Journées de méthodologie statistique de l'Insee 2022

Redressements de l'enquête complémentaire sur les descendants d'immigrés de troisième génération d'origine non-européenne (TeO 2 - G3)

Olivier GUIN (*), Pierre TANNEAU (**), Willy THAO KHAMSING (**)

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

olivier.guin@insee.fr
pierre.tanneau@insee.fr
willy.thao-khamsing@insee.fr

Mots-clés : Sondage indirect, correction de la non-réponse, partage des poids, troncature des poids.

Domaine concerné : Théorie des sondages aval – Pondération et repondération, calage sur marges.

Résumé

L'enquête « TeO 2-G3 » est une enquête à visée expérimentale et complémentaire à l'enquête Trajectoires et Origines 2 (TeO 2) qui s'intéresse à une population ayant jusqu'à présent peu fait l'objet d'études statistiques : les descendants d'immigrés de troisième génération (c'est-à-dire les petits-enfants d'immigrés), et plus particulièrement ceux d'origine non-européenne, ou G3 non-européens. La question centrale soulevée par cette population est celle de la persistance de l'influence de l'origine sur la trajectoire sociale des individus. En effet, depuis 2008 (date de la précédente édition de l'enquête TeO), davantage d'enfants des descendants d'immigrés de deuxième génération (G2) d'origine non-européenne sont devenus adultes, et il s'agit d'étudier si les inégalités observées pour la deuxième génération se maintiennent ou disparaissent à la génération suivante.

Il n'existe pas de base de sondage issue de sources usuelles permettant d'identifier les individus G3 non-européens. Dans le cadre des enquêtes TeO 2 et TeO 2 - G3, ces derniers sont atteints par deux biais distincts :

- Soit directement, lors de l'enquête TeO 2, parmi les répondants qui s'avéreraient être des G3 non-européens selon leurs propres déclarations relativement aux origines migratoires de leurs parents et grands-parents ;
- Soit par sondage indirect, en constituant une base de sondage lors de l'interrogation des G2 d'origine non-européenne dans le cadre de l'enquête TeO 2. Ainsi, un bloc de questions dédié permet d'identifier lors de l'enquête les enfants de ces individus susceptibles d'être des G3 non-européens, et de demander aux enquêtés de fournir les coordonnées de ces enfants. Ces informations permettent donc de constituer une base de sondage dédiée, utilisée pour l'enquête complémentaire TeO 2 - G3 dont le questionnaire est identique à celui de l'enquête TeO 2.

Ce protocole particulier de collecte oriente fortement et complexifie la chaîne globale des redressements de l'enquête complémentaire. Ainsi, les étapes de redressement de l'échantillon de cette enquête s'articulent avec celles de l'enquête principale.

L'objet de cet article est de décrire de façon détaillée la méthodologie de redressement de l'enquête TeO 2 - G3, ainsi que les redressements complémentaires appliqués à l'enquête TeO 2 dans l'objectif de travailler conjointement sur les G3 non-européens de ces deux enquêtes.

Abstract

The « TeO 2 - G3 » survey is an experimental survey that complements the *« Trajectoires et Origines »* 2 (TeO 2) survey. It focuses on a population that has not been the subject of many statistical studies to date : the descendants of third-generation immigrants of non-European origin, or non-European G3. Since the previous edition of the TeO survey, more children of the descendants of second-generation immigrants of non-European origin have become adults, and the aim is to study whether the inequalities observed for the second generation are maintained or disappear in the following generation. The G3 are identified either directly, through their responses on the nationality and country of birth of their parents and grandparents ; or by indirect sampling, by constructing a complementary sample as the main survey is collected. This particular collection protocol strongly influences and complicates the overall chain of adjustments of the complementary survey.

1. L'enquête Trajectoires et Origines 2 et l'enquête complémentaire TeO 2 - G3

1.1. Deux enquêtes distinctes mais articulées

L'enquête Trajectoires et Origines 2 2019-2020 (TeO 2) est une enquête dont la maîtrise d'ouvrage est assurée conjointement par l'Institut national d'études démographiques (Ined) et l'Institut national de la statistique et des études économiques (Insee), visant à étudier la diversité des populations en France et la situation des populations d'origine immigrée. Il s'agit de la deuxième édition d'une enquête réalisée en 2008-2009 dont la méthodologie est décrite dans Algava et Lhommeau (2013).

