ÉCHANTILLONNAGE SPATIAL VIA DES DISTANCES SOCIO-ÉCONOMIQUES : COMPARAISON DE MÉTHODES POUR LE TIRAGE D'UN ÉCHANTILLON-MAÎTRE

Samuel GIVOIS (*), Thomas MERLY-ALPA (**)

(*) Ensae, stage à l'Insee, Direction de la méthodologie et de la coordination statistique et internationale
(**) Insee, Direction de la méthodologie et de la coordination statistique et internationale

samuel.givois@ensae-paristech.fr
thomas.merly-alpa@insee.fr

Mots-clés : échantillonnage spatial, équilibrage

Résumé

Cet article présente des travaux réalisés dans le cadre d'un stage de deuxième année de l'Ensae et qui se sont inscrits dans la réflexion portant sur la méthode de tirage du futur échantillon-maître.

L'échantillon-maître actuel a été sélectionné à l'aide d'un tirage équilibré sur des variables corrélées avec les concepts d'intérêt pour les enquêtes auprès des ménages, en utilisant la méthode du cube. Des premiers travaux ont permis de montrer qu'utiliser le cube local était susceptible d'apporter des gains en matière de précision.

Ce travail a consisté à explorer plus en détail la possibilité offerte par les méthodes d'échantillonnage spatial. Le point central de ces méthodes est d'introduire un mécanisme de répulsion réduisant la probabilité d'inclusion double d'unités proches au sens d'une certaine distance. En présence d'autocorrélation spatiale positive, ce mécanisme permet théoriquement une réduction de la variance de l'estimateur Horvitz-Thompson.

Une des pistes de réflexion a été d'élargir la notion de distance au-delà de la seule distance géographique (utilisée de façon naturelle). En effet, un échantillon « spatialement » réparti dans l'espace socio-économique est approximativement équilibré sur les variables utilisées pour la distance. Avec cette approche, on est en mesure d'introduire un nombre plus important de variables auxiliaires qu'avec la méthode du cube, naturellement limitée par la phase d'atterrissage et la taille de l'échantillon cible.

Une comparaison, en matière de précision, de plans de sondage basés sur le cube local et le pivot local a donc été entreprise. Différentes distances ont été employées, basées sur des variables socio-démographiques et/ou géographiques. La distance utilisée la plus simple est la distance euclidienne. Plusieurs variantes ont été testées, avec normalisation ou après projection dans un espace factoriel obtenu à la suite d'une analyse en composantes principales (ACP). La distance de Mahalanobis, permettant une décorrélation des variables, a également été utilisée.

Pour demeurer le plus fidèle possible au tirage de l'échantillon-maître, ces simulations ont reposé sur un échantillonnage à probabilités inégales, proportionnelles au nombre de résidences principales de chacune des unités (l'ensemble des unités formant une partition du territoire de France métropolitaine). Quatre tailles d'échantillon cible ont été testées permettant de contrôler la stabilité des résultats.

Ces travaux ont permis de confirmer que le cube local est globalement préférable au cube simple avec les variables mobilisées pour équilibrer l'échantillon. Le pivot local apparaît quant à lui beaucoup plus sensible à la distance utilisée pour répartir l'échantillon. La précision des méthodes reposant sur le pivot local varie ainsi du pire au meilleur. Son emploi est donc plus incertain, surtout si les variables utilisées devaient changer.

Bibliographie

[1] Chauvet, G. (2009). Stratified balanced sampling. *Survey Methodology*, 35(1):115–119.

[2] Christine, M. et Faivre, S. (2009). Octopusse : un système d'échantillon-maître pour le tirage des échantillons dans la dernière enquête annuelle de recensement. *Actes des Journées de Méthodologie Statistique de 2009*.

[3] Deville, J.-C. et Tille, Y. (1998). Unequal probability sampling without replacement through a splitting method. *Biometrika*, 85(1):89–101.

[4] Deville, J.-C. et Tillé, Y. (2004). Efficient balanced sampling : the cube method. *Biometrika*, 91(4):893–912.

[5] Favre-Martinoz, C. et Merly-Alpa, T. (2016). Utilisation des méthodes d'échantillonnage spatialement équilibré pour le tirage des unités primaires des enquêtes ménages de l'insee. *SFDS - 9ème Colloque Francophone sur les Sondages*.

[6] Favre-Martinoz, C. et Merly-Alpa, T. (2017). Constitution et tirage d'unités primaires pour des sondages en mobilisant de l'information spatiale. *SFDS - 49èmes Journées de Statistique*.

[7] Grafström, A. et Lundström, N. L. (2013). Why well spread probability samples are balanced. *Open Journal of Statistics*, 3(1):36–41.

[8] Grafström, A., Lundström, N. L. et Schelin, L. (2012). Spatially balanced sampling through the pivotal method. *Biometrics*, 68(2):514–520.

[9] Grafström, A. et Tillé, Y. (2013). Doubly balanced spatial sampling with spreading and restitution of auxiliary totals. *Environmetrics*, 24(2):120–131.

[10] Guggemos, F. (2009). Simulations de tirages de zones d'action pour les enquêtes de l'insee. *Actes des Journées de Méthodologie Statistique de 2009*.

[11] Le Gleut, R. (2017). Analyse factorielle et sondage - utilisation de méthodes d'échantillonnage spatial. *SFDS - 49èmes Journées de Statistique*.

[12] McLachlan, G. J. (1999). Mahalanobis distance. *Resonance*, 4(6):20–26.

[13] Tillé, Y. et Wilhelm, M. (2017). Probability sampling designs : Principles for choice of design and balancing. *Statistical Science*, 32(2):176–189.