Constitution et Tirage d'Unités Primaires pour des sondages en mobilisant de l'information spatiale

Cyril Favre-Martinoz [1] & Thomas Merly-Alpa [2]

[1] *INSEE, Département des Méthodes Statistiques,* cyril.favre-martinoz@insee.fr
[2] *INSEE, Département des Méthodes Statistiques,* thomas.merly-alpa@insee.fr

Résumé. L'objectif de cette communication est d'évaluer, dans le cadre du tirage des unités primaires des enquêtes ménages de l'INSEE, l'apport de l'information géographique sur deux éléments : la constitution d'unités primaires et le tirage équilibré. Nous commençons par décrire notre méthode de constitution d'unités primaires comme groupes de communes, à l'aide d'un algorithme solution au problème du voyageur de commerce. Nous détaillons ensuite la méthode de tirage spatialement équilibré (Grafström et Tillé (2013)). La propriété d'équilibrage spatial est importante dans la mesure où elle permet de s'assurer que les unités primaires tirées sont éloignées géographiquement. En effet, tirer des unités proches aurait des conséquences néfastes en termes de précision pour des variables d'intérêt spatialement corrélées. Une étude par simulation est présentée pour mettre en évidence les gains supplémentaires apportés sur certaines variables d'intérêt par la méthode de tirage spatialement équilibrée par rapport à la méthode du cube. Nous étudierons enfin les gains apportés en termes d'équilibrage spatial par la méthode spatialement équilibrée, puis nous présenterons une approximation par Monte-Carlo des probabilités d'inclusion doubles pour la méthode spatialement équilibrée dans le but de présenter des estimations de variance.

Mots-clés. Tirage spatialement équilibré, algorithme du voyageur de commerce, estimation de variance, auto-corrélation spatiale.

Abstract. This article aims to assess, regarding sampling of primary units for INSEE household surveys, the contribution of geographic information on two elements: constitution of primary units and balanced sampling. We begin by describing our method of constitution of primary units as groups of municipalities, using an algorithm to solve the traveling salesman problem. Then we detail the spatially balanced drawing method (Grafström and Tillé (2013)). We provide a simulation study to show the additional gains on certain variables of interest due to spatial dispersion with respect to the cube method. Finally we present a Monte-Carlo analysis of variance estimators for this sampling method.

Keywords. Spatial balanced sampling, traveling salesman problem algorithm, variance estimation, spatial auto-correlation.

1 Introduction

En France, une grande partie des enquêtes auprès des ménages réalisées par l'INSEE sont issues d'un échantillon appelé Échantillon-Maître, c'est à dire un ensemble de zones qui sont sélectionnées pour plusieurs années. C'est un ensemble de zones de collecte sélectionnées pour représenter au mieux la population française tout en permettant de limiter le coût de la collecte des enquêtes.

On se placera ici dans une population U d'unités primaires de taille $N = 5128$. On souhaite estimer le total de la variable d'intérêt y, noté $t_y = \sum_{i \in U} y_i$. De la population, on tire un échantillon S, de taille (espérée) $n = 488$[1] selon un plan de sondage p. Un estimateur classique de t_y est l'estimateur de Horvitz-Thompson, $\hat{t}_{y\pi} = \sum_{i \in S} d_i y_i$, où $d_i = 1/\pi_i$ désigne le poids de sondage de l'unité i et π_i désigne sa probabilité d'inclusion dans l'échantillon de l'unité primaire i. Les probabilités d'inclusion des unités primaires ont été fixées proportionnellement à la taille de celles-ci en termes de résidences principales, comme pour le précédent Échantillon-Maître.

2 Constitution des unités primaires

Pour déterminer un Échantillon-Maître, il est nécessaire de définir les unités primaires, c'est à dire une partition géographique du territoire français qui va servir de base de sondage. Nous allons présenter ici un découpage basé sur la brique communale, mais d'autres approches sont possibles, comme un découpage carroyé, actuellement étudié également à l'Insee (voir un exemple de mise en oeuvre au Portugal : Santos and Schoenmakers (2013)). Les caractéristiques souhaitées des unités primaires, dites UP par la suite, sont les suivantes :

- les UP sont composées d'au moins 2 500 résidences principales, afin de ne pas réinterroger le même ménage plusieurs fois ;
- les UP sont d'étendue minimale afin de limiter les déplacements ;
- idéalement, les UP sont assez hétérogènes en intra, afin de limiter l'effet de grappe issu de leur sélection.

Tester l'ensemble des partitions d'un ensemble à n éléments distincts n'est pas envisageable d'un point de vue computationel. Le cardinal de cet ensemble correspond au n-ième nombre de Bell. Or le 50^e nombre de Bell est déjà de l'ordre de 10^{47}. Afin de minimiser l'étendue moyenne des nouvelles UP, nous proposons d'utiliser l'algorithme bien

[1] Ce chiffre correspond au nombre d'unités primaires hors exhaustives tirés pour l'Echantillon-Maître actuel

connu du voyageur de commerce (voir, par exemple Applegate et al. (2003)) en mobilisant la matrice de distance par la route entre toutes les communes d'un département. Cet algorithme permet d'approximer à partir d'un point de départ fixé le trajet le plus court pour visiter l'ensemble des communes d'un département.