Depuis la première édition de l'enquête TeO, davantage d'enfants des descendants de deuxième génération d'immigrés d'origine non-européenne sont devenus adultes. Un des objectifs de l'enquête TeO 2 est de dénombrer les descendants d'immigrés de troisième génération (notés G3 dans la suite du texte), c'est-à-dire les personnes ayant au moins un grand-parent immigré sans qu'eux-mêmes ou leurs parents soient immigrés, et d'étudier si l'influence des origines sur les trajectoires sociales observées pour les descendants d'immigrés de deuxième génération d'origine non-européenne se maintient, s'atténue voire s'accentue à la génération suivante.

L'enquête principale (plus de 27 000 répondants) ne permet d'atteindre qu'un nombre limité de G3 d'origine non-européenne (134 répondants). Pour compléter l'échantillon de cette enquête et afin de permettre des études sur la situation des G3 d'origine non-européenne, un volet complémentaire à l'enquête TeO 2 a été réalisé, identifié sous l'acronyme « TeO 2 – G3 ». Cette enquête complémentaire à visée expérimentale bénéficie d'un avis d'examen favorable du Comité du Label, tandis que l'enquête TeO 2 (ci-après désignée sous le terme d'enquête principale) bénéficie du label d'intérêt général et de qualité statistique.

Une description générale de ces deux enquêtes et de leur échantillonnage sont présentés dans Paliod et Merly-Alpa et Thao Khamsing (2022).

Ces deux enquêtes posent des défis méthodologiques en termes de méthodologie de redressement d'enquête à la fois dans les opérations dédiées (correction de la non-réponse, partage des poids, calage sur marge) mais aussi dans l'articulation de ces opérations. Cet article propose une description du schéma de redressements de ces deux enquêtes sur le champ des descendants d'immigrés de troisième génération d'origine non-européenne. Ainsi, il s'agit de construire un ensemble de poids permettant de travailler conjointement sur les observations de ces deux enquêtes sur leur champ commun, à savoir les G3 d'origine non-européenne.

1.2. L'échantillonnage des G3 dans ces deux enquêtes

Les G3 non-européens sont définis comme les individus respectant les conditions de champ de TeO 2, à savoir :

- Âgés de 18 ans à 59 ans ;
- Résidant en France métropolitaine ;
- Vivant en logement ordinaire ;

Et :

- Ayant au moins un grand-parent immigré d'origine extra-européenne, donc au moins un parent descendant d'immigré de deuxième génération (G2) d'origine non-européenne, sans qu'eux-mêmes ou leurs parents soient immigrés. Tout individu ayant un parent G2 d'origine non-européenne et l'autre immigré est exclu du champ, car un tel individu pourrait être classé comme un G2 ou un G3 non-européen.
- Nés en France métropolitaine, de même que le parent G2.

Il n'est pas possible de constituer une base de sondage issue de sources usuelles pour cette population d'intérêt, contrairement aux immigrés et descendants d'immigrés de deuxième génération. Ainsi la stratégie déployée pour cibler les G3 s'effectue à deux niveaux :

1. Le questionnaire de l'enquête TeO 2 permet d'identifier les G3 grâce à leurs réponses aux questions sur la nationalité et le pays de naissance de leurs quatre grands-parents (et de leurs parents). Ces questions n'étaient pas présentes dans la première édition de l'enquête, contrairement à celles portant sur la nationalité et le pays de naissance des parents qui sont communes aux deux éditions de l'enquête.

Compte tenu des flux migratoires passés et de la taille de l'échantillon de l'enquête principale, cette identification conduit à repérer principalement des descendants d'immigrés de troisième génération d'origine européenne. Ces individus sont estimés être assez nombreux dans l'enquête principale pour faire l'objet d'une exploitation sans nécessiter de surreprésentation particulière dans l'échantillon complémentaire.

2. Un échantillon complémentaire de G3 non-européens est constitué par sondage indirect.

La sélection des G3 non-européens s'effectue à partir d'une base d'individus dont les coordonnées ont été collectées auprès de leur(s) parent(s) interrogé(s) lors de l'enquête principale.

Ainsi, à la fin du questionnaire, tous les individus enquêtés lors de l'enquête principale qui sont descendants d'immigrés de deuxième génération et dont l'un des parents est un immigré non-européen, ayant au moins un enfant né en France, âgé de 18 ans et plus et résidant en France métropolitaine (ces enfants sont dits « éligibles »), sont invités à communiquer les coordonnées de chacun de leurs enfants respectant ces conditions.

Ainsi, les individus G3 sont atteints par deux biais distincts :

- Soit directement, parmi les répondants à l'enquête principale ;
- Soit par sondage indirect, avec constitution d'une base de sondage complémentaire au fur et à mesure de la passation de l'enquête principale.

