groupe\_enq=X}(i) + \sum_{i \in U} Y_i^{fiscal} \, \mathbb{1}_{groupe\_rep=X}(i) - \sum_{i \in exh \oplus R} w_i Y_i^{fiscal} \, \mathbb{1}_{groupe\_rep=X}(i)$$

❷ Gestion des estimations négatives (respectivement positives) « à tort » sur les agrégats de niveau groupe relatifs aux variables élémentaires ainsi calculés.

Comme évoqué précédemment, les cas d'estimations négatives à tort s'avèrent assez rares au niveau groupe, et concernent deux groupes de variables bien distincts :

- d'une part les variables à occurrences rares : pour de telles variables, un agrégat négatif à tort est simplement révélateur du manque de robustesse de l'estimation, et il a donc été décidé de considérer ces estimations négatives à tort comme non significatives et de ne pas les diffuser ;

- d'autre part les variables de ventilation agrégée du chiffre d'affaires ainsi que les achats et variations de stocks correspondants : les variables « ventes de marchandise », « production de biens » et « production de services » constituent des variables spécifiques du dispositif Esane, tant au niveau des traitements dont elles font l'objet – elles sont fortement affectées par Redi – qu'en ce qui concerne leur importance dans les statistiques produites par le système – elles sont des composantes essentielles de la valeur ajoutée, et les agrégats sectoriels relatifs à ces variables servent de base à l'établissement de la matrice [secteurs ⊗ branches] des chiffres d'affaires, et donc au calcul des chiffres d'affaires branches –, elles font donc l'objet d'un traitement particulier.

  Les agrégats négatifs à tort concernant l'une de ces trois variables sont gérés selon la procédure suivante : mise à zéro de la variable négative, report du montant négatif ainsi traité sur la variable correspondant à la branche principale de l'entreprise et ajustement en

[17] i.e. par calcul d'un estimateur, via la procédure décrite précédemment, à partir des données individuelles.
[18] i.e. des variables n'intervenant qu'en tant que composantes et jamais comme solde dans les équations comptables.

conséquence des variables d'achats et de variations de stock correspondantes. Cette procédure spécifique permet d'obtenir, pour les trois variables « ventes de marchandise », « production de biens » et « production de services », des agrégats sectoriels positifs ou nuls pour tous les groupes – point essentiel en vue de la constitution de la matrice [secteurs ⊗ branches] des chiffres d'affaires – tout en conservant inchangés les agrégats de chiffre d'affaires et de valeur ajoutée.

À l'issue de cette correction spécifique des agrégats négatifs à tort pour les trois variables « ventes de marchandise », « production de biens » et « production de services », les éventuels agrégats négatifs à tort relatifs aux variables d'achats correspondantes sont considérés comme non diffusables.

❸ Une fois l'ensemble des estimations portant sur des variables élémentaires réalisées, calcul de tous les agrégats résultant d'une équation comptable. Par exemple, le chiffre d'affaires par groupe est estimé via la somme des agrégats groupes « ventes de marchandise » ⊕ « production de biens » ⊕ « production de services ».

❹ pour les niveaux de nomenclature plus agrégés (divisions, sections, etc.), les estimations s'en déduisent par agrégations des estimations de niveau groupe.

Pour ces niveaux de diffusion, la gestion des rares estimations problématiques s'effectue donc directement au niveau des agrégats issus de la procédure d'estimation par différence, et de manière différente selon les variables. Elle se traduit :

- par l'introduction de biais[19] dans les estimations relatives à la ventilation agrégée du chiffre d'affaires et aux variables d'achats et de variations de stock correspondantes. À noter que, par construction, cette procédure n'affecte pas les estimations de chiffre d'affaires et de valeur ajoutée – variables centrales du système – qui restent donc sans biais ;

- par une très légère[20] perte d'information pour les autres variables, pour lesquelles l'information jugée non significative n'est pas diffusée.

