Journées de méthodologie statistique de l'Insee 2022

Échantillonnage de l'enquête Trajectoires et Origines 2

Thomas MERLY-ALPA (*), Nicolas PALIOD (**), Willy THAO KHAMSING (***)

(*) Ined, Service des enquêtes et sondages
(**) Insee, Direction de la méthodologie et de la coordination statistique et internationale
(***) Insee, Direction des statistiques démographiques et sociales

thomas.merly-alpa@ined.fr
nicolas.paliod@insee.fr
willy.thao-khamsing@insee.fr

Mots-clés. (6 maximum) : enquête démographique et sociale, échantillonnage de populations rares, plan de sondage complexe, sondage indirect.

Domaines. théorie des sondages, échantillonnage, enquêtes.

Résumé

L'enquête Trajectoires et Origines 2 2019/2020 réalisée par l'Ined et l'Insee est une réplique de la première enquête « Trajectoires et Origines » de 2008-2009. Cette enquête étudie la diversité des populations en France et la situation des populations d'origine immigrée.

L'objectif de l'enquête est de produire des statistiques sur l'ensemble de la population mais aussi sur des groupes de populations parfois mal connues (les immigrés et leurs descendants, en les distinguant selon la zone géographique d'origines). Pour ce faire, l'échantillonnage de l'enquête repose sur plusieurs sous-échantillons bien identifiés, sélectionnés dans l'enquête annuelle de recensement 2018. La constitution de l'échantillon des immigrés est stratifiée par zone géographique d'origines, avec l'objectif de conserver une bonne représentation globale des immigrés tout en limitant la dispersion des poids. L'échantillon des enfants d'immigrés (descendants d'immigrés), stratifié par zone géographique d'origines des parents, est plus difficile à construire, car cette information n'est pas disponible initialement dans la base de sondage. Il nécessite un appariement entre l'enquête annuelle de recensement et les fichiers anonymisés de l'État-Civil pour identifier les individus de la base de sondage susceptibles d'être descendants d'immigrés. Puis, une consultation en mairie des bulletins d'État-Civil de naissance de ces individus est réalisée afin de recueillir des informations sur le lieu de naissance de leurs parents, et pose à cet égard des difficultés de mise en place de l'échantillonnage.

Enfin, nouveauté de cette édition, l'enquête s'intéresse également aux troisièmes générations. Une partie de cette population d'intérêt - les individus de troisième génération non européenne - a été enquêtée dans une enquête expérimentale (TeO2-G3). Ils ont été préalablement identifiés

dans l'enquête TeO2 par sondage indirect, via l'interrogation de leurs parents à travers des questions dédiées au repérage de ces derniers.

L'objet de cet article est de décrire de façon détaillée la méthodologie d'échantillonnage de l'enquête TeO2 et TeO2-G3.

Abstract

The Trajectories and Origins 2 2019/2020 (TeO2) survey is a statistical survey conducted by both Ined (French Institute for Demographic Studies) and Insee (French National Statistical Institute). It is the second edition of a survey conducted in 2008-2009 which studies the diversity of populations in France and the situation of immigrant populations. It aims at producing national statistics on the diversity of the entire population but also on groups of populations that are sometimes poorly known (immigrants and their descendants). This paper describes the design of the immigrant sample, of the immigrants' children sample (descendants of immigrants) and of the third generation's sample which is a novelty of the current edition.

1 Enquête TeO2 et volet complémentaire TeO2-G3

L'enquête TeO2 (Trajectoires et Origines 2) est une réédition d'une enquête menée conjointement par l'Ined et l'Insee (dont la méthodologie est détaillée par Algava et Lommeau (2013)) ayant été largement réutilisée (plus de 220 chercheur-e-s en France et à l'étranger, par des administrations, donnant lieu à plus de 150 publications scientifiques [1] : articles de revues scientifiques, chapitres, ouvrages...) et ayant bénéficié d'une forte couverture médiatique. Cette enquête a pour but de fournir des informations fiables sur l'intégration comme processus d'accès aux ressources économiques et à la reconnaissance sociale, sur les inégalités selon l'origine et sur les questions discriminations dans la société française comme obstacle à l'égalité sociale entre groupes. Elle s'intéresse à l'articulation entre l'origine et les autres catégories de distinction dans la société française (genre, âge, milieu social, configuration familiale, lieu de vie etc.) afin d'analyser les processus d'intégration, de discrimination et de construction identitaire concernant toute la population dans la société française. Dans le but de mesurer l'impact des origines sur les trajectoires sociales et l'accès aux ressources au fil des générations, un des enjeux est également de savoir si les inégalités observées pour les enfants d'immigrés d'origine non européenne, se maintiennent ou s'atténuent à la génération suivante.

