Utilisation des méthodes d'échantillonnage spatialement équilibré pour le tirage des unités primaires des enquêtes ménages de l'Insee

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

[1] *Direction de la Méthodologie et de la Coordination Statistique et Internationale, division Sondage, INSEE, 18 Boulevard Adolphe Pinard, 75014 Paris, cyril.favre-martinoz@insee.fr*

[2] *Direction de la Méthodologie et de la Coordination Statistique et Internationale, division Sondage, INSEE, 18 Boulevard Adolphe Pinard, 75014 Paris, thomas.merly-alpa@insee.fr*

Résumé. L'objectif de cette communication est de comparer dans le cadre du tirage des unités primaires des enquêtes ménages de l'INSEE deux méthodes de tirage : le tirage équilibré (Deville et Tillé, 2004) et le tirage spatialement équilibré (Grafström et Tillé (2013)). Dans ce document, nous présentons les avantages et inconvénients d'un tirage équilibré spatialement par rapport à un tirage uniquement équilibré. La propriété d'équilibrage spatial est importante dans la mesure où elle permet de s'assurer que le tirage de groupes d'unités primaires éloignées géographiquement. En effet, tirer des unités proches aurait des conséquences néfastes en termes de précision pour des variables d'intérêt spatialement corrélées. Dans cet article, nous commençons par rappeler le principe du tirage équilibré par la méthode du cube, puis nous détaillons la méthode de tirage spatialement équilibré. Une étude par simulation est présentée pour mettre en évidence les gains supplémentaires apportées sur certaines variables d'intérêt par la méthode de tirage spatialement équilibrée par rapport à la méthode du cube. Nous étudierons enfin les gains apportées en termes d'équilibrage spatial de la méthode spatialement équilibrée, puis nous présenterons une approximation par Monte-Carlo des probabilités d'inclusion double pour la méthode spatialement équilibrée dans le but de présenter des estimations de variance.

Mots-clés. Tirage spatialement équilibré, estimation de variance, auto-corrélation spatiale.

1 Introduction

En France, une grande partie des enquêtes auprès des ménages réalisées par l'INSEE sont issues d'un échantillon appelé Échantillon-Maître. Avant la mise en place d'un recensement rotatif en janvier 2004, cet échantillon restait fixe pendant les périodes inter-censitaires. Désormais, les communes de moins de 10 000 habitants sont recensées exhaustivement tous les 5 ans. Pour ce faire, chaque petite commune s'est vue affecter un groupe de rotation compris entre 1 et 5.

Ce changement de méthodologie permet non seulement de lisser la charge financière liée au recensement sur plusieurs années, mais également de disposer de données plus fraiches. En effet, chaque année 7000 petites communes sont recensées et une enquête par sondage au taux de 8% environ est menée dans 900 grandes communes.

Une des conséquences de ce changement de méthodologie est la nécessité de construire des unités primaires particulières appelées ZAE (Zones d'Action Enquêteurs). Ces zones ont été construits en respectant les contraintes suivantes :

- il s'agit de zones fixes et stables dans le temps de façon à leur allouer un enquêteur de façon durable.

- elles doivent comporter des communes des cinq groupes de rotation de façon à ce qu'un échantillon de logements puisse être sélectionné parmi les logements recensés l'année précédente.

- elles doivent également respecter une certaine contrainte de taille afin que la charge de travail de l'enquêteur soit suffisante et que les échantillons tirés une année donnée ne se recouvrent pas.

Pour plus de détails sur la constitution de ces zones, le lecteur pourra se référer à l'article de Christine et Faivre (2009).

Dans l'Échantillon-Maître, les unités primaires ont été tirées proportionnellement à leur taille en termes de nombre de logements principaux via un tirage équilibré sur des totaux régions et stratifié par région. Dans cet article, on cherche à comparer le plan mis oeuvre pour constituer l'Échantillon-Maître actuel, i.e un plan de sondage équilibré avec un plan de sondage spatialement équilibré.

Dans la suite, on se placera dans une population U d'unités primaires de taille $N = 3704$. 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 $p(S)$. Un estimateur classique de t_y est l'estimateur par dilatation, aussi appelé 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.

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

2 Tirage équilibré

Le tirage équilibré 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 équilibré sur p variables auxiliaires. Autrement dit les estimateurs Horvitz-Thompson des totaux des variables d'intérêt sont égaux aux totaux de ces variables d'intérêt dans la population :

$$\sum_{i \in S} \frac{\boldsymbol{x}_i}{\pi_i} = \sum_{i \in U} \boldsymbol{x}_i \tag{2.1}$$

L'algorithme permettant de réaliser un tel tirage est appelé algorithme du cube. Pour décrire le principe de l'algorithme, il est opportun d'avoir recours à la représentation géométrique suivante. Un échantillon est un des sommets d'un N-cube, noté C. L'ensemble des p contraintes, rappelées par l'équation (2.1) définit un hyperplan de dimension $N - p$, noté Q. On note $K = Q \cap C$, l'intersection du cube et de l'hyperplan. Une représentation graphique du problème en dimension 3, issue de l'article de Deville et Tillé (2004) est donnée ci-dessous.

