



HAL
open science

Pour les enquêtes auprès des ménages, l’Insee rénove ses échantillons

Patrick Sillard, Sébastien Faivre, Nicolas Paliod, Ludovic Vincent

► To cite this version:

Patrick Sillard, Sébastien Faivre, Nicolas Paliod, Ludovic Vincent. Pour les enquêtes auprès des ménages, l’Insee rénove ses échantillons. *Courrier des Statistiques*, 2020, 4, pp.81-100. <hal-05315203>

HAL Id: hal-05315203

<https://insee.hal.science/hal-05315203v1>

Submitted on 15 Oct 2025

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.




Distributed under a Creative Commons CC BY 4.0 - Attribution - International License

POUR LES ENQUÊTES AUPRÈS DES MÉNAGES, L'INSEE RÉNOVE SES ÉCHANTILLONS

Patrick Sillard, Sébastien Faivre**, Nicolas Paliod***, Ludovic Vincent*****

Les enquêtes auprès des ménages de l'Insee sont menées, majoritairement, en face-à-face, via un réseau d'enquêteurs implanté sur tout le territoire. Leur échantillonnage s'appuie sur deux zonages structurants : les secteurs de l'enquête Emploi et, pour toutes les autres enquêtes, les unités primaires de l'échantillon-maître. Depuis les années soixante, l'Insee utilisait la même base de sondage pour ses deux échantillons : le recensement de la population, qui a ainsi imprimé son rythme quasi-décennal aux refontes de l'échantillon-maître et de l'échantillon Emploi. Au début des années deux-mille, l'institut a rénové la méthode de recensement, devenue collecte annuelle sur des échantillons rotatifs de communes et de logements. Une nécessaire adaptation pour l'échantillon-maître a été menée en 2009 ; à la même époque, l'enquête Emploi choisissait d'utiliser des sources fiscales comme base de sondage. Fort de ces deux expériences, l'institut a engagé son projet de renouvellement de l'échantillon-maître et de l'échantillon Emploi en 2017 : utilisant comme base de sondage les fichiers fiscaux, adoptant des méthodes d'équilibrage spatial, le chantier a abouti en 2019 à la coordination des deux grands échantillons de l'Insee, une première dans l'histoire des enquêtes ménages de la statistique publique. L'article décrit les choix méthodologiques opérés : ils seront mis en œuvre pour les enquêtes en face-à-face pendant les dix prochaines années.

 *INSEE conducts face-to-face most of its household surveys, via a network of interviewers located throughout the country. Their sampling is based on two structuring zoning: the sectors of the Labour force survey and, for the other surveys, the primary units of the master sample. Since the 1960s, INSEE has been using the same sampling frame for the both samples: the population census, which has thus given its almost decennial rhythm to the redesign of the master sample and the Labour force survey's sample. In the early two-thousand's, the institute renovated the census method, which became annual collection on rotating samples of municipalities and housings. A necessary adaptation for the master sample was carried out in 2009; at the same time, the Labour force survey chose to base its sampling frame on Tax sources. On the strength of these two experiences, the institute began to renew the master sample and the Labour force survey's sample in 2017: using Tax files as a sampling frame, and adopting spatial balancing methods, the project led in 2019*

* Chef du département des Méthodes statistiques, direction de la Méthodologie et de la coordination statistique et internationale (DMCSI), Insee,
patrick.sillard@insee.fr

** Chef de la division Sondages, DMCSI, Insee,
sebastien.favre@insee.fr

*** Expert méthodologue, division Sondages, DMCSI, Insee,
nicolas.paliod@insee.fr

**** Chef du projet Nautile, division Sondages, DMCSI, Insee,

to the coordination of the two large INSEE samples, a first in the history of household surveys in official statistics. The article describes the methodological choices made: they will be implemented for face-to-face surveys over the next ten years.

Pour construire les échantillons des enquêtes auprès des ménages, l'Insee s'appuie sur un échantillon-maître : sa constitution est structurante pour un institut statistique, car les choix méthodologiques qui sont opérés à cette occasion conditionnent la qualité des enquêtes menées pendant les dix années qui suivront et l'organisation de leur collecte. Longtemps associé aux recensements de la population, l'échantillon-maître a dû d'abord s'adapter aux évolutions méthodologiques du recensement rénové. La disponibilité des sources fiscales est ensuite apparue comme une alternative sérieuse. Leur utilisation pour l'enquête Emploi a confirmé dès 2009 l'intérêt pour l'Insee d'investir dans un nouvel échantillon-maître basé sur les fichiers fiscaux¹ : gains en précision des enquêtes, mais aussi efficacité de la collecte en face-à-face, facilitation du multi-mode, coordination accrue entre les échantillons pour limiter la charge statistique sur les ménages, etc. Cet article résume les étapes qui ont conduit l'Insee à rénover l'échantillon-maître qui sera mis en œuvre à compter de 2020.

ÉCHANTILLON ET QUALITÉ DES ENQUÊTES AUPRÈS DES MÉNAGES

La statistique publique documente les faits économiques et sociaux à l'aide de nombreuses sources de données. Parmi celles-ci, les enquêtes auprès des ménages sont pratiquées de longue date². Elles permettent, en une seule opération, la production d'un vaste panorama des caractéristiques des individus et de leurs interactions avec la thématique d'enquête : l'emploi, la sécurité et les conditions de vie, la santé, les revenus et le patrimoine, le logement, les transports sont autant de domaines qui sont l'objet de grandes enquêtes

« En s'appuyant sur un échantillon de quelques milliers de ménages, on peut prétendre inférer une connaissance en population générale de quelques dizaines de millions. »

de la statistique publique. Leur apport sur la connaissance des faits sociaux est irremplaçable : en s'appuyant sur un échantillon de quelques milliers de ménages, on peut prétendre inférer une connaissance en population générale de quelques dizaines de millions. Mais ceci sous réserve de s'inscrire dans un cadre de référence scientifique rigoureux : celui de la théorie probabiliste des sondages. Ce cadre associe des principes de sélection

aléatoire des individus dans l'enquête, à une méthodologie d'estimation de statistiques sans biais et de précision connue, fondée sur cette enquête. Par exemple, l'enquête Emploi en continu³ de l'Insee, avec un échantillon de 90 000 logements⁴, permet de connaître le taux de chômage trimestriel à $\pm 0,3$ point de pourcentage, dans un intervalle de confiance à 95 %.

1. L'Insee produit annuellement des Fichiers démographiques sur les logements et les individus (Fidéli) à partir des données fiscales sur les personnes physiques (essentiellement impôt sur le revenu, taxe d'habitation, taxe foncière).
2. L'Insee mène des enquêtes auprès des ménages depuis plus de 60 ans.
3. EEC, ou enquête Emploi dans la suite de l'article.
4. L'unité d'observation dans la base de sondage est le logement tandis que l'unité d'intérêt est le ménage, résidant dans ce logement.

Ce principe de sélection aléatoire est le seul qui garantit l'inscription des analyses de l'enquête dans le cadre théorique précédent. Cependant il exige de disposer *a priori* d'une base de sondage, c'est-à-dire d'une liste des ménages ou individus de l'ensemble de la population, ou à défaut, d'un sous-ensemble, lui-même issu d'une sélection aléatoire dans la population générale. Quand aucune base de sondage n'est disponible, des méthodes alternatives comme la méthode des quotas, ont été développées. Elles sont très peu utilisées dans le champ de la statistique publique, car les estimateurs qu'elles permettent de construire ne vérifient pas spontanément les « bonnes » propriétés évoquées *supra*. En effet, ceux-ci ne sont non biaisés et de précision connue que sous l'hypothèse, généralement invérifiable, que conditionnellement aux variables sur lesquelles les quotas sont construits, les réponses apportées aux questionnaires sont aléatoires (Deville, 1991). Le principe de sélectionner des ménages ou des individus aléatoirement dans une base de sondage est général aux enquêtes de la statistique publique, quel que soit le mode

« Le principe de sélectionner des ménages ou des individus aléatoirement dans une base de sondage est général aux enquêtes de la statistique publique, quel que soit le mode de collecte retenu. »

de collecte retenu : intermédié en face-à-face ou par téléphone, auto-administré par questionnaire internet ou papier.