2.2.3. Méthode d'estimation pour les niveaux infra-groupe

La procédure d'estimation retenue pour les niveaux infra-groupe est la suivante :

❶ Pour les variables élémentaires, ventilation de l'agrégat niveau groupe $\hat{Y}^{G}_{diff}$, issu de la procédure décrite au paragraphe précédent, selon la « structure Horvitz-Thompson » propre à chaque variable élémentaire. Plus précisément, pour un groupe G et un domaine D ⊂ G donnés, on estime le total de Y sur le domaine D via la formule suivante :

$$\hat{Y}^{D}_{ratio} = \hat{Y}^{G}_{diff}\frac{\hat{Y}^{D}_{HT}}{\hat{Y}^{G}_{HT}} = \hat{Y}^{G}_{diff}\frac{\sum_{i\in exh\oplus R} w_i Y_i^{Redi}\, \mathbb{1}_{Domaine\_enq=D}(i)}{\sum_{i\in exh\oplus R} w_i Y_i^{Redi}\, \mathbb{1}_{Groupe\_enq=G}(i)}$$

Remarque 1 : les agrégats considérés comme non significatifs et de ce fait non diffusés au niveau groupe, suite à la procédure décrite au paragraphe précédent, ne sont bien évidemment pas concernés par ces calculs, et ne donnent lieu à aucune diffusion de niveau infra-groupe.

Remarque 2: lorsque l'on travaille sur une variable Y à occurrences rares, il arrive que l'estimateur retenu au niveau groupe soit l'estimateur par différence et que ce dernier soit non nul tandis que l'estimateur d'Horvitz-Thompson post-calage de niveau groupe est nul. Dans une telle situation, il est

[19] À la hausse pour les variables mises à zéro, à la baisse pour les variables sur lesquelles on a reporté un montant négatif…
[20] Ce cas ne concerne en effet que 0,01 % du nombre total d'estimations niveau groupe.

impossible de calculer une clef de répartition à partir de l'échantillon pour ventiler l'estimateur du niveau groupe au niveau infra groupe. Afin de pallier ce problème, on utilise dans ce cas une clef de répartition calculée sur l'ensemble du champ :

$$\hat{Y}^{D}_{ratio} = \hat{Y}^{G}_{diff}\,\frac{\hat{Y}^{D}_{U}}{\hat{Y}^{G}_{U}} = \hat{Y}^{G}_{diff}\,\frac{\sum_{i\in U} Y_i^{fiscal}\ \mathbb{1}_{Domaine\_rep=D}(i)}{\sum_{i\in U} Y_i^{fiscal}\ \mathbb{1}_{Groupe\_rep=G}(i)}$$

❷ Une fois l'ensemble des estimations portant sur des variables élémentaires réalisées, calcul de tous les agrégats résultant d'une équation comptable.

Pour les agrégats de niveaux infra-groupe, la gestion des estimations problématiques de l'estimateur par différence passe donc par un abandon pur et simple de ce type d'estimateur, au profit d'un estimateur procédant par ventilation des agrégats de niveau groupe suivant des structures observées dans l'enquête.

En ce qui concerne la qualité des estimations, le nouvel estimateur proposé est asymptotiquement sans biais[21]. Nous avons d'ailleurs vérifié cette propriété en comparant, pour des estimations de niveau sous-classe – classement sectoriel le plus fin selon les cinq caractères du code APE en NAF Rév.2 –, ces nouveaux estimateurs avec les estimateurs par différence pour quelques variables importantes du dispositif Esane[22] : quelle que soit la variable considérée, les estimations issues des deux approches sont bien globalement proches – dans 80 % des cas, l'ampleur des différences constatées reste inférieure à ± 10 % du total considéré (± 20 % pour le nombre d'entreprises), inférieure à ± 15 % du total considéré dans 90 % des cas (± 30 % pour le nombre d'entreprises) – et les cas de divergence importante concernent précisément les secteurs de petite taille.
Quant à la précision des estimations, le nouvel estimateur mobilisant de manière moins intensive les informations administratives de niveau fin, on s'attend théoriquement à une diminution de la précision des estimations de niveau infra-groupe. Afin de quantifier cette perte de précision potentielle, nous avons calculé, toujours pour le même groupe de variables, la variance[23] des estimations de niveau sous-classe selon les deux approches. Le tableau 2 donne les résultats de la comparaison des écarts-types qui s'en déduisent, pour les sous-classes comptant au minimum 10 unités répondantes[24] :

Tableau 2 : ratio des écarts-types « estimateur par différence » / « estimateur par ventilation »