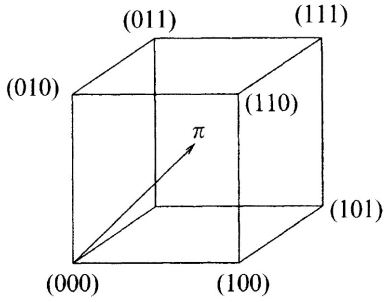

FIGURE 2.1: Représentation graphique du cube pour $N = 3$

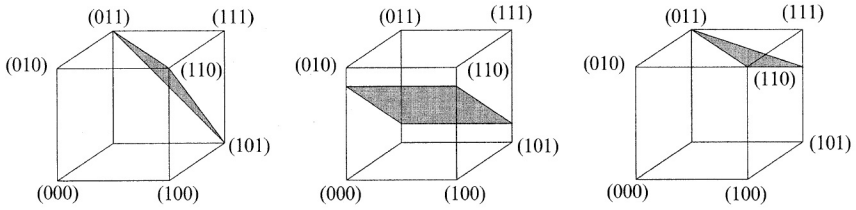

FIGURE 2.2: Représentation graphique des différentes configurations possibles

L'algorithme du cube se décompose en deux phases. La première phase, dite « phase de vol », est une marche aléatoire, qui part du vecteur des probabilités d'inclusion et évolue dans K. Cette marche aléatoire s'arrête lorsqu'elle a atteint

un sommet π^* de K. A l'issu de cette première phase, le sommet π^* n'est pas nécessairement un sommet du cube C. Soit q, le nombre de composantes non entières dans le vecteur π^*. Si q est nulle, la procédure d'échantillonnage est terminée, sinon il faut procéder à la deuxième étape, appelée "phase d'atterrissage". Elle consiste à relâcher le moins possible les contraintes d'équilibrage jusqu'à l'obtention d'un échantillon.

L'implémentation de ce algorithme est disponible sous SAS grâce à la macro FAST CUBE ou sous R dans le package « BalancedSampling » de Grafström et Lisic (2016).

3 Tirage équilibré spatialement

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 méthode de tirage spatialement étalé appelé GRTS (Generalized Random Tesselation Sampling) de Stevens et Olsen (2004) et la méthode de tirage équilibré (CUBE) introduite par Deville et Tillé (2004) présentée en Partie 2.

Dans leur article, Grafström et Tillé (2013) montrent que, sous l'hypothèse de l'existence d'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.

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

$$E_p E_M \left\{ \left(\hat{t}_{y\pi} - t_y \right)^2 \right\} = E_p \left[\left\{ \left(\sum_{i \in S} \frac{\boldsymbol{x}_i}{\pi_i} - \sum_{i \in U} \boldsymbol{x}_i \right)^\top \boldsymbol{\beta} \right\}^2 \right] + \sum_{i \in U} \sum_{j \in U} \sigma_i \sigma_j \rho_{ij} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j}, \tag{3.1}$$

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 (3.1) disparaît, ce qui conduit à une variance anticipée égale à :

$$E_p E_M \left\{ \left(\hat{t}_{y\pi} - t_y \right)^2 \right\} = \sum_{i \in U} \sum_{j \in U} \sigma_i \sigma_j \rho_{ij} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j}$$

Cette dernière expression montre que pour minimiser la variance anticipée sous un modèle linéaire intégrant de la corrélation spatiale entre les unités, il faut choisir des probabilités d'inclusion double les plus faibles possibles lorsque le coefficient de corrélation entre les unités i et j, $\rho_{ij} \in [0; 1]$ est élevé. En général, les unités proches géographiquement sont corrélées et on pourrait par exemple supposer que la corrélation entre les unités i et j, $\rho_{ij} = \rho^{d(i,j)}$, où $d(i, j)$ représente la distance entre les unités i et j. Sous cette hypothèse, le tirage d'un échantillon dispersé permet de réduire la variance anticipée sous le modèle.

L'idée générale de l'algorithme de tirage spatialement équilibré est de construire un cluster de $p + 1$ unités géographiquement proches, puis d'appliquer la phase de vol du cube sur ce cluster. Cela conduit à statuer sur la sélection ou non d'une unité dans ce cluster en respectant les p contraintes localement dans ce cluster. Ensuite, les probabilités sont modifiées localement, ce qui assure que les probabilités d'inclusion des unités proches sont réduites, limitant ainsi la probabilité qu'une de ses unités proches soient sélectionnées par la suite dans l'étape suivante de l'algorithme. Puis on répète la procédure : on sélectionne une unité, on crée un cluster de $p + 1$ unités autour de l'unité sélectionnée et on applique la phase de vol du cube. On répète le processus tant qu'il reste plus de $p + 1$ unités. Pour finir, on applique la phase d'atterrissage classique du cube.

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.

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. Pour cela, nous avons effectué $K = 10000$ tirages selon les deux méthodes, avec différentes tailles d'échantillon $n = 488; 427; 366$. Le tirage des ZAE est stratifié selon les 13 grandes régions françaises. Le positionnement spatial des ZAE est indiqué via les coordonnées géographiques du centroïde de la ZAE. Soient N_h le nombre de ZAE dans la grande région h et n_h^1, n_h^2, n_h^3 les nombres de ZAE à tirer dans la grande région h pour les trois tailles d'échantillon n_1, n_2, n_3. Les valeurs de ces allocations sont spécifiées dans le tableau suivant :

