Journées de méthodologie statistique de l'Insee 2022

LA DISJONCTION DES ÉCHANTILLONS DES ENQUÊTES AUPRÈS DES MÉNAGES DE LA THÉORIE À LA PRATIQUE

Nicolas PALIOD (*)

(*) Insee, Direction de la méthodologie et de la coordination statistique et internationale

nicolas.paliod@insee.fr

Mots-clés : Disjonction d'échantillons, marquage, échantillonnage, charge de collecte, tirages à deux degrés

Domaine concerné : Théorie des sondages amont – Échantillonnage et Bases de sondage

Résumé

La gestion de la charge de collecte auprès des ménages fait partie des bonnes pratiques de la statistique européenne. Le choix réalisé par l'Insee est de retirer de la base de sondage les échantillons tirés pour des enquêtes passées. Ce choix est valable sur le plan méthodologique tant que les unités de la base de sondage ont toutes la même probabilité d'être sélectionnées à chaque nouveau tirage d'échantillon.

La disjonction des échantillons au sein d'un millésime donné de la base de sondage pose peu de difficultés sur le plan théorique. Elle implique le tirage d'un échantillon complémentaire à chaque fois qu'un échantillon tiré pour une enquête nécessite une restriction de champ ou une surreprésentation d'une partie de la population. En outre, le recours à un échantillon d'unités primaires pour les enquêtes en face à face conduit à une gestion de la disjonction des échantillons par unité primaire pour les tirages à 2 degrés. Le taux de logements marqués, c'est-à-dire retirés de la base de sondage suite à des tirages antérieurs, diffère par unité primaire.

Ce mécanisme de disjonction était suffisant jusqu'à une période récente puisque les tirages d'échantillons étaient historiquement réalisés dans les enquêtes annuelles de recensement. Le fait que les enquêtes annuelles de recensement soient disjointes assurait qu'un logement donné ne pouvait être échantillonné qu'une fois tous les 5 ans, à condition de garantir la disjonction des échantillons au sein d'un millésime donné de la base de sondage.

Depuis quelques années, l'Insee s'est orienté vers une utilisation accrue des sources fiscales pour le tirage d'échantillons d'enquêtes auprès des ménages, au détriment des enquêtes annuelles de recensement. Cette évolution s'accompagne de nouveaux enjeux autour de la disjonction des échantillons, en raison du caractère exhaustif des sources fiscales d'une part et de la possibilité nouvelle de réaliser des tirages d'individus dans celles-ci d'autre part.

D'abord, la possibilité de tirer des logements ou des individus pose la question de l'interaction entre les disjonctions d'échantillons de logements et d'échantillons d'individus.

Ensuite, l'exhaustivité de la base de sondage conduit à suivre le marquage des unités d'un millésime de la base de sondage au suivant. Cela nécessite une gestion particulière des logements neufs et des

naissances ou arrivées depuis l'étranger d'individus. Ces unités, absentes par nature du millésime précédent de la base de sondage, n'ont aucune chance d'avoir été tirées par le passé et doivent faire l'objet d'un tirage d'échantillon (non enquêté), afin que chaque unité de la base de sondage ait la même probabilité d'avoir été tirée par le passé. Cette propriété est en effet essentielle pour permettre de continuer à disjoindre les échantillons sans introduire de biais d'estimation.

Enfin, la gestion de la disjonction des échantillons d'individus présente une problématique supplémentaire par rapport aux échantillons de logements : les individus sont mobiles et déménagent. Donc, lors du passage à un nouveau millésime, les individus peuvent changer d'unités primaires. Conditionnellement à l'échantillon d'unités primaires sélectionné pour les tirages à deux degrés, les individus d'une même unité primaire n'ont plus la même probabilité d'avoir été tirés pour un échantillon dans le millésime précédent. Ainsi, il n'est pas possible de disjoindre les échantillons d'individus lors du passage à un nouveau millésime de la base de sondage, sans traitement statistique supplémentaire.

Cet article présente l'ensemble des solutions adoptées par l'Insee aux problématiques précédemment évoquées : hypothèses-clés pour que la disjonction d'échantillons ne s'accompagne pas de biais d'estimation, gestion des entrées de champ (logements neufs, naissances, arrivées de l'étranger) lors du passage d'un millésime de la base de sondage à un autre, gestion des déménagements d'individus d'une année à l'autre. Il s'achève sur une présentation de la politique actuelle de disjonction des échantillons à l'Insee et propose quelques éléments chiffrés pour évaluer son efficacité.

Abstract en anglais

In order to limit response burden in household surveys, Insee withdraws units sampled in former surveys from its sampling frame. This solution requires that all remaining units have the same probability to be drawn in next sample selections. Interaction between one-stage sampling and two-stage sampling impose conditions on probabilities by primary unit. Nevertheless, the building of dwellings and individuals' moves from a primary unit to another change the structure of individuals' probability to have been selected in past samples in a given primary unit. This article suggests a solution that enables to continue to withdraw dwellings or individuals sampled for former surveys from the sampling frame without distortion for next sample drawings.