	Chiffre d'affaires	Valeur ajoutée	Excédent brut d'exploitation	Nombre d'entreprises	Effectif salarié	Frais de personnel	Total de l'actif	Total du passif
Moyenne	0,90	0,89	0,84	0,89	0,85	0,89	0,86	0,86
Max	11,53	9,58	4,01	1,53	7,79	10,04	6,04	6,20
Q99	2,85	2,96	3,09	1,42	2,64	3,05	3,33	3,49
Q95	1,53	1,46	1,54	1,15	1,25	1,33	1,82	1,87
Q90	1,21	1,18	1,19	1,07	1,08	1,13	1,20	1,22
Q75	1,00	1,00	1,01	1,00	0,99	1,00	1,00	1,00
Médiane	0,90	0,89	0,86	0,95	0,88	0,88	0,86	0,85
Q25	0,68	0,68	0,52	0,81	0,69	0,68	0,56	0,56
Q10	0,37	0,39	0,22	0,63	0,40	0,37	0,30	0,27
Q5	0,18	0,21	0,13	0,47	0,25	0,26	0,16	0,13
Q1	0,07	0,06	0,02	0,08	0,00	0,05	0,05	0,03
Min	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00

[21] Comme tout estimateur par substitution qui se respecte…
[22] Il s'agit des variables « nombre d'entreprises », « chiffre d'affaires », « valeur ajoutée », « excédent brut d'exploitation », « frais de personnel », « effectif salarié en ETP », « total de l'actif » et « total du passif ».
[23] La procédure de calcul de variance prend en compte la variance d'échantillonnage liée au plan de sondage de l'ESA, à la correction de la non-réponse par repondération selon la technique des groupes de réponses homogènes et au calage effectués sur la partie échantillonnée de l'enquête, mais pas le traitement de la non-réponse par imputation dans la partie exhaustive, ni la gestion par imputation des liasses fiscales manquantes.
[24] Les groupes comportant une unique sous-classe, et pour lesquels les estimations sont par construction identiques entre les deux approches, ont été exclus de l'analyse.

On observe donc bien généralement une diminution de la précision des estimations quand on passe à l'estimateur par ventilation, mais celle-ci reste globalement limitée avec des augmentations moyennes et médianes des écart-types de l'ordre de 10 à 15 %.

Conclusion

Pour faire face aux problèmes d'estimations négatives à tort rencontrées lors des premiers tests, la procédure d'estimation du système Esane initialement envisagée – estimation par différence pour toutes les variables et à tous les niveaux de diffusion – a été fortement amendée, avec d'une part l'abandon du principe d'estimation directe pour toutes les variables, au profit d'estimations directes pour les seules variables élémentaires et indirectes pour les variables soldes, et d'autre part et surtout une différenciation de la méthode d'estimation retenue selon le niveau de détail des statistiques produites :

- estimations par différence avec gestion des cas problématiques, assez rares, au niveau des agrégats pour les statistiques de niveau groupe, les estimations de niveaux supérieurs s'en déduisant par agrégation ;

- estimations de niveau infra-groupe par ventilation des agrégats de niveau groupe selon des clefs de répartition issues de l'enquête.

Cette nouvelle mécanique d'estimation assure par construction une cohérence parfaite des estimations entre les différents niveaux de nomenclature ainsi que le respect des égalités comptables reliant les variables, et garantit un signe valide aux agrégats obtenus. Elle implique en revanche une légère perte d'information – certaines statistiques, jugées non significatives, n'étant pas diffusées – et impose surtout l'abandon de l'estimateur par différence au niveau infra-groupe au profit d'un estimateur mobilisant moins d'informations administratives de niveau fin – puisque les structures permettant le passage du niveau groupe aux niveaux infra-groupe sont issues de l'enquête seule –, ce qui se traduit par une perte de précision mesurée des statistiques de niveau infra-groupe. Il semble cependant que cette perte potentielle de précision soit le prix à payer pour obtenir des estimations diffusables et cohérentes entre elles au sein d'un système aussi riche et complexe qu'Esane.

Bibliographie

[1] Brion Ph., « L'utilisation combinée de données d'enquête et de données administratives pour la production des statistiques structurelles d'entreprises », *Actes des 10èmes Journées de Méthodologie Statistique*, mars 2009, Insee.

[2] Haag O., « Reengineering French structural business statistics : redesign of the annual survey », *paper presented at the Q2010 conference*, Helsinki, mai 2010.

[3] Haag O., « Esane : À la recherche d'une cohérence maximale des données multi-sources sur les entreprises par le biais de micro et macro contrôles », *Actes des 11èmes Journées de Méthodologie Statistique*, janvier 2012, Insee.

[4] Bauer P., Brilhault G., Gros E., « Le plan de sondage de l'ESA (enquête sectorielle annuelle du futur dispositif Esane) », *Actes des 10èmes Journées de Méthodologie Statistique*, mars 2009, Insee.

[5] Gros E., « Quality improvement of individual data and statistical outputs based on combined use of administrative and survey data », *UN/ECE work session on statistical data editing*, Ljubljana, mai 2011.