L'UTILISATION COMBINÉE DE DONNÉES D'ENQUÊTE ET DE DONNÉES ADMINISTRATIVES POUR LA PRODUCTION DES STATISTIQUES STRUCTURELLES D'ENTREPRISES

Philippe BRION (*)

(*) Insee, Direction des statistiques d'entreprises

Introduction

L'Insee met en place à l'heure actuelle un nouveau dispositif de production des statistiques structurelles d'entreprises, intitulé ESANE (Élaboration des Statistiques Annuelles d'Entreprises). Ce dispositif est destiné à utiliser au maximum les données administratives disponibles sur les entreprises :

- déclarations annuelles sur les bénéfices adressées par les entreprises à la Direction Générale des Impôts (il faut noter que ces déclarations peuvent être utilisées directement car les informations comptables demandées par l'administration fiscale française font référence au Plan Comptable Général français, tout comme les variables statistiques) ;
- déclarations annuelles de données sociales (DADS), contenant des données sur les effectifs employés et les rémunérations, établies pour le compte des organismes de protection sociale;
- déclarations douanières.

Il est relativement facile d'utiliser, en France, les données administratives concernant les entreprises en raison du rôle de répertoire inter-administratif joué par SIRENE : les différentes administrations transmettant les données mentionnées ci-dessus utilisent l'identifiant de SIRENE. De plus, à l'exception de certaines grandes unités pour lesquelles un profilage est pratiqué (à savoir la définition d'unités de collecte spécifiques), il n'y a pas de problème d'unité spécifique à chaque administration : c'est l'unité légale telle que définie dans SIRENE qui est la référence.

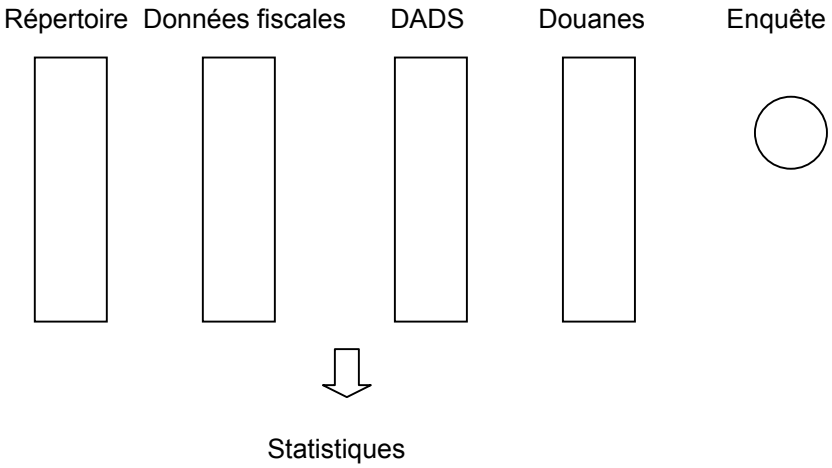
L'utilisation conjointe des trois sources administratives (fiscale, « sociale » et douanière) n'est cependant pas suffisante pour répondre à l'ensemble des besoins exprimés en matière de statistiques structurelles d'entreprises. En particulier, les comptables nationaux souhaitent disposer d'une ventilation du chiffre d'affaires selon les différentes activités de l'entreprise, cette ventilation servant ensuite à établir les comptes de branche. Or cette information n'est pas disponible dans la source fiscale : elle doit donc faire l'objet d'une interrogation directe des entreprises, par l'intermédiaire d'une enquête statistique. Cette enquête statistique (dite ESA, enquête sectorielle annuelle, ou EAP, enquête annuelle de production, sur le seul secteur de l'industrie) sera réalisée sur un échantillon d'entreprises. Elle servira également à obtenir des informations non disponibles dans les sources administratives et souhaitées pour produire les statistiques structurelles d'entreprises, sur certains types de dépenses, ou encore sur des sujets sectoriels, comme par exemple la superficie des magasins de vente au détail dans le cas du commerce.