Introduction

Comme tout institut national statistique, l'Insee cherche à limiter la réinterrogation des unités enquêtées par le passé. Cette exigence est autant nécessaire pour réguler la charge de collecte auprès des unités enquêtées que pour limiter la non-réponse qui pourrait découler d'une lassitude liée à des réinterrogations régulières des enquêtés.

Cet enjeu est très connu pour les enquêtes auprès des entreprises, étant donné que certaines strates de tirage présentent des taux de sondage élevés qui impliquent des réinterrogations régulières pour des entreprises à chiffre d'affaires ou à nombre d'employés élevés. La littérature est pléthorique pour la coordination négative d'échantillons visant à limiter le nombre d'unités tirées pour plusieurs enquêtes auprès des entreprises proches dans le temps (voir Hesse (1999) ou Ohlsson (1995) qui présentent un panorama de ces méthodes).

Elle est moins fournie pour les enquêtes auprès des ménages pour lesquelles les taux de sondage sont moindres, ce qui implique moins de recouvrement entre deux échantillons. Néanmoins, les problématiques induites par la coordination négative des échantillons est différente entre les enquêtes auprès des ménages et les enquêtes auprès des entreprises. En effet, les enquêtes auprès des ménages utilisent encore beaucoup le mode de collecte du face à face bien qu'ayant de plus en plus recours à d'autres modes de collecte. Cette articulation des enquêtes en face à face et

d'enquêtes n'utilisant pas ce mode de collecte induit des tirages à 2 degrés pour les premières et des tirages à 1 degré pour les dernières. L'utilisation de ces deux types de tirages, et en particulier celle des tirages à 2 degrés, conduit à des enjeux spécifiques quant à la coordination négative d'échantillons.

D'abord, nous présentons la solution retenue pour la coordination négative des échantillons des enquêtes auprès des ménages : la disjonction des échantillons. Ensuite, nous détaillons les hypothèses nécessaires quant à la mise en œuvre de la disjonction des échantillons. Nous introduisons alors les différents types d'échantillons tirés à l'Insee et montrons qu'ils nécessitent le tirage d'échantillons complémentaires non mis en collecte pour éviter de déformer la base utilisée ultérieurement pour de futurs tirages. Les problématiques induites par les changements de millésime de base de sondage sont ensuite abordées, en se focalisant sur les tirages d'individus. Enfin, nous expliquons la solution retenue par l'Insee pour disjoindre les échantillons d'individus en tenant compte du changement de millésime de base de sondage.

1. La disjonction des échantillons : une solution pour coordonner négativement les échantillons des enquêtes auprès des ménages tirés par l'Insee

Depuis le déploiement du dispositif Fidéli (Fichiers démographiques sur les logements et les individus) permettant d'exploiter au mieux les fichiers issus des sources fiscales (voir Crenner (2018)), l'Insee tire la grande majorité de ses échantillons pour les enquêtes auprès des ménages dans les sources fiscales. Les bases issues de Fidéli présentent en effet les bonnes propriétés d'une base de sondage, comme l'évoquent Merly-Alpa, Pendoli et Vincent (2018).

Avant l'utilisation des sources fiscales, seuls des échantillons de logements pouvaient être tirés par l'Insee. L'utilisation du dispositif Fidéli permet désormais de sélectionner aléatoirement des échantillons de logements ou des échantillons d'individus.

Afin de limiter la charge de collecte pour les ménages, l'Insee coordonne négativement les échantillons en procédant à une disjonction de la base de sondage avec les échantillons préalablement tirés. Concrètement, cela implique que, pour un tirage d'échantillon de logements, les logements sélectionnés au cours de tirages d'échantillons de logements par le passé sont retirés de la base de sondage en amont du nouveau tirage. De même, pour un tirage d'échantillon d'individus, les individus sélectionnés au cours de tirages d'échantillons d'individus par le passé sont retirés de la base de sondage en amont du tirage. Par construction, cette méthode coordonne négativement les échantillons et évite de tirer à nouveau une unité déjà sélectionnée par le passé.

Actuellement, les tirages d'échantillons de logements ne sont pas disjoints des tirages d'échantillons d'individus et réciproquement, à l'exception notable de l'échantillon de logements de l'enquête Emploi en Continu dont le volume élevé conduit à le disjoindre de la base de sondage pour le tirage d'échantillons d'individus. La disjonction des échantillons de logements avec les échantillons d'individus fait l'objet de complexités méthodologiques, plus particulièrement lorsqu'on souhaite éviter de tirer des logements d'individus sélectionnés lors d'un tirage d'individus, qui sortent du cadre de cet article mais sont introduits dans le dernier chapitre de Vincent et al. (2021).