Le questionnaire de l'enquête complémentaire est identique à celui de l'enquête principale, ce qui permet de travailler conjointement sur les deux enquêtes. Par ailleurs, le questionnaire permet de vérifier que les répondants à l'enquête complémentaire, identifiés comme des G3 non-européens *ex-ante* à partir des réponses de leur(s) parent(s) enquêté(s), le sont bien d'après leurs propres réponses.

1.3. Une différence de champ assumée

On notera que ce sondage indirect crée une différence de champ particulière. Ainsi, pour être échantillonnés, les G3 de l'enquête complémentaire doivent non seulement vérifier les conditions de champ de l'enquête, mais celles-ci doivent également être vérifiées par au moins un de leurs parents, interrogé dans le cadre de l'enquête principale, qui aurait communiqué leurs coordonnées. Ainsi, ce parent doit notamment être toujours en vie, résider en France métropolitaine et être âgé de 18 à 59 ans. Le champ de l'enquête complémentaire est donc plus restrictif que celui de l'enquête principale.

Cette différence de champ entre les G3 de l'enquête principale et ceux de l'enquête complémentaire, qui découle des différences d'échantillonnage et est assimilable à un biais de sélection lié à un défaut de couverture de l'enquête complémentaire, est partiellement prise en compte à l'aide d'un partage des poids détaillé ci-après. Le choix a été fait d'assumer cette différence de champ plutôt que de se restreindre à un champ commun qui serait celui des G3 dont les parents respectent les conditions de champ de l'enquête TeO 2, qui présente moins d'intérêt sociologique. Cette restriction de champ aurait par ailleurs conduit à une réduction de l'effectif total de G3 enquêtés.

2. Schéma global de redressement de TeO 2 et TeO 2 - G3 sur le champ des G3 non-européens

Ce protocole particulier de collecte oriente fortement la chaîne globale des redressements des deux enquêtes sur leur champ commun qui est celui des G3 non-européens.

Le point de départ du redressement de l'enquête complémentaire TeO 2 -G3 et de l'enquête TeO 2 sur le champ des G3 non-européens correspond aux poids finaux de l'enquête principale obtenus après différentes étapes de redressement décrites dans Guin et Tanneau et Thao Khamsing (2022). Les poids utilisés en entrée des redressements décrits ici correspondent donc à ceux des G3 de l'enquête principale, $r^{G3<G3}$, ainsi qu'à ceux des parents des G3 de l'enquête complémentaire, $r_{3,5}^{G2}$ [1].

Afin de calculer les poids des individus G3 de l'enquête complémentaire atteints par leurs parents G2 non-européens, on réalise les opérations suivantes :

- Une étape spécifique de correction de la non-réponse (dite « ACC ») permet de corriger le biais de sélection due au refus de parents G2 de transmettre les coordonnées de certains de leurs enfants G3 ;
- Pour les G3 atteints par sondage indirect à partir des G2 non-européens répondants à l'enquête principale et ayant accepté de fournir les coordonnées de leurs enfants, une opération de partage des poids (dite « PPSI ») permet de prendre en compte le nombre de liens du G3 avec la base de sondage indirect.

  Ainsi, le G3 peut être atteint par un ou deux de ses parents interrogés dans le cadre de l'enquête principale ; et, s'il n'est atteint que par un parent, l'autre peut ou non être lui-même un G2 non-européen, dont l'interrogation dans le cadre de l'enquête principale aurait également permis l'identification du G3 comme éligible ;
- Enfin, une correction de la non-réponse (dite « CNR G3 ») permet le transfert des poids des G3 non-européens échantillonnés mais non répondants au questionnaire sur les répondants.

À ce stade on dispose donc de deux échantillons pondérés : l'échantillon $r^{G3<G2}$ des G3 non-européens répondants atteints par sondage indirect via leurs parents G2, et l'échantillon $r^{G3<G3}$ des G3 non-européens répondants à l'enquête principale. Les poids des G3 atteints directement par le questionnaire de l'enquête principale sont déjà corrigés de la non-réponse.

- Un partage des poids supplémentaire (dit « PPLM ») permet finalement, à partir de ces deux ensembles de pondérations, la construction d'un jeu unique de poids composites pour l'échantillon global r^{G3} des G3 non-européens répondants à l'une ou l'autre des deux enquêtes ;
- Enfin, une étape de troncature est réalisée pour limiter la dispersion des poids.

1 Ces parents G2 ont été échantillonnés dans les sous-échantillons de l'enquête principale (3) et (5), décrits dans Merly-Alpa et Paliod et Thao Khamsing (2022), qui correspondent respectivement aux descendants d'immigrés de deuxième génération et à la population générale.

