Tirage coordonné d'échantillons : une application à l'Échantillon-Maître Nautile et à l'enquête Emploi

Thomas Merly-Alpa [1], Ludovic Vincent [2], Laurent Costa [3], Clément Guillo [4], Nicolas Paliod [5], Martin Chevalier [6], Thomas Deroyon [7]

[1] *Insee, 88 avenue Verdier, 92120 Montrouge, thomas.merly-alpa@insee.fr*
[2] *ludovic.vincent@insee.fr* [3] *laurent.costa@insee.fr* [4] *clement.guillo@insee.fr*
[5] *nicolas.paliod@insee.fr* [6] *martin.chevalier@insee.fr* [7] *thomas.deroyon@insee.fr*

Résumé. L'Institut National de la Statistique et des Études Économiques (Insee) renouvelle son Échantillon-Maître ainsi que l'échantillon de l'enquête Emploi en Continu (EEC) dans les prochaines années. Dans ce cadre, une coordination de ces deux tirages a été étudiée afin de limiter les déplacements des enquêteurs. Cette communication présente les grands principes des deux tirages puis expose la méthode mise en œuvre pour la coordination géographique des deux enquêtes. Les unités primaires de l'Échantillon-Maître sont des regroupements de communes et sont tirées par échantillonnage spatialement équilibré à un niveau régional, en mobilisant des variables socio-démographiques diverses, directement ou résumées par une analyse factorielle. L'échantillon de l'EEC est un ensemble de grappes compactes d'une vingtaine de logements également sélectionnées par sondage spatialement équilibré à l'aide de variables *proxy* de la situation vis-à-vis du marché du travail. Enfin, la coordination de ces échantillons se fait par l'introduction d'unités de coordination (UC), englobant les unités primaires, et permettant un équilibrage des deux échantillons de façon parallèle. Les UC sont des zones agrégeant plusieurs unités primaires avec une étendue minimale. Elles sont sélectionnées par sondage indirect *via* les unités primaires, et les grappes de l'EEC sont ensuite sélectionnées dans les UC, dans une dernière phase. Le tirage des unités primaires est équilibré sur un jeu de variables dites indirectes permettant d'assurer que les UC sont représentatives pour les variables d'équilibrage de l'EEC, et ainsi de garantir une bonne précision à l'enquête.

Mots-clés. Échantillonnage équilibré, équilibrage spatial, coordination d'échantillons, Échantillon-Maître, sondage indirect.

Abstract This paper details the coordination of the sampling process of the new Master Sample and the LFS (Labour Force Survey) at the French National Statistical Institute (Insee). As the two sampling frames are different, coordination units (UC) are introduced to allow a spatial coordination, in order to facilitate data collection. The coordination process requires specific variables to simultaneously balance the two samples ; this step is achieved using generalized weight sharing method.

Keywords. Balanced sampling, spatial sampling, sample coordination, Master sample, indirect sampling.

1 Contexte : le projet Nautile

L'Institut National de la Statistique et des Études Économiques (Insee) réalise la plupart de ses enquêtes auprès des ménages en face-à-face. Afin de disposer d'un réseau d'enquêteurs stable et de limiter leurs déplacements, la logique de l'Échantillon-Maître est mise en oeuvre : tous les dix ans, un nouvel ensemble de zones est sélectionné et est mobilisé pour le tirage des logements et des individus.

L'Échantillon-Maître actuellement en vigueur à l'Insee est Octopusse, introduit par Faivre et Christine (2009), et basé sur le recensement rotatif de la population ; ce système de recensement est en lui-même un sondage, comme décrit par Godinot (2005), ce qui complexifie la plupart des sujets de sondage. En particulier, cela demande la construction d'unités primaires spécifiques dont la part recensée chaque année est suffisante pour assurer le tirage des enquêtes de l'institut.

Grâce à une plus grande disponibilité et une meilleure qualité des sources administratives issues des services fiscaux, une alternative a émergé pour les bases de sondage : le Fichier Démographique sur les Logements et les Individus (FIDELI), nouveau nom du Répertoire Statistique des Logements présenté dans Lollivier (2015), qui regroupe les informations fiscales, corrigées des doublons et des règles de gestion administrative. Il a donc été décidé de baser le prochain Échantillon-Maître sur ce fichier, qui présente les caractéristiques d'une bonne base de sondage, en particulier l'exhaustivité.