2. Les hypothèses nécessaires à la disjonction des échantillons

Les parties 2 et 3 s'appliquent de manière indifférenciée à la disjonction des échantillons de logements et à la disjonction des échantillons d'individus. Le terme « unité » s'appliquera donc indifféremment aux individus ou aux logements dans ces deux parties.

La disjonction des échantillons, également appelée opération de marquage, permet de retirer de la base de sondage les unités sélectionnées pour des enquêtes précédentes afin de constituer une base

dite de tirage. Les unités qui sont dans la base de tirage ont donc une probabilité d'y figurer, fonction de leurs probabilités de sélection dans chacun des échantillons participant à la disjonction.

Plus précisément, soit *i* une unité et $p_{P,i}$ sa probabilité d'avoir été tirée au cours de l'ensemble des tirages passés. Alors, sa probabilité de figurer dans la base de tirage pour le prochain tirage est $1-p_{P,i}$. Lors du tirage d'un échantillon dans la base de tirage, cette unité aura, conditionnellement aux tirages passés, une probabilité $p_{F,i}$ de figurer dans l'échantillon. La probabilité de l'unité *i* de figurer dans cet échantillon est $\left(1-p_{P,i}\right)p_{F,i}$ en tenant compte du tirage de l'échantillon considéré et de tous les tirages concernés par la disjonction.

Une telle construction des pondérations serait extrêmement complexe, car le poids de tirage d'une unité dans une enquête dépendrait du tirage de tous les tirages d'échantillons passés.

La solution retenue est donc de ne pas utiliser ce poids de tirage mais de calculer un poids par post-stratification à partir des volumes de la base de sondage et non de la base de tirage. Par exemple, pour un tirage à 1 degré, dans une strate *s* contenant N_s unités dans la base de sondage et $N_s^{'}$ unités dans la base de tirage et dans laquelle n_s unités sont sélectionnées, on utilisera le poids post-stratifié $\frac{N_s}{n_s}$ plutôt que le poids d'Horwitz-Thompson $\frac{1}{1-p_{P,i}}\frac{N_s^{'}}{n_s}$.
Une condition suffisante pour que cette solution aboutisse à des estimations sans biais est que la probabilité $1-p_{P,i}$ soit homogène[1] au sein de chaque strate *s*.

Les enjeux de cette méthode se rapprochent de la technique présentée par McKenzie et Gross (2001), si ce n'est que nous n'avons pas recours à l'utilisation d'un numéro aléatoire pour la disjonction des échantillons.

Pour un tirage à 2 degrés, dans le croisement d'une unité primaire *up* et d'une strate *s* contenant $N_{s,up}$ unités dans la base de sondage et $N_{s,up}^{'}$ unités dans la base de tirage et dans lequel $n_{s,up}$ unités sont sélectionnées, on utilisera le poids post-stratifié $w_{up}\frac{N_{s,up}}{n_{s,up}}$ plutôt que le poids d'Horwitz-Thompson $\frac{1}{1-p_{P,i}}\frac{N_{s,up}^{'}}{n_{s,up}}$. w_{up} est le poids d'Horwitz-Thompson de l'unité primaire. On peut noter que, dans le cas d'un estimateur d'Horwitz-Thompson, w_{up} est inclus dans le poids $\frac{1}{1-p_{P,i}}$.
Une condition suffisante pour que cette solution aboutisse à des estimations sans biais pour un échantillon sélectionné par 2 degrés de tirage est que la probabilité $1-p_{P,i}$ soit homogène au sein de chaque strate *s* d'une unité primaire *up* donnée.

3. Les principales catégories de tirages d'échantillons à l'Insee

La majorité des tirages d'échantillon à l'Insee sont des tirages :

- stratifiés à 1 degré, dans l'ensemble de la base de sondage ;
- stratifiés à 2 degrés, pour les tirages métropolitains, les unités primaires utilisées étant les 541 zones de collecte de l'échantillon-maître Nautile sélectionnées parmi les 5 064 unités primaires partitionnant la France métropolitaine, comme présenté dans Vincent et al. (2021) ou Sillard et al. (2020).

[1] Dans l'ensemble de ce document, on entend par « homogénéité », le fait que l'ensemble des unités concernées aient une probabilité de figurer dans la base de tirage égale en espérance. Dans le cas d'un tirage à 1 degré, cette espérance concerne également le tirage d'unités primaires, tandis que dans le cas d'un tirage à 2 degrés, il s'agit d'une espérance conditionnelle à l'échantillon d'unités primaires utilisé.

3.1. Disjonction d'échantillons pour les tirages à 1 degré

D'un tirage à l'autre, le champ de la base de sondage et les strates de tirage changent puisque les enquêtes n'ont pas nécessairement les mêmes populations d'intérêt. Puisque les strates sont mouvantes entre deux tirages d'échantillons, pour que les mêmes unités présentes dans la base de tirage pour une strate donnée du nouveau tirage aient la même probabilité d'y figurer, il est nécessaire d'avoir tiré par le passé des échantillons qui ne surreprésentent aucune partie de la population. Il est donc nécessaire de coupler le tirage de l'échantillon enquêté avec le tirage d'un échantillon complémentaire afin que le cumul de ces deux échantillons ait la même structure que la base de tirage utilisée pour le tirage de l'enquête concernée.

