LE DISPOSITIF ESANE, OU COMMENT L'UTILISATION COMBINEE DE DONNEES ADMINISTRATIVES ET DE DONNEES D'ENQUETE PERMET D'AMELIORER LA QUALITE DES DONNEES INDIVIDUELLES ET DES STATISTIQUES

Emmanuel GROS[1]

[1] *Insee, 18 bd Adolphe Pinard, 75675 PARIS CEDEX 14 – emmanuel.gros@insee.fr*

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 (déclarations annuelles sur les bénéfices adressées par les entreprises à la direction générale des Impôts, déclarations annuelles de données sociales), 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 et ouvre de nouvelles perspectives, tant en termes de contrôle des données individuelles que lors de la phase d'estimation, via la mise en œuvre de procédures de calage et l'utilisation d'estimateurs composites spécifiques.

La première partie de cet article présente le dispositif Esane et détaille le mécanisme de contrôle de cohérence des données individuelles ainsi que les estimateurs retenus. La seconde partie présente une évaluation de l'impact des améliorations méthodologiques sur les statistiques produites.

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 de ce dispositif :

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

[1] Cf. Brion (2011) 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.

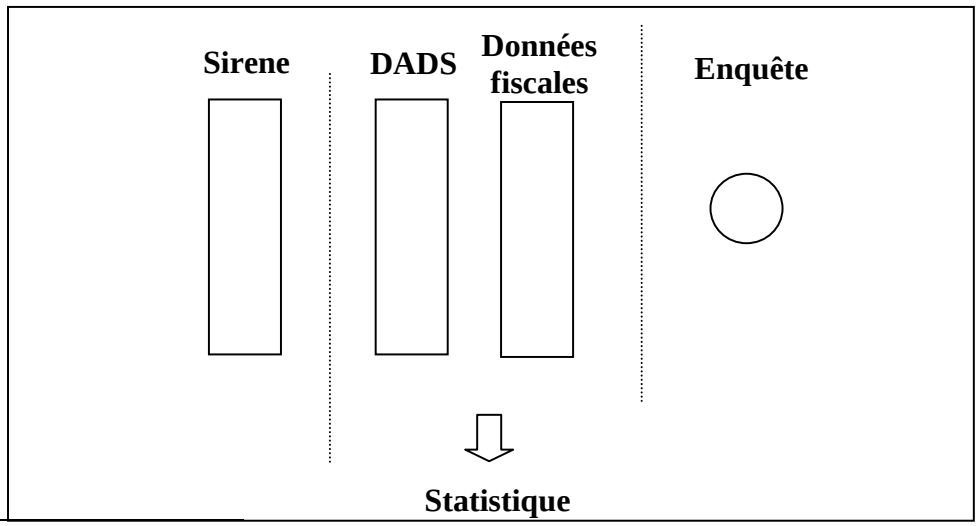
Le répertoire inter-administratif Sirene sert de cadre de référence à ce dispositif – l'unité légale telle que définie dans ce répertoire constitue en effet l'unité de référence[3] tant du point de vue administratif que statistique – et permet une exploitation conjointe aisée de ces données provenant 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 (cf. Haag, 2009, pour plus de détails sur le sujet) – 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.

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[5] ;

- 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[6] 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

- Puis un score, mesurant l'importance de l'écart entre les deux sources relatives à une variable donnée, 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

[5] La ventilation agrégée du chiffre d'affaires se déduit en effet immédiatement des résultats de l'enquête par agrégation de la ventilation détaillée demandée.

[6] Processus REDI, pour « RÉconciliation des Données Individuelles », décrit plus en détail dans Haag (2012).

la source de référence. Les autres unités font l'objet d'une vérification par un gestionnaire, qui détermine alors la « bonne[7] » 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.

À l'issue de cette phase de réconciliation des données individuelles, on dispose donc, pour chaque variable concernée[8], d'une variable réconciliée[9] 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[10] 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.

---

[7] 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.

[8] i.e. les variables communes contrôlées, ainsi que les variables liées à ces dernières.