Par ailleurs, l'échantillon de l'Enquête Emploi en Continu (EEC) arrive également à son terme. Ce renouvellement commun des deux échantillons offre la possibilité de les coordonner. L'objectif de cette coordination est de faire en sorte que les logements enquêtés dans le cadre de l'EEC soient à proximité de ceux collectés pour les autres enquêtes, afin de créer de grandes zones d'activité pour les enquêteurs simplifiant l'organisation du travail.

2 L'Échantillon-Maître Nautile : plan de sondage

2.1 Constitution des unités primaires

On rappelle le principe de l'Échantillon-Maître : il s'agit de réaliser un échantillon géographique de premier degré permettant de concentrer la collecte de plusieurs enquêtes dans les mêmes zones ; le tirage des échantillons relatifs à ces enquêtes se fait alors au second degré au sein des zones sélectionnées. Étant donné qu'il s'agit d'un premier degré, on appelle les zones des unités primaires.

Il est nécessaire tout d'abord de constituer une partition du territoire en unités primaires. La méthode mise en œuvre ici consiste à regrouper des communes afin de créer des zones suffisamment peuplées pour s'assurer qu'on ne réinterroge pas deux fois les mêmes individus tout au lond de la durée de vie de l'Échantillon-Maître, mais suffisamment compactes pour limiter les temps de déplacement lors d'une enquête. Elle se base sur un algorithme solution du problème du voyageur de commerce par Applegate et al. (2003) pour la réalisation d'un chemin optimal de parcours de chaque département français, chemin qui est ensuite découpé en unités primaires. Favre-Martinoz et Merly-Alpa (2017) présentent plus en détail la méthode qui permet de construire les 5 128 unités primaires.

2.2 Optimisation de l'équilibrage

L'Échantillon-Maître étant mobilisé pour le tirage de la plupart des enquêtes auprès des ménages de l'Insee, il apparaît pertinent de chercher à équilibrer ce tirage sur un ensemble de variables socio-économiques le plus large possible. Le tirage se fait à un niveau régional (anciennes régions), pour permettre la diffusion de résultats sur ce domaine, *via* la méthode de sondage équilibré stratifié avec mise en commun des phases d'atterrissage introduite par Chauvet (2009).

Par ailleurs, des travaux préliminaires de Favre-Martinoz et Merly-Alpa (2016) ont permis de montrer que la méthode de sondage spatialement équilibré apportait des bénéfices importants pour le tirage d'un Échantillon-Maître. En effet, comme la méthode permet de limiter la sélection d'unités proches géographiquement et donc partageant des caractéristiques socio-économiques proches, elle améliore la précision de variables, en particulier celles n'ayant pas pu être incluses dans les variables d'équilibrage ; elle favorise également une moindre détérioration de la précision des variables dans le temps.

Il aurait été possible d'utiliser d'autres méthodes d'échantillonnage spatial telles que le pivot spatial, mais les gains en précision semblaient moins homogènes sur l'ensemble des variables suivant la définition de la distance choisie selon Givois et Merly-Alpa (2018). Le choix de réaliser un échantillonnage spatialement équilibré de 541 unités primaires a donc été fait. La question restante est alors le choix des variables d'équilibrage, décrite en détail par Guillo et Merly-Alpa (2018) et résumée ici.

Du fait de l'utilisation de la commune comme brique de constitution des unités primaires, nous disposons de nombreuses variables mobilisables, issues du recensement de la population française et de diverses sources administratives. On sait cependant que l'intégration d'un trop grand nombre de variables d'équilibrage peut en dégrader la qualité ; en particulier, les premières simulations réalisées avec l'intégralité du jeu de variables d'équilibrage conduisaient à ne pas respecter la contrainte de taille fixe et ainsi à avoir un nombre de zones d'enquête variable.