L'interrogation intermédiée en face-à-face est, à ce jour, le mode dominant⁵ pour les enquêtes auprès des ménages⁶ gérées par l'Insee. Ce mode comporte en effet des avantages,

même s'il est plus onéreux. Il permet par exemple de mieux s'assurer de l'identité de la personne interrogée. Il permet surtout d'apprécier si l'enquêté comprend bien le contexte de l'enquête, s'il interprète convenablement les questions, et de prévenir, par l'échange direct avec l'enquêteur, ses éventuelles réserves. Il facilite aussi l'administration d'un questionnement parfois plus long que les autres modes de collecte. Le face-à-face est aussi, naturellement, plus adapté pour atteindre des personnes qui n'auraient pas accès à internet ou qui répondent peu au téléphone. Les autres modes de collecte, téléphone et internet, se développent, mais le questionnement en face-à-face restera vraisemblablement, pour les années prochaines, un mode important dans les enquêtes de la statistique publique.

1 APPORTS, FONCTIONNEMENT ET LIMITES D'UN ÉCHANTILLON-MAÎTRE

La collecte en face-à-face soulève des questions logistiques d'accès aux enquêtés et de limitation des déplacements des enquêteurs. En effet, les ménages susceptibles d'être enquêtés sont répartis sur tout le territoire national : s'ils étaient sélectionnés en population générale à chaque enquête, les coûts de déplacement des enquêteurs seraient prohibitifs⁷.

-
5. L'article a été rédigé avant les événements liés à la crise de la Covid-19. Il est possible que la part du face-à-face dans les modes de collecte s'en trouve durablement réduite.
 6. La question ne se pose pas dans les mêmes termes en statistique d'entreprises, où la collecte des enquêtes est en général en mode auto-administré. Voir (Haag et Husseini-Skalitz, 2019).
 7. En application du cadre d'emploi des enquêteurs de l'Insee, instauré en 2013, les temps de déplacement font partie intégrante du temps de travail, ce qui justifie l'importance accordée à leur optimisation.

Pour pallier ces inconvénients, on prédéfinit des zones géographiques sur lesquelles les ménages enquêtés ont vocation à être sélectionnés, quelle que soit l'enquête. Pour répondre au principe de sélection aléatoire, il convient que les zones géographiques soient elles-mêmes sélectionnées au hasard : on va donc couvrir le territoire national avec une partition géographique de zones, appelées **unités primaires**. Les unités primaires sélectionnées définissent alors les **zones de l'échantillon-maître** de l'Insee. Lorsqu'une enquête est organisée, les ménages à enquêter sont sélectionnés au hasard parmi ceux qui résident dans les zones de l'échantillon-maître (*figure 1*).

Afin de limiter la charge de réponse, lorsqu'un ménage est sélectionné pour une enquête, on cherche à faire en sorte qu'il ne puisse plus l'être par la suite (sauf évidemment dans le cadre d'une enquête panel). Mais de ce fait, les zones de l'échantillon-maître s'épuisent progressivement et il convient, au bout d'une dizaine d'années, de sélectionner un nouvel échantillon-maître. On verra par la suite que ce renouvellement devient également nécessaire pour des raisons d'efficacité de l'échantillonnage initial : en dix ans, la population évolue, se déplace, se concentre différemment sur le territoire ; les critères de représentativité initialement retenus se dégradent progressivement.

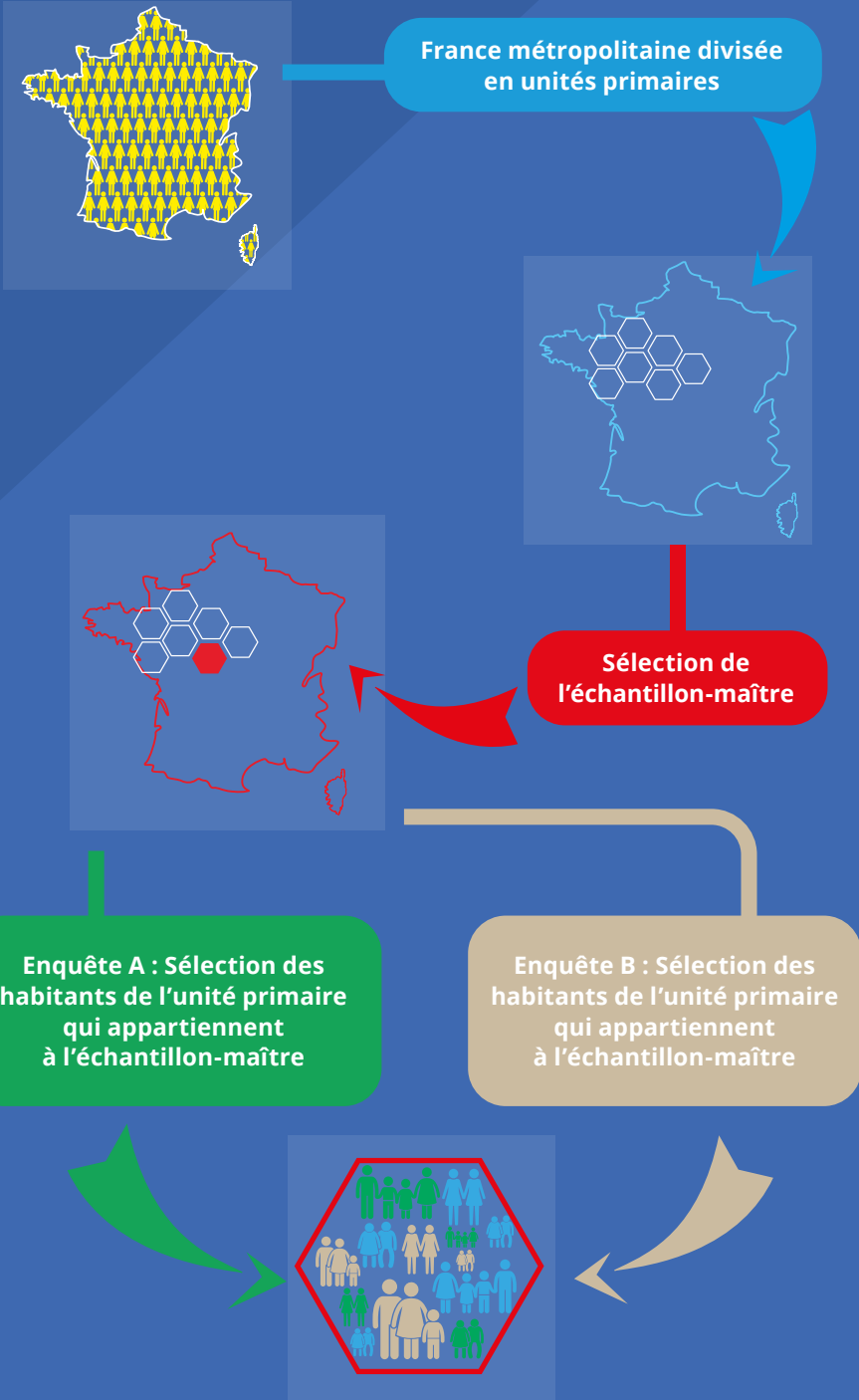
Depuis les années soixante, la **base de sondage** qui permettait de constituer les unités primaires des enquêtes de l'Insee était issue du recensement de la population. Ce recensement, appelé « général » jusqu'en 1999 parce qu'il était exhaustif, a vu sa méthodologie changer à partir de 2004 avec le recensement rénové ou en continu : le recensement quasi-décennal a laissé la place à une collecte tournante, sur cinq ans, et distinguant les communes selon leur taille ; en dessous du seuil de 10 000 habitants, les communes sont recensées exhaustivement tous les cinq ans ; au-dessus de ce seuil, la collecte a lieu chaque année sur un échantillon de 8 % de logements. Cette évolution a conduit à faire évoluer fortement la méthode de constitution de l'échantillon-maître dès 2010 (Christine et Favier, 2009).

UN RENOUVELLEMENT NÉCESSAIRE, DANS UN CONTEXTE NOUVEAU

En 2017, l'Insee disposait d'un échantillon-maître dont les zones, sélectionnées en 2010, s'épuisait peu à peu, et étaient susceptibles, à terme, de moins bien représenter le territoire. Un projet a donc été lancé, visant non seulement à renouveler cet échantillon-maître, mais aussi à en revisiter la méthodologie pour l'adapter à de nouvelles contraintes et à de nouvelles exigences. Car entre-temps, l'institut avait engrangé une expérience nouvelle : l'utilisation des sources fiscales depuis 2009, pour le tirage de l'échantillon de l'enquête Emploi, en lieu et place du recensement (voir *infra*).