[9] Très originalement baptisée « variable Redi »

[10] 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.

Tout d'abord, le fait de disposer de données administratives exhaustives permet l'utilisation de techniques de calage – cf. Deville et Särndal (1992) –, 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 :

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

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 :

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

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[11], cet estimateur par différence devrait conduire en général à une amélioration de la qualité des estimations sectorielles[12] par rapport à l'estimateur usuel de type Horvitz-Thompson post-calage $\sum_{i\in exh\oplus R} w_i Y_i^{Redi}\ \mathbb{1}_{APE\_enq=X}(i)$.

[11] Hors effet Redi, ce cas de figure arrive d'autant plus fréquemment qu'on se place à un niveau sectoriel plus agrégé…
[12] On trouvera dans Bauer, Brilhault et Gros (2009) une estimation *ex ante* du gain de précision attendu.

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[13].

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

- d'une part l'estimation de variables fortement affectées par le processus Redi, et pour lesquelles la valeur finale Y_j^{Redi} diffère fortement de la valeur de la source fiscale Y_j^{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[14] – 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 Rév.2) ou supérieur –, ces estimations problématiques s'avèrent assez rares[15], 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.

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

[13] Symétriquement, on observe un problème similaire dans le cas d'estimations positives de quantités censément négatives, telle la variable « Perte comptable de l'exercice ».

[14] 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…

[15] 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.

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, décrite dans le paragraphe suivant. 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 »[16] uniquement pour les variables « élémentaires »[17], et d'en déduire les estimations pour les variables non élémentaires via les égalités comptables.

La procédure d'estimation finalement mise en œuvre est la suivante :

⇒ pour les agrégats de niveau groupe et supérieurs, 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 la procédure 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

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

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 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 les estimations de niveau infra-groupe, la méthode retenue consiste à ventiler les estimations de niveau groupe selon une clef de répartition issue de l'enquête seule, selon la procédure 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 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 :

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

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

2. Évaluation de ces améliorations méthodologiques

2.1. Impact du processus Redi

Nous présentons ici les résultats du processus Redi, relatifs aux variables « chiffre d'affaires », « ventes de marchandise », « production de biens » et « production de services » pour la campagne Esane 2010.

Tout d'abord, notons que concernant la variable « chiffre d'affaires », les données administratives et les données d'enquête sont globalement cohérentes. En effet, 60 % des entreprises reportent des chiffres d'affaires identiques – à l'unité près – dans les deux sources, et seules 14 % des unités présentent une différence relative supérieure[18] à 15 %. Qui plus est, les contrôles manuels et les rappels téléphoniques menés par les gestionnaires pour les unités détectées par la procédure de vérification sélective viennent confirmer le choix qui a été fait de la source de référence dans Redi, tant pour le chiffre d'affaires que pour sa ventilation agrégée : en effet, la valeur finale de Redi renseignée par un gestionnaire correspond à celle de la source de référence dans plus de 95 % des cas pour le chiffre d'affaires, et dans plus de 80 % des cas pour sa ventilation agrégée.

Au niveau individuel, les résultats sont les suivants :

- pour le chiffre d'affaires, la valeur issue des sources fiscales a été retenue pour 97 % des unités, représentant 98 % du chiffre d'affaires total. Les 3 % restants, pour lesquels le chiffre d'affaires issu de l'enquête a été préféré, correspondent généralement à des unités pour lesquelles les données fiscales n'étaient pas disponibles ;

- pour la ventilation agrégée du chiffre d'affaires en « ventes de marchandise », « production de biens » et « production de services », la ventilation détaillée issue de l'enquête, lorsqu'elle est disponible, est choisie, à moins qu'elle ne soit moins détaillée que la ventilation agrégée observée dans les sources fiscales. En conséquence, la structure dérivée de l'enquête a été retenue pour 86 % des unités, représentant 87 % du chiffre d'affaires total.

[18] En valeur absolue.

