LE TRAITEMENT DES CHANGEMENTS DE CONTOURS DES ENTREPRISES PAR PARTAGE DES POIDS

Arnaud Fizzala

Insee - 88 avenue Verdier CS 70058 92541 Montrouge Cedex, France - arnaud.fizzala@insee.fr

Résumé

Les statistiques d'entreprises françaises font actuellement l'objet d'une refonte majeure. En effet, ces dernières tendent à s'appuyer de plus en plus sur une notion économique d'entreprise plutôt que sur la notion juridique d'unité légale qui était utilisée jusqu'ici à l'Insee. Le profilage des entreprises, permettant d'établir la liste d'unités légales composant chaque entreprise, est réalisé manuellement pour les plus grandes sociétés et sur la base d'algorithmes pour les autres.

Depuis 2016, le tirage des échantillons de deux grandes enquêtes entreprises de l'Insee - l'enquête sectorielle annuelle et l'enquête annuelle de production - est réalisé au niveau des entreprises profilées. Au moment du tirage, les contours des entreprises profilées sont provisoires, et ce n'est que quelques mois plus tard que l'information actualisée de ces contours peut être utilisée. La mise à jour de ces contours pose plusieurs problèmes méthodologiques, à commencer par la définition de l'échantillon d'entreprises effectivement échantillonné et mis en collecte après mise à jour des contours et le calcul d'un poids d'extrapolation pour ces entreprises. Pour réaliser ces opérations, nous nous sommes appuyés sur la théorie du sondage indirect et sur la méthode généralisée du partage des poids.

L'article présentera dans une première partie une comparaison de deux versions de partage des poids permettant de gérer la mise à jour des contours des entreprises, et dans une deuxième partie une comparaison de deux possibilités d'adaptation à un partage des poids des méthodes de traitement des valeurs influentes utilisées jusqu'ici dans ces enquêtes.

Mots-clés. Enquêtes économiques et dans les entreprises, Sondages indirects, Winsorisation, Profilage.

Abstract

The French businesses statistics are currently undergoing a major change. Indeed, they tend to be more and more based on an economic notion of enterprise rather than on the concept of legal unit. The profiling of an enterprise, that is making the list of legal units belonging to it, is done manually for the largest companies and based on algorithms for the others.

Starting with the reference year 2016, the drawing of two major Insee business surveys' sample has been conducted at the enterprise level. At the time of the draw profiled units' composition are provisory, and it is only few months later that up-to-date composition can be used. Doing the

updating involves several methodological problems, starting with the calculation of an extrapolation weight. We have used the Indirect sampling theory and the generalised weight share method to handle these problems.

The paper will present in a first part a comparison of two versions of the weight share method to manage the updating of the compositions of enterprises, and in a second part a comparison of two adaptations of the influential values treatment used at the legal unit level in these surveys.

Keywords. Business surveys, Indirect sampling, Winsorization, Profiling.

1. Résumé long

Les statistiques structurelles d'entreprises, photographie annuelle de la population des entreprises appartenant au système productif, constituent le matériau de base pour de nombreux utilisateurs de statistiques : les statisticiens au premier rang desquels les comptables nationaux, l'Union européenne pour la réponse au règlement SBS (structural business statistics), les cabinets ministériels, les fédérations professionnelles, les chercheurs, les économistes, etc. La production de ces statistiques se base sur le système d'information dénommé Esane (élaboration des statistiques annuelles d'entreprises), dont le principe est de conjuguer l'utilisation de données administratives exhaustives et de données d'enquête (Brion 2011).

En France, comme dans de nombreux autres pays européens, les statistiques d'entreprises font actuellement l'objet d'une refonte majeure. En effet, les enquêtes auprès des entreprises, qui étaient jusqu'à présent basées sur l'observation d'unités légales (UL) correspondant à une vision purement juridique de l'entreprise, tendent à s'appuyer de plus en plus sur le concept économique d'entreprise, définie comme « la plus petite combinaison d'unités légales qui constitue une unité organisationnelle de production de biens et services jouissant d'une certaine autonomie de décision, notamment pour l'affectation de ses ressources courantes » pour reprendre la définition donnée dans le règlement européen 696/93 du 15 mars 1993 et dans le décret n°2008-1354 pris en application de la loi de modernisation de l'économie de 2008 (Hecquet, 2010). À cette fin, une importante opération méthodologique de « profilage », consistant à analyser les groupes complexes pour identifier en leur sein, par-delà leur organisation juridique en sociétés, des entreprises pertinentes en tant qu'acteurs économiques, est en cours à l'Insee. Pour les groupes les plus importants, cette opération est réalisée manuellement par des experts spécialisés en étroite coopération avec ces groupes ; pour les autres, le profilage est automatique. Les unités obtenues après profilage sont dénommées « entreprises profilées » (EP), et correspondent en pratique à une liste d'unités légales, ces dernières étant répertoriées dans le répertoire français des « entreprises » Sirene géré par l'Insee.

Depuis l'année de référence 2016, le tirage des échantillons des deux principales enquêtes sur lesquelles s'appuie le système Esane - l'enquête sectorielle annuelle (ESA) et l'enquête annuelle de production (EAP) - est réalisé au niveau des entreprises profilées (Gros et Le Gleut, 2017). Lorsqu'une EP est tirée, toutes[1] les UL du champ des enquêtes qui lui sont rattachées sont sélectionnées dans l'échantillon d'UL correspondant. On envoie alors un questionnaire aux UL de cet échantillon, et les réponses des EP sont ensuite « reconstituées » à partir des retours de questionnaires des UL.