En recherche de gains de productivité et de minimisation de ses coûts, la statistique publique se devait aussi de trouver des solutions méthodologiques permettant d'optimiser la précision de ses enquêtes. Car les réflexions s'inscrivaient dans un contexte européen qui évoluait. Le nouveau règlement-cadre européen sur les statistiques sociales IESS (*Integrated European Social Statistics*) fixe désormais un cadre général aux productions statistiques des États membres (Cases, 2019). Il impose, entre autres, des niveaux de précision à atteindre (ou des effectifs minimum pour les échantillons d'enquête) pour certains indicateurs-clés, obtenus par estimation sur données d'enquête, comme le taux de chômage ou le taux de pauvreté, au niveau national ou au niveau régional selon le cas.

Figure 1. Le procédé d'échantillon-maître



Dans les parties qui suivent, la méthodologie de ce nouvel échantillon-maître métropolitain⁸ est présentée, en particulier s'agissant du choix de la base de sondage, de la définition des unités primaires de l'échantillon-maître et de leur tirage. Un dernier paragraphe expose les travaux réalisés sur le renouvellement de l'échantillon de l'enquête Emploi que l'Insee a décidé, pour la première fois en 2019, de tirer de manière coordonnée à l'échantillon-maître.

❶ LE CHOIX DE LA BASE DE SONDRAGE : OPPORTUNITÉS ET ARBITRAGES

L'échantillon d'enquête élémentaire consiste en une liste de ménages ou d'individus que l'enquêteur doit être en mesure d'interroger. Il faut pour cela qu'il puisse disposer d'éléments lui permettant d'entrer en contact avec le ménage sélectionné : adresse de résidence, caractérisation du logement, numéro de téléphone, courriel, etc. La base de sondage doit donc comprendre, outre une liste des individus et ménages, des variables permettant de les contacter.

Elle peut aussi utilement contenir des variables de caractéristiques socio-démographiques des ménages. En effet, si des variables de revenu, de profession, de situation par rapport au marché du travail, etc. sont disponibles dans la base de sondage, alors, dans la mesure où elles

« Le choix de la base de sondage résulte donc d'un arbitrage entre fraîcheur, couverture par rapport à l'exhaustif et qualité de la description des ménages et individus qui y figurent. »

sont corrélées aux variables d'intérêt de l'enquête, elles permettent d'élaborer un plan de sondage qui optimise la précision des estimations issues d'enquêtes, à ressource de collecte donnée.

Le choix de la base de sondage résulte donc d'un arbitrage entre fraîcheur, couverture par rapport à l'exhaustif et qualité de la description des ménages et individus qui y figurent. Nous allons voir comment la bascule a été progressivement opérée entre le choix du recensement et celui de la source fiscale.

❷ JUSQU'EN 2009, LE RECENSEMENT, GÉNÉRAL PUIS RÉNOVÉ...

Lorsque l'Insee réalisait des recensements généraux, le dernier datant de 1999, la base de sondage des enquêtes ménages de l'Insee était constituée simultanément à ces opérations⁹. Les unités primaires étaient sélectionnées à l'issue du recensement exhaustif et les coordonnées (nom, adresse, etc.) des habitants de ces zones étaient conservées pour que ceux-ci puissent, le cas échéant, être interrogés. Un nouveau recensement venait rafraîchir une base de sondage qui, vieillissant au rythme des déménagements, était, au bout d'une décennie, menacée d'obsolescence. Le dernier échantillon-maître constitué sur ce format a perduré, pour le tirage d'enquêtes, jusqu'en 2009. Cette année-là, après dix années d'usage de la base issue du dernier recensement exhaustif de 1999, la nouvelle base de sondage a été ancrée sur le dispositif du recensement en continu.

8. Les DOM font l'objet d'un traitement complémentaire spécifique qui n'est pas abordé dans cet article.

9. La périodicité quasi-décennale des échantillons-maîtres est en grande partie liée à celle des recensements généraux de population.

Comme auparavant, une liste d'unités primaires avait été constituée : mais cette fois, les personnes susceptibles d'être enquêtées une année donnée étaient celles qui avaient été *effectivement recensées* dans ces unités, lors de la *précédente* campagne annuelle du recensement. Au cours des années deux-mille-dix, la plupart des enquêtes auprès des ménages ont été tirées dans cette base. Ce dispositif bénéficiait de la fraîcheur de l'information acquise peu avant le tirage dans le cadre de la dernière campagne annuelle de recensement.

Mais un inconvénient est apparu au fil du temps, lié au mode de tirage des communes dans le recensement. En dessous de 10 000 habitants, l'enquête annuelle de recensement est fondée sur le schéma d'un échantillon de communes rotatif, destiné à les couvrir exhaustivement en 5 ans¹⁰. Cet échantillon annuel de communes résulte d'un tirage équilibré sur des données du recensement de 1999. Après une vingtaine d'années d'exercice, les évolutions démographiques et sociales ont éloigné cet échantillon de la situation équilibrée originelle. Ce défaut, dont l'ampleur dépend de la variable d'intérêt, se traduit en général par une hausse de l'erreur moyenne quadratique d'un rapport 2 à 20, par rapport à un tirage dans une base exhaustive (Favre-Martinoz, 2015).

La fin des années 2010 se rapprochant, le renouvellement de l'échantillon-maître de 2009 s'imposait donc et les premiers travaux méthodologiques ont été lancés en 2017.

... EN 2017, LE CHOIX DES SOURCES FISCALES

L'Insee disposait, depuis une décennie déjà, d'une solide expérience dans l'usage des sources fiscales pour le tirage d'échantillon d'enquête. En effet, l'échantillon de l'enquête Emploi de 2009 avait été tiré grâce aux fichiers de la taxe d'habitation (Loonis, 2009).

« L'échantillon de l'enquête Emploi de 2009 avait été tiré grâce aux fichiers de la taxe d'habitation. »

Ces données offrent une vision quasi-exhaustive des ménages résidant en France et de leurs habitations. Ces logements sont également l'objet de la taxe foncière et du suivi cadastral¹¹. Ils sont donc localisés très finement, à la parcelle cadastrale, soit une précision de quelques mètres des coordonnées géographiques associées. Ceci permet ainsi de constituer des grappes pour l'enquête Emploi, par proximité géographique. C'est

cette particularité qui a justifié, en 2009, de fonder l'échantillon de l'enquête Emploi sur les sources fiscales.

D'autres enquêtes ménages ont été tirées depuis lors dans la source fiscale, notamment celles ayant trait au patrimoine des ménages. L'usage de ces sources se généralisant, l'Insee a mis en place, en 2016, une production annuelle issue de la mise en cohérence statistique des sources fiscales. Les fichiers ainsi créés — appelés Fidéli ou fichiers démographiques sur les logements et les individus — constituent une base d'information de premier choix. Par rapport aux données brutes issues de la gestion fiscale, la source Fidéli porte une attention particulière aux questions de double-compte et de couverture de champ, notamment

10. En dessous du seuil de 10 000 habitants, pour une commune donnée, l'opération de recensement est quinquennale et exhaustive : tous les logements et les habitants de la commune sont recensés la même année.

11. Il s'agit du suivi par les services fiscaux des biens soumis à propriété privée, donc à la taxe foncière.

en rapprochant au maximum les concepts qui y sont utilisés de ceux du recensement. Par exemple, une identification des communautés¹² est réalisée dans le cadre du processus Fidéli, tandis que celles-ci ne sont pas à proprement parler repérées dans les données fiscales brutes (Merly-Alpa et Sillard, 2019).

Par ailleurs, Fidéli comprend un grand nombre de variables intéressantes pour procéder au sondage puis à la collecte des enquêtes : variables de contact (adresse des logements, courriels et coordonnées téléphoniques), variables de description des logements et de composition des ménages, variables de revenu. Celles-ci peuvent être utilisées *ex ante*, pour l'équilibrage ou la stratification, et *ex post*, pour les redressements. Les principales différences entre Fidéli et le recensement portent sur les variables de profession et de catégorie sociale, de diplôme et de revenus : les premières ne figurent que dans le recensement, tandis que la dernière ne figure que dans Fidéli. Les revenus sont connus pour être corrélés avec de nombreuses variables socio-démographiques, donc leur présence dans la base de sondage constitue une avancée importante. Au final, il apparaît donc que Fidéli constitue un candidat solide pour servir de base de sondage principale, en substitution du recensement. C'est le choix qui a été fait, pour la construction de l'échantillon-maître et de l'échantillon de l'enquête Emploi en 2020.