Pour les tirages stratifiés à 1 degré, le tirage de l'échantillon complémentaire consiste en un tirage stratifié à 1 degré dans la base de tirage privée de l'échantillon enquêté. L'allocation n_s^{comp} utilisée pour ce tirage dans la strate s se calcule de la manière suivante, en notant *{1,…,h…,H}* les différentes strates de tirage :

$$n_s^{comp} = max_h \left(\frac{n_h}{N_h} \right) N_s - n_s$$

On montre aisément que le taux de sondage de l'échantillon complet (échantillon enquêté et échantillon complémentaire) est le même pour chaque strate et donc que toutes les unités de la base de tirage « initiale » ont la même probabilité d'être tiré dans cet échantillon. Ainsi, toutes les unités ont également la même probabilité de figurer dans la base de tirage « finale » qui sera celle utilisée pour le prochain tirage d'échantillon.

La figure 1 illustre cette procédure. La base de tirage disponible pour les futurs tirages est la partie de la base ayant un fond bleu.

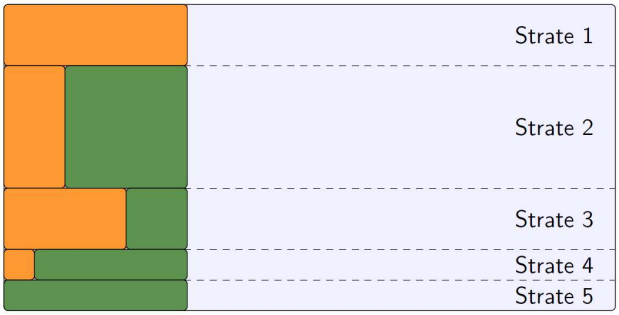

Figure 1 : Tirage d'un échantillon complémentaire (en vert) pour rééquilibrer la base de tirage suite à la sélection d'un échantillon enquêté (en orange). La strate 5 correspond aux unités hors-champ pour le tirage de l'échantillon enquêté.

3.2. Disjonction d'échantillons pour les tirages à 2 degrés

Les tirages à 2 degrés mobilisent des unités primaires sélectionnées aléatoirement avec une probabilité π_{up}. Les poids de tirage associés à ces unités primaires sont leur poids d'Horwitz-Thompson :

$$w_{up} = \frac{1}{\pi_{up}}$$

Les allocations pour les tirages de 2e degré sont des allocations dites autopondérées qui permettent d'obtenir le même poids pour l'ensemble des unités d'une strate donnée, quelle que soit leur unité primaire. Si on souhaite tirer n_s unités dans la strate *s*, l'allocation $n_{s,up}$ tirée dans le croisement de la strate *s* et de l'unité primaire *up* se calcule comme :

$$n_{s,up} = n_s \frac{w_{up} N_{s,up}}{\sum_{i \in EM} w_i N_{s,i}}$$

où $N_{s,up}$ est le nombre d'unités de l'intersection de la strate *s* avec l'unité primaire *up*.

On peut montrer assez simplement que le tirage d'un échantillon complémentaire ayant pour allocations $n_{s,up}^{comp}$ dans l'intersection de la strate *s* avec l'unité primaire *up* où :

$$n_{s,up}^{comp} = \left(max_h \left(\frac{n_h}{N_h} \right) N_s - n_s \right) \frac{w_{up} N_{s,up}}{\sum_{i \in EM} w_i N_{s,i}}$$

permet de tirer chaque unité dans l'échantillon complet (enquêté + complémentaire) avec la même probabilité au sein d'une unité primaire donnée quelle que soit sa strate $h \in \{1, \ldots, H\}$.

Concrètement, cela revient à rééquilibrer chaque unité primaire de l'échantillon de 1er degré comme une petite de base de sondage, ainsi que cela a été présenté en partie 3.1 pour les tirages à 1 degré. Ainsi, dans chaque unité primaire, toutes les unités ont la même probabilité de figurer dans la base de tirage pour le prochain tirage d'échantillon à 2 degrés.

Cela permet d'obtenir des estimations sans biais à partir des poids post-stratifiés par strate *s* croisée à l'unité primaire *up* pour les unités *i* tirées dans cette intersection :

$$w_i = w_{up} \frac{N_{s,up}}{n_{s,up}}$$

3.3. Gestion infra-annuelle d'une base de tirage

Au cours d'une année, les tirages à 1 degré alternent avec les tirages à 2 degrés.

La satisfaction de l'hypothèse clé de la disjonction des échantillons, à savoir l'homogénéité de la probabilité des unités d'avoir été sélectionnées au cours d'un tirage d'échantillon passé développée en partie 2, n'est pas évidente. On donne ci-dessous l'intuition à la base des démonstrations qui garantissent la satisfaction de cette propriété.

