JMS Journées de méthodologie statistique de l'Insee 2018

Échantillonnage spatial via des distances socio-économiques : comparaison de méthodes pour le tirage d'un échantillon-maître

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

(*) Ensae

(**) 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

Cette étude a été réalisée en amont de la sélection d'un nouvel échantillon-maître (*i.e* un échantillon au premier degré) par l'Insee pour ses enquêtes auprès des ménages. Cet échantillon doit pouvoir être exploité pendant plusieurs années. L'Insee souhaite qu'il ait les meilleures propriétés possibles, et a envisagé pour cela d'utiliser de nouveaux résultats issus de la théorie de l'échantillonnage spatial. Cette étude a consisté à explorer les nouvelles possibilités offertes par la théorie puis à mettre en œuvre une série de simulations afin de confronter les méthodes candidates avec la méthode de référence, un tirage équilibré, effectué région par région. Les méthodes proposées reposent sur le pivot local et le cube local, en utilisant différentes distances. On montre que plusieurs méthodes ont un meilleur comportement que la méthode de référence, ce qui nous a amené à suggérer de les prendre en considération lors de la décision finale.

Abstract

This study was conducted prior to the selection of a new master sample (*i.e* a first-degree sample) by the French Institute of Statistics (INSEE) for its household surveys. This sample must be operational for several years. INSEE wants it to have the best possible properties, and has planned to use new results from spatial sampling theory for this purpose. This work consisted in exploring the new possibilities offered by the theory and in implementing a series of simulations in order to compare the candidate methods with the reference method, a balanced sampling, carried out region by region. The studied methods rely on the local pivotal method and the local cube method, using different distances. Results show that several methods behave better than the reference method, which has led us to suggest that they should be taken into account in the final decision.

Introduction

L'Insee produit un éventail d'enquêtes auprès des ménages afin d'offrir au public un panorama de la société française. Ces enquêtes peuvent être régulières ou plus ponctuelles et portent sur des sujets divers, comme l'évolution des loyers ou l'insertion professionnelle des jeunes adultes. La plupart d'entre elles ont pour point commun la base de sondage au sein de laquelle est sélectionnée la population interrogée *in fine*. Pour ces enquêtes, les échantillons sont en effet choisis via des méthodes de sondage aléatoires au sein d'un sous-ensemble de la population française (lui-même un échantillon) appelé échantillon-maître (EM).

L'échantillon-maître actuel a été sélectionné en 2009 et devrait cesser de fonctionner en 2019, du fait de la dégradation naturelle de sa représentativité, liée à l'évolution différenciée de la structure sociale sur le territoire. De plus, l'opportunité d'utiliser de nouvelles sources de données, essentiellement administrative, a renforcé l'intérêt de ce renouvellement. L'Insee a donc prévu la sélection d'un nouvel échantillon pour le début de l'année 2018.

Le nouvel échantillon-maître, comme ses prédécesseurs, est destiné à une exploitation de plusieurs années. On souhaite donc qu'il possède de bonnes propriétés permettant de conserver une confiance raisonnable dans les estimations en aval. Lors des précédentes opérations, l'échantillonnage s'est appuyé sur de l'information auxiliaire, avec laquelle il a été équilibré. Depuis la sélection du dernier EM, des avancées ont été faites dans la théorie de l'échantillonnage spatial, dont les propriétés semblent prometteuses. L'Insee a envisagé d'en tirer les fruits, si un gain de qualité est confirmé pour ce cas d'utilisation.

Deux méthodes d'échantillonnage spatial en particulier doivent théoriquement permettre d'améliorer les propriétés du processus de tirage. Ce sont le pivot local et le cube local, reposant tous deux sur une distance et visant à réduire la probabilité que des unités proches soient sélectionnées simultanément. Ce comportement est susceptible d'améliorer la précision des estimations réalisées sur les échantillons, en présence d'autocorrélation spatiale positive pour les variables concernées.

Ces méthodes ont été initialement conçues pour fonctionner avec une distance géographique. Cependant, il est possible de contourner cet usage en choisissant d'utiliser une distance dans l'espace des variables socio-démographiques. Une telle approche suppose de savoir quelles variables retenir et quel moyen utiliser pour calculer une distance (qui sont multiples).

L'objectif de cette étude a été de déterminer les méthodes d'échantillonnage qui semblaient les plus appropriées pour la sélection du futur échantillon-maître. Nous commencerons par présenter plus précisément le contexte de cette étude afin d'expliciter les attentes autour de l'échantillon-maître. Nous exposerons ensuite les méthodes d'échantillonnage spatial apparues depuis la dernière opération de sélection de l'EM, ainsi que les données mobilisées pour comparer les différentes méthodes entre elles. Nous proposerons alors une démarche empirique permettant cette comparaison, dont nous présenterons les principaux résultats. Nous discuterons enfin ces résultats au regard de la théorie et des données mobilisées, et nous présenterons une application réalisée au cours de l'étude.

1 Contexte de l'étude

1.1 Les enquêtes auprès des ménages

L'Insee réalise de façon régulière des enquêtes nationales auprès des ménages français, réalisées par sondage sur des échantillons de la population. Ces enquêtes portent sur des questions démographiques, économiques ou sociales et sont la plupart du temps considérées d'intérêt général par le CNIS [1]. On peut citer par exemple l'enquête Loyers et Charges qui permet de mesurer trimestriellement l'évolution des loyers, ou l'enquête Emploi du Temps, qui vise à restituer la façon dont les individus organisent leur temps (une liste plus complète des enquêtes est donnée en annexe A).

Ces enquêtes sont en général menées par la Direction des Statistiques Démographiques et Sociales (DSDS) de l'Insee. Quelques autres le sont par les services statistiques du Ministère du Travail, de l'Emploi, de la Formation Professionnelle et du Dialogue Social (Dares) ou du Ministère des Solidarités et de la Santé (Drees). Les échantillons exploités lors de ces enquêtes sont fournis par la Direction de la Méthodologie et de la Coordination Statistique et Internationale (DMCSI) de l'Insee.

1.2 Le système existant

Une part importante de ces enquêtes sont aujourd'hui échantillonnées à partir des données du recensement de la population. Depuis 2009, un système coordonné, appelé Octopusse [2], permet de réaliser ces sélections en garantissant qu'un ménage déjà interrogé pour une enquête ne puisse pas être réinterrogé pendant 5 ans, réduisant ainsi le « fardeau de réponse ». Des raisons spécifiques conduisent d'autres enquêtes à être échantillonnées hors système. Ainsi, l'enquête emploi en continu, qui permet notamment de calculer le taux de chômage au sens du BIT, est basée sur un fonctionnement aréolaire et donc sélectionnée dans un système à part. Quelques autres enquêtes sont quant à elles issues de sources tierces, essentiellement administratives. Ces dernières, dans le cas des échantillons en France métropolitaine, sont sélectionnées dans une base de sondage proche du futur échantillon-maître.

Le système Octopusse ne sélectionne pas les échantillons des enquêtes ménages au sein de la population totale. Cette sélection est faite au sein d'un sous-ensemble de la population, lui-même sélectionné aléatoirement par sondage, appelé Échantillon-Maître (EM). L'échantillon de logement final utilisé pour une enquête est donc issu d'un échantillonnage à deux degrés (voir Ardilly et Tillé (2006)). Ici, les logements sont des « unités secondaires », sélectionnées au sein de zones les englobant (agrégations de logements), elles-même choisies aléatoirement et appelées « unités primaires ».

L'emploi d'une procédure en deux étapes pour l'échantillonnage est antérieure au système Octopusse. Depuis les années 60 jusqu'en 1999, l'Insee exploitait les données des recensements exhaustifs de la population. À chaque nouveau recensement, une liste exhaustive des logements était ainsi constituée. Ces données étaient complétées entre deux campagnes de recensement par des sources annexes, permettant de prendre en compte les nouveaux logements. L'actualisation ponctuelle de ces données était accompagnée de la sélection d'un nouvel échantillon-maître, qui avait alors une durée de vie moyenne de 8 ans (correspondant au délai moyen entre deux recensements).

L'utilisation d'un tirage à deux degrés est justifié par les gains en termes d'organisation et de coût de collecte qu'il permet de réaliser. L'Insee effectue une part importantes de ses enquêtes en face à face en s'appuyant sur un réseau d'enquêteur maillant le territoire. Elle cherche à apporter

1. Conseil National de l'Information Statistique

2. pour Organisation Coordonnée de Tirages Optimisés Pour une Utilisation StatiStique des Échantillons

le plus de stabilité possible à ce dispositif. La concentration du réseau apportée par un tirage au premier degré permet de s'approcher de cet objectif (Christine et Faivre, 2009). À ce gain organisationnel, est opposé une légère perte de précision, due à « l'effet grappe ».

Depuis 2004, le recensement n'est plus exhaustif mais s'effectue de façon rotative avec un fonctionnement différencié entre grandes (plus de 10 000 habitants) et petites communes. Les petites communes sont recensées de façon exhaustive tous les 5 ans tandis que 8 % de la population des grandes communes est couverte chaque année (soit 40 % sur 5 ans). Ce lissage dans le temps permet un rafraîchissement continu des données disponibles. Le système Octopusse a été conçu pour en tirer profit. L'échantillon-maître sur lequel il repose a été sélectionné en 2009 et conçu pour durer 10 ans. Son exploitation doit donc s'achever au cours de l'année 2019. Par construction, cet échantillon au premier degré intègre les contraintes liées aux rotations du recensement.

Les unités primaires de l'EM d'Octopusse sont appelées « Zones d'Action Enquêteurs »(ZAE). En effet, elles ont été conçues pour qu'un enquêteur puisse la couvrir de façon raisonnable. D'une part, elles contiennent suffisamment de logements pour permettre la conduite de toutes les enquêtes prévues, et d'autre part, elles sont suffisamment « compactes » (en termes d'étendue géographique) pour que l'enquêteur en place puisse tenir la charge. Enfin, chaque ZAE contient des logements de chacun des groupes de rotation du nouveau recensement, afin de bénéficier du rafraîchissement annuel des données [3].

Si ce fonctionnement permet effectivement une mise à jour régulière des informations disponibles, il tend structurellement à se dégrader du fait de l'évolution démographique depuis la constitution de l'EM, qui n'est généralement pas la même selon les ZAE. Un calage annuel permet de réviser les poids de sondage des ZAE et prendre ainsi en compte ces tendances. Cependant, ce procédé n'est pas suffisant. Une étude interne à l'Insee a ainsi montré qu'entre 1999 et 2011, 25 % des ZAE les plus dynamiques avaient vu leur nombre de résidences principales s'accroître au minimum de 29 %, tandis que les 25 % les moins dynamiques connaissaient une croissance de 5 % au maximum. Il est donc nécessaire de renouveler l'EM à la fin de sa période d'exploitation pour retrouver une structure plus fidèle à la population totale.

1.3 Le nouveau système

Parallèlement au besoin de constituer un nouvel échantillon-maître, l'Insee a considéré la possibilité d'exploiter de nouvelles sources de données, essentiellement administratives. Leur utilisation apporte deux avantages par rapport au système existant. D'une part, alors que les données issues du recensement ne sont rafraîchies que partiellement chaque année (20 % des petites communes et 8 % des grandes communes), ces nouvelles informations sont renouvelées de façon exhaustive tous les ans. D'autre part, le recensement ne permet qu'un échantillonnage au niveau du logement, alors que certaines enquêtes nécessitent une sélection au niveau des individus. Enfin, ces sources doivent permettre une sélection d'échantillon mobilisant des données fiscales, que ne contient pas le recensement.

Cette évolution s'est traduite par le lancement d'un nouveau projet, Nautile [4] (projet de refonte de l'échantillonnage des enquêtes ménages), devant remplacer à terme Octopusse. Ce système doit s'articuler avec un autre projet de l'Insee, Fideli [5], destiné à fournir l'information exploitable à partir de diverses sources administratives. Trois groupes de travail ont été créés à l'Insee autour du projet Nautile afin d'étudier des questions jugées stratégiques, et en particulier

3. Ce qui fait que le processus d'échantillonnage actuel est en réalité à trois degrés (ZAE, groupe de rotation et logement). Pour plus de précisions, voir Christine et Faivre (2009).

4. Voir aussi la contribution JMS 2018 associée : « Le projet Nautile (Nouvelle Application Utilisée pour le Tirage des Individus et des Logements des Enquêtes) », par Ludovic VINCENT et Sébastien FAIVRE

5. Pour Fichiers démographiques sur les logements et les individus.

un groupe « Méthodologie » chargé d'instruire les questions liées aux méthodes d'échantillonnage. C'est en périphérie de ce groupe de travail que s'est inscrite cette étude.

1.4 Nouvelles « unités primaires » (UP)

L'utilisation de données administratives renouvelées en intégralité chaque année ne nécessite plus l'utilisation des ZAE, construites pour s'adapter aux contraintes du nouveau recensement. Il n'est plus indispensable en particulier d'inclure 5 groupes de rotation, et il est donc possible de construire de nouvelles unités primaires qui soient plus compactes géographiquement (et donc plus facile à couvrir par un enquêteur). Des travaux permettant de constituer ces nouvelles unités, que nous abrégerons dans la suite du rapport par « UP » étaient en cours de validation à l'Insee lors de cette étude.

Les nouvelles UP sont construites sur une base communale [6]. Une UP est un ensemble de communes, contenant un nombre minimal de résidences principales (pour réduire le risque de « fardeau de réponse »). Ces unités ont été construites de façon à être les plus compactes possibles (Favre-Martinoz et Merly-Alpa, 2017). Le découpage du territoire en UP respecte les frontières départementales, et donc celles des anciennes régions (avant le 1er janvier 2016), chacune d'entre elles hébergeant un établissement régional de l'institut. En effet, les divisions enquêtes ménages (DEM), en charge de la gestion locale des enquêtes, ont été conservées, suite à la création des nouvelles régions au 1er janvier 2016.

1.5 Nouvel échantillon-maître

L'exploitation de l'actuel échantillon-maître a débuté en 2009 et doit s'achever en 2019. C'est pourquoi l'Insee a souhaité sélectionner son successeur en 2018, au sein des nouvelles UP. Depuis la dernière opération de sélection, de nouvelles méthodes d'échantillonnage prenant en considération la dimension spatiale de l'information ont été développées. L'Insee a donc souhaité actualiser le processus de tirage de l'EM, si ces nouvelles approches permettaient d'améliorer la qualité des estimations pour les enquêtes ménage.

L'objectif de cette étude, à partir de ces nouvelles théories, était de proposer une ou plusieurs méthodes d'échantillonnage qui présentent de bonnes propriétés pour la sélection du nouvel échantillon-maître. On cherche en particulier à présenter des stratégies susceptibles de faire mieux que le procédé utilisé lors de la dernière opération de sélection de l'EM. Les fondements théoriques de ces nouvelles méthodes sont présentées dans la partie suivante.

2 Méthodes de sondage étudiées

L'objectif de cette étude était de comparer différentes méthodes d'échantillonnage pour la sélection du nouvel échantillon-maître. On a notamment cherché à savoir si les avancées récentes dans la théorie de l'échantillonnage spatial étaient susceptibles de faire mieux que la méthode employée lors du dernier tirage. On présente dans cette partie les différentes méthodes de tirage étudiées.

2.1 Quelques notations

On se place dans le cadre de la théorie des sondages. Les informations sur la population, celles dont on dispose où celles qu'on cherche à estimer, sont supposées déterministes. En revanche, le processus de tirage est aléatoire, caractérisé par une loi de probabilité sur tous les sous-ensembles

6. Voir aussi la contribution JMS 2018 associée : « Un nouvel échantillon-maître pour 2020 et pour Nautile », par Clément GUILLO et Thomas MERLY-ALPA

(ou échantillons) possibles de la population. On note classiquement U la population totale de taille N, S l'échantillon aléatoire de taille n.

En pratique, on spécifie le loi suivi par l'échantillonnage à l'aide des probabilités d'inclusion. La probabilité d'inclusion simple π_k d'une unité $k \in U$ est la probabilité qu'elle appartienne à l'échantillon, soit $\pi_k = \mathbb{P}(k \in S)$. La probabilité d'inclusion double π_{kl} est la probabilité que les unités $k, l \in U$ appartiennent simultanément à l'échantillon. On a donc $\pi_{kl} = \mathbb{P}(k, l \in S)$, et par convention, $\pi_{kk} = \pi_k$. Pour des plans de sondage complexes, il peut ne pas être possible de calculer explicitement les probabilités d'inclusion double.

