

LE RÔLE DES ESTIMATEURS DANS LA PRODUCTION DE STATISTIQUES À PARTIR DE DONNÉES ADMINISTRATIVES ET DE DONNÉES D'ENQUÊTES : L'EXEMPLE DES STATISTIQUES STRUCTURELLES D'ENTREPRISES FRANÇAISES

Philippe Brion¹

RÉSUMÉ

L'Insee mène un projet de rénovation des statistiques structurelles d'entreprises françaises s'appuyant sur une utilisation intensive des données administratives (données fiscales, données d'emploi et de salaires, données douanières) complétée par une enquête menée sur un échantillon d'entreprises. Le papier est centré sur la méthode de production d'estimations préconisée pour la partie échantillonnée (en particulier pour les statistiques sectorielles), qui s'appuie sur les deux types de données obtenues (administratives et données d'enquête).

MOTS CLÉS : Données administratives; enquête; estimateurs; statistiques d'entreprises.

ABSTRACT

Insee is renewing French structural business statistics, using administrative data in an intensive way (tax data, data about salaries and wages, customs data) and completing them with a survey conducted on a sample of enterprises. The paper focusses on the method of production of estimates for the sampled part (particularly for sector-based statistics), which relies on the two kinds of data (administrative and survey data).

KEY WORDS: Administrative data, Business statistics, Estimators, Survey.

1. LA RÉNOVATION DES STATISTIQUES STRUCTURELLES D'ENTREPRISES FRANÇAISES

1.1 Une production de statistiques fondée sur l'utilisation de différentes sources

L'Insee (Institut National de la Statistique et des Études Économiques) a démarré en 2005 un projet de rénovation des statistiques structurelles d'entreprises françaises, intitulé RESANE (REfonte des Statistiques ANnuelles d'Entreprises), destiné en particulier à utiliser au maximum les données administratives disponibles (pour plus de détails sur le système qui était en cours jusqu'à présent, voir Brion, 2007).

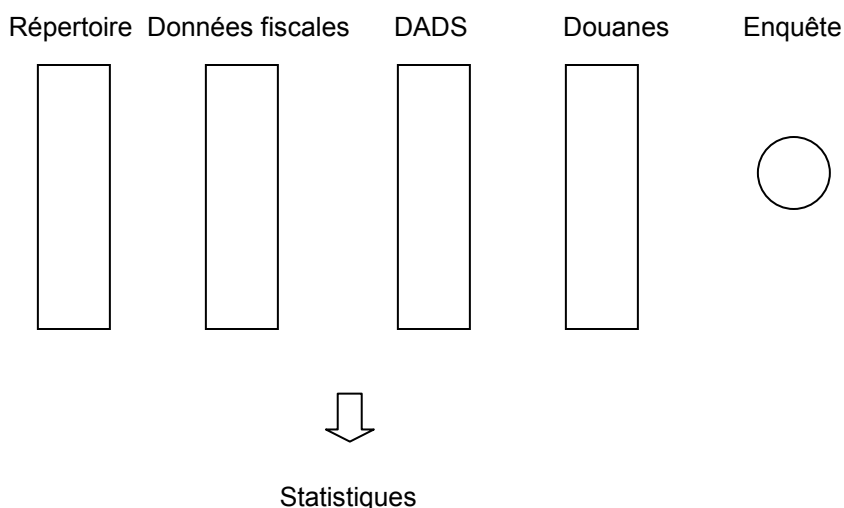
Plus précisément, trois types de données administratives seront utilisées (figure 1) :

- déclarations annuelles sur les bénéficiaires 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.

¹ Philippe Brion, Direction des statistiques d'entreprises, Insee, 18 bd A. Pinard, 75675 Paris cedex 14, France, philippe.brion@insee.fr

Il est relativement facile d'utiliser de manière conjointe ces données en raison du rôle de répertoire inter-administratif joué en France par le répertoire d'entreprises SIRENE géré par l'Insee: en effet, toute entreprise enregistrée dans les fichiers des différentes administrations citées précédemment l'est avec l'identifiant du répertoire (dit n° Siren). De plus, il n'y a pas de problème d'unités statistiques différentes, l'unité légale (telle que définie dans le répertoire) étant la référence pour les différentes administrations qui fournissent les données utilisées.

Figure 1 – Les différentes composantes du système de production des statistiques structurelles d'entreprises



L'utilisation conjointe de ces trois sources administratives (fiscale, « sociale », 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 à établir les comptes de branche. Or cette information n'est disponible dans aucune des sources administratives détaillées précédemment, l'administration fiscale s'intéressant par exemple au total du chiffre d'affaires d'une entreprise mais non à sa décomposition ; elle doit donc faire l'objet d'une interrogation directe des entreprises.

