

Enquête Trajectoires et Origines 2

De la conception à la réalisation

Documents de travail

N° 2022-02 – Juillet 2022





Institut national de la statistique et des études économiques

F2022/02

Enquête Trajectoires et Origines 2

De la conception à la réalisation

Willy THAO KHAMSING¹ Olivier GUIN² Thomas MERLY-ALPA³ Nicolas PALIOD⁴

Juillet 2022

Direction des Statistiques Démographiques et Sociale - Timbre F001
88, avenue Verdier - CS 70058 - 92541 MONTRouGE CEDEX - France
Tél. : 33 (1) 87 69 62 82 - E-mail : dg75-f001@insee.fr - Site Web Insee : <http://www.insee.fr>

Ces documents de travail ne reflètent pas la position de l'Insee et n'engagent que leurs auteurs.

Working papers do not reflect the position of INSEE but only their author's views.

Remerciements : les auteurs remercient toute l'équipe projet TeO2 pour son investissement tout au long de cette enquête et tout particulièrement Claudie Becquet. Ils remercient également Jérôme Lê, Sylvie Le Mînez et Christel Colin, pour leurs commentaires et suggestions, qui ont permis d'aboutir à ce document de travail. Ce travail a également bénéficié des contributions d'Odile Rouhban et Pierre Tanneau.

¹Insee, Cellule Statistiques et études sur l'immigration, willy.thao-khamsing@insee.fr

²Insee, Divison Sondages, olivier.guin@insee.fr

³Ined, Service des enquêtes et sondages, thomas.merly-alpa@ined.fr

⁴Insee, Divison Sondages, nicolas.paliod@insee.fr

Enquête Trajectoires et Origines 2

De la conception à la réalisation

Résumé

L'enquête « Trajectoires et Origines 2 » 2019/2020 est une enquête coréalisée par l'Insee, l'institut national de la statistique et des études économiques et l'Ined, l'institut national des études démographiques. Elle étudie la diversité des populations en France et la situation des populations d'origine immigrée. Ce document décrit de façon détaillée l'élaboration et la méthodologie de réalisation de la 2^{ème} édition de cette enquête, qui fait suite à la 1^{ère} édition qui avait eu lieu en 2008-2009.

Ce document est structuré en 4 parties. La première partie présente le contexte et la mise en place de cette enquête. La 2^{ème} partie présente dans le détail la méthode d'échantillonnage et le plan de sondage. Le déroulement et le bilan de l'enquête sont décrits dans la 3^{ème} partie. Le redressement fait l'objet de la 4^{ème} partie. Enfin, des annexes complètent ce document.

Mots-clés : enquête démographique et sociale, immigrés, protocole d'enquête, échantillonnage, plans de sondage complexes, sondage indirect, correction de la non réponse, méthode généralisée du partage des poids, liens multiples, liens pondérés, calage sur marge

Abstract

The *Trajectoires et Origines 2* (TeO2) survey is coproduced by Insee, the French National Statistical Institute and Ined, the French Institute for Demographic Studies. This survey aims at studying the diversity of populations in France and the situation of immigrant populations. This document describes in detail the development and methodology for carrying out the 2nd edition of this survey, which follows on from the 1st edition of the survey which took place in 2008-2009.

This document is structured in 4 parts. The first part presents the context and the implementation of this survey. The 2nd part presents in detail the sampling method and the sampling design. The progress and outcome of the survey are described in part 3. The adjustments of this survey is the subject of the 4th part. Finally, appendices complete this document.

Keywords : demographic and social survey, immigrants, data collection protocol, sampling, indirect sampling, complex sample design, non-response adjustment, generalized weight share method, multiple links, weighted links, marginal calibration

TABLE DES MATIERES

1. Contexte et mise en place de l'enquête TeO2.....	7
1.1 Genèse de l'enquête TeO2	7
1.1.1 Une première édition de l'enquête TeO largement exploitée	7
1.1.2 Une deuxième édition de l'enquête TeO très attendue	7
1.2 Objectifs de l'enquête.....	8
1.2.1 Objectifs principaux : un principe de continuité.....	8
1.2.2 Les évolutions concernant l'échantillon	9
1.2.3 Les principales évolutions concernant les thématiques de l'enquête	11
1.3 Organisation de l'enquête.....	11
1.3.1 Les acteurs et la comitologie.....	11
1.3.2 Le calendrier.....	13
1.4 Mise en place du protocole de l'enquête.....	13
1.4.1 Le suivi des personnes qui ont déménagé	13
1.4.2 L'interrogation des personnes « non francophones »	14
1.4.3 L'interrogation des personnes de la « 3 ^{ème} génération » d'origines non européennes.....	15
2. Plan de sondage et échantillonnage	16
2.1 La méthode de constitution de l'échantillon de TeO2	16
2.1.1 Champ de l'enquête et unités enquêtées.....	16
2.1.2 La méthode d'échantillonnage de TeO2.....	17
2.2 Les 1^{ères} générations et la population générale	23
2.2.1 Le choix des groupes d'origines et allocations des G1	23
2.2.2 La construction de la base de sondage des immigrés, des Domiens et de la population générale	24
2.2.3 Le tirage des G1 et de la population générale	25
2.2.4 Echantillon complémentaire et taille finale de l'échantillon de la vague 1	26
2.3 Les 2^{èmes} générations	27
2.3.1 Constitution de la base de sondage des descendants	28
2.3.2 Le tirage des relevés mairies	29
2.3.3 Bilan des relevés mairie	31
2.3.4 Allocations	32
2.3.5 Détermination du groupe d'origines des G2	33
2.3.6 Structure de la base de sondage des G2	35
2.3.7 Constitution des strates de tirage.....	36
2.3.8 Tirage pour la collecte des G2 et pondération	37
2.4 Les 3^{èmes} générations non européennes : sondage indirect, collecte par lots	40
3. Déroulement et bilan de la collecte de TeO2.....	42

3.1 Vague 1 de TeO2 : immigrés, Domiens et population générale	42
3.1.1 Bilan de collecte de la vague 1 : juillet-décembre 2019.....	42
3.1.2 Bilan de l'enquête de rattrapage des non francophones : janvier-août 2020	44
3.2 Vague 2 TeO2 : descendants d'immigrés et de domiens	45
3.2.1 Bilan de l'opération mairie : juin - août 2019.....	45
3.2.2 Le déroulé de la collecte de la vague 2 : janvier-novembre 2020	46
3.2.3 Bilan de la participation à la vague 2	48
3.3 L'enquête expérimentale TeO2-G3	52
3.3.1 Déroulement de l'opération TeO2-G3 : mars 2020 - janvier 2021	52
3.3.2 Bilan de la participation à l'enquête TeO2-G3.....	53
3.4 Retour sur certaines spécificités de la collecte de l'enquête TeO2	54
3.4.1 Des temps de passation du questionnaire variables	54
3.4.2 Bilan sur le suivi des personnes ayant déménagé.....	55
3.4.3 Répartition géographique de l'échantillon et des répondants	57
4. Pondération et redressement de l'échantillon TeO2	62
4.1 Définition de la non-réponse totale	62
4.2 Chaîne de redressements de TeO2 (hors calage)	62
4.2.1 Chaîne de redressements des immigrés et des Domiens (G1).....	63
4.2.2 Chaîne de redressements des descendants d'immigrés et de Domiens.....	65
4.2.3 Chaîne de redressements des « autres individus ».....	66
4.2.4 Chaîne de redressements globale de l'enquête TeO2 (hors calage)	67
4.3 Correction de la non-réponse de TeO2	69
4.3.1 Correction de la non-réponse de TeO2 – vague 1	69
4.3.2 Correction de la non-réponse de TeO2 – vague 2	72
4.4 Partage des poids de TeO2	77
4.4.1 Enjeux du partage des poids de TeO2	77
4.4.2 Partage des poids de TeO2 – G1.....	77
4.4.3 Partage des poids de TeO2 – G2.....	78
4.4.4 Partage des poids de TeO2 – G0.....	78
4.5 Calage sur marge de l'enquête TeO2	78
4.5.1 Principe du calage et application à l'enquête TeO2	78
4.5.2 Calage sur marges des immigrés	80
4.5.3 Calage sur marge des Domiens.....	83
4.5.4 Calage sur marge des « autres ».....	85
Annexes.....	87
Annexe 1 – Calendrier détaillé des tests et collecte	87
Annexe 2 – Simulation de l'impact de l'intégration de nouveaux groupes d'origines sur la précision globale	93
Annexe 3 – Méthodologie échantillon complémentaire vague 1	94
Annexe 4 – Schéma du processus d'échantillonnage des G2	95

Annexe 5 – Estimation de l’allocation par cible de répondants et origines géographiques	96
Annexe 6 – Estimation du nombre d’individus de 3ème génération maghrébine et non européenne.....	97
Annexe 7 – Modèles de correction de la non-réponse – vague 1	100
Annexe 8 – Modèles de correction de la non-réponse – vague 2	104
Annexe 9 – Partage des poids	108
Annexe 10 – Vérifications sur les pondérations.....	110

1. Contexte et mise en place de l'enquête TeO2

1.1 Genèse de l'enquête TeO2

1.1.1 Une première édition de l'enquête TeO largement exploitée

L'enquête « Trajectoires et Origines 2 » (TeO2, 2019-2020) est une réédition de l'enquête « Trajectoires et Origines » (TeO, 2008-2009). Cette enquête est réalisée dans le but de combler un déficit de données sur les immigrés et les enfants nés en France de parent(s) immigré(s), dans un contexte où la diversité croissante de la société française suscite de nombreux débats.

La première édition de l'enquête a donné lieu à une production scientifique importante et sert de référence pour les chercheurs et administrations travaillant sur les questions de la place de l'origine dans les processus d'intégration, de discrimination et de construction identitaire au sein de la société française.

Elle a fait l'objet :

- d'un nombre important de publications et textes scientifiques. Près de 150 ont été recensés fin 2015, dont 52 articles de revues scientifiques, 3 ouvrages et 11 chapitres d'ouvrage, 22 publications institutionnelles (telles que *Insee Première*, *Population & Sociétés*, *Infos migrations*, *Dares Analyses*, etc.), 49 documents de travail et 8 thèses,
- de plus de 220 demandes d'utilisation effectuées et validées, sur une grande variété de sujets, dont certaines émanant de chercheurs d'autres pays européens ou nord-américains depuis la mise à disposition des données de l'enquête sur le réseau Quêtelet et auprès du Centre d'Accès Sécurisé aux Données (CASD),
- de la publication fin 2015, aux éditions de l'Ined, d'un ouvrage collectif qui condense les résultats les plus importants de l'enquête et a eu un retentissement important dans les médias. On a ainsi pu recenser 25 articles de presse autour des résultats de TeO au moment de la sortie de l'ouvrage entre janvier et mi-février 2016⁵.

Elle a également été fortement mobilisée dans l'ouvrage *Insee Références « Immigrés et descendants d'immigrés »* paru en octobre 2012, que ce soit dans les dossiers ou les fiches de l'ouvrage.

1.1.2 Une deuxième édition de l'enquête TeO très attendue

L'enquête TeO (2008-2009) a servi de référence pour les travaux sur les immigrés et leurs enfants dans la recherche et dans la statistique publique. Dix ans après la première enquête, les thèmes de l'immigration et de la diversité de la population française sont toujours au cœur du débat public en France. Dans un contexte social qui a évolué depuis dix ans, la réédition de l'enquête était donc attendue par les pouvoirs publics, la société civile et les scientifiques. D'où la nécessité de mettre à jour les résultats de la première enquête et d'alimenter en données objectives le débat public sur ces thématiques afin d'en mesurer et comprendre les évolutions. En effet sans la réédition de l'enquête, il aurait existé un décalage important

⁵ Beauchemin C., Hamel C., Simon P. (éd.), 2015, *Trajectoires et origines, enquête sur la diversité des populations en France*, Paris, Ined (collection Grandes Enquêtes). Cf. par ailleurs le site consacré par l'Ined à l'enquête (<https://teo.site.ined.fr/>) et l'onglet « Résultats/Publi » dans lequel toutes les publications mentionnées sont référencées.

entre la place grandissante de ces questions dans le débat public et le niveau des informations statistiques disponibles.

