
NAUTILE : UN NOUVEL ÉCHANTILLON-MAÎTRE

Laurent COSTA, Clément GUILLO, Nicolas PALIOD, Thomas MERLY-ALPA, Ludovic VINCENT,
Martin CHEVALIER, Thomas DERUYON

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

laurent.costa@insee.fr
clement.guillo@insee.fr
nicolas.palioid@insee.fr
thomas.merly-alpa@insee.fr
ludovic.vincent@insee.fr
martin.chevalier@insee.fr
thomas.deroyon@insee.fr

Mots-clés : NAUTILE, échantillonnage, sondage

Résumé

1) Un nouvel échantillon-maître pour 2020 et pour Nautille

Mots-clés : NAUTILE, échantillonnage, équilibrage spatial, analyse en composantes principales

Dans le cadre de la construction du nouvel Échantillon-Maître prévu pour 2020, de nouvelles unités primaires ont été construites. Ces unités primaires sont basées sur des briques communales, en intégrant des distances entre communes par la route à l'algorithme du voyageur du commerce. Ainsi, ce travail préalable a permis le découpage du territoire français en 5155 unités primaires.

Cet article s'intéresse au tirage de ces unités et en particulier au choix de variables d'équilibrages pertinentes.

Un tirage équilibré, réalisable par la méthode du cube, permet de limiter grandement la variance des estimateurs sous réserve que les variables choisies soient suffisamment corrélées avec les variables d'intérêt des enquêtes ménages.

En outre, la méthode du tirage spatialement équilibré, retenue par la suite, offre une variabilité moindre des estimateurs. Ceci dans la mesure où la variance de l'estimateur d'Horvitz-Thompson s'écrit, pour un tirage équilibré :

$$\sum_{i \in U} \sum_{j \in U} \sigma_i \sigma_j \rho_{ij} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j}$$

Avec $y_i = \mathbf{x}_i^T \beta + \epsilon_i, \forall i \in U$, $cov(\epsilon_i, \epsilon_j) = \sigma_i \sigma_j \rho_{ij}, \forall (i, j) \in U^2$ et ρ_{ij} une fonction décroissante de la distance entre les unités i et j .

Le tirage sera effectué au niveau régional avec mise en commun des phases d'atterrissage. Après avoir choisi les variables candidates pour l'équilibrage, nous avons pu déterminer plusieurs scénarios à tester :

- Des scénarios « exhaustifs » regroupant l'ensemble des variables d'équilibrage.
- Des scénarios « parcimonieux » dans lequel le choix des variables d'équilibrage est fait de manière raisonnée.
- Des scénarios « résumés » utilisant les premiers axes d'une analyse en composante principale portant sur l'ensemble des variables candidates.

Les simulations sont analysées sur la base des coefficients de variation estimés par la méthode de Monte-Carlo et sur le respect des contraintes de taille fixe. Il ressort que les scénarios exhaustifs ne respectent pas suffisamment la contrainte de taille fixe, que les scénarios parcimonieux ne sont pas performants sur les variables de zonage urbain/rural et que les scénarios résumés respectent les contraintes de tailles fixes mais manquent de précision sur certaines variables.

L'idée suivante est donc d'étudier des scénarios « hybrides » mêlant des variables choisies parcimonieusement et les premiers axes d'une analyse en composante principale (ACP) réalisée sur les variables restantes.

Des considérations sur l'ordre des variables, le type de variable utilisées (niveaux versus taux), le choix des variables participantes à l'ACP, nous amène à choisir un jeu de variable optimal au regard des méthodes utilisées.

Ce jeu est utilisé dans le tirage de l'Échantillon-Maître.

2) Le renouvellement de l'échantillon Emploi : améliorations et évolutions

Mots-clés : sondage, échantillonnage, sondage spatial, enquête, simulations

L'Enquête Emploi en Continu (EEC) actuelle a été tirée en 2009 dans les fichiers fiscaux pour une durée de 9 ans de par sa construction : un tirage de secteurs et de grappes qui sont des regroupements de logements proches. Cet échantillon arrive donc à épuisement et conduit à un renouvellement.