Pour les tirages à 2 degrés, on a vu dans la partie 2 que cette propriété d'homogénéité de la probabilité d'une unité de figurer dans la base de tirage devait être appréhendée au niveau de l'unité primaire. Si on réalise le tirage d'échantillons complémentaires lors des tirages à 2 degrés comme présenté en partie 3.2, cette propriété est satisfaite. La réalisation d'un tirage à 1 degré dans la base de tirage intégrant l'ensemble des unités primaires ne mettra pas à mal cette propriété puisque le tirage d'un échantillon complémentaire conduira à tirer les unités restantes dans la base de tirage avec la même probabilité quelle que soit l'unité primaire et donc, *a fortiori*, au sein d'une unité primaire. Ainsi, lorsqu'on utilise une base de tirage pour un tirage à 2 degrés, l'alternance de tirages à 2 degrés et de tirages à 1 degré qui a précédé ce tirage permettent bien de disposer pour chaque unité primaire d'une base de tirage dans laquelle les unités ont la même probabilité d'y figurer.

Pour les tirages à 1 degré, la satisfaction de cette propriété semble moins évidente. En effet, les allocations autopondérées présentées en partie 3.2 conduisent à tirer les échantillons à 2 degrés uniquement dans les unités primaires sélectionnées. En outre, parmi cet échantillon d'unités primaires, elles conduisent à utiliser un taux de sondage plus élevé dans les unités primaires ayant une faible probabilité d'inclusion. Ainsi, conditionnellement à l'échantillon d'unités primaires sélectionné pour les tirages à 2 degrés, lorsqu'on a déjà réalisé des tirages à 2 degrés, la base de tirage restant pour les tirages à 1 degré ne garantit plus que chaque unité ait la même probabilité d'y figurer. En particulier, la probabilité de rester dans la base de tirage des enquêtes à 1 degré est plus faible pour une unité appartenant à une unité primaire de l'échantillon de 1er degré. Cependant, si on raisonne en espérance sur l'ensemble des échantillons d'unités primaires qui auraient pu être tirés,

on peut montrer que les unités ont bien la même probabilité d'être dans la base de tirage utilisée pour les tirages à 1 degré, y compris après avoir effectué des tirages à 2 degrés. Cette démonstration nécessite cependant que les tirages à 2 degrés soient autopondérés.

Des travaux sont en cours à l'Insee pour démontrer rigoureusement que l'alternance de tirages à 1 degré et à 2 degrés selon la méthodologie présentée dans la partie 3.1 et dans la partie 3.2 permettent d'obtenir une base de tirage pour les tirages à 1 degré dans laquelle chaque unité a la même probabilité de figurer et une base de tirage pour les tirages à 2 degrés dans laquelle, pour une unité primaire donnée, chaque unité a la même probabilité de figurer[2].

4. Les problématiques de la disjonction des échantillons d'individus lors d'un changement de millésime de la base de sondage

Lors du changement de millésime de base de sondage, la gestion de la base de tirage présentée lors de la partie 3 est mise à mal par l'apparition de nouvelles unités dans le nouveau millésime et par la disparition d'unités de l'ancien millésime. Par exemple, pour les bases de logements, des logements sont construits tandis que d'autres sont détruits.

Cette problématique, bien connue des enquêtes entreprises, peut être traitée par de nombreuses façons dépendant de la méthode retenue pour la coordination négative d'échantillons comme celle présentée dans McKenzie et Gross (2001).

Dans notre cas, la disparition d'unités ne pose aucun problème puisque cette étape est déterministe. Donc les unités restant dans la base de tirage conservent la même probabilité d'y figurer qu'avant le passage au nouveau millésime.

Par contre, l'apparition de nouvelles unités est plus complexe à gérer. Comme illustré par la figure 2 pour une unité primaire donnée, seules les unités présentes à la fois dans le millésime N et dans le millésime N-1 peuvent être marquées dans le nouveau millésime. Les unités qui apparaissent dans le nouveau millésime ont une probabilité de 1 de figurer dans la base de tirage, ce qui met à mal les hypothèses garantissant le bon fonctionnement de la disjonction des échantillons.

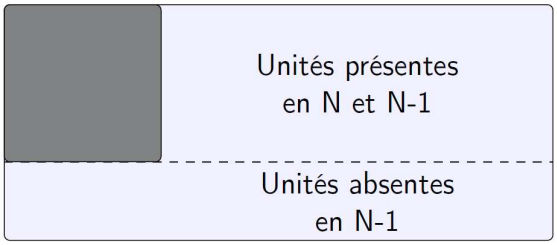

Figure 2 : Passage au nouveau millésime de la base de sondage pour une unité primaire donnée. Les unités en gris sont celles ayant été marquées par le passé suite au tirage des échantillons et disjointes des tirages. La partie en fond bleu désigne la base de tirage avant toute opération de rééquilibrage liée au passage au nouveau millésime.