Le souhait d'un renouvellement de l'enquête a été exprimé à plusieurs reprises :

- En 2010, le rapport du Comité pour la mesure de la diversité et l'évaluation des discriminations (COMEDD) recommandait la réalisation d'une enquête, renouvelée périodiquement, dédiée à l'étude des discriminations et des inégalités.
- Le souhait de renouveler l'enquête TeO a été exprimé lors de la réunion de la commission « démographie et questions sociales » du CNIS d'octobre 2015, ainsi que par le Défenseur des droits lors de la conférence de presse de présentation des résultats de l'enquête de 2008-2009 le 8 janvier 2016 à l'Ined.
- L'enquête TeO2 a, par ailleurs, été inscrite dans le Plan national de lutte contre le racisme et l'antisémitisme (2018-2020), élaboré par la DILCRAH (Délégation Interministérielle à la Lutte contre le Racisme, l'Antisémitisme et la Haine anti-LGBT).
- Un indicateur du fort intérêt pour la première édition de l'enquête a été le nombre de demandes d'utilisation des données effectuées auprès du réseau Quêtelet qui a dépassé les 200. Comme pour la première édition, il est probable que TeO2 sera largement utilisée par la recherche universitaire et les services statistiques. Parmi les services statistiques ministériels (SSM), le DSED (ministère de l'Intérieur), la DARES (ministère du Travail), la DREES (ministère de la Santé), l'ONZUS (devenu ONPV au sein de l'Agence nationale de la cohésion des territoires) ont été les principaux utilisateurs de TeO et exploiteront certainement la deuxième édition. En outre, au-delà de l'Ined et de l'Insee, d'autres utilisateurs attendus sont le Défenseur des droits, l'Institut national de la jeunesse et de l'éducation populaire (INJEP), l'Institut d'aménagement et d'urbanisme Ile-de-France, Santé Publique France (fusion de l'InVs, l'Inpes, Eprus et Adalis), le Département des études de la prospective et des statistiques (DEPS) du ministère de la Culture, la Direction générale de la cohésion sociale (DGCS).

1.2 Objectifs de l'enquête

1.2.1 Objectifs principaux : un principe de continuité

L'enquête TeO2 reprend les principaux objectifs de la première édition réalisée en 2008-2009. Elle s'intéresse à l'articulation entre l'origine et les autres catégories de distinction dans la société (le genre, la classe sociale, l'âge, le quartier, la religion, etc.) afin d'analyser les processus d'intégration, de discrimination et de construction identitaire concernant l'ensemble de la population. L'étude des évolutions intergénérationnelles de ces processus est au cœur des objectifs de l'enquête. Un des enjeux essentiels est de mettre à jour les données produites par l'enquête de 2008-2009, avec un recul d'une dizaine d'années. Le principe de continuité a donc prévalu dans les choix de conception de l'enquête, qu'il s'agisse de l'échantillon ou du questionnaire. Un certain nombre d'innovations ont cependant été introduites.

1.2.2 Les évolutions concernant l'échantillon

Les populations d'intérêt de l'enquête sont des personnes résidant en France métropolitaine⁶ : les immigrés (personnes dites de la « première génération » à vivre en France) et les personnes nées en France dont au moins un parent est immigré (personnes dites de la « deuxième génération » à vivre en France) ; les personnes nées dans un département d'Outre-mer et les personnes nées en France métropolitaine dont au moins un parent est né dans un département d'Outre-mer ; les personnes n'appartenant à aucun des groupes ci-dessus (« population ni immigrée ni descendante d'immigrée, ni originaire des Dom »). En conséquence, comme pour la 1^{ère} édition de l'enquête, l'échantillon est constitué de plusieurs sous-échantillons : un échantillon de la population générale (qui comprendra donc des personnes appartenant à toutes les populations d'intérêt de l'enquête), un échantillon d'immigrés, un échantillon de descendants d'immigrés, un échantillon de personnes nées dans un Dom et un échantillon de descendants de Domiens. Ces quatre derniers échantillons visent à assurer une représentation minimale des groupes dont on pense qu'ils peuvent être soumis à des discriminations du fait de leur origine, quelle que soit leur nationalité. À la différence de la 1^{ère} édition cependant, l'enquête TeO2 s'intéresse également spécifiquement à des personnes qui ont soit une ascendance migratoire au-delà de la 3^{ème} génération, soit ont au moins un parent né Français à l'étranger (y compris dans une ancienne colonie).

1.2.2.1 Identification et interrogation des personnes de la « 3^{ème} génération »

Depuis 2008, davantage d'enfants des descendants d'immigrés sont devenus adultes et la question de la persistance de l'influence des origines sur leurs trajectoires sociales se pose. Un des objectifs de l'enquête TeO2 a été d'identifier les personnes de la « 3^{ème} génération », c'est-à-dire les personnes ayant au moins un grand parent immigré, sans être elles-mêmes ni immigrées, ni descendantes d'immigrés. 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. Ce questionnement s'appuie sur les résultats de TeO1. L'enquête TeO1 qui a été réalisée auprès des 1^{ères} et 2^{èmes} générations a montré que la situation socio-économique et/ou le ressenti des discriminations des enfants nés en France d'au moins un parent d'origine maghrébine, africaine subsaharienne, asiatique ou turque s'amélioraient peu ou se dégradaient par rapport aux immigrés des mêmes origines. Il s'agit également d'éclairer les situations de discriminations ou de racisme perçues par une partie de la population majoritaire dans la première enquête, qui pourraient être expliquées par la présence, au sein de cette population, de personnes d'origines migratoires plus anciennes.

Pour cela la stratégie déployée pour atteindre ces populations s'est effectuée à deux niveaux :

1. le questionnaire de l'enquête TeO2, permet d'identifier les enquêtés dits de la « 3^{ème} génération » grâce à l'introduction de questions sur la nationalité et le pays de naissance des 4 grands-parents. Ces questions sont nouvelles, elles n'existaient pas dans la 1^{ère} édition de l'enquête.
2. l'identification ci-dessus devrait conduire, dans l'échantillon de la population générale, à repérer principalement des personnes de 3^{ème} génération d'origine européenne (compte tenu des flux

⁶ Dans les Dom, les enquêtes « Migrations, famille et vieillissement » ont pour objectif l'analyse et la mesure des mutations sociodémographiques à l'œuvre dans chacun de ces départements : importance et caractéristiques des migrations, évolution des structures familiales, changement dans les comportements de fécondité, place et perspectives de la jeunesse, importance de la précarité sociale, évolution des formes traditionnelles de solidarités en lien avec le vieillissement de la population. Elles abordent aussi les questions de discriminations vécues par les natifs des Dom quand ils les ont quittés ou par les non-natifs s'étant installés dans les Dom.

migratoires passés et de la taille de l'échantillon de la population générale). Pour mieux connaître ceux d'origine non européenne, une enquête spécifique, complémentaire à l'enquête principale, a été menée, auprès d'un échantillon de personnes de 3^{ème} génération d'origine non européenne, pour atteindre des effectifs suffisants. La méthodologie mise en œuvre pour ce faire est décrite dans la partie 2 du document.

Encadré 1 : Le contexte institutionnel de l'enquête complémentaire auprès des petits-enfants d'immigrés (TeO2 - G3).

En septembre 2016, la maîtrise d'ouvrage de TeO2 s'est interrogée sur l'opportunité d'introduire dans le questionnaire des questions visant à identifier les personnes de la « 3^{ème} génération ». À cet égard, un débat a eu lieu au CNIS (séance du 29 novembre 2016). L'avis d'opportunité du CNIS a demandé de faire preuve de prudence concernant le « repérage » de la 3^{ème} génération et d'expliquer les finalités d'un tel repérage. Le bureau du Cnis, le 27 juin 2018, s'est prononcé en faveur d'un sur-échantillonnage de personnes de 3^{ème} génération non européenne. Conformément à ces décisions, les concepteurs de l'enquête ont explicité devant les membres du Comité du Label les finalités d'un questionnement des personnes de 3^{ème} génération et ont mis au point et testé à la fois le repérage des 3^{èmes} générations dans l'enquête principale et le protocole d'une enquête spécifique auprès des 3G d'origine non européenne. L'enquête principale a reçu l'avis de conformité du Comité du Label, le label d'intérêt général et de qualité statistique pour les années 2019 et 2020. L'enquête complémentaire à visée expérimentale auprès des petits-enfants d'immigrés, a reçu un avis d'examen favorable délivré par le Comité du Label et s'est également vue conférer le caractère obligatoire.

1.2.2.2 Interrogation des enfants nés en France de parent(s) « Français de l'étranger »

Dans TeO1, les personnes nées en France de parent(s) né(s) de nationalité française à l'étranger (ex : rapatriés d'Algérie ou d'autres territoires coloniaux, expatriés etc.), sans qu'aucun ne soit immigré, étaient considérés comme hors champ. Elles n'ont donc répondu qu'à un questionnaire tronqué. Ainsi, pour environ 3 000 fiches-adresses (FA), l'entretien débutait mais était interrompu (questionnaire court) dès que l'enquête répondait que son ou ses parents né(s) à l'étranger étaient nés Français⁷. Dans l'enquête TeO2, il a été décidé de poursuivre le questionnaire complet. Les trajectoires des personnes nées en France de parent(s) rapatrié(s) ou expatrié(s) constituent un domaine de recherche encore peu étudié. Un des enjeux est de pouvoir comparer les trajectoires des personnes nées en France de parent(s) « Français de l'étranger » à celles des personnes nées en France de parent(s) immigré(s) dont les parents sont nés dans le même pays.

Une des difficultés consiste à identifier ces individus à travers le questionnaire des enquêtes.

Une des principales sources mobilisées a été l'Échantillon démographique permanent (EDP)⁸. Toutefois, le nombre d'enfants nés en France de parent(s) rapatrié(s) ou expatriés est faible dans l'EDP. TeO2 présente ainsi un avantage comparatif en termes d'effectif de cette population et de richesse du questionnaire.

Récemment, le recensement de la population (données compilées et harmonisées) et certaines éditions de l'enquête emploi en continu (données empilées) ont été mobilisés pour étudier cette population⁹. Ces sources présentent des effectifs supérieurs à l'EDP et à l'enquête TeO2. L'apport de cette dernière est surtout de mieux renseigner le parcours migratoire de ces personnes.

⁷ Le protocole de l'enquête s'arrête à la fin du module 3 du questionnaire (nationalité et origine des parents). L'enquêté a tout de même répondu au module 1 (Tableau des habitants du logement) et au module 2 (Revenus).

⁸ Couto MP., « Repenser l'intégration socioéconomique des pieds-noirs en France à l'aide des données de l'Échantillon Démographique Permanent de l'INSEE », document de travail CREST n°2014-49, 2014

⁹ Kerjosse R., Lê J., « Les personnes nées françaises à l'étranger ont un profil plus proche des natifs que des immigrés », Insee Première n°1829, Décembre 2020

1.2.2.3 De nouveaux groupes d'origines : Chinois, réfugiés

Afin de tenir compte de l'évolution de la structure par origine de la population et d'apporter des connaissances nouvelles sur des groupes méconnus, il a été décidé d'introduire deux nouveaux groupes à surreprésenter dans l'échantillon des immigrés par rapport aux groupes d'origines déjà surreprésentés dans la 1^{ère} édition de l'enquête :

- un groupe « Chine », dont la population immigrée a crû rapidement au cours des dix dernières années ;
- un groupe constitué d'individus immigrés originaires de pays d'où sont originaires un grand nombre de réfugiés (pays qui comptent plus de 20% de réfugiés dans TeO1 et dont l'effectif est supérieur à 1 000 individus dans l'enquête annuelle de recensement (EAR) de 2013) : Angola, Sri Lanka, Thaïlande, Pakistan, Haïti, Russie, pays de l'ex Yougoslavie hors UE. L'objectif visé était d'augmenter les effectifs de réfugiés déjà présents dans les groupes d'origine d'intérêt (notamment les deux Congo compris dans le groupe « Afrique centrale et golfe de Guinée » ; ou le Vietnam, le Laos et le Cambodge inclus dans l'Asie du Sud-Est).

1.2.3 Les principales évolutions concernant les thématiques de l'enquête

En réponse aux objectifs de l'enquête, le questionnaire couvre un grand nombre de thèmes et comprend de nombreux éléments rétrospectifs relatifs aux trajectoires. Le recueil des éléments constitutifs de l'origine, aussi bien géographique que sociale, culturelle que résidentielle, fait l'objet d'une attention particulière. Les principaux thèmes abordés sont : l'environnement familial et social (entourage familial, histoire matrimoniale, relations sociales) ; l'accès aux différentes ressources de la vie sociale (éducation, logement, emploi, santé, vie citoyenne) ; les différentes dimensions des origines et appartenances (lien avec le pays d'origine, religion, langues, image de soi et regard des autres). Le thème des discriminations est abordé de façon transversale dans différents modules, et fait également l'objet d'un module spécifique.

Par rapport à l'enquête TeO de 2008-2009, des développements ont été introduits pour permettre de nouvelles études sur les sujets suivants :

- l'analyse de la ségrégation résidentielle et de ses effets sur les processus de discrimination scolaire et professionnelle ;
- les conséquences sur les trajectoires socio-économiques des migrants de leurs caractéristiques pré-migratoires ;
- l'impact des trajectoires légales (statut administratif) des migrants sur leurs conditions d'intégration ;
- les comportements de santé des populations immigrées et leur intégration au sein du système de soins français.