L'entrée en vigueur en 2019 de ce nouvel échantillon a été l'occasion de reprendre la méthode développée il y a une dizaine d'année et de l'améliorer sur certains points : réduction de la variabilité de la taille des grappes, choix des variables d'équilibrage, mesure de la distance intra-grappes, dispersion géographique des secteurs, etc.

Cette étude passe premièrement par une reconstruction algorithmique des secteurs Emploi (et de leurs grappes) sur l'ensemble du territoire métropolitain pour avoir la base la plus fraîche possible. Cette reconstruction s'établit à partir des fichiers Fidéli (dont les différences avec les fichiers TH seront détaillées) et conjointement au projet Nautile sur le renouvellement de l'Échantillon-Maître (EM) des ménages dans la mesure où la construction de ces grappes et secteurs est restreinte à l'intérieur des nouvelles Unités Primaires (UP) créés dans le cadre du projet. De plus, on s'efforce à les construire à l'intérieur d'un même Iris d'une commune (quand c'est possible) afin de réduire au mieux leur dispersion géographique.

Les grappes construites doivent respecter plusieurs contraintes. On retrouve, par exemple, que :

- tous les logements au sein d'une même grappe doivent être facilement et rapidement accessibles par la route ;
- la taille de ces grappes est bornée (entre 17 et 24 logements *a priori*) ;
- tous les logements d'un même étage doivent être regroupés dans la même grappe par souci de collecte.

Une fois la base établie, des simulations de tirage spatialement équilibré de ces secteurs selon différents scénarios sont réalisées en mobilisant plusieurs variables d'équilibrage ; en particulier, ont été intégrées des variables proxy du statut d'activité (chômeur, actif occupé, etc.) à partir des variables fiscales, en plus des variables socio-démographiques plus classiques (tranches d'âge, etc).

Les CV obtenus sur les variables d'équilibrage et sur d'autres variables sont alors comparés, entre eux et avec la méthode de tirage mise en œuvre en 2009, pour déterminer le meilleur scénario possible et mener au nouveau tirage de l'enquête.

Enfin, des éléments relatifs au protocole d'enquête de l'EEC (semaines de référence, trimestres d'entrée, gestion des logements neufs et non principaux) ont été instruits : des premiers éléments de contexte et de réponse sont présentés en dernière partie de la communication.

3) Coordination spatiale d'échantillons : application à l'EEC et l'Échantillon-Maître

Mots-clés : échantillonnage, équilibrage, sondage indirect, partage des poids, enquête Emploi, enquêtes ménages

Les enquêtes en face à face de l'Insee auprès des ménages reposent sur la concentration spatiale des logements interrogés, afin de limiter les coûts de collecte. D'un point de vue technique, cette concentration est rendue possible par le tirage de zones à l'intérieur desquelles aura lieu pendant plusieurs années la collecte des enquêtes. Pour l'enquête Emploi, ces zones portent le nom de secteurs. Dans le cadre du projet Nautile, les zones tirées pour les enquêtes ménages autres que l'enquête Emploi sont dénommées unités primaires et forment l'échantillon-maître.

Le tirage de l'échantillon-maître est une phase importante dans le processus d'enquête à l'Insee, au même titre que le tirage des secteurs pour l'enquête Emploi. En effet, pour les enquêtes ménages comme pour l'enquête Emploi, il est possible de minimiser la variance des indicateurs phares dès l'échantillonnage. En outre, les choix effectués pour ces tirages ont des répercussions directes tant sur la phase de collecte, que dans les traitements statistiques de la chaîne aval.