Nous créons un jeu d'unités primaires de la façon suivante : à partir du point de départ de l'algorithme, on parcourt les communes situées sur le chemin proposé jusqu'à respecter la taille minimale souhaitée en termes de résidences principales. On continue à parcourir le chemin pour construire les unités primaires suivantes. Afin de limiter la dépendance au point de départ et à la solution obtenue, on effectue 1 000 réalisations de l'algorithme du voyageur de commerce et on choisit ensuite le jeu d'unités primaires qui minimise l'étendue moyenne.

Nous avons donc constitué 5128 unités primaires qui partitionnent l'ensemble des départements. Ces unités primaires sont d'étendue moyenne 25% inférieure aux UP actuellement utilisées à l'INSEE.

3 Tirage spatialement équilibré

Le tirage équilibré (Deville et Tillé (2004)) est une procédure dont le but est de fournir un échantillon respectant les deux contraintes suivantes :

- les probabilités d'inclusion sont respectées ;
- l'échantillon est approximativement équilibré sur p variables auxiliaires.

La méthode de tirage spatialement équilibré a été proposée par Grafström et Tillé (2013). L'idée sous-jacente de cette méthode est de combiner une généralisation de la méthode du pivot spatial (Grafström et al., 2012) et la méthode de tirage équilibré via l'algorithme du Cube.

Dans leur article, Grafström et Tillé (2013) postulent un modèle linéaire entre la variable d'intérêt et les variables auxiliaires de la forme :

$$y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \epsilon_i, \ \forall i \in U$$

où $\boldsymbol{x}_i^\top$ est un vecteur contenant les valeurs prises par l'unité i sur les p variables auxiliaires, $\boldsymbol{\beta} \in \mathbb{R}^p$ est le vecteur des coefficients de régression et les ϵ_i sont des variables aléatoires suivant une loi normale centrée de variance sous le modèle $Var_M(\epsilon_i) = \sigma_i^2$ et de covariance sous le modèle :

$$\forall\, (i,j) \in U^2,\ i \neq j,\ cov_M(\epsilon_i, \epsilon_j) = \sigma_i \sigma_j \rho_{ij}.$$

On suppose que ρ_{ij} est une fonction décroissante de la distance entre les unités i et j. Par exemple, on pourrait supposer que $\rho_{ij} = \rho^{d(i,j)}$, où $d(i,j)$ représente la distance entre les unités i et j.

La variance anticipée sous le plan (supposé non informatif) et le modèle de l'estimateur Horvitz-Thompson $\hat{t}_{y\pi} = \sum_{i \in S} \frac{y_i}{\pi_i}$ s'écrit :

$$E_p E_M \left\{ \left(\hat{t}_{y\pi} - t_y \right)^2 \right\} = E_p \left[\left\{ \left(\sum_{i \in S} \frac{\boldsymbol{x}_i}{\pi_i} - \sum_{i \in U} \boldsymbol{x}_i \right)^\top \boldsymbol{\beta} \right\}^2 \right] + \sum_{i \in U} \sum_{j \in U} \sigma_i \sigma_j \rho_{ij} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j}. \tag{1}$$

En utilisant un plan de sondage équilibré, respectant :

$$\sum_{i \in S} \frac{\boldsymbol{x}_i}{\pi_i} = \sum_{i \in U} \boldsymbol{x}_i,$$

le premier terme de l'équation (1) disparaît. Ainsi pour minimiser le deuxième terme restant de la variance anticipée, il faut choisir des probabilités d'inclusion double π_{ij} les plus faibles possibles lorsque le paramètre $\rho_{ij} \in [0;1]$ est élevé. Sous cette hypothèse, le tirage d'un échantillon dispersé permet de réduire la variance anticipée sous le modèle.

4 Descriptif des simulations mises en oeuvre

L'objectif est de comparer dans ces simulations les méthodes de tirage équilibré et de tirage spatialement équilibré. La méthode de tirage spatialement équilibrée décrite ci-dessus est disponible dans le package R appelé "BalancedSampling" de Grafström et Lisic (2016). Ce Package développé en C++ permet d'appliquer l'algorithme très rapidement, cependant la phase d'atterrissage ne peut s'effectuer que par suppression successive des contraintes.

Nous avons effectué $K = 10000$ tirages selon les deux méthodes, avec différentes tailles d'échantillon n. Le tirage des Unités primaires est stratifié selon les 13 grandes régions françaises. Le positionnement spatial des UP est indiqué via les coordonnées géographiques du centroïde de la plus grande commune (en termes de résidences principales) de l'unité primaire.