Pour ce qui est des agrégats, la différence entre les estimateurs de chiffres d'affaires Horvitz-Thompson basés sur les données d'enquêtes et ceux utilisant les données administratives s'élève à 33 milliards d'euros. Bien qu'importante, cette différence ne représentent que 1 % du chiffre d'affaires total, ce qui confirme la cohérence globale entre les deux sources d'information pour cette variable. En ce qui concerne la ventilation agrégée du chiffre d'affaires, on observe des divergences plus importantes entre données d'enquêtes et données fiscales, comme le montre le tableau 2. Par exemple, la production de services représente 18,2 % du chiffre d'affaires dans l'industrie selon les sources fiscales, mais seulement 6,1 % avec les données d'enquête.

Tableau 2 : Estimateurs du total pour le chiffre d'affaires et sa ventilation agrégée (en million d'euros) et structure de ladite ventilation agrégée (en %)

Secteur	Estimateur Horvitz Thompson utilisant les données d'enquête				Secteur	Estimateur Horvitz Thompson utilisant les données fiscales			
	Chiffre d'affaires	Ventes de marchandises	Production de biens	Production de services		Chiffre d'affaires	Ventes de marchandises	Production de biens	Production de services
Industrie agro-alimentaire	147 333	15 127 10,3%	130 355 88,5%	1 851 1,3%	Industrie agro-alimentaire	140 979	20 057 14,2%	113 788 80,7%	7 135 5,1%
Construction	255 177	2 297 0,9%	198 483 77,8%	54 397 21,3%	Construction	247 922	8 675 3,5%	60 097 24,2%	179 374 72,4%
Commerce	1 297 781	1 221 823 94,1%	27 370 2,1%	48 588 3,7%	Commerce	1 282 840	1 130 678 88,1%	75 239 5,9%	77 006 6,0%
Industrie	802 142	114 980 14,3%	638 456 79,6%	48 706 6,1%	Industrie	808 901	119 997 14,8%	541 839 67,0%	147 057 18,2%
Services	659 840	21 591 3,3%	2 246 0,3%	636 004 96,4%	Services	646 305	69 444 10,7%	49 110 7,6%	527 803 81,7%
Transport	175 411	2 098 1,2%	1 131 0,6%	172 182 98,2%	Transport	177 493	3 465 2,0%	448 0,3%	173 576 97,8%
Total	3 337 684	1 377 915 41,3%	998 041 29,9%	961 728 28,8%	Total	3 304 439	1 352 315 40,9%	840 522 25,4%	1 111 951 33,7%

Secteur	Estimateur Horvitz Thompson utilisant les données Redi			
	Chiffre d'affaires	Ventes de marchandises	Production de biens	Production de services
Industrie agro-alimentaire	141 217	13 439 9,5%	126 087 89,3%	1 700 1,2%
Construction	250 834	2 153 0,9%	246 635 98,3%	2 076 0,8%
Commerce	1 289 953	1 200 038 93,0%	29 463 2,3%	60 452 4,7%
Industrie	807 933	107 915 13,4%	672 624 83,3%	27 393 3,4%
Services	649 826	23 357 3,6%	2 940 0,5%	623 529 96,0%
Transport	176 027	2 536 1,4%	1 209 0,7%	172 288 97,9%
Total	3 315 789	1 349 437 40,7%	1 078 957 32,5%	887 439 26,8%

Source : données Esane 2010

L'impact de Redi sur les agrégats est cohérent avec le choix de la source de référence et la procédure de vérification sélective détaillés au §1.2 : pour chaque secteur, l'estimateur du chiffre d'affaires total utilisant les données Redi est proche de l'estimateur fondé sur les données fiscales, tandis que la structure de sa ventilation agrégée est relativement similaire à celle observée dans l'enquête.