Région	N_h	n_h^1	n_h^3	n_h^3
Ile-de-France	361	84	74	63
Grand-Est	414	45	39	34
Hauts-de-France	416	52	45	39
Normandie	286	28	24	21
Centre	192	22	19	16
Bourgogne-Franche-Comté	256	25	22	19
Pays de la Loire	195	25	22	19
Bretagne	186	26	23	20
Nouvelle Aquitaine	413	49	43	37
Occitanie	349	40	35	30
Auvergne-Rhône-Alpes	473	59	52	44
PACA	146	30	26	22
Corse	19	3	3	2

TABLE 1: Allocations par grande région suivant la taille d'échantillon global

Les variables d'équilibrage utilisées et rangées par ordre d'importance décroissante sont les suivantes (avec leur dénomination utilisée dans la suite) :

- nombre de résidences en zone urbaine : nres_rural
- nombre de résidences en zone périurbaine : nres_peri
- nombre de résidences en zone rurale : nres_urbain
- total du revenu fiscal en 2004 : revenufisc04
- nombre de personnes de moins de 20 ans au recensement de 1999 : age1
- nombre de personnes entre 20 et 59 ans au recensement de 1999 : age2
- nombre de personnes de 60 ans et plus au recensement de 1999 : age3
- nombre de familles monoparentales : monoparental
- nombre de familles de grande taille (quatre enfants et plus) : grande_taille
- nombre de propriétaires de leur logement : proprietaire
- nombre d'étrangers : etranger
- nombre de logements HLM : hlm

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).

5 Résultats en termes de précision

Comme mesure du biais des estimateurs Horvitz-Thompson $\hat{t}_{y\pi}$, on calcule le biais relatif Monte-Carlo :

$$BR_{MC}\left(\hat{t}_{y\pi}\right) = \frac{E_{MC}\left(\hat{t}_{y\pi}\right) - t_y}{t_y} \times 100,$$

où

$$E_{MC}(\hat{t}_{y\pi}) = \frac{1}{K}\sum_{k=1}^{K}\hat{t}_{y\pi}^{(k)},$$

et $\hat{t}_{y\pi}^{(k)}$ l'estimation Horvitz-Thompson obtenu sur l'échantillon tiré lors de la $k-$réplication.

De même, comme mesure du coefficient de variation relatif, on calcule

$$CV_{MC}\left(\hat{t}_{y\pi}\right) = \frac{\sqrt{E_{MC}\left(\left[\hat{t}_{y\pi} - E_{MC}(\hat{t}_{y\pi})\right]^2\right)}}{t_y} \times 100,$$

et l'erreur quadratique moyenne Monte Carlo par :

$$EQM_{MC}\left(\hat{t}_{y\pi}\right) = E_{MC}\left(\left[\hat{t}_{y\pi} - t_y\right]^2\right) = \frac{1}{K}\sum_{k=1}^{K}\left(\hat{t}_{y\pi}^{(k)} - t_y\right)^2.$$

5.1 Sur les variables d'équilibrage

Les tables 1 et 2 résument les résultats obtenus en termes de biais, de CV et d'EQM sur les différentes variables d'intérêt considérées pour les trois tailles d'échantillon fixées. La colonne RATIO correspond au rapport de l'EQM du tirage spatialement équilibré sur l'EQM du tirage équilibré.

	n=488							n=427						
	Cube			Cube spatial			RATIO	Cube			Cube spatial			RATIO
	BR	CV	EQM	BR	CV	EQM	EQM	BR	CV	EQM	BR	CV	EQM	EQM
nres_rural	0,00	1,15	2,45E+09	0,01	1,17	2,50E+09	1,02	0,03	1,32	3,22E+09	0,00	1,36	3,42E+09	1,06
nres_peri	0,01	1,08	2,37E+09	0,01	1,09	2,43E+09	1,03	-0,01	1,25	3,15E+09	-0,01	1,28	3,31E+09	1,05
nres_urbain	0,00	0,52	3,06E+09	-0,01	0,52	3,12E+09	1,02	0,00	0,59	3,95E+09	0,01	0,59	3,92E+09	0,99
revenufisc04	0,00	0,18	1,22E+18	0,00	0,19	1,32E+18	1,09	0,01	0,41	6,19E+18	0,00	0,40	6,09E+18	0,98
age1	0,00	0,18	5,30E+08	-0,01	0,20	6,14E+08	1,16	0,00	0,43	2,82E+09	0,00	0,42	2,81E+09	0,99
age2	0,00	0,11	8,45E+08	0,00	0,11	9,20E+08	1,09	0,00	0,38	1,00E+10	0,00	0,37	9,73E+09	0,97
age3	0,00	0,19	3,96E+08	0,00	0,19	4,33E+08	1,09	0,01	0,43	2,15E+09	0,00	0,43	2,14E+09	1,00
monoparental	0,00	0,21	1,16E+07	0,00	0,22	1,27E+07	1,10	0,00	0,42	4,63E+07	0,00	0,42	4,57E+07	0,99
grande_taille	0,00	0,58	4,24E+06	0,00	0,63	4,93E+06	1,16	0,00	0,78	7,48E+06	-0,01	0,83	8,44E+06	1,13
proprietaire	0,00	0,21	5,83E+08	0,00	0,21	6,18E+08	1,06	0,01	0,44	2,62E+09	0,00	0,44	2,64E+09	1,01
etranger	0,00	1,48	2,58E+07	-0,02	1,53	2,78E+07	1,08	0,03	1,74	3,57E+07	0,01	1,79	3,76E+07	1,06
hlm	0,00	0,61	1,44E+09	0,01	0,64	1,58E+09	1,10	0,01	0,74	2,11E+09	0,01	0,77	2,28E+09	1,08