Il faut noter également que l'un des produits les plus importants des statistiques structurelles d'entreprises est la publication de résultats sectoriels : ceux-ci s'appuient sur le classement sectoriel des entreprises, classement qui se réfère à la NAF (nomenclature d'activités française). Si l'on dispose d'un code d'activité principale (dit code APE) pour chaque entreprise dans le répertoire d'entreprises SIRENE, on ne peut cependant pas l'utiliser directement pour produire les statistiques d'entreprises : ce code peut avoir été déterminé quelques années auparavant (par exemple à partir de la déclaration de l'entreprise, au moment où celle-ci s'est enregistrée dans SIRENE) et ne pas avoir été mis à jour depuis.

Or la ventilation du chiffre d'affaires des entreprises enquêtées par l'enquête statistique du dispositif ESANE permet de déterminer leur activité principale selon une approche économique de leurs activités (qui passe par l'application d'un algorithme), et non plus selon une approche déclarative. C'est ce classement sectoriel obtenu par l'enquête, et non celui de SIRENE, qui doit être utilisé pour produire les statistiques. Le problème est qu'on va devoir ensuite utiliser conjointement des données obtenues sur un échantillon (en particulier ce classement sectoriel), et des données administratives exhaustives.

La problématique développée dans ce papier est donc celle de la meilleure utilisation possible de l'information disponible (données administratives, enquête) pour produire les statistiques structurelles d'entreprises (figure 1). Plus précisément, on se limitera aux problèmes statistiques posés par la partie échantillonnée.

Figure 1 : le dispositif ESANE



1. Comment produire des statistiques à partir d'un matériau composite ?

Le matériau qu'on va récupérer peut être représenté sous la forme d'une base de données rectangulaire incomplète :
- une base de données complète pour les variables administratives (aux données manquantes près), sur le champ, soit plus de deux millions d'entreprises ;
- une base de données limitée à l'échantillon pour les données de l'enquête : si les grandes entreprises sont enquêtées de manière exhaustive, les petites et moyennes entreprises seront sondées, et la taille de l'échantillon est de l'ordre de 120 000, soit environ 5% de la population totale.

Il n'est pas immédiat d'utiliser une telle base de données. Une méthode envisageable consiste à « boucher les trous » pour arriver à un rectangle complet : on l'appellera imputation de masse (en effet on impute les valeurs des variables de l'enquête pour environ 95% des unités). Ce n'est pas la méthode qui a été choisie, en raison des questions soulevées au paragraphe suivant ; des estimateurs statistiques combinés, plus complexes, doivent être mis en place. Ils sont présentés dans la partie 1.2.

1.1. Les problèmes posés par l'imputation de masse

L'imputation de masse consiste à proposer des valeurs, pour chaque entreprise non échantillonnée dans l'ESA, pour chacune des variables du questionnaire de l'ESA.

Pour cela, il faut s'appuyer sur des caractéristiques disponibles dans le répertoire ou dans les données administratives obtenues, en y cherchant les variables qui sont les plus liées à celles qu'on veut imputer. Un cas particulier important est celui du code d'activité principale APE : l'idée est d'observer sur l'échantillon la matrice de passage entre la valeur du code dans le répertoire SIRENE et celle obtenue à l'enquête, et d'appliquer des probabilités de transition aux unités non échantillonnées pour déterminer la valeur du code à la période de référence (donc celle de l'enquête).