Pour conclure, notons l'efficacité, en termes de « sélectivité », de Redi en tant que procédure de vérification sélective des données. En effet, pour la campagne Esane 2010, seules 4 % des unités de l'échantillon – soit environ 6 300 unités – présentent des incohérences sérieuses entre les données d'enquêtes et les données fiscales à l'issue des contrôles de Redi, mais ces unités expliquent plus de 40 % de l'écart absolu observé sur le chiffre d'affaires…

2.2. Impact de la nouvelle procédure d'estimation

Dans cette section, nous évaluons l'impact des changements méthodologiques mis en œuvre dans Esane en termes d'estimation, à savoir l'utilisation de techniques de calage et d'estimateurs spécifiques. Pour ce faire, nous avons calculé, à partir des donnés de la campagne Esane 2010, des estimateurs selon la méthode mise en œuvre dans l'ancien système[19], à savoir des estimateurs de type Horvitz-Thompson avec des poids finaux prenant en compte la correction de la non-réponse totale par repondération et l'étape de winsorisation, mais pas le calage sur données administratives. Nous avons ensuite évalué la précision de ces estimateurs, ainsi que celle des estimateurs du système Esane, sur un petit nombre de variables clefs du dispositif : nombre d'entreprises, chiffre d'affaires et sa ventilation agrégée – ventes de marchandise, production de biens et production de services –, salaires, valeur ajoutée et excédent brut d'exploitation.
Les calculs de précision réalisés prennent en compte, pour les estimateurs relatifs à l'ancien système, l'erreur d'échantillonnage de l'enquête – liée au plan de sondage stratifié et à la correction de la non-réponse totale par repondération selon la méthode des groupes de réponse homogènes – ainsi que l'impact de la procédure de winsorisation. Pour les estimateurs du système Esane, sont également prises en compte la procédure de calage et l'utilisation d'estimateurs spécifiques.

Le tableau 3 donne les résultats de la comparaison des coefficients de variation ainsi calculés pour les six grands secteurs couverts par le dispositif Esane – en vert gras, estimations dont la précision est accrue avec les nouveaux estimateurs, en rouge italique, estimations dont la précision est dégradée avec les nouveaux estimateurs.

Tableau 3 : comparaison des coefficients de variation du système Esane et de l'ancien système

Secteur	CV des estimations relatives à l'ancien système							
	Nombre d'entreprises	Chiffre d'affaires	Ventes de marchandises	Production de biens	Production de services	Salaires	Valeur ajoutée	EBE
IAA	3,9%	0,5%	1,2%	0,5%	0,8%	6,7%	0,8%	0,5%
Construction	0,9%	1,1%	9,1%	1,1%	5,7%	1,5%	1,4%	3,1%
Commerce	1,1%	0,4%	0,4%	2,3%	0,9%	1,1%	0,6%	1,2%
Industrie	1,3%	0,1%	0,3%	0,1%	0,2%	2,5%	0,1%	0,2%
Services	0,5%	0,4%	1,3%	3,5%	0,4%	1,2%	0,5%	1,0%
Transport	2,1%	0,4%	4,3%	1,0%	0,4%	3,6%	0,5%	0,6%
Total	0,40%	0,20%	0,37%	0,28%	0,32%	0,67%	0,26%	0,52%
Sector	**CV des estimations relatives au système Esane**							
	Nombre d'entreprises	Chiffre d'affaires	Ventes de marchandises	Production de biens	Production de services	Salaires	Valeur ajoutée	EBE
IAA	**3,2%**	**0,3%**	*2,1%*	**0,3%**	*13,3%*	**3,3%**	**0,4%**	**0,3%**
Construction	**0,3%**	**0,6%**	*25,6%*	**0,8%**	*51,4%*	**0,4%**	**1,2%**	*3,3%*
Commerce	**0,5%**	**0,1%**	**0,1%**	*4,1%*	*1,5%*	**0,4%**	**0,4%**	**1,0%**
Industrie	**1,3%**	**0,1%**	*0,3%*	**0,1%**	*1,0%*	**1,4%**	**0,1%**	**0,2%**
Services	**0,3%**	**0,2%**	*4,5%*	*14,7%*	**0,3%**	**0,2%**	**0,3%**	**0,7%**
Transport	**0,6%**	**0,2%**	*5,9%*	*2,3%*	**0,1%**	**1,2%**	**0,1%**	**0,2%**
Total	**0,1%**	**0,0%**	**0,1%**	**0,2%**	**0,2%**	**0,0%**	**0,1%**	**0,4%**