1.3 Organisation de l'enquête

1.3.1 Les acteurs et la comitologie

TeO2 est le fruit d'une co-maîtrise d'ouvrage entre l'Insee et l'Ined. Les deux parties sont intervenues conjointement à tous les stades de conception, d'exécution et d'exploitation de l'enquête. L'Insee a assuré

la maîtrise d'œuvre d'organisation statistique, méthodologique et informatique de l'enquête. L'Ined a piloté les questions d'opportunités de recherche, de la conception du questionnaire à l'exploitation de l'enquête. Une comitologie a été mise en place pour permettre la réalisation de l'enquête de sa conception jusqu'à son exploitation. Celle-ci est présentée dans l'encadré 2.

Encadré 2 : Comitologie de l'enquête TeO2 et financement

- Un comité de pilotage (COPIL) a été constitué pour prendre les décisions stratégiques. Il précise le contour du projet et s'assure de l'atteinte des objectifs en termes de réalisation, coût et délai. Il a eu pour mission de :

- fixer les calendriers et suivre le bon déroulement des différentes étapes ;
- instruire les demandes de moyens et ressources alloués au projet ;
- valider les propositions faites par l'équipe projet et décider des solutions à mettre en œuvre ;
- prendre les décisions structurantes (en termes de périmètre du projet, de dates butoirs, de moyens).

Il s'est réuni régulièrement à l'initiative des maîtrises d'ouvrage.

- Un comité de suivi a été constitué pour instruire les questions techniques et préparer les éléments pour la prise de décision par le comité de pilotage. Il a eu pour mission de :

- coordonner toutes les phases de l'enquête,
- assurer le suivi de l'exécution, notamment le travail des équipes de projet informatique et statistique,
- veiller au respect du calendrier de travail et des coûts associés,
- préparer les réunions du comité de pilotage du projet.

Le comité de suivi s'est réuni régulièrement, au moins une fois par mois, de la phase de conception à la fin de la collecte, à l'initiative du CPOS (Chef de Projet en Organisation Statistique) de l'Insee.

- Un conseil scientifique a été mis en place comme instance de consultation. Présidée par la directrice de l'Ined, il est composé des deux maîtres d'ouvrage et de quatre membres reconnus pour leur compétence scientifique sur les thématiques de l'enquête. Il peut être sollicité par l'un ou l'autre des deux maîtres d'ouvrage en cas de difficulté dans la préparation ou la conduite de l'enquête.
- Un groupe technique dédié à l'échantillonnage dont le but a été d'instruire l'avancée de la constitution de l'échantillon et faire partager les constats et les contraintes pour faciliter les arbitrages au COPIL. Ce groupe n'était pas décisionnaire et était co-animé par la division Sondages de l'Insee et le chef de projet statistique (CPS). Le secrétariat était assuré par le CPS.
- Un groupe de conception, piloté par l'Ined, a été chargé de la révision du questionnaire. Plusieurs groupes thématiques ont été mis en place pour réfléchir à l'évolution des différents modules du questionnaire.
- Le groupe d'exploitation est composé des membres du groupe de conception et d'autres chercheurs, chargés d'études et représentants de bailleurs. Il intervient à l'issue de la collecte pour aider à l'apurement des données et la constitution des fichiers d'étude. Il sera ensuite chargé de l'exploitation de l'enquête et de la production d'un ouvrage collectif. Il est animé par les responsables scientifiques de l'Ined.

Le financement

Le coût total de l'enquête est estimé à 5,7 millions d'euros dont 3,8 millions d'euros liés à la collecte. Outre l'Insee et l'Ined, les organismes ci-dessous ont participé financièrement à l'opération.

- Institut National de la jeunesse et de l'éducation populaire (INJEP)
- Caisse nationale des Allocations familiales (CNAF)
- Agence nationale de la cohésion des territoires (ANCT)
- Défenseur Des Droits (DDD)
- Département des statistiques, des études et de la documentation (DSED) de la Direction générale des étrangers en France (SSM Immigration)
- France Stratégie
- Ministère de la Culture
- Ministère du Travail / Direction de l'animation de la recherche, des études et des statistiques (DARES)
- Délégation Interministérielle à la Lutte Contre le Racisme et l'Antisémitisme et la Haine anti-LGBT (DILCRAH)
- Secrétariat d'Etat à l'égalité Femmes / Hommes
- Direction générale de la cohésion sociale (DGCS)

1.3.2 Le calendrier

Du début du projet à la mise à disposition des données sur le réseau français des centres de données pour les sciences sociales Quételet-PROGEDO Diffusion, l'enquête se déroule sur 6 ans, de septembre 2016 à la fin 2022. Les principales dates de préparation et de réalisation de l'enquête sont résumées dans l'encadré 3. Le calendrier complet est en annexe 1.

1.4 Mise en place du protocole de l'enquête

Pour l'essentiel, l'enquête a été réalisée en face-à-face, sous Capi, en une seule visite. Les enquêteurs ont bénéficié de formations, de supports de formation, de documents de collecte, de fiches de procédures, de vidéos présentant le protocole de collecte. Trois points particuliers sont cependant à souligner concernant le protocole : le suivi des enquêtés qui ont déménagé, l'interrogation des enquêtés non francophones, et l'interrogation des personnes de la 3^{ème} génération d'origine non européenne.

La collecte s'est déroulée en plusieurs vagues : une première vague auprès des immigrés, des personnes nées dans les Dom et de la population générale en 2019 ; une deuxième vague auprès des descendants d'immigrés et de Domiens en 2020 ; l'enquête expérimentale auprès des descendants d'immigrés en 2020-début 2021. La 2^{ème} vague de collecte de TeO2 s'est donc déroulée en 2020 et a donc dû être adaptée aux conséquences de l'épidémie de Covid-19. La collecte de l'enquête a été interrompue lors du 1^{er} confinement (16 mars au 11 mai), a repris courant juillet, après une expérimentation du protocole sanitaire, et son protocole a été pour partie revu à l'automne 2020, tout d'abord pour pallier le manque de fiches-adresses (FA) collectées dans certaines régions en raison du 1^{er} confinement, puis en raison du second confinement, avec une passation par téléphone. Ces aspects seront développés dans la partie 3.

1.4.1 Le suivi des personnes qui ont déménagé

TeO2 est une enquête nominative qui cible des personnes et non des ménages comme le font d'autres enquêtes de l'Insee. Lorsqu'une personne a été tirée dans la base de sondage (EAR 2018), l'enquêteur cherche à la joindre et dispose pour cela de son nom/prénom et adresse en 2018. Or, certaines populations cibles de TeO2 sont plus mobiles et déménagent plus fréquemment que les autres. Dans ce cas, les enquêteurs ont pour consigne de chercher la nouvelle adresse des personnes pour qu'elles puissent être interrogées.

Les expériences précédemment menées à l'Insee ont montré le faible gain obtenu par le recours au fichier de la Poste et les enquêtes qui s'en étaient servies ont abandonné cette option. La pratique à l'Insee était donc de ne plus utiliser ce fichier.

Pour retrouver les personnes ayant déménagé, les enquêteurs ont eu recours¹⁰ à :

1. des informations du terrain (habitants actuels du logement, voisinage, mairie, etc.) pour retrouver les adresses des personnes qui ont déménagé,
2. la consultation de l'annuaire, la recherche internet,

¹⁰ En termes de document de collecte, un « vademecum du suivi des enquêtés ayant déménagé » a été rédigé : le document présentait aux enquêteurs l'ensemble des méthodes auxquelles ils pouvaient recourir pour retrouver un enquêté qui a déménagé. Il contenait aussi des messages argumentaires types à utiliser pour les contacts pris par téléphone et courriel.

3. les coordonnées (adresse mail et téléphone) récupérées grâce à l'appariement de la base de sondage TeO2 avec les fichiers fiscaux (taxe d'habitation). Cet appariement est techniquement possible et l'utilisation des données des fichiers d'imposition des personnes est juridiquement autorisée. Cela a fait l'objet d'une déclaration CNIL générique par l'Insee couvrant l'ensemble des enquêtes.

1.4.2 L'interrogation des personnes « non francophones »

Les immigrés constituent une cible privilégiée de l'enquête TeO2 et pouvoir répondre à une enquête comme TeO2 demande une bonne maîtrise de la langue française. Il était donc essentiel de mettre en œuvre une méthodologie permettant d'enquêter dans des conditions satisfaisantes la population immigrée, quel que soit son niveau en français. Comme pour l'enquête TeO1, pour aider à la passation du questionnaire en situation de mauvaise compréhension, les enquêteurs avaient à leur disposition un document papier (cartes-codes) contenant environ 70 questions et leur modalités traduites¹¹ en 10 langues (allemand, anglais, arabe, chinois, espagnol, khmer, portugais, russe, turc, vietnamien). Il a toutefois été décidé de modifier sur certains aspects le protocole d'interrogation des personnes non francophones utilisé lors de la première édition de l'enquête car celui-ci avait été jugé insatisfaisant sur certains points¹². Après avoir envisagé différentes solutions¹³, il a été décidé le protocole suivant :

- En cas de mauvaise compréhension du français, le principe du recours à des tiers traducteurs (famille, entourage, voisins) était l'option privilégiée ;
- En l'absence d'un tiers pouvant traduire (cas d'enquêtés non francophones isolés), il était prévu de réaliser l'enquête dans la langue maternelle de l'enquêté, plutôt qu'en ayant recours à un traducteur extérieur comme cela avait été prévu dans TeO1. Cette opération a pris la forme d'une enquête de rattrapage réalisée par l'Ined. Le protocole prévu était le suivant :
 - Les enquêteurs Insee étaient pourvus d'une fiche de contact qu'ils pouvaient présenter aux enquêtés. Cette fiche de contact a été traduite en 22 langues : anglais, turc, arabe, vietnamien, portugais, espagnol, mandarin, laotien, russe, tamoul, serbe, croate, allemand, khmer, bengali, thaï, pashto, roumain, albanais, géorgien, lituanien et tchèque.
 - Les fiches-adresses concernées ont été identifiées et transmises (avec les informations complémentaires recueillies : la langue maternelle et les coordonnées de la personne) à l'Ined en fin de période de collecte de la première phase. L'Insee a également transmis préalablement des informations à l'Ined sur la localisation et les langues parlées des enquêtés afin de permettre d'anticiper le recrutement des enquêteurs en langue étrangère.
 - L'Ined devait réaliser les entretiens en face-à-face à partir de l'outil Blaise/Capi permettant la saisie des données informatiques en direct. Il a été autorisé la réalisation par téléphone des entretiens de l'enquête de rattrapage, du fait de la dispersion géographique des enquêtés non francophones par rapport au lieu de résidence des enquêteurs de l'Ined.

¹¹ Les traductions ont été réalisées par un prestataire extérieur. Pour en contrôler la qualité et corriger les erreurs, une prestation de rétrotraduction a également été réalisée par un autre prestataire.

¹² Lors de l'enquête TeO1, en cas de non maîtrise de la langue, enquêteurs et enquêtés pouvaient s'appuyer sur des documents support : la lettre-avis, la fiche de mots-clés et les cartes-codes ont été traduits en 10 langues. Si l'enquêté éprouvait des difficultés liées à la langue, il lui était proposé de recourir à un interprète. Si 1 084 entretiens ont nécessité une traduction partielle ou complète, le recours à des interprètes professionnels n'a concerné que 128 entretiens. Une des raisons était la difficulté de mettre en place des rendez-vous à 3 (enquêteur, enquêté et traducteur).

¹³ L'utilisation sous CAPI de fichiers sonores (traduction orale des questions) ou encore de traductions écrites a notamment été envisagée. Toutefois, ces options sont coûteuses et sont difficiles à valider, car les concepteurs de l'enquête ne maîtrisent pas toutes les langues étrangères. De plus, sous CAPI, l'utilisation de caractères spéciaux étrangers n'est pas toujours possible.

1.4.3 L'interrogation des personnes de la « 3^{ème} génération » d'origines non européennes

Ce volet complémentaire à visée expérimentale (TeO2 - G3) consistait à administrer le même questionnaire que celui de l'enquête TeO2 à un échantillon élargi, constitué des personnes ayant au moins un grand-parent immigré d'origine non européenne.

Les personnes de la « 3^{ème} génération » d'origines non européennes ont été identifiées via leurs parents - descendants d'immigrés (G2) - interrogés dans l'enquête TeO2. Le questionnaire comporte en effet des questions sur les enfants des enquêtés qu'ils résident ou non dans le logement de leurs parents. Les parents devaient ensuite accepter de fournir les coordonnées de leurs enfants pour une interrogation ultérieure. Les questions sur les coordonnées sont non obligatoires, le temps est laissé aux parents d'échanger avec leurs enfants.

