Recueil du Symposium 2018 de Statistique Canada
Combiner pour vaincre : innovations dans l'utilisation de multiples sources de données

Passer du recensement aux sources fiscales : un changement de base de sondage pour une meilleure coordination des échantillons de l'Insee

Thomas Merly-Alpa et Ludovic Vincent[1]

Résumé

L'Institut National de la Statistique et des Études Économiques (Insee) a développé une alternative au recensement pour ses bases de sondage basée sur les informations fiscales corrigées des règles de gestion administrative. Ce changement entraîne la modification des variables disponibles pour le tirage des échantillons et pour la collecte ; en parallèle l'utilisation de méthodes d'échantillonnage avancées telles que l'équilibrage spatial pour la sélection des zones de collecte permet d'améliorer la précision des enquêtes de l'institut. Enfin, l'utilisation d'une base unique a permis de réaliser un tirage coordonné de l'enquête Emploi avec l'Échantillon-Maître de l'Insee, et ainsi faciliter le travail des enquêteurs.

Mots Clés : Échantillonnage, bases administratives, Échantillon-Maître, équilibrage spatial, coordination d'échantillons, sondage indirect.

1. Introduction

Grâce à une plus grande disponibilité et une meilleure qualité des sources administratives issues des services fiscaux, l'Institut National de la Statistique et des Études Économiques (Insee) a développé une alternative pour ses bases de sondage : le Fichier Démographique sur les Logements et les Individus (Fidéli), qui regroupe les informations fiscales, corrigées des doublons et des règles de gestion administrative.

L'utilisation de Fidéli à la place du recensement de la population, traditionnellement utilisé comme base de sondage à l'Insee, permet de diminuer l'étendue des zones d'enquêtes et d'améliorer leur plan de sondage. Cependant, ce changement entraîne la disparition d'informations, la modification de concepts et l'apparition de nouvelles variables, notamment pour le repérage des enquêtes qui pourra se baser sur une géolocalisation plus précise.

Le renouvellement de l'Échantillon-Maître, prévu pour 2020 dans le cadre du projet de Nouvelle Application Utilisée pour le Tirage des Individus et des Logements des Enquêtes (Nautile), ainsi que celui de l'échantillon de l'enquête Emploi, sont basés sur ce fichier, qui présente les caractéristiques d'une bonne base de sondage, en particulier l'exhaustivité.

La coordination de ces deux tirages a été étudiée afin de limiter les déplacements des enquêteurs. Les unités primaires de l'Échantillon-Maître sont tirées par échantillonnage spatialement équilibré ; l'échantillon de l'enquête Emploi est un ensemble de grappes compactes d'une vingtaine de logements également sélectionnées par sondage spatialement équilibré sur des variables proxy de la situation vis-à-vis du marché du travail. Enfin, leur coordination se fait par l'introduction d'unités de coordination sélectionnées par sondage indirect via les unités primaires, les grappes emploi étant sélectionnées dans les unités de coordination dans une dernière phase.

2. L'utilisation des sources fiscales comme base de sondage

[1]Thomas Merly-Alpa, INSEE, 88 avenue Verdier Montrouge, France, 92120 (thomas.merly-alpa@insee.fr); Ludovic Vincent, INSEE, 88 avenue Verdier Montrouge, France, 92120 (ludovic.vincent@insee.fr)

2.1 La collecte des enquêtes à l'Insee

L'Insee réalise chaque année différentes enquêtes auprès des ménages (sur les loyers, les conditions de vie, etc.). La particularité de ces enquêtes réside dans le fait que nombre d'entre elles se font en face-à-face et demandent la présence d'un enquêteur. Pour pouvoir réaliser ces enquêtes, l'Insee a mis en place une méthodologie consistant :

- à sélectionner un échantillon de zones de collecte (les « unités primaires ») d'étendue acceptable représentant l'ensemble du territoire français
- puis, au sein de chacune des zones sélectionnées, et pour chaque enquête ménages, à échantillonner les logements qui seront interrogés.

Cet échantillon de zones est appelé « Échantillon-Maître ».

L'Échantillon-Maître actuel, introduit par Faivre et Christine (2009), a été mis en place en 2009 pour une durée de vie de 10 ans. Les zones constituées devaient respecter de nombreuses contraintes :

- contenir des logements de chacun des groupes de rotation du recensement de la population,
- être les moins étendues possibles, afin de limiter les déplacements des enquêteurs,
- être de taille suffisamment grande (en nombre de logements) afin d'éviter d'interroger deux fois un même logement sur une période de 5 ans.