Au moment du tirage des échantillons de l'année de référence T, en novembre T-1, les contours des entreprises profilées sont provisoires, et c'est plus tard, en mars T, que l'information actualisée de ces contours est disponible. Il est naturel de souhaiter utiliser les contours les plus à jour pour

[1] En pratique, toutes les UL ne sont pas forcément interrogées notamment lorsque l'EP appartient à la partie exhaustive de l'échantillon, on se contente parfois des UL au chiffre d'affaires les plus important pour les EP composées d'un grand nombre d'UL.

élaborer les résultats qui concernent l'année T. La méthode généralisée de partage des poids (MGPP), décrite dans *Indirect Sampling* de Pierre Lavallée (Lavallée, 2007), semble toute indiquée pour gérer cette mise à jour de contours. Avec la règle suivante : l'échantillon d'EP après mise à jour des contours est constitué de l'ensemble des EP dont au moins une UL du contour mis à jour appartient à l'échantillon initial d'UL. De fait, les probabilités d'inclusion des EP dans l'échantillon mis à jour suivant cette règle sont complexes à déterminer puisqu'une EP, en fonction des UL qui la composent, peut avoir plusieurs possibilités d'appartenir à l'échantillon final. La MGPP permet de s'affranchir du calcul de ces probabilités d'inclusion tout en associant aux EP appartenant à l'échantillon final des poids d'estimation ayant de bonnes propriétés, i.e. permettant de construire des estimateurs sans biais.

La première partie de l'article sera consacrée à la comparaison de deux versions de la MGPP qui ont été testées sur les données Esane 2016 :

– MGPP avec liens classiques ;

– MGPP avec liens pondérés par le chiffre d'affaires des unités légales.

La version classique est habituellement utilisée dans les enquêtes ménages et la deuxième version, moins habituelle, est a priori davantage adaptée aux statistiques d'entreprises, car elle permet de mieux tenir compte de « l'importance économique » des unités constituant les liens[2].

La mise en place de la MGPP nécessite des adaptations des traitements post-collecte utilisés dans le système Esane. Ces adaptations font l'objet de plusieurs études inscrites au programme de travail de l'année 2018 de la Division Sondages de l'Insee. La deuxième partie de l'article devrait se focaliser sur l'adaptation de la méthode de Winsorisation (Deroyon, 2015) qui était utilisée jusqu'en 2015 pour traiter les valeurs influentes non entachées d'erreurs (representative outliers) à un partage des poids. Jusqu'en 2015, les seuils de winsorisation pour l'ESA et l'EAP étaient déterminés par la méthode de Kokic et Bell (Kokic et Bell, 1994), qui se base sur l'hypothèse d'un plan de sondage aléatoire simple stratifié, ce qui ne correspond plus au cadre dans lequel nous nous retrouvons après un partage des poids.

L'étude en question sera menée à l'été 2018, et nous envisageons deux possibilités :

- Faire comme si l'échantillon d'EP après partage des poids était obtenu par tirage aléatoire simple stratifié dans les strates mises à jour, i.e. définies sur la base des caractéristiques des EP après mise à jour des contours, et appliquer la méthode de Kokic et Bell telle quelle.

- Se ramener au tirage initial en winsorisant une variable « transformée » permettant d'exprimer l'estimateur du total d'une variable Y après partage des poids comme l'estimateur par expansion d'une variable Z_j sur les entreprises de l'échantillon initial (tiré selon un plan de sondage aléatoire simple stratifié).

La deuxième possibilité présente l'avantage de respecter le cadre théorique de calcul des seuils de winsorisation selon la méthode de Kokic et Bell, mais la variable winsorisée ne sera pas « directement » la variable dont on cherchait initialement à détecter et traiter les valeurs influentes. Il est donc possible que l'on manque certaines de ces valeurs. A l'inverse la première possibilité permet de travailler directement avec la « bonne » variable mais les seuils obtenus ne devraient pas être optimaux puisque les hypothèses de Kokic et Bell ne seront pas respectées.

L'étude consistera à réaliser des simulations de tirage d'échantillons et à comparer les performances des deux types de winsorisation pour estimer les totaux de variables fiscales. Ces dernières présentant le double avantage d'être disponibles pour la quasi-totalité des entreprises du champ et d'être très corrélées aux variables d'intérêt des enquêtes.

[2] La MGPP avec liens pondérés est décrite dans l'ouvrage de Pierre Lavallée (2007), ainsi que dans l'article de Deville et Lavallée (2006).

Bibliographie

P. Brion (2011). *Esane, le dispositif rénové de production des statistiques structurelles d'entreprises*, Courrier des statistiques n°130.

J-C Deville, P. Lavallée (2006). *Sondage indirect : Les fondements de la méthode généralisée du partage des poids*, Techniques d'enquête, Vol. 32, N o 2, pp. 185196.

E. Gros, R. Le Gleut (2017). *The impact of profiling on sampling*, presentation à l'European Establishment Statistics Workshop.

T. Deroyon (2015). *Traitement des valeurs atypiques d'une enquête par winsorization - application aux enquêtes sectorielles annuelles*, Acte des Journées de Méthodologie Statistique.

V. Hecquet (2010). *Quatre nouvelles catégories d'entreprises – une meilleure vision du système productif*, Insee Première n°1321.

P.N. Kokic, P.A. Bell (1994), *Optimal winsorizing cut-offs for a stratified finite population estimator*, Journal of Official Statistics, vol. 10, n° 4: 419-435.

P. Lavallée (2007). *Indirect Sampling*, Springer Series in Statistics.