TABLE 2: Résultats des simulations en termes de biais, CV et EQM pour les variables d'équilibrage $n = 488$; 427

	n=366						
	Cube			Cube spatial			RATIO
	BR	CV	EQM	BR	CV	EQM	EQM
nres_rural	-0,01	1,55	4,41E+09	0,00	1,58	4,57E+09	1,03
nres_peri	0,02	1,47	4,40E+09	0,01	1,47	4,40E+09	1,00
nres_urbain	-0,01	0,68	5,25E+09	0,01	0,68	5,26E+09	1,00
revenufisc04	0,00	0,42	6,74E+18	0,00	0,43	7,03E+18	1,04
age1	0,00	0,43	2,91E+09	0,00	0,44	3,02E+09	1,04
age2	0,00	0,39	1,04E+10	0,00	0,39	1,07E+10	1,03
age3	0,00	0,48	2,61E+09	0,01	0,48	2,68E+09	1,03
monoparental	0,00	0,45	5,16E+07	0,00	0,46	5,41E+07	1,05
grande_taille	0,01	0,85	8,92E+06	0,02	0,89	9,80E+06	1,10
proprietaire	0,00	0,49	3,18E+09	0,00	0,50	3,31E+09	1,04
etranger	-0,02	2,00	4,72E+07	0,02	2,10	5,20E+07	1,10
hlm	-0,02	0,84	2,73E+09	0,01	0,88	3,03E+09	1,11

TABLE 3: Résultats des simulations en termes de biais, CV et EQM pour les variables d'équilibrage pour $n = 366$

Pour l'ensemble des variables d'équilibrage 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 : ils sont inférieurs à 1% pour l'ensemble des variables pour toutes les tailles d'échantillon, hormis pour les variables nres_rural, nres_peri et etranger pour lesquelles l'équilibrage est légèrement dégradé. Néanmoins le coefficient de variation de ces variables demeure inférieur à 2%. De façon plus générale, on constate que la qualité de l'équilibrage se dégrade avec la diminution de la taille de l'échantillon, ce qui s'explique par le fait que celui-ci est plus contraint : on garde le même nombre de contraintes p tout en diminuant les degrés de liberté. Enfin, on constate que l'EQM est légèrement plus élevé 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.

5.2 Sur les variables d'intérêt

	n=488							n=427						
	Cube			Cube spatial			RATIO	Cube			Cube spatial			RATIO
	BR	CV	EQM	BR	CV	EQM	EQM	BR	CV	EQM	BR	CV	EQM	EQM
PSDC99	0,00	0,08	1,62E+09	0,00	0,09	1,79E+09	1,10	0,00	0,38	3,47E+10	0,00	0,37	3,36E+10	0,97
PRES99	0,00	0,09	2,09E+09	0,00	0,09	2,08E+09	1,00	0,01	0,38	3,38E+10	0,00	0,37	3,25E+10	0,96
NAIS99	0,00	0,33	3,28E+08	0,00	0,33	3,17E+08	0,97	0,01	0,51	7,80E+08	-0,01	0,50	7,55E+08	0,97
DEC99	0,00	0,51	4,24E+08	0,01	0,48	3,79E+08	0,89	0,00	0,68	7,46E+08	0,01	0,66	7,15E+08	0,96
PSDC62	0,00	0,73	7,42E+10	0,01	0,68	6,47E+10	0,87	0,00	0,91	1,16E+11	0,00	0,82	9,37E+10	0,80
PRES62	0,00	0,73	7,15E+10	0,01	0,68	6,19E+10	0,87	0,00	0,92	1,12E+11	0,00	0,82	8,89E+10	0,79
NAIS62	0,00	0,91	2,27E+09	0,01	0,85	1,94E+09	0,85	-0,01	1,11	3,32E+09	-0,01	0,99	2,67E+09	0,80
DECES62	0,00	0,98	1,05E+09	0,01	0,91	9,06E+08	0,86	0,01	1,18	1,51E+09	-0,01	1,06	1,23E+09	0,81
chomage_2007	0,02	0,68	1,20E+08	-0,01	0,61	9,66E+07	0,80	0,00	0,83	1,81E+08	-0,01	0,77	1,54E+08	0,85
aucun_dipl_12	0,00	0,70	2,44E+09	-0,01	0,65	2,08E+09	0,85	0,00	0,87	3,70E+09	0,00	0,80	3,19E+09	0,86
bac_dipl_12	0,00	0,47	1,06E+09	0,00	0,44	9,15E+08	0,87	0,01	0,64	1,96E+09	-0,01	0,60	1,72E+09	0,88
RTLIT15	0,06	21,03	1,77E+10	-0,20	18,83	1,41E+10	0,80	-0,13	23,32	2,17E+10	0,34	20,86	1,75E+10	0,81
C12_PMEN_CS1	0,06	3,34	8,82E+08	0,00	3,06	7,38E+08	0,84	0,07	3,71	1,09E+09	-0,02	3,38	9,00E+08	0,83
C12_PMEN_CS2	0,02	0,97	1,03E+09	-0,01	0,88	8,49E+08	0,83	0,01	1,13	1,41E+09	-0,03	1,03	1,16E+09	0,82
C12_PMEN_CS3	0,01	1,05	4,55E+09	0,01	0,95	3,71E+09	0,82	0,05	1,22	6,20E+09	-0,04	1,09	4,93E+09	0,80
C12_PMEN_CS4	-0,02	0,69	3,49E+09	0,01	0,64	3,01E+09	0,86	0,00	0,84	5,17E+09	-0,03	0,76	4,27E+09	0,83
C12_PMEN_CS5	0,01	0,72	1,92E+09	0,00	0,66	1,64E+09	0,85	-0,01	0,86	2,74E+09	-0,01	0,80	2,39E+09	0,87
C12_PMEN_CS6	-0,03	0,82	9,23E+09	-0,01	0,75	7,78E+09	0,84	0,01	0,97	1,31E+10	0,01	0,90	1,13E+10	0,86
C12_PMEN_CS7	0,00	0,37	2,46E+09	-0,01	0,35	2,17E+09	0,88	-0,01	0,56	5,56E+09	0,01	0,53	5,04E+09	0,91
C12_PMEN_CS8	0,02	1,08	5,86E+08	0,00	1,05	5,48E+08	0,93	-0,01	1,26	7,94E+08	-0,01	1,22	7,40E+08	0,93

