LA MISE A JOUR DE REPERTOIRES D'ENTREPRISES

Philippe BRION[1] (*)

(*) Insee, Direction des statistiques d'entreprises

Résumé

Une des tâches fondamentales que doit assurer un Institut de statistique est de gérer différentes bases de sondage, celles-ci étant ensuite utilisées pour différents processus de production statistique. Chaque base de sondage est potentiellement affectée par un certain nombre de défauts : problèmes de sous-couverture, problèmes de sur-couverture (avec des unités cessées mais dont on ne sait pas qu'elles le sont), problèmes de doublons, mauvaise qualité des informations utilisables pour la stratification.

Le travail de mise à jour en continu des bases de sondage a fait l'objet de peu de développements théoriques, et l'objectif de ce papier est de donner quelques éléments destinés à mieux le formaliser, particulièrement dans le domaine des statistiques d'entreprises (pour lequel c'est le répertoire d'entreprises qui sert de base de sondage). En effet, il existe souvent des métadonnées, ou d'autres informations, qui permettent d'avoir une idée de la « qualité » des variables utilisées pour mettre au point les plans de sondage ; ces informations peuvent aider à orienter le travail de mise à jour, en fonction des opérations de production de statistiques que l'institut statistique s'apprête à mettre en place, ainsi que des moyens disponibles pour les travaux de mise à jour.

Abstract

This paper deals with the question of the updating of sampling frames, especially for business statistics for which the business register is generally used. There has been few methodological works on the question of the choices to make when updating the register with limited means. Some metadata can be helpful to guide this work of updating, taking into account the main topics the statistical office wants to study later on. This paper gives some first elements on this question.

Mots-clés

Échantillonnage, base de sondage, répertoire d'entreprises

Introduction

Une des tâches fondamentales que doit assurer un Institut de statistique est de gérer différentes bases de sondage, celles-ci étant ensuite utilisées lors de la mise en place des processus de production statistique. Chaque base de sondage est potentiellement affectée par un certain nombre de défauts : problèmes de sous-couverture, problèmes de sur-couverture (avec des unités cessées mais dont on ne sait pas qu'elles le sont), problèmes de doublons, mauvaise qualité des informations

[1] Philippe.brion@insee.fr

disponibles qui sont candidates pour être des variables utilisées pour la stratification de divers plans de sondage.

La manière d'utiliser au mieux les bases de sondage, avec les imperfections qui les caractérisent, a fait l'objet de nombreux papiers méthodologiques (pour plus de détails sur le sujet, voir par exemple [1]). En revanche, le travail de mise à jour d'une base de sondage, nécessairement limité par les moyens humains qu'on peut y consacrer, a, à ma connaissance, fait l'objet de très peu de développements formalisés. Pourtant, le statisticien confronté à ce problème est rarement totalement démuni, puisqu'il dispose d'informations permettant de cibler, de manière plus ou moins précise, les « zones » qu'il serait bon de contrôler en priorité : par exemple, informations sur la fraîcheur des données contenues dans la base de sondage (à partir du moment où on a conservé trace de celles-ci, bien sûr), informations sur la plus ou moins grande fiabilité des caractéristiques contenues dans la base de sondage (par exemple suite à une enquête qui peut opérer une sorte de contrôle qualité de certaines caractéristiques).

L'objectif de ce papier est de proposer de premiers éléments de formalisation, tenant compte des éléments quantifiés disponibles, pour guider le travail des responsables de la gestion d'une base de sondage, en se centrant sur le domaine des statistiques d'entreprises. Les éléments présentés ici devront bien sûr être complétés, comme il est indiqué dans la partie 3. La première partie du papier présente le problème, et la formalisation adoptée, alors que la deuxième partie aborde les questions de classement sectoriel, en se limitant toutefois à un exemple très simplifié.

1. Formalisation du problème

Si l'on s'intéresse au domaine des statistique d'entreprises, la base de sondage est généralement constituée à partir d'un répertoire d'entreprises. Celui-ci contient des éléments d'identification (identifiant de l'entreprise, raison sociale, adresse), mais également des critères permettant de stratifier les plans de sondage : souvent, la stratification croise un critère qualitatif (classement de l'entreprise dans un secteur d'activités, en référence à la nomenclature d'activités, NAF au niveau français ou NACE au niveau européen) et un critère de taille établi à partir d'une variable quantitative (chiffre d'affaires, ou effectif salarié).