Si cette enquête sert de référence pour les travaux sur les immigrés et leurs enfants dans la recherche et dans la statistique publique, sa précédente édition remontait à 2009, et une mise à jour des résultats une décennie après la première enquête est très attendue et souhaitée par les pouvoirs publics, la société civile et les scientifiques, afin de fournir des informations détaillées sur des populations au cœur de débats politiques et sociaux que sont les immigrés et leurs descendants. Construire une enquête sur la diversité des populations en France demande ainsi un échantillonnage adapté pour s'assurer de mesurer l'ensemble des situations avec la précision nécessaire pour l'analyse des résultats et garantir la comparabilité avec les résultats de l'enquête précédente.

Cette enquête est réalisée auprès d'individus et concerne cinq sous-échantillons distincts :

1. Immigrés : individus nés à l'étranger, avec une nationalité étrangère à la naissance ;
2. Domiens : individus nés dans les départements d'Outre-Mer français ;
3. Descendants d'immigrés : individus non immigrés dont au moins un des parents est né à l'étranger, avec une nationalité étrangère à la naissance ;

1. cf. `https://teo1.site.ined.fr/`

4. Descendants de domiens : individus non domiens dont au moins un des parents est né dans les départements d'Outre-Mer français ;
5. Population générale : individus appartenant à la population générale, y compris ceux concernés par les autres sous-échantillons ;

L'enquête, en préparation depuis début 2017, a une collecte longue de juillet 2019 à octobre 2020. Cette collecte se répartit en plusieurs vagues :
— 1^{er} juillet 2019 au 31 décembre 2019 : immigrés, domiens et population générale ;
— 1^{er} janvier 2020 au 31 novembre 2020 : enfants d'immigrés, de domiens.

L'enquête expérimentale TeO2-G3 porte sur les individus de troisième génération non européenne avec comme objectif d'étudier la question de la persistance de l'influence des origines sur leurs trajectoires sociales. En effet, depuis 2008 (date de la précédente édition de l'enquête TeO), davantage d'enfants des descendants d'immigrés sont devenus adultes et il s'agit également de savoir si les inégalités observées pour la deuxième génération, d'origine non européenne, se maintiennent ou disparaissent à la génération suivante. Cette enquête a également une collecte longue qui s'est étalée du 10 mars 2020 au 31 janvier 2021.

2 Les premières générations et la population générale

Les individus enquêtés lors de la première vague appartiennent aux sous-échantillons des immigrés, des domiens, ainsi qu'au sous-échantillon en population générale. Au total, près de 27 000 individus sont enquêtés en vague 1, répartis de la façon suivante :
— 18 277 au titre du sous-échantillon Immigrés
— 1 646 au titre du sous-échantillon Domiens
— 6 928 au titre du sous-échantillon Population Générale

De nombreuses questions se sont posées pour la réalisation du tirage de ces échantillons. En ce qui concerne les immigrés, les objectifs de l'enquête amènent à cibler certaines origines pour pouvoir commenter des chiffres concernant certains groupes d'origines (par exemple, les conditions d'emploi des immigrés d'Afrique Subsaharienne par rapport aux immigrés de pays de l'Union Européenne). Il est néanmoins impossible d'assurer qu'une diffusion sera possible sur l'ensemble des groupes d'origines : il faut les cibler, et assurer qu'un volume de répondants suffisant sera atteint. Par ailleurs, il faut que la diffusion sur le groupe constitué de l'ensemble des immigrés soit possible, et donc que la dispersion des poids issue du tirage ne soit pas trop importante.

Dans la première édition de l'enquête TeO, 9 groupes d'origines étaient ciblés, et environ 1 200 individus d'autres origines étaient enquêtés ; pour cette nouvelle édition, les concepteurs de l'enquête souhaitaient pouvoir intégrer de nouveaux groupes d'origines tels que la Chine et les pays dont les ressortissants composent une partie importante des personnes bénéficiant du statut de réfugié en France. Cependant, rajouter trop de groupes conduit, par voie de conséquence, et dans un contexte de taille d'échantillon totale contrainte, à réduire le nombre d'immigrés des pays non-ciblés, et ainsi à augmenter leur poids, ce qui a un impact sur la précision des estimations sur l'ensemble des immigrés.