Un processus d'échantillonnage peut être vu comme la transformation aléatoire d'un vecteur de probabilités d'inclusion appartenant à $[0,1]^N$ en un vecteur appartenant à $\{0,1\}^N$ représentant l'échantillon sélectionné *in fine* (il doit donc contenir exactement n uns et $N-n$ zéros). Cette approche est commune à l'ensemble des méthodes exploitées dans cette étude.

2.2 Estimateur Horvitz-Thompson

Les enquêtes par sondage cherchent à réaliser des estimations sur des variables Y, appelées variables d'intérêt. Les estimations classiques sont le total, la moyenne, des ratios ou des quantiles. Pour les trois premiers, l'estimateur classique est basé sur l'estimateur Horvitz-Thompson du total. Cet estimateur est sans biais, pour peu que les probabilités d'inclusion π_k soient toutes strictement positives. Pour une variable Y dont on souhaite estimer le total t_y, l'estimateur Horvitz-Thompson est défini par :

$$\hat{t_y} = \sum_{k \in S} \frac{y_k}{\pi_k} \tag{1}$$

2.3 Tirage équilibré

Lors des précédentes opérations de sélection d'un échantillon-maître, la méthode utilisée reposait sur la théorie de l'échantillonnage équilibré sur des variables auxiliaires. L'idée centrale est d'exploiter de l'information disponible sur l'intégralité de la population afin d'améliorer la précision des estimations.

Cette approche fait partie de la famille des méthodes de sondage fondées sur un modèle (Tillé et Wilhelm, 2017). On suppose en effet qu'une relation lie la variable d'intérêt y pour laquelle on souhaite proposer des estimations par sondage, à un ensemble de variables auxiliaires x (vectoriel), disponibles sur l'ensemble de la population. Ce modèle peut par exemple être une simple relation linéaire :

$$y = {}^t x.\beta + \epsilon \tag{2}$$

L'aléa n'est donc plus uniquement contenu dans le processus d'échantillonnage, mais également dans la réalisation des variables observées. En notant $\mathrm{cov}_M(\epsilon_k, \epsilon_l) = \sigma_{kl}$ et avec la convention $\pi_{kk} = \pi_k$, on peut montrer (Grafström et Tillé, 2013) que, sous le plan de sondage et sous le modèle, la variance anticipée de l'estimateur Horvitz-Thompson du total $\hat{t_y}$ est égale à :

$$E_p E_M[(\hat{t_y} - t_y)^2] = E_p\left[\left({}^t\left(\sum_{k \in S} \frac{x_k}{\pi_k} - \sum_{k \in U} x_k\right).\beta\right)^2\right] + \sum_{k \in U}\sum_{l \in U} \sigma_{kl} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_k \pi_l} \tag{3}$$

En équilibrant sur les variables auxiliaires, c'est-à-dire en cherchant à obtenir l'égalité :

$$\sum_{k \in S} \frac{x_k}{\pi_k} = \sum_{k \in U} x_k \tag{4}$$

on réduit donc cette variance. Le tirage équilibré a été utilisé lors de la sélection du précédent échantillon-maître (Guggemos, 2009).

En pratique, si l'idée d'équilibrage sur des variables auxiliaires est ancienne, sa mise en œuvre effective et de façon exacte est plutôt récente, avec l'introduction de la méthode du cube par Deville et Tillé (2004). Auparavant, il était nécessaire de simuler une multitude d'échantillons avant de sélectionner les plus équilibrés (méthodes dites « réjectives »). Une description de l'algorithme du cube est présentée en annexe B. Chauvet et Tillé (2006) ont introduit une implémentation efficace, qui a été utilisée lors du précédent tirage de l'EM.

2.4 Équilibrage pour un échantillonnage stratifié

Étant donné l'organisation de la collecte par l'Insee, décentralisée au niveau des DEM [7], on a envisagé que le tirage de l'échantillon-maître serait stratifié par région. Lors du précédent tirage, Guggemos (2009) procédait à 22 tirages indépendants pour chacune des anciennes régions de France métropolitaine. Cette façon de faire à deux inconvénients majeurs. Premièrement, elle introduit une limitation sur le nombre de variables pouvant être mobilisées pour équilibrer. En effet, il n'est pas possible d'équilibrer s'il y a plus de variables auxiliaires que d'unités à tirer. Or, pour certaines régions peu peuplées, comme la Corse, l'allocation (*i.e* le nombre de zones à tirer) peut être très faible. Deuxièmement, l'équilibre parfait n'est jamais atteint mais toujours approximatif. Accumulées pour les 22 régions, les erreurs d'approximation sont plus importantes que si l'équilibrage est réalisé à l'échelle nationale.

Chauvet (2009) a proposé un algorithme, basé sur la méthode du cube, permettant de conserver un processus de tirage stratifié, mais équilibré sur l'ensemble de la population. Les étapes de cette algorithme sont les suivantes :

1. Réaliser une phase de vol sur chacune des strates.
2. Pour les unités sur lesquelles aucune décision n'a été prise (les probabilités d'inclusion mises à jour appartiennent à l'intervalle $]0, 1[$), une nouvelle phase de vol est effectuée, sans séparation par strate.
3. La phase d'atterrissage est réalisée dans chacune des strates, pour garantir la taille fixe par strate.

2.5 Autocorrélation spatiale

Dans le cadre des approches basées sur un modèle, équilibrer sur des variables auxiliaires n'est pas l'unique stratégie permettant de réduire la variance des estimateurs. On y suppose en effet que les données ne sont plus déterministes mais la réalisation d'une variable aléatoire. Mais on peut également supposer que les résultats observés ne sont pas indépendants de leur localisation géographique, ce que peut révéler l'autocorrélation spatiale.

Une variable est autocorrélée spatialement lorsqu'elle est corrélée avec elle-même selon la position géographique des observations (on peut faire un parallèle avec la notion d'autocorrélation dans le cadre de l'étude des séries temporelles, auquel cas on parle d'autocorrélation temporelle). Cette autocorrélation peut-être positive (les valeurs proches le sont aussi géographiquement), négative (les valeurs opposées s'attirent) ou nulle (les valeurs sont distribuées aléatoirement sur le territoire).

Différents indicateurs permettent de quantifier la présence d'autocorrélation spatiale. Les indices de Moran et de Geary en rendent compte de façon globale. D'autres indicateurs, plus récents, se penchent sur l'autocorrélation spatiale à un niveau plus local, comme les indicateurs LISA [8].

7. pour Division Enquêtes Ménages
8. « Local Indicators of Spatial Association », introduits en 1995 par Luc Anselin

2.6 Méthode du pivot local

En présence d'autocorrélation spatiale positive, on espère mieux prendre en compte la variabilité des situations et améliorer *in fine* la qualité des estimations en répartissant les unités sélectionnées. Une façon d'obtenir cet étalement spatial est d'introduire un mécanisme de répulsion : des unités proches géographiquement doivent avoir peu de chances d'être présentes simultanément dans l'échantillon ; lorsqu'une unité est sélectionnée, les unités voisines doivent voir leur chance d'être sélectionnées réduite.

Grafström *et al.* (2012) ont proposé une méthode, le pivot local, permettant d'offrir un système de répulsion. Le procédé reprend l'approche introduite par Deville et Tille (1998) avec la méthode du pivot. L'idée du pivot est de mettre itérativement à jour les probabilités d'inclusion d'un couple d'unités. À chaque étape, une décision est prise (sélection ou non) pour au moins une unité, ce qui permet de garantir la sélection en au plus N itérations. Une présentation plus précise du pivot est proposée en annexe C.

La méthode du pivot propose une règle de décision. Elle nécessite aussi un critère d'ordre dans le choix des couples d'unité successifs. Le pivot local se base sur la proximité géographique. À chaque étape, une unité est choisie aléatoirement parmi les unités pour lesquelles aucune décision n'a été prise. Puis, l'unité qui lui est la plus proche est également choisie. Les probabilités d'inclusion sont mises à jour pour ces deux unités, et le procédé est ensuite réitéré jusqu'à la sélection de l'échantillon.

Le caractère répulsif de cette méthode vient de la règle de mise à jour des probabilités d'inclusion. En effet, sur les deux unités confrontées, l'une voit sa chance d'être sélectionnée réduite d'autant que l'autre y gagne. L'inclusion double des unités proches géographiquement est donc réduite. On peut ainsi espérer améliorer les estimations. Cependant, rien ne garantit qu'on compense dans ce cas le fait de ne pas équilibrer sur l'information auxiliaire.

2.7 Échantillonnage doublement équilibré

Il semble donc judicieux de cumuler les gains apportées par l'équilibrage sur des variables auxiliaires et l'étalement spatial pour obtenir le meilleur estimateur. On reprend les notations de l'équation 3. En présence d'autocorrélation spatiale positive pour la variable Y, et en supposant que les variables auxiliaires ne captent pas la totalité de cette corrélation, la covariance σ_{kl} sera d'autant plus élevée que les unités k et l seront proches spatialement. On voit donc qu'en réduisant la probabilité d'inclusion des unités voisines, on peut encore réduire la variance de l'estimateur via le terme résiduel.

Grafström et Tillé (2013) ont introduit une méthode permettant d'obtenir cet équilibrage double (sur variables auxiliaires et spatial). Le procédé repose à la fois sur le cube et sur le pivot local. En supposant que p variables auxiliaires sont utilisées pour équilibrer le tirage, à l'issue d'une phase de vol sur n unités, une décision a été prise pour au moins $n - p$ unités. L'idée du cube local est de procéder itérativement à des phases de vol sur $p + 1$ unités, plus proches au sens d'une certaine distance. À l'issue de chacune de ces phases, une décision a été prise sur au moins une unité. Le procédé est arrêté lorsqu'il ne reste plus que p unités pour lesquelles aucune décision n'est prise. Une phase d'atterrissage est alors enclenchée pour aboutir à l'échantillon final.

Dans une première étude s'intéressant à l'apport des méthodes spatiales pour le tirage de l'échantillon-maître, Favre-Martinoz et Merly-Alpa (2016) ont montré que la répartition spatiale, en plus de l'équilibrage, permettait de gagner en précision sur l'estimation des variables d'intérêt (mais sans effet notable sur l'information auxiliaire). Une autre étude, interne à l'Insee, a également montré que plus l'autocorrélation spatiale est élevée, plus les méthodes spatiales apportent des gains en termes de précision conséquents.

2.8 Étalement dans l'espace des variables auxiliaires

Si l'étalement géographique des unités sélectionnées est susceptible d'améliorer la précision des estimations, on peut aussi supposer que conserver la répartition de chacune des variables auxiliaires permet d'avoir un meilleur estimateur. C'est l'idée que défendent Grafström et Lundström (2013) pour qui les échantillons bien répartis dans l'espace des variables auxiliaires sont approximativement équilibrés. En cherchant cette fois à réduire la probabilité d'inclusion double de deux unités proches selon une distance dans l'espace des variables auxiliaires, on espère de nouveau une réduction de la variance via le terme résiduel. Cette approche offre l'avantage de pouvoir traiter le modèle plus général $Y =^t f(X).\beta$ (même si elle est moins exacte pour le modèle linéaire simple). Il existe de multiples façons de définir une distance dans l'espace des variables auxiliaires. Nous présenterons les distances retenues pour cette étude dans la partie décrivant l'approche empirique.

3 Données

Pour comparer les méthodes d'échantillonnage présentées dans la partie précédente, il est nécessaire de récupérer des données disponibles sur la totalité de la population, sur lesquelles nous pourrons faire des estimations. De plus, les méthodes équilibrées ou étalées dans un espace socio-démographiques nécessitent l'usage de variables auxiliaires. Les données mobilisées pour l'étude sont présentées dans cette partie.

3.1 Données communales

Les premières données exploitées sont la composition communale des « Unités Primaires » (UP) du jeu de test. Dans le cadre de cette étude, les travaux sont faits sur un jeu d'UP non définitif, le découpage final du territoire nécessitant un processus de validation, encore en cours au moment de ces travaux. Cependant, les règles de constitution des zones devraient rester inchangées. Les résultats obtenus sur ces données de test doivent donc rester valides. Au total, on dispose de 5 213 unités primaires, partitionnant le territoire de la France métropolitaine (voir graphique 1).

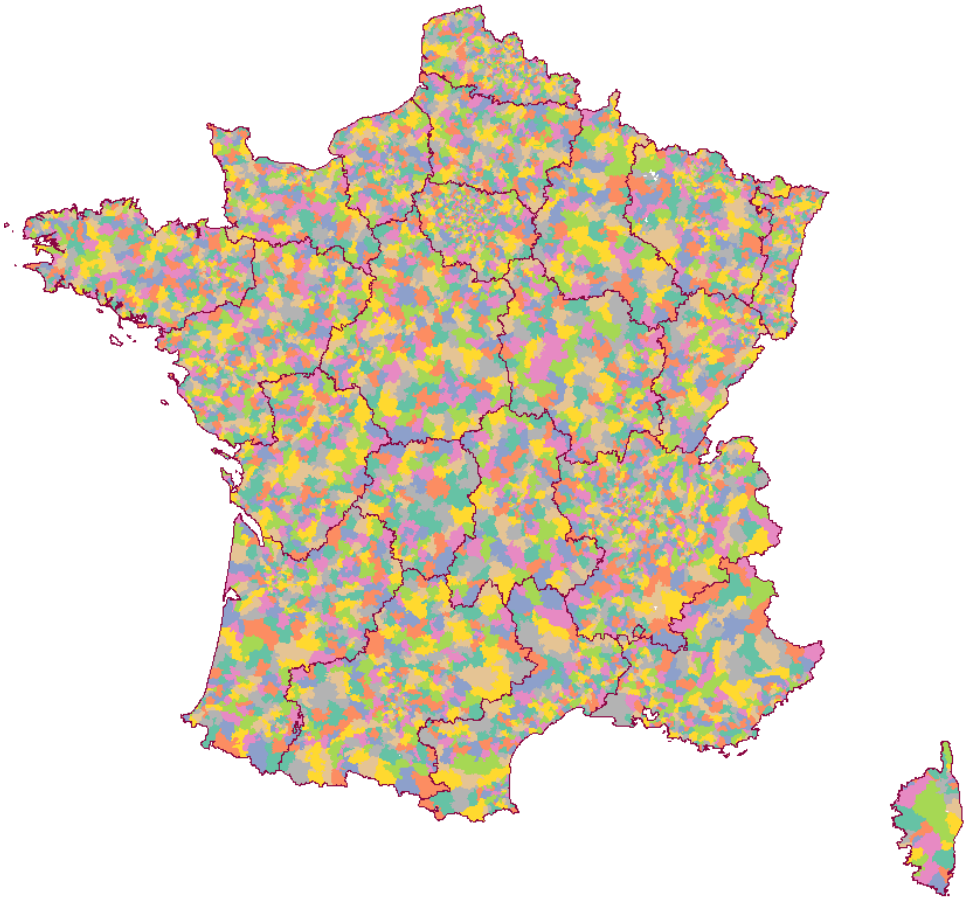
Les 5 213 zones concernent 36 575 communes, soit l'intégralité des communes de France métropolitaine et de Corse au 1er janvier 2014, à l'exception de quelques communes difficiles d'accès, comme certaines îles. Les données supplémentaires mobilisées seront donc choisies de façon à être adaptées à cette géographie.

3.2 Données obtenues grâce au recensement

La première source d'information exploitée est le recensement de la population. Les données d'une année N sont définitivement diffusées l'année $N + 2$, actualisée à la géographie la plus récente. Les données disponibles en géographie 2014 sont donc celles de 2012. Ces données sont récupérées au niveau communal puis agrégée au niveau UP.

De l'exploitation principale du recensement, on récupère les données par commune sur les tranches d'âge et sur le niveau de diplôme des non scolarisés de 15 ans et plus. La répartition de ces variables selon les unités primaires (après jointure et agrégation) sont données dans le tableau 1.

On récupère également des données issues de l'exploitation complémentaire du recensement, théoriquement moins précises puisque non exhaustives. On souhaite exploiter les informations sur la composition des ménages et sur la catégorie socioprofessionnelle de la personne de référence du ménage. Les variables en question sont décrites succinctement dans le tableau 2.