🕒 CONSTRUIRE LES UNITÉS PRIMAIRES : CONTRAINTE DE TAILLE ET DE GÉOGRAPHIE

L'Insee emploie un réseau d'environ 850 enquêteurs qui peuvent raisonnablement assurer la collecte d'une vague d'enquête auprès des ménages, répartie sur environ 500 unités primaires. Les enquêteurs se voient attribuer une zone de collecte, constituée généralement autour d'une unité primaire de l'échantillon-maître. Ces zones doivent donc être d'étendue adaptée pour pouvoir être couvertes par un enquêteur. Si différentes configurations ont été étudiées, il est apparu plus simple, y compris pour l'organisation sur le terrain, de constituer les unités primaires sous forme de réunions de communes contiguës.

Environ 500 000 ménages sont interrogés par l'Insee chaque année. Afin de limiter la charge de collecte pesant sur les ménages, les unités primaires sont aussi conçues pour éviter de ré-interroger un ménage qui resterait dans le même logement sur une période de cinq ans.

Compte tenu de ces différentes contraintes, les unités primaires de l'échantillon des enquêtes ménages comprennent au minimum 2 500 logements, assurant ainsi une réserve en cas d'utilisation plus large.

Enfin, pour faciliter la gestion des enquêteurs et de leur zone de collecte, chaque unité primaire est intégralement contenue dans un département. Ceci permet en outre de déconcentrer à ce niveau le problème d'optimisation sous-jacent.

12. Pour le recensement, la communauté est « un ensemble de locaux d'habitation relevant d'une même autorité gestionnaire, dont les habitants partagent à titre habituel un mode de vie commun » : maisons de retraite, cités universitaires, communautés religieuses, casernes, prisons, etc.

① SUIVRE DES CHEMINS POUR TROUVER LE PLUS COURT

Le problème a été formulé ainsi : pour chaque département, on parcourt les barycentres des communes du département, en partant de l'un d'entre eux sélectionné au hasard, en adoptant le plus court chemin revenant au point de départ et ne passant qu'une fois par commune. Ce chemin fermé, qui ne dépend que de la commune de départ, est calculé par un algorithme solution au problème du voyageur de commerce¹³ (Applegate *et alii*, 2006). On peut répéter le calcul autant de fois qu'il y a de communes dans le département.

Puis, pour chaque chemin calculé, on détermine les unités primaires associées, en parcourant les communes le long du chemin et en saturant progressivement la contrainte sur le nombre de logements de l'unité primaire, qui doit être supérieur à 2 500. Lorsqu'une unité de plus de 2 500 logements est constituée, on ouvre la constitution d'une nouvelle unité en passant à la commune suivante sur le chemin (*figure 2*). À chaque chemin, correspond donc un découpage du département en unités primaires respectant la contrainte sur le nombre de logements.

On définit ensuite **l'extension géographique d'une unité primaire**, par la moyenne des distances le long de la route entre le centre de la commune principale¹⁴ et celui des autres communes de l'unité, pondérées par le nombre de logements des communes non principales. Au niveau d'un département, la moyenne de l'extension géographique des unités primaires qui le recouvrent, détermine **une valeur-objectif**, qui caractérise le découpage obtenu et le chemin qui a été utilisé.

L'extension géographique associée à un découpage s'entend aussi comme une distance moyenne d'accès aux logements des unités primaires ainsi constituées, depuis leur centre. La question intéresse directement les enquêteurs sur le terrain : l'extension géographique de la zone qu'ils devront parcourir pour leur collecte aura un impact direct sur le coût et la durée de leurs déplacements ; il convient donc de la minimiser.

Finalement, parmi l'ensemble des chemins calculés, on retient celui dont la valeur-objectif est la plus petite : ce chemin correspond à une extension géographique moyenne des unités primaires, la plus petite parmi les différents découpages construits pour le département considéré. En moyenne, du fait de l'application du schéma d'optimisation précédent, l'extension géographique des unités primaires métropolitaines a diminué de 25 % par rapport à l'échantillon-maître de 2010.

La méthode ne permet cependant pas de prendre en compte certaines spécificités locales liées à la présence d'une frontière nationale, à la saturation de certains réseaux ou la fermeture de certaines voies, lesquelles peuvent moduler l'appréciation de l'extension géographique. La carte des unités primaires a donc été amendée ponctuellement après examen par les directions régionales de l'Insee. *In fine*, les travaux ont conduit à la partition de la France métropolitaine en 5 128 unités primaires (*figure 3*).

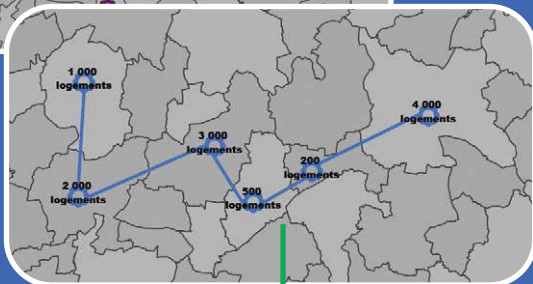
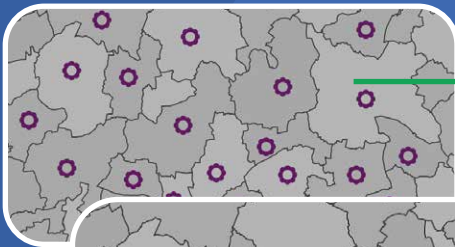
13. Problème d'optimisation qu'on pourrait résumer ainsi : étant donné une liste de villes, et des distances entre toutes les paires de villes, quel est le plus court chemin qui visite chaque ville une et une seule fois et qui se termine dans la ville de départ ?

14. Déterminée comme étant la commune de l'unité primaire comprenant le plus grand nombre de logements.

Figure 2. Comment constituer des unités primaires le long d'un chemin donné

1

Les points figurent les centres géographiques des communes.
Des distances le long de la route et centre à centre ont été calculées pour tous les couples de communes du département.

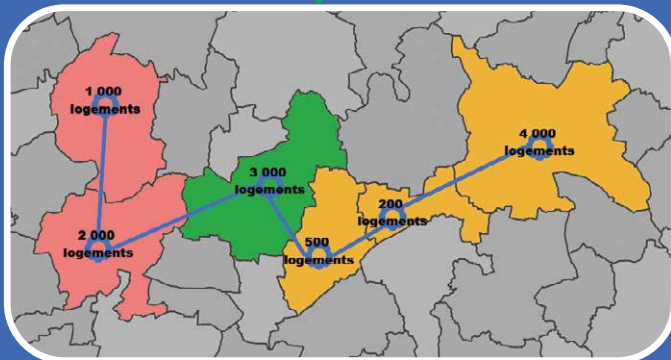


2

On part d'une première commune, sélectionnée au hasard : dans l'exemple, cette commune comprend 1 000 logements.

On cherche les chemins qui partent de cette commune et passent par toutes les autres communes du département, mais une fois seulement.

Parmi ces chemins, l'algorithme permet de calculer celui qui est le plus court.



3

Puis, le long de ce chemin, on constitue des regroupements – les unités primaires – en saturant successivement la contrainte de 2 500 logements par unité.

Dans l'exemple, le long de ce chemin, trois unités sont constituées : la première en rouge, la deuxième en vert et la troisième en jaune.