Le protocole prévoyait un module de « prise de contact et sélection des personnes de 3^{ème} génération » à la toute fin du questionnaire pour :

1. vérifier l'éligibilité de leurs enfants
2. collecter leurs coordonnées (adresses postale et électronique, téléphone).

Les FA éligibles ont été traitées comme une enquête « clone », avec livraison de lots successifs, quatre au total, entre mars 2020 et janvier 2021. Il a été un temps envisagé une collecte « au fil de l'eau », c'est-à-dire avec la création et la mise en collecte, en temps réel, des nouvelles FA. Un inconvénient potentiel de ce choix était d'augmenter la charge de travail des enquêteurs en vague 2 sans possibilité de prévision et gestion de cette charge.

Un effort particulier a été réalisé pour former les enquêteurs à une enquête au questionnaire long, abordant de nombreux sujets, dont certains ne sont pas familiers des enquêteurs et concernent des données sensibles (questions sensibles sur la religion, les opinions politiques, les origines et le ressenti d'appartenance, les trajectoires migratoires et toutes les étapes administratives qui y sont liées). L'encadré ci-après présente la réalisation des vidéos TeO2.

Encadré 4 : Focus sur les vidéos TeO2

Pour les besoins de l'enquête, six courtes séquences vidéos ont été produites. Initialement destinées à la formation des gestionnaires d'enquête et des enquêteurs, l'équipe de conception a proposé de mettre à disposition deux d'entre elles au grand public : une sur la présentation de l'enquête et une sur le traitement des données sensibles. Ces deux vidéos sont mises en ligne sur le [site de l'Insee](#), le [site TeO2](#) et sur la [chaîne YouTube de l'Insee](#). Elles devaient permettre aux enquêtés de comprendre les objectifs de l'enquête, ses enjeux.

Le bilan de l'opération a été très positif. Ces vidéos ont été très appréciées, que ce soit par les enquêteurs ou enquêtés et jugés très utiles par les gestionnaires pour leur formation auprès des enquêteurs. Sur internet les vidéos comptent de nombreuses vues.

La liste des 6 vidéos produites entre mars et septembre 2019 figure ci-dessous :

- séquence 1 : présentation de l'enquête
- séquence 2 : nationalités
- séquence 3 : trajectoires légales des migrants
- séquence 4 : le protocole d'interrogation des non francophones
- séquence 5 : le protocole d'interrogation des 3^{èmes} générations
- séquence 6 : données sensibles

2. Plan de sondage et échantillonnage

2.1 La méthode de constitution de l'échantillon de TeO2

2.1.1 Champ de l'enquête et unités enquêtées

Le champ de l'enquête est l'ensemble des personnes vivant en logement ordinaire¹⁴ en France métropolitaine¹⁵ âgées de 18 à 59 ans au moment de l'enquête¹⁶. En effet, l'enquête portant de façon approfondie sur l'insertion sociale et professionnelle, le champ de la population enquêtée a été réduit aux personnes d'âge actif strict¹⁷ (18-59 ans), afin d'en privilégier la représentativité.

L'enquête se limite de façon pragmatique aux logements ordinaires, bien qu'il existe des collectivités spécifiques aux immigrés : en 1992, la collecte qui avait été réalisée dans les foyers pour l'enquête Mobilité Géographique et Insertion Sociale (MGIS), conduite par l'Ined avec le soutien de l'Insee, s'était avérée très difficile et la plupart des personnes résidant en collectivités n'appartiennent pas aux classes d'âge concernées par l'enquête.

L'enquête vise un échantillon de 26 500 répondants pour 48 600 fiches-adresses¹⁸ correspondant à des logements ordinaires. Ce sont cependant des individus spécifiquement identifiés qui sont recherchés au sein de ces logements.

L'échantillon de l'enquête est composé pour permettre des analyses sur **5 grands groupes de personnes** :

- les immigrés (personnes dites de « première génération ») ;
- les personnes nées en France dont au moins un parent est immigré (descendants d'immigrés, personnes pouvant aussi être dites de « deuxième génération ») ;
- les personnes nées dans un département d'Outre-mer ;
- les personnes nées en France métropolitaine dont au moins un parent est né dans un département d'Outre-mer ;
- les personnes n'appartenant à aucun des groupes ci-dessus (« population majoritaire »), parmi lesquelles certaines ont une ascendance migratoire et présentent un intérêt particulier pour TeO :
 - les personnes dont au moins un parent est né Français à l'étranger, y compris dans une colonie (dit « enfants de Français de l'étranger »),
 - les petits-enfants d'immigré(s) (dits de « troisième génération »), d'origines européennes et non européennes, pour lesquels un effort de surreprésentation doit être réalisé.

¹⁴ En 2020, selon les données de l'enquête annuelle de recensement, 1,7 % des personnes de 18 à 59 ans vivent en logement non ordinaire en France métropolitaine. Cette proportion monte à 3% parmi les immigrés. Il s'agit généralement d'immigrés arrivés récemment, avec une surreprésentation de ceux originaires d'Afrique, d'Asie et d'Europe de l'Est.

¹⁵ La question d'étendre l'enquête TeO aux Doms a été posée. Du fait de la décision de renouveler l'enquête « Migrations, Famille et Vieillesse » (MFV) dans les quatre Dom en 2020, il a été décidé de ne pas étendre TeO2 aux Doms. Pour rappel, le champ géographique de l'enquête TeO1 portait également sur la France Métropolitaine.

¹⁶ Dans TeO1, les contraintes d'échantillonnage n'avaient pas permis de dépasser l'âge de 49 ans pour les enfants d'immigré(s) (recul temporel insuffisant des fichiers d'état civil anonyme permettant d'identifier les enfants d'immigré(s)). Le champ d'âge couvert pour les immigrés (18-59 ans) était différent de celui des enfants d'immigré(s) (18-49 ans) du fait de contraintes d'échantillonnage. Ces contraintes sont moins fortes dix ans plus tard du fait de l'avancée des âges des enquêtés. Pour TeO2, il a été décidé d'étendre la tranche d'âge des enfants d'immigré(s) à 59 ans.

¹⁷ La classe « 15-64 ans » désigne habituellement dans les enquêtes et statistiques officielles, les personnes d'« âges actifs ». L'enquête TeO2 s'intéresse aux personnes de « 18-59 ans » (« âges actifs stricts »). Cela évite les effets de bord sur les études et les transitions emploi/retraite.

¹⁸ Le chiffre de 48 600 FA correspond à l'allocation initiale. Plus de 2000 FA ont été ajoutées pour combler un déficit de l'échantillon initial correspondant aux individus nés en 1979, 1980, 1989 et 1990 absents de la base de sondage. L'allocation finale de l'enquête TeO2 a été de 50 730 FA. Ce point est détaillé dans la partie 2.2.4.

Pour constituer ces groupes d'analyse, **6 sous-échantillons** ont été prévus :

- un sous-échantillon de personnes immigrées (01)
- un sous-échantillon de personnes originaires des DOM (02)
- un sous-échantillon de descendants d'immigrés (03)
- un sous-échantillon de descendants de domiens (04)
- un sous-échantillon en population générale (05)
- un sous-échantillon des petits enfants d'immigrés (06)

Tableau 1 : Effectifs répondants visés pour chacun des sous-échantillons

		SOUS-ECHANTILLONS						
		Immigrés	Enfants de personne(s) nées à l'étranger	Domiens	Enfants de Domien(s)	G3 d'origine non-européenne'	Population générale*	
C A T E G O R I E S D' A N A L Y S E	Immigrés (G1)	9 000					3 400 population générale au total	
	Enfants d'immigré(s) (G2)		9 000					
	Domiens (G1)			800				
	Enfants de Domiens (G2)				800			
	« Population majoritaire »	G3 d'origine européenne						
		G3 d'origine non-européenne						500
		Enfants de Français de l'étranger		3 000				
Autres								

* Le sous-échantillon population générale est représentatif de l'ensemble de la population vivant en France métropolitaine. Les individus de ce sous-échantillon peuvent donc être immigrés ou descendants d'immigrés, Domiens ou descendants de Domiens, nés de parents nés Français à l'étranger, petits-enfants d'immigrés ou issu de la « population majoritaire ». Les probabilités d'appartenir à deux sous-échantillons nécessiteront le recours à la technique du partage des poids (cf. partie 4).

2.1.2 La méthode d'échantillonnage de TeO2

À l'issue du bilan méthodologique de TeO1 réalisé au début du projet TeO2 (cf. encadré ci-après), il a été décidé de reconduire le principe général de la méthode d'échantillonnage de la première édition de l'enquête TeO¹⁹. Cette méthode consiste en la mise en œuvre d'un dispositif exceptionnel 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 les enquêtes annuelles de recensement, qui constituent la base de sondage de l'enquête TeO. Bien que coûteuse et risquée à l'origine, cette opération est apparue comme un succès. La méthode de constitution de l'enquête TeO2 a toutefois été adaptée pour répondre aux nouveaux objectifs de cette édition de l'enquête (nouveaux groupes échantillonnés pour les immigrés, échantillonnage *ad hoc* 3^{ème} génération ...).

¹⁹ La méthode de constitution de l'échantillon de TeO1 est détaillée dans 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

Encadré 5 : Suite du bilan sur l'échantillonnage de TeO1

Extension de la tranche d'âge des enfants d'immigré(s) au-delà de 49 ans

Dans TeO1, le champ d'âge couvert pour les immigrés (18-59 ans) était différent de celui des enfants d'immigré(s) (18-49 ans) du fait de contraintes d'échantillonnage (recul temporel insuffisant des fichiers d'état civil anonyme permettant d'identifier les enfants d'immigré(s)). Ces contraintes sont moins fortes dix ans plus tard du fait de l'avancée en âge des enquêtés. Pour TeO2, il a ainsi été décidé d'étendre la tranche d'âge des enfants d'immigré(s) à 59 ans.

Bases de sondage alternatives à l'enquête annuelle de recensement (EAR) 2018 et non retenues pour l'enquête TeO2

1. La première solution, celle d'une enquête filtre, a été écartée pour des questions de coût et de faisabilité. Elle consiste à faire une enquête téléphonique pour demander le lieu de naissance et la nationalité à la naissance des parents de la personne.
2. La seconde solution aurait été de mobiliser l'enquête emploi en continu (EEC). Cela aurait impliqué de mobiliser 10 vagues de sorties de l'EEC pour pouvoir couvrir toutes les origines avec des effectifs suffisants. Par ailleurs, l'Insee s'engage auprès des répondants à l'enquête emploi à ne plus les interroger au terme des six interrogations trimestrielles pour d'autres enquêtes.
3. La troisième solution était celle du Répertoire statistique des logements, devenu le Fichier démographique d'origine fiscale sur les logements et les personnes (Fideli). Ce dernier répertorie l'ensemble des personnes connues des services fiscaux et assure une bonne couverture de la population tout en disposant des informations nominatives des personnes (le nom et le prénom). Il présente l'avantage d'être chaîné sur 2 années, ce qui permet de retrouver les personnes qui ont déménagé. Pour autant, Fideli ne dispose pas de certaines informations indispensables pour constituer l'échantillon des immigrés et de leurs enfants : nationalité française ou non des personnes à leur naissance, filiation...
4. La question d'un éventuel appariement avec la BRPP (Base des répertoires des personnes physiques) s'est posée dans l'optique de diminuer le volume de recherches en mairie nécessaire à la constitution de l'échantillon des enfants d'immigré(s). Cette piste n'a pas été retenue pour la raison suivante : le bulletin de naissance permet *a priori* de repérer les enfants d'immigré(s) potentiels, mais les informations sur la filiation sont supprimées dans le processus de production de la BRPP. Seuls le numéro de sécurité sociale, le numéro d'acte d'état civil, un identifiant interne plus précis que le n° de sécurité sociale, la date de naissance, le lieu de naissance de la personne sont conservés. En accord avec la CNIL, les autres informations (identité des parents, lieu de naissance des parents...) sont purgées pour les individus qui ont plus de 3 ans d'âge.