Graphique 1 – Découpage du territoire en 5 213 unités primaires de test (base communale, respect des anciennes frontières régionales) - *Sources : IGN, Insee*

Variable	Description	Min.	Q1	Q2	Q3	Max.
Tranche d'âge (nombre de personnes âgées de ...)						
P12_POP1519	15 à 19 ans	135	342	424	600	33 776
P12_POP2024	20 à 24 ans	85	251	326	498	61 693
P12_POP2539	25 à 39 ans	393	1 077	1 301	1 754	119 291
P12_POP4054	40 à 54 ans	689	1 374	1 609	2 137	73 731
P12_POP5564	55 à 64 ans	553	891	1 007	1 345	41 397
P12_POP6579	65 à 79 ans	271	787	957	1 276	48 752
P12_POP80P	80 ans et plus	65	328	442	623	27 546
Niveau de diplôme (nombre de personnes titulaires d'un diplôme ...)						
P12_NSCOL15P_DIPL0	Aucun diplôme	195	704	916	1 297	45 587
P12_NSCOL15P_CEP	Certificat d'études primaires	123	520	656	864	22 634
P12_NSCOL15P_BEPC	Brevet des collèges	121	276	332	468	20 500
P12_NSCOL15P_CAPBEP	CAP ou BEP	359	1 329	1 508	1 939	45 276
P12_NSCOL15P_BAC	Baccalauréat	476	794	922	1 229	51 401
P12_NSCOL15P_BACP2	Supérieur ou égal à Bac+2	249	550	712	974	49 513
P12_NSCOL15P_SUP	Diplôme du supérieur	128	336	491	839	89 274

Tableau 1 – Variables issues de l'exploitation principale du recensement de la population (données 2012, géographie 2014) - *Source : Insee*

Variable	Description	Min.	Q1	Q2	Q3	Max.
Composition du ménage (nombre de ménages dont la famille principale est composée de ...)						
C12_MENHSEUL	Homme seul	127	300	376	533	57 741
C12_MENFSEUL	Femme seule	148	376	481	727	65 880
C12_MENSFAM	Autre sans famille	5	46	63	93	14 455
C12_MENCOUPSENF	Couple sans enfant	405	851	959	1 233	46 361
C12_MENCOUPAENF	Couple avec enfant(s)	254	816	986	1 288	36 734
C12_MENFAMMONO	Famille monoparentale	64	180	230	362	19 537
Catégorie socioprofessionnelle (nombre de ménages dont la personne de référence est ...)						
C12_MEN_CS1	Agriculteur exploitant	0	16	42	85	504
C12_MEN_CS2	Artisan, Commerçant, Chef d'entreprise	32	134	173	237	8 745
C12_MEN_CS3	Cadre ou exerce une Profession intellectuelle supérieure	35	151	240	448	47 303
C12_MEN_CS4	Profession intermédiaire	115	339	451	635	40 731
C12_MEN_CS5	Employé	86	227	296	460	33 711
C12_MEN_CS6	Ouvrier	61	483	623	827	25 387
C12_MEN_CS7	Retraité	328	950	1 144	1 532	59 035
C12_MEN_CS8	Autre sans activité professionnelle	12	75	108	190	39 209

Tableau 2 – Variables issues de l'exploitation complémentaire du recensement de la population (données 2012, géographie 2014) - *Source : Insee*

On récupère enfin des données sur le territoire, à savoir les tranches d'unité urbaine (la catégorie de taille d'unité urbaine à laquelle appartient chaque commune). Pour chacune de ces catégories, on compte le nombre de résidences principales au sein des UP.

3.3 Données issues de sources fiscales

À partir de sources fiscales, on récupère également des données, de 2014 cette fois, sur le revenu et le logement (tableau 3). On en extrait aussi des données sur le sexe (nombre d'hommes et nombre de femmes).

Variable	Description	Min.	Q1	Q2	Q3	Max.
Revenu fiscal (montant des ...)						
REVENU_TOTAL	Revenu fiscal total	38 229 250	88 696 451	109 484 603	154 919 485	6 516 276 428
SALAIRES	Salaires	19 261 297	51 396 946	67 862 248	99 379 077	4 450 092 986
PENSIONS	Pensions et retraites	13 406 499	26 724 057	31 612 202	45 101 510	1 667 046 521
ALLOC_CHOMAGE	Allocations chômage	1 236 934	2 804 035	3 423 507	4 799 062	250 526 538
BEN_IND	Bénéfices industriels	275 446	1 676 225	2 160 332	2 888 715	91 009 061
BEN_AGRI	Bénéfices agricoles	-1 302 427	197 001	776 009	1 912 215	32 771 113
BEN_NON_COM	Bénéfices non commerciaux	283 001	1 853 028	2 706 676	4 501 193	598 703 762
Logement (nombre de ...)						
NB_RES_PRINC	Résidences principales	2 500	2 689	3 011	4 196	235 818
NB_RES_SEC	Résidences secondaires	0	49	151	420	31 960
NB_LOG_OCCAS	Logements occasionnels	0	5	10	22	3 854
NB_LOG_VAC	Logements vacants	25	196	278	432	28 966
SURFACE_TOTALE	Surface totale (m^2) des logements	170 945	268 368	300 369	387 822	14 915 233
VALEUR_LOCATIVE	Valeur locative (euros) des logements	705 084	1 367 359	1 824 545	2 886 164	159 418 355
PROPRIETAIRES	Propriétaires	467	1 963	2 189	2 754	83 042
LOCATAIRES	Locataires	231	580	797	1 404	146 472

Tableau 3 – Variables issues de sources fiscales (données 2014, géographie 2014) - *Source : Insee*

3.4 Données complémentaires

On récupère le nombre de HLM par communes en 2015, disponible depuis la loi SRU.

On récupère également des données hors champ a priori, afin d'observer le comportement des différentes méthodes de sondage sur ces données. En l'occurrence, on se procure les résultats du premier tour des dernières élections présidentielles, dont :

— INSCRITS : nombre d'inscrits sur les listes électorales ;
— VOTANTS : nombre de votants ;
— ABSTENTIONS : nombre d'abstentionnistes ;
— BLANCS : nombre de votes blanc ;
— NULS : nombre de votes nuls.

On récupère enfin le nombre de voix pour chacun des candidats.

3.5 Données géographiques

On va par ailleurs chercher à répartir spatialement les échantillons, ce qui nécessite des coordonnées géographiques. Celles-ci sont constituées à partir de la base de données GEOFLA® fournie par l'IGN [9]. La version récupérée date de 2014, avec un découpage au niveau communal. Pour chaque commune, vue comme un polygone, on récupère les coordonnées (abscisse X et ordonnée Y) de son centroïde. Ces coordonnées sont issues de la projection Lambert-93, basée sur le système géodésique RGF93 [10].

Pour adapter ces coordonnées au niveau unités primaires, on choisit d'attribuer à chaque UP les coordonnées de sa commune la plus peuplée. On aurait également pu envisager de calculer le barycentre de la zone, pondéré par le nombre de résidence principales.

3.6 Typologie des variables

L'ensemble des variables récupérées sont disponibles sur l'intégralité de la population et vont donc permettre de construire des méthodes d'échantillonnage qu'on espère plus précise, ou être utiles pour comparer les différentes méthodes entre elles (constituant en quelque sorte des variables « témoins »). La question qui se pose est de savoir quelles informations privilégier pour construire les modèles. Théoriquement, une variable intéressante dans le processus d'échantillonnage (pour équilibrer ou au sein d'une distance) est une variable corrélée avec les variables d'intérêt pour les enquêtes en aval.

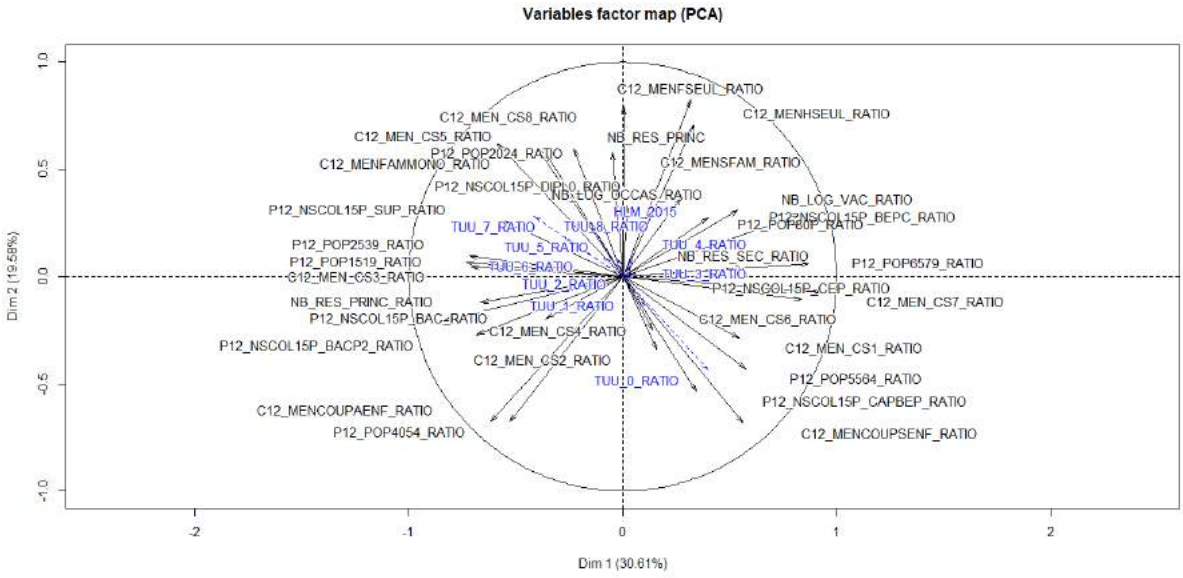
On décide, dans le prolongement des travaux de Guggemos (2009) et Favre-Martinoz et Merly-Alpa (2016), et au vu des enquêtes concernées (listées en annexe A), de mobiliser a minima les informations sur le territoire, le revenu total, les tranches d'âge, la composition des ménages, le nombre de propriétaires et le nombre de logements sociaux. Cela fait un total de 25 variables, qui monte à 26 en prenant en considération les probabilités d'inclusion (qui sont proportionnelles au nombre de résidences principales). Les tirages équilibrés sont limités sur le nombre de variables auxiliaires à mobiliser. Il ne semble donc pas raisonnable d'ajouter des variables supplémentaires. On dira dans la suite du rapport que ces variables appartiennent à la famille « équilibrage ».

Il n'existe en revanche pas de limite de dimension pour les méthodes utilisant une distance. On choisit donc d'utiliser en supplément les informations sur le niveau de diplôme, la catégorie socioprofessionnelle et le type de résidence (principale, secondaire, ...) pour ces méthodes. Les variables en question, servant dans les distances mais pas à équilibrer, sont classées dans la famille « distance ». L'ensemble des autres variables est appelé famille « d'intérêt ».

3.7 Lien entre les variables

On souhaite éclairer les liens que peuvent avoir entre elles les différentes données sociales récupérées. On effectue pour cela une analyse en composante principale sur les variables issues du recensement : l'âge, le revenu, la CSP, le niveau de diplôme. On observe aussi comment se positionnent les tranches d'unité urbaine ; on les projette en variables supplémentaires de l'analyse (graphique 2).

9. Institut national de l'information géographique et forestière
10. Réseau géodésique français 1993, système géodésique officiel en France métropolitaine



Graphique 2 – Nuage des variables d'une ACP sur quelques variables sociales - *Source : Insee*

La moitié de l'inertie est résumée dans le premier plan factoriel. Le premier axe, qui porte 30 % de l'inertie, oppose des zones plus âgées (sur la droite) à des zones plus jeunes (sur la gauche). Les zones corrélées positivement avec le premier axe sont également moins diplômées que les zones qui lui sont corrélées négativement. Enfin, Les zones à gauche sont souvent plus composées de catégories socioprofessionnelles supérieures que celles de droite.

En regardant la répartition des zones sur ce premier axe (graphique 3), on constate que les zones qui lui sont corrélées négativement sont plutôt les zones urbaines au sens large (Paris, Toulouse, Lyon, ...) ainsi que la frontière nord-est du pays. Les zones corrélées positivement forment une diagonale du sud-ouest vers le nord-est, avec en supplément la Corse et les cœurs de la Bretagne et de la Normandie. Ces zones sont plus généralement des zones rurales, éloignées des pôles urbains. On peut noter une autocorrélation spatiale positive sur ce premier axe : les zones proches géographiquement ont tendance à se ressembler.

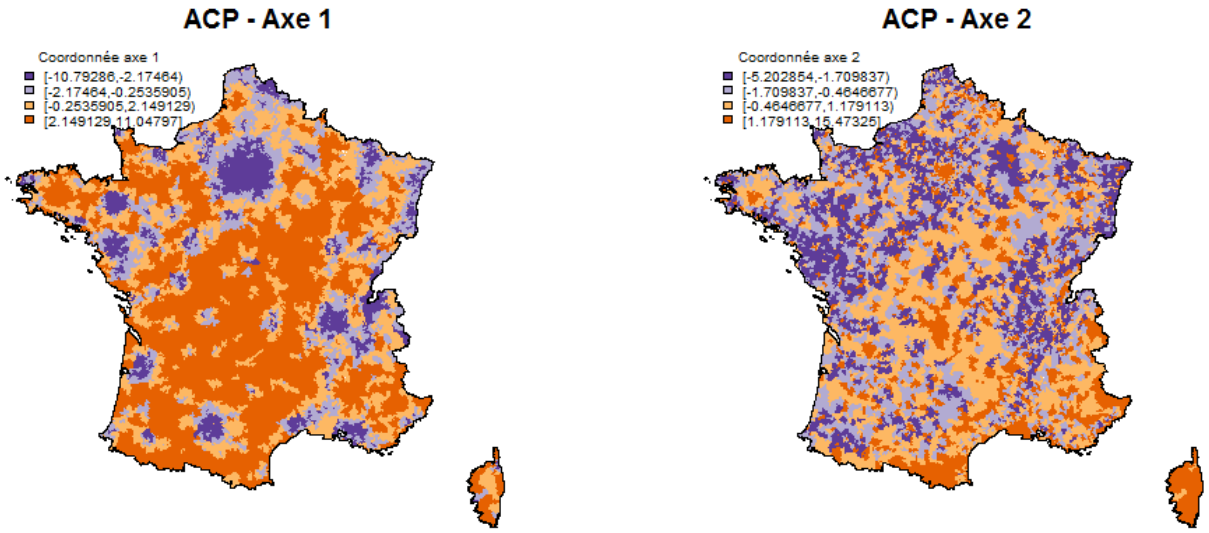
En observant l'indice de Moran sur les variables d'équilibrage (graphique 13, en annexe D), on constate que l'autocorrélation spatiale correspond à une réalité. Pour toutes ces variables, les unités proches géographiquement ont significativement des valeurs proches.

Le second axe porte près de 20 % de l'inertie. Il oppose essentiellement des zones ou les personnes ont plus tendance à vivre seules (corrélation positive) qu'en famille (avec ou sans enfants). Les premières sont également plus urbaines et possèdent plus de logements HLM. Ces zones semblent correspondre aux zones périurbaines. Mais, si l'autocorrélation spatiale existe encore, elle est beaucoup moins nette pour cet axe.

A priori, on devrait donc améliorer les estimations sur le niveau de diplôme ou la catégorie socioprofessionnelle, ces concepts étant corrélés avec les variables « d'équilibrage ».

4 Approche empirique

Nous avons présenté les méthodes d'échantillonnage envisagées pour la sélection du nouvel échantillon-maître, ainsi que les données disponibles, que ce soit pour construire des modèles ou les comparer entre eux. Dans cette partie, nous exposons l'implémentation mise en œuvre pour comparer les différentes méthodes.



Graphique 3 – Coordonnées des zones sur les deux premiers axes de l'ACP - *Sources : IGN, Insee*

4.1 Plan de sondage

Ici, la base de sondage est constituée de $N = 5213$ unités primaires. On cherche à sélectionner une sous-population de taille $n = 567$ (100%), 508 (87,5%), 439 (75%) ou 306 (50%) [11] qui soit la plus représentative possible de la population totale. L'objectif est d'estimer des indicateurs sur un certain nombre de variables Y.

Chacune des unités primaires étant une agrégation d'unités secondaires (qui seront tirées *in fine* pour les enquêtes ménages), on tient compte de l'hétérogénéité en leur attribuant des probabilités d'inclusion proportionnelles à leur taille. L'unité finale exploitable est la résidence principale [12]. On note res_k le nombre de résidences principales de l'unité primaire $k \in U$. On définit alors la probabilité d'inclusion $\pi_k = n \times \frac{res_k}{\sum_{l \in U} res_l}$.

Pour certaines unités, cette grandeur est potentiellement supérieure à 1. On dit qu'elles sont « exhaustives ». On fixe alors arbitrairement leur probabilité d'inclusion à 1 (ce qui revient à les inclure automatiquement dans l'échantillon final). On déduit ensuite ces unités du nombre n qu'on souhaite échantillonner et on recalcule de nouveau les probabilités d'inclusion. On réitère ce processus jusqu'à ce que toutes les unités aient une probabilité d'inclusion inférieure ou égale à 1.