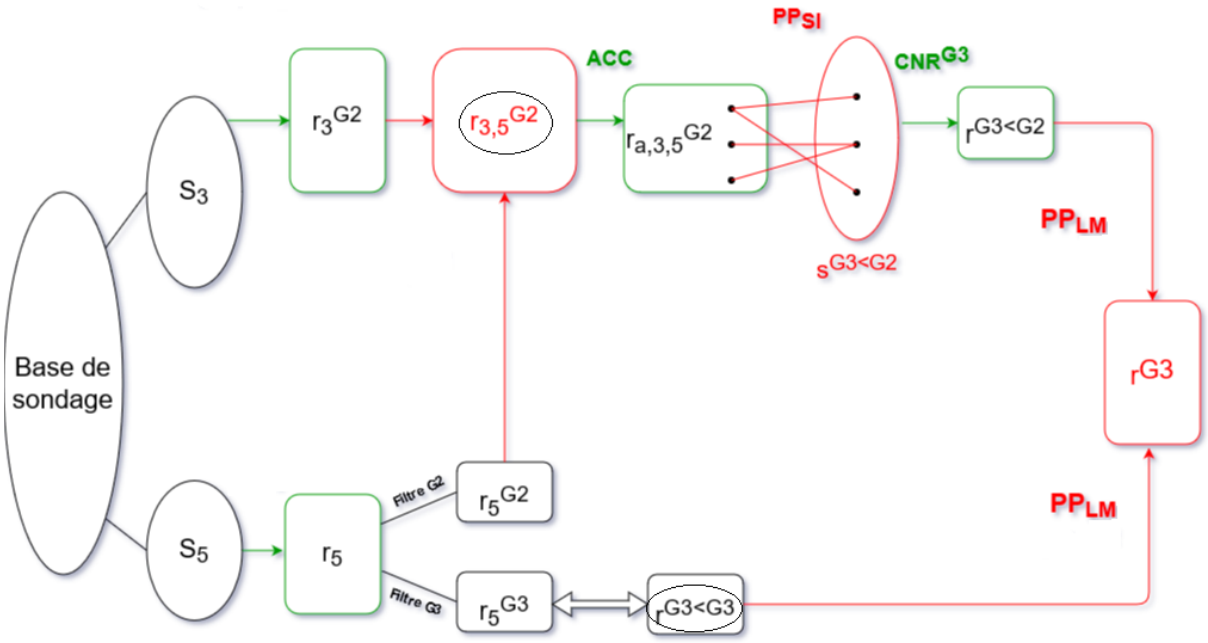

Figure 1: Redressements de l'enquête complémentaire TeO 2 - G3 et de l'enquête TeO 2 sur le champ des G3 non-européens

Note : Les poids utilisés comme points de départ de ces redressements, issus des redressements de l'enquête TeO 2, sont entourés sur le schéma.

Dans les parties suivantes, nous revenons plus en détails sur chacune de ces étapes.

3. Redressements de l'enquête complémentaire

3.1. CNR ACC : Correction de l'absence de transmission des coordonnées des éligibles par les parents

Au cours du questionnaire de l'enquête principale, les enquêtés dont les enfants sont considérés comme éligibles sont invités à communiquer leurs coordonnées. On notera que, dans le contexte de faible échantillon de G3 éligibles, l'interrogation de fratries a été autorisée.

Après identification des enfants éligibles à l'aide des réponses déjà apportées au questionnaire, une première question portant sur l'ensemble de ses enfants éligibles permet au parent enquêté de donner ou non son accord de principe sur la communication de leurs coordonnées.

Dans le cas d'une réponse positive à cette question générale, l'enquêteur demande, séparément pour chaque enfant, des informations complémentaires sur le second parent (afin de vérifier que celui-ci n'est pas lui-même immigré) puis les coordonnées des enfants dont le second parent n'est pas immigré lui-même.

Un même parent peut donc adopter des comportements de transmission des coordonnées différents pour chacun de ses enfants. C'est le cas pour 5 % des parents interrogés ayant plus d'un enfant éligible.

Il apparaît donc nécessaire, pour rendre compte de façon plus fine du comportement de transmission des coordonnées, de construire un modèle de correction de la non-réponse pour lequel l'unité d'intérêt ne serait ni un parent G2 ni un enfant G3, mais plutôt – de façon un peu formelle – un couple constitué d'un parent G2 et de l'un de ses enfants G3[2]. On compte ainsi 980 couples, qui

2 De sorte qu'un même G2 peut être associé à différents couples.

seront désormais désignés par la notation G2*G3. On notera qu'une faible proportion des questionnaires de l'enquête principale ayant permis l'identification de G3 éligibles ont été apurés, notamment faute de qualité suffisante. Ainsi, pour 11 éligibles dont un répondant à l'enquête complémentaire, il n'est pas possible de constituer un couple avec le G2 correspondant. Ces cas sont supprimés des bases éligibles et répondants à l'enquête complémentaire avant tout redressement.

