Sélection des logements pour les enquêtes ménages dans les grandes villes : tirage équilibré versus systématique en présence de non-réponse.

Lionel Delta [1]

[1] *Insee, Département des Méthodes Statistiques, 92120 Montrouge, lionel.delta@insee.fr*

Résumé. L'Insee a récemment renouvelé son échantillon-maître, c'est-à-dire l'échantillon de zones, appelées unités primaires, au sein desquelles seront tirés pendant dix ans les logements interrogés dans le cadre de différentes enquêtes de la statistique publique. Ces différentes zones ont été tirées selon un plan de sondage équilibré spatialement réparti permettant des gains de précision assez important par rapport à d'autres méthodes envisageables. Après ce premier degré de tirage, il était donc envisagé que le second degré de tirage, celui correspondant à la sélection des logements, soit également équilibré lorsqu'il concerne un nombre significatif de logements au sein d'une même unité primaire, généralement les grandes villes.

En se basant sur des simulations de tirages à partir de différents plans de sondage y compris la simulation de phénomènes de non-réponse, nous comparons les performances en matière de précision de différents tirages équilibrés à des tirages systématiques avec tri préalable de la base de sondage selon les mêmes variables auxiliaires.

Les contraintes liées à la taille des échantillons de logements au sein de chacune des unités primaires ainsi que l'existence de non-répondants sont de nature à limiter voire inverser les gains de précisions attendus, en particulier lorsqu'on effectue un calage sur marges. Cette étude permettait donc de justifier le choix final de conserver un tirage systématique au second degré et pose d'autres questions comme celle de la justification de la préférence pour cette option même en l'absence de non-réponse.

Mots-clés. Tirage systématique, échantillonnage équilibré, échantillonnage spatial, Non-réponse, ...

Abstract. INSEE produces a range of household surveys to provide the public with an overview of French society. For these surveys, the samples are chosen by random sampling methods within of a subset of the French population (itself a sample) called "échantillon-maître (EM).

The constitution of the EM constitutes, as such, a first polling stage in which zones (common only or aggregation of municipalities) containing dwellings and called primary units (UP) are selected.

In many UPs, the relatively small fixed size (20) of pull-out samples leaves only a few margins on the choice of pulling method. In other larger UPs, corresponding to large

cities in France, the sample size (greater than 50) is large enough to raise the question of whether to select housing with so-called balanced methods. The goal is to improve the accuracy of the estimators.

This study provides some tools, using both Monte-Carlo simulations and calibrated sampling methods, to estimate the dispersion of indicators of the estimators calculated at the end of the surveys.

The comparison of the results shows that, given the data structure and the relatively small size of the samples, balancing is no more efficient than drawing with fewer constraints on the sample. This is true because so-called calibrated estimators are used to be more precise because they provide a post-collection correction with auxiliary information.

We also sought to simulate non-response behavior among the dwellings surveyed using different methods. The conclusions remain the same as without nonresponse, in the sense that balancing does not seem preferable.

Keywords. Systematic sampling, Balanced sampling, ...

1 Les enquêtes de l'Insee auprès des ménages

Les échantillons constitués pour les enquêtes de l'Insee ne sont pas tirés directement dans la population totale des ménages mais le sont dans un sous-ensemble de la population appelé Echantillon-Maître pendant environ une décennie.

L'Insee a récemment constitué un nouvel EM comprenant un certain nombre d'unités primaires (UP), des zones géographiques. Les unités secondaires (US) que sont les logements seront alors tirées au sein de ces zones. Cette étude s'inscrit donc dans le cadre des travaux nécessaires à la définition du plan de sondage de deuxième degré.

La sélection des logements au sein d'une UP donnée appartenant à l'EM se fait indépendamment du tirage dans les autres UP. Malgré tout et pour des raisons de réprésentativité de l'échantillon, on souhaite que les logements, toutes UP confondues, aient in fine la même probabilité d'appartenir à l'échantillon finalement enquêté.

Une manière de faire est de mettre en place un plan de sondage qui est dit autopondéré : les UP sont tirées proportionnellement à leur nombre de logements, puis un nombre fixe de logements (de l'ordre de 20) est tiré dans chaque UP. Or, certaines UP sont très grandes et doivent être sélectionnées automatiquement dans l'EM (compte-tenu de leur nombre de logements, elles auraient une probabilité d'inclusion spontanée supérieure à 1). En pratique il s'agit d'UP dites exhaustives au sens où une seule grande ville suffit à constituer une UP. Pour préserver l'équipondération entre tous les logements du territoire

français, le nombre de logements à tirer dans ces UP doit être supérieur à 20.