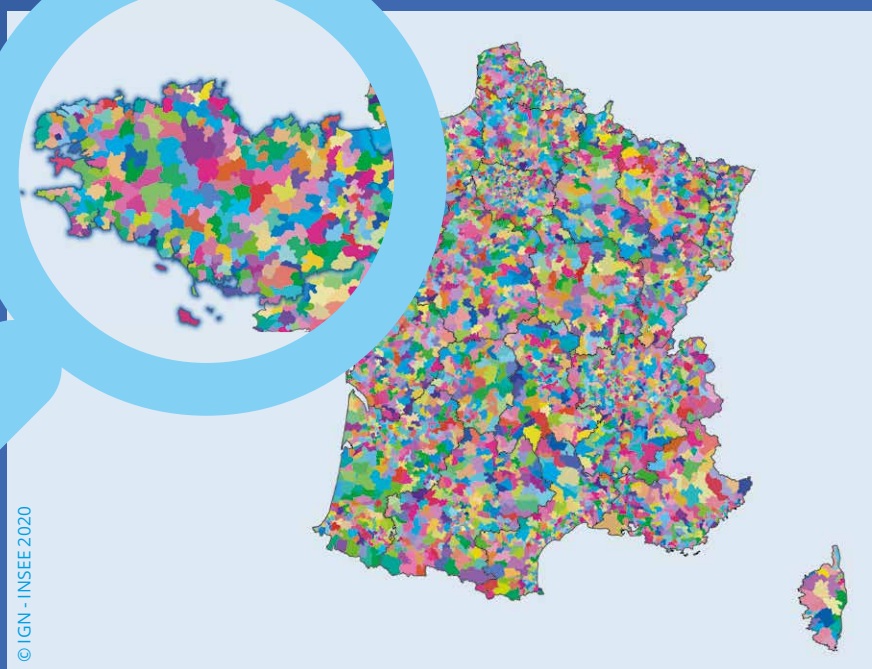
OPTIMISER LA PRÉCISION AVEC LE TIRAGE DE L'ÉCHANTILLON-MAÎTRE...

Un des enjeux principaux, pour l'Insee, est de maîtriser la précision de ses enquêtes. Comme on l'a vu, celles-ci sont tirées, en pratique, en deux étapes non simultanées. La première étape, dite de « tirage de premier degré » consiste à sélectionner les unités primaires de l'échantillon-maître. Cette première étape a lieu une fois tous les dix ans. Puis, pour chaque enquête, on tire, dans le cadre du « tirage de deuxième degré », parmi les ménages résidents de chaque unité primaire de cet échantillon-maître, les quelques dizaines qui seront effectivement interrogés au titre de cette enquête. La précision de l'enquête résulte de ces deux étapes de tirage : la sélection des unités primaires et celle des ménages parmi les résidents de ces unités primaires.

Les deux étapes de tirage sont indépendantes, donc la variance d'un estimateur, par exemple de moyenne ou de total fondé sur l'échantillon d'enquête recueilli, se décompose en deux termes distincts qui s'additionnent.

Plus précisément, on montre (Tillé, 2019) que la variance d'un estimateur de moyenne ou de total se décompose en deux termes : un terme de variance, dit de « premier degré » lié à la variance des totaux de la variable, observés au sein de chaque unité primaire ; à ce premier terme s'ajoute un second, dit de « deuxième degré », lié à la somme des variances

Figure 3. Avec le nouvel échantillon-maître, la métropole compte 5 128 unités primaires



intra-unités primaires. La variance de premier degré est indépendante de celle de second degré ; la variance de second degré est, elle aussi, essentiellement indépendante de celle de premier degré, hormis par la présence, dans son expression, d'un facteur d'échelle inversement proportionnel au taux de sondage de premier degré : si on réduit le nombre d'unités primaires sélectionnées, toutes choses égales par ailleurs, la variance de second degré augmente en proportion. Mais cette interaction est modeste et maîtrisable, de sorte qu'on peut donc découpler l'optimisation de variance. Ce qui s'avère très pratique puisque les tirages de premier et second degré ne sont pas réalisés simultanément.

L'optimisation du tirage de premier degré se résume donc à construire un plan de sondage optimal, c'est-à-dire qui minimise la variance de premier degré, associée aux quelque 500 unités primaires à sélectionner parmi les 5128 du territoire métropolitain.

📍 ...EN UTILISANT LES STATISTIQUES CONNUES SUR LES UNITÉS-PRIMAIRES...

Pour construire ce plan, on dispose des variables de la base de sondage (Fidéli), mais aussi de toutes les variables issues des productions de statistiques sociales de l'Insee au niveau de chaque commune, sommables sur des regroupements de communes. C'est d'ailleurs un des

« D'autres solutions auraient pu être envisagées, mais en complexifiant la production d'un ensemble de statistiques s'y rapportant. »

intérêts majeurs de constituer les unités primaires sous la forme de regroupements de communes, tandis que d'autres solutions auraient pu être envisagées, mais en complexifiant la production d'un ensemble de statistiques s'y rapportant.

De la sorte et en pratique, on dispose d'un grand nombre d'informations connues *a priori* sur les unités primaires : effectifs par sexe et âge issus du recensement, catégorie socio-professionnelle, diplôme, scolarisation, activité, composition des ménages, salaires, revenus, bénéfices industriels, agricoles et commerciaux, nombres d'assujettis à l'impôt sur la fortune, caractéristiques du logement, régions d'appartenance, etc.

Or la connaissance de ces statistiques est déterminante pour la variance de premier degré. En effet, le premier degré de sondage est, par nature, non affecté par d'éventuelles non-réponses, puisque toute unité primaire de l'échantillon-maître sera couverte par la collecte en toute circonstance. De ce fait, les plans de sondages les plus intéressants sont les plans équilibrés sur certaines variables : dans ces plans, les échantillons vérifient systématiquement l'égalité des estimateurs de moyennes avec leurs vraies valeurs sur l'ensemble de la population, ces vraies valeurs étant connues par ailleurs, par exemple issues du recensement.

📍 ... POUR PRIVILÉGIER L'ÉQUILIBRAGE (OU S'EN APPROCHER)...

Des algorithmes d'échantillonnage permettent de ne sélectionner que des plans équilibrés, ou éventuellement approchés sur le plan du strict équilibrage si celui-ci n'est pas rigoureusement possible. C'est le cas de la méthode du Cube (Deville et Tillé, 2004). L'équilibrage est intéressant dès qu'il existe une corrélation statistique entre une variable d'intérêt de l'enquête et les variables d'équilibrage : si celles-ci sont explicatives des variables d'intérêt, au sens d'une régression linéaire, alors on montre que la variance des estimateurs issus du sondage est de l'ordre du niveau de variabilité des résidus de la régression, et non de celle de la variable d'intérêt. Il convient donc de choisir des variables d'équilibrage

qui expliquent au mieux les futures variables d'intérêt des enquêtes. Néanmoins, pour l'élaboration du plan, il convient de limiter le jeu de variables d'équilibrage qui, de facto, limitent le champ de l'aléatoire dans la sélection des unités primaires.

Or le nombre d'unités primaires est relativement faible (5 128), de même que le nombre d'unités sélectionnées (environ 500). Par conséquent, le caractère surdéterminant des contraintes d'équilibrage, essentiellement lié à leur nombre, est assez vite atteint. Par exemple, introduire trop de variables d'équilibrage conduirait à ne plus pouvoir respecter le principe d'une taille d'échantillon-maître fixe en termes de nombre d'unités primaires.

Une analyse en composantes principales (ACP) a permis de réduire les variables d'équilibrage aux 15 premiers axes, représentant 99 % de l'inertie de l'ensemble des variables disponibles. Cette représentation de la quasi-totalité de l'information sur la population par l'échantillon obtenu se fait cependant au détriment de l'équilibrage parfait de celui-ci sur un nombre restreint de variables. C'est pourquoi, on complète ces axes de quelques variables clés, comme la population, ou celles permettant d'assurer un nombre fixe d'unités de l'échantillon-maître dans chaque région. Le plan est finalement équilibré au niveau régional, avec une phase d'atterrissage au niveau national permettant de libérer les contraintes qu'on ne peut pas respecter au niveau régional (Chauvet, 2009).

La sélection du plan, donc des variables retenues pour l'équilibrage, se fonde sur l'optimisation empirique du coefficient de variation des variables d'intérêt. Techniquement, celui-ci peut-être approché en simulant, par une méthode de Monte-Carlo¹⁵, un grand nombre de fois le plan de sondage retenu, la moyenne et la variance empiriques des estimations produites lors de chaque simulation convergeant vers la valeur vraie des estimateurs. Ces estimateurs sont calculés aux niveaux national et régional, du fait des contraintes du règlement IESS à ces deux niveaux géographiques. Les allocations retenues, ainsi que le plan de sondage de premier degré, visent donc à respecter les contraintes européennes.