Le poids d'un couple G2*G3 s'obtient par simple transfert du poids du parent G2, puisque la sélection des enfants G3 à partir d'un parent G2 ne résulte pas d'un tirage aléatoire mais se fait de façon purement déterministe[3]. Il avait été envisagé dans un premier temps de réaliser un échantillonnage parmi les individus identifiés comme « éligibles » mais leur faible nombre, comparativement à la cible initiale de 500 répondants, a obligé à mettre en collecte la totalité des éligibles dont les coordonnées ont été recueillies.

Dans ce contexte, les couples correspondant à un refus de transmission de coordonnées transfèrent leurs poids vers les couples correspondant à un accord.

À cette étape, un premier modèle de correction de la non-réponse est appliqué à l'ensemble des 980 couples G2*G3, et vise à transférer les poids des couples G2*G3 pour lesquels le parent G2 refuse la transmission de coordonnées sur les couples pour lesquels les coordonnées sont effectivement obtenues.

3.2. PPSI : Prise en compte du nombre de lien entre les éligibles et la base de sondage de l'enquête complémentaire

i. Pour les G3 dont un seul parent est interrogé dans l'enquête principale

Pour les individus G3 identifiés une seule fois parmi les couples G2*G3, on est en mesure de déterminer si le second parent aurait pu être interrogé dans l'enquête principale, et si cette interrogation aurait permis d'identifier l'individu G3 comme éligible. Il s'agit donc de déterminer si le second parent respecte les conditions de champ de l'enquête (en vie, âgé de 18 à 59 ans, résidant en France métropolitaine), et si ce parent est lui-même descendant d'immigrés de deuxième génération d'origine non-européenne.

Le second parent de chacun des G3 éligibles appartient à l'une des catégories suivantes : il peut s'agir soit du conjoint actuel du parent G2 répondant à l'enquête principale ; soit du premier conjoint cohabitant du parent G2 répondant ; soit d'un autre conjoint.

- Dans les deux premiers cas, les réponses du parent interrogé lors de l'enquête principale permettent d'identifier si celui-ci respecte les conditions de champ de l'enquête, et de connaître le lieu et la nationalité de naissance des deux parents du second parent ;
- Dans le cas où ce dernier est un « autre conjoint » du parent enquêté, les réponses permettent uniquement de déterminer que le second parent n'est pas lui-même immigré. Pour ces cas pour lesquels nous ne disposons pas d'une information complète, qui représentent 8 % de l'ensemble des couples G2*G3, une imputation est réalisée. Pour déterminer si ces « autres conjoints » auraient permis d'identifier leurs enfants comme des G3 éligibles, nous retenons une loi de Bernoulli de même probabilité que pour les « premiers conjoints ».

À ce stade, on a affecté aux couples G2*G3 décrits plus haut des poids corrigés du refus de transmission de coordonnées. Ce que l'on souhaite désormais, c'est construire des poids pour les unités de la population d'intérêt, i.e. les individus G3. Une opération de partage des poids est ainsi nécessaire.

3 De fait, le volet complémentaire vise à interroger de façon exhaustive tous les enfants éligibles de ces parents G2

En reprenant les notations utilisées par Lavallée (2009), U^A désigne le champ décrit par les couples l'ensemble des couples répondants G2*G3 lors de la collecte des coordonnées, U^B est la population d'intérêt, $i \in r^B$ est un G3 non-européen atteint par un couple G2*G3 répondant lors de la collecte des coordonnées noté $j \in U^A$. $L_{j,i}$ indique la présence d'un lien entre *j* et *i :*

$$L_{j,i} = \begin{cases} 1 \text{ si i et j ont un lien} \\ 0 \text{ sinon} \end{cases}$$

et

$$L_i^B = \sum_{j \in U^A} L_{j,i}$$

est le nombre de liens entre l'individu *i* de r^B et U^A .