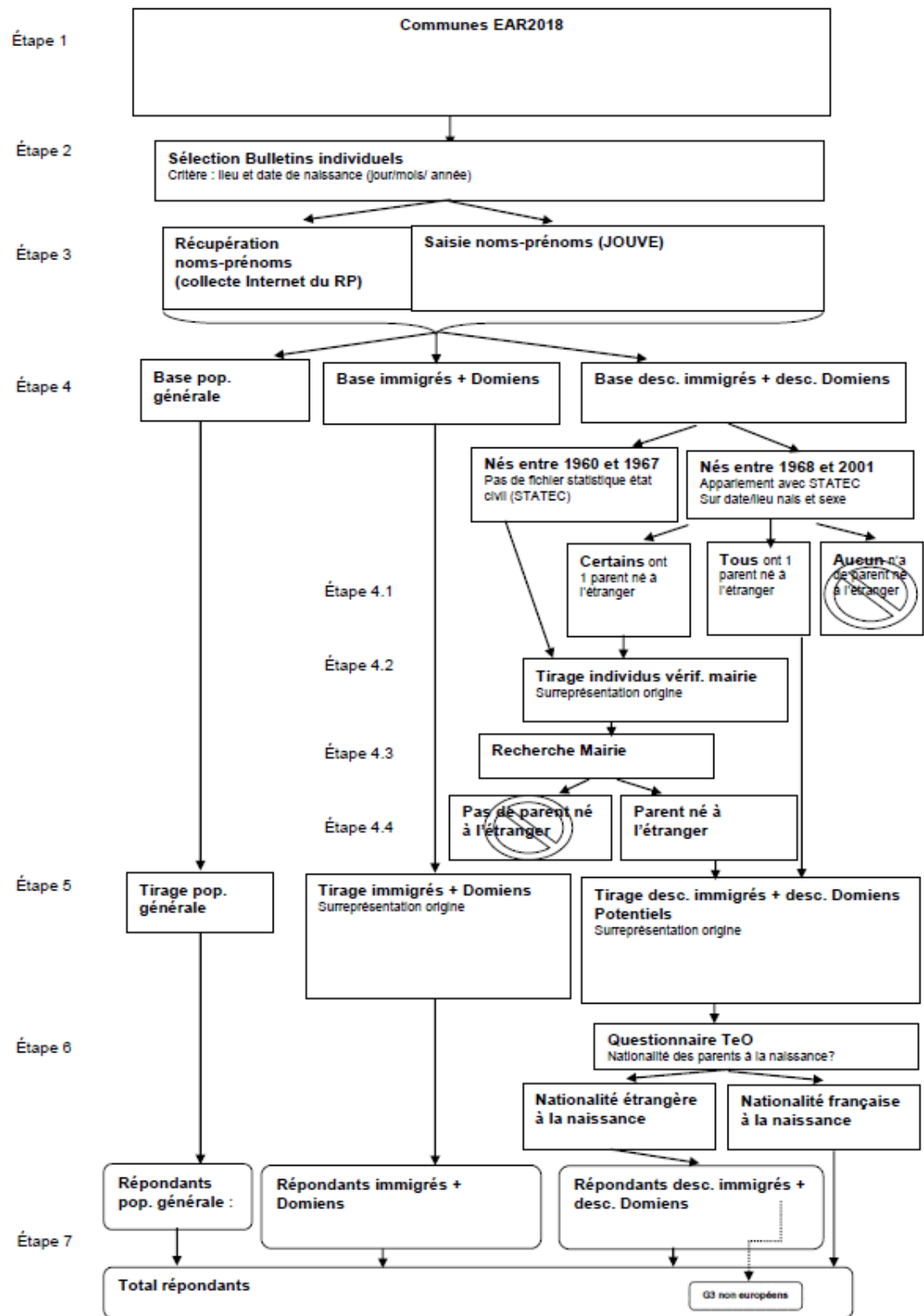
L'EAR 2018 a été retenue comme base de sondage pour TeO2.

Inclusion du 1er janvier parmi les dates de naissance sélectionnées pour le tirage des individus

Les individus de la base de sondage de TeO1 étaient sélectionnés à partir de l'EAR 2007 selon des critères liés à la naissance (lieu et date). Initialement, la sélection des jours de naissance avait été choisie arbitrairement, avec l'hypothèse que cela assurait une sélection aléatoire des individus. Or, il s'est avéré qu'une forte concentration d'immigrés (en particulier parmi ceux originaires de pays dans lesquels l'enregistrement des naissances à l'état civil est déficient) déclarent être nés le 1er janvier à défaut d'une date précisément connue. Ces personnes immigrées déclarant être nées le 1er janvier ont été rajoutées *a posteriori* à la base de sondage de TeO1. Dans la constitution de la base de sondage de TeO2, il a été décidé de retenir *a priori* dans le plan de sondage les personnes nées le 1er janvier.

Le schéma ci-dessous décrit les principales étapes de constitution de l'échantillon.

Schéma 1 : Les principales étapes de constitution de l'échantillon



Les étapes du schéma sont brièvement décrites ci-dessous. La méthodologie d'échantillonnage est, quant à elle, détaillée dans les parties 2.2 à 2.4.

Étape 1 : choix d'une base de sondage

La base de sondage utilisée pour l'enquête TeO2 est l'enquête annuelle de recensement 2018 (EAR 2018) : l'ensemble des personnes qui vivent dans les communes de moins de 10 000 habitants recensées cette année-là (une sur cinq) et environ 8% de la population vivant dans des communes de plus de 10 000 habitants. Il s'agit de la source la plus récente qui soit disponible. Cela permet d'avoir des adresses récentes, et donc de limiter les biais et les difficultés de contact liés aux déménagements. Par ailleurs, la collecte de l'enquête est principalement réalisée dans les zones recensées qui font partie de l'échantillon maître²⁰ : cela couvre presque totalement les zones urbaines mais une partie seulement des zones rurales.

Étape 2 : sélection des bulletins individuels (BI) sur le champ de TeO2

L'échantillon de l'enquête est constitué à partir :

- des personnes qui ont été recensées en 2018,
- vivant en métropole,
- nées principalement entre le 20 et le 25 de chaque mois (bornes incluses) et le 1^{er} janvier entre 1960 et 2001²¹,
- résidant dans une zone couverte par l'enquête EAR qui correspond aux zones de l'échantillon maître (EM) plus quelques zones hors EM.

Étape 3 : récupération des informations à caractère personnel des individus

L'enquête TeO2 portant auprès d'individus, il est indispensable de construire une base de sondage nominative et de recueillir leurs noms et prénoms. Ces données figurent dans les réponses aux questionnaires Internet de l'EAR 2018, mais pour les bulletins papiers, elles ne sont pas systématiquement saisies²². C'est pourquoi la saisie spécifique des noms et prénoms a été programmée dans ce deuxième cas : elle couvre près de 350 000 personnes recensées.

²⁰ Les zones de l'échantillon maître correspondent à l'ensemble des zones géographiques sur lesquelles les ménages enquêtés ont vocation à être sélectionnés

²¹⁷ Le choix des jours est repris de l'expérience de TeO1. Lors de la saisie des données du recensement courant, les noms et prénoms des enquêtés nés le 1^{er} janvier ou entre le 20 et le 25 de chaque mois ont été saisis manuellement dans le cadre d'une sous-traitance avec la société Jouve spécifique à l'enquête TeO2 alors que, pour les besoins usuels du recensement, cette saisie est uniquement optique sur tous les BI. La saisie manuelle assure une qualité des noms et prénoms supérieure à celle de la saisie optique. Pour les besoins de l'enquête TeO2 (enquête auprès des individus), il a paru utile d'avoir l'information sur les noms et prénoms la plus fiable en vue du repérage des enquêtés. Pour TeO2, une extension des jours de naissance a été nécessaire à cause de contraintes d'échantillonnage. Ce point est développé dans les parties 2.2 et 2.3

²²⁸ La collecte par internet du recensement, introduite en 2015, simplifie l'échantillonnage de TeO2. Le contexte a évolué depuis la première édition de TeO. Il est désormais possible de récupérer les noms et prénoms des bulletins individuels via le canal de la collecte par internet du recensement (récupération des informations nominatives sans coût financier). Le taux de réponse par internet pour les bulletins individuels était de 56 % en 2018 et de 46 % en 2016 au moment des premières réflexions sur l'échantillon.

Étape 4 : Constitution de 5 bases de sondage distinctes

À cette étape, 5 bases ont été constituées pour permettre le tirage de l'échantillon :

- Immigrés + Natifs des DOM : bases directement constituées à partir de l'EAR 2018, toutes les informations nécessaires y figurant
- Enfants de parents nés à l'étranger et dans les DOM : l'EAR 2018 ne renseignant pas le lieu de naissance des parents, les bases correspondantes sont constituées à partir de l'EAR 2018 en excluant dans un premier temps les personnes immigrées et les natives des DOM, – cf. points 4.1 & 4.2
- Population générale : base directement constituée à partir de l'EAR 2018, représentative de l'ensemble de la population et à partir de laquelle sont identifiées les personnes de la population « majoritaire ».

La base population générale recouvre la « population majoritaire » mais également les immigrés, les Domiens et leurs enfants.

Pour en faciliter la lecture, dans le schéma, les immigrés et les Domiens ont été regroupés à cette étape. Les enfants d'immigré(s) et de Domiens sont également mis ensemble pour la même raison.

Étape 4.1 : appariement avec le fichier anonyme d'état civil en vue de l'identification des descendants d'immigrés et de Domiens potentiels

Les étapes 4.1 à 4.4 concernent uniquement les échantillons des enfants d'immigré(s) et de Domiens.

Afin d'identifier les descendants d'immigrés et de Domiens potentiels dans la base de sondage, l'Insee a procédé à une consultation préalable de la base des bulletins anonymes de naissance qu'il collecte depuis 1968 pour les personnes du champ de l'enquête enquêtées à l'EAR 2018 et n'étant ni immigrées, ni nées dans un Dom. La méthodologie de cette étape est détaillée dans la partie 2.3.1.

Étape 4.2 : tirage d'individus permettant de constituer l'échantillon d'individus faisant l'objet d'une recherche en mairie (consultation de leur bulletin individuel de naissance, sur lequel figurent les pays de naissance de leurs parents).

Afin de limiter les coûts des recherches de relevés d'état civil en mairie (déplacement des enquêteurs, perte de temps), un tirage d'individus a été effectué. Il a permis de constituer un échantillon d'individus faisant l'objet d'une recherche en mairie (consultation de leur bulletin individuel de naissance, sur lequel figurent les pays de naissance de leurs parents). Cet échantillon est composé des sous-populations :

- des individus nés entre 1960 et 1967
- des correspondants ayant au moins un parent né à l'étranger ou dans un DOM en fonction de cibles prédéfinies

La méthodologie de cette étape est détaillée dans la partie 2.3.2.

Étape 4.3 : recherches en mairie

L'opération en mairie de consultations des bulletins d'état civil de naissance de ces personnes a dû être autorisée par le procureur de Paris.

Au total, la recherche en mairie a concerné :

- les individus tirés dans la sous-population pour laquelle l'appariement avec le fichier anonyme d'état civil avait permis d'identifier au moins un correspondant (personne née le même jour dans la même commune avec le même sexe),
- les individus nés entre 1960 et 1967 pour lesquels l'Insee ne dispose pas de fichier statistique d'état civil.

Étape 4.4 : sélection des individus dont au moins un parent est né à l'étranger

Parmi les individus recherchés en mairie, seuls les individus dont il a été validé qu'au moins un parent est né à l'étranger ou dans un DOM ont été retenus.

Étape 5 : Tirage des sous-échantillons

Parmi les personnes des sous-échantillons « Domiens », « enfants de Domiens » et population générale, les fiches-adresses sont tirées aléatoirement (« tirage simple »).

En revanche, les sous-échantillons d'immigrés et d'enfants d'immigré(s) sont stratifiés selon l'origine, avec des taux de sondage variables par origine pour permettre d'atteindre les objectifs fixés (effectifs par origine). Le détail de la méthodologie de tirage est décrit dans les parties 2.2 et 2.3.

Étape 6 : Identification des enfants d'immigré(s)

Pour les personnes dont au moins un parent est né à l'étranger, c'est la nationalité de naissance du/des parent(s) qui permet de déterminer parmi les personnes nées en France celles qui sont des enfants d'immigré(s). C'est à partir du questionnaire de l'enquête TeO2 que l'on recueille cette information. À l'issue de l'interrogation sur la nationalité des parents, deux cas de figure se présentent :

- L'individu répondant a au moins un parent né de nationalité étrangère : il s'agit d'un enfant d'immigré(s) ;
- L'individu répondant a deux parents de nationalité française à la naissance : il s'agit d'un enfant de « Français nés à l'étranger » (ex : les enfants de rapatriés). Il appartient également au champ de l'enquête TeO2, mais il est reversé, lors des analyses, dans la population « majoritaire ».

Étape 7 : Petits-enfants d'immigré(s) d'origine non européenne

Pour enquêter les 3^{èmes} générations non européennes, il est nécessaire de constituer un sous-échantillon spécifique pour atteindre un effectif suffisant de petits-enfants d'immigrés d'origine non européenne. La base de sondage est constituée lors de la collecte de l'enquête, en collectant les coordonnées de ces petits-enfants d'immigrés d'origine non européenne auprès de leur parent descendant d'immigrés non européens répondant à l'enquête TeO2. La méthodologie est détaillée dans la partie 2.4.