Pour étudier si l'intégration de nouveaux groupes réduit trop la précision globale, nous allons évaluer la précision obtenue sur deux variables issues de TeO1 (proportion d'individus apportant une aide financière régulière à l'étranger, proportion d'individus sans diplôme) si l'on séparait la population "Autre pays" en plusieurs groupes afin de réaliser d'autres surreprésentations. On étudie deux cas d'exploitation de l'échantillon résiduel :

1. un groupe avec une taille d'échantillon n_1, ce qui réduit la taille de l'échantillon pour le groupe résiduel à environ $n_R \approx 1\,200 - n_1$;

2. deux groupes, avec des tailles d'échantillon égales $n_1 = n_2$, ce qui réduit la taille de l'échantillon pour le groupe résiduel à environ $n_R \approx 1\,200 - (n_1 + n_2)$;

Les graphiques suivants montrent les résultats obtenus en termes de précision lorsque n_1, la taille d'échantillon souhaitée pour le(s) groupe(s) ajouté(s), varie.

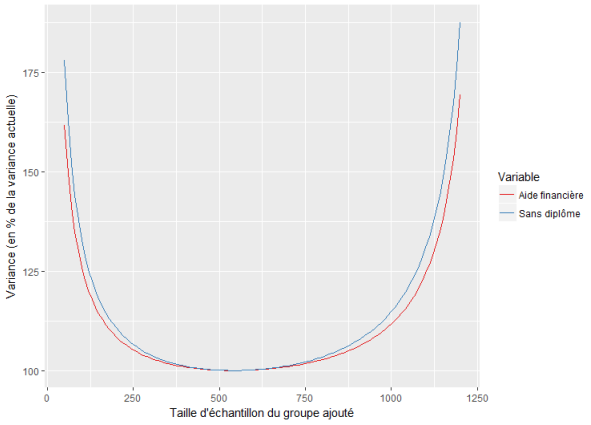


FIGURE 1 – Un groupe

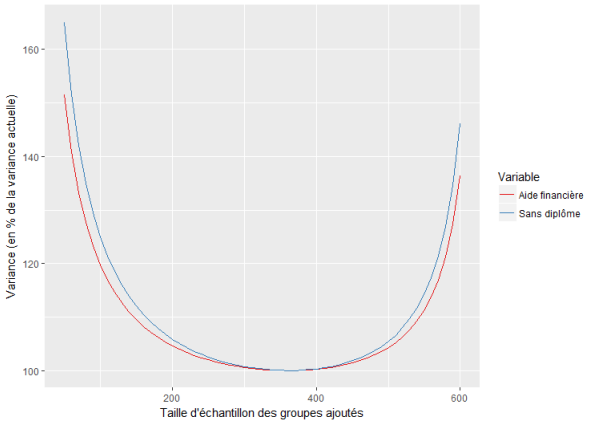


FIGURE 2 – Deux groupes

On constate alors qu'il est possible de rajouter :

1. un groupe comportant entre 400 et 700 répondants ;
2. deux groupes comportant environ 350 répondants chacun.

sans que cela ne modifie trop la précision obtenue sur les deux variables considérées. Le tableau ci-après résume les tailles des sous-échantillons et les origines choisies, en prenant en compte les objectifs de diffusion (divers selon les origines), ainsi que des taux de réponse anticipés à partir de la précédente édition et d'informations sur la collecte des enquêtes en face-à-face à l'Insee.

Origine	Répondants attendus	Taux de réponse	Nombre de FA
Algérie	1000	47%	2126
Maroc et Tunisie	1000	43%	2304
Afrique sahélienne	800	41%	1936
Afrique guinéenne ou centrale	800	46%	1730
Espagne, Italie, Portugal	1000	55%	1820
Autres pays de l'UE28	800	52%	1528
Turquie	800	52%	1533
Asie du Sud-Est	800	51%	1559
Chine	700	49%	1431
Pays avec de nombreux réfugiés	700	48%	1451
Autres pays	600	42%	1426

TABLE 1 – Objectifs et allocations TeO2 – populations immigrées

La base de sondage utilisée est l'Enquête Annuelle de Recensement 2018 (EAR 2018), car elle permet de donner une information précise sur la nationalité à la naissance, ce que ne permet pas la source fiscale, par exemple. L'enquête TeO étant une enquête individuelle, il est indispensable de construire une base de sondage nominative. Via des opérations d'enrichissement complémentaire (saisie automatique et manuelle), les noms et prénoms de certains individus appartenant au

champ de l'enquête (nés entre 1960 et 2001 inclus, résidant en France métropolitaine, en logement ordinaire) ont été ajoutés à la base de sondage. Comme pour la précédente édition l'opération de saisie manuelle des noms et prénoms a été uniquement réalisée pour les individus nés certains jours de l'année, le 20 au 25 de chaque mois. Pour cette nouvelle édition le 1^{er} janvier a également été inclus dans l'opération de saisie manuelle des noms et prénoms. En effet, cette date a un statut atypique car concernant de nombreux immigrés d'origines particulières n'ayant pas d'état civil complet à leur arrivée en France. Lorsque les dates de naissance sont incertaines, certains immigrés déclarent par défaut être nés le 01/01 (cela concerne 41% des immigrés d'origine turque, par exemple).