D'après ce qui a été explicité en partie 1.2 sur le filtre utilisé pour l'échantillonnage des G3 de l'enquête complémentaire, le calcul du nombre de liens est immédiat :

$$L_i^B = \begin{cases} 2 \text{ si le second parent aurait pu permettre l'identification du G3 comme éligible} \\ 1 \text{ sinon} \end{cases}$$

Sous cette définition des liens, le calcul du poids de l'individu répondant i est alors donné par la formule

$$w_i^{B,ACC} = \sum_{j \in r^A} \frac{L_{j,i}}{L_i^B} w_j^{A,ACC}$$

où r^A désigne l'ensemble des couples *j* associés à un accord de transmission de coordonnées, $w_j^{A,ACC}$ le poids corrigé de la non-réponse (au sens de la non-transmission des coordonnées du G3 éligible) de l'unité répondante G2*G3 *j* et $w_i^{B,ACC}$ le poids de l'individu *i* déduit des liens et des poids $w_j^{A,ACC}$ par partage des poids.

ii. Pour les G3 dont les deux parents sont interrogés dans l'enquête principale

Dans certains cas, un même G3 est identifié comme éligible par le biais de ses deux parents, tous deux interrogés dans le cadre de l'enquête principale. Ces cas de double identification d'un G3 s'expliquent par les contraintes d'échantillonnage de l'enquête principale, qui conduisent à une forte probabilité d'interroger les deux membres d'un couple cohabitant pour certaines régions de résidence et origines.

Ce cas concerne 11 individus, soit 22 doublons G2*G3. Dans la quasi-totalité de ces cas (10 individus sur les 11 concernés), un seul parent transmet effectivement les coordonnées de l'enfant.

Soit $w_{1,i}^{A,ACC}$ et $w_{2,i}^{A,ACC}$ les poids corrigés de la non-réponse (au sens de la non-transmission des coordonnées du G3 éligible) du premier et du second couple G2*G3 obtenus pour un même individu G3, et $w_i^{B,ACC}$ le poids de cet individu *i* déduit des liens et des poids précédents. Le partage des poids devrait distinguer deux cas.

Ainsi, on devrait avoir :

$$w_i^{B,ACC} = \frac{w_{1,i}^{A,ACC}}{2}$$

lorsque seul le premier parent transmet effectivement les coordonnées de l'individu *i*. Cet individu a deux liens avec la base de sondage puisque, par définition, son second parent a effectivement permis son identification comme éligible.

Et :

$$w_i^{B,ACC} = \frac{w_{1,i}^{A,ACC} + w_{2,i}^{A,ACC}}{2}$$

lorsque les deux parents de l'individu *i* transmettent effectivement ses coordonnées, ce qui permet de conserver une information sur le poids des deux parents.

Compte tenu du faible effectif concerné, l'usage d'une formule unique a été privilégié. Pour ces 11 cas, le partage des poids consiste donc à effectuer la moyenne des poids après correction de la non-réponse des deux couples G2*G3 correspondant au même individu G3.

3.3. CNR G3 : Correction de la non-réponse des G3 échantillonnés

À l'issue de l'étape précédente, on transfère simplement le poids des couples G2*G3 vers les G3 correspondants, échantillonnés en vue de l'enquête complémentaire. Ainsi, à cette étape chaque individu G3 est identifié de manière unique, avec un poids prenant en compte les caractéristiques connues de ses deux parents.

Un second modèle de correction de la non-réponse est appliqué afin de corriger le biais de sélection induit par la non-réponse à l'enquête pour les 360 G3 échantillonnés. Il s'agit de transférer le poids des G3 échantillonnés non-répondants à l'enquête complémentaire sur les 240 répondants à cette même enquête.

Même si l'on suspecte des effets différenciés des facteurs explicatifs de la non-réponse suivant l'origine des G3, la faible taille de l'échantillon ne permet pas, comme dans l'enquête principale, de construire des modèles spécifiques par origines ou groupes d'origines. En effet, cela aurait nécessairement conduit à la création de strates de tailles trop petites, affaiblissant la puissance statistique des modèles utilisés.

Une autre approche aurait consisté à n'utiliser qu'un seul modèle commun à tous les individus de l'échantillon, et à croiser certaines des variables explicatives de la non-réponse avec des indicatrices d'appartenance aux différents groupes d'origines. Cela aurait impliqué un nombre élevé de croisements et, là encore, le faible effectif de certaines strates aurait nuit à la robustesse statistique du modèle.

Finalement, l'origine des individus a été introduite comme une variable à part entière dans l'unique modèle de correction de la non-réponse appliqué à l'ensemble des éligibles. On notera que l'application de modèles de CNR spécifiques aux origines des individus, justifiée pour l'enquête principale par l'objectif d'adhérer aux enjeux de diffusion, est moins nécessaire pour l'enquête complémentaire en raison de la restriction de son champ aux seuls G3 non-européens.