[2] Le fait que l'échantillon d'unités primaires majoritairement utilisé à l'Insee surreprésente certaines régions, tout en ayant la contrainte d'enquêter à peu près le même nombre d'unités par unité primaire, conduit à ne pas utiliser la formule d'allocations autopondérées présentée dans la partie 3.2. Les allocations utilisées pour les tirages de second degré sont des allocations autopondérées régionalement. Elles garantissent que les unités tirées dans une région et dans une strate données aient le même poids de tirage. Mais cette homogénéité n'est plus garantie nationalement. On peut montrer que, en ayant une gestion régionale de la base pour les tirages à 1 degré, c'est-à-dire en calculant les allocations régionalement pour chaque strate et en post-stratifiant les pondérations par région croisée à la strate de tirage, tout ce qui précède reste valable au sein d'une région donnée. Cela permet de rééquilibrer la base de tirage régionalement et non plus nationalement. Les manipulations de formules sont alors plus complexes, mais cela présente l'avantage majeur de ne pas épuiser la base de sondage dans les DROM lorsqu'une enquête qui ne concerne que la France métropolitaine fait l'objet d'un tirage d'échantillon et réciproquement. C'est d'autant plus pratique que les échantillons tirés en France métropolitaine et dans les DROM sont parfois sélectionnés avec des plans de sondage très différents.

Afin de disposer d'une base de tirage présentant les propriétés nécessaires pour réaliser des estimations sans biais, une solution simple est de tirer dans chaque unité primaire un échantillon d'unités présentes dans le millésime N et absentes du millésime N-1 au même taux de sondage que le taux d'unités marquées dans la strate des unités présentes en N et en N-1 dans cette unité primaire. Ce rééquilibrage de la base de tirage est illustré par la figure 3.

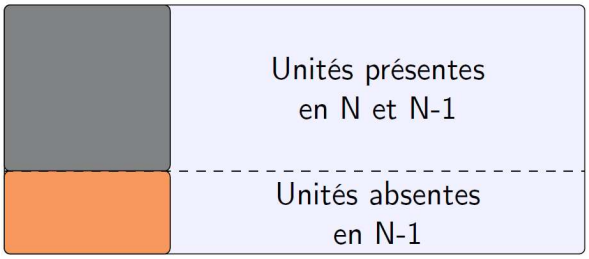

Figure 3 : Rééquilibrage d'une unité primaire donnée en tirant un échantillon d'unités apparues dans le nouveau millésime (en orange).

Ainsi, après rééquilibrage, on a l'intuition que la base de tirage du nouveau millésime a les mêmes propriétés que celle de l'ancien millésime puisque le taux d'unités marquées dans chaque unité primaire dans le nouveau millésime est identique au taux d'unités marquées dans l'unité primaire correspondante dans l'ancien millésime. La démonstration plus rigoureuse n'est pas présentée ici.

Cette gestion est efficace pour rééquilibrer la base de tirage de logements après apparition des logements neufs. Elle est également tout à fait valable pour traiter des situations telles que des retours d'individus qui vivaient à l'étranger ou des naissances dans la base de tirage d'individus.

Néanmoins, une problématique supplémentaire se pose pour les tirages d'individus. Contrairement aux logements, les individus sont mobiles et peuvent changer d'unité primaire. Or les tirages à 2 degrés conduisent à tirer des individus avec un taux différencié selon les unités primaires. Conditionnellement à une unité primaire donnée, les individus qui s'installent dans cette unité primaire n'ont pas la même probabilité d'être marqués que les individus de cette unité primaire qui n'ont pas déménagé. Or, tout notre raisonnement dans la partie 3.3 pour justifier qu'on puisse utiliser des pondérations post-stratifiées malgré l'alternance de tirages à 1 degré et de tirages à 2 degrés est basé sur 2 éléments. D'une part, l'ensemble des individus d'une unité primaire doivent avoir la même probabilité d'avoir été tirés par le passé. D'autre part, le nombre d'individus marqués dans une unité primaire suite à des tirages à 2 degrés doit provenir d'allocations autopondérées. Les déménagements ne permettent de garantir aucune de ces 2 conditions pour la base du tirage du nouveau millésime.

La figure 4 illustre cette situation pour une unité primaire UP1 dans laquelle s'installent des individus provenant de l'UP2, dont la proportion d'individus tirés pour des échantillons antérieurs est plus importante que pour l'UP1.