Pour certaines origines pour lesquelles la surreprésentation était importante, se restreindre à ces quelques jours contraignait trop fortement les possibilités de tirage de l'échantillon. Pour lever cette contrainte d'échantillonnage, il a été décidé d'étendre la sélection des jours naissance pour certains groupes d'origines. Les noms et prénoms des individus concernés sont issus de saisies automatiques (par lecture optique). Par ailleurs, afin d'assurer une disjonction entre les sous-échantillons pour éviter non seulement qu'un même individu soit échantillonné plusieurs fois, mais aussi que plusieurs individus d'un même logement soient enquêtés, afin de limiter la charge de collecte, le tirage du sous-échantillon en population générale a lui-aussi été réalisé au sein d'une population dont les noms ont été saisis de façon automatisée.

Une fois les allocations définies pour l'ensemble des strates de tirage (i.e. des groupes d'origines), celles-ci ont été ventilées région par région au prorata de la taille de la strate. Des ajustements ont été nécessaires pour s'assurer que la taille de l'échantillon ne dépassait pas celle de la base, d'une part, mais aussi pour réduire la taille de l'échantillon en Île-de-France, région qui aurait sans cela concentrait une part trop importante de la collecte, ce qui aurait entraîné un risque sur les taux de réponse. L'échantillonnage se fait ensuite à probabilités inégales, en suivant la formule suivante pour un individu k donné :

$$\pi_k = \frac{w_{EAR,k}\ w_{ZAE,k}\ w_{JNAI,k}\ w_{IndLog,k}\ w_{ssech05,k}}{\sum_{l \in strate,reg} w_{tot,l}} n_{strate,reg}$$

Chacun des termes de cette formule correspond à la prise en compte d'une étape du tirage afin de limiter la dispersion des poids de sondage. Plus spécifiquement :

— $n_{strate,reg}$ est l'allocation par strate et par région ;

— $w_{EAR,k}$ poids lié au tirage de l'EAR 2018 : pour mémoire, le recensement français est rotatif et concerne environ 10% de la population chaque année, avec un fonctionnement différencié entre les petites communes et les plus grandes, où certaines adresses (comme les nouvelles constructions) sont surreprésentées. On pourra se reporter à Godinot (2006) pour une description plus précise du fonctionnement des EAR. Il est donc nécessaire de prendre en compte cette phase pour le tirage.

— $w_{ZAE,k}$ poids lié au tirage de l'Échantillon-Maître (EM) Octopusse : l'EM Octopusse décrit dans Faivre, Christine (2009) qui est en vigueur à l'Insee pour la période 2010 - 2020, est un ensemble de zones géographiques (communes ou groupes de commune) dans lesquels se concentre la collecte des enquêtes. Cela concerne également l'enquête TeO2, sauf pour certaines strates où l'EM tiré en 2009 ne couvrait pas suffisamment les communes dans lesquelles vivent la majorité des individus concernés. Dans ces cas, le poids $w_{ZAE,k}$ est uniformément égal à 1, et l'échantillonnage se fait sur l'ensemble du territoire de la région concernée.

— $w_{JNAI,k}$ poids lié au jour de naissance : comme indiqué plus haut, seuls certains jours de naissance sont mobilisés pour l'échantillonnage, et il faut prendre en compte cette restriction de champ, en faisant l'hypothèse d'un sondage aléatoire. Par ailleurs, pour

certaines strates, comme celle déjà évoquée des immigrés d'origine turque, la spécificité du 1^{er} janvier force un traitement à part de ce jour. Il est alors considéré comme exhaustivement échantillonné ; son poids est forcé à 1, et les individus nés ce jour là n'ont pas un poids visant à représenter les autres individus de leur strate nés les autres jours, car on suppose qu'il existe une différence intrinsèque au fait d'avoir sa date de naissance fixée administrativement au 1^{er} janvier.