On introduit également une étape de stratification de façon à garantir une quantité d'unités par région proportionnelle à la taille de ces régions. De nouveau, le réseau d'enquêteur existant explique ce besoin, mais également la gestion des enquêtes par les différents établissements régionaux de l'Insee. On définit une allocation par région (*i.e* le nombre d'unités devant y être tirées) selon la formule $n_r \approx n \times \frac{\sum_{k \in r} res_k}{\sum_{k \in U} res_k}$ pour la région r.

11. En effet, pour améliorer le processus de collecte, la question s'est posée de couvrir une unité primaire par plus d'un enquêteur

12. Logement occupé de façon habituelle et à titre principal par une ou plusieurs personnes qui constituent un ménage (définition Insee)

4.2 Estimation des quantiles

Si on souhaite retenir une méthode de sondage qui a de bonnes propriétés pour les estimations du total d'un certain nombre de variables d'intérêt, on veut aussi avoir une idée de son comportement sur la distribution de ces variables. Il peut en effet arriver que l'estimation d'un quantile soit une information pertinente pour une enquête, par exemple le premier décile de revenu. Dans le cadre de cette étude, on choisit de se focaliser sur les trois quartiles. Cela suppose de disposer d'une méthode pour les estimer.

Étant donné une répartition empirique d'une variable aléatoire, il existe un très grand nombre de façon d'en calculer des quantiles empiriques. Hyndman et Fan (1996) en exposent neuf, qu'ils considèrent devoir appartenir à tout logiciel faisant du traitement statistique des données. Les trois premiers sont basés sur une fonction de distribution empirique discontinue, tandis que les six suivants se fondent sur une extrapolation linéaire, continue par morceau. Dans le cadre de cette étude, c'est le calcul des quantiles de type 4 qui est utilisé même s'il ne présente pas les meilleures propriétés, la généralisation avec des poids inégaux étant plus aisée.

Pour une population de taille n, avec des poids (w_i), et pour une variable X ordonnée (on note $w_{(i)}X_{(i)}$ les valeurs ordonnées), on commence par définir une fonction de répartition par l'ensemble des couples $(X_{(i)}, \sum_{j=1}^{i} w_{(j)})$, qu'on relie par un segment. Pour un ordre $\alpha \in]0,1[$, on définit alors le quantile empirique comme l'inverse de cette fonction, à l'exception des ordres inférieurs à $w_{(1)}$ pour lesquels on retourne 0.

4.3 Estimation par Monte-Carlo

Pour estimer les différents indicateurs permettant de comparer les plans de sondage envisagés, on utilise le procédé de Monte-Carlo. On réalise pour chaque méthode $Q = 10\ 000$ simulations. Pour un estimateur t_y (du total ou d'un quantile d'une variable Y), on estime alors son biais relatif Monte-Carlo par :

$$\hat{BR}_{MC}(\hat{t_y}) = \frac{1}{Q}\sum_{q=1}^{Q}\left(\frac{\hat{t_{y(q)}} - t_y}{t_y} \times 100\right) \tag{5}$$

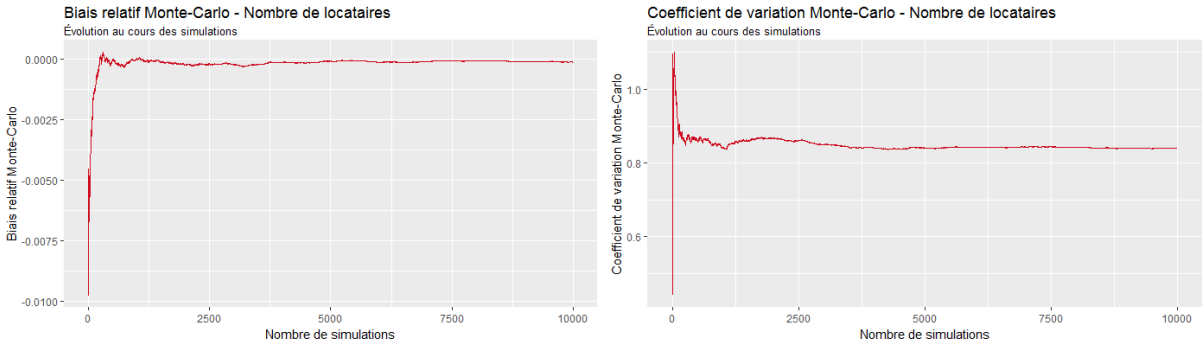
Le coefficient de variation est lui estimé par :

$$\hat{CV}_{MC}(\hat{t_y}) = \frac{1}{Q}\sum_{q=1}^{Q}\left(\frac{\left[\hat{t_{y(q)}} - \frac{1}{Q}\sum_{r=1}^{Q}\hat{t_{y(r)}}\right]^2}{t_y} \times 100\right) \tag{6}$$

et l'erreur quadratique moyenne par :

$$\hat{EQM}_{MC}(\hat{t_y}) = \frac{1}{Q}\sum_{q=1}^{Q}\left(\hat{t_{y(q)}} - t_y\right)^2 \tag{7}$$

Le nombre Q de simulations est choisi de façon à réduire l'erreur d'estimation commise. Cette erreur est inversement proportionnelle à $\sqrt{Q}$. En choisissant $Q = 10\ 000$, on espère ainsi réduire l'erreur d'estimation d'un facteur 100. Le graphique 4 illustre la convergence des estimateurs Monte-Carlo du biais relatif et du coefficient de variation relatif.



Graphique 4 – Évolution des estimateurs Monte-Carlo du biais relatif et du coefficient de variation relatif au cours des simulations, pour l'estimateur Horvitz-Thompson du nombre de locataires avec la méthode du cube

4.4 Plans testés

La méthode de référence, utilisée lors du précédent tirage, est une sélection, région par région, équilibrée sur variables auxiliaires CUBE (renommé *cube*, simple ou classique, par la suite). On essaye également un tirage stratifié par région mais équilibré au niveau national CUBESTRATIFIED (Chauvet, 2009), renommé *cube stratifié*. On compare dans un premier temps ces méthodes avec les méthodes spatialement réparties, en utilisant les coordonnées géographiques, X et Y, des unités primaires : le cube doublement équilibré région par région (*cube local*) ; le cube doublement équilibré, eu niveau national (*cube local stratifié*) ; les trois implémentations du pivot local. Toutes ces méthodes sont implémentées dans la librairie R « BalancedSampling » (Grafström et Lisic, 2016).

Méthode	Abbrév.	Description	Références
Cube local	LCUBE	Tirage doublement équilibré, région par région	Grafström et Tillé (2013)
Cube local stratifié	LCUBESTRATIFIED	Tirage doublement équilibré, avec une phase de vol nationale	Chauvet (2009)
Pivot local 1	LPM1	Plus proches voisins mutuels	Grafström *et al.* (2012)
Pivot local 2	LPM2	Plus proches voisins unilatéral	%
Pivot local 2 - Arbre KD	LPM2_KDTREE	Plus proches voisins unilatéral avec arbre KD	

Tableau 4 – Les cinq implémentations des méthodes de tirage faisant intervenir une distance

On complète ces premières comparaisons avec un ensemble de jeu de distances faisant intervenir à la fois des informations sociales, économiques et démographiques, et la distance géographique. En effet, on veut à la fois profiter des gains apportés par l'étalement dans l'espace des variables auxiliaires, et de la part inobservable captée éventuellement par la distance géographique (en présence d'autocorrélation spatiale positive).

La première distance utilisée, la plus simple, est la distance euclidienne sur les variables brutes (*i.e* sans transformations). On travaille également sur cette distance en lui adjoignant le couple de coordonnées (X, Y). Une première façon de transformer ces variables et de les harmoniser est de les centrer et de les réduire. Pour une variable Z et un individu i, étant donné une population totale de n individus, on considère ainsi la variable transformée :

$$\tilde{Z}_i = \frac{Z_i - \bar{Z}}{S_Z} \text{ avec } \bar{Z} = \frac{1}{n}\sum_{j=1}^{n} Z_j \text{ et } S_Z = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(Z_j - \bar{Z})^2} \tag{8}$$

À cette transformation, on essaye également d'adjoindre le couple de coordonnées (X, Y) ainsi que le couple de coordonnées transformé $(\tilde{X}, \tilde{Y})$.

En normalisant les variables auxiliaires, on permet d'égaliser leur influence, afin qu'une variable prenant de grandes valeurs ne l'emporte pas sur les autres. Mais il peut aussi être souhaitable d'ôter l'influence qu'elles peuvent avoir les unes sur les autres. L'analyse en composante principale permet cela. On effectue donc une analyse en composante principale sur les variables appartenant aux deux familles « équilibrage » et « distance ». On récupère ensuite la projection sur les axes principaux dans le nuage des individus. On effectue de nouveau cette analyse en ajoutant cette fois les coordonnées géographiques (X, Y). On s'intéresse alors à deux types de coordonnées : celles sur les deux premiers axes principaux et celles sur l'ensemble des axes (avec autant de coordonnées que le nombre de variables utilisées dans l'ACP). Pour l'analyse où la géographie n'intervient pas, on essaye également les méthodes auxquelles on adjoint le couple (X, Y) et sa version centrée réduite $(\tilde{X}, \tilde{Y})$.

Enfin, en plus de dé-corréler les variables, on souhaite que cette dernière version soit également normalisée. On utilise pour cela la distance de Mahalanobis. Pour un ensemble de variables X, de moyenne empirique μ et de matrice de variance-covariance empirique Σ, la distance de Mahalanobis entre deux unités k et l est définie par :

$$dist_M(k,l) = \sqrt{{}^t(X_k - X_l)\Sigma^{-1}(X_k - X_l)} \tag{9}$$

Une façon de la calculer est de transformer les coordonnées initiales et de prendre simplement la distance euclidienne dans ce nouvel espace. Cette transformation consiste à calculer la racine carrée de Σ (qui est symétrique définie positive) et de la multiplier à gauche des données centrées réduites. L'unité k est alors définie par le vecteur $\Sigma^{-\frac{1}{2}}(X_k - \mu)$. C'est ce procédé qui est utilisé pour les simulations de cette étude.

Une autre façon de faire est de passer par la projection des unités dans l'ensemble des plans factoriels résultants d'une analyse en composantes principales, à condition de réduire les coordonnées sur chaque axe. En effet, le processus de construction des axes factoriels implique une décorrélation (ces axes étant orthogonaux selon des variables centrées réduites). L'étape de réduction permet de normaliser. Ces coordonnées sont équivalentes au cas précédent, à un changement de base près.

De nouveau, on effectue deux transformations : la première avec uniquement l'information auxiliaire, et la seconde avec la géographie. On adjoint également à la première transformation les deux versions du couple (X, Y).

Toutes ces distances faisant intervenir de l'information auxiliaire sont testées sur deux versions des variables : une première version avec des totaux et une seconde version avec des ratios. En effet, la plupart des variables sont catégorielles et correspondent à la grandeur de la population appartenant à chaque catégorie. Or, les probabilités d'inclusion ayant déjà une influence dans la sélection des unités les plus peuplées, on attend des méthodes d'étalement qu'elles permettent une sélection plus variée des zones selon la structure de leur population.

En résumé, on effectue donc des simulations avec la méthode du cube et du cube stratifié. On mobilise aussi les méthodes avec distance sur (X, Y), centré réduit ou non, et sur 17 distances « mixtes », à la fois sur les totaux et les ratios des variables auxiliaires, ce qui correspond à 36 distances différentes (résumées dans le tableau 5). Chacune des distances est testées pour les cinq implémentations de méthodes de tirage avec distance, ce qui fait 180 simulations. Au total, 182 processus d'échantillonnage sont comparés.

4.5 Mesure de l'étalement spatial

Pour rendre compte de l'étalement spatial d'un processus d'échantillonnage, il est possible d'utiliser un indicateur basé sur les polygones de Voronoï. Étant donné un ensemble de points sur un plan, on procède à un découpage de l'espace en traçant l'ensemble des points équidistants entre deux points. On obtient ainsi un ensemble de zones, chacune d'entre elles contenant un unique point (il y a donc autant de zones que de points initiaux). Pour un autre point du plan,

Distance	Variables transformées		Géographie adjointe	
	Variables auxiliaires	Variables auxiliaires et (X, Y)	(X, Y)	(X, Y) Centré réduit
		Sans information auxiliaire		
Euclidienne			XY	XY_NORM
		TOTAL		
Euclidienne	EUCL_TOTAL		EUCL_TOTAL_XY	
Euclidienne centré réduit	EUCL_TOTAL_NORM		EUCL_TOTAL_NORM_XY	EUCL_TOTAL_XY_NORM
ACP axes 1 et 2	ACP_TOTAL	ACP_TOTAL_MIX	ACP_TOTAL_XY	ACP_TOTAL_XY_NORM
ACP tous axes	ACP_TOTAL_FULL	ACP_TOTAL_MIX_FULL	ACP_TOTAL_FULL_XY	ACP_TOTAL_FULL_XY_NORM
Mahalanobis	MAHAL_TOTAL	MAHAL_TOTAL_MIX	MAHAL_TOTAL_XY	MAHAL_TOTAL_XY_NORM
		RATIO		
Euclidienne	EUCL_RATIO		EUCL_RATIO_XY	
Euclidienne centré réduit	EUCL_RATIO_NORM		EUCL_RATIO_NORM_XY	EUCL_RATIO_XY_NORM
ACP axes 1 et 2	ACP_RATIO	ACP_RATIO_MIX	ACP_RATIO_XY	ACP_RATIO_XY_NORM
ACP tous axes	ACP_RATIO_FULL	ACP_RATIO_MIX_FULL	ACP_RATIO_FULL_XY	ACP_RATIO_FULL_XY_NORM
Mahalanobis	MAHAL_RATIO	MAHAL_RATIO_MIX	MAHAL_RATIO_XY	MAHAL_RATIO_XY_NORM

Tableau 5 – Résumé des 36 distances testées lors des simulations

présent à l'intérieur d'une cellule, le point central de cette cellule est le point initial dont il est le plus proche. Un point présent sur une frontière est à équidistance des points centraux des zones adjacentes à cette frontière.

Appliqué à ce cadre, on construit un diagramme de Voronoï à partir des unités sélectionnées dans l'échantillon. Pour chacune de ces n unités tirées, on calcule la somme δ_i des probabilités d'inclusion des unités pour lesquelles elle est l'unité tirée la plus proche. Ce qui revient, pour les n cellules du diagramme de Voronoï, à calculer la somme de probabilités d'inclusion des unités incluses dans la cellule. En espérance, δ_i est égal à 1. On définit alors l'indicateur « de Voronoï » par :

$$\Delta = \frac{1}{n-1}\sum_{i\in S}(\delta_i - 1)^2 \tag{10}$$

qui correspond à la variance empirique de δ_i.

Plus un échantillon est étalé spatialement, plus cette variance sera faible.

5 Résultats

Au total, on simule 182 méthodes d'échantillonnage, pour 4 tailles d'échantillon différentes. On souhaite les comparer selon leur comportement sur l'estimateur Horvitz-Thompson du total en priorité, puis sur l'estimation des trois quartiles. Pour ces quatre estimateur, on observe trois indicateurs, le biais relatif (%), le coefficient de variation (%) et l'erreur quadratique moyenne. Dans l'idéal, on souhaiterait qu'une méthode se démarque sur tous ces indicateurs.

Pour faire émerger les meilleures méthodes, une première approche, un peu rudimentaire, a consisté à regarder le nombre d'apparition de chaque méthode dans différent palmarès. Par exemple, dans le cas du coefficient de variation de l'estimateur du total, un classement des méthodes est réalisé sur chacune des variables disponibles (la meilleure méthode est celle ayant le coefficient de variation le plus faible). En se donnant un seuil, par exemple le top 10, on regarde combien de fois chaque méthode appartient à ce palmarès. Une bonne méthode doit alors être présente plus souvent que les autres dans ce palmarès.

Cette approche permet de repérer les méthodes meilleures sur le plus grand nombre de variables. En revanche, elle occulte l'aspect quantitatif d'une éventuelle domination. Une méthode peut en effet être souvent meilleure, mais être très mauvaise sur certaines variables. Pour compléter cette analyse, on a produit un autre indicateur destiné à palier ce manque. Il s'agit, en fonction de l'indicateur, de prendre la norme euclidienne de chacune des méthodes selon les valeurs prises pour chacune des variables. En reprenant l'exemple du coefficient de variation du total, une méthode est ainsi définie par un coefficient de variation sur chacune des variables disponibles. On en définit la norme en prenant la racine carrée de la somme des carrés de chacun des coefficients de variation. Ainsi, une méthode est globalement moins dispersée selon cet

indicateur, si elle a une norme plus faible que les autres.