Une piste pour réduire le nombre de variables d'équilibrage a été de synthétiser l'information contenue dans l'ensemble des variables par des méthodes d'analyse de données. L'application d'une analyse en composantes principales (ACP) au jeu de données au niveau unités primaires a ainsi permis de faire émerger une quinzaine d'axes expliquant 99 % de l'inertie du nuage de données. L'équilibrage sur ces axes, en suivant la logique de Le Gleut (2017), permet alors de gagner en précision dans l'estimation des variables du nuage ainsi que de celles qui y sont corrélées, tout en respectant bien mieux la contrainte de taille fixe.

L'approche finale retenue est une approche hybride. En effet, la méthode mobilisant l'ACP permet de gagner en précision mais de façon moins marquée que l'équilibrage direct sur les variables considérées. Ainsi, deux groupes de variables ont été identifiés : les variables principales, sur lesquelles l'équilibrage se fait sans transformation (par exemple, les revenus, la pyramide des âges...), et les variables plus secondaires (par exemple, les zonages géographiques, les types de revenus...), qui sont intégrées *via* les axes de l'ACP.

3 Tirage de l'enquête Emploi

3.1 Description de l'enquête

L'enquête Emploi en Continu, ou EEC, est une enquête visant à observer à la fois de manière structurelle et conjoncturelle la situation des personnes sur le marché du travail. Il s'agit du pendant français de la "Labour Force Survey" (LFS). En France, il s'agit de la seule source fournissant une mesure des concepts d'activité, de chômage, d'emploi et d'inactivité tels qu'ils sont définis par le Bureau International du Travail (BIT).

Cette enquête est aréolaire : son échantillon est un ensemble de zones géographiquement compactes appelées grappes, regroupées en secteurs. Une grappe est un ensemble d'une vingtaine de logements, et chaque secteur contient six ou sept grappes. Chaque trimestre, une grappe de chaque secteur est interrogée exhautivement dans un délai de deux semaines ; au bout de six interrogations de la même grappe, elle est remplacée par la suivante au sein du même secteur. Ce mode d'interrogation, mis en place par Loonis (2009) pour une durée de 9 ans, est reconduit pour le futur échantillon EEC.

3.2 Constitution des grappes et secteurs

La grappe Emploi est un ensemble d'une vingtaine de logements les plus proches possibles afin de permettre à un enquêteur de collecter l'ensemble des questionnaires en face-à-face dans les délais très contraints : en effet, comme chaque grappe est affectée à une semaine de référence, il est nécessaire de collecter l'information dans les deux semaines qui suivent afin d'éviter des effets mémoire néfastes à la qualité de la réponse. Par

ailleurs, une autre contrainte est l'intégration complète des logements d'un même étage d'un immeuble à la même grappe, afin de faciliter le travail de repérage. De même, il est préférable pour le repérage de réduire le nombre d'immeubles différents inclus dans une grappe. Il s'agit donc de trouver un découpage de l'ensemble des logements en France qui respecte toutes ces contraintes.

La méthode mise en œuvre pour l'échantillon actuel, détaillée par Loonis (2009), est de suivre un chemin entre logements basé sur l'ordre de tri des variables liées au cadastre et au bâti des logements. Elle permet de construire des grappes respectant les contraintes, mais au prix d'une variabilité non négligeable dans leurs nombres de logements. Ici, il s'agit d'améliorer cette approche en multipliant les chemins possibles, *via* l'utilisation des coordonnées géographiques des logements comme base du problème du voyageur de commerce. On suit alors chaque chemin solution en constituant pas-à-pas des grappes.

Cette méthode a été améliorée sur plusieurs aspects. Tout d'abord, afin de limiter le temps machine nécessaire, et pour contraindre l'étendue géographique des grappes, cette constitution est réalisée au sein des iris (Ilots Regroupés pour l'Information Statistique), qui sont des zones infracommunales de quelques milliers d'habitants. Ce découpage, qui suit les frontières naturelles, permet également de limiter les temps de déplacements effectifs (pas de grappes de part et d'autre d'un fleuve, par exemple).

D'autre part, la contrainte au niveau des étages demande un traitement spécifique des grands immeubles : il est plus simple de commencer par essayer de constituer des grappes d'une vingtaine de logements au sein de chaque immeuble, et ensuite de le rattacher au chemin général. Cependant, même avec cette adaptation, il reste des grands immeubles et il est donc nécessaire de laisser un peu de liberté à la constitution des grappes en permettant de « sauter » certains immeubles du chemin pour éviter des effets d'accumulation. Ces améliorations sont décrites en détail par Costa et al (2018).

