L'IMPACT DU PROFILAGE SUR LA REFONTE DU PLAN DE SONDAGE DES ENQUÊTES SECTORIELLES ANNUELLES

Ronan LE GLEUT, Thomas MERLY-ALPA

Insee, Direction de la méthodologie et de la coordination statistique et internationale

ronan.le-gleut@insee.fr

Mots-clés : profilage, sondage en grappes, optimisation d'un plan de sondage

Résumé

Dans de nombreux pays de l'Union Européenne, les statistiques d'entreprise sont en grand changement. En effet, afin de répondre au règlement européen SBS (Structural Business Statistics), les instituts nationaux de statistiques se sont engagés à fournir à Eurostat des agrégats basés sur la notion économique d'entreprise profilée (EP), qui correspond à la plus petite combinaison d'unités légales (UL) qui constitue une unité organisationnelle de production de biens et de services jouissant d'une certaine autonomie de décision.

A l'Insee, cela se traduit en particulier par une modification du plan de sondage des Enquêtes Sectorielles Annuelles, dont l'un des objectifs est de déduire l'activité principale d'une entreprise via la ventilation en branches de son chiffre d'affaires (CA). Dans ce nouveau contexte, les unités statistiques (EP) sont différentes des unités de collecte (UL), les réponses étant toujours collectées au niveau des UL pour une diffusion en entreprises. Le nouveau plan de sondage peut ainsi être vu comme un sondage en grappes, où une EP (grappe) est sélectionnée puis toutes les UL qui la composent sont interrogées.

La refonte du plan de sondage a tout d'abord conduit à revoir les critères d'exhaustivité de l'enquête. Pour cela, les seuils historiques utilisés précédemment (en termes de CA, d'effectifs et de total de bilan) ont été conservés, en les modulant par un taux de couverture du CA à couvrir par activité. Les EP composées de plus de 20 UL, de plus de 100 salariés et de plus de 50 000 k€ de CA sont également forcées dans l'exhaustif.

Malgré ces critères, certaines EP ont encore un CA atypique par rapport aux autres unités de leur strate de tirage. Trois méthodes sont alors exploitées afin d'identifier des unités atypiques dans chaque strate, et de les forcer dans l'exhaustif :

- la contribution d'une unité à la dispersion du CA dans la strate ;
- la méthode de Kokic et Bell (1994) détectant les unités influentes ;
- l'algorithme des centres mobiles (k-means).

La combinaison des deux premières méthodes permet de détecter des unités influentes dans chaque strate en prenant en compte à la fois le taux de sondage de la strate et le CA des unités. La troisième méthode intervient en validation afin d'être sûr de ne pas avoir manqué quelques cas atypiques au sein de leur activité.

L'étape suivante de la refonte de plan de sondage a consisté à revoir la méthode de calcul des allocations. Tout d'abord, la stratification a été définie par le croisement de l'APE de l'EP et sa

tranche d'effectifs (9 tranches). Les domaines de diffusion considérés pour l'optimisation sont l'APE, le croisement groupe (activité sur 3 positions) x tranche d'effectifs, et le groupe.

La première contrainte du calcul des allocations portait sur le nombre d'UL à interroger. La seconde contrainte portait sur la précision des estimations des totaux de CA sur les trois domaines de diffusion. Afin de respecter toutes ces contraintes, nous avons généralisé l'algorithme de Koubi et Mathern (2009) en introduisant des contraintes de coûts pour respecter le nombre fixe d'UL à enquêter. L'algorithme ne permettant pas de combiner trois domaines de diffusion en même temps, nous avons calculé les allocations optimisées sur chacun des trois domaines de diffusion, puis nous avons calculé une moyenne pondérée des trois jeux d'allocations.

Les résultats attendus sur cette refonte de l'échantillonnage concernaient essentiellement les deux contraintes précédemment détaillées pour le calcul des allocations à savoir :

- respecter le bon nombre d'UL à enquêter ;
- évaluer la précision du nouveau plan de sondage au niveau EP et UL en comparaison avec la précision de l'ancien plan de sondage en UL.

Sur le premier point, malgré l'introduction de contraintes de coûts dans le calcul d'allocations, il n'est possible de maîtriser la taille de l'échantillon en UL qu'en espérance. En effet, le nombre exact d'UL à interroger reste aléatoire et peut varier d'un échantillon à l'autre, car il dépend du tirage effectué au niveau EP. Le calcul analytique de la variabilité de ce volume d'UL à enquêter a permis de démontrer que ce nombre est assez stable, notamment grâce au fait que les plus grosses EP (e.g. plus de 20 UL) sont toutes forcées dans l'exhaustif.

En termes de précision des estimations des totaux de CA, il en résulte que la moyenne pondérée des trois jeux d'allocations permet d'avoir un compromis pour la diffusion en EP sur les trois domaines d'intérêt considérés. Au niveau UL, la précision est améliorée dans 50 % des cas et détériorée également dans 50 % des cas par rapport à l'ancien plan de sondage. Cependant, le nouveau plan de sondage permet de limiter les cas des coefficients de variation très élevés que l'on pouvait rencontrer avec l'ancien plan de sondage.

Bibliographie

[1] Bethel J., « Sample allocation in multivariate surveys », *Survey Methodology*, vol 15, n° 1, pp 47-57, 1989.

[2] Council Regulation (EEC) 696 / 93, « The statistical units for the observation and analysis of the production system in the Community », 1993.

[3] Dalenius T., Hodges Jr J.L., « Minimum variance stratification », *Journal of the American Statistical Association*, vol 54, n° 285, pp 88-101, 1959.

[4] Falorsi P.D., Righi P., « A balanced sampling approach for multi-way stratification designs for small area estimation », *Survey Methodology*, vol 34, n° 2, pp 223-234, 2008.

[5] Gunning P., Horgan J.M., « A new algorithm for the construction of stratum boundaries in skewed populations », *Survey Methodology*, vol 30, n° 2, pp 159-166, 2004.

[6] Kokic P.N., Bell P.A., « Optimal Winsorizing cutoffs for a stratified finite population estimator », *Journal of Official Statistics*, vol 10, n° 4, pp 419-435, 1994.

[7] Koubi M., Mathern S., « Résolution d'une des limites de l'allocation de Neyman », *Acte des Journées de Méthodologie Statistique de l'Insee*, Paris, 2009.

[8] Lavallée P., Hidiroglou M.A., « On the stratification of skewed populations », *Survey Methodology*, vol 14, n° 1, pp 33-43, 1988.

[9] Merly-Alpa T., Rebecq A., « Optimisation d'une allocation mixte », *9ème colloque francophone sur les Sondages*, Gatineau, 2016.