Il existera ainsi, à côté des fichiers administratifs utilisés, une enquête statistique réalisée sur échantillon (contrairement aux données administratives, a priori exhaustives). Cette enquête servira à obtenir les informations non disponibles dans les sources administratives et souhaitées pour produire les statistiques structurelles d'entreprises : outre la ventilation du chiffre d'affaires déjà mentionnée, des questions seront posées sur certains types de dépenses, ou sur des sujets spécifiques à un secteur donné. La ventilation du chiffre d'affaires selon les activités de l'entreprise constitue une des informations principales apportée par l'enquête : outre l'utilisation déjà mentionnée (par les comptables nationaux), elle sert à déterminer le code d'activité principale de l'entreprise au moment de l'enquête, à partir d'un algorithme spécifique.

Le classement de l'entreprise dans un des postes de la nomenclature d'activités française NAF (dérivée de la NACE européenne) qui en résulte est la base des statistiques sectorielles. En effet, on ne peut utiliser la valeur disponible dans le répertoire d'entreprises SIRENE, cette valeur pouvant avoir été établie quelques années auparavant et ne plus être à jour. De plus, au moment où une entreprise se crée, ce code est déterminé selon la déclaration de l'entreprise, et peut être de moins bonne qualité que celui résultant de l'enquête (qui découle d'une approche plus économique au travers de l'étude de la ventilation du chiffre d'affaires). C'est la valeur réelle, au moment de l'enquête, qui sert de base aux statistiques sectorielles.

Plus précisément, les statistiques relatives à un secteur donné X s'écrivent donc comme, pour le chiffre d'affaires CA par exemple :

$$\sum_U CA(i) 1_{APEenq=X}(i),$$

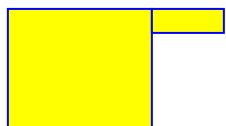
où la variable $1_{APEenq=X}(i)$ est l'indicatrice de l'appartenance au secteur X au moment de l'enquête. C'est sur les données de l'échantillon qu'on devra donc s'appuyer pour les produire.

1.2 Les problèmes posés par l'utilisation conjointe de plusieurs sources

Le système de production envisagé va donc conduire à une base de données rectangulaire incomplète (figure 2) :

- une base de données complète pour les variables administratives (aux données manquantes près), sur un champ constitué de plus de deux millions d'entreprises ;
- une base de données limitée à l'échantillon pour les données d'enquête : si les grandes entreprises seront systématiquement enquêtées, la catégorie des petites et moyennes entreprises sera sondée et la taille de l'échantillon global sera de l'ordre de 150 000 entreprises.

Figure 2 - La base de données obtenue avec les données collectées



Comment utiliser une telle base de données ? Une méthode destinée à produire des statistiques à partir de ce matériau, dite « imputation de masse », consiste à “boucher les trous”, et donc à imputer les données du morceau manquant du rectangle de la figure 2. Il faut noter que cette imputation concernerait alors plus de 90% des entreprises, même si ces dernières sont des petites unités. On peut envisager d'autres types de méthodes, reposant sur des estimateurs statistiques combinés.

Une autre question est posée par l'utilisation de plusieurs sources de données : comment mener le travail de contrôle-redressement des données, celles-ci n'étant pas disponibles aux mêmes périodes ?

2. LES ESTIMATEURS UTILISÉS POUR LA PARTIE ÉCHANTILLONNÉE

Cette partie est consacrée à la partie échantillonnée de la population d'entreprises : la partie exhaustive fait l'objet de traitements plus spécifiques qui ne seront pas abordés ici.

2.1 Deux méthodes utilisables pour produire des estimations statistiques

Disposant d'un matériau composé de données administratives exhaustives et de données d'enquête recueillies sur échantillon, deux méthodes peuvent être proposées pour produire des statistiques, comme cela a été indiqué précédemment : l'imputation de masse, et l'utilisation d'estimateurs statistiques combinés (détaillés dans les parties suivantes).

L'imputation de masse semble a priori séduisante : disposer d'une base de données rectangulaire (comparée à la base de données incomplète présentée dans la figure 2) permet de produire des estimations statistiques de manière simple. Cependant, les propriétés statistiques des estimations produites à partir de cette méthode posent des problèmes de variance, mais aussi de biais. Les comparaisons chiffrées réalisées sur l'efficacité comparée des deux méthodes ont conduit à préférer la méthode s'appuyant sur les estimateurs combinés (Brion, 2007).