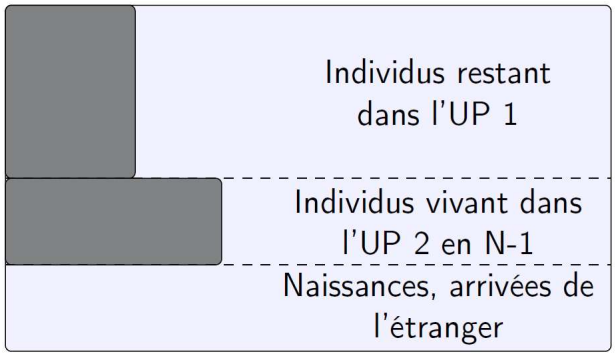

Figure 4 : Passage au nouveau millésime de la base de sondage d'individus pour une unité primaire UP 1. Les individus en gris sont ceux ayant été marqués par le passé suite au tirage des échantillons et disjoints des tirages. La partie en fond bleu désigne la base de tirage avant toute opération de rééquilibrage liée au passage au nouveau millésime.

5. Une disjonction des échantillons limitée à certains individus lors d'un changement de millésime de la base de sondage

Pour rééquilibrer la base de tirage du nouveau millésime en tenant compte des déménagements et en faisant en sorte que la structure de la base de tirage du nouveau millésime ait les mêmes propriétés que la base de tirage de l'ancien millésime, plusieurs options peuvent être envisagées.

La propriété que nous souhaitons atteindre est que l'ensemble des individus d'une unité primaire donnée soient marqués au même taux que ceux qui étaient déjà présents dans cette unité primaire l'année précédente. On peut ainsi envisager de tirer un échantillon complémentaire pour les individus venant d'une unité primaire avec un taux de marquage plus faible et démarquer aléatoirement des individus venant d'une unité primaire avec un taux de sondage plus élevé que celui de l'unité primaire dans laquelle ils s'installent. Par exemple, dans la figure 4, cela reviendrait à démarquer (i.e. rendre disponible pour un tirage ultérieur) aléatoirement des individus de l'UP 2 s'étant installés dans l'UP 1.

Néanmoins, cette solution est coûteuse car il y a plus de 5 000 unités primaires sur le territoire. Donc une telle gestion des déménagements conduirait à tirer un échantillon de marquage ou de démarquage dans chaque croisement d'unité primaire d'arrivée et d'unité primaire de départ. En ordre de grandeur, cela conduirait à tirer un échantillon dans 25 000 000 de strates pour environ 65 000 000 d'individus. En pratique, toutes les combinaisons d'unités primaires ne font pas l'objet de déménagements de l'une vers l'autre, et on pourrait se restreindre à un ordre de grandeur d'environ 250 000 croisements car toutes les unités primaires hors de l'échantillon-maître utilisé pour la grande majorité des tirages font l'objet d'un taux de marquage très proche. Il est cependant difficile d'automatiser une telle opération, et la gestion des arrondis d'allocations liée à l'utilisation d'un tel ordre de grandeur de strates rend incertaine sa réussite.

Une autre solution a été retenue. Il s'agit de ne continuer à marquer que des individus qui restent dans la même unité primaire. En pratique, tout se passe comme si les individus qui s'installent dans une unité primaire étaient traités comme des individus revenant de l'étranger ou des naissances. Concrètement, cela signifie que si un individu a été tiré pour un échantillon précédent, dès lors qu'il change d'unité primaire, on démarque cet individu. Si on reprend l'exemple de la figure 4, cela signifie qu'on démarque tous les individus tirés qui se sont installés dans l'UP 1 en provenance de l'UP 2. Cela est illustré par la figure 5.

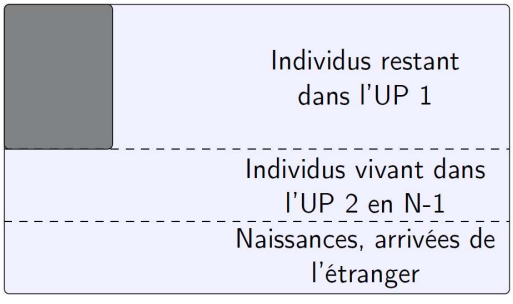

Figure 5 : Passage au nouveau millésime de la base de sondage d'individus pour une unité primaire UP 1 en démarquant les individus en provenance d'autres unités primaires et qui étaient marqués dans le millésime précédent.

Il ne reste alors plus qu'à rééquilibrer la base de tirage d'individus en tirant, dans chaque unité primaire, des individus ayant changé d'unité primaire ou n'étant pas présents dans le millésime précédent. Le calcul des allocations s'effectue en tirant ces individus au même taux de sondage que le taux auquel sont marqués les individus présents dans cette unité primaire à la fois dans l'ancien et dans le nouveau millésime de la base de sondage. Cette solution est illustrée par la figure 6.