— $w_{IndLog,k}$ poids lié au tirage d'un individu par logement, afin de limiter la charge de collecte : ce poids vise à rééquilibrer la probabilité de sélection d'individus appartenant à des ménages ayant un nombre différent de membres appartenant au champ de l'enquête ;

— $w_{ssech05,k}$ poids spécifique pour le tirage en population générale, dont l'échantillonnage est réalisé dans un second temps une fois que chaque strate concernant les populations immigrées et domiennes est échantillonnée ; le poids prend en compte le mécanisme de disjonction, plus fortement marqué dès lors que le ménage est de taille plus importante, ceci afin d'éviter de sous-représenter des individus en couple avec une personne immigrée, par exemple, les questions de cohabitation et de mise en couple étant centrales dans l'enquête ;

— Enfin, $\sum_l w_{tot,l}$ est la somme des poids $w_{tot,k} = w_{EAR,k}\, w_{ZAE,k}\, w_{JNAI,k}\, w_{IndLog,k}\, w_{ssech05,k}$ mise au dénominateur pour garantir le respect de l'allocation de tirage.

3 Les deuxièmes générations : consultation des registres d'État civil en mairie

Un individu descendant d'immigré ou de domien appelé « G2 » vérifie les deux conditions suffisantes : il n'est pas immigré (né étranger à l'étranger) ou né dans les DOM, mais l'un de ses parents (ou les deux) l'est. Échantillonner cette population est un défi, comme l'indiquent Algava et Lhommeau (2009) pour TeO1. Un dispositif exceptionnel avait été mis en œuvre pour échantillonner les enfants d'immigré(s), pour lesquels il avait été nécessaire de constituer une base de sondage ad hoc, ces derniers ne pouvant être directement identifiés dans le recensement. En effet, on ne dispose pas facilement de l'ensemble des informations nécessaire au tirage : est-ce que l'individu est un descendant d'immigré ? Si oui, de quelle origine (au sens de l'origine du parent immigré) ? Cette dernière information est importante car, de la même façon que pour les immigrés, l'analyse des résultats est faite sur l'ensemble de la population mais aussi par groupe d'origines.

L'approche utilisée mobilise également l'EAR 2018 même si l'information complète n'est pas disponible (sauf, éventuellement, pour les individus cohabitant avec leurs parents, mais on ne souhaite pas introduire un biais aussi important dans le tirage de l'échantillon). Pour identifier les individus descendants d'immigrés potentiels [2] dans la population, nous utilisons les fichiers d'État Civil de la Population, qui contiennent l'information sur le lieu de naissance des parents de chaque individu. Il en existe deux versions :

— Une version anonymisée, accessible de façon informatique à l'Insee et par les statisticiens publics, qui ne permet pas d'identifier directement un individu mais donne des éléments sur l'ensemble des personnes du sexe choisi nées dans une commune un jour donné. Ce

2. L'opération en mairie permet de collecter le lieu de naissance des parents. Mais elle ne suffit pas à déterminer de façon certaine parmi les personnes nées en France celles qui sont enfants d'immigré(s) car il est nécessaire de connaître la nationalité des parents à la naissance. C'est en effet à partir du questionnaire de l'enquête TeO2 que l'on recueille cette information à l'issue de l'interrogation des individus répondants sur la nationalité des parents. L'opération en mairie constitue donc un étape dans l'identification des enfants d'immigrés.

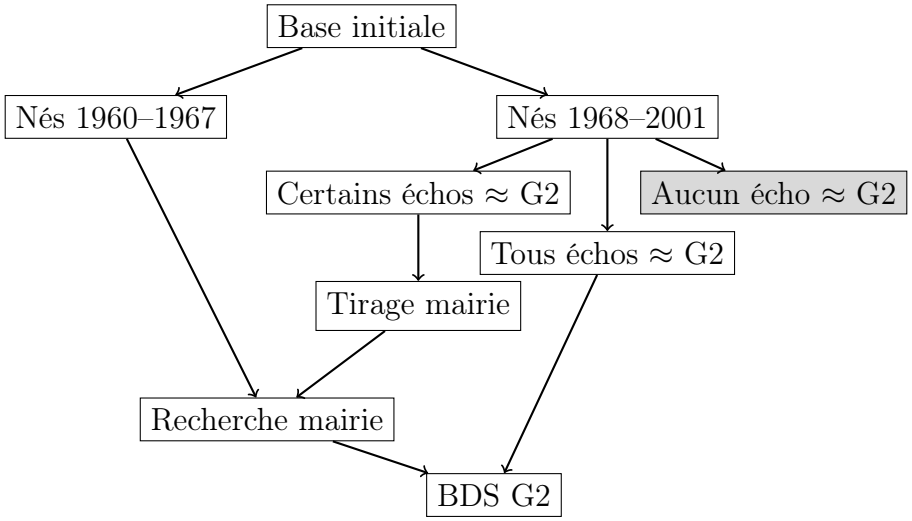
fichier couvre uniquement les années 1968 et après.