2.2 Les estimateurs statistiques combinés, principes

De manière générale, les estimateurs produits à partir des variables collectées sur l'échantillon de l'enquête (pour une variable Y , par exemple) s'écrivent comme :

$$\sum_S w_i Y_i.$$

La disponibilité de données administratives permet d'améliorer la précision de ces estimateurs, grâce à l'utilisation de techniques de calibrage (Deville, Särndal, 1992). Pour les statistiques structurelles d'entreprises, la variable chiffre d'affaires est considérée comme importante, et liée à beaucoup d'autres variables : elle est donc utilisée pour déterminer les équations de calibrage. Plus précisément, on cherche à ce que l'échantillon extrapolé « retrouve » le chiffre d'affaires de certains secteurs tel que connu par les données administratives :

$$\sum_S w_i CA(i) 1_{APErep=X}(i) = \sum_U CA(i) 1_{APErep=X}(i),$$

où on utilise cette fois un classement sectoriel issu du répertoire $1_{APErep=X}(i)$, seul disponible sur l'ensemble des unités.

La question qui se pose à ce stade est celle du niveau de la nomenclature auquel procéder sans que les poids initiaux ne soient trop déformés. Les études menées sur des données d'enquêtes passées ont montré qu'il devrait se situer au niveau trois caractères de la nomenclature.

On utilisera encore, par la suite, la notation w_i pour le poids issu du calage (et non w'_i , par exemple), afin de ne pas surcharger les notations.

Pour les statistiques sectorielles cependant, on peut utiliser de façon plus poussée l'existence des deux informations (classement sectoriel connu dans le répertoire $1_{APErep=X}(i)$, et classement sectoriel « réel » connu au moment de l'enquête pour les unités de l'échantillon $1_{APEenq=X}(i)$), et proposer un estimateur par différence.

Pour toute variable administrative Y (le chiffre d'affaires, mais aussi le total des investissements, l'effectif employé, etc.), le total sectoriel peut être estimé par l'estimateur par différence suivant :

$$\sum_U Y(i) 1_{APErep=X}(i) + \sum_S w_i Y(i) (1_{APEenq=X}(i) - 1_{APErep=X}(i)),$$

l'échantillon servant à estimer le biais qu'on aurait en utilisant de manière directe les données exhaustives disponibles (administratives, plus classement sectoriel tel que connu dans le répertoire).

Pour certaines statistiques mixant données échantillonnées et données d'enquêtes, on peut être conduit à utiliser l'enquête comme clé de ventilation appliquée aux statistiques sectorielles précédentes : c'est par exemple le cas du passage secteur branche, où on estime le chiffre d'affaires relatif à chaque branche au sein d'un secteur donné ; pour ce faire, on part du chiffre d'affaires estimé pour le secteur, qu'on ventile par branche grâce aux données de l'enquête.

2.3 Le contrôle-redressement des données

Les différents types de données collectées ne seront pas disponibles au même moment. Les questionnaires de l'enquête relative aux résultats de l'année n seront envoyés au début de l'année $n+1$, et les retours seront échelonnés sur une période plus ou moins longue. Les fichiers administratifs seront, eux, disponibles de façon globale : par exemple, en octobre $n+1$ pour les fichiers fiscaux, avec en plus une première livraison en juin ou juillet comprenant les entreprises traitées en priorité par la Direction Générale des Impôts.

Le travail de contrôle et redressement des données est alors plus complexe que dans le cas d'une seule enquête, pour plusieurs raisons.

Lors de l'arrivée des premiers questionnaires de l'enquête, on ne s'appuiera plus sur les variables disponibles dans les fichiers administratifs pour procéder à des contrôles (alors que c'était le cas dans le système qui existait jusqu'à présent) : on pourra, cependant, utiliser la donnée administrative de l'année précédente, dans le cas où elle existe. Par ailleurs, le contrôle des données administratives sera effectué de manière autonome.

Une phase de contrôle de cohérence supplémentaire, entre sources, opérée sur, a priori, au moins une variable commune (le chiffre d'affaires) sera réalisée, appelée « réconciliation des données individuelles ». Cette phase est nécessaire pour s'assurer que l'on « connecte » bien des données relatives à la même unité : même si, comme cela a été indiqué précédemment, l'utilisation de l'identifiant du répertoire SIRENE facilite l'exploitation des données administratives, on peut se trouver confronté à des questions de déclarations fiscales ne correspondant pas au contour de l'unité, dans certains cas où plusieurs déclarations existent pour cette unité et pour une année donnée, par exemple. Lors de cette phase de réconciliation des données individuelles, un certain nombre de données d'entreprises vont être expertisées de manière manuelle par les gestionnaires. On reviendra ci-après sur les conséquences que cette phase a sur les estimateurs à utiliser.