Pour sélectionner les US au sein de chacune des UP, une option est d'avoir recours à des méthodes de tirage équilibré qui génèrent des échantillons à la fois aléatoires et permettant d'estimer sans biais les différents totaux sur les variables d'équilibrage choisies dans la population cible. En découle alors des gains de précision pour toutes les variables d'intérêt corrélées avec les variables d'équilibrage retenues.
Pour cette raison, un tirage équilibré des logements au sein des unités primaires semble être une option tout à fait pertinente pour constituer les échantillons de ménages à enquêter.

Dans les UP non exhaustives, le nombre d'US tirées (20 logements) n'est pas suffisant pour tirer de manière équilibrée. Dès lors que l'on en tire plus comme c'est le cas au sein des UP exhaustives, on peut étudier ce qu'apporte l'équilibrage.

Néanmoins, même pour la plupart de ces UP exhaustives, les échantillons de ménages sélectionnés in fine sont de taille relativement modeste ce qui pourrait être de nature à relativiser l'intérêt d'un tirage équilibré comparativement à d'autres méthodes de tirage. En outre, les enquêtes font apparaître un phénomène de non-réponse qui dégrade la précision des estimations réalisées à partir des échantillons de répondants.

2 Le principe du tirage équilibré

Un échantillonnage équilibré permet spécifiquement d'être représentatif du total de différentes variables auxiliaires quantitatives, connues a priori sur toutes les unités de la population.

On considère une population U de N individus et une variable d'intérêt Y (prenant sur la population les valeurs y_k, $k \in U$) dont on cherche à estimer le total Y sur la population. On suppose par ailleurs qu'on dispose de p variables auxiliaires (la j^{eme} variable prenant sur la population les valeurs x_{kj}, $k \in U$). Un échantillon s est dit équilibré sur les p variables auxiliaires si, pour chacune des ces p variables, l'estimateur de Horvitz-Thompson du total de la variable auxiliaire sur l'échantillon est exactement égal au total de cette de variable sur la population, c'est-à-dire si, pour tout $j = 1..p$:

$$\sum_{k \in s} \frac{x_{kj}}{\pi_k} = \sum_{k \in U} x_{kj}$$

où π_k désigne la probabilité d'inclusion de l'unité k.

Soit S l'ensemble des échantillons possibles sur U. Un plan de sondage p sur S est dit équilibré si l'ensemble des échantillons tels que $p(s) > 0$ sont des échantillons équilibrés.

2.1 Les méthodes du cube et du cube local

La méthode du cube, conçue par Jean-Claude Deville et Yves Tillé à la fin des années 1990, a notamment été utilisée pour le recensement rénové de la population. Cette méthode permet de réaliser des tirages d'échantillons équilibrés sur plusieurs variables auxiliaires tout en respectant les probabilités d'inclusion définissant le plan de sondage.

La méthode du cube local est une variante de la méthode du cube qui permet, outre des échantillons équilibrés sur différentes variables auxiliaires, d'éviter la sélection d'unités voisines; la notion de voisinage étant mesurée par le biais d'une certaine distance sur ces variables spatiales, souvent la distance euclidienne.

L'autocorrélation spatiale peut être causée par des processus inobservés ou difficilement quantifiables tels que les phénomènes d'interaction entre les décisions des agents. Il est préférable d'en tenir compte pour capter au mieux la dispersion initiale des variables d'intérêt pour ne pas perdre en précision. Pour y remédier, il semble approprié de ne retenir parmi les échantillons possibles que ceux qui présentent une répartition spatiale convenable en ce sens que ses unités sont suffisamment dispersées géographiquement.

Dans le cadre des enquêtes de l'Insee, on a tout intérêt à éviter d'interroger des ménages voisins d'autant plus que les tailles d'échantillons au sein d'une UP sont relativement faibles ce qui limite le nombre de contraintes d'équilibrage que l'on peut retenir. L'équilibrage spatial peut servir à capter partiellement les effets de celles qui, parmi les variables auxiliaires, présentent une certaine auto-corrélation spatiale, permettant éventuellement de faire l'économie d'un certain nombre de contraintes d'équilibrage.

3 Résultats

Nous avons alors réalisé de nombreuses simulations afin de procéder à la comparaison des différentes méthodes en fonction notamment du coefficient de variation et de l'erreur quadratique moyenne des estimateurs calés du total et de la proportion. Ces indicateurs ont été estimés à l'aide du procédé de Monte-Carlo. Nous avons également cherché à apprécier dans quelle mesure, la prise en compte de la non-réponse pouvait ou non remettre en question l'efficacité relative de chacune des méthodes.

Nous avons pu montrer que compte tenu de la structure des données et de la taille relativement modeste des échantillons, l'équilibrage n'est pas plus performant qu'un tirage systématique comme on pourrait le penser une fois prise en compte l'étape du calage sur marges; cette analyse est valable que l'on prenne ou pas en compte la non-réponse.