2.2 Les 1^{ères} générations et la population générale

Cette partie précise la méthode d'échantillonnage des populations enquêtées lors de la première vague de collecte.

Les individus enquêtés lors de la vague 1 (juillet 2019 – décembre 2019) appartiennent aux sous-échantillons des immigrés, des domiens, ainsi qu'au sous-échantillon de la population générale.

Par la suite, les populations immigrés et Domiens pourront être appelées la « 1^{ère} génération » ou « G1 » par souci de simplification.

2.2.1 Le choix des groupes d'origines et allocations des G1

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 établir des cibles par groupes d'origines pour pouvoir commenter des résultats concernant ces différents groupes (par exemple, les conditions d'emploi des immigrés d'Afrique subsaharienne par rapport aux immigrés de pays de l'Union Européenne). Afin de pouvoir diffuser sur chacun des groupes d'origines retenu, il faut à la fois avoir un nombre suffisant de fiches-adresses mises en collecte, compte tenu des taux de réponse escomptés, mais aussi éviter une trop grande dispersion des poids.

Dans la première édition de l'enquête TeO, 9 groupes d'origines étaient ciblés (c'est-à-dire faisant l'objet d'un tirage spécifique et d'un objectif de répondants), et environ 1 200 individus d'autres origines étaient enquêtés (groupe « résiduel » : « Autres pays ») ; pour cette nouvelle édition, un objectif majeur pour le sous-échantillon des immigrés était de 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.

Dans cette perspective, la division Sondages de l'Insee a mené des simulations sur les conséquences de l'intégration de groupes d'origines supplémentaires sur la précision globale de l'enquête (présentée en annexe 2). De ces simulations ont découlé deux recommandations importantes. La première concernait le nombre de groupes d'origine à cibler : compte tenu du volume de FA disponibles, ce nombre ne peut excéder 10 groupes d'étude et 1 groupe résiduel, soit un de plus que pour TeO1. La deuxième recommandation concernait la taille de ces groupes : pour des raisons de précision, il ne faudrait pas descendre en dessous d'une taille visée de répondants de 800²³. Compte tenu des incertitudes sur les taux de réponse, viser une taille inférieure ferait peser une forte incertitude sur les possibilités d'exploitation. Il était possible que la taille finale de certains groupes soit inférieure à 800 répondants et c'est pourquoi il était important de ne pas descendre en-dessous de cet objectif.

Une fois les cibles par groupe d'origines déterminées, le nombre de FA à mettre en collecte (allocation) a été calibré par groupe d'origines en tenant compte à la fois du taux de collecte par groupe d'origines de

²³ En pratique, dans TeO2, il est recommandé de mener des analyses sur des groupes comptant au moins 400 observations (analyse de précision sur l'estimation d'une proportion simple, annexe 2). Cet objectif de 800 répondants par groupe doit permettre de mener des analyses plus fines, par exemple ventilées par sexe.

l'enquête TeO1 et du taux de collecte global anticipé pour TeO2 (54%²⁴). Le tableau ci-après résume pour chaque groupe d'origines retenu, les tailles des objectifs de diffusion (cible), ainsi que le nombre de FA mises en collecte initialement prévu.

Tableau 2 : Allocation par groupes d'origines initialement prévue

Groupe d'origines	Immigrés	
	Effectifs d'Immigrés répondants souhaités	Nombre de fiches-adresses sélectionnées
1. Algérie	1 000	1998
2. Maroc Tunisie	1 000	1885
3. Afrique sahélienne	800	1737
4. Afrique centrale et du golfe de Guinée	800	1567
5. Portugal-Espagne-Italie	1 000	1637
6. Autre UE 28	800	1354
7. Turquie	800	1377
8. Asie du Sud-Est	800	1479
9. Chine	700	1294
10. Pays avec de nombreux réfugiés	700	1294
11. Autres pays	600	1198
Total	9 000	16665

2.2.2 La construction de la base de sondage des immigrés, des Domiens et de la population générale

Pour rappel, la base de sondage utilisée est l'EAR 2018. Afin d'obtenir une base de sondage nominative, les noms et prénoms de certains individus appartenant au champ de l'enquête²⁵ ont été ajoutés²⁶ 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 de nombreux immigrés d'origines particulières n'avaient 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).

²⁴ Le taux de collecte est le rapport entre le nombre de FA réussies ou bien collectées et le nombre de FA mises en collecte, résultant d'informations sur la collecte des enquêtes en face-à-face à l'Insee et des spécificités de TeO2.

- suite à la mise en place du nouveau cadre d'emploi des enquêteurs (NCEE), le taux de collecte des enquêtes à l'Insee a baissé. Pour TeO2 on anticipait également une baisse du taux de collecte par rapport à TeO1 lié au passage au NCEE.
- le taux de collecte de l'Île-de-France (IdF) était autour de 50 %. Etant donné que l'échantillon de TeO2 en IdF est volumineux (lié à la concentration des immigrés et enfants d'immigrés dans cette région), il était important de prendre en compte ce paramètre dans l'estimation du taux de collecte anticipé.
- le caractère sensible de l'enquête TeO2 par rapport aux autres enquêtes de l'Insee
- des personnes a priori plus difficiles à atteindre (population immigrée)
- un protocole individu complexe : le suivi des déménagements n'existe pas dans la plupart des enquêtes ménages
- le caractère obligatoire de l'enquête

²⁵ Pour rappel le champ de l'enquête est constitué des individus nés entre 1960 et 2001 inclus, résidant en France métropolitaine, en logement ordinaire

²⁶ via des opérations d'enrichissement complémentaire des noms et prénoms (saisie automatique et manuelle)

Pour certaines origines pour lesquelles la surreprésentation nécessaire 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 de naissance pour certains groupes d'origines (« Chine », « Asie du Sud-Est » et « Afrique centrale et du golfe de Guinée »). Ainsi, les jours utilisés pour le tirage ne se limitent pas aux jours saisis spécifiquement pour TeO2 (1^{er} janvier et les jours du 20 au 25 de chaque mois) mais incluent également les jours du 14 au 19 inclus de chaque mois. Les noms et prénoms des individus concernés par ces jours de naissance supplémentaires 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. Les jours de naissance utilisés sont alors ceux du 10 au 19 de chaque mois.

Enfin, pour la base de sondage des Domiens, il n'y avait pas de contrainte d'échantillonnage particulière. La base de sondage se limite aux jours saisis spécifiquement pour TeO2 (1^{er} janvier et les jours du 20 au 25 de chaque mois).

2.2.3 Le tirage des G1 et de la population générale

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 des strates. 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 (de 25% à 20%), région qui aurait sans cela concentré une part trop importante de la collecte, ce qui aurait entraîné une surcharge pour la Direction régionale concernée, et donc un risque sur les taux de collecte. L'échantillonnage se fait ensuite à probabilités inégales, en suivant la formule suivante pour un individu k donné :

$$\pi_k = \frac{w_{EAR}w_{ZAE}w_{JNAI}w_{IndLog}w_{ssech}}{\sum_{strate,reg} w_{tot}}$$

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} poids de tirage du logement de l'individu dans l'enquête annuelle de recensement (EAR 2018)²⁷

²⁷ 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. Il est donc nécessaire de prendre en compte cette phase pour le tirage. Pour une description plus précise du fonctionnement des EAR, on pourra se reporter à Godinot (2005). Pour comprendre le recensement de la population, Insee Méthodes, hors-série - mai 2005, <https://insee.fr/fr/information/2579979>.

- W_{ZAE} poids lié au tirage de l'Échantillon-Maître (EM)²⁸
- W_{JNAI} 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 1er janvier.
- W_{IndLog} poids lié au tirage d'un individu par logement : 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 ;
- $W_{Sousech05}$ poids spécifique pour le tirage en population générale, dont l'échantillonnage est réalisé dans un second temps, une fois chaque strate concernant les populations immigrées et domiennes enquêtée ; 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 ;
- $\sum W_{tot}$ est la somme des poids totaux mise au dénominateur pour garantir le respect de l'allocation de tirage.

2.2.4 Echantillon complémentaire et taille finale de l'échantillon de la vague 1

Le tirage réalisé dans un premier temps ne couvrait pas la totalité des individus du champ : les individus nés en 1979, 1980, 1989 et 1990 étaient absents de la base de sondage. Étant donné leur caractère atypique dans les exploitations par tranche de cinq ans, il était nécessaire de réaliser un échantillonnage complémentaire. La méthodologie de tirage de cet échantillon complémentaire reprend en grande partie celle décrite au-dessus. Elle est détaillée en annexe 3.

À l'issue du tirage, comme indiqué dans le dernier point ci-dessus, certains individus échantillonnés appartiennent au même logement que des individus ayant été échantillonnés dans le cadre de l'échantillon principal. On compte ainsi 170 individus appartenant à un tel logement. Il a été décidé à cette étape de ne pas modifier la méthodologie de tirage, mais de réaliser ex post un tirage au sein de ces logements d'un des deux individus à probabilités égales (et de modifier le poids de l'individu choisi en conséquence). Ce tirage conduit à :

²⁸ L'EM Octopusse est en vigueur à l'Insee pour la période 2010 – 2020. Il s'agit d'un ensemble de zones géographiques (communes ou groupes de communes) 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} est uniformément égal à 1, et l'échantillonnage se fait sur l'ensemble du territoire de la région concernée. L'EM Octopusse est décrit dans 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.

- retirer de l'échantillon principal 76 individus ;
- retirer de l'échantillon complémentaire 94 individus.

Le tableau ci-après résume les tailles des objectifs de diffusion (cible), ainsi que le nombre de FA finalement mises en collecte par sous-échantillon et pour chaque groupe d'origine retenue du sous-échantillon « immigrés ».

Tableau 3 : Allocation finale G1 par sous-échantillon (groupes d'origines)

Sous-échantillon (origines)	Effectifs répondants souhaités	Nombre de fiches-adresses sélectionnées
01 – Immigrés	9000	18 844
1. Algérie	1 000	2126
2. Maroc Tunisie	1 000	2304
3. Afrique sahélienne	800	1730
4. Afrique centrale et du golfe de Guinée	800	1936
5. Portugal-Espagne-Italie	1 000	1820
6. Autre UE 28	800	1528
7. Turquie	800	1533
8. Asie du Sud-Est	800	1559
9. Chine	700	1431
10. Pays avec de nombreux réfugiés	700	1451
11. Autres pays	600	1426
02 – Domiens	800	1 646
05 – Population générale	3 400	6 928

Au final, près de 27 500 individus sont enquêtés en vague 1, répartis de la façon suivante :

- 18 844 au titre du sous-échantillon Immigrés (01)
- 1 646 au titre du sous-échantillon Domiens (02)
- 6 928 au titre du sous-échantillon Population générale (05)

2.3 Les 2^{èmes} générations

Cette partie revient en détail sur l'échantillonnage des « 2^{èmes} générations » c'est-à-dire des descendants d'immigrés et de Domiens (cf. schéma en annexe 4).

Un descendant d'immigré est une personne née en France dont au moins un parent est immigré. Un enfant de Domiens ou descendant de Domiens vérifie la condition suivante : un de ses parents (ou les deux) l'est. Par la suite, ces populations pourront être appelées la « 2^{ème} génération » ou « G2 » par souci de simplification.

Échantillonner cette population est un défi. Un dispositif exceptionnel avait été mis en œuvre dans l'enquête TeO1 pour échantillonner les enfants d'immigré(s) et de Domiens, 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 nécessaires 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 groupe d'origine par groupe d'origine.

Dans TeO2, nous avons utilisé la même approche que pour TeO1 pour échantillonner les enfants d'immigrés et de Domiens. Comme pour les 1^{ères} générations et la population générale, nous avons utilisé l'EAR 2018 comme base de sondage de départ même si l'information complète n'est pas disponible pour repérer les descendants. En effet, il n'existe pas de base de sondage d'individus dans laquelle on dispose du lieu de naissance des parents, ce qui aurait permis de définir une population de descendants de personnes nées à l'étranger ou dans les DOM. Il était donc nécessaire de construire une telle base de sondage. Toutefois, une telle base de sondage est plus large que la population cible de descendants d'immigrés ou de Domiens. En effet, avec uniquement l'information du lieu de naissance des parents, il est seulement possible d'identifier les descendants de personnes nées à l'étranger ou dans les DOM qui « potentiellement » pourraient être des enfants d'immigrés ou de Domiens. Par exemple, le parent de l'individu pourrait très bien être une personne née à l'étranger sans pour autant être immigrée car de nationalité française à la naissance. C'est le cas notamment des enfants de « Français à l'étranger » (ex : les enfants de rapatriés, d'expatriés...).