Les variables d'équilibrage utilisées sont les suivantes : nombre de résidences par type d'espace (rural, urbain, périurbain), le revenu fiscal, l'âge en trois classes, le type de logemement (hlm ou non) et le type de ménage (monoparental, grande taille, autres).

Le choix de ces variables d'équilibrage pour le tirage des unités primaires d'un Echantillon Maître est présenté en détail dans l'article de Guggemos (2009).

Les variables d'intérêt considérées sont : la population en résidences principales, le nombre de décès et le nombre de naissances pour différentes années de recensement, le

nombre de lis en résidence de tourisme en 2015, le nombre de chômeurs en 2007 et le nombre de personnes par catégories socio-professionnelles en 2012.

5 Résultats en termes de précision

Pour l'ensemble des variables (d'équilibrage et d'intérêt) et quelle que soit la taille d'échantillon considérée, on observe des biais relatifs très nettement inférieurs à 1% pour les deux méthodes utilisées. Les coefficients de variation relatifs sont également très faibles. On constate cependant que l'EQM est légèrement plus élevée pour le tirage spatialement équilibré que pour le tirage équilibré quel que soit la taille de l'échantillon. Cette légère perte s'explique également par un ajout de contraintes dans l'équilibrage. En effet, dans l'équilibrage spatialement équilibré, on a ajouté les coordonnées spatiales des unités primaires indirectement dans l'équilibrage, ce qui contraint davantage la méthode de tirage.

Les coefficients de variation associés aux variables d'intérêts sont très faibles (inférieurs à 1%)[2]. En plus des gains intrinsèques de l'équilibrage, on observe un gain lié à l'équilibrage spatial. Ces gains sont d'autant plus importants que l'autocorrélation spatiale des variables est forte. C'est notamment le cas pour les totaux de la variable de chômage (voir, par exemple la communication de Floch et Le Saout (2015), dans laquelle un test de Moran est mise en oeuvre pour tester l'autocorrélation spatiale du taux de chômage, de la catégorie socio-professionnelle des cadres ainsi que du nombre de lits en résidence de tourisme en 2015.

6 Résultats en termes d'équilibrage spatial

Un autre objectif de la méthode de tirage spatialement équilibré est d'obtenir un échantillon d'unités primaires géographiquement dispersé, afin de bien couvrir tout le territoire français.

Afin de comparer les résultats en termes de dispersion spatiale de l'échantillon, nous allons nous intéresser au polygones de Voronoi. Le polygone de Voronoi associé à une unité primaire tirée regroupe l'ensemble des points du plan plus proches de cette unité primaire, que de toutes les autres unités primaires tirées. On note δ_i[3] le total des probabilités d'inclusion des unités primaires contenu dans le polygone i. On peut montrer que l'espérance sous le plan de δ_i est égale à 1 (voir, par exemple, Grafström et al. (2012)). Ainsi en moyenne, sous le plan, un polygone de Voronoi regroupe une masse de probabilité égale à 1. On définit ensuite l'indicateur de dispersion spatiale suivant :

[2]Ces coefficients de variation ne tiennent compte que du premier degré de tirage, pour évaluer la précision d'une enquête ménage, il faudrait également tenir compte de variance induite par la sélection des ménages

[3]Remarque : l'indice i du polygone n'est pas rattaché à l'unité i de l'échantillon.

$$\Delta = \frac{1}{n} \sum_{i \in S} (\delta_i - 1)^2$$

L'indicateur suivant sera appelé indicateur de Voronoi. Cet indicateur correspond à de la dispersion empirique des δ_i. Ainsi plus l'indicateur Δ est faible, plus les δ_i sont proches de 1 et plus le tirage est spatialement réparti.

La valeur de Δ a été estimée par la moyenne par Monte-Carlo sur l'ensemble des $K = 10000$ échantillons tirés selon la méthode du cube ou selon la méthode du cube spatialement équilibré. On constate alors, comme attendu que le tirage spatialement équilibré est moins dispersé spatialement que le tirage issu de la méthode du cube classique.

Bibliographie

Applegate, D., Cook, W., and Rohe, A. (2003). Chained Lin-Kernighan for large traveling salesman problems. *INFORMS Journal on Computing*, 15(1), 82–92.

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

Floch, J. M. et Le Saout, R. (2015). Econométrie spatiale : une introduction pratique. *Actes des Journées de Méthodologie Statistique de 2015, INSEE*.

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

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

Grafström, A. and Lisic, J. (2016). BalancedSampling: Balanced and Spatially Balanced Sampling. R package version 1.5.2.
http://CRAN.R-project.org/package=BalancedSampling.

Guggemos, F (2009). Simulation de tirages de zones d'action enquêteurs pour les enquêtes ménages de l'Insee. *Actes des Journées de Méthodologie Statistique de 2009, INSEE*.

Santos, A. and Schoenmakers, B.-J. (2013). Using european grid (etrs89-laea-pt-1k) as the foundation for the new portuguese sampling infrastucture. European Forum for Geostatistics.