Ainsi, le système actuel est très adhérent au recensement de la population, dont le processus est décrit dans Godinot (2005), tant d'un point de vue statistique (utilisation des variables du recensement pour la stratification par exemple) que pour la collecte (utilisation de l'image de l'adresse complétée sur le bulletin de collecte du recensement).

2.2 Une nouvelle source : Fideli

Pour 2019, la question du renouvellement de l'Échantillon-Maître s'impose à l'Insee. En prévision de celui-ci, l'Insee, en s'appuyant sur des premières analyses menées auparavant par Hallépée, Pendoli et Sautory (2018), a entrepris des travaux sur l'équilibre des groupes de rotation du recensement, constitués à partir des résultats du recensement exhaustif de 1999. Ces travaux ont montré des évolutions socio-démographiques différenciées d'un groupe à l'autre, qui augmentent année après année. Ainsi, le déséquilibre croissant des groupes de rotations du recensement au cours du temps, pénalise directement la qualité de l'Échantillon-Maître actuel et de tout Échantillon-Maître issu de la même méthode de tirage.

Depuis le dernier Échantillon-Maître, une nouvelle source, Fidéli (Fichier démographique des logements et des individus) est apparue, ouvrant de nouvelles opportunités, tant pour l'échantillonnage des enquêtes que pour la constitution des zones ou le repérage des unités à interroger.

Fidéli, présenté dans Lollivier (2015), est un fichier d'individus et de logements issu des fichiers fiscaux, apurés, complétés par le répertoire des communautés et celui des résidences hôtelières, et enrichis d'informations de géolocalisation (coordonnées, zonage) et d'informations sur les revenus (issues du Fichier Localisé Social et Fiscal – Filosofi). Ces différents traitements permettent d'obtenir chaque année un fichier présentant les propriétés d'une bonne base de sondage :

- l'exhaustivité : contrairement au recensement, la source fiscale concerne l'ensemble de la population, ce qui présente un avantage conséquent pour la précision des enquêtes auprès des ménages
- l'unicité (absence de doublons) : les travaux effectués par l'Insee sur les fichiers fiscaux permettent d'apurer les données et d'assurer (autant que possible) l'unicité de chaque individu et chaque logement, facilitant ainsi la collecte et la conception des plans de sondage, et améliorant également la précision des enquêtes.
- l'acualité des données : la remontée des données fiscales à l'Insee et les traitements effectués permettent d'obtenir en 18 mois une base de sondage complète, ce qui représente un avantage par rapport au recensement qui n'était aussi « rapide » que pour la dernière enquête annuelle de recensement, soit 1/5 de la population.

Par ailleurs, le passage à Fidéli offre la possibilité de sélectionner des individus. En effet, l'utilisation des fichiers fiscaux permet d'avoir les variables d'identification et des informations suffisantes sur chaque personne pour repérer avec précision les individus d'intérêt. L'Insee a ainsi décidé de tirer les nouveaux échantillons des futures enquêtes

ménages du service statistique public dans Fidéli.

2.3 De nouvelles variables pour l'échantillonnage et la collecte

Le passage du recensement à Fidéli ne se limite pas à l'abandon des 5 groupes de rotation pour une base exhaustive. Ainsi, comme tout changement de source, cela s'accompagne de la disparition de certaines variables, de l'apparition d'autres et de certains changements de concepts. Par exemple, des données telles que le niveau de diplôme ou la catégorie socio-professionnelle des salariés ne pourront plus être utilisées directement, car elles n'existent pas dans les fichiers fiscaux. L'utilisation d'un proxy ou le recours à d'autres sources, à des niveaux moins fins seront nécessaires. À l'inverse, d'autres variables pourront être utilisées par les enquêtes comme les détails sur les revenus. Il sera également possible, grâce aux variables de géolocalisation, d'apprécier plus précisément la position des unités de collecte.

Enfin, certaines variables (ou concepts) sont présentes dans les deux sources, mais leur définition peut légèrement différer, ou leur qualité être plus ou moins bonne (par exemple, la variable sur le logement social, ou le statut d'une résidence – principale, secondaire.). Globalement, les variables à disposition dans Fidéli sont beaucoup plus nombreuses que dans le recensement, et leur qualité est appelée à s'améliorer année après année.

L'impact de ces modifications sur les enquêtes se situe à trois niveaux :

- En amont du tirage, la stratification de l'échantillon portera sur des variables différentes, plus proches ou plus éloignées des sujets d'étude en fonction des enquêtes. De plus, la traduction de « logement ordinaire au sens du recensement » comme unité d'intérêt, ou plus généralement le champ d'interrogation des enquêtes pourra changer par rapport à ce qui est fait actuellement.