2.3.1 Constitution de la base de sondage des descendants

Pour identifier les individus descendants potentiels²⁹ dans la population, nous utilisons les fichiers d'état civil, 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 fichier couvre uniquement les années 1968 et après.
- Des versions complètes, accessibles uniquement auprès des mairies des communes de naissance des individus après accord des procureurs de la République.

L'idée de la construction de cette base de sondage est de mobiliser autant que possible les fichiers anonymisés afin de limiter les recherches directes dans les registres des communes de naissance (opération terrain coûteuse). Dans cette perspective, un appariement de l'enquête annuelle de recensement de 2018 avec les fichiers de l'état civil anonymisés (base STATEC) a été réalisé. La base à apparier correspond aux individus ayant entre 18 et 59 ans, dont les dates de naissance sont le 1^{er} janvier et du 10 au 25 inclus de chaque mois³⁰. Cela correspond à 2 116 688 personnes dont sont exclues les personnes immigrées ou domiennes, qui, par définition, ne peuvent être G2.

²⁹ 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 une étape dans l'identification des enfants d'immigrés.

³⁰ Le choix des jours de naissance a été étendu afin de permettre des sur-représentations de certaines origines dans le but d'atteindre les objectifs cibles pour chaque groupe d'origines

Ainsi, pour chacune des personnes nées en métropole entre 1968 et 2001³¹ et n'étant ni immigrée, ni née dans un Dom présente dans l'enquête annuelle de recensement 2018 (EAR), a été recherché l'ensemble de ses correspondants « anonymes » (c'est-à-dire les personnes nées le même jour dans la même commune avec le même sexe) dans les fichiers anonymisés de l'état civil (base STATEC). Le lieu de naissance des parents des correspondants associés pouvant être très divers, 3 situations sont distinguées :

- tous les correspondants liés à l'individu de l'EAR ont au moins un parent né à l'étranger ou dans les DOM.
- seuls certains correspondants liés à l'individu de l'EAR ont au moins un parent né à l'étranger ou dans les DOM
- aucun correspondant lié à l'individu de l'EAR n'a de parents nés à l'étranger ou dans les DOM

Pour les personnes de l'EAR nées entre 1960 et 1967, aucun appariement avec STATEC n'était possible.

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. Si au moins un correspondant (mais pas tous) a un parent né à l'étranger ou dans un DOM, la recherche en mairie du lieu de naissance de l'individu a été organisée pour vérifier cette information. Une opération préalable permettant de surreprésenter certaines origines a été organisée (cf partie 2.3.2).

Dans le troisième cas, la personne n'est pas descendante d'immigrés ni de natifs des DOM. Il est de fait inutile d'organiser une recherche en mairie.

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 (nationalité à la naissance des parents) ne sera disponible dans les registres d'état civil³². Dans ce cas, il n'est pas non plus nécessaire d'aller consulter l'état civil en mairie. Le questionnaire de l'enquête TeO2 qui contient l'information du lieu de naissance et de la nationalité à la naissance des parents permettra de distinguer si la personne est un descendant d'immigrés ou bien est née de parents nés Français à l'étranger (cf. étape 6 schéma 1).

2.3.2 Le tirage des relevés mairies

La base est ensuite filtrée pour se limiter aux individus vivant dans une commune de l'Échantillon-Maître (EM). Ne pas se restreindre à l'EM pouvait complexifier largement la collecte et pouvait également remettre en question la surreprésentation des origines rares dans le tirage pour la collecte. En effet il était difficile d'étendre la sélection hors EM pour certaines origines uniquement (avant les relevés mairie, on ne pouvait pas savoir avec certitude quelle origine était concernée). Étendre au-delà de l'EM nous exposait au risque de réduire potentiellement la base de tirage notamment en raison des cas où l'opération mairie n'était pas concluante (par exemple pour les cas où à l'issue de l'opération mairie il s'avère que les individus n'ont pas de parents nés à l'étranger).

³¹ Le champ des sous-échantillons 03 et 04 concerne les 18-59 ans. Cependant, l'appariement n'intègre que les personnes nées à partir de 1968 car la base STATEC n'existait pas avant cette date.

³² Tous les cas sont néanmoins conservés pour alimenter la BS pour le 1er tirage afin de limiter la dispersion des poids.

Afin de limiter les coûts, un nombre limité de relevés en mairie a été réalisé. La taille maximale était fixée à environ 100 000 relevés comme dans l'enquête TeO1. Il fallait donc réaliser un échantillonnage d'individus pour lesquels un relevé en mairie serait 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 termes de nombre de répondants pour les groupes d'origines soient atteints. Pour cela, il a été réalisé autant que possible un tirage à probabilités égales (modulo les étapes antérieures de sélection telles que l'EAR, décrites en partie 2.2.), sauf en ce qui concerne certaines origines dites « rares » (« Afrique Sahélienne », « Afrique Centrale », « Asie du Sud-Est », « Turquie »). 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 correspondant correspond à une origine dite « rare », sont systématiquement inclus dans l'échantillon de relevés mairies. De plus pour ces origines, les jours utilisés pour le tirage ne se limitent pas aux jours saisis spécifiquement pour TeO2 (1^{er} janvier et les jours du 20 au 25 de chaque mois) mais incluent également les jours du 10 au 19 inclus de chaque mois.

Un premier tirage a été effectué sur l'ensemble de la base de sondage restreinte à l'échantillon-maître et à certains jours de naissance, sans séparation selon le résultat de STATEC. L'idée était de tirer globalement dans l'ensemble de cette population pour garder des poids aussi peu dispersés que possible, ce qui sera utile au moment de l'étape du tirage pour la mise en collecte des G2.

Ce premier tirage a abouti à la sélection de 161 200 individus répartis de la manière suivante :

- 24 982 individus nés entre 1960 et 1967. Pour rappel, pour ces individus, l'information sur le lieu de naissance des parents est manquante dans notre base de sondage. Pour savoir s'ils sont dans le champ décrit par les sous-échantillons 03 (descendants d'immigrés) et 04 (descendants de Domiens), une opération de relevés mairie a été réalisée pour vérifier le lieu de naissance des parents. Seuls les individus dont les parents sont nés à l'étranger ou dans les DOM figureront dans la base de sondage des sous-échantillons 03 et 04.
- 75 032 individus nés entre 1968 et 2001 dont seulement certains échos³³ présentent au moins un parent né à l'étranger ou dans les DOM. Pour rappel, pour ces individus l'information sur le lieu de naissance des parents est incertaine dans notre base de sondage. Ces individus sont potentiellement dans le champ des sous-échantillons 03 et 04. Pour savoir s'ils le sont effectivement, une opération de relevés mairie a été réalisée pour vérifier le lieu de naissance des parents. Seuls les individus dont les parents sont nés à l'étranger ou dans les DOM figureront dans la base de sondage des sous-échantillons 03 et 04.
- 8 252 individus nés entre 1968 et 2001 dont tous les échos présentent au moins un parent né à l'étranger ou dans les DOM. Pour eux, il n'y a aucun doute : ils sont dans le champ des sous-échantillons 03 et 04.
- 52 934 individus dont aucun écho n'a de parents nés à l'étranger. Les parents sont soit nés en métropole, soit d'origine inconnue (pays de naissance d'un des parents non renseigné ou un des

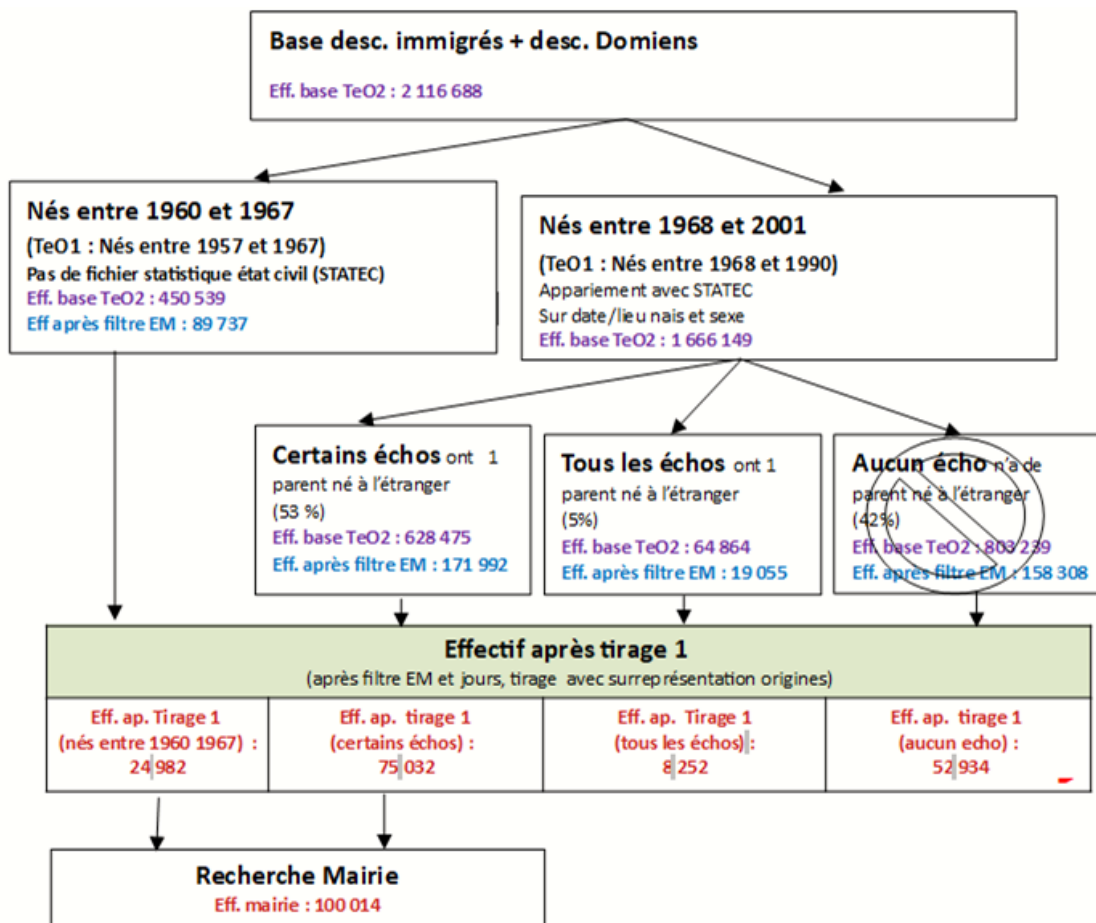
³³ ou également appelés correspondants

parents inconnus). Ces individus ne seront pas dans la base de sondage des sous-échantillons 03 et 04 puisqu'il ne s'agit pas de descendants d'immigrés ou de domiens³⁴

Au final, la vérification des lieux de naissance des parents dans les registres d'état civil présents en mairie ou « opération mairie » a concerné 100 014 individus.

Le schéma ci-après résume le processus d'échantillonnage des G2 jusqu'à l'étape des relevés mairie :

Schéma 2 : Processus échantillonnage G2 (relevés mairie)



2.3.3 Bilan des relevés mairie

L'opération de relevés mairie concernait 100 014 individus. Parmi ceux-ci, 94 881 ont abouti et sont exploitables, après reprise manuelle des commentaires des enquêteurs Insee. Ces reprises manuelles ont concerné le sexe, le nom, le prénom, la date de naissance de l'individu et le pays de naissance des parents. Pour les 5 133 individus qui n'ont pas pu être retrouvés en mairie, une opération de correction de la non-réponse totale est nécessaire. La méthodologie de redressement est présentée dans la partie 4.

³⁴ sauf pour quelques personnes à la marge parmi les échos d'origine inconnue, appelés échos « douteux » mais qu'on est en incapacité de repérer parmi ces 52 934 individus.

Parmi les 94 881 relevés mairie retrouvés, ceux-ci se répartissent de la manière suivante :

Tableau 4 : Croisement des résultats des variables du lieu de naissance du père et de la mère après reprise des relevés mairie exploitables

Naissance du père	Naissance de la mère	France métropolitaine	DOM	Étranger	Inconnu ³⁵¹
France métropolitaine		63 563	323	5 741	81
DOM		526	716	91	2
Étranger		8 337	99	13 549	24
Inconnu		1 364	114	262	89

Seuls les 29 784 individus dans les cases à fond vert seront donc intégrés à la base de sondage des sous-échantillons 03 et 04.

Donc, la base de sondage des sous-échantillons 03 et 04 est constituée :

- de 29 784 individus issus des relevés mairie et pour lesquels il a été vérifié qu'au moins un des parents était né à l'étranger ou dans les Dom ;
- de 8 252 individus dont tous les échos dans STATEC présentent au moins un parent né à l'étranger ou dans les DOM.