Source : données Esane 2010

[19] Plus précisément, dans l'ancien dispositif, deux systèmes coexistaient en parallèle : d'une part, une exploitation de données administratives, et d'autre part, des enquêtes statistiques indépendantes – les Enquêtes Annuelles auprès des Entreprises – au sein desquelles étaient collectées de nombreuses informations également disponibles dans les sources administratives. On se concentre ici sur le système basé sur les seules enquêtes.

Comme on peut le constater, l'utilisation conjointe de techniques de calage et d'estimateurs spécifiques – estimateurs par différence à ce niveau de nomenclature – conduit en général à une amélioration de la précision des estimations sectorielles, à l'exception notable des estimations relatives à la ventilation agrégée du chiffre d'affaires. Cette première impression est confirmée par la comparaison des coefficients de variation pour les estimations sectorielles relatives au niveau groupe de la nomenclature présentée dans le tableau 4. En effet, toujours exception faite de la ventilation agrégée du chiffre d'affaires, les nouveaux estimateurs par différence conduisent systématiquement à une diminution moyenne des coefficients de variation, de 7 % à 35 %, et améliore la précision des estimations pour plus de 80 % des groupes. Qui plus est, on n'observe une diminution notable de la qualité des estimations que dans 10% des cas.

Tableau 4 : ratios « CV Esane » / « CV ancien système », niveau groupe

	Nombre d'entreprises	Chiffre d'affaires	Ventes de marchandises	Production de biens	Production de services	Salaires	Valeur ajoutée	EBE
Moyenne	0,78	0,68	3,64	2,82	9,13	0,65	0,78	0,93
Max	2,20	2,26	130,01	98,94	325,05	3,36	6,34	10,09
Q99	1,89	1,97	35,68	47,93	104,51	3,04	5,13	9,93
Q95	1,02	1,02	11,24	9,01	45,11	1,37	1,80	2,65
Q90	0,99	0,97	6,36	4,43	18,59	1,07	1,12	1,40
Q75	0,93	0,89	2,54	1,40	5,23	0,89	0,92	0,98
Médiane	0,81	0,73	1,34	0,86	1,00	0,68	0,72	0,71
Q25	0,63	0,47	0,78	0,66	0,73	0,41	0,46	0,40
Q10	0,46	0,24	0,40	0,32	0,52	0,20	0,19	0,20
Q5	0,37	0,11	0,30	0,15	0,37	0,13	0,08	0,09
Q1	0,17	0,00	0,00	0,09	0,23	0,03	0,00	0,00
Min	0,16	0,00	0,00	0,07	0,16	0,02	0,00	0,00

Source : données Esane 2010

On observe le même type de résultats pour des estimations au niveau le plus fin de la nomenclature – niveau sous-classe –, même si dans ce cas l'amélioration de la précision induite par la nouvelle procédure d'estimation est moins fréquente et nettement moins importante (cf. tableau 5). Cette moindre performance des estimateurs Esane de niveau infra-groupe – estimateurs par « ventilation » – s'explique par le fait que ces derniers mobilisent de manière moins intensive les informations administratives de niveau fin que les estimateurs par différence utilisés aux niveaux groupe et supérieurs.