Toutefois, c'est seulement une fois prise en compte le calage sur marges que l'avantage comparatif du tirage systématique est clairement identifié. Sans calage, les différentes méthodes équilibrées affichent des niveaux de précision généralement comparables à ceux

du tirage systématique voire meilleurs sur certains estimateurs.

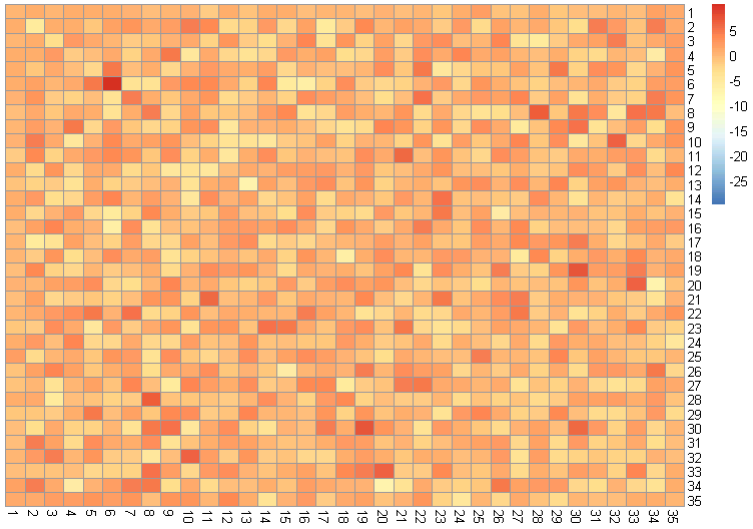

Figure 1: **Probabilités d'inclusion double selon les tranches de revenus, en écarts relatifs à l'indépendance (tirage systématique)**

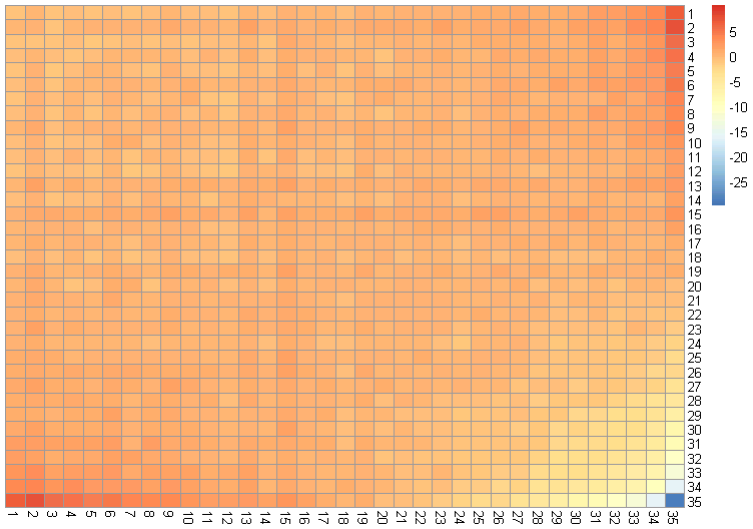

Figure 2: **Probabilités d'inclusion double selon les tranches de revenus, en écarts relatifs à l'indépendance (méthode du cube équilibré sur les niveaux de revenus)**

Tout se passe comme si les échantillons obtenus à partir d'un tirage systématique ont des combinaisons d'unités plus favorables à la repondération. En particulier, le calcul des probabilités d'inclusion double suggère que l'algorithme du cube tend à générer des associations privilégiées d'individus, principalement entre ménages à revenus très élevés et ménages sans revenus comme on le voit sur la figure 2 où l'analyse est faite selon un découpage de la population en 35 tranches de revenus. On y voit des sortes d'effets "granulaires" que l'on ne maîtrise pas bien à ce jour et qui s'avèrent manifestement pénalisants pour les petites tailles d'échantillon.

On suggère donc d'utiliser un tirage systématique, en choisissant les variables de tri les plus pertinentes selon l'enquête visée, pour sélectionner les logements à interroger. Cette étude pourrait être complétée en approfondissant sur les méthodes de correction de la non-réponse, afin notamment d'évaluer si la précision relativement moindre des tirages équilibrés n'est pas propre à l'utilisation du calage sur marges.

Bibliographie

Ardilly, P. (2006), *Les techniques de Sondage*, Editions TECHNIP.

Chauvet, G. et Tillé, Y. (2006), *A fast algorithm for balanced sampling*, Computational Statistics, 21(1):53–62.

Costa, L. et Merly-Alpa, T., *L'échantillonnage équilibré*, Note méthodologique Insee

Deville, J-C, & Tillé, Y., *Echantillonnage équilibré par la méthode du cube, variance et estimation de variance.*, INSEE Méthodes.

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

Tillé, Y. (2011), *Dix années d'échantillonnage équilibré par la méthode du cube : une évaluation.*, Techniques d'enquêtes, Vol. 37, No 2, pp. 233-246, Statistique Canada