Ces deux indicateurs ont permis une première identification des méthodes les plus performantes (il y sera fait référence de façon occasionnelle). Mais ils ont pour principal désavantage d'être difficile à restituer dans leur globalité. Pour permettre une interprétation plus aisée, une troisième approche a été développée qui sera principalement utilisée ici. Celle-ci consiste à effectuer une analyse en composantes principales (ACP) pour un indicateur considéré. Prenons l'exemple du coefficient de variation de l'estimateur Horvitz-Thompson du total. On effectue une ACP où les individus sont les 182 méthodes d'échantillonnage testées. Chacune de ses méthodes est caractérisée par la valeur du coefficient de variation estimé pour les 77 variables socio-démographiques disponibles, qu'on peut interpréter comme des coordonnées.

5.1 Estimation du total

On compare dans un premier temps tous les modèles en bloc, afin de faire émerger les méthodes globalement les meilleures sur l'estimateur Horvitz - Thompson du total. En particulier, on recherche des méhodes qui feraient mieux que la méthode de référence, c'est-à-dire la méthode du cube classique, équilibré strate par strate (CUBE). On comparera plus finement les méthodes retenues dans un second temps.

On s'intéresse d'abord au biais relatif (en %). L'estimateur Horvitz-Thompson est théoriquement sans biais, on s'attend donc à ce que les valeurs obtenues soient proches de 0. Quelque soit la méthode et la variable étudiée, le biais relatif est en valeur absolu inférieur à $0,4\%$. Ce résultat est confirmé par l'observation de la norme de chaque méthode dans l'espace des biais relatifs (tableau 6).

Min	Median	Moyenne	Max
0.08731	0.20433	0.22923	0.59605

Tableau 6 – Moments de la norme euclidienne des méthodes d'échantillonnage dans l'espace des biais relatifs

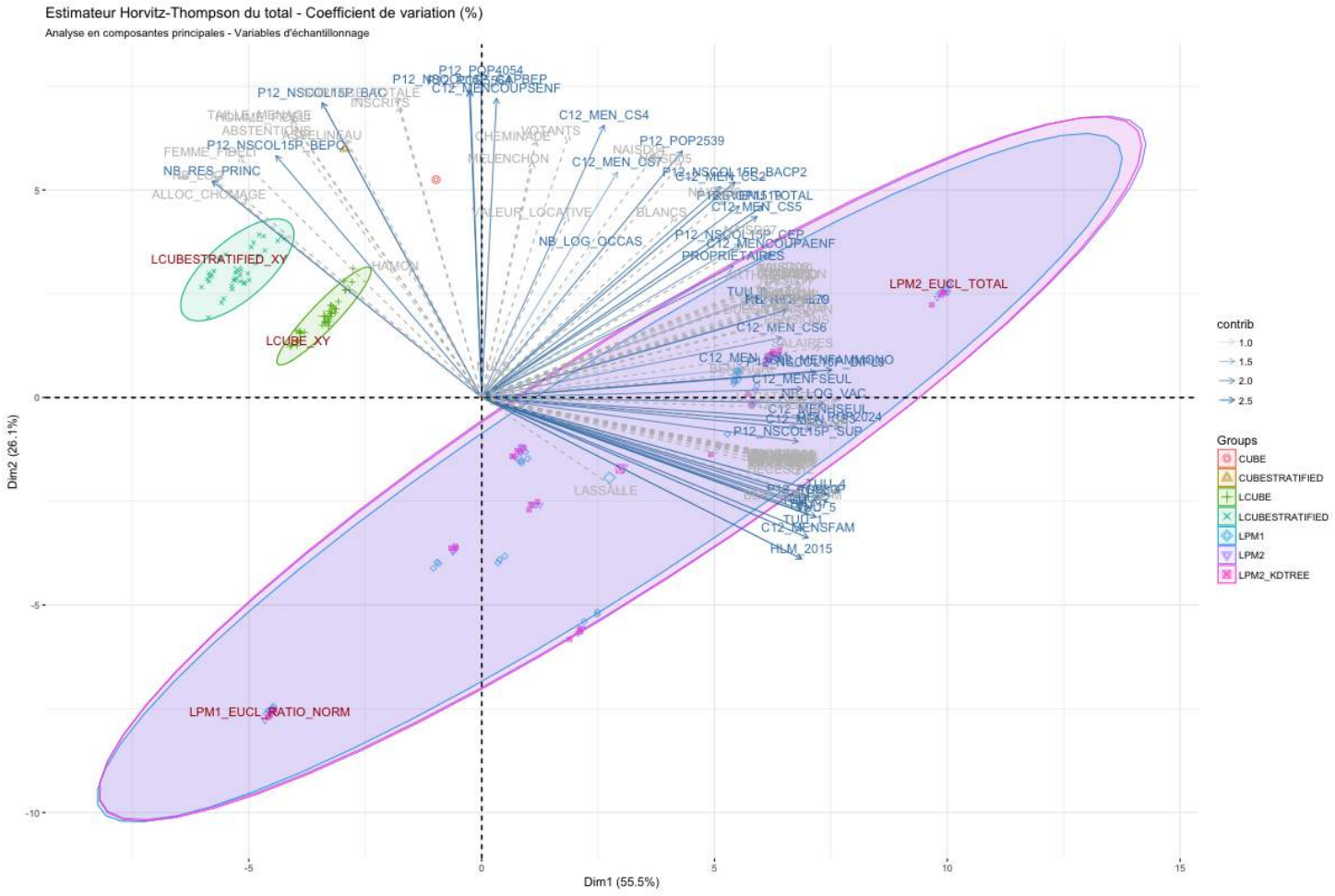
De même, l'observation des palmarès ou en composantes principales, ne fait pas émerger de méthode en particulier. Il est notamment difficile de différencier les grandes familles de méthodes (cube, cube local ou pivot local). Les différentes méthodes ont le comportement attendu en termes de biais dans l'estimation Horvitz-Thompson du total.

On souhaite alors comparer la précision de cette estimation. En effet, plus une méthode sera précise, plus le risque d'obtenir un échantillon éloigné des vrais totaux sera faible. On s'intéresse pour cela au coefficient de variation relatif (en %). En effet, étant donné la faiblesse des biais estimés, les résultats sur cet indicateur sont proches de ceux sur l'erreur quadratique moyenne.

On cherche à visualiser le comportement en termes de précision des différentes méthodes en utilisant une analyse en composantes principales (graphique 5). Les deux premiers axes portent plus de 80 % de l'inertie. Les variables contribuant le plus au premier axe sont majoritairement liées à la tranche d'unité urbaine, à la composition du ménage et à l'âge. C'est-à-dire des variables utilisées pour l'équilibrage. Si l'âge contribue également au second axe, c'est aussi le cas de façon importante pour le diplôme et la catégorie socioprofessionnelle, utilisées uniquement dans les distances.

Toutes les coordonnées initiales sont positives (par définition du coefficient de variation). Une méthode ayant une coordonnée négative sur un axe est plus précise pour les variables ayant fortement contribué à cet axe. Elle est moins précise lorsque cette coordonnée est positive.

Les trois familles de méthodes (cube local LCUBE, cube local avec équilibrage national LCUBESTRATIFIED et pivot local LPM) suivent trois directions distinctes. Pour les méthodes



Graphique 5 – Analyse en composantes principales sur les coefficients de variation de l'estimateur Horvitz-Thompson du total, sur les variables ayant servi à l'échantillonnage, pour les 182 méthodes testées

du pivot, la méthode de recherche du plus proche voisin (LPM1, LPM2 ou LPM2_KDTREE) semble peu déterminante par rapport au choix de la distance. En effet, la distance utilisée a un fort impact sur le comportement de ces méthodes, qui sont moins concentrées dans le plan factoriel que les différentes variantes du cube local.

Les méthodes les plus éloignées dans la direction sud-ouest sont a priori globalement plus précises (c'est-à-dire sur l'ensemble des variables). Ainsi, les méthodes du cube et du cube stratifié sont globalement moins bonnes que leurs variantes locales. Au sein de ces dernières, les méthodes où les coordonnées géographiques (X et Y) ont été adjointes forment des groupes à part, meilleurs sur l'axe 1 (et sur l'axe 2 pour le cube local strate par strate LCUBE)

Pour les méthodes du pivot, un groupe se détache pour 5 distances (les 3 algorithmes du pivot y sont présents) :

1. ACP_RATIO_FULL : projection sur tous les axes d'une ACP construite sur les ratio de variables ;
2. ACP_RATIO_FULL_XY_NORM : projection sur tous les axes d'une ACP construite sur les ratio de variables, et coordonnées géographiques centrées et réduites ;
3. ACP_RATIO_MIX_FULL : projection sur tous les axes d'une ACP construite sur les ratio de variables et incluant les coordonnées géographiques ;

4. EUCL_RATIO_NORM : distance euclidienne sur les ratios de variables, centrés et réduits ;
5. EUCL_RATIO_XY_NORM : distance euclidienne sur les ratios de variables et les coordonnées géographiques, centrés et réduits ;

Il s'agit donc de méthodes basées sur les ratios de variables, où la géographie peut intervenir à condition d'avoir été centrée et réduite. Les méthodes du pivot local étalées uniquement sur la géographie (LPM_XY) ont quant à elles des coordonnées positives sur les deux axes, mais plus faibles que les cubes simples (CUBE et CUBESTRATIFIED) sur l'axe 2. L'étalement géographique est ainsi susceptible d'être plus précis pour certains indicateurs que l'équilibrage sur variables auxiliaires.

Parmi les meilleurs méthodes du pivot, on choisit de retenir celle utilisant la distance euclidienne centrée réduite (EUCL_RATIO_NORM), la plus simple à construire. On se focalise sur la première version du pivot (LPM1) qui est la plus rigoureuse. On souhaite comparer ce plan de sondage avec les méthodes équilibrées sur variables auxiliaires (CUBE et CUBESTRATIFIED) et étalées sur la géographie (LCUBE_XY et LCUBESTRATIFIED_XY). En effet, la distance ayant peu d'impact sur la précision pour le cube local, on retient de nouveau la plus simple : la distance géographique. Elle offre aussi l'avantage de prendre en compte la dimension spatiale des données.

5.2 Comparaison des méthodes retenues

On propose à présent d'analyser plus en détail les cinq plans de sondage retenus. On observe pour cela la valeur du coefficient de variation de l'estimateur Horvitz-Thompson du total pour chacune de ces méthodes sur l'ensemble des variables disponibles (tableau 7) [13]

On constate tout d'abord que les ordres de grandeurs des coefficients de variation sont les mêmes pour les cinq méthodes. Une variable sur laquelle il est difficile d'être précis la sera pour toutes les méthodes. C'est par exemple le cas des tranches d'unité urbaine, du nombre de HLM ou du nombre de résidences secondaires. Ces variables sont nulles ou prennent des valeurs très faibles pour une majorité des unités primaires. On trouve deux exceptions, le nombre de résidences principales et le nombre de logements, pour lesquelles le pivot est nettement plus précis que les autres. Ces deux variables sont très fortement liées aux probabilités d'inclusion, que le pivot vérifie de façon exacte (contrairement au cube, qui ne les respecte qu'en espérance).

À l'exception des tranches d'unité urbaine, pour lesquelles le cube stratifié (CUBE_SD) peut être meilleur, et du nombre de vote nuls, dominé par le cube strate par strate (CUBE), la meilleure méthode est soit le cube local équilibré nationalement (LCUBE_SD_XY), soit le pivot (LPM1_ERN). Sur un total de 77 variables, le pivot est 51 fois la méthode la plus précise et le cube local stratifié 19 fois. Le pivot est en particulier performant sur les variables de « distance » (qui ont été mobilisées dans sons processus d'échantillonnage) et les variables « d'intérêt ». On peut supposer que les informations supplémentaires intégrées dans la distance permettent d'améliorer encore la précision sur des variables tierces. Le pivot est parfois meilleur que les méthodes du cube sur les variables d'équilibrage, pourtant censées les estimer de façon exacte. Il faut garder à l'esprit que le cube se termine par une phase d'atterrissage où les contraintes sont progressivement relâchées.

Cependant, le pivot est 15 fois la moins bonne méthode alors que le cube local (stratifié ou non) ne l'est jamais. Le pivot est systématiquement moins précis sur les tranches d'unités urbaines, exceptée celle de Paris, avec un coefficient de variation 2 à 3 fois supérieur que celui du cube stratifié (local ou non). Il est également fortement imprécis sur le nombre de HLM par rapport aux autres méthodes. Dans une moindre mesure, il est aussi moins bon sur le nombre de

13. Pour alléger les notations, on a renommé la méthode LPM1_EUCL_RATIO_NORM en LPM_ERN, CUBESTRATIFIED en CUBE_SD et LCUBESTRATIFIED_XY en LCUBE_SD_XY.