La méthode d'équilibrage retenue *in fine* est une variante de l'algorithme du Cube, dite « échantillonnage doublement équilibré », adaptée aux situations dans lesquelles les variables d'intérêt sont spatialement corrélées, ce qui est le cas de la plupart des variables socio-économiques. En effet, dans ce type de situation, toutes choses égales par ailleurs, deux unités proches vont apporter plutôt moins d'information statistique que deux unités plus éloignées (Loonis et De Bellefon, 2018). On aura donc intérêt à sélectionner les unités modulo une légère force de répulsion entre unités proches, engendrant un échantillon mieux réparti spatialement (**figure 4**). C'est l'objet de l'algorithme d'équilibrage spatial, mis au point par (Grafström et Tillé, 2013), et appliqué dans le cas présent. Des simulations montrent que, pour des variables spatialement corrélées ne contribuant pas à l'équilibrage, comme le nombre de chômeurs ou de cadres, les gains de l'équilibrage spatial peuvent représenter jusqu'à 20 %, en termes d'erreur quadratique moyenne des estimations, par rapport à l'équilibrage simple (Favre-Martinoz et Merly-Alpa, 2016).

L'échantillon-maître fixe les zones géographiques sur lesquelles sont collectées la plupart des enquêtes auprès des ménages de l'Insee. Toutefois, pour des raisons liées principalement au protocole d'enquête, l'enquête Emploi est collectée sur des zones distinctes de celles

15. Le terme désigne une famille de méthodes algorithmiques visant à calculer une valeur numérique approchée en utilisant des techniques probabilistes.

de l'échantillon-maître. Il paraissait judicieux de rapprocher ces zones de collecte pour faciliter l'activité des enquêteurs. C'est ce qui a été réalisé pour la première fois lors du présent exercice.

LE RENOUVELLEMENT DE L'ÉCHANTILLON DE L'ENQUÊTE EMPLOI

Parmi les enquêtes ménages, l'enquête Emploi se distingue par un temps de collecte très court (trois semaines dans le trimestre) imposant à l'enquêteur sur le terrain un parcours rapide *in situ*. C'est ce qui justifie, entre autres, de tirer son échantillon par grappes de logements, géographiquement proches.

« En 2009, l'échantillon de l'enquête Emploi a été tiré de manière indépendante de l'échantillon-maître. »

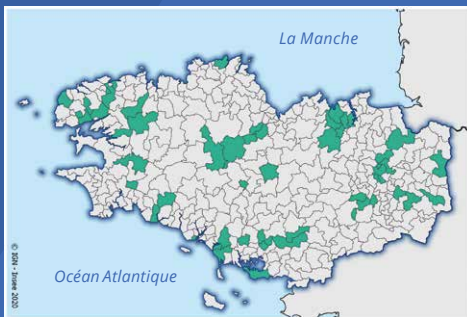
En 2009, l'échantillon de l'enquête Emploi a été tiré de manière indépendante de l'échantillon-maître. Le parti retenu lors de ce renouvellement (Loonis, 2009) a été de s'appuyer sur les sources fiscales pour constituer des **secteurs de collecte**, chacun composé de six **grappes contenant chacune environ 20 logements** quasiment contigus. La France métropolitaine est découpée en 232 000

secteurs Emploi, composés de 1,4 millions de grappes, regroupant 28,3 millions de résidences principales. Un échantillon de secteur a été tiré. Tous les logements d'une grappe échantillonnée sont interrogés, pendant six trimestres consécutifs (pour plus de détail, voir **encadré 1**), puis remplacés par les logements d'une autre grappe du même secteur. L'échantillon de l'enquête Emploi a ainsi une durée de vie de neuf ans.

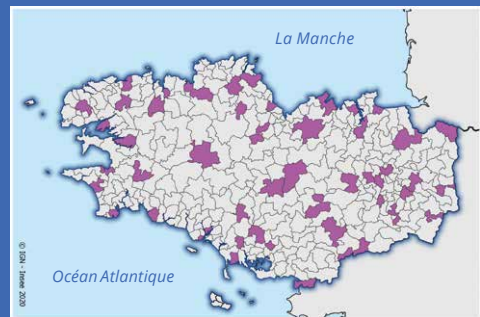
En 2019, la volonté d'optimiser l'activité de collecte des enquêteurs de l'Insee a conduit à envisager une coordination des échantillons de l'EEC et des autres enquêtes auprès des ménages.

Figure 4. Effet de l'équilibrage spatial sur la sélection des unités primaires

Tirage équilibré



Tirage spatialement équilibré



Encadré 1. La constitution des grappes et des secteurs de l'enquête Emploi en 2019

L'enquête Emploi en continu (EEC) interroge les ménages échantillonnés pendant six trimestres consécutifs : on peut ainsi apprécier l'évolution de la situation d'emploi des personnes, tout en limitant l'attrition que provoquerait un nombre d'interrogations plus important. Chaque interrogation se déroule sur une période relativement courte (3 semaines), ce qui nécessite de réduire au maximum l'étendue de la zone de collecte pour un enquêteur donné. D'où l'intérêt de constituer des **grappes d'une vingtaine de logements** géographiquement très proches, et d'interroger tous les logements d'une grappe en même temps. À la fin de la période de six trimestres, une grappe « entrante » s'y substitue : partant de l'idée que la situation face à l'emploi dans deux grappes voisines est proche, dès la constitution de l'échantillon, on regroupe les grappes dans des **secteurs**, au sein desquels s'opère le renouvellement de l'échantillon.

Pour constituer les grappes, on mobilise une logique de proximité entre les logements, analogue à celle retenue entre les communes lors de la constitution de l'échantillon-maître. Sauf qu'ici on s'appuie sur une localisation très précise, disponible indirectement dans les sources fiscales : les coordonnées géographiques des logements, à partir des parcelles cadastrales.

Dans les immeubles collectifs, on facilite le travail des enquêteurs, en constituant des grappes comprenant soit toutes les résidences principales d'un même étage, soit aucune. Un « chemin » est ensuite créé entre chaque étage d'immeuble ou maison individuelle, au sein d'un même Iris*. Pour garantir le principe de faible étendue géographique, l'étendue moyenne des grappes est pénalisée dans une fonction de coût. Une fois le chemin parcouru et les grappes associées constituées** dans un Iris, le processus d'association est rejoué, en sélectionnant d'autres points de départ du chemin : pour chaque Iris considéré, on retient *in fine* le chemin qui aboutit à la plus faible étendue des grappes.

Une fois les grappes constituées, celles-ci sont rassemblées, toujours au sein des Iris, en **secteurs** de six à sept grappes, sur un principe similaire à celui déjà appliqué pour les communes dans les unités primaires ou pour les logements dans les grappes***. On détermine donc différents chemins optimaux pour différents points de départ parmi les barycentres des grappes et **retient, parmi les chemins calculés, celui qui confère la plus petite étendue géographique aux secteurs.**

L'échantillon devant être utilisé pendant 9 ans, il est rafraîchi chaque année en associant les nouveaux logements aux grappes ainsi constituées et en mettant à jour le statut des logements (résidences principales, secondaires ou vacantes). Là aussi, la source fiscale est la seule à pouvoir offrir ce service.

Ce schéma aboutit à constituer 1,4 millions de grappes, regroupant 28,3 millions de résidences principales en 232 000 secteurs. L'étendue géographique étant pénalisée dans l'algorithme de sélection, elle est nettement diminuée par rapport à l'échantillon de 2009. Par exemple, la longueur médiane passe de 1 km pour l'ancien échantillon à 0,6 km pour le nouveau.

Note : le détail des opérations sera l'objet d'un document de travail de l'Insee, à paraître en 2020.

* Îlots regroupés pour l'information statistique, maille élémentaire de diffusion des résultats du recensement correspondant à des zones de 2 000 logements en moyenne. Ce découpage concerne les communes de 5 000 habitants ou plus, en dessous de ce seuil, la commune est assimilée à un Iris.

** Une phase de consolidation finale permet d'affecter les résidences principales résiduelles.

*** Afin de limiter les secteurs comprenant 7 grappes, la contrainte d'appartenance des secteurs à l'Iris est assouplie, ce qui conduit, dans certains cas, à constituer des secteurs à cheval sur deux Iris.