On peut montrer (voir [1] pour plus de détails) que ceci conduit non seulement à générer de la variance due à l'imputation, mais également des biais pour les statistiques sectorielles (du fait qu'il faudrait imputer le code APE au moment de l'enquête conditionnellement à un certain nombre de caractéristiques, par exemple relatives à la taille, ce qui est impossible en pratique). Ces biais ont été évalués, et, par exemple, sur un secteur économique comme celui du commerce, sur 119 secteurs élémentaires (à savoir au niveau 700 de la NAF révision 1), 15 présentaient un biais dû à ce type d'imputation supérieur à 10% de la valeur de la grandeur pour une valeur à estimer comme le chiffre d'affaires total du secteur élémentaire. La perspective d'utiliser la méthode d'imputation de masse a été abandonnée. Il faut mentionner que d'autres inconvénients liés à cette méthode existent, comme la possibilité, par exemple, qu'un chercheur mène des travaux sur un morceau de fichier uniquement composé de données imputées ; dans [2], on trouve également des éléments sur les conséquences potentielles de l'imputation de masse sur les relations entre variables.

1.2. Les estimateurs combinés

1.2.1. La modification des poids de sondage de l'enquête à l'aide des données administratives

On peut, au départ, partir des estimateurs basiques résultant du plan de sondage. Ils sont, pour l'estimation de totaux, du type : $\sum_s wiX_i$, où wi est le poids de sondage associé à l'unité i.

Le fait de disposer de données administratives exhaustives permet d'améliorer la précision de certains estimateurs grâce à l'utilisation de techniques de calibrage [3] ; plus précisément, on ajuste les poids de façon que l'échantillon extrapolé « retrouve » des données connues par la source administrative.

En pratique, la variable « chiffre d'affaires » semble une variable importante pour le calage (plus précisément, un des résultats importants des statistiques structurelles d'entreprises est le chiffre d'affaires par secteur) ; on va donc ajuster les poids d'échantillonnage de telle façon que le chiffre d'affaires d'un secteur X, tel que défini par le répertoire (et non le secteur réel au moment de l'enquête, impossible à délimiter dans le répertoire en raison des changements de secteur), soit retrouvé par l'échantillon extrapolé :

$\sum_s wi\, CA(i)\ 1_{APErep=X}(i) = \sum_U CA(i) 1_{APErep=X}(i)$, APErep étant le code APE connu dans le répertoire et CA(i) le chiffre d'affaires.

Dans la suite du papier, on conserve la notation w_i pour les poids résultant de la phase de calage, pour ne pas alourdir les formules (on aurait pu utiliser une notation w_i', par exemple).

Lors du calage, les poids sont modifiés. Dans la pratique, on évite que les coefficients multiplicateurs des poids ne soient trop dispersés, et en particulier ne prennent des valeurs trop importantes On

trouve dans [4] les résultats des études menées sur ce sujet, résultats qui montrent qu'il est possible de procéder à un tel calage au niveau 3 caractères de la NAF (mais pas de manière plus fine).

1.2.2. Le cas particulier des statistiques sectorielles

Les statistiques sectorielles sont produites à partir du code APE déterminé par l'ESA (au travers de la ventilation du chiffre d'affaires), et non à partir du code APE connu dans le répertoire (c'est-à-dire au lancement de l'enquête). Plus précisément, une grandeur sectorielle comme le chiffre d'affaires du secteur X s'écrit :

$$\sum_{U} CA(i)\ 1_{APEenq=X}(i)$$

Le problème est qu'on ne connaît la valeur du code APE « au moment de l'enquête » que pour les unités échantillonnées.

Le chiffre d'affaires d'un secteur X donné peut donc être estimé, dans un premier temps, par l'estimateur « basique » :

$\sum_{s} wi\, CA(i)\ 1_{APEenq=X}(i)$, où APEenq est la valeur du code APE résultant de l'enquête, et les w_i résultant du calage précédent.

La bonne corrélation entre les deux variables $CA(i)1_{APEenq}(i)$ et $CA(i)1_{APErep}(i)$ assure l'efficacité (au sens erreur quadratique moyenne) de l'estimateur ci-dessus au niveau où on a calé. En revanche, il n'est pas certain que sur des sous-domaines (en l'occurrence des secteurs définis à niveau plus fin que trois caractères de la NAF), l'estimateur par calage généralisé soit efficace.