VARIABLE	CUBE	CUBE_SD	LCUBE_XY	LCUBE_SD_XY	LPM1_ERN	PIRE	MEILLEUR
Variables utilisées pour équilibrer ou dans la constitution des distances							
TUU_0	1,354	1,113	1,395	1,096	2,580	LPM1_ERN	LCUBE_SD_XY
TUU_1	3,404	1,970	3,459	2,028	7,045	LPM1_ERN	CUBE_SD
TUU_2	4,542	2,456	4,637	2,560	7,825	LPM1_ERN	CUBE_SD
TUU_3	5,751	2,918	5,812	3,054	8,374	LPM1_ERN	CUBE_SD
TUU_4	4,505	2,554	4,600	2,596	6,480	LPM1_ERN	CUBE_SD
TUU_5	4,315	2,260	4,393	2,288	5,726	LPM1_ERN	CUBE_SD
TUU_6	4,320	2,543	4,355	2,794	5,411	LPM1_ERN	CUBE_SD
TUU_7	1,151	0,903	1,113	0,890	1,858	LPM1_ERN	LCUBE_SD_XY
TUU_8	1,680	1,789	0,982	1,113	0,897	CUBE_SD	LPM1_ERN
REVENU_TOTAL	0,531	0,529	0,456	0,473	0,362	CUBE	LPM1_ERN
P12_POP1519	0,520	0,515	0,457	0,480	0,420	CUBE	LPM1_ERN
P12_POP2024	0,581	0,530	0,514	0,489	0,551	CUBE	LCUBE_SD_XY
P12_POP2539	0,521	0,514	0,442	0,458	0,352	CUBE	LPM1_ERN
P12_POP4054	0,502	0,506	0,439	0,471	0,284	CUBE_SD	LPM1_ERN
P12_POP5564	0,487	0,489	0,437	0,467	0,317	CUBE_SD	LPM1_ERN
P12_POP6579	0,497	0,474	0,454	0,448	0,414	CUBE	LPM1_ERN
P12_POP80P	0,587	0,488	0,556	0,450	0,657	LPM1_ERN	LCUBE_SD_XY
C12_MENHSEUL	0,511	0,444	0,452	0,397	0,478	CUBE	LCUBE_SD_XY
C12_MENFSEUL	0,497	0,442	0,438	0,398	0,459	CUBE	LCUBE_SD_XY
C12_MENSFAM	0,584	0,494	0,512	0,415	0,679	LPM1_ERN	LCUBE_SD_XY
C12_MENCOUPSENF	0,471	0,470	0,431	0,454	0,300	CUBE	LPM1_ERN
C12_MENCOUPAENF	0,576	0,547	0,518	0,510	0,412	CUBE	LPM1_ERN
C12_MENFAMMONO	0,556	0,529	0,471	0,461	0,479	CUBE	LCUBE_SD_XY
PROPRIETAIRES	0,525	0,505	0,489	0,484	0,389	CUBE	LPM1_ERN
HLM_2015	2,513	1,538	2,388	1,149	3,964	LPM1_ERN	LCUBE_SD_XY
Variables utilisées dans la constitution des distances							
NB_RES_PRINC	0,420	0,457	0,352	0,420	0,000	CUBE_SD	LPM1_ERN
NB_RES_SEC	8,096	8,137	7,913	7,210	6,721	CUBE_SD	LPM1_ERN
NB_LOG_OCCAS	6,284	6,446	6,072	5,988	5,510	CUBE_SD	LPM1_ERN
NB_LOG_VAC	1,009	0,943	0,982	0,906	0,924	CUBE	LCUBE_SD_XY
P12_NSCOL15P_DIPL0	0,820	0,797	0,755	0,731	0,679	CUBE	LPM1_ERN
P12_NSCOL15P_CEP	0,723	0,698	0,690	0,676	0,592	CUBE	LPM1_ERN
P12_NSCOL15P_BEPC	0,640	0,650	0,582	0,626	0,447	CUBE_SD	LPM1_ERN
P12_NSCOL15P_CAPBEP	0,582	0,585	0,544	0,565	0,398	CUBE_SD	LPM1_ERN
P12_NSCOL15P_BAC	0,540	0,564	0,477	0,523	0,306	CUBE_SD	LPM1_ERN
P12_NSCOL15P_BACP2	0,645	0,644	0,575	0,583	0,453	CUBE	LPM1_ERN
P12_NSCOL15P_SUP	0,851	0,830	0,717	0,687	0,729	CUBE	LCUBE_SD_XY
C12_MEN_CS1	3,313	3,309	3,258	3,203	3,201	CUBE	LPM1_ERN
C12_MEN_CS2	0,906	0,926	0,860	0,875	0,770	CUBE_SD	LPM1_ERN
C12_MEN_CS3	0,911	0,880	0,768	0,730	0,778	CUBE	LCUBE_SD_XY
C12_MEN_CS4	0,645	0,646	0,581	0,591	0,459	CUBE_SD	LPM1_ERN
C12_MEN_CS5	0,659	0,661	0,584	0,596	0,507	CUBE_SD	LPM1_ERN
C12_MEN_CS6	0,759	0,744	0,718	0,706	0,609	CUBE	LPM1_ERN
C12_MEN_CS7	0,503	0,492	0,460	0,475	0,367	CUBE	LPM1_ERN
C12_MEN_CS8	0,938	0,821	0,886	0,805	0,894	CUBE	LCUBE_SD_XY
Variables n'ayant pas servi à l'échantillonnage							
NB_LOG	0,422	0,458	0,354	0,422	0,019	CUBE_SD	LPM1_ERN
SURFACE_TOTALE	0,478	0,497	0,430	0,471	0,242	CUBE_SD	LPM1_ERN
VALEUR_LOCATIVE	0,614	0,634	0,508	0,546	0,451	CUBE_SD	LPM1_ERN
SALAIRES	0,613	0,585	0,524	0,507	0,485	CUBE	LPM1_ERN
PENSIONS	0,544	0,532	0,494	0,499	0,440	CUBE	LPM1_ERN
ALLOC_CHOMAGE	0,599	0,611	0,532	0,562	0,424	CUBE_SD	LPM1_ERN
BEN_IND	1,215	1,231	1,174	1,159	1,101	CUBE_SD	LPM1_ERN
BEN_AGRI	5,410	5,564	5,393	5,440	5,392	CUBE_SD	LPM1_ERN
BEN_NON_COM	1,247	1,197	1,167	1,100	1,312	LPM1_ERN	LCUBE_SD_XY
TAILLE_MENAGE	0,472	0,497	0,403	0,455	0,170	CUBE_SD	LPM1_ERN
LOCATAIRES	0,570	0,515	0,490	0,444	0,544	CUBE	LCUBE_SD_XY
HOMME_FIDELI	0,463	0,488	0,398	0,451	0,162	CUBE_SD	LPM1_ERN
FEMME_FIDELI	0,446	0,478	0,376	0,440	0,108	CUBE_SD	LPM1_ERN
INSCRITS	0,616	0,635	0,574	0,608	0,445	CUBE_SD	LPM1_ERN
VOTANTS	0,641	0,652	0,599	0,624	0,482	CUBE_SD	LPM1_ERN
ABSTENTIONS	0,692	0,722	0,649	0,687	0,542	CUBE_SD	LPM1_ERN
BLANCS	0,990	0,988	0,952	0,948	0,889	CUBE	LPM1_ERN
NULS	3,196	3,297	3,251	3,200	3,211	CUBE_SD	CUBE
ARTHAUD	1,222	1,227	1,178	1,177	1,119	CUBE_SD	LPM1_ERN
ASSELINEAU	1,039	1,053	0,988	1,000	0,942	CUBE_SD	LPM1_ERN
CHEMINADE	1,520	1,509	1,484	1,497	1,422	CUBE	LPM1_ERN
DUPONT_AIGNAN	1,107	1,099	1,028	1,021	0,992	CUBE	LPM1_ERN
FILLON	0,978	0,957	0,933	0,889	0,949	CUBE	LCUBE_SD_XY
HAMON	0,906	0,918	0,814	0,831	0,776	CUBE_SD	LPM1_ERN
LASSALLE	2,391	2,417	2,078	2,032	2,489	LPM1_ERN	LCUBE_SD_XY
LE_PEN	0,929	0,906	0,912	0,883	0,809	CUBE	LPM1_ERN
MACRON	0,780	0,783	0,707	0,724	0,653	CUBE_SD	LPM1_ERN
MELENCHON	0,817	0,816	0,742	0,769	0,677	CUBE	LPM1_ERN
POUTOU	1,091	1,092	1,031	1,038	1,010	CUBE_SD	LPM1_ERN
NAISD04	0,716	0,710	0,635	0,648	0,551	CUBE	LPM1_ERN
NAISD15	0,786	0,787	0,703	0,714	0,649	CUBE_SD	LPM1_ERN
DECESD04	0,823	0,774	0,811	0,760	0,832	LPM1_ERN	LCUBE_SD_XY
DECESD15	0,823	0,782	0,811	0,769	0,839	LPM1_ERN	LCUBE_SD_XY

Lecture : l'estimateur Horvitz-Thompson du nombre de propriétaires a un coefficient de variation de $0,525$ % avec la méthode CUBE ; la méthode la plus précise pour cette variable est LPM1_ERN.

Tableau 7 – Coefficient de variation de l'estimateur Horvitz-Thompson du total, pour les cinq méthodes d'échantillonnage sélectionnées, sur toutes les variables

personnes âgées, le nombre de ménages autre sans famille, le nombre d'électeurs de Lasalle et le nombre de décès. Cette méthode semble donc moins précise sur des variables prenant des valeurs faibles en général. On peut toutefois remarquer que la défaillance sur les tranches d'urbaine a un faible impact sur la majorité des autres variables, puisque le pivot est très souvent le plus précis.

Le cube classique, stratifié ou non, est quant à lui le plus souvent la moins bonne méthode. Le cube et le cube stratifié sont chacun 31 fois la méthode la moins précise. Le cube stratifié fait plutôt mieux que le cube simple sur les variables d'équilibrage, qu'on peut expliquer par le fait que le cube stratifié est équilibré au niveau national, alors que le cube simple est équilibré région par région. En revanche, les deux méthodes sont plus difficiles à départager sur les variables de « distance », et le cube stratifié semblent même plutôt moins bon sur les variables « d'intérêt ». Quoi qu'il en soit, ces deux méthodes sont relativemen proches, à l'exception de la précision sur les tranches d'unité urbaine. De nouveau, on peut noter qu'une faible précision sur ces variables a peu d'impact sur la majorité des autres.

Méthode et distance	100%	87%	75%	50%
CUBE	0.3147	0.3131	0.3097	0.2857
CUBESTRATIFIED	0.3206	0.3195	0.3159	0.2922
LCUBE_XY	0.2831	0.2843	0.2833	0.2687
LCUBESTRATIFIED_XY	0.2822	0.2828	0.2820	0.2644
LPM1_ACP_RATIO_FULL	0.3141	0.3136	0.3107	0.2884
LPM1_ACP_RATIO_FULL_XY_NORM	0.3134	0.3131	0.3100	0.2881
LPM1_ACP_RATIO_MIX_FULL	0.3131	0.3128	0.3097	0.2874
LPM1_EUCL_RATIO_NORM	0.3141	0.3138	0.3105	0.2885
LPM1_EUCL_RATIO_XY_NORM	0.3133	0.3127	0.3097	0.2883

Lecture : la variance des aires des polygones de Voronoi est estimée à 0.2831 pour le cube local (LCUBE_XY) sur un échantillon couvrant 100% du réseau d'enquêteurs.

Tableau 8 – Étalement géographique au sens de l'indicateur « de Voronoï » des méthodes retenues pour les quatre tailles d'échantillon

On souhaite également observer l'étalement géographique des différentes méthodes, à l'aide des aires des polygones de Voronoï (tableau 8). Les méthodes du cube local sont particulièrement bien étalées, l'indicateur étant plus proche de 0 pour ces méthodes, quelque soit la taille d'échantillon. En revanche, le cube stratifié apparaît moins bon que les autres méthodes. Cela peut provenir du fait qu'intervient dans l'algorithme, une phase de survol du cube au niveau national. L'« étalement géographique » introduit par la stratification par région y est donc peut-être affaibli. On peut d'ailleurs constater de façon plus générale que les écarts sont plutôt faibles sur cet indicateur. Les échantillons sont en effet artificiellement répartis par la stratification. Cela permet d'éviter les pires des cas où, par exemple, l'ensemble des unités primaires non exhaustives seraient sélectionnées en Île-de-France.

5.3 Analyse de la distribution

Nous avons pu extraire les méthodes les plus précises sur l'estimation de totaux de variables. On souhaite compléter l'analyse en observant le comportement des méthodes sur la restitution de la distribution des variables. On s'intéresse pour cela aux estimations des trois quartiles. On reprend une approche globale à l'aide d'une analyse en composantes principales.

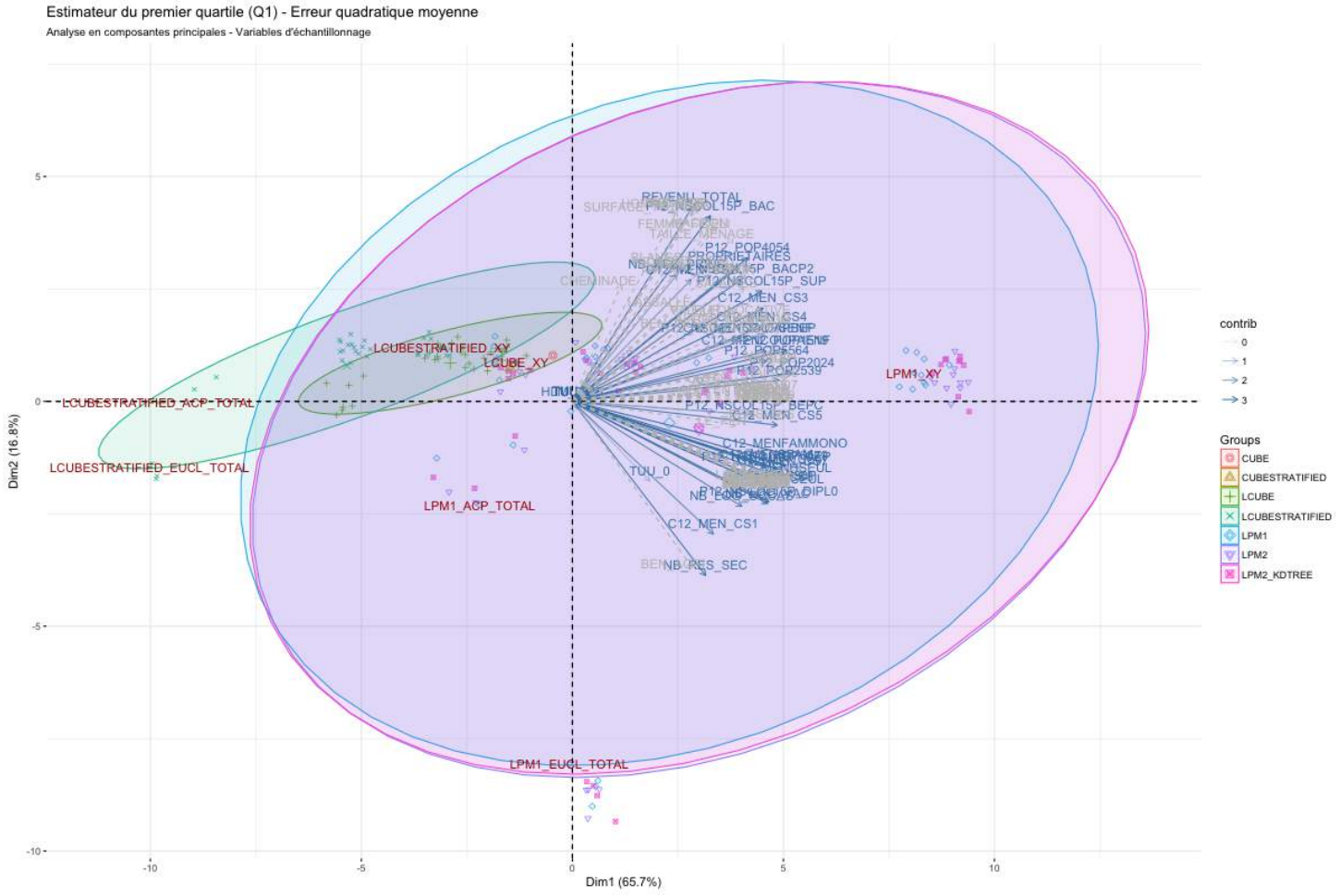
Premier quartile (Q1) Le biais relatif de l'estimateur du premier quartile varie de -5,724 % pour la méthode LPM2_KDTREE_MAHAL_TOTAL_XY sur la variable C12_MEN_CS1, à 1,96 % pour la méthode LPM1_EUCL_RA-TIO_NORM_XY sur la variable CHEMINADE.

La norme euclidienne dans l'espace des biais relatif pour les 182 méthodes est sensiblement supérieure à 0 et a une distribution compacte (tableau 9).

0%	25%	50%	75%	100%
6.853092	7.020875	7.068700	7.127479	7.263255

Tableau 9 – Distribution de la norme euclidienne dans l'espace des biais relatifs

Contrairement à l'estimateur Horvitz-Thompson du total, l'estimateur du premier quartile est donc biaisé en général. Comme pour le total, il ne semble pas se dégager de schéma facilement interprétable dans la répartition des méthodes selon le biais relatif. Étant donné le caractère biaisé de l'estimateur, on se concentre plutôt ici sur l'erreur quadratique moyenne, qui permet de rendre compte à la fois du biais et de la variance de l'estimateur (graphique 6).



Graphique 6 – Analyse en composantes principales sur les erreurs quadratiques moyennes de l'estimateur du premier quartile (Q1), pour les 182 méthodes d'échantillonnage testées

Les deux premiers axes factoriels portent plus de 80 % de l'inertie totale du nuage. Les 3 implémentations du pivot local ont à peu près le même comportement. Ces méthodes sont plus sensibles que les méthodes du cube local à la distance utilisée pour étaler les unités (elles sont plus dispersées dans le premier plan factoriel). Cette propriété est moins vraie que ce qui a pu être observé pour l'estimateur du total (les méthodes du cube local varient plus dans ce cas).

Toutes les variables sont corrélées positivement avec le premier axe factoriel. On peut interpréter le graphique de la façon suivante : plus une méthode de sondage est positionnée dans la direction est, et proche de l'axe des abscisses, plus elle est globalement précise (*i.e* sur l'ensemble des variables) dans l'estimation du premier quartile.

La famille des méthodes du cube local stratifié est celle qui va le plus loin vers la gauche, et présente a priori les meilleurs candidats pour la précision sur le premier quartile. Les distances euclidienne sur les totaux (EUCL_TOTAL) et sur les deux premiers axes d'une ACP sur les totaux (ACP_TOTAL) semblent particulièrement bien fonctionner. En revanche, le pivot uniquement étalé sur la géographie (LPM_XY) fait partie des moins bonnes méthodes pour estimer avec précision le premier quartile.

On peut noter que le cube local simplement étalé sur la géographie (LCUBESTRATIFIED_XY) est tout de même proche de l'axe des abscisses, et plutôt corrélé négativement avec le premier axe factoriel (cette méthode est donc globalement plus précise que la moyenne sur le premier quartile). En tout cas, cette méthode paraît la meilleure parmi les autres méthodes sélectionnées (tableau 10), même si elles sont relativement proches.

Méthode et distance	Axe 1	Axe 2
CUBE	-0.4638	1.0231
CUBESTRATIFIED	-1.4746	0.7462
LCUBE_XY	-1.3228	0.8802
LCUBESTRATIFIED_XY	-3.0842	1.2047
LPM1_EUCL_RATIO_NORM	-1.7866	0.9581

Tableau 10 – Coordonnées des méthodes retenues sur les deux premiers axes d'une ACP dans l'espace des erreurs quadratiques moyennes de l'estimateur du premier quartile (Q1)

Médiane Les résultats sur l'estimation de la médiane sont assez similaires aux résultats sur le premier quartile, avec pour principale nuance que la précision des méthodes du cube local stratifié (LCUBESTRATIFIED) est moins nettement supérieure aux autres. Il y a de nouveau un biais pour cet estimateur, qui varie de -1,73 % pour la méthode LPM1_MAHAL_RATIO_XY sur la variable CHEMINADE, à 3,301 % pour la méthode LPM2_ACP_RATIO_XY sur la variable NB_LOG_OCCAS. On se concentre donc une nouvelle fois sur l'erreur quadratique moyenne (graphique 9, en annexe E).