— Des versions complètes, accessibles uniquement avec accord des administrations compétentes auprès des mairies de naissance des individus.

L'idée de la construction de l'échantillon est de mobiliser autant que possible les fichiers anonymisés en recherchant les « échos » de chaque individu recensé, c'est-à-dire l'ensemble des personnes du même sexe nées dans la même commune et le même jour. Il y a trois cas de figure possible :

1. Soit tous les échos de l'individu ont un parent né à l'étranger ;
2. Soit certains des échos de l'individu ont un parent né à l'étranger ;
3. Soit aucun des échos de l'individu n'a de parent né à l'étranger

La recherche complémentaire en mairie n'est intéressante que dans le second cas, pour déterminer si l'individu en question fait bien partie de la population des descendants d'immigrés. En effet, dans le troisième cas, il n'y a aucune chance ; dans le premier cas, s'il n'est pas sûr que l'individu soit descendant d'immigré (le parent né à l'étranger peut être de nationalité française à la naissance), aucune information complémentaire ne sera disponible dans les registres, si ce n'est le lieu de naissance exact des parents de l'individu lorsqu'un individu a plusieurs échos avec un parent né à l'étranger. Le schéma ci-après résume le processus d'échantillonnage :



L'utilisation des fichiers de l'État Civil permet ainsi de se rapprocher de la population ciblée, les descendants d'immigrés. Mais l'échantillon couvrira un champ plus large que cette population, puisqu'il intègrera des individus dont un des parents est né à l'étranger sans être immigré pour autant.

Afin de limiter les coûts, un nombre limité de relevés en mairie peut être réalisé. Il faut donc réaliser un échantillonnage d'individus pour lesquels un relevé en mairie sera fait sous cette contrainte, et avec pour objectif multiple d'assurer, d'une part, que la dispersion des poids de sondage des individus soit minimale, et, d'autre part, que les objectifs en terme de nombre de répondants pour les groupes d'origine soient atteints. Pour cela, on réalise autant que possible un tirage à probabilités égales (modulo les étapes antérieures de sélection telles que l'EAR, décrites en partie 2), sauf en ce qui concerne certaines origines dites « rares ». En effet, pour certaines origines, le nombre de descendants d'immigrés est faible dans la population ; il est donc crucial de réaliser un relevé en mairie dès lors qu'il y a une chance de pouvoir atteindre un individu correspondant à ce groupe. Ainsi, les individus pour lesquels un écho correspond à une origine dite « rare » sont systématiquement inclus dans l'échantillon de relevés mairies.

En tout, 100 014 individus ayant potentiellement un parent né à l'étranger ou dans les DOM sont ainsi échantillonnés et font l'objet de relevés en mairie. 75 032 d'entre eux sont des individus

nés entre 1968 et 2001 et dont l'appariement avec les fichiers de l'État Civil présente seulement certains échos avec au moins un parent né à l'étranger ou dans les DOM. Les 24 982 individus restants sont nés entre 1960 et 1967 et aucun appariement avec les fichiers de l'État Civil n'a été possible.

À l'issue de l'opération de relevés en mairie, toutes les informations nécessaires à la constitution de la base de sondage des individus de nationalité française et ayant au moins un parent né à l'étranger ou dans les DOM sont disponibles. Elle est composée de :

1. 29 784 individus pour lesquels les relevés en mairie ont conclu que ces individus appartenaient au champ de ce sous-échantillon ;
2. 8 252 individus dont l'appariement avec les fichiers de l'État Civil avait assuré que ces individus avaient au moins un parent né à l'étranger.