TABLE 4: Résultats des simulations en termes de biais, CV et EQM pour les variables d'intérêt pour $n = 488; 427$

	n=366						
	Cube			Cube spatial			RATIO
	BR	CV	EQM	BR	CV	EQM	EQM
PSDC99	0,00	0,38	3,62E+10	0,00	0,39	3,71E+10	1,02
PRES99	0,00	0,39	3,53E+10	0,00	0,39	3,61E+10	1,02
NAIS99	0,00	0,53	8,42E+08	0,01	0,53	8,44E+08	1,00
DEC99	-0,02	0,74	8,82E+08	0,00	0,72	8,45E+08	0,96
PSDC62	0,01	0,99	1,36E+11	0,01	0,92	1,18E+11	0,86
PRES62	0,01	0,99	1,31E+11	0,01	0,92	1,13E+11	0,86
NAIS62	0,01	1,20	3,89E+09	0,01	1,10	3,32E+09	0,85
DECES62	0,00	1,28	1,80E+09	-0,01	1,18	1,51E+09	0,84
chomage_2007	0,00	0,89	2,07E+08	-0,01	0,83	1,81E+08	0,87
aucun_dipl_12	0,00	0,94	4,39E+09	0,02	0,88	3,82E+09	0,87
bac_dipl_12	-0,01	0,69	2,30E+09	0,01	0,67	2,12E+09	0,92
RTLIT15	-0,27	25,45	2,58E+10	0,22	22,83	2,09E+10	0,81
C12_PMEN_CS1	0,04	4,06	1,31E+09	0,06	3,71	1,09E+09	0,83
C12_PMEN_CS2	-0,01	1,24	1,69E+09	0,01	1,14	1,42E+09	0,84
C12_PMEN_CS3	-0,01	1,34	7,39E+09	-0,04	1,21	6,02E+09	0,81
C12_PMEN_CS4	0,00	0,89	5,91E+09	0,00	0,85	5,37E+09	0,91
C12_PMEN_CS5	-0,02	0,92	3,12E+09	0,00	0,88	2,88E+09	0,92
C12_PMEN_CS6	-0,01	1,07	1,59E+10	0,03	0,99	1,35E+10	0,85
C12_PMEN_CS7	0,00	0,61	6,69E+09	0,01	0,59	6,21E+09	0,93
C12_PMEN_CS8	0,01	1,39	9,60E+08	-0,03	1,36	9,19E+08	0,96

TABLE 5: Résultats des simulations en termes de biais, CV et EQM pour les variables d'intérêt pour $n = 366$

Pour l'ensemble des variables d'intérêt (voir la description en annexe) et quelle que soit la taille d'échantillon considérée, on observe des biais relatifs nettement inférieurs à 1% pour les deux méthodes utilisées. Les coefficients de variation sont également très faibles (inférieurs à 1%) pour l'ensemble des variables d'intérêt hormis pour les variables RTLIT15[2] et C12_PMEN_CS1[3]. Ces deux dernières sont très peu corrélées aux variables d'équilibrage, ce qui explique leurs CV élevés. Les variables d'équilibrages sont principales issues du recensement de 1999, par conséquent, on observe que pour les variables PSDC99, PRES99, NAIS99, DEC99[4] des CV bien plus faibles pour ces variables que pour les mêmes variables prises en 1962. Ceci s'explique par le fait que le fait que les variables calculées

[2]Cette variable correspond au nombre de lits disponibles en résidence de tourisme en 2015

[3]Cette variable indique le nombre de ménages dont la personne de référence est Agriculteur exploitant en 2012

[4]Ces variables correspondent respectivement à la population sans double compte, à la population en résidence principale, au nombre de naissances et au nombre de décès en France au recensement de 1999

en 1999 sont naturellement plus corrélées et également mieux expliquées par les variables d'équilibrage que les variables calculées 27 ans plus tôt. En plus des gains intrinsèques de l'équilibrage, on observe un gain lié à l'équilibrage spatial. En effet, pour la totalité des variables, le ratio entre l'EQM du tirage équilibré sur l'EQM du tirage spatialement équilibré est inférieur à 1. Ces gains en termes d'EQM observées 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 (chomage_2007) (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 (C12_PMEN_CS3) ainsi que du nombre de lits en résidence de tourisme en 2015 (RTLIT15).