Une autre difficulté liée à l'utilisation de sources multiples est relative au travail de contrôle-redressement. Il est envisagé d'utiliser des méthodes de vérification sélective ciblant les unités sur lesquelles il est nécessaire de mener un expertise manuelle, ces méthodes s'appuyant sur des fonctions de score du type (Lawrence, McKenzie, 2000 ; Hedlin, 2003) :

$$w_i (z_i - y_i),$$

où z_i est une valeur « attendue » pour une variable, tandis que y_i est la valeur brute transmise par l'entreprise. Le score ainsi calculé est un impact potentiel de l'erreur attendue, et c'est en fonction de la valeur des scores par rapport à des seuils qu'on détermine si un questionnaire doit être expertisé de manière manuelle.

Le fait que les poids w_i soient revus après réception (tardive) des données administratives, lors de la phase de calibrage, entraîne que les valeurs des scores calculées lors de la phase de contrôle initiale (pendant le premier semestre de l'année $n+1$) peuvent être révisées ; il est nécessaire d'avoir une étape finale de recalcul de ces scores. Les calculs menés sur des données d'enquêtes passées, et sur le secteur des services, ont cependant montré que seuls 5% des poids initiaux étaient multipliés par un facteur supérieur à 1.6 sur ce secteur lors de la phase de calibrage. Le classement des unités par rapport aux seuils définis pour les scores ne devrait donc pas être beaucoup modifié ; mais on peut noter que la remarque précédente (à savoir le recalcul des scores) s'applique de fait à toutes les enquêtes, en raison des non réponses totales qui conduisent à une révision des poids.

2.4 Les estimateurs finaux

Comme indiqué dans la partie 2.3, le module de réconciliation des données individuelles conduira à l'expertise manuelle d'un certain nombre de données d'entreprises, et à la proposition de choix en cas de divergences : soit privilégier la donnée administrative, soit privilégier la donnée du questionnaire, soit proposer une tierce valeur. Ceci doit être pris en compte, *in fine*, dans les statistiques produites.

Pour les statistiques sectorielles, l'estimateur final utilisé sera du type :

$$\sum_U Y_{fiscal}(i) 1_{APE_{rep}=X}(i) + \sum_S w_i (Y_{vrai}(i) 1_{APE_{enq}=X}(i) - Y_{fiscal}(i) 1_{APE_{rep}=X}(i))$$

où $Y_{vrai}(i)$ est la valeur finale arbitrée pour une variable Y et une entreprise i , et $Y_{fiscal}(i)$ est la valeur connue par la source administrative (ici, fiscale).

Il faut noter que c'est au travers des poids w_i que les non réponses totales seront prises en compte. Sur ce point, l'utilisation du répertoire sera fondamentale (pour plus de détails sur la manière de procéder, voir Brion, Caron, Pietri-Bessy, 2005).

Ainsi, l'utilisation conjointe de données administratives et de données d'enquêtes permet un contrôle de qualité mutuel de ces fichiers, et, en particulier, évite aux estimations de subir des conséquences graves résultant, par exemple, d'une perte de qualité des sources administratives.

RÉFÉRENCES

- Brion, Ph. (2007). « Redesigning the French structural business statistics, using more administrative data ». *Proceedings of the Third International Conference on Establishment Surveys, Montreal*.
- Brion, Ph., Caron, N., Pietri-Bessy, P. (2005). « Redresser la non réponse totale dans les enquêtes auprès des entreprises : les pièges à éviter. Illustration avec l'enquête innovation ». *Communication aux Journées de Méthodologie Statistique 2005, Insee*.
- Deville, J.-C., Särndal, C.-E (1992). « Calibration estimators in survey sampling ». *Journal of the American Statistical Association*, 87, pp. 376-382.
- Hedlin D (2003). « Score functions to reduce business survey editing at the U.K. office for national statistics ». *Journal of Official statistics*, vol 19, n°2, pp. 177-199.
- Lawrence D., McKenzie R. (2000). « The general application of significance editing ». *Journal of Official Statistics*, vol. 16, n°3, pp. 243-253.