Mais l'existence de deux codes APE, celui ex ante du répertoire (connu de façon exhaustive) et celui venant de l'enquête statistique (disponible uniquement sur échantillon), conduit à proposer d'utiliser un estimateur par différence, pour un secteur donné X :

$$\sum_{U} CA(i)1_{APErep=X}(i) + \sum_{s} wi\, CA(i)(1_{APEenq=X}(i) - 1_{APErep=X}(i))$$

Quatre remarques peuvent être formulées à propos de cet estimateur :

- il est évidemment égal à l'estimateur $\sum_{s} wi\, CA(i)\ 1_{APEenq=X}(i)$ pour le niveau où l'on a calé (trois caractères de la NAF) ;
- il est a priori plus efficace que l'estimateur $\sum_{s} wi\, CA(i)\ 1_{APEenq=X}(i)$ pour un secteur fin, car les résidus de la régression de la variable $CA(i)(1_{APEenq=X}(i) - 1_{APErep=X}(i))$ sur les variables de calage $CA(i)\ 1_{APErep=X}(i)$ sont plus petits que ceux de la variable $CA(i)1_{APEenq=X}(i)$ sur ces mêmes variables de calage ;
- il peut paraître « instable » : en fait, il n'est pas plus instable que l'estimateur qui était utilisé pour estimer le chiffre d'affaires d'un secteur à partir de l'EAE, qui était $\sum_{s} wi\, CA(i)\ 1_{APEenq=X}(i)$: c'est la statistique sectorielle qui, par essence même, est instable ;

- il est utilisable pour toute variable administrative, comme par exemple l'investissement ; en revanche, la remarque précédente (égalité de l'estimateur « basique » et de l'estimateur par différence au niveau où on a calé) n'est plus valable pour une variable qui n'a pas été utilisée pour le calage (c'est l'estimateur par différence qui est le plus efficace à ce niveau).

L'information apportée par l'enquête statistique est donc utilisée ici pour opérer un découpage sectoriel des données administratives exhaustives, et l'estimateur par différence permet de fournir des statistiques sectorielles plus précises. A priori, il ne semble pas exister d'autre variable catégorielle que le code APE bénéficiant du même statut, à savoir l'existence de deux valeurs proches, l'une exhaustive mais imparfaite, l'autre obtenue par l'enquête et donc considérée comme la valeur de référence.

1.2.3. D'autres types de statistiques

Pour les statistiques autres que celles issues de variables venant de sources administratives et relatives à un secteur, les estimateurs ne seront pas nécessairement du même type que celui proposé pour les statistiques sectorielles.

Pour certaines variables disponibles uniquement dans l'enquête statistique, on s'appuiera sur l'estimateur pondéré classique $\sum_s wiX_i$.

En revanche, le problème se complique quand on veut estimer des « sous-statistiques sectorielles ». Le passage secteur branches en est un cas particulier important.

Passage secteurs-branches

Ce qu'il faut estimer est le chiffre d'affaires total de la branche b pour les entreprises du secteur X.

Ce qui est proposé ici est de s'appuyer sur le chiffre d'affaires total du secteur X obtenu par l'estimateur par différence présenté ci-dessus, et d'utiliser l'ESA comme clé de ventilation entre branches, au travers de la quantité :

$$\frac{\sum_s wi\, CA(b,i) 1_{APEenq=X}(i)}{\sum_s wi\, CA(i) 1_{APEenq=X}(i)}$$, CA(b,i) étant le CA de la branche b de l'entreprise i.

La valeur du CA de la branche b du secteur X est donc différente de celle qui serait obtenue par l'estimateur $\sum_s wiCA(b,i)1_{APEenq=x}$ (sauf au niveau où on a calé).

Cette méthode peut être appliquée à toute ventilation d'une estimation sectorielle (obtenue sur une variable comptable) : par exemple si l'on veut estimer l'investissement total, au sein d'un secteur donné du commerce, selon le type de fournisseurs de l'entreprise (cette dernière information étant apportée par l'enquête statistique).