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 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 :

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

L'indicateur suivant sera appelé indicateur de Voronoi. Cet indicateur correspond à de la variance des δ_i. Ainsi plus la variance de Δ est faible, plus les δ_i sont proches de 1, plus le tirage est spatialement réparti. Le cas extrême serait de faire un découpage géographique a priori de l'espace en n polygones regroupant chacun une masse de probabilité égale à 1 (dans notre cas, on pourrait imaginer un regroupement de communes dont la somme des probabilités d'inclusion serait égale à 1), et de tirer une et une seule unité dans chaque polygone ainsi constitué. Afin d'illustrer le principe des polygones de Voronoi, nous donnons l'exemple d'un tirage équilibré spatialement dans la région Rhône-Alpes-Auvergne dans la Figure 6.1. Cette figure représente le découpage en unités primaires de la région Rhône-Alpes-Auvergne. Les communes principales des unités primaires sélectionnées sont représentées en rouge et les figures géométriques noires correspondent

aux polygones de Voronoi construits pour cette réalisation de tirage.

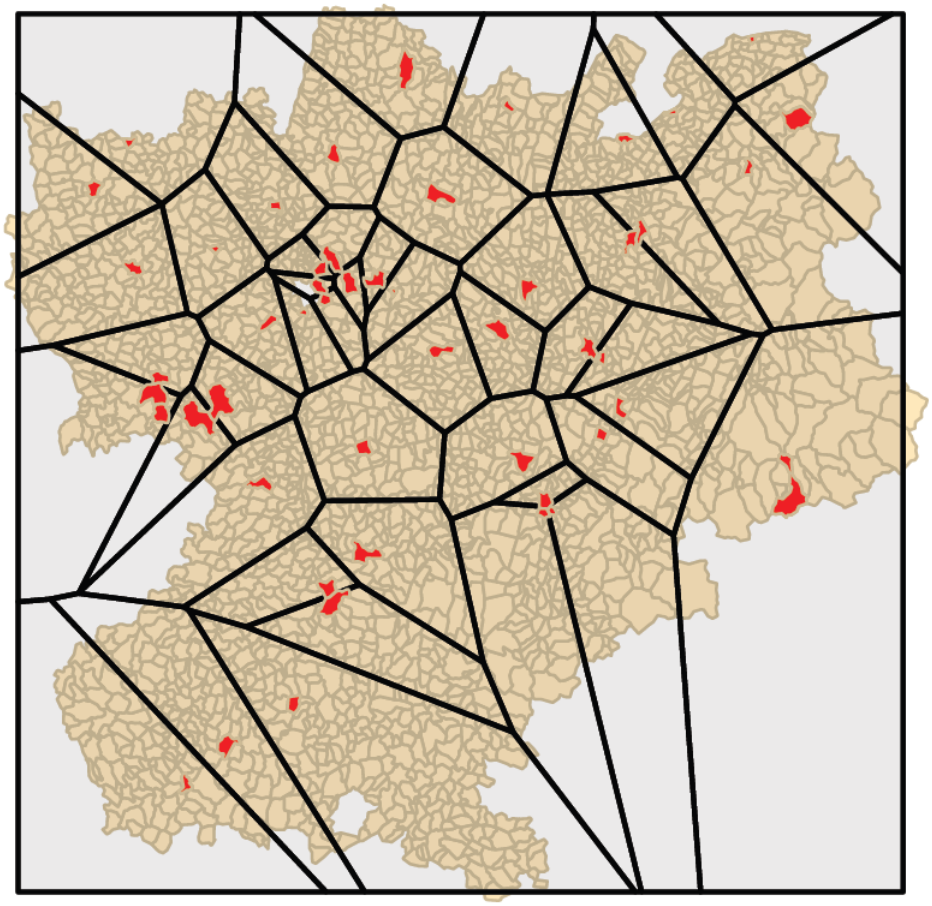

FIGURE 6.1: Représentation graphique des polygones de Voronoi pour la région Rhône-Alpes-Auvergne.

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. Les résultats sont résumés dans la table 5.

	$n = 488$		$n = 427$		$n = 366$	
	Cube	Cube spatial	Cube	Cube spatial	Cube	Cube spatial
Δ	0.35	0.27	0.48	0.38	0.73	0.59

TABLE 6: Valeur de l'indicateur de Voronoi pour différentes tailles d'échantillon

On vérifie que le tirage spatialement équilibré est dispersé spatialement que le tirage issu de la méthode du cube classique. L'étude d'autres critères tels que l'inertie des coordonnées géographiques ou un indicateur de Moran pourrait être envisagée pour confirmer ces résultats.

7 Estimation de variance

Les méthodes de tirage équilibré respectent les probabilités d'inclusion d'ordre 1 fixé par l'utilisateur, cependant les probabilités d'inclusion double nécessaires pour produire des estimations de variance sont inconnues. Une solution alternative consiste à obtenir des approximations des π_{ij} au moyen de méthodes Monte Carlo ; on se reportera à Fattorini (2006) et Wu et Thompson (2008) pour plus de détails. Dans notre étude, nous proposons une approximation de ces probabilités d'inclusion double au moyen de $P = 10^6$ réplications Monte-Carlo du tirage spatialement équilibré. En effet, dans le cas d'un plan de sondage de taille fixe, la variance de l'estimateur Horvitz-Thompson est donnée par :

$$V = -\frac{1}{2} \sum_{i \in U} \sum_{j \in U, j \neq i} (\pi_{ij} - \pi_i \pi_j) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2$$

Cette variance peut être estimée via une approximation des probabilités d'inclusion d'ordre 2 par :

$$\widehat{V} = -\frac{1}{2} \sum_{i \in S} \sum_{j \in S, j \neq i} \frac{(\hat{\pi}_{ij} - \pi_i \pi_j)}{\hat{\pi}_{ij}} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 \tag{7.1}$$

où $\hat{\pi}_{ij} = \frac{1}{P} \sum_{p=1}^{P} I_{ij}^{(p)}$ et $I_{ij}^{(p)}$ est l'indicatrice de présence simultanée des unités i et j des l'échantillon tiré lors de la réplication p.