Une première idée pouvait être de sélectionner les secteurs emploi au sein des unités primaires de l'échantillon-maître. Mais d'une part ceci aurait conduit à l'épuisement rapide de certaines unités primaires trop petites ; d'autre part, le niveau de précision sur les estimateurs issus de l'enquête Emploi aurait été insuffisant au regard de ce qui est attendu, notamment dans le règlement européen IESS. Cette solution a donc été vite abandonnée au profit d'une solution consistant à définir des voisinages des unités primaires sélectionnées au sein desquels les secteurs de l'EEC ont vocation à être échantillonnés. Ces zones voisines constituent des **unités de coordination**.

Une analyse détaillée des différentes options a été menée, notamment s'agissant de la taille et du nombre d'unités primaires constituant les unités de coordination, et s'appuyant sur des considérations liées à la précision des deux types d'enquêtes. Elle a conduit à retenir, par rapport à la taille minimale des unités primaires qui est de 2 500 logements, une taille plancher des unités de coordination à 10 000 logements-résidences principales.

Sur le plan du mécanisme de sélection, les unités de coordination apparaissent comme indirectement sélectionnées, une fois les unités primaires de l'échantillon-maître tirées. Plusieurs unités primaires peuvent être à l'origine d'une même unité de coordination. On est donc dans le contexte du sondage indirect (Deville et Lavallée, 2006) qui nécessite de procéder à un partage des poids.

“ Une fois les unités de coordination définies, celles-ci contiennent les secteurs emploi qui font l'objet, comme les unités primaires de l'échantillon-maître, d'un échantillonnage spatialement équilibré. ”

Une fois les unités de coordination définies, celles-ci contiennent les secteurs emploi qui font l'objet, comme les unités primaires de l'échantillon-maître, d'un échantillonnage spatialement équilibré. Les secteurs sont sélectionnés indépendamment d'une région à l'autre (stratification

régionale) ce qui permet d'ajuster convenablement les allocations régionales en fonction des contraintes du règlement européen IESS.

Comme pour l'échantillon-maître, des simulations de Monte-Carlo permettent d'éclairer le paramétrage du plan de sondage, par analyse des précisions obtenues sur des *proxies* des variables d'intérêt de l'enquête Emploi, à partir des variables de la base de sondage. C'est le cas par exemple de la variable d'activité dont un *proxy*, fondé sur les données fiscales, est l'indicatrice de somme non nulle des revenus d'activité.

FINALISATION ET OPTIMISATION CONJOINTE DES DEUX ÉCHANTILLONS

Le schéma d'ensemble requiert une optimisation, puisque les deux échantillons sont fortement connectés du fait de la coordination instaurée : une variation du nombre d'unités primaires retenues a une incidence sur la précision des enquêtes ménages, comme sur celle de l'enquête Emploi. Le nombre des secteurs a lui aussi une incidence sur la précision de cette dernière. Ainsi, plusieurs leviers sont possibles pour, par exemple, jouer sur la précision de l'enquête Emploi : le nombre d'unités primaires sélectionnées, la taille des unités de coordination, le nombre de secteurs. Des simulations sont mobilisées pour évaluer l'impact des différents paramètres en termes de précision. Ces résultats sont comparés aux niveaux de précisions à atteindre au regard du règlement cadre IESS.

Tous les moyens mis en œuvre pour améliorer la précision permettent *in fine* d'améliorer de manière spectaculaire, d'un facteur 5 à 6, le niveau de précision du premier degré du nouvel échantillon-maître, par rapport à celui fondé sur les enquêtes annuelles de recensement.

En rejouant les tirages simultanés des deux échantillons, on peut procéder à des estimations de précision convergeant vers les vraies valeurs, connues, pour des variables de la base de sondage. Des dizaines de millions d'échantillons d'unités primaires et de secteurs emploi ont été tirés, dans de nombreuses configurations différentes, pour parvenir à un choix de paramétrage. Celui-ci portait sur la configuration des zones, en particulier sur les seuils plancher de nombre de logements par unités primaires et unités de coordination, sur les variables d'équilibrage, et sur les allocations, c'est-à-dire le nombre de zones sélectionnées dans les échantillons retenus.

Les derniers paramètres cible ont été les allocations régionales : elles ont permis d'affiner les *scenarii* présentés au comité de direction de l'Insee, ainsi qu'aux directions régionales qui gèrent les enquêteurs. Initialement proportionnelles à la taille de la région (en termes de nombre de résidences principales), elles ont été dans un second temps renforcées dans les petites régions, au détriment des plus grandes, afin d'éviter une concentration excessive des secteurs emploi au voisinage d'un nombre trop restreint d'unités primaires. Des simulations de tirage ont permis de vérifier que la précision attendue sur les variables phares de l'enquête Emploi apparaissait compatible avec les besoins de diffusion. L'exercice de validation finale a débouché sur un échantillon de 541 unités primaires, desquelles découlent la sélection de 524 unités de coordinations et finalement 2 944 secteurs pour l'enquête Emploi¹⁶.

Naturellement, la variance totale dépend aussi du second degré qui dépend lui-même des allocations (nombre de ménages tirés dans chaque unité primaire). La répartition de variance entre premier et second degré étant très diverse, il convient de ne pas trop réduire le nombre d'unités primaires au motif que la précision du premier degré serait très favorable. La prudence nécessaire a conduit à retenir un nombre d'unités primaires en légère baisse par rapport à l'échantillon antérieur, passant de 567 à 541.

Spontanément, à nombre de secteurs sélectionnés identiques, les gains attendus des évolutions de méthode sur la précision de l'EEC sont modestes, puisque l'échantillon antérieur avait déjà été construit sur les fichiers fiscaux (Loonis, 2009). Toutefois la possibilité, nouvelle en raison de la disponibilité de la source Fidéli, d'équilibrer le tirage des secteurs sur des *proxies* de variables de chômage, par exemple, a constitué une amélioration substantielle. À l'opposé, la coordination entre les deux échantillons génère un surcroît de variance sur l'enquête Emploi. Compte tenu des gains attendus sur le premier degré des enquêtes ménages, il a été possible d'ajuster les allocations d'unités primaires par région de façon à assurer un niveau de précision de l'EEC par région compatible avec les contraintes du règlement IESS, malgré la perte de précision spontanée liée à la coordination. En outre, il a été possible de réduire un peu le nombre de secteurs de l'EEC, car dans le bilan de précision de l'enquête, le gain marginal associé aux unités primaires est nettement plus favorable¹⁷ que celui associé aux secteurs, en raison de l'effet de grappe, caractéristique de ces derniers.

16. Les allocations régionales seront présentées dans un document de travail de l'Insee, à paraître en 2020.

17. Le ratio des élasticités des coefficients de variation de moyennes, estimées par l'EEC, au nombre d'unités primaires d'une part, et de secteurs d'autre part, vaut 4.

UN PROJET SERVANT UNE STRATÉGIE D'EXTENSION DES MODES DE COLLECTE

La rénovation des échantillons des enquêtes auprès des ménages et de l'enquête Emploi est un projet stratégique pour l'Insee. Elle répond à la volonté de pouvoir mener des enquêtes, sur la base d'une méthodologie solide et éprouvée, adaptée au mode de collecte en face-à-face.

Les modes alternatifs (internet, téléphone, papier) se développent, en raison de leur souplesse, de leur faible coût et parce qu'ils permettent d'atteindre mieux certaines catégories de ménages : ils ne requièrent pas, *stricto sensu*, d'échantillon-maître. Du reste, l'Insee renforce activement sa capacité à administrer les enquêtes en multimode (Cotton et Dubois, 2019 ; Koumarios et Sigaud, 2019). Mais, dans un pays étendu comme la France¹⁸, le principe d'un échantillon-maître reste le plus adapté à une collecte en face-à-face, que ce soit en mode unique ou comme composante d'une enquête multimodale.

Le projet a aussi été exigeant. Mené en mode agile, avec une interaction forte entre les concepteurs et les développeurs, il a mobilisé trois experts de la direction de la Méthodologie à plein temps pendant 18 mois. Le déploiement des enquêteurs sur les nouvelles zones a été étudié et déterminé par les directions régionales et la direction des statistiques démographiques et sociales, en concertation avec les organisations syndicales. Au-delà des aspects méthodologiques et organisationnels, une application spécifique a été développée : Nautille¹⁹ standardise et sécurise les tirages dans l'échantillon-maître, et marque les logements sélectionnés pour une enquête. Cette application est compatible avec les outils de gestion des métadonnées actives²⁰. Elle alimentera par ses échantillons les applications de gestion d'enquête, de calcul de charge des enquêteurs et d'administration des contacts avec les ménages, pour les dix ans qui viennent.