Comme pour le calcul des allocations pour les sous-échantillons d'immigrés et de domiens présenté dans le tableau 1, le nombre d'unités à échantillonner par origine des parents se déduit de la cible du nombre escompté de répondants ainsi que du taux de réponse par origine des parents dans TeO1. Cependant, le calcul du volume de l'échantillon à tirer pour les descendants d'immigrés présente une spécificité : la cible de répondants de ces sous-échantillons concerne les enfants d'immigrés et de domiens tandis que le champ de ces sous-échantillons est plus étendu puisqu'il concerne l'ensemble des enfants d'au moins un parent né à l'étranger, quelle que soit la nationalité à la naissance de ce dernier, ou dans les DOM. Il est donc nécessaire de tenir compte du taux de répondants qui s'avèrent avoir au moins avoir un parent immigré ou domien. Là encore, ce taux par origine est extrait des résultats de TeO1. Ces calculs aboutissent à un échantillon de 24 703 individus à sélectionner. Compte-tenu de la capacité de collecte du réseau d'enquêteurs, un facteur proportionnel a été appliqué pour tirer 23 917 individus. Les allocations sont présentées ci-après.

Origine	Répondants attendus	Taux de réponse	Taux d'enfants d'immigrés et de domiens	Nombre d'unités à tirer
Algérie	1600	57%	49%	5535
Maroc et Tunisie	1200	58%	59%	3405
Afrique sahélienne	800	56%	75%	1841
Afrique guinéenne ou centrale	800	56%	69%	1983
Espagne, Italie	800	62%	97%	1285
Portugal	800	60%	99%	1301
Autres pays de l'UE28	800	62%	87%	1433
Turquie	800	58%	99%	1346
Asie du Sud-Est	800	60%	73%	1765
Autres pays	600	61%	37%	2568
DOM	800	55%	100%	1455

TABLE 2 – Objectifs et allocations TeO2 – populations descendant d'immigrés et de domiens

Pour être en mesure d'effectuer le tirage de cet échantillon, une dernière information est nécessaire : l'origine associée aux individus dans la base de sondage. Cette dernière étape ne coule pas de source bien que dérivant de l'ensemble de la procédure d'appariement et de recherche en mairie décrite précédemment. Deux sujets nécessitent une vigilance particulière. D'abord, il s'agit d'affecter à un individu l'origine d'un de ses parents lorsque ses deux parents sont nés à l'étranger ou dans les DOM et sont d'origines différentes, afin de déterminer la strate de tirage à laquelle sera affectée l'individu. Ensuite, parmi les 8 252 individus de la base de sondage n'ayant

pas fait l'objet de relevés en mairie car on avait la certitude qu'un de leurs parents était né à l'étranger, certains peuvent avoir plusieurs échos dans l'État Civil avec des parents certes nés à l'étranger ou dans les DOM, mais dans une zone géographique différente selon les échos. Pour ces individus, on ne dispose donc que d'une probabilité d'être associés à une origine.

Pour pallier cet écueil, le tirage a été réalisé dans 14 strates. 11 strates correspondent aux origines mentionnées dans le tableau 2. Les individus dont on connaît l'origine des parents avec certitude sont affectés à ces 11 strates. Les 3 strates restantes sont des strates dans lesquelles on ne dispose que de la probabilité d'un individu d'avoir des parents d'une origine donnée. Certaines origines ayant été fortement surreprésentées, les individus ayant une probabilité non nulle d'avoir des parents originaires d'une zone géographique « rare » ont été séparés des autres individus. Compte-tenu du fait que la taille de l'allocation prévue pour la strate de personnes d'origine Afrique guinéenne ou centrale était inférieure à taille de la base de sondage, les individus ayant une probabilité non nulle d'avoir des parents nés dans cette zone ont été intégrés d'office à l'échantillon. Ainsi, les 3 strates supplémentaires sont :

1. une strate exhaustive d'individus pour laquelle l'origine n'est pas connue avec certitude mais ayant une probabilité non nulle d'être d'origine Afrique centrale ou guinéenne ;
2. une strate d'individus pour laquelle l'origine n'est pas connue avec certitude, ayant une probabilité non nulle d'être associés à une origine surreprésentée mais ayant une probabilité nulle d'avoir au moins un parent né en Afrique centrale ou guinéenne ;
3. une strate d'individus pour laquelle l'origine n'est pas connue avec certitude et ayant une probabilité nulle d'être associés à une origine surreprésentée.