De façon indépendante, on calcule également sur un million de réplications Monte-Carlo une approximation de la variance, puis nous comparons ces deux approximations. L'objectif est de vérifier que l'on peut obtenir une estimation non biaisée et de variance raisonnable de la variance à partir de l'approximation des probabilités d'inclusion d'ordre 2. Dans le cas du tirage spatialement équilibré, les probabilités d'inclusion d'ordre 2 sont très proches de 0, ce qui peut conduire à des estimations de variance instables. Pour les estimations de variances produites à partir de l'expression (7.1), nous avons écarté les couples d'unités primaires pour lesquelles les probabilités d'incluses double étaient égales à 0 pour $\hat{V}$, puis tous les couples d'unités primaires dont les probabilités d'inclusion étaient inférieures à 10^{-3}, 10^{-4} et 10^{-5}. Les estimateurs résultants seront appelés $\widehat{V}_{seuil,10^{-3}}$, $\widehat{V}_{seuil,10^{-4}}$, $\widehat{V}_{seuil,10^{-5}}$. Cependant, le fait d'écarter certains couples d' unités primaires de l'estimation introduit un biais au profit d'une réduction de la variance des estimateurs de variance.

Une autre façon d'estimer la variance de l'estimateur issu d'un plan de sondage spatialement équilibré provient de l'article de Grafström et Tillé (2013). Ils suggérent de combiner l'estimateur de variance proposé par Stevens et Olsen (2003) dans le cas de tirages spatialement dispersés avec l'estimateur de variance de Deville et Tillé (2005) utilisé dans le cas de plans de sondage équilibrés. Cela conduit dans le cas d'un tirage stratifié et spatialement équilibré à un estimateur de variance de la forme :

$$\hat{V}_{SB} = \frac{p+1}{p} \sum_{h=1}^{H} \frac{n_h}{n_h - p} \sum_{i \in S_h} (1 - \pi_i) \left(\frac{e_i}{\pi_i} - \bar{e}_i \right)^2$$

où

$$\bar{e}_i = \frac{\sum_{j \in G_i} (1 - \pi_j) \frac{e_j}{\pi_j}}{\sum_{j \in G_i} (1 - \pi_j)}$$

et G_i est l'ensemble des $p + 1$ unités les plus proches géographiquement de l'unité i appartenant à la même strate. L'unité i est incluse dans cet ensemble. Par conséquent, $\bar{e}_i$ est une moyenne calculée localement autour de l'unité i sélectionnée.

Les résultats obtenus en termes de variance sont présentés dans la table 6. Dans la première colonne, nous avons indiqué pour les 5 estimateurs proposés le ratio de la moyenne des variances estimées sur 5000 réplications Monte-Carlo sur la variance calculée à l'aide de 1000000 de réplications Monte-Carlo(indépendantes des 5000 premières).

Les estimations issues d'une formule de Yates-Grundy sont d'autant plus stables que le seuil introduit sur les probabilités d'inclusion double est faible. Ce gain en stabilité se fait au détriment du biais : dans le cas d'un seuil à 10^{-3}, on observe en moyenne une sous-estimation quasi systématique de la variance. L'estimateur proposé par Grafström et Tillé a des performances très satisfaisantes, hormis pour des variables d'intérêt parfaitement expliquées par les variables d'équilibrage. C'est le cas notamment des variables PSDC99, PRES99, NAIS99, DEC99. Pour le reste des variables, le ratio de la moyenne des variances estimées $\hat{V}_{SB}$ sur la variance estimée par Monte-Carlo est proche de 1 et l'écart-type associé à ces estimations de variance est beaucoup plus faible (de l'ordre d'un facteur 2) que l'estimation directe par l'estimateur de Yates-Grundy $\hat{V}$.