18. D'autres pays, comme l'Australie et le Canada, disposent d'un échantillon-maître. Et pour leurs enquêtes en face à-face, la plupart des pays procèdent par sélection initiale de zones géographiques, puis de ménages au sein de ces zones (OECD, 2019).

19. Nautille (Nouvelle application utilisée pour le tirage des individus et des logements des enquêtes) succède ainsi à Octopusse (Organisation coordonnée de tirages optimisés pour une utilisation statistique des échantillons) qui a servi aux échantillonnages des enquêtes de 2009 à 2019.

20. RMéS, le référentiel de métadonnées statistiques de l'Insee, a fait l'objet d'un article dans le numéro N2 du Courrier des statistiques (Bonnans, 2019).

BIBLIOGRAPHIE

APPLEGATE, David L., BIXBY, Robert E., CHVÁTAL, Vašek et COOK, William J., 2006. *The traveling salesman problem: a computational study*. Princeton University Press, Princeton Series in Applied Mathematics. ISBN : 978-0691129938.

BONNANS, Dominique, 2019. RMÉS, le référentiel de métadonnées statistiques de l'Insee. In : *Courrier des statistiques*. [en ligne]. 27 juin 2019. N°N2, pp. 46-57. [Consulté le 16 juin 2020]. Disponible à l'adresse : <https://www.insee.fr/fr/statistiques/fichier/4168396/courstat-2-6.pdf>.

CASES, Chantal, 2019. IESS : l'Europe harmonise ses statistiques sociales pour mieux éclairer les politiques. In : *Courrier des statistiques*. [en ligne]. 19 décembre 2019. N°N3, pp. 127-139. [Consulté le 16 juin 2020]. Disponible à l'adresse : <https://www.insee.fr/fr/statistiques/fichier/4254233/courstat-3-10.pdf>.

CHAUVET, Guillaume, 2009. Stratified balanced sampling. In : *Survey Methodology*. [en ligne]. Juin 2009. Statistique Canada, Catalogue n°12-001-X, vol. 35, n°1, pp. 115-119. [Consulté le 16 juin 2020]. Disponible à l'adresse : <https://www150.statcan.gc.ca/n1/pub/12-001-x/12-001-x2009001-eng.pdf>.

CHRISTINE, Marc et FAIVRE, Sébastien, 2009. OCTOPUSSE : un système d'Échantillon-Maître pour le tirage des échantillons dans la dernière Enquête Annuelle de Recensement. In : *X^e Journées de Méthodologie Statistique de l'Insee*. [en ligne]. 23-25 mars 2009. [Consulté le 16 juin 2020]. Disponible à l'adresse : http://www.jms-insee.fr/2009/S14_3_ACTE_FAIVRE_JMS2009.PDF.

COTTON, Franck et DUBOIS, Thomas, 2019. Pogues, un outil de conception de questionnaires. In : *Courrier des statistiques*. [en ligne]. 19 décembre 2019. N°N3, pp. 17-28. [Consulté le 16 juin 2020]. Disponible à l'adresse : <https://www.insee.fr/fr/statistiques/fichier/4254216/courstat-3-3.pdf>.

DEVILLE, Jean-Claude, 1991. Une théorie des enquêtes par quotas. In : *Techniques d'Enquête*. [en ligne]. 16 décembre 1991. Statistique Canada, vol. 17, pp. 177-195. [Consulté le 16 juin 2020]. Disponible à l'adresse : <https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/1991002/article/14504-fra.pdf?st=1bW9pHnv>.

DEVILLE, Jean-Claude et LAVALLÉE, Pierre, 2006. Indirect sampling : The foundations of the generalized weight share method. In : *Survey Methodology*. [en ligne]. Décembre 2006. Statistique Canada, catalogue n°12-001, vol. 32, n° 2, pp. 165-176. [Consulté le 16 juin 2020]. Disponible à l'adresse : <https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2006002/article/9551-eng.pdf?st=i92lk5a4>.

DEVILLE, Jean-Claude et TILLÉ, Yves, 2004. Efficient balanced sampling : The cube method. In : *Biometrika*. [en ligne]. 1^{er} décembre 2004. Vol. 91, n°4, pp. 893-912. [Consulté le 16 juin 2020]. Disponible à l'adresse : <https://academic.oup.com/biomet/article-pdf/91/4/893/1249543/914893.pdf>.

FAVRE-MARTINOZ, Cyril, 2015. *Étude préliminaire de l'impact de la prise en compte des groupes de rotation du RP pour l'échantillonnage des enquêtes ménages*. Insee, note interne n°776/DG75-L110/DE/CFM/SF.

FAVRE-MARTINOZ, Cyril et MERLY-ALPA, Thomas, 2016. Utilisation des méthodes d'échantillonnage spatialement équilibré pour le tirage des unités primaires des enquêtes ménages de l'Insee. In : *9^e Colloque Francophone sur les Sondages*. [en ligne]. 14 octobre 2016. Société française de Statistique (SFDS) et Université du Québec en Outaouais (UQO), Gatineau, Canada. [Consulté le 16 juin 2020]. Disponible à l'adresse : <http://sondages2016.sfds.asso.fr/wp-content/uploads/2016/10/Session21-MERLYALPA.pdf>.

GRAFSTRÖM, Anton et TILLÉ, Yves, 2013. Doubly balanced spatial sampling with spreading and restitution of auxiliary totals. In : *Environmetrics*. [en ligne]. Vol. 24, n°2, pp. 120-131. [Consulté le 16 juin 2020]. Disponible à l'adresse : <https://core.ac.uk/download/pdf/43674147.pdf>.

KOUMARIANOS, Heïdi et SIGAUD, Eric, 2019. Eno, un générateur d'instruments de collecte. In : *Courrier des statistiques*. [en ligne]. 19 décembre 2019. N°N3, pp. 17-28. [Consulté le 16 juin 2020]. Disponible à l'adresse : <https://www.insee.fr/fr/statistiques/fichier/4254218/courstat-3-4.pdf>.

LOONIS, Vincent, 2009. La construction du nouvel échantillon de l'Enquête Emploi en Continu à partir des fichiers de la Taxe d'Habitation. In : *Journées de Méthodologie Statistique*. [en ligne]. 23 mars 2009. Insee, X^e journées, Paris, Session 2 – Autour de l'Enquête Emploi. [Consulté le 16 juin 2020]. Disponible à l'adresse : http://www.jms-insee.fr/2009/S02_5_PRESENTATION_LOONIS_JMS2009.PDF.

LOONIS, Vincent et DE BELLEFON, Marie-Pierre, 2018. *Manuel d'analyse spatiale – Théorie et mise en œuvre pratique avec R*. [en ligne]. 29 octobre 2018. Insee. Eurostat. Collection Insee Méthodes, n°131, [Consulté le 16 juin 2020]. Disponible à l'adresse : <https://www.insee.fr/fr/statistiques/fichier/version-html/3635442/imet131.pdf>.

MERLY-ALPA, Thomas et SILLARD, Patrick, 2019. *The use of the French administrative dataset Fidéli as Sampling Frame for Household Surveys*. [en ligne]. 4-5 juin 2019, ESS workshop Administrative data for social statistics, Valence, Espagne. [Consulté le 16 juin 2020]. Disponible à l'adresse : https://ec.europa.eu/eurostat/cros/system/files/mr-patrick-sillard_french-admin-dataset-fideli_inseepdf.pdf.

OECD, 2019. Sampling and Weighting. In : *Technical Report of the Survey of Adult Skills (PIAAC) (3rd Edition)*. [en ligne]. Section 4, chapitres 14-16. [Consulté le 16 juin 2020]. Disponible à l'adresse : http://www.oecd.org/skills/piaac/publications/PIAAC_Technical_Report_3rd_Edition_2019_Section4_Chapters14-16.pdf.

TILLÉ, Yves, 2019. *Théorie des sondages : Échantillonnage et estimation en populations finies*. [en ligne]. Dunod, 2^e édition, ISBN : 978-2-10-079355-6. [Consulté le 16 juin 2020]. Disponible à l'adresse : <https://www.insee.fr/fr/statistiques/fichier/4254233/courstat-3-10.pdf>.