Enfin, ces grappes sont regroupées en secteurs. Le rôle du secteur dans le plan de sondage de l'EEC est de permettre une amélioration de la précision longitudinale de l'estimation des taux de chômage et d'emploi lors des trimestres où l'on arrête la collecte d'une grappe pour en commencer une nouvelle du même secteur. L'objectif est donc de créer des secteurs homogènes d'un point de vue socio-économique afin de minimiser la perte en précision liée à ce changement de grappe d'interrogation. On dispose finalement de 1 395 866 grappes qui composent 231 966 secteurs.

3.3 Construction des variables d'équilibrage

Contrairement à ce qui a été présenté pour l'Échantillon-Maître plus haut, l'échantillon de l'EEC n'a pas vocation à être représentatif sur un large panorama de sujets. L'enquête

Emploi se concentre sur les thématiques liées au marché du travail, et il est nécessaire d'être précis uniquement sur ces sujets, éventuellement avec des déclinaisons selon des domaines de diffusion.

L'approche usuellement suivie dans ce genre de cas est d'utiliser des variables socio-démographiques liées à la commune (taux de pauvreté, richesse de la commune, urbanisation...) comme variables d'équilibrage. Nous proposons deux améliorations.

Tout d'abord, comme évoqué plus haut, nous utilisons la méthode de sondage spatialement équilibré pour sélectionner les 2 944 secteurs de l'échantillon. Cette méthode est efficace ici car on sait que les caractéristiques par rapport à l'emploi sont soumis à une autocorrélation spatiale significativement positive comme étudié par Floch et Le Saout (2015).

D'autre part, les variables au niveau communal sont remplacées par des variables *proxy* des concepts de l'enquête, c'est à dire des variables au niveau individu ou logement construites à partir des informations de la base pour se rapprocher au maximum du concept d'intérêt de l'enquête. Nous approximons ainsi le statut de chômeur en fonction de la perception ou non d'allocations liées à la recherche d'emploi, ainsi que celui d'actif en fonction de la perception de salaires ou de revenus *via* une activité en tant qu'indépendant. Ces variables sont alors utilisées pour l'équilibrage spécifique de l'échantillon.

4 La coordination des deux échantillons

4.1 Logique

Le tirage de l'Échantillon-Maître d'une part, et de l'échantillon de l'EEC d'autre part, consiste en la sélection de zones géographiques au sein desquelles les enquêteurs vont réaliser les entretiens. Les échantillons mobilisés actuellement n'ont pas été tirés de concert : il n'y a aucune raison qu'un secteur Emploi tiré soit proche ou éloigné d'une unité primaire sélectionnée. Le seul élément pris en compte est la disjonction : pour éviter de réinterroger les mêmes ménages, l'échantillon de l'EEC est retiré des unités primaires sélectionnées qu'il intersecte.

Cette gestion indépendante pose plusieurs problèmes. Tout d'abord, elle implique des déplacements d'assez longue durée lorsqu'un enquêteur doit atteindre un secteur Emploi isolé. Dans les cas les plus extrêmes, cela peut même entraîner un défaut de collecte pour les secteurs trop isolés. Par ailleurs, la dispersion géographique des zones de collecte limite les possibilités de remplacement entre enquêteurs, en cas d'indisponibilité de longue durée (maladie...).

Une solution pour pallier ces problèmes est donc de concentrer la collecte, c'est-à-dire de faire en sorte que les secteurs et les unités primaires tirés soient proches ; l'inconvénient de cet approche est l'effet de grappe engendré, qui risque de détériorer la précision des estimations issues des enquêtes. Il faut donc trouver une méthode qui assure une bonne qualité des chiffres produits.

4.2 Constitution des unités de coordination

Plusieurs méthodes sont envisageables pour réaliser la coordination de deux tirages. La première approche consiste à tout simplement tirer les secteurs au sein des unités primaires : cela garantit une proximité très forte entre les deux tirages. Cependant, cette méthode conduit à voir que le risque d'épuisement des zones (en raison de la règle de non-réinterrogation) est très fort et que la précision de l'EEC diminue fortement, du fait de la concentration du tirage.