A l'INSEE, la base de sondage utilisée est SIRUS, répertoire statistique adossé au répertoire inter-administratif SIRENE. Il faut noter que l'identifiant disponible dans ce dernier est utilisé par l'ensemble des administrations françaises ayant des contacts avec les entreprises. Ceci a pour conséquence d'avoir une très bonne couverture du monde des entreprises, et ainsi d'éviter les problèmes de sous-couverture et de doublons ; ce papier n'abordera donc pas ces questions de sous-couverture ou de doublons, et sera centré sur les questions de mise à jour des valeurs des informations disponibles.

En revanche, à côté des problèmes de « qualité » des caractéristiques contenues dans le répertoire, les problèmes de sur-couverture sont présents. Certaines entreprises cessées sont toujours présentes dans le répertoire, et d'autres qui sont enregistrées n'ont même jamais démarré d'activité. Des travaux ont été menés par le passé, pour estimer le nombre de ces unités « fausses actives » [2].

Pour essayer d'atténuer les conséquences néfastes de ces défauts du répertoire, des opérations de nettoyage sont menées chaque année, en fonction d'objectifs spécifiés a priori (par exemple, vérification du bon classement dans la NAF pour les unités contenues dans deux postes de la nomenclature qui semblent présenter une certaine porosité, ou vérification des informations d'une catégorie d'entreprises sur laquelle on veut ensuite mener une investigation statistique). Mais, à ma connaissance, il n'existe pas de démarche formalisée, permettant de justifier d'une manière quantifiée la mise en place de telles opérations qualité.

Cependant, des informations sont disponibles sur la qualité des éléments contenus dans le répertoire : ancienneté des signalements (information qui vient s'ajouter à la date de création de l'entreprise), et contrôle qualité opéré par la confrontation entre les données obtenues lors d'enquêtes par sondage et celles disponibles dans le répertoire. Il semble possible d'utiliser ces informations dans une approche globale destinée à améliorer la qualité des processus de production statistique (cette problématique

rejoint celle qui consiste à ne pas se limiter au seul objectif de réduire l'erreur d'échantillonnage, voir par exemple [3]). Cette approche globale peut également travailler sur la question des arbitrages à opérer entre moyens consacrés aux opérations statistiques stricto sensu, et moyens consacrés aux infrastructures comme les répertoires.

2. Exemple : mise à jour du code d'activité

2.1. Le code d'activité principale

Celui-ci, appelé code APE, fait référence à la nomenclature d'activités françaises (NAF). Chaque entreprise enregistrée dans le répertoire dispose d'une valeur pour ce code : quand l'entreprise se crée, cette valeur est déterminée au vu de la déclaration de l'entreprise, et ensuite cette valeur peut être mise à jour, en fonction de demandes de l'entreprise elle-même, ou suite à des éléments recueillis par certaines enquêtes statistiques (essentiellement les enquêtes structurelles du dispositif Esane [4]).

La qualité de l'information disponible dans le répertoire est donc très variable : certaines entreprises ont un code APE dont la mise à jour est relativement fraîche (en particulier les entreprises situées au-dessus d'un seuil en termes d'emploi salarié, fixé pour la plupart des secteurs à 20 salariés ; pour ces entreprises, des enquêtes statistiques fournissent régulièrement des informations sur leurs différentes activités), d'autres, essentiellement les plus petites, ont une valeur, dans le répertoire, qui peut dater de plusieurs années, voire plus, et ne plus nécessairement correspondre à leur situation présente.

Ceci a pour conséquence que les statistiques qui peuvent être produites sur les différents secteurs d'activité à partir de simples comptages issus du répertoire ont une qualité limitée : des dispositifs comme Esane permettent d'avoir des estimations de meilleure qualité, prenant en compte les mouvements démographiques affectant les différents secteurs d'activité.

2.2. Un exemple simplifié, avec deux secteurs d'activité

On suppose ici qu'il n'y a que deux secteurs d'activité qui composent l'économie et dont on cherche à estimer différentes caractéristiques, le secteur *A* et le secteur *B*. Pour chaque entreprise, on dispose d'une information, dans le répertoire, la classant dans *A* ou *B*, et également d'une information sur la « qualité » de ce code, qui est donc la probabilité, pour chaque entreprise *i*, d'être bien classée : *Pi*.

On se fixe comme objectif d'estimer le total d'une variable *Y* (par exemple le chiffre d'affaires) du secteur *A*, et on va étudier les conséquences de la qualité plus ou moins bonne du classement sectoriel sur différents estimateurs statistiques.

2.2.1. Dans un premier temps, on procède à un recensement

L'estimation du total de la variable *Y* se fait ici en utilisant la valeur du code du répertoire. On n'utilise pas l'enquête pour réévaluer la valeur du code APE, à l'instar de ce qui est fait dans Esane par exemple ; l'enquête – ici un recensement – n'est utilisée que pour obtenir la valeur de la variable *Y* pour chaque entreprise.