Tableau 5 : ratios « CV Esane » / « CV ancien système », niveau sous-classe

	Nombre d'entreprises	Chiffre d'affaires	Ventes de marchandises	Production de biens	Production de services	Salaires	Valeur ajoutée	EBE
Moyenne	0,94	0,91	5,34	4,98	10,63	1,00	1,02	1,18
Max	17,08	2,25	352,95	197,15	918,85	9,75	14,20	17,63
Q99	1,21	1,82	73,30	103,03	140,78	3,48	4,39	6,08
Q95	1,06	1,11	18,12	17,61	34,31	1,50	1,59	2,51
Q90	1,02	1,04	6,24	6,21	15,49	1,16	1,20	1,67
Q75	1,00	1,00	1,98	1,71	3,35	1,04	1,03	1,09
Médiane	0,98	0,97	1,11	1,00	1,20	0,98	0,97	0,96
Q25	0,89	0,84	0,99	0,95	0,98	0,85	0,85	0,84
Q10	0,74	0,61	0,90	0,74	0,78	0,60	0,58	0,59
Q5	0,60	0,39	0,65	0,52	0,62	0,42	0,45	0,41
Q1	0,19	0,08	0,23	0,19	0,32	0,03	0,10	0,08
Min	0,00	0,00	0,00	0,07	0,14	0,00	0,00	0,00

Source : données Esane 2010

Reste le cas de la ventilation agrégée du chiffre d'affaires, pour laquelle la nouvelle procédure d'estimation mise en œuvre dans Esane conduit généralement à une détérioration, parfois importante, de la qualité des estimations sectorielles, et ce quel que soit le niveau d'agrégation considéré. Ceci traduit le fait que l'estimateur par différence – qui constitue la base de la procédure d'estimation dans Esane – est inapproprié pour l'estimation de ces variables : en effet, ces variables sont fortement affectées par le processus Redi et leurs valeurs finales post-Redi Y_i^{Redi} peuvent fréquemment différer des valeurs issues des sources fiscales Y_i^{fiscal}, puisque l'enquête constitue la source de référence du processus Redi pour ces variables. Dès lors, l'hypothèse $Y_i^{Redi} \mathbb{1}_{APE\_enq=X}(i) \approx Y_i^{fiscal} \mathbb{1}_{APE\_rep=X}(i)$, qui est vérifiée pour la très grande majorité des variables du dispositif Esane et qui justifie l'utilisation d'estimateurs par différence, ne tient pas pour ces variables spécifiques. Des études méthodologiques complémentaires restent à mener afin de définir quel estimateur alternatif pourrait être utilisé pour ces variables[20].

Conclusion

Après trois années de fonctionnement, la première évaluation du nouveau système de production des statistiques structurelles d'entreprises françaises tend à valider la majorité des choix méthodologiques effectués. La comparaison des différentes sources d'information permet la mise en place de contrôles de cohérence sur quelques variables clefs qui améliorent la qualité des estimations en réduisant les biais dus aux erreurs de réponse. En ce qui concerne les nouveaux estimateurs statistiques, ils conduisent à une amélioration de la précision des estimations, à la fois au niveau agrégé et dans la très grande majorité des estimations sectorielles de niveau groupe, sauf pour les variables de ventilation agrégée du chiffre d'affaires et les achats et variations de stock liés à celles-ci. Pour ces variables spécifiques, des études méthodologiques complémentaires restent à mener afin de déterminer le meilleur estimateur à utiliser.

Bibliographie

Bauer P., Brilhault G. et Gros E. (2009). Le plan de sondage de l'ESA (enquête sectorielle annuelle du futur dispositif Esane), article présenté aux dixièmes Journées de Méthodologie Statistique, Paris, France.

Brion Ph. (2011). Esane, le dispositif rénové de production des statistiques structurelles d'entreprises, Courrier des statistiques n°130, Insee, Paris, France.

Deville J.-C. et Särndal C.-E. (1992). Calibration estimators in survey sampling, Journal of the American Statistical Association, 87, pp. 376-382

Haag O. (2009). Reengineering French structural business statistics : redesign of the annual survey, article présenté à la conférence Q2010, Helsinki, Finlande.

Haag O. (2012). Esane : À la recherche d'une cohérence maximale des données multi-sources sur les entreprises par le biais de micro et macro contrôles, article présenté aux onzièmes Journées de Méthodologie Statistique, Paris, France.

[20] Ainsi que pour les variables d'achats et de variations de stock qui leur sont liées et pour lesquelles les estimateurs par différence sont également inadaptés.