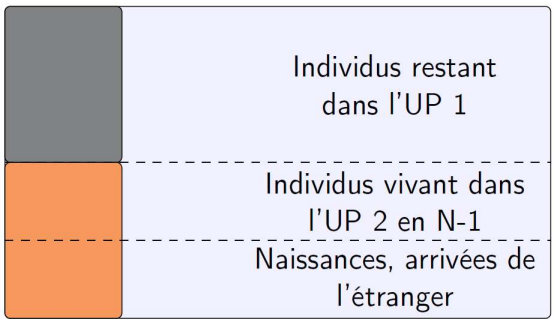

Figure 6 : Rééquilibrage de l'unité primaire UP 1 en démarquant les individus en provenance d'autres unités primaires et qui étaient marqués dans le millésime précédent puis en tirant un échantillon d'individus (en orange) qui sont arrivés dans l'unité primaire quelle qu'en soit la raison (déménagement, naissance, retour de l'étranger).

Dans le millésime 2020 de la base de sondage issue de Fidéli, cette solution permet de conserver le marquage de 89,9 % des unités qui étaient marquées dans la base de sondage 2019. Le suivi imparfait des individus d'un millésime à l'autre conduit dans tous les cas à ne retrouver dans la base de sondage 2020 que 96,3 % des individus marqués dans la base de sondage 2019. Le fait de retenir cette solution pour la gestion des déménagements conduit ainsi à cesser de marquer 6,4 % des unités qui étaient disjointes de la base de tirage 2019.

Conclusion

La coordination négative des échantillons pour les enquêtes auprès des ménages de l'Insee consiste à disjoindre de la base de sondage les échantillons précédemment tirés. Les pondérations des échantillons sont calculées par post-stratification sur la base de sondage, soit par strate pour les tirages à 1 degré, soit par strate croisée à l'unité primaire pour les tirages à 2 degrés. Sur le plan méthodologique, cette solution est tout à fait valide tant que des échantillons complémentaires sont tirés afin que le cumul de l'échantillon enquêté et l'échantillon complémentaire ait la même structure que la base de sondage.

Les tirages à 2 degrés nécessitent d'utiliser des allocations autopondérées pour ne pas biaiser les estimateurs pour des échantillons tirés ultérieurement avec 1 degré de tirage. Le passage d'un millésime de la base de sondage au millésime suivant conduit à tirer un échantillon d'unités qui n'étaient pas présentes dans le millésime précédent de la base de sondage.

La complexité nouvelle induite par l'introduction des tirages d'individus à l'Insee est liée au déménagement d'individus entre les différentes unités primaires découpant le territoire. La solution retenue conduit à ne pas conserver la mémoire des tirages du millésime précédent pour les individus qui changent d'unité primaire entre les 2 millésimes. Cela implique le démarquage de 6,4 % des unités tirées par le passé alors qu'elles ont été retrouvées dans le nouveau millésime 2020 de la base de sondage.

Cet article n'évoque nullement la disjonction des tirages d'individus et des tirages de logements de manière jointe. Si, sur le plan théorique, cela implique des complexités supplémentaires, des solutions semblent exister pour éviter de tirer les individus des logements sélectionnés en raisonnant à partir des tirages par grappe. La réciproque est plus complexe, car un logement ne peut être

marqué dès qu'un de ses occupants est tiré, car le cas échéant, les logements des familles nombreuses auraient une probabilité plus forte d'être marqués. Par contre, les déménagements présentent à nouveau des difficultés méthodologiques. Au final, la principale difficulté d'une gestion jointe de la disjonction des tirages d'individus et des tirages de logements reste sans doute le développement de solutions potentiellement complexes dans un cadre applicatif.

Bibliographie

[1] Crenner E., « Fidéli : un fichier démographique d'origine fiscale au service des utilisateurs », *Séminaire de méthodologie statistique de l'Insee*, novembre 2018.

[2] Hesse C., « Sampling co-ordination : A review by country », *Documents de travail de l'Insee*, Technical Report E9908, 1999.

[3] McKenzie R., Gross B., « Synchronised sampling », *Proceedings of ICES II The Second International Conference on Establishment Surveys*, pp 237-244, 2001.

[4] Merly-Alpa T., Pendoli P.A., Vincent L., « Passer du recensement aux sources fiscales pour le nouvel échantillon-maître de l'Insee : le projet Nautile », *Séminaire de méthodologie statistique de l'Insee*, novembre 2018.

[5] Ohlsson E., « Coordination of samples using permanent random numbers », *Business Survey Methods, Wiley, New York,* chapter 9, pp. 153-169, 1995.

[6] Paliod N., « Disjonction d'échantillons d'individus dans les tirages de l'Insee », *11e Colloque International Francophone sur les Sondages*, 2021.

[7] Sillard P., Faivre S., Paliod N., Vincent L., « Pour les enquêtes auprès des ménages, l'Insee rénove ses échantillons », *Courrier des statistiques*, 4, pp. 81-100, 2020.

[8] Vincent L., Chevalier M., Costa L., Delta L., Deroyon T., Favre-Martinoz C., Givois S., Guillo C., Merly-Alpa T., Paliod N., Pendoli P.A., Sauvaget T., « Document de travail Nautile », *Série des documents de travail Méthodologie Statistique de l'Insee*, publication à venir.