3.4. PPLM : Mise en cohérence des champs des deux enquêtes

Au total, nous disposons donc de deux échantillons de G3 non-européens : les 240 répondants à l'enquête complémentaire, et les 134 répondants à l'enquête principale qui remplissent d'après leurs propres réponses les conditions d'identification des G3 non-européens. Une dernière étape de partage des poids permet de mettre en cohérence les poids des individus de ces deux enquêtes.

Par définition, l'ensemble des G3 de l'enquête complémentaire auraient pu être interrogés dans le cadre de l'enquête principale. Ainsi, chacun d'eux respecte nécessairement les conditions de champ de l'enquête principale, qui sont un sous-ensemble des conditions de champ de l'enquête complémentaire.

En reprenant les notations précédentes, on a donc deux liens entre les répondants à l'enquête complémentaire et les champs décrits par ces deux échantillons. Pour ces individus, l'opération de partage des poids consiste donc à diviser par deux chacun de leurs poids.

4. Redressements de l'enquête principale et concaténation des observations

4.1. PPLM : Mise en cohérence des champs des deux enquêtes

Contrairement aux répondants de l'enquête complémentaire, la totalité des répondants de l'enquête principale ne respectent pas les conditions de champ de l'enquête complémentaire.

Ainsi, comme évoqué en partie 1.3, seuls les G3 de l'enquête principale dont au moins l'un des parents aurait pu être échantillonné dans l'enquête principale et permettre leur identification comme éligible respectent les conditions de champ de l'enquête complémentaire. Les réponses de ces G3 à l'enquête principale permettent de déterminer, pour chacun d'entre eux, si au moins un de leurs parents est toujours en vie au moment de la collecte, âgé de 18 à 59 ans, résidant en France métropolitaine, et ayant au moins un parent immigré d'origine non-européenne.

Les G3 non-européens répondants à l'enquête principale ont donc deux liens avec les champs décrits par les enquêtes principale et complémentaire si au moins un de leurs parents aurait pu permettre de les identifier comme « éligibles » à une interrogation par l'enquête complémentaire, et un seul lien dans le cas contraire. Ainsi, l'opération de partage des poids consiste pour ces individus à diviser leur poids par le nombre de liens : selon le cas, soit conserver leur poids, soit le diviser par deux.

4.2. Finalisation des pondérations des deux enquêtes sur le champ des G3

À cette étape, la forte dispersion des poids, inhérente à celle des poids de tirage de l'enquête TeO 2, conduit à ce que 4 individus représentent à eux-seuls environ 16 % du poids total des G3 non-européens. Une étape de troncature des poids est donc appliquée à l'ensemble des observations afin de réduire la variance des résultats.

Compte tenu de la distribution des poids avant troncature, et afin de limiter l'importance des poids extrêmement élevés sans introduire un biais supplémentaire trop important, un seuil de troncature par le haut au 98^{e} percentile est retenu.

Enfin, aucun calage sur marge n'est possible en raison de l'absence d'une source de référence sur cette population. Ainsi, l'enquête emploi en continu (EEC) ne permet par exemple d'identifier que les G3 résidant chez au moins un de leurs parents. L'enquête principale aurait pu constituer une base de calage pour l'enquête complémentaire, mais son effectif de G3 non-européens est inférieur à celui de l'enquête complémentaire.

5. Résultats numériques

5.1. Étapes de correction de la non-réponse

i. Correction de l'absence de transmission des coordonnées des éligibles par les parents

Les variables retenues pour expliquer la non-transmission par les G2 interrogés dans l'enquête principale des coordonnées de leur enfant éligible sont les suivantes : origine, tranche d'âge du G3, lieu d'habitation du G3 par rapport au G2, nombre de pièces du logement et niveau de diplôme du G2.

Les variables sélectionnées pour ce modèle sont celles apparaissant comme significatives lors d'une régression logistique (p-value inférieure à 0,05).

Le tableau suivant présente une synthèse des résultats numériques obtenus. Le nombre de GRH est égal à 4, la taille d'un GRH est supérieure ou égale à 135 (donc plus grande que 100, qui est la taille minimale généralement recommandée) et la probabilité de réponse estimée la plus faible est 0,17

(donc assez éloignée de 0, afin d'éviter d'avoir des poids trop élevés après correction de la non-réponse).

Nombre de GRH	Taille du plus petit GRH	Probabilité de réponse la plus faible	Probabilité de réponse la plus forte
4	135	0,17	0,67

Tableau 1: CNR – Transmission des coordonnées des éligibles par leur(s) parent(s)

ii. Correction de la non-réponse des G3 échantillonnés

Les variables retenues pour expliquer la non-réponse à l'enquête complémentaire des G3 sont les suivantes : région de résidence, sexe, tranche d'âge et origine du G3 ; lieu d'habitation du G3 par rapport au G2 interrogé dans l'enquête principale, et niveau de diplôme du G2.