	Ratio de la moyenne des variances estimées sur la variance Monte Carlo					Ecart-type des estimateurs de variance				
	$\widehat{V}$	$\widehat{V}_{seuil,10^{-3}}$	$\widehat{V}_{seuil,10^{-4}}$	$\widehat{V}_{seuil,10^{-5}}$	$\hat{V}_{SB}$	$\widehat{V}$	$\widehat{V}_{seuil}$	$\widehat{V}_{seuil}$	$\widehat{V}_{seuil}$	$\hat{V}_{SB}$
PSDC99	1,31	1,09	1,20	1,32	0,00	1,36	0,94	1,28	1,35	0,00
PRES99	1,27	1,00	1,17	1,28	0,21	1,18	0,79	1,11	1,17	0,02
NAIS99	1,16	0,73	1,05	1,14	0,68	0,67	0,47	0,62	0,67	0,06
DEC99	1,14	0,50	0,95	1,10	0,85	0,53	0,20	0,42	0,51	0,09
PSDC62	1,13	0,60	0,96	1,10	1,01	0,61	0,24	0,36	0,61	0,08
PRES62	1,14	0,60	0,97	1,11	1,03	0,61	0,25	0,37	0,60	0,09
NAIS62	1,11	0,50	0,92	1,09	1,02	0,58	0,21	0,31	0,58	0,10
DECES62	1,13	0,48	0,94	1,10	1,01	0,44	0,17	0,36	0,43	0,09
chomage_2007	1,08	0,61	0,95	1,07	1,01	0,39	0,29	0,37	0,39	0,09
aucun_dipl_12	1,09	0,58	0,95	1,08	1,04	0,41	0,30	0,39	0,41	0,11
bac_dipl_12	1,06	0,48	0,91	1,04	1,03	0,38	0,21	0,33	0,38	0,12
RTLIT15	1,06	0,30	0,86	1,02	1,16	1,42	0,60	1,30	1,40	0,78
C12_PMEN_CS1	1,12	0,13	0,79	1,06	1,15	0,52	0,12	0,43	0,51	0,16
C12_PMEN_CS2	1,07	0,36	0,87	1,04	1,13	0,37	0,17	0,32	0,36	0,11
C12_PMEN_CS3	1,16	0,72	1,06	1,15	1,06	0,58	0,42	0,54	0,57	0,10
C12_PMEN_CS4	1,11	0,50	0,93	1,09	1,00	0,46	0,24	0,41	0,46	0,11
C12_PMEN_CS5	1,06	0,46	0,90	1,04	1,10	0,34	0,20	0,30	0,33	0,12
C12_PMEN_CS6	1,10	0,44	0,90	1,06	1,13	0,35	0,21	0,31	0,35	0,11
C12_PMEN_CS7	1,13	0,59	0,97	1,11	1,00	0,40	0,22	0,33	0,40	0,12
C12_PMEN_CS8	1,02	0,71	0,93	1,00	0,92	0,42	0,26	0,33	0,42	0,10

TABLE 7: Résultats des simulations sur l'approximation de variance pour $n = 488$.

8 Conclusion

La méthode de tirage spatialement équilibrée permet d'obtenir des résultats plus précis, notamment pour des variables d'intérêt spatialement auto-corrélées au prix d'une légère perte au niveau des variables d'équilibrage. En plus de ce gain en précision, cette méthode permet d'obtenir des échantillons plus dispersés spatialement couvrant une plus grande partie du territoire français. Enfin, malgré une proportion importante de probabilités d'inclusion double très faibles, les première estimations de variance proposées dans cet article sont plutôt satisfaisantes, notamment celles obtenues via l'approximation de variance donnée par Grafström et Tillé (2013). Enfin, cette méthode est fonctionnelle et efficace d'un point de vue computationnelle, dans la mesure où elle est implémentée de façon efficace en C++ dans le package « BalancedSampling » en R.

Bibliographie

Christine, M. et Faivre, S. (2009). OCTOPUSSE : un système d'Echantillon-Maître pour le tirage des échantillons dans la dernière Enquête Annuelle de Recensement. *Actes des Journées de Méthodologie Statistique de 2009, INSEE*.

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

Deville, J. C. and Tillé, Y. (2005). Variance approximation under balanced sampling. *Journal of Statistical Planning and Inference*, 128(2), 569–591.

Fattorini, L. (2006). Applying the Horvitz–Thompson criterion in complex designs : A computer-intensive perspective for estimating inclusion probabilities, *Biometrika*, 93, 269–278.

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*.

Stevens, D. L. and Olsen, A. R. (2003). Variance estimation for spatially balanced samples of environmental resources. *Environmetric*s, 14(6), 593–610.

Thompson, M.E. and Wu, C. (2008). Simulation-based Randomized Systematic PPS Sampling Under Substitution of Units. *Survey Methodology*, 34, 3–10.

Annexe : liste des variables d'intérêt considérées dans les simulations

Les variables d'intérêt considérées sont :

- population en résidence principale au recensement de 1962 et 1999 : PRES62, PRES99
- population sans double compte au recensement de 1962 et 1999 : PSDC62, PSDC99
- nombre de naissances au recensement de 1962 et 1999 : NAIS62, NAIS99
- nombre de décès au recensement de 1962 et 1999 : DECE62, DECE99
- nombre de chômeurs en 2007 : chomage_2007
- nombre de lits disponibles en résidence de tourisme en 2015 : RTLIT15
- population par niveau de diplôme (aucun diplôme, BAC) en 2012 : aucun_dipl_12, bac_dipl_12
- nombre de ménages dont la personne de référence est Agriculteur exploitant en 2012 : C12_MEN_CS1
- nombre de ménages dont la personne de référence est Artisan, Commerçant, Chef d'entreprise en 2012 : C12_MEN_CS2
- nombre de ménages dont la personne de référence est Cadre ou exerce une Profession intellectuelle supérieure en 2012 : C12_MEN_CS3
- nombre de ménages dont la personne de référence exerce une Profession intermédiaire en 2012 : C12_MEN_CS4
- nombre de ménages dont la personne de référence est Employé en 2012 : C12_MEN_CS5
- nombre de ménages dont la personne de référence est Ouvrier en 2012 : C12_MEN_CS6
- nombre de ménages dont la personne de référence est Retraité en 2012 : C12_MEN_CS7
- nombre de ménages dont la personne de référence est Autre sans activité professionnelle en 2012 : C12_MEN_CS8