Le premier plan factoriel porte cette fois près de 90 % de l'inertie. Les distances ACP_TOTAL et EUCL_TOTAL sont encore performantes pour le cube local stratifié (LCUBESTRATIFIED), tandis que les méthodes du pivot étalées sur la géographie (LPM_XY) sont toujours décevantes. De nouveau, les cinq méthodes retenues sont très proches, le cube local équilibré au niveau national (LCUBESTRATIFIED_XY) étant légèrement plus à gauche que les autres (coordonnée de -2,28 sur le premier axe lorsque les autres sont plutôt aux alentours de -1), et donc meilleur.

Troisième quartile De nouveau, l'estimateur du troisième quartile est biaisé et on se focalise sur l'étude de l'erreur quadratique moyenne (graphique 10, en annexe E). Plus de 90 % de l'inertie est comprise dans le premier plan factoriel, et le premier axe en particulier en concentre presque 78 %. Encore une fois, les méthodes les plus à gauche sur le plan sont les plus précises dans l'estimation du troisième quartile.

Les distances ACP_TOTAL et EUCL_TOTAL appliquées au cube local stratifié sont parmi les plus corrélées négativement avec le premier axe. Cependant, d'autres méthodes sont à présent plus à gauche, comme le pivot local couplé à la distance ACP_TOTAL. Le pivot étalé sur la géographie (LPM_XY) reste décevant pour cet estimateur. Les cinq méthodes sélectionnées sont une nouvelle fois proches. Le cube double (LCUBESTRATIFIED_XY) se situe de nouveau légèrement plus à gauche.

6 Discussion et application

On propose à présent de discuter les résultats obtenus, au regard de la théorie et des données utilisées. On expose également une application réalisée durant l'étude, prenant en compte les premiers résultats obtenus au moment de sa mise en œuvre.

6.1 Discussion

Il sort des résultats que les tirages équilibrés au niveau national sont plus précis sur les variables d'équilibrage que les tirages équilibrés par région (ou strate). Ce résultat est conforme à la théorie. En effet, le tirage strate par strate est susceptible d'accumuler des erreurs d'arrondi préjudiciables pour des estimations au niveau national. En revanche, il est potentiellement meilleur pour des estimations régionales (ce qui n'était pas testé ici). De plus, équilibrer au niveau national doit aussi permettre l'utilisation d'un plus grand nombre de variables auxiliaires. Pour le cube région par région, le nombre de variables auxiliaires était supérieur au nombre d'unités à tirer dans un certain nombre de strate. Ceci signifie que l'algorithme entame dès le départ la phase d'atterrissage, où les contraintes d'équilibrage sont progressivement retirées.

En revanche, la différence entre les deux approches est beaucoup moins claire sur les variables n'ayant pas servi à équilibrer. Elles affichent une précision de même ordre de grandeur et sont l'une et l'autre, parfois moins bonne, parfois meilleure. Ce résultat est plus surprenant. On aurait en effet pu supposer qu'un gain en précision sur les variables d'équilibrage aurait entraîné une variance moindre sur les autres variables. Visiblement, la part d'imprécision captée par le tirage strate par strate était maximale avec les données utilisées. En particulier, le gain apporté par une meilleure précision sur les tranches d'aire urbaine, pour lesquelles les deux approches diffèrent le plus, ne semble pas avoir beaucoup d'effet. On suggère quand même de privilégier le cube stratifié qui, quand il est précis, l'est beaucoup plus que le cube strate par strate.

Un second résultat est conforme à la théorie, le gain à étaler spatialement les unités sélectionnées. En introduisant un mécanisme de répartition spatiale dans le processus d'échantillonnage, on permet pour un certain nombre de variables de capter une part de l'hétérogénéité inobservé, liée à une autocorrélation spatiale positive. Hormis pour quelque variables auxiliaires servant à équilibrer le tirage, le cube étalé est presque systématiquement meilleur (plus précis) que le cube simple. Il constitue donc un excellent candidat pour le tirage du prochain échantillon-maître.

Un bon candidat peut également venir de la famille des méthodes du pivot local, à condition de choisir convenablement l'espace et la distance permettant l'étalement. On a en effet observé que ces méthodes étaient très sensibles au choix de la distance, et qu'elles pouvaient rivaliser avec le cube local tout comme être nettement moins bonne que le cube simple. Une distance basée sur les ratios de variables centrés et réduits paraît fonctionner correctement dans le cadre de ces simulations et pourrait être envisagée pour le futur tirage de l'EM. Cependant, contrairement au cube local, il n'est pas systématiquement meilleur que la méthode de référence, le cube classique strate par strate. De plus, le cube local se comporte légèrement mieux dans l'estimation des quantiles, et il permet une répartition plus homogène des unités sur le territoire.

Il serait nécessaire de renforcer cette étude en faisant varier les jeux de variables utilisés. Il n'est en effet pas à exclure que les résultats obtenus soient sensibles aux informations mobilisées, en particulier les méthodes du pivot local retenues, qui ne reposent que sur de l'information auxiliaire. Le cube étalé semble offrir cet avantage d'être arbitraire dans le choix des variables complémentaires (X et Y) pour capter une part de l'inobservable. La difficulté qui peut se présenter est l'existence de données sur la population totale corrélées avec les concepts traités dans les enquêtes ménages.

A posteriori, le choix des tranches d'unité urbaine pour traduire le type de territoire ne semble pas la plus adaptée. En effet, on a pu observer qu'une perte de précision sur ces variables avait un impact limité sur l'estimation des autres variables. De plus, elles sont découpées en de

nombreuses modalités, ce qui limite le nombre de variables auxiliaires utilisables pour équilibrer. On peut suggérer de ne retenir que certaines modalités spécifiques, comme les zones rurales, ou de regrouper les modalités. Une autre possibilité est d'exploiter la notion d'aire urbaine, axée sur la position par rapport à un (ou des) pôle(s) urbain(s).

6.2 Une application : tirage pour l'enquête EHIS

Au cours de l'étude, un besoin similaire au tirage de l'échantillon-maître a émané pour l'enquête EHIS (« European Health International Survey ») 2019. Pour fournir aux chargés de collecte la liste des communes potentiellement impactées par l'enquête, il a été émis le souhait de tirer un échantillon au premier degré de 250 unités primaires.

La base de sondage est cette fois basée sur les nouvelles unités primaires, construites sur la géographie au 1er janvier 2016 (en particulier, les régions sont passées de 22 à 13). Les concepts d'intérêt de l'enquête sont *a priori* corrélés avec l'âge et le revenu. On récupère donc les informations nécessaires au sein du recensement 2014 (géographie 2016), de sources administratives (2014 également) et de l'IGN.

La méthode choisie pour l'échantillonnage est un tirage équilibré au niveau national et spatialement réparti (équivalent à la méthode LCUBESTRATIFIED_XY simulée durant l'étude). On choisit d'équilibrer, dans l'ordre, sur le revenu fiscal total, les pensions (retraites), les tranches d'âge, et le territoire. On construit cette fois-ci une variable sur le territoire plus synthétique (URBAIN, COURONNE, PERIURBAIN et RURAL), et basée non plus sur les unités urbaines mais sur la position par rapport à un pôle urbain, à l'aide de la notion d'aire urbaine. Les probabilités sont de nouveau déterminées proportionnellement au nombre de résidences principales des zones, grâce au même algorithme que celui implémenté pour l'étude.

On constate que la typologie du territoire est de nouveau difficile à respecter (tableau 11). En effet, les zones rurales sont légèrement surreprésentées alors que les zones périurbaines sont légèrement sous-représentées. En revanche, les résultats sont plutôt bons sur les autres variables d'équilibrage. L'échantillon obtenu a été validé par les responsables de l'enquête.

Variable	Total dans la population	Estimateur Horvitz-Thompson du total	Biais relatif (%)
PI_REG	291	291,00	0,00
REVENU_TOTAL_FDL14	949 825 270 250	949 122 127 735,43	-0,07
PENSIONS_FDL14	266 335 415 127	266 495 631 992,00	0,06
P14_POP1529	11 444 179	11 384 288,17	-0,52
P14_POP3044	12 374 045	12 370 614,68	-0,03
P14_POP4559	12 741 418	12 751 554,69	0,08
P14_POP6074	9 786 932	9 802 580,57	0,16
P14_POP7589	5 273 009	5 276 417,75	0,06
P14_POP90P	683 565	685 473,56	0,28
URBAIN	18 354 876	18 360 420,90	0,03
COURONNE	5 201 937	5 254 546,34	1,01
PERIURBAIN	2 800 614	2 653 131,09	-5,27
RURAL	1 292 011	1 381 339,67	6,91

Tableau 11 – Biais relatif (%) des estimateurs Horvitz-Thompson du total pour les variables d'équilibrage utilisées dans la sélection de l'échantillon EHIS

7 Conclusion

Début 2018, l'Insee a procédé à l'échantillonnage du futur échantillon-maître, qui permettra d'alimenter l'application Nautile et de réaliser la sélection des enquêtes auprès des ménages. Le précédent échantillon-maître a été sélectionné à l'aide d'un tirage équilibré région par région. Cependant, de nouvelles avancées dans la théorie de l'échantillonnage spatial étaient susceptibles d'améliorer la qualité des estimations. En particulier, le pivot local et le cube local permettent de prendre en compte le comportement spatial des données. De plus, ces méthodes peuvent être réadaptées en utilisant une distance dans l'espace des variables socio-démographiques.

Nous avons proposé dans le cadre de cette étude de comparer 182 processus d'échantillonnage reposant sur le cube, le pivot local et le cube local, mobilisant des distances mêlant les coordonnées géographiques et des variables auxiliaires. Les simulations ont été réalisées sur un jeu d'unités primaires de test, le découpage final étant en cours de validation pendant l'étude, complété par des données disponibles sur la totalité de la population. La comparaison des méthodes a reposé sur l'analyse du biais, du coefficient de variation et de l'erreur quadratique moyenne des estimateurs Horvitz-Thompson du total et des trois quartiles. Ces indicateurs ont été estimés à l'aide du procédé de Monte-Carlo.

Nous avons pu montrer qu'au moins deux méthodes sont préférables à la méthode de référence, utilisée lors de la précédente opération. La première méthode est le cube local réparti spatialement, prenant en compte la stratification au niveau régional mais avec équilibrage au niveau national sur les variables auxiliaires. Cette méthode est systématiquement plus précise que la méthode de référence dans l'estimation de totaux, et équivalente pour l'estimation de quantiles.

L'autre méthode intéressante est basée sur le pivot local. On a en effet pu montrer qu'avec une distance bien choisie, il était possible d'obtenir un processus d'échantillonnage généralement plus précis que le benchmark. Cependant, ce n'est pas vrai pour la totalité des variables, contrairement au cube local. De plus, les méthodes du pivot local retenues sont légèrement moins bonnes que le cube local dans l'estimation des quantiles, et les échantillons obtenus sont généralement répartis de façon moins homogène sur le territoire. Le pivot local présente cependant l'avantage de respecter exactement les probabilités d'inclusion, et donc le nombre total de résidences principales.

On a donc été amené à suggérer d'utiliser le cube local ou le pivot local (avec une distance basée sur les ratios de variables auxiliaires) pour sélectionner le nouvel échantillon-maître. Cette étude pourrait être complétée en faisant varier le jeu de variables auxiliaires mobilisées, afin notamment d'observer le comportement de la méthode du pivot local retenue.

Références

Ardilly, P. et Tillé, Y. (2006). Multi-stage sampling. *Sampling Methods : Exercises and Solutions*, pages 159–207.

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

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

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.*

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

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

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.*

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.*

Grafström, A. et Lisic, J. (2016). Balancedsampling : balanced and spatially balanced sampling. r package version 1.5.2.

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

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

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

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

Hyndman, R. J. et Fan, Y. (1996). Sample quantiles in statistical packages. *The American Statistician*, 50(4):361–365.

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

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

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

A Enquêtes de l'Insee auprès des ménages

Dans le cadre de cette étude, on retient 12 enquêtes auprès des ménages (tableau A). 5 sont récurrentes, c'est-à-dire qu'elles sont réalisées au moins une fois par an et 7 plus ponctuelles, pour lesquelles les échantillonnages successifs peuvent être espacés de plusieurs années. Elles sont sélectionnées au sein de l'échantillon-maître Pour chacune de ces enquêtes, on retient 3 variables, a priori corrélées aux concepts étudiés.

Enquête	Objectif	Variables corrélées
Enquêtes récurrentes		
Loyers et Charges	Mesurer l'évolution des loyers	Nombre de locataires Surface de logement Loyer
Technologies de l'information et de la communication (TIC)	Décrire l'équipement et les usages des TIC par les ménages	Territoire rural Accès au haut débit Âge
Conjoncture auprès des ménages (CAMME)	Suivre l'opinion que portent les ménages sur leur environnement économique	Âge Revenu CSP
Statistiques sur les ressources et conditions de vie (SRCV)	Production d'indicateurs structurels sur la répartition des revenus, de la pauvreté et de l'exclusion	Revenu Allocations Logements sociaux
Cadre de vie et sécurité (CVS)	Connaître les faits de délinquance dont les ménages et leurs membres auraient pu être victimes	Territoire Âge CSP
Enquêtes ponctuelles		
Enquête sur la formation des adultes (AES)	Mesurer l'accès des adultes à des activités de formation	Âge Sexe Diplôme
Budget de famille (BDF)	Restituer la comptabilité des ménages : dépenses et ressources	Revenu Âge CSP
Emploi du temps (EDT)	Collecter les données sur la façon dont les individus organisent leur temps	Composition du ménage Logement Revenu
Histoire de vie et patrimoine	Décrire les actifs financiers, immobiliers et professionnels des ménages	CSP Revenu Composition du ménage
Conditions de travail / Risques psycho-sociaux	Étudier l'organisation et les rythmes de travail, ainsi que les contraintes psychosociales	CSP Diplôme Revenu
Mobilité des personnes	Connaître les déplacements des ménages et leur usage des moyens de transport	Territoire Composition du ménage CSP
Entrée dans la vie adulte (EVA)	Étudier l'insertion professionnelle des jeunes sortis du système éducatif	Âge Diplôme CSP

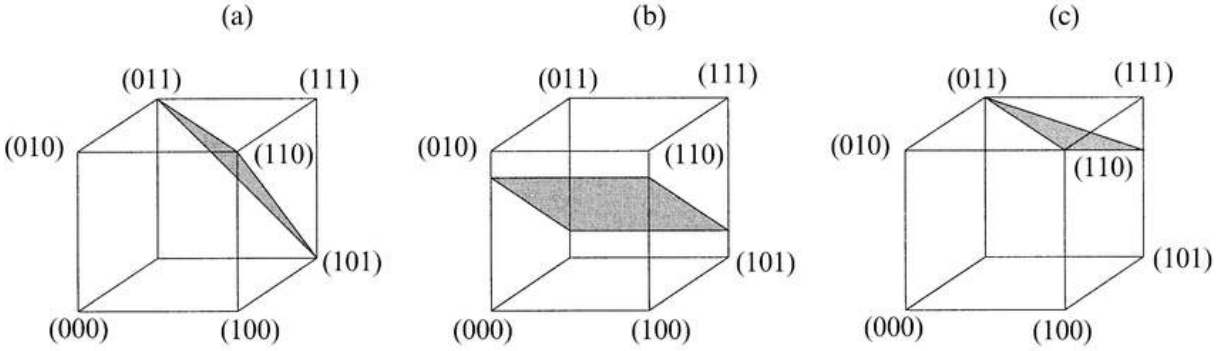
Tableau 12 – Liste des enquêtes ménage échantillonnées par l'Insee, retenues dans le cadre de cette étude, et les variables qu'on considère être d'intérêt

B La méthode du cube

La méthode du cube a été introduite par Deville et Tillé (2004) et permet de réaliser des tirages d'échantillon équilibrés sur des variables auxiliaires, tout en respectant les probabilités d'inclusion définissant le plan de sondage. Cette méthode est caractérisée par une représentation géométrique de la population et des contraintes d'équilibrage.

