PASSER DU RECENSEMENT AUX SOURCES FISCALES POUR LE NOUVEL ÉCHANTILLON-MAÎTRE DE L'INSEE : LE PROJET NAUTILE (NOUVELLE APPLICATION UTILISÉE POUR LE TIRAGE DES INDIVIDUS ET DES LOGEMENTS DES ENQUÊTES)

Ludovic VINCENT[1] & Sébastien FAIVRE[2]
[1] *Insee, 88, avenue Verdier, 92120 Montrouge, ludovic.vincent@insee.fr*
[2] *Insee, 88, avenue Verdier, 92120 Montrouge, sebastien.faivre@insee.fr*

Résumé. Actuellement, les enquêtes ménages de l'INSEE sont tirées, pour la majorité, dans le recensement de la population (RP). Pour réaliser ces enquêtes, la plupart en face-à-face, l'Insee a mis en place une méthodologie consistant à sélectionner un échantillon de zones (appelé Échantillon-Maître (EM)), puis, au sein de chaque zone sélectionnée à échantillonner les logements interrogés. En 2019, le renouvellement de l'EM, mis en place en 2009 pour 10 ans, s'impose à l'Insee.

Pour cela, l'Insee a mené des études montrant un déséquilibre croissant des groupes de rotations du RP au cours du temps, dégradant la qualité de tout EM issu de la méthode actuelle. Par ailleurs, une nouvelle source, Fidéli (Fichier démographique des logements et des individus), a ouvert de nouvelles opportunités. Ce fichier, mis à jour chaque année à partir des fichiers fiscaux, présente les propriétés d'une bonne base de sondage : exhaustivité, unicité, fraîcheur des données. L'utilisation de Fidéli à la place du RP permet de diminuer l'étendue des zones d'enquêtes et d'améliorer leur plan de sondage. Elle rend également possible le tirage d'individus. Cependant, ce changement entraîne la disparition d'informations, la modification de concepts et l'apparition de nouvelle variables, engendrant des conséquences sur l'échantillonnage et le redressement des enquêtes qui seront à étudier. De même, passer à Fidéli modifie considérablement le repérage des enquêtes qui pourra se baser sur une géolocalisation plus précise.

Cette communication présente les raisons et conséquences du passage du RP à Fidéli comme base de sondage pour les enquêtes ménages de l'Insee.

Mots-clés. Échantillonnage, enquêtes ménages, échantillon-maître, sondage en grappes, équilibrage spatialement réparti

Abstract. To realise their social survey, the French national institute of statistic uses the Master sample design: a group of areas representing the French territory is drawn, and for each survey, sample are extracted from these areas. Up to now, this surveys were selected from the annual census survey. But a new data base has appeared, Fidéli, built on administrative data and tax sources. This database is annual, exhaustive and redundancy-free. Its use permits a reduction in the size of areas, and improves the Master sampling design. But the data issued from these sources is different, and some of them are new or are missing ; moreover, some concepts (for example, the primary residence) could change, leading to adjustement of sample designs. Furthermore, the tracking system for intervied people must be reevaluated.

Keywords. Sampling, Social survey, Master sample, cluster sampling, spatial sampling.

1. Un changement de source pour le nouvel échantillon-Maître

L'Insee réalise chaque année différentes enquêtes auprès des ménages. 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 » (EM).

L'EM 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 (RP),

• 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, tant d'un point de vue statistique (utilisation des variables du RP pour la stratification par exemple) que pour la collecte (utilisation de l'image de l'adresse complétée sur le bulletin de collecte du RP).

Pour 2019, la question du renouvellement de l'EM 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 RP, constitués à partir des résultats du recensement exhaustif de 1999.

Ces travaux ont montré des évolutions socio-démographiques différentiées d'un groupe à l'autre, qui augmentent année après année. Ainsi, le déséquilibre croissant des groupes de rotations du RP au cours du temps, pénalise directement la qualité de l'EM actuel et de tout EM issu de la même méthode de tirage.

Depuis le dernier EM, une nouvelle source, Fidéli (Fichier démographique des logements et des individus – ex-RSL pour répertoire statistique des logements) 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 RP, 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.

• la fraîcheur 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 RP qui n'était aussi « rapide » que pour la dernière EAR, 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.

Le comité de direction de l'Insee a ainsi décidé de tirer les nouveaux échantillons des futures enquêtes ménages du service statistique public dans Fidéli.

2. Un EM plus petit et plus précis

L'utilisation de Fidéli dès la constitution des zones de tirage permet de s'affranchir des groupes de rotation du RP, et ainsi de diminuer la surface des zones d'enquêtes tout en améliorer la qualité des zones sélectionnées (et donc des futures enquêtes).

Une nouvelle méthode de constitution et de sélection des zones, validées localement par les directions régionales de l'Insee, a ainsi conduit à l'élaboration d'un échantillon-Maître de 541 unités primaires, à l'étendue plus restreinte que l'EM actuel, mais plus prometteur quant à la précision attendue pour les enquêtes. Le tirage de l'EM est détaillé dans Guillo et Merly-Alpa (2018)

Le passage à Fidéli à également permis d'étudier la coordination de l'échantillon-Maître avec l'échantillon de l'enquête emploi en continu (EEC), facilitant la gestion de la collecte, particulièrement dans les zones rurales. La constitution des nouvelles zones de collecte de l'EEC est présentée dans Costa, Merly-Alpa et Chevalier (2018), et les travaux sur la coordination des deux échantillons dans Paliod, Chevalier et Deroyon (2018).

L'ensemble des travaux méthodologiques ayant conduit à l'EM et au nouvel échantillon de l'EEC sont présentés par ailleurs dans ce colloque (Tirage coordonné d'échantillons : une application à l'Échantillon-Maître Nautile et à l'enquête Emploi par Thomas Merly-Alpa).

3. Des changements à prévoir pour les futurs enquêtes de l'Insee

Le passage du RP 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 RP, 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 RP 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 évolué 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 RP 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.

D'autres impacts, moins liés à la nouvelle source, ont été repérés pour les futures enquêtes.

Ainsi, la modification des zones de collectes peut engendrer un aléa supplémentaire au moment de l'estimation des indicateurs. De même, ce changement pourra complexifier la collecte des enquêtes quand les deux EM coexisteront. De plus, les panels, dont les échantillons porteront en partie sur les deux sources verront leurs chaînes de traitements avales se complexifier.

Enfin, l'abandon du RP a permis la simplification du plan de sondage, engendrant une amélioration du calcul de la précision des enquêtes, jugé comme trop complexe actuellement.

Tous ces changements nécessiteront un investissement conséquent pour de nombreux acteurs de l'Insee, qu'ils soient experts sondages, concepteurs d'enquêtes, responsables de la collecte ou même enquêteurs.

Bibliographie

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.

Hallépée, S., Pendoli, P.A. et Sautory, O. (2018), La repondération 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, https://www.cnis.fr/wp-content/uploads/2017/09/DC_2015_1re_reunion_COM_Territoires_RLS.pdf

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.

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.

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.