2.2.1.1. Estimation du total de la variable sans mise à jour

L'estimateur qu'on utilise est donc biaisé, et d'autant plus que l'erreur concernant la valeur des codes APE est importante. Par rapport à la formalisation « classique » utilisée en théorie des sondages, on se situe ici dans une approche légèrement différente : on considère que c'est la grandeur qu'on

cherche à estimer qui est aléatoire, alors que l'estimateur, dans le cas d'un recensement, a une valeur fixe.

Plus précisément, le paramètre qu'on cherche à estimer vaut :

$$\theta = \sum_A 1(i \in A) Y_i + \sum_B (1 - 1(j \in B)) Y_j$$

où $1(i \in A)$ est la variable aléatoire qui vaut 1 si l'entreprise *i* classée dans le répertoire dans le secteur *A* a une activité « actuelle » qui est bien celle-là.

La sommation, sur chacun des deux sous-ensembles A et B tels que définis à partir du répertoire (ce ne sont donc pas les « vrais » ensembles des unités qui ont, effectivement, les codes d'activité A et B), retient donc, pour A, les unités qui sont bien classées, et pour B celles qui sont mal classées (et donc ont en fait une valeur du code APE égale à A).

L'estimateur utilisé consiste à sommer les chiffres d'affaires sur le secteur *A* tel que connu dans le répertoire, et vaut donc :

$$Estim = \sum_A Y_i$$

Si l'on étudie le risque de cet estimateur, en écrivant une pseudo erreur quadratique moyenne, celle-ci vaut :

$$E(\text{estimat-}\theta)^2 = E(\sum_A (1 - 1(i \in A)) Y_i - \sum_B (1 - 1(j \in B)) Y_j)^2$$

soit :

$$\left[\sum_A (1 - P_i) Y_i - \sum_B (1 - P_j) Y_j\right]^2 + \sum_{A+B} P_i (1 - P_i) Y_i^2$$

si l'on suppose que les variables aléatoires liées au classement sectoriel sont indépendantes entre entreprises, et si l'on nomme *Pi* la probabilité que le code APE de l'entreprise *i* soit exact.

2.2.1.2. On procède à la mise à jour du code APE sur un nombre déterminé d'entreprises

Pour un ensemble d'entreprises déterminées à l'avance, on procède à une mise à jour du code APE. Si l'on nomme *A1* et *B1* les sous-populations mises à jour, il reste deux sous populations *A2* et *B2* sur lesquelles la formule précédent donnant le risque de l'estimateur doit être calculée (en effet, sur *A1* et *B1*, les valeurs de *Pi* passent à 1).

La question est donc de choisir au mieux les populations *A1* et *B1.* On peut disposer d'éléments donnant des ordres de grandeur des valeurs *Pi* et *Yi* : pour les premières, on dispose par exemple d'informations issues d'Esane indiquant les secteurs où l'instabilité, concernant le classement sectoriel, est la plus forte (et à l'inverse des secteurs pour lesquels le classement est quasi-certain). Pour les valeurs *Yi*, qui sont obtenues à partir du recensement opéré, on peut disposer d'une valeur antérieure qui donne un bon proxy.

Ensuite, il faut donc déterminer les sous-populations A1 et B1, de façon à minimiser la grandeur :

$$\left[\sum_{A2}(1-P_i)Y_i-\sum_{B2}(1-P_j)Y_j\right]^2+\sum_{A2+B2}P_i(1-P_i)Y_i^2$$

A ma connaissance, il n'existe pas de solution simple à cette question. Des simulations menées sur des jeux d'essai permettraient sans doute de voir si, en se limitant aux *(1-Pi)Yi²* , on dispose d'un critère approché relativement pertinent pour opérer la sélection des unités à mettre à jour en priorité, en fonction d'un budget fixé à l'avance par exemple.

2.2.2. On procède à un sondage

Pour simplifier, on va se situer dans le cadre d'une sondage aléatoire simple. Mais, vu que l'on utilise le classement sectoriel fourni par le répertoire, on va limiter le sondage à la partie *A*, dans laquelle on sélectionne un échantillon de taille *n*.

Sur cet échantillon, on collecte la valeur de la variable *Y*. On a maintenant deux niveaux d'aléatoire : celui dû au classement sectoriel (qu'on appellera ici premier niveau), et celui dû au sondage (deuxième niveau). L'estimateur utilisé est l'estimateur classique de Horvitz-Thompson résultant du sondage sur *A*.