D'autres méthodes existent mais ne sont pas nécessairement compatibles avec les plans de sondages utilisés. La méthode de coordination basée sur le pivot spatial présentée par Matei et Grafström (2016) répond à notre problématique, mais demande des changements importants sur les plans de sondage choisis, ce qui peut altérer les qualités statistiques des échantillons.

La piste suivie ici est de constituer des zones plus étendues englobant les unités primaires et de considérer la coordination comme le tirage au sein de cette zone, appelée unité de coordination (UC), d'unités primaires et de secteurs Emploi. La méthode mise en œuvre pour construire ces zones est l'agrégation d'unités primaires de sorte à garantir une taille minimale à chaque UC de 10 000 résidences principales. L'agrégation choisie est celle qui réduit l'étendue des UC constituées, afin de s'assurer que la coordination entraîne bien une baisse des temps de déplacement. On aboutit à un jeu de 1 646 UC découpant la France.

4.3 Équilibrage indirect

Une fois la piste des unités de coordination explorée, deux alternatives sont possibles. On peut décider d'échantillonner des UC, puis des unités primaires et des secteurs au sein des UC tirées (méthode directe) ; ou on peut décider d'échantillonner des unités primaires, en déduire les UC sélectionnées et tirer les secteurs au sein de ces UC (méthode indirecte, en référence au sondage indirect des UC *via* les unités primaires, concept introduit par Deville et Lavallée (2006)).

Plusieurs arguments et résultats conduisent à favoriser la méthode indirecte. Tout d'abord, il semble difficile de combiner un équilibrage des unités primaires avec la contrainte

de tirer une unité primaire au sein de chaque unité de coordination. Or, si l'on autorise des UC sélectionnées à ne pas contenir d'unité primaire tirée, on réduit les avantages de la coordination. D'autre part, les travaux de simulation de tirage menés montrent que la précision de nombreuses variables au niveau de l'Échantillon-Maître est meilleure dans la méthode indirecte que pour la méthode directe. Ces travaux sont détaillés par Paliod et al (2018). Par ailleurs, les calculs analytiques de précision pour les enquêtes tirés dans l'Échantillon-Maître sont plus difficiles à réaliser dans le cadre de la méthode directe, car ils impliquent une phase supplémentaire pour les enquêtes ménages ainsi que le tirage d'une unique unité au sein d'une strate – l'UC.

La méthode indirecte demande néanmoins une amélioration. En effet, le tirage indirect des UC ne garantit pas la qualité de l'échantillon d'UC obtenu. Or, comme décrit plus haut, le tirage des secteurs de l'EEC est équilibré sur des variables spécifiques. Pour permettre à cet équilibrage d'être de bonne qualité, il est nécessaire que l'univers de tirage, c'est-à-dire l'ensemble des UC sélectionnées par sondage indirect, possède des caractéristiques similaires à la population totale.

Pour garantir cette propriété, il faut se placer dans le cadre du partage des poids défini par Deville et Lavallée (2006), car chaque unité de coordination peut être atteinte par n unités primaires différentes (correspondant au nombre de liens), et en particulier peut être captée plusieurs fois si plusieurs de ces unités primaires sont sélectionnées. L'idée mise en place est alors d'introduire des variables transformées dites variables indirectes permettant l'équilibrage de l'univers des UC sélectionnées. Pour chaque variable X, pour chaque unité primaire i parmi l'ensemble UP_j des unités primaires composant une unité de coordination j, on introduit la variable transformée à partir de sa valeur sur l'UC X_j comme suit :

$$\tilde{X}_i = \frac{\pi_i}{\sum_{k \in UP_j} \pi_k} X_j \tag{1}$$

où π_i est la probabilité de sélection simple de l'unité primaire i dans le plan de sondage de l'Échantillon-Maître qui représente ici la force du lien dans le cadre de la méthode généralisée de partage des poids. Équilibrer sur les $\tilde{X}_i$ le tirage de l'Échantillon-Maître permet alors de garantir que les unités de coordination seront de bonne qualité pour le tirage en phase suivante des secteurs Emploi ; les simulations de tirage, décrites par Paliod et al (2018), montrent que cela n'affecte que très peu la qualité de l'équilibrage des unités primaires.