Les renouvellements simultanés de l'échantillon-maître pour les enquêtes ménages de l'Insee et de l'échantillon de l'enquête Emploi ont rendu possible l'étude d'un tirage coordonné des deux échantillons. La coordination spatiale des échantillons permet de limiter l'éloignement des zones de collecte de l'échantillon-maître et de l'échantillon Emploi, offrant ainsi la possibilité de réduire les déplacements des enquêteurs au sein des zones de collecte.

Diverses stratégies sont possibles pour un tirage spatialement coordonné :

- tirer les secteurs Emploi dans les unités primaires de l'échantillon-maître ;
- tirer les secteurs Emploi à proximité des unités primaires de l'échantillon-maître, sans contraindre pour autant que les secteurs soient tirés dans les unités primaires de celui-ci.

Les risques d'épuisement rapide de l'échantillon-maître dans le cas de la première option conduisent à étudier les autres possibilités de coordination spatiale. Des unités de coordination regroupant plusieurs unités primaires ont alors été construites afin que les secteurs Emploi soient également tirés dans les unités primaires voisines de celles appartenant à l'échantillon-maître. Cela permet de s'assurer de la proximité spatiale de l'échantillon-maître et de l'échantillon Emploi, tout en garantissant un moindre recouvrement entre les deux échantillons. En respectant cette contrainte, il est possible de tirer des unités de coordination :

- de manière directe, puis de tirer les unités primaires et les secteurs Emploi de manière indépendante dans les unités de coordination déjà tirées ;
- de manière indirecte en tirant d'abord les unités primaires de l'échantillon-maître, ce qui permet d'accéder indirectement aux unités de coordination auxquelles elles appartiennent puis de tirer l'échantillon Emploi dans les unités de coordination ainsi sélectionnées.

Ces options aboutissent à des méthodes d'échantillonnage différentes. D'un côté, il s'agit d'un tirage équilibré des unités de coordination sur des variables d'équilibrage tant liées aux enquêtes ménages qu'à l'enquête Emploi. De l'autre, c'est le tirage des unités primaires qui est équilibré sur les variables liées aux enquêtes ménages. Dans ce dernier cas, l'utilisation d'une méthode de partage des poids adaptée permet également l'équilibrage du tirage des unités primaires sur les variables liées à

l'enquête Emploi, par le biais des unités de coordination. Cette étude compare notamment les résultats de ces deux méthodes.

L'étude présente également la sensibilité des scénarios de tirage de l'échantillon-maître à plusieurs paramètres, tels que la taille des unités de coordination, les nombres d'unités de l'échantillon-maître et de l'échantillon Emploi. Coordonner spatialement les tirages des deux échantillons fait dépendre la précision de l'échantillon Emploi des caractéristiques du tirage de l'échantillon-maître des enquêtes ménages. Dès lors, bien choisir ces caractéristiques en arbitrant entre coûts de collecte d'une part et précision des échantillons d'autre part se révèle central.

L'étude présente enfin les conclusions retenues dans le cadre du projet Nautile de tirage du nouvel échantillon-maître.

Bibliographie

- [1] Favre-Martinoz C., Merly-Alpa T., « Utilisation des méthodes d'échantillonnage spatialement équilibré pour le tirage des unités primaires des enquêtes ménages de l'INSEE », 9^{ème} Colloque Francophone sur les Sondages – Gatineau, 14 octobre 2016.
 - [2] Grafström A., Tillé Y., « Doubly balanced spatial sampling with spreading and restitution of auxiliary totals », *Environmetrics* 24, issue 2, 120-131, 2013.
 - [3] Breidt F.J., Chauvet G., « Improved variance estimation for balanced samples drawn via the Cube method. *Journal of Statistical Planning and Inference* », vol 141, pp. 479 – 487, 2011.
 - [4] Loonis V., « La construction du nouvel échantillon de l'enquête emploi en continu à partir des fichiers de la taxe d'habitation. », Journées de méthodologie statistique de l'Insee, 2009.
-