On note N la taille de la population, n la taille de l'échantillon, $\pi = (\pi_1, ..., \pi_N)$ le vecteur des probabilités d'inclusion. On se place dans l'espace $\mathbb{R}^N$, chaque dimension pouvant être vue comme représentant une unité de la population. On y représente l'ensemble des échantillons possibles (mais de probabilité éventuellement nulle) par un hypercube C de côté 1. Un sommet du cube peut ainsi être représenté par un vecteur de taille N, contenant des 0 et de 1, en fonction de la présence ou non des unités dans l'échantillon.

Les auteurs montrent que pour p variables auxiliaires $X_1, ..., X_p$, la contrainte d'équilibrage peut être représentée par un sous-espace Q de $\mathbb{R}^N$ de dimension $N - p$. De plus, l'intersection de C et Q est non nul puisque le vecteur des probabilités d'inclusion π appartient à ces deux espaces. On peut se représenter les deux objets à l'aide du graphique 7.



Graphique 7 – Représentation géométrique du cube C et de l'espace des contraintes Q pour une population de taille 3 - *Source - Deville et Tillé (2004)*

Le tirage équilibré avec la méthode du cube repose sur deux étapes : une phase « de vol » et une phase « d'atterrissage ». L'idée de la phase de vol est de se déplacer dans l'intersection de C et de Q en s'approchant le plus possible d'un des sommets du cube, ou à défaut d'une de ses arêtes ou une de ses faces (*i.e* le plus proche d'un échantillon définitif). Ainsi, durant cette étape, les contraintes d'équilibrage sont parfaitement respectées. Pour l'implémenter, les auteurs introduisent ce qu'ils appellent une « martingale équilibrée », qui permet d'évoluer dans le cube (c'est-à-dire mettre à jour le vecteur des probabilités d'inclusion), tout en garantissant qu'en espérance le vecteur des probabilités mises à jour vaille π et que lorsqu'une des phases du cube est atteinte, elle ne soit plus quittée. En règle générale, un sommet du cube ne peut pas être atteint à l'issue de cette étape. Aucune décision peut n'avoir été prise pour au plus p unités (leurs probabilités d'inclusion mises à jour sont encore comprise entre 0 et 1).

Pour la phase d'atterrissage, dans le cadre de cette étude, on choisit la stratégie consistant à supprimer successivement des contraintes, en commençant par la moins *prioritaire*, et à relancer une phase de vol.

C La méthode du pivot local

La méthode du pivot a été introduite par Deville et Tille (1998) et permet la sélection d'échantillons en au plus N étapes (avec N le nombre d'unités dans la population). Cette stratégie repose sur la mise à jour conjointe des probabilités d'inclusion pour deux unités (étant donné une règle de sélection des couples d'unités).

On note $\pi = (\pi_1, ..., \pi_N)$ le vecteur des probabilités d'inclusion. L'idée est de mettre à jour ce vecteur étape par étape, jusqu'à l'obtention d'un vecteur composé uniquement de 0 et de 1 (en fonction de la sélection ou non de chacune des unités dans l'échantillon). À chaque étape, après avoir sélectionné un couple d'unités $(k, l) \in \{1, ..., N\}^2$, on met à jour leurs probabilités d'inclusion selon la règle de décision :

— Si $\pi_k + \pi_l \leq 1$: $\begin{pmatrix} \pi_k \\ \pi_l \end{pmatrix}$ devient $\begin{pmatrix} \pi_k + \pi_l \\ 0 \end{pmatrix}$ avec probabilité $\frac{\pi_k}{\pi_k + \pi_l}$, et $\begin{pmatrix} 0 \\ \pi_k + \pi_l \end{pmatrix}$ sinon.

— Si $\pi_k + \pi_l > 1$ $\begin{pmatrix} \pi_k \\ \pi_l \end{pmatrix}$ devient $\begin{pmatrix} 1 \\ \pi_k + \pi_l - 1 \end{pmatrix}$ avec probabilité $\frac{1-\pi_l}{2-\pi_k-\pi_l}$, et $\begin{pmatrix} \pi_k + \pi_l - 1 \\ 1 \end{pmatrix}$ sinon

Grafström *et al.* (2012) ont repris cette stratégie et l'ont adaptée à une approche spatiale de l'échantillonnage, qu'ils appellent pivot local. Leur apport consiste dans la stratégie de sélection des couples d'unités successifs. Dans le papier, ils proposent deux algorithmes d'échantillonnage, qui diffèrent essentiellement dans la façon de choisir les couples de plus proches voisins. La trame de cet algorithme est la suivante :

1. Choix d'une unité parmi celles pour lesquelles aucune décision (sélection ou non) n'a été prise.
2. Recherche du plus proche voisin de cette unité. Si le choix n'est pas satisfaisant, l'algorithme reprend à l'étape 1.
3. Les probabilités d'inclusion des deux unités sont mises à jour selon la règle du pivot.

L'algorithme s'achève lorsqu'une décision a été prise pour l'intégralité des unités.

La première version du pivot est plus exigeante que la seconde. En effet, à l'étape 2 de l'algorithme, on estime convenable un couple d'unités qui sont mutuellement plus proches voisines. Dans la seconde version, il suffit qu'une des deux unités soit la plus proche voisine de l'autre.

L'intérêt de cette méthode est de permettre de réduire la probabilité d'inclusion double de deux unités proches selon une certaine distance, puisque la mise à jour des probabilités d'inclusion renforce une des deux unités tout en affaiblissant l'autre. Ce mécanisme de répulsion est bénéfique en présence d'autocorrélation spatiale positive.

D Autocorrélation spatiale

On calcule l'indice de Moran pour les variables dites d'« équilibrage », afin d'évaluer la présence potentielle d'autocorrélation spatiale.

La question que traite l'indice I de Moran est celle de savoir si une variable X co-varie de la même manière pour des unités voisines. Étant donné n unités, liées par des poids w_{ij}, en notant X_i la valeur de X pour l'unité i et $\bar{X} = \sum_{i=1}^{n} X_i$, on définit l'indice par :

$$I = \frac{n}{\sum_{i \neq j} w_{ij}} \frac{\sum_i \sum_j w_{ij}(X_i - \bar{X})(X_j - \bar{X})}{\sum_i (X_i - \bar{X})^2} \tag{11}$$

En pratique, les poids utilisés ici sont l'inverse des distances entre deux unités primaires. Ainsi, plus deux unités sont proches, plus leur poids mutuel est important.

Sous l'hypothèse de non autocorrélation spatiale, cet indice a pour espérance $\frac{-1}{n-1}$ et une variance non détaillée ici. Ceci permet de construire un test statistique : si la p-valeur obtenue est très faible, on peut rejeter avec une confiance raisonnable l'hypothèse de non autocorrélation.

Variable	I de Moran	Espérance	Écart-type	p-valeur
TUU_0	0.077434	-0.000192	0.000525	0
TUU_1	0.012855	-0.000192	0.000525	0
TUU_2	0.011277	-0.000192	0.000524	0
TUU_3	0.006949	-0.000192	0.000524	0
TUU_4	0.009260	-0.000192	0.000522	0
TUU_5	0.009321	-0.000192	0.000520	0
TUU_6	0.011139	-0.000192	0.000508	0
TUU_7	0.027873	-0.000192	0.000509	0
TUU_8	0.224181	-0.000192	0.000515	0
REVENU_TOTAL	0.082464	-0.000192	0.000518	0
P12_POP1519	0.037115	-0.000192	0.000517	0
P12_POP2024	0.026582	-0.000192	0.000512	0
P12_POP2539	0.059917	-0.000192	0.000517	0
P12_POP4054	0.063236	-0.000192	0.000520	0
P12_POP5564	0.046980	-0.000192	0.000520	0
P12_POP6579	0.034827	-0.000192	0.000519	0
P12_POP80P	0.025097	-0.000192	0.000518	0
C12_MENHSEUL	0.029746	-0.000192	0.000514	0
C12_MENFSEUL	0.031046	-0.000192	0.000517	0
C12_MENSFAM	0.041408	-0.000192	0.000511	0
C12_MENCOUPSENF	0.030577	-0.000192	0.000518	0
C12_MENCOUPAENF	0.076697	-0.000192	0.000520	0
C12_MENFAMMONO	0.054002	-0.000192	0.000520	0
PROPRIETAIRES	0.037277	-0.000192	0.000517	0
HLM_2015	0.014507	-0.000192	0.000484	0

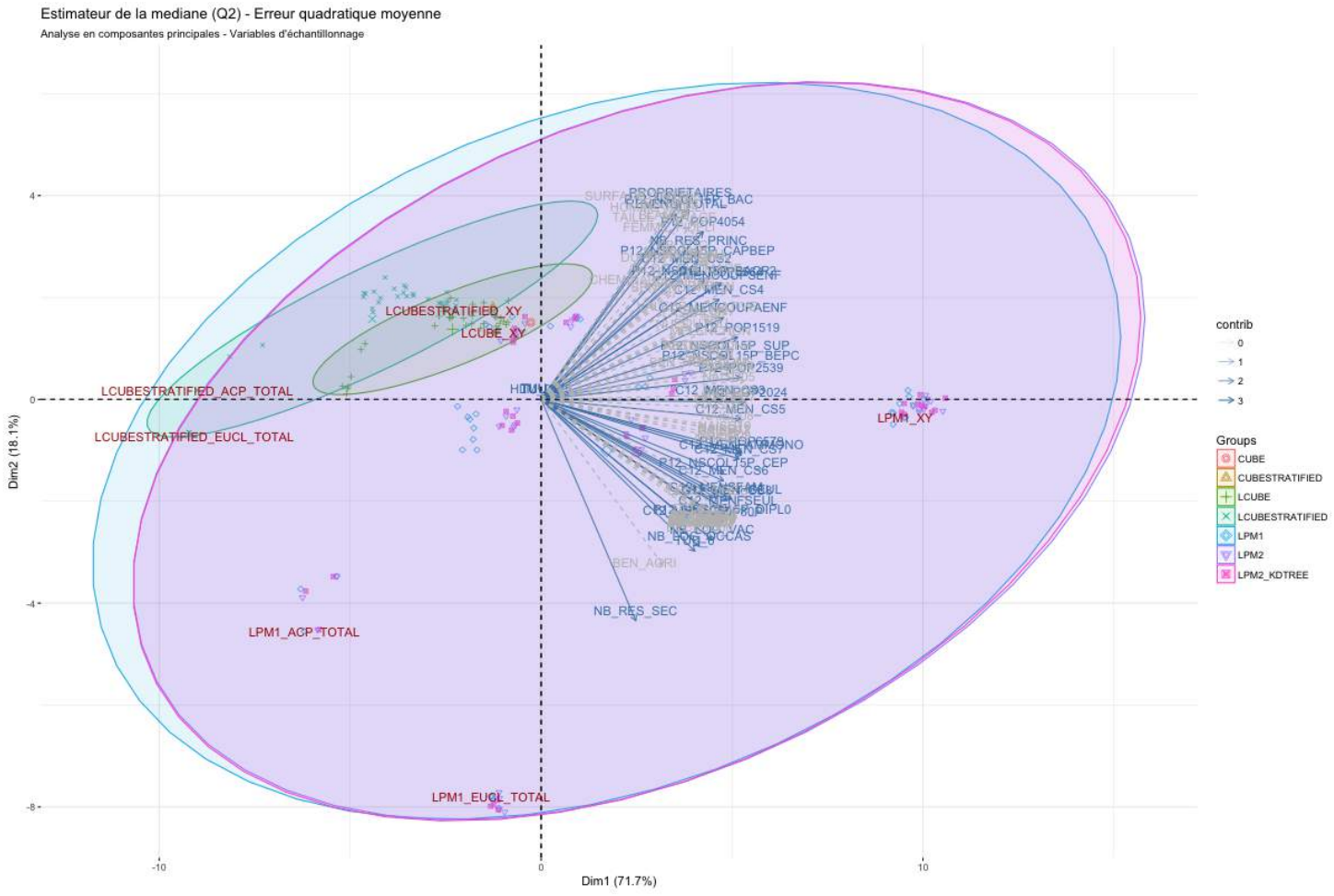
Tableau 13 – Indice de Moran pour les variables « d'équilibrage »

E Résultats : compléments

Biais relatif de l'estimateur Horvitz-Thompson du total

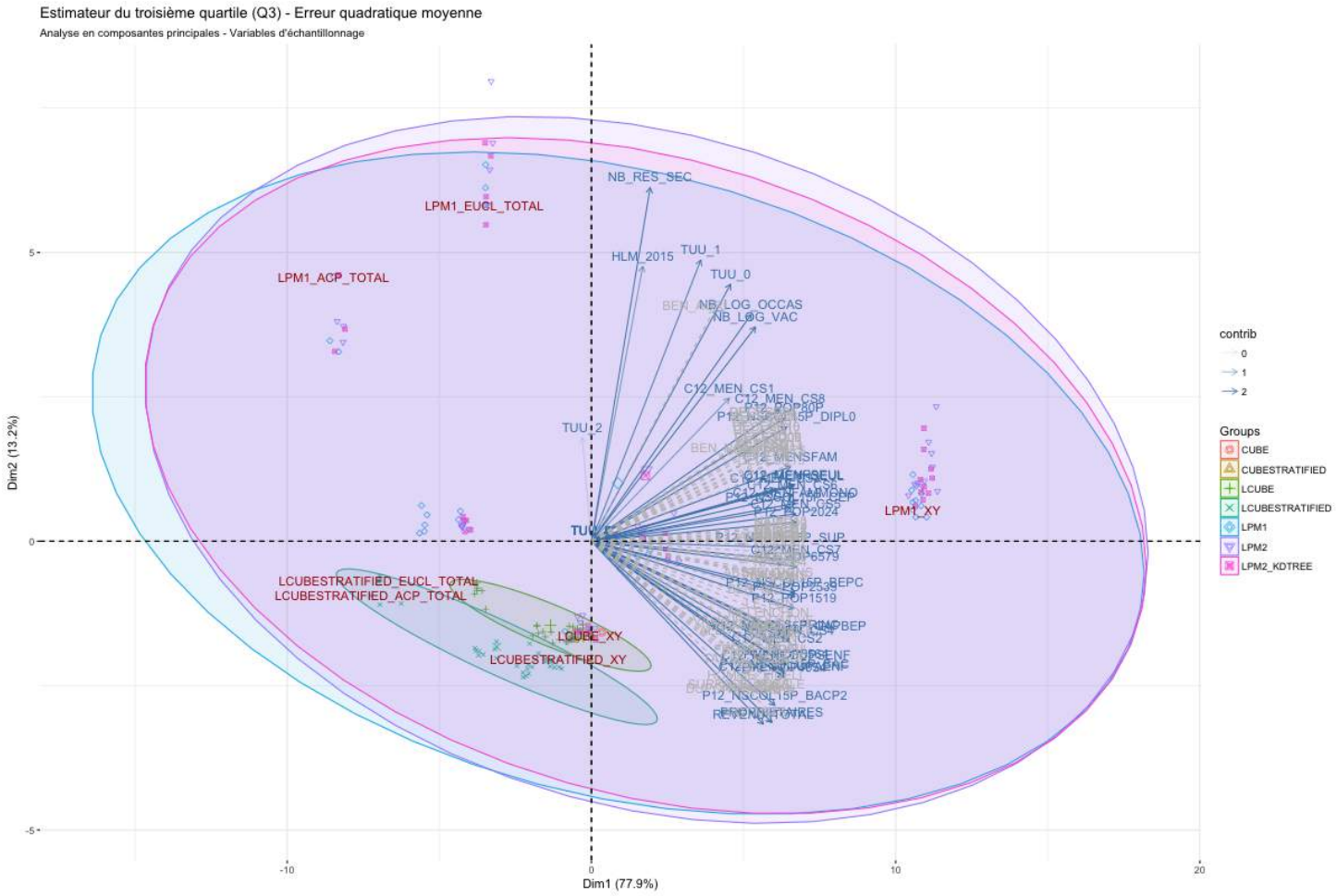
Graphique 8 – Analyse en composantes principales sur le biais relatif de l'estimateur Horvitz-Thompson du total, pour les 182 méthodes d'échantillonnage testées

Erreur quadratique moyenne de l'estimateur de la médiane (Q2)



Graphique 9 – Analyse en composantes principales sur les erreurs quadratiques moyennes de l'estimateur de la médiane (Q2), pour les 182 méthodes d'échantillonnage testées

Erreur quadratique moyenne de l'estimateur du troisième quartile (Q3)



Graphique 10 – Analyse en composantes principales sur les erreurs quadratiques moyennes de l'estimateur du troisième quartile (Q3), pour les 182 méthodes d'échantillonnage testées