Dans chaque strate, la probabilité d'inclusion des individus est calculée itérativement pour aboutir à la formule suivante faisant intervenir les poids $w_{k,mairie}$ issus du tirage des relevés en mairie :

$$\pi_k = (n_{strate} - n^{exh}_{strate}) \frac{c_{idf} w_{k,mairie}}{\sum_{strate, i \notin exh} c_{idf} w_{i,mairie}}$$

c_{idf} est un coefficient qui sous-représente les individus qui vivent en Ile-de-France afin de limiter la charge de collecte dans cette région à 20% de l'échantillon total. n_{strate} désigne l'allocation par strate et n^{exh}_{strate} le nombre d'unités tirées exhaustivement dans cette strate. En effet, les probabilités ainsi définies visent à compenser les déséquilibres importants des poids après relevés en mairie. Lorsque les individus avaient potentiellement des parents nés dans une zone géographique « rare », ils ont été fortement surreprésentés à l'étape de la sélection des unités pour les relevés en mairie. Or, une partie de ces individus ont des parents nés à l'étranger mais dans des zones géographiques moins surreprésentées. Ils ont donc, avant cette étape de tirage, un poids plus faible que les individus dont au moins un parent est né à l'étranger mais qui, suite aux appariements avec les fichiers de l'État Civil, avaient une probabilité nulle d'avoir un parent né dans une zone géographique « rare ». En outre, l'absence d'information issue de l'État Civil pour les individus nés entre 1960 et 1967 implique que ceux qui sont dans le champ de l'enquête et dans une strate surreprésentée ont un poids bien plus élevé que les individus de leur strate après les relevés en mairie mais avant la présente sélection d'unités.

Les allocations par strate sont d'abord calculées pour les strates dont l'origine des individus n'est pas connue avec certitude. Puis, le nombre d'individus attendus par origine est calculé pour chacune de ces strates d'origine incertaine, afin d'en déduire les allocations nécessaires pour atteindre les objectifs du tableau 2.

4 Les troisièmes générations non européennes : sondage indirect

Pour les premières et deuxièmes générations, les individus sont directement sélectionnés dans des bases de sondage avec la méthodologie présentée précédemment. Le sondage des troisièmes générations est différent, puisqu'il n'est pas possible de constituer une base de sondage issue de sources usuelles pour cette population d'intérêt. Ainsi la stratégie déployée pour cibler les individus de troisième générations s'effectue à deux niveaux :

1. le questionnaire de l'enquête TeO2, permet d'identifier les enquêtés dits de « troisième génération » grâce à l'introduction de questions sur la nationalité et le pays de naissance des 4 grands-parents. Cette identification conduit, dans le sous-échantillon de la population générale, à repérer principalement des personnes de troisième génération d'origine européenne (compte tenu des flux migratoires passés et de la taille de l'échantillon de la population générale). Les individus de troisième génération d'origine européenne sont estimés être assez nombreux dans le sous-échantillon de la population générale pour faire l'objet d'une exploitation sans nécessiter de surreprésentation particulière ;
2. La méthode retenue pour surreprésenter les individus de troisièmes générations non européennes est le sondage indirect. Leur sélection s'effectue à partir d'une base d'individus dont les coordonnées ont été collectées auprès de leur parent (individus de deuxième génération non européenne interrogés dans l'enquête TeO2).

Le volume d'individus de troisième génération non européenne dont les coordonnées ont été récupérées étant inférieur à la capacité prévue pour la collecte auprès de ces individus, tous ont été enquêtés, soit 362 individus. Il n'y a donc pas eu d'étape de sélection supplémentaire. Les complexités liées à ce sous-échantillon résident surtout dans les traitements post-collecte, qu'il s'agisse de partage de poids ou de l'interaction des étapes de redressement de ce sous-échantillon avec celles qui concernent les individus de deuxième génération (la méthodologie est décrite par Guin et Thao Khamsing (2021)).

Bibliographie

[1] Algava, E. et Ménard, C. (2007). A la recherche de la 2 eme génération pour l'enquête Trajectoires et Origines, Actes Colloque francophone sur les sondages 2007, Insee

[2] Algava, E. et Lhommeau, B. (2009). T'es où TeO ? À la recherche de la Deuxième Génération pour l'Enquête Trajectoires et Origines, Actes des Journées de Méthodologie Statistique de 2009, Insee

[3] Algava, E. et Lhommeau, B. (2013). À l'origine de l'enquête TeO : enjeux de l'échantillonnage, collecte et pondérations de l'enquête, Document de Travail Insee, `http://www.epsilon.insee.fr/jspui/bitstream/1/16994/1/f1304.pdf`

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

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

[6] Guin, O. et Thao Khamsing, W. (2021). Partage des poids pour l'enquete Trajectoires et Origines 2, *Actes 11e Colloque International Francophone sur les Sondages, 2021*

[7] Thao Khamsing, W., Guin, O., Merly-Alpa, T., Paliod, N. (2021), Enquête Trajectoires et Origines 2 - de la conception à la réalisation, document de travail Insee (F2021/03) (à paraître)