5 Conclusions et perspectives

Cette communication présente une méthode permettant de coordonner deux échantillons de zones géographiques avec un impact limité sur la précision des deux enquêtes, *via* l'introduction d'unités de coordination et leur tirage indirect équilibré. Des pistes d'études complémentaires existent : la question de la constitution d'unités de coordination plus homogènes ou plus hétérogènes peut avoir un impact important sur la précision, par exemple ; la méthode de calcul des probabilités d'inclusion en seconde phase des secteurs Emploi peut impliquer différentes propriétés en termes de taille fixe, de répartition des secteurs entre UC et d'équipondération des secteurs.

Par ailleurs, la méthode d'équilibrage indirect présentée ici pourrait avoir d'autres applications, par exemple dans le cadre de tirage de petits échantillons que l'on souhaite équilibrer. L'idée pourrait être par exemple de tirer un échantillon de salariés équilibré sur les variables transformées comme dans (1) pour obtenir par sondage indirect un petit échantillon d'entreprises équilibré. Ces autres applications restent à explorer.

Bibliographie

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

Chauvet G. (2009), Stratified balanced sampling, *Survey Methodology*, Vol. 35, No. 1, pp. 115-119.

Costa, L., Merly-Alpa, T. et Chevalier, M. (2018), Le renouvellement de l'échantillon Emploi : améliorations et évolutions, *Actes des Journées de Méthodologie Statistique de 2018, Insee.*

Deville, J.-C., Lavallée, P. (2006), Sondage indirect : les fondements de la méthode généralisée du partage des poids. *Techniques d'enquête*, Vol. 32, No 2, p. 185.

Faivre, S et Christine, M. (2009). Le projet OCTOPUSSE de nouvel Échantillon-Maître de l'Insee, *Actes des Journées de Méthodologie Statistique de 2009, Insee.*

Favre-Martinoz, C. et Merly-Alpa, T. (2016), Utilisation des Méthodes d'Échantillonnage Spatialement Équilibre pour le Tirage des Unites Primaires des Enquêtes Ménages de l'Insee, *9eme Colloque Francophone sur les Sondages, Gatineau.*

Favre-Martinoz, C. et Merly-Alpa T. (2017), Constitution et Tirage d'Unités Primaires

pour des sondages en mobilisant de l'information spatiale, *49èmes Journées de Statistique, Avignon*.

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

Givois, S. et Merly-Alpa, T. (2018), Échantillonnage spatial *via* des distances socio-économiques : comparaison de méthodes pour le tirage d'un Échantillon-Maître, *Actes des Journées de Méthodologie Statistique de 2018, Insee*.

Godinot, A. (2005). Pour comprendre le recensement de la population, *Insee Méthodes*, hors série - mai 2005, `https://insee.fr/fr/information/2579979`.

Guillo, C. et Merly-Alpa, T. (2018), Un nouvel Échantillon-Maître pour 2020 et pour Nautile, *Actes des Journées de Méthodologie Statistique de 2018, Insee*.

Le Gleut, R. (2017), Analyse Factorielle et Sondage - Utilisation de Méthodes d'Échantillonnage Spatial, *49èmes Journées de Statistique, Avignon*.

Lollivier S. (2015). Le répertoire statistique des logements, *Commission Territoires du CNIS*, `https://www.cnis.fr/wp-content/uploads/2017/09/DC_2015_1re_reunion_COM_Territoires_RLS.pdf`.

Loonis, V. (2009), La construction du nouvel échantillon de l'Enquête Emploi en Continu à partir des fichiers de la Taxe d'Habitation, *Actes des Journées de Méthodologie Statistique de 2009, Insee*.

Matei, A. et Grafström, A. (2016), Coordination des échantillons dans l'échantillonnage spatial, *9eme Colloque Francophone sur les Sondages, Gatineau*.

Paliod, N., Chevalier, M. et Deroyon, T. (2018), Coordination spatiale d'échantillons : application à l'EEC et l'Échantillon-Maître, *Actes des Journées de Méthodologie Statistique de 2018, Insee*.