Les variables sélectionnées pour ce modèle sont celles apparaissant comme significatives lors d'une régression logistique (p-value inférieure à 0,05).

Le tableau suivant présente une synthèse des résultats numériques obtenus. Le nombre de GRH est égal à 4, la taille d'un GRH est supérieure ou égale à 51 (donc plus faible que 100, qui est la taille minimale généralement recommandée, ce qui est inévitable compte tenu du faible nombre de G3 échantillonnés) et la probabilité de réponse estimée la plus faible est 0,28 (donc assez éloignée de 0, afin d'éviter d'avoir des poids trop élevés après correction de la non-réponse).

Nombre de GRH	Taille du plus petit GRH	Probabilité de réponse la plus faible	Probabilité de réponse la plus forte
4	51	0,28	0,88

Tableau 2: CNR des G3 de l'enquête complémentaire

5.2. Éléments de dispersion des poids

Nous présentons ici quelques éléments de distribution des poids, en distinguant selon l'enquête d'interrogation des G3, à trois étapes du schéma de redressement décrit :

1. Les poids initiaux, correspondant aux poids $r^{G3<G3}$ des G3 répondants à l'enquête principale et aux poids $r_{3,5}^{G2}$ des parents des G2 répondants à l'enquête complémentaire à l'issue des redressements de l'enquête principale non spécifiques aux G3 ;
2. Les poids avant troncature ;
3. Les poids r^{G3} finaux.

Enquête	Poids minimal	Premier quartile	Poids médian	Troisième quartile	Poids maximal	Poids moyen	Écart-type
TeO 2	34	251	506	1 718	26 418	2 702	4 838
TeO 2 - G3	26	185	298	508	18 529	539	1 347
Ensemble	26	197	346	707	26 418	1 314	3 245

Tableau 3 : Distribution des poids initiaux

Enquête	Poids minimal	Premier quartile	Poids médian	Troisième quartile	Poids maximal	Poids moyen	Écart-type
TeO 2	17	186	323	1 718	26 418	1 905	3 856
TeO 2 - G3	35	203	475	934	15 697	933	1 750
Ensemble	17	196	431	955	26 418	1281	2 735

Tableau 4 : Distribution des poids avant troncature

Enquête	Poids minimal	Premier quartile	Poids médian	Troisième quartile	Poids maximal	Poids moyen	Écart-type
TeO 2	17	186	323	1 718	8 871	1 625	2 552
TeO 2 - G3	35	203	475	934	8 871	876	1 332
Ensemble	17	196	431	955	8 871	1 144	1 895

Tableau 5 : Distribution des poids finaux

Les poids initiaux présentent des différences significatives entre les deux enquêtes. Ainsi, l'écart de la moyenne des poids initiaux entre les G3 non-européens de l'enquête principale et ceux de l'enquête complémentaire est de 5 avant tout redressement spécifique aux G3. Les étapes de correction de la non-réponse et de partage des poids permettent de diminuer cet écart moyen, qui est inférieur à 2 pour les poids finaux.

La troncature permet de diviser par près de 3 le poids maximal des G3. Cette dernière étape des redressements, qui ne concerne que 7 individus, affecte le poids de certains G3 des deux enquêtes.

Enfin, les poids finaux obtenus sur le champ des G3 présentent une distribution relativement importante, avec notamment la subsistance de quelques poids élevés. Si une telle amplitude des poids n'est pas commune dans les enquêtes réalisées auprès des ménages, elle est comparable à celle des poids de l'enquête TeO 2 sur l'ensemble de ses observations. Elle ne résulte pas des opérations de redressement, mais plutôt des poids de tirage eux-mêmes qui s'expriment comme les produits de différents facteurs dont les dispersions se cumulent.

Bibliographie

[1] 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 n°F1304, Insee.

[2] Lavallée, P. (2009). « Indirect sampling », Springer Science & Business Media.

[3] Guin, O., Tanneau, P., Thao Khamsing, W. (2022). « Redressements de l'enquête Trajectoires et Origines 2 », Journées de méthodologie statistique de l'Insee.

[4] Merly-Alpa, T., Paliod, N., Thao Khamsing, W. (2022). « Échantillonnage de l'enquête Trajectoires et Origines 2 », Journées de méthodologie statistique de l'Insee.

[5] Thao Khamsing, W., Guin, O., Merly-Alpa, T., Paliod, N. (2022), « Enquête Trajectoires et Origines 2 – de la conception à la réalisation », Document de Travail n°F2022-XX, Insee. (à paraître)