Si l'on s'intéresse à la pseudo erreur quadratique moyenne, on peut la décomposer comme suit (ici encore, le paramètre à estimer *θ* est aléatoire :

$$E(\text{estimat-}\theta)^2=E_1E_2(estimat-E_2(estimat)+E_2(estimat)-\theta)^2$$

Ce qu'on peut encore écrire :

$$E(\text{estimat-}\theta)^2=E_1(\text{var}iancesond)+E_1(\sum_A Y_i-\theta)^2 \quad ,$$

puisque *E2(estimat)* vaut la somme de la variable *Y* sur *A.*

Le premier terme est l'espérance, relativement au niveau d'aléatoire constitué par le classement sectoriel, de la variance due au sondage aléatoire dans A, alors que le second terme est celui qui a été calculé dans la section précédente (2.2.1.1.).

On peut penser que, suite à la mise à jour d'un certain nombre d'unités dans le répertoire, la valeur du premier terme ne sera pas beaucoup modifiée, dans le cas d'un sondage aléatoire et de la mise à jour d'un caractère qualitatif comme le code APE ; c'est donc sur le deuxième terme qu'il faudra agir, et on est donc ramenés au cas précédent.

3. Prolongements à donner

Deux grands axes peuvent être dégagés quant aux prolongements qui pourraient être donnés aux développements présentés ci-dessus : affiner les travaux sur les types de variables (en prenant en compte des plans de sondage plus complexes que celui présenté dans la partie précédente), et adopter une approche globale du problème, prenant en compte des éléments de coût.

3.1. Différents types de variables

La partie précédente a étudié le cas d'une variable qualitative, le code APE. Mais le répertoire contient également des variables quantitatives, certaines étant particulièrement utiles pour définir des catégories de taille d'entreprises.

Dans ce cas, l'hypothèse formulée dans la partie 2.2.2. ne tient plus : on est face à la question des *stratum jumpers* (changements de strates), ce qui aura une effet sur la variance d'échantillonnage.

Dans la formule présentée en 2.2.2., le premier terme sera sans doute réduit si la mise à jour est bien ciblée, et la formalisation devient plus complexe.

De même, on peut chercher à travailler sur les unités « fausses actives ». Dans [2] par exemple, on trouve des travaux de modélisation de la probabilité d'être active en fonction de différents critères (secteur, catégorie juridique, nombre de salariés, région, date de création). Ces éléments peuvent également être pris en compte pour orienter les travaux de mise à jour.

Enfin, la production des statistiques d'entreprises est confrontée à une question difficile, qui est celle de l'unité à prendre en compte : le répertoire d'entreprises enregistre des unités appelées unités légales, alors que la réalité économique, particulièrement dans la cadre de grands groupes où les structures juridiques sont complexes, nécessite de passer par un concept d'entreprise qui n'est pas nécessairement identique à l'unité légale. Cette question fait l'objet à l'heure actuelle de beaucoup de développements nouveaux (voir par exemple [5] pour une approche formalisée de cette question), et la problématique de la constitution de ces unités « entreprises » rejoint, en termes de qualité de production des statistiques finales, celle qui a été présentée ci-dessus.

3.2. Vers une approche globale ?

Pour un office statistique, chaque opération a un coût, et on peut postuler un modèle de coût :

$$C = mC1 + nC2$$

où chacune des *m* mises à jour a un coût *C1* alors que le coût unitaire d'enquête vaut *C2* (la taille de l'échantillon est *n*).

La question posée est celle de l'arbitrage entre mises à jour (qui ont un impact sur l'erreur totale) et taille de l'échantillon (qui a un impact sur l'erreur d'échantillonnage).

Pour avancer sur cette question, il serait nécessaire de modéliser une fonction de gain du type :

$$gain = f(m) + \frac{V}{n}$$

La difficulté réside bien entendu dans la détermination de la fonction *f(m)* ; si l'on arrive à « approximer » par une fonction « classique », on devrait pouvoir, par des techniques d'optimisation, déterminer des arbitrages à opérer entre *m* et *n*.

Bibliographie

[1] Sautory O., « Les enjeux méthodologique liés à l'usage de bases de sondage imparfaites », *Recueil du Symposium 2013 de Statistique Canada*.
[2] Mariotte H., « Estimating the number of falsely active legal units in the French Business Register », *papier présenté à la treizième table ronde sur les répertoires,* Paris, 27 septembre-1er octobre 1999.
[3] Lyberg L., « La qualité des enquêtes », *Techniques d'enquête*, vol 38, N°2, décembre 2012.
[4] Brion Ph., « Esane, le dispositif rénové de production des statistiques structurelles d'entreprises », *Courrier des statistiques n°130*, Insee, 2011.
[5] Brion Ph., Deroyon T., Gros E., « A first assessment of profiling in France », *présentation à l'ENBES workshop – the unit problem in business statistics*, Genève, 10 novembre 2014.