1.2.4. Le gain de précision apporté

Les résultats des études quantifiées présentés dans [4] montrent que le dispositif proposé permettra un gain de précision net pour l'estimation d'un certain nombre de statistiques. Il faut noter que ces calculs prennent en compte la variance d'échantillonnage uniquement ; dans la pratique, viendront s'ajouter les effets dus aux problèmes de collecte (erreurs de mesure), et également aux données manquantes dans les sources administratives. En fonction des résultats obtenus, une diminution de l'échantillon a été décidée, sous la forme d'une division par deux de la partie échantillonnée (voir [4] pour plus de détails).

On ne peut cependant dire que le nouveau dispositif permettra de gagner sur tous les tableaux : si, pour les variables s'appuyant sur les sources administratives, on aura certainement une précision meilleure, pour certaines variables obtenues par l'enquête statistique, les estimations pourront être moins précises qu'avec l'ancien dispositif en raison de la réduction de la taille de l'échantillon d'enquête.

2. Deux questions posées par le dispositif : la prise en compte des non réponses et la cohérence entre données d'enquête et données administratives

Les estimateurs proposés dans la partie précédente ne peuvent cependant être utilisés tels quels :

- il existe de la non réponse à l'enquête statistique ;
- la source administrative n'est pas parfaite (de même que les données issues de l'enquête) ; il existe une phase de confrontation des deux types de sources, destinée à produire des statistiques bénéficiant le plus possible des potentialités que ces deux types de sources offrent.

2.1. La prise en compte des non réponses

Une fois que l'on dispose d'un fichier définitif des données administratives et d'un état définitif des unités de l'échantillon considérées comme en non réponse totale, on devra prendre en compte cette dernière au travers d'une modification des poids de sondage.

Le problème des enquêtes d'entreprises réalisées par courrier est l'incertitude, en cas de non retour du questionnaire, entre une situation de « vraie » non réponse et une situation de non réponse pour cause de cessation. On ne dispose pas, dans la base de sondage, d'un indicateur d'activité relatif à chaque unité. L'idée est alors d'utiliser une méthode analogue à celle décrite dans [5], à savoir utiliser au maximum l'information disponible dans des sources autres que l'enquête statistique (par exemple les données TVA sur la présomption d'activité) pour définir des catégories d'entreprises non répondantes : celles dont on sait qu'elles sont actives, celles dont on sait qu'elles sont cessées, et celles pour lesquelles on ne sait rien. In fine, les poids obtenus « estiment » le contour recherché des unités actives. En revanche, il faudra revenir ici sur la question du niveau de la nomenclature sur lequel on peut caler, et sur l'utilisation de variables autres que le chiffre d'affaires sectoriel pour caler (en particulier le nombre d'entreprises « sectoriel » de la base de sondage, ou le nombre d'entreprises présumées actives - par exemple au travers d'une source du type DGI -, toujours au niveau d'un secteur dans la base de sondage). Ce point (choix des variables, niveau de la nomenclature) est à étudier lors de la phase de repondération, et, le cas échéant, on pourrait être conduit à opérer un calage à un niveau supérieur de la nomenclature.

2.2. La réconciliation entre données d'enquête et données administratives

Un travail de réconciliation des données individuelles, entre variables d'enquête et variables administratives, est opéré dans le processus ESANE. On n'entrera pas ici dans les détails de cette phase (et en particulier pour la phase de traitement des sources administratives préalable à cette phase de réconciliation), mais il faut noter qu'à l'issue de ce travail, certaines valeurs des variables administratives, ou de l'enquête statistique, peuvent être considérées comme fausses (chiffre

d'affaires, autre variable fiscale, ventilation en branches, etc.). Elles peuvent, in fine, être modifiées, automatiquement ou suite à l'intervention d'un gestionnaire, pour être rendues cohérentes entre elles.