- En cours de collecte, le repérage des unités interrogées ne pourra s'appuyer sur les mêmes données qu'elles utilisent actuellement, issues des bulletins de recensement. Si les informations de collecte du recensement ne seront plus disponibles, Fidéli met à disposition plusieurs adresses de qualité différente, mais complémentaires, qui permettront de trouver le logement. Par ailleurs, des informations supplémentaires (mail, numéro de téléphone, coordonnées géographique…) pourraient être mises à disposition, permettant d'améliorer le repérage des unités, en faisant évoluer la méthode associée.

- En aval de la collecte, la base de sondage peut être utilisée pour le redressement des données. Ainsi, les enquêtes utilisant les variables du recensement pour la correction de la non-réponse totale devront revoir leur procédure pour les adapter à la nouvelle base de sondage. Les variables de calage, souvent utilisée à des niveaux plus agrégés (niveau communal et non niveau de l'unité échantillonnée), ne devraient pas subir de modifications. Cela étant, Fidéli pourrait apporter de nouvelles variables, utiles pour certaines enquêtes dans l'amélioration du calage.

3. La coordination de l'Échantillon-Maître et de l'enquête Emploi

3.1 L'Échantillon-Maître Nautile

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. Ainsi, il est nécessaire tout d'abord de constituer une partition du territoire en unités primaires. La méthode mise en oeuvre 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 long 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.

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. 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.

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 conduisent 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, développée par Guillo (2018), 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 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 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.

3.2 L’enquête Emploi

L'enquête Emploi 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. 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 exhaustivement 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.

Contrairement à ce qui a été présenté plus haut, l'échantillon de l'enquête Emploi 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 ; deux améliorations sont possibles dans ce but. Tout d'abord, comme évoqué plus haut, nous utilisons la méthode de sondage spatialement équilibré pour sélectionner les 2 944 secteurs de l'échantillon. 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. On utilise ici la perception d’allocations compensatrices à la recherche d’emploi comme proxy de la situation de chômeur, même si les deux situations ne sont pas évidemment pas équivalentes. Le choix des variables est détaillé par Costa (2018).

3.3 Une coordination des deux échantillons

Le tirage de l'Échantillon-Maître d'une part, et de l'échantillon de l'enquête Emploi 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 Emploi 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. 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 cette 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; plusieurs méthodes possibles sont décrites dans Matei (2016).

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, d'unités primaires et de secteurs Emploi. Deux alternatives sont alors possibles. On peut décider d'échantillonner des unités de coordination, puis des unités primaires et des secteurs au sein des unités de coordination tirées (méthode directe) ; ou on peut décider d'échantillonner des unités primaires, en déduire les unités de coordination sélectionnées et tirer les secteurs au sein de ces unités de coordination (méthode indirecte, en référence au sondage indirect des unités de coordination 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 unités de coordination 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.

La méthode indirecte demande néanmoins une amélioration. En effet, le tirage indirect des unités de coordination ne garantit pas la qualité de l'échantillon d'unités de coordination 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 unités de coordination 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 par Paliod (2018) est alors d'introduire des variables transformées dites variables indirectes permettant l'équilibrage de l'univers des unités de coordination sélectionnées ; ces variables sont construites en considérant le nombre de liens permettant d'atteindre une unité de coordination.

4. Conclusion

La transition du recensement aux sources fiscales pour la constitution des bases de sondage de ses enquêtes auprès des ménages est une opportunité pour l'Insee de rénover et d'améliorer ses méthodes d'échantillonnage (équilibrage spatial, nouvelles variables proxy des sujets d'enquête, coordination), son organisation de la collecte (regroupement des unités primaires et des secteurs Emploi, présence de nouvelles informations pour le contact) mais aussi les post-traitements des enquêtes (nouvelles variables, nouvelles marges de calage). Cependant, cette évolution reste un défi tant organisationnel que méthodologique, dont certains aspects restent encore à instruire ; d'autant plus que la stabilité des sources administratives fiscales n'est pas complètement assurée : passage au prélèvement à la source dès janvier 2019, fin possible de la taxe d'habitation…

Bibliographie

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

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

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

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

Favre-Martinoz, C., et T. Merly-Alpa (2017), « Constitution et tirage d'unités primaires pour des sondages en mobilisant de l'information spatiale », 49èmes Journées de Statistique, Avignon.

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

Godinot, A. (2005), « Pour comprendre le recensement de la population », Insee Méthodes, hors série - mai 2005.

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

Hallépée, S., P.A. Pendoli et O. Sautory (2018), « La re-pondération des enquêtes annuelles de recensement pour une diffusion complémentaire du RP », Actes des Journées de Méthodologie Statistique de 2018, Insee.

Lollivier, S. (2015). « Le répertoire statistique des logements », Commission Territoires du CNIS.

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

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