Il est proposé d'utiliser ici encore un estimateur par différence, mais en partant cette fois des données fiscales avant la phase de réconciliation (par exemple le CA fiscal) et en utilisant l'échantillon pour inférer. De cette façon, l'enquête statistique fournit un « contrôle qualité », par sondage, des données administratives, ainsi que de ses propres données (le raisonnement est présenté ici sur le CA mais peut être étendu à toute variable administrative), contrôle qualité qui est pris en compte dans les estimations finales. Une dégradation, ou une amélioration, de la qualité de la source administrative ne devraient donc pas avoir de conséquences sur les statistiques produites.

Précisément, ce dont on dispose dans la source administrative vaut, pour le CA par exemple :

$$\sum_U CAfiscal(i) 1_{APErep=X}(i)$$

Ce qu'on cherche à estimer vaut :

$\sum_U CAvrai(i) 1_{active}(i) 1_{APEenq=X}(i)$, CAvrai étant la valeur arbitrée après la phase de réconciliation (donc valeur de l'enquête statistique, valeur issue de la source fiscale, ou tierce valeur).

L'estimateur final proposé est alors un estimateur par différence s'appuyant sur l'échantillon (pondéré par les nouveaux poids obtenus à l'étape précédente) :

$$\sum_U CAfiscal(i) 1_{APErep=X}(i) + \sum_s wi\,(CAvrai(i) 1_{active}(i) 1_{APEenq=X}(i) - CAfiscal(i) 1_{APErep=X}(i)).$$

La variable 1_{active} est indiquée ici car elle est importante dans le cadre de l'estimation du nombre d'entreprises.

La formule présentée ci-dessus pour le chiffre d'affaires s'applique à n'importe quelle variable administrative qui aura été éventuellement corrigée lors de la phase de réconciliation.

Conclusion

Le dispositif proposé permet une utilisation complète des informations disponibles dans les différentes sources (enquête, données administratives).

Par rapport au dispositif précédent (voir par exemple [1] pour une description de ce dispositif), de gros progrès ont été réalisés pour unifier les statistiques structurelles d'entreprises : auparavant, on avait en parallèle des enquêtes (EAE, enquêtes annuelles d'entreprise) et une base de données s'appuyant sur la source fiscale (SUSE), qui fournissaient chacune leurs propres résultats. Qui plus est, on avait pour une statistique comme le nombre d'entreprises trois « sources » possibles : les EAE, SUSE et le répertoire SIRENE, chacune ayant ses caractéristiques propres et donc ses limites.

Il faut noter également le rôle prépondérant du répertoire d'entreprises, véritable colonne vertébrale du dispositif. Ce répertoire est utilisé pour tirer les échantillons, procéder à l'opération de calage, mais également pour récupérer les données administratives, au travers de son identifiant unique. L'Insee travaille actuellement sur la mise en place d'un répertoire statistique qui devrait, plus encore qu'aujourd'hui, permettre la mise en relation d'informations provenant de différentes sources, et donc offrir des potentialités supplémentaires par rapport à celles existant à l'heure actuelle.

Bibliographie

[1] Brion Ph., « Redesigning the French structural business statistics, using more administrative data», *Proceedings of the third International Conference on Establishment Surveys*, Montréal, 2007.
[2] Kroese A.H., Renssen R.H., « New applications of old weighting techniques – constructing a consistent set of estimates based on data from different sources », *Proceedings of the second International Conference on Establishment Surveys*, Buffalo, 2000.
[3] Deville J.C., Särndal C.E., « Calibration estimators in survey sampling », *Journal of the American Statistical Association*, 1992, n°87, pp 376-382.
[4] Bauer P., Brilhault G., Gros E., « Le plan de sondage de l'ESA (enquête sectorielle annuelle du futur dispositif Esane) », *Communication aux JMS2009,* Insee.
[5] Brion Ph., Caron N., Pietri-Bessy P., « Redresser la non réponse totale dans les enquêtes auprès des entreprises : les pièges à éviter – illustration avec l'enquête innovation », *Communication aux JMS2005,* Insee.