2.3.4 Allocations

L'opération en mairie a permis de constituer la base de sondage pour le tirage du sous-échantillon (03) des descendants d'immigrés et (04) de Domiens. Avant le tirage de l'échantillon de la vague 2 (descendants) il est nécessaire de calibrer les allocations de FA par groupe d'origines. Pour rappel, la cible concerne un nombre de descendants d'immigrés et de Domiens, tandis que la base de sondage est une base de descendants de personnes nées à l'étranger ou dans les DOM. Donc, pour atteindre la cible de répondants par origine, il est nécessaire de tenir compte à la fois du fait que, parmi les FA tirées, il y aura des non-répondants et du fait qu'un certain nombre d'individus tirés sont des enfants de personnes nées à l'étranger mais pas immigrées. Le taux de réponse et le taux « d'enfants d'immigrés et de Domiens » ont été estimés à partir de TeO1. La méthodologie d'estimation du nombre de FA mises en collecte est présentée en annexe 5.

³⁵ Lieu de naissance non-renseigné ou parent ne figurant nulle part (ni sur le bulletin de naissance, ni en mention en marge de reconnaissance)

Tableau 5 : Allocations de tirage par groupe d'origines pour les sous-échantillons 03 et 04

Groupe d'origines des parents	Cible d'enfants d'immigrés et de Domiens répondants	Taux de réponse	Taux d'enfants d'immigrés et de Domiens	Nombre idéal de FA tirées	Nombre de FA à tirer en pratique ³⁶
Algérie	1600	57 %	49 %	5 729	5 535
Maroc Tunisie	1200	58 %	59 %	3 507	3 405
Afrique sahélienne	800	56 %	75 %	1 905	1 841
Afrique centrale et guinéenne	800	56 %	69 %	2 070	1 983
Italie Espagne	800	62 %	97 %	1 330	1 285
Portugal	800	60 %	99 %	1 347	1 301
Autres pays de l'UE 28	800	62 %	87 %	1 483	1 433
Turquie	800	58 %	99 %	1 393	1 346
Vietnam Laos Cambodge	800	60 %	73 %	1 826	1 765
Autres pays	600	61 %	37 %	2 658	2 568
DOM	800	55 %	100 %	1 455	1 455
Total	9 800			24 703	23 917

2.3.5 Détermination du groupe d'origines des G2

Les allocations définies dans la partie précédente sont des allocations par zone géographique. Cependant, un individu peut avoir deux parents d'origines géographiques différentes. Dans ce cas, comment sera exploitée sa réponse pour la diffusion de l'enquête ? Sera-t-il intégré dans les exploitations concernant les deux origines ou seulement dans les exploitations d'une seule des deux ?

Cette question est structurante pour le tirage. En effet, pour atteindre la cible du nombre de descendants d'immigrés et de Domiens répondants, il est nécessaire d'affecter la bonne origine aux individus dès la phase de tirage.

L'origine d'un individu est définie par convention de la manière suivante :

- si un des deux parents est né dans les DOM, l'individu est considéré comme descendant de Domien ;

³⁶ Cette colonne tient compte du nombre d'individus disponibles dans les bases de sondage pour chaque origine et du fait qu'on ne tire que 23 917 FA et non les 24 703 FA comme cela serait nécessaire pour atteindre les cibles

- sinon, si les deux parents sont nés à l'étranger, l'individu est considéré comme descendant de l'origine géographique du père ;
- sinon, si un des deux parents est né à l'étranger, l'individu est considéré comme de l'origine géographique de ce parent ;
- sinon, l'individu n'est pas dans le champ des sous-échantillons 03 et 04.

Pour les 29 784 individus de la base de sondage qui ont fait l'objet de relevés mairie et dont un des deux parents est né à l'étranger ou dans les DOM (cf. partie 2.3.3), le lieu de naissance de leurs parents est connu et les règles ci-dessus permettent de déterminer l'origine géographique de l'individu et donc le groupe d'origine d'appartenance sans ambiguïté pour le tirage.

Pour les 8 252 individus dont tous les échos présentent au moins un parent né à l'étranger ou dans les DOM (cf. partie 2.3.2), on est sûr qu'ils sont dans le champ des sous-échantillons 03 et 04. Mais aucune vérification n'a permis de déterminer le lieu de naissance de leur parent. Pour rappel, pour des raisons de coût, ces individus n'ont pas fait partie de l'opération mairie. Toutefois, pour ces individus, les règles ci-dessus permettent de définir une origine géographique pour chacun des échos de l'individu. On est donc en mesure de définir la probabilité d'un individu d'appartenir à un groupe d'origines donnée : il s'agit du nombre d'échos de l'individu associés à ce groupe d'origine divisé par le nombre total d'échos de l'individu. Cette probabilité peut être égale à 1 pour un groupe d'origine lorsque l'individu n'a qu'un écho dans STATEC ou lorsque tous les échos de l'individu sont du même groupe d'origine. Donc, parmi ces 8 252 individus, on distingue :

- 4 867 individus qui n'ont qu'un écho dans STATEC ou dont tous les échos sont du même groupe d'origine. On est certain du groupe d'origine associée à ces individus ;
- 3 385 individus qui ont deux échos ou plus associés à des groupes d'origine différents. On ne peut associer qu'une probabilité inférieure à 1 à l'individu d'être associé à un groupe d'origine donnée.

Au final, ces individus se répartissent de la manière suivante selon les groupes d'origines.

Tableau 6 : Nombre d'individus par groupe d'origine présents dans la base de sondage

Groupe d'origines	Individus issus des relevés mairie	Individus n'ayant pas fait l'objet de relevés mairie		Nombre total d'individus
		Individus d'origine certaine	Individus d'origine incertaine ³⁷	
Algérie	6 876	814	556	8 246
Maroc Tunisie	5 297	668	507	6 472
Afrique sahélienne	1 720	352	333	2 405
Afrique centrale et guinéenne	1 496	237	250	1 983
Italie Espagne	2 923	352	150	3 425
Portugal	2 083	475	262	2 820
Autres pays de l'UE 28	1 833	210	103	2 146
Turquie	1 422	425	199	2 046
Vietnam Laos Cambodge	1 761	333	210	2 304
Autres pays	2 502	713	491	3 706
DOM	1 871	288	324	2 483
Total	29 784	4 867	3385	38 036

Note de lecture : en espérance, 8 246 individus d'origine algérienne sont présents dans la base de sondage. Pour 6 876 individus, on a pu vérifier cette origine grâce aux relevés mairie. Pour 814 individus, on a pu s'assurer qu'ils étaient d'origine algérienne grâce aux appariements STATEC. Enfin, la somme des probabilités d'être d'origine algérienne pour les 3 385 individus appariés avec STATEC dont on ne connaît pas avec certitude l'origine est égale à 556.

2.3.6 Structure de la base de sondage des G2

Les allocations étant fournies par groupe d'origine, il est nécessaire de distinguer les personnes dont on est sûr du lieu de naissance des parents de ceux pour lesquels on n'en est pas certain.

Ainsi la base de sondage est composée de :

- 8 252 individus nés entre 1968 et 2001 dont tous les échos présentent au moins un parent né à l'étranger ou dans les DOM (cf. partie 2.3.2). Les différents échos associés à un individu peuvent tous concerner la même région du monde (il peut d'ailleurs n'y avoir qu'un écho), auquel cas on est certain du groupe d'origine des parents. Mais ces échos peuvent renvoyer à différentes régions du

³⁷ Pour cette colonne, on ne connaît pas l'origine géographique de l'individu de manière certaine. Cette colonne correspond donc à la somme des probabilités des 3 385 individus concernés d'être d'une origine donnée. Par exemple, la somme des probabilités de ces 3 385 individus d'être d'origine algérienne est de 556 (arrondi à l'unité pour plus de lisibilité).

monde, auquel cas le groupe d'origine des individus est incertain. La manière dont ces 8 252 individus ont été séparés en individus dont on est sûr du groupe d'origines des parents et en individus dont on ne connaît pas avec certitude le groupe d'origines des parents, est expliquée à la partie 2.3.7. En pratique, ce sont 4 867 individus dont on connaît le groupe d'origines des parents de manière certaine et 3 385 individus pour qui il existe une incertitude.

- 29 784 individus ayant fait l'objet de relevés mairie et dont il a été confirmé qu'ils avaient au moins un parent né à l'étranger ou dans les DOM (cf. partie 2.3.2). Il n'y a pas de doute sur l'origine des parents.

Cette différence d'information selon les individus conduit à utiliser deux bases de sondage disjointes. En effet, les allocations de la partie 2.3.4 étant fixées par groupe d'origine des parents, on doit distinguer pour le tirage les individus dont on connaît le groupe d'origines des parents de manière certaine de ceux dont on ne connaît que la probabilité d'être d'un certain groupe d'origines.

Pour les besoins du tirage, la base de sondage est scindée en deux :

- une base de sondage avec les 34 651 (29 784 + 4 867) individus dont on connaît le groupe d'origines des parents de manière certaine. À ces individus, on est capable d'associer un groupe d'origines pour leurs parents ;
- une base de sondage avec les 3 385 individus dont on ne connaît pas le groupe d'origines des parents de manière certaine. À ces individus, on ne peut associer que des probabilités que les parents soient d'un groupe d'origines données.

Pour ces deux bases de sondage, la définition du groupe d'origines associée à un individu et le mode de calcul de la probabilité du groupe d'origines sont expliqués dans les parties suivantes.

2.3.7 Constitution des strates de tirage

On souhaite effectuer un tirage stratifié par groupe d'origines.

Pour la base de 34 651 individus dont on connaît le groupe d'origines des parents de manière certaine, on utilise naturellement les strates de tirage correspondant aux cibles de répondants par groupe d'origines. On génère donc les strates suivantes : Algérie, Maroc Tunisie, Afrique sahélienne, Afrique centrale et guinéenne, Italie Espagne, Portugal, Autres pays de l'UE 28, Turquie, Vietnam Laos, Cambodge, Autres pays, DOM.

Pour la base de 3 385 individus dont on ne connaît pas le groupe d'origines de manière certaine, on ne peut pas stratifier par groupe d'origines puisqu'à un individu sont associés plusieurs groupe d'origines. Dans le tableau 6 du nombre d'individus présents dans la base de sondage (partie 2.3.5), le nombre d'individus d'Afrique centrale et guinéenne est exactement égal à l'allocation de tirage pour les individus d'Afrique centrale et guinéenne (partie 2.3.4). Il est donc nécessaire de tirer toutes les personnes qui ont une probabilité non nulle d'être associées à cette origine géographique. Ces individus forment donc une strate exhaustive qu'on nommera *Origines incertaines – Afrique centrale et guinéenne*.

Par ailleurs, pour le tirage des relevés mairie, les personnes qui avaient une probabilité non nulle d'être d'origine d'Afrique sahélienne, d'Asie du Sud-Est, de Turquie ou d'un DOM ont été surreprésentées. Les

poids avant tirage des G2 des individus concernés par cette surreprésentation sont donc similaires après tirage des 8 252 individus dont tous les échos présentaient des parents nés à l'étranger. Afin de limiter la dispersion des poids pour les individus d'un de ces groupes d'origines, on construit donc une strate spécifique pour les individus qui ont une probabilité non nulle d'être d'origine d'Afrique sahélienne, d'Asie du Sud-Est, de Turquie ou d'un DOM et une probabilité nulle d'être d'origine d'Afrique centrale et guinéenne. On nommera cette strate *Origines incertaines - origines surreprésentées*.

Enfin, une troisième strate a été construite, celle des individus ayant une probabilité nulle d'être d'origine d'Afrique sahélienne, d'Afrique centrale et guinéenne, d'Asie du Sud-Est, de Turquie ou d'un DOM. On nommera cette strate *Origines incertaines – origines sous-représentées*.

2.3.8 Tirage pour la collecte des G2 et pondération

Pour chacune des bases de sondage :

- une base de 34 651 individus dont on connaît l'origine des parents de manière certaine ;
- une base de 3 385 individus dont on ne connaît pas l'origine des parents de manière certaine.

un tirage systématique stratifié³⁸ a été effectué. Les bases ont été triées par strate, région de gestion et code commune.

Le tirage a d'abord été réalisé dans la base de sondage des individus d'origine incertaine. En fonction de la probabilité des individus tirés dans cette base d'avoir des groupes d'origines différents, de nouvelles allocations par strate ont été calculées pour le tirage des individus dont le groupe d'origines est connue de manière certaine.

Les allocations de tirage sont les suivantes :

³⁸ La partie 2.3.7 a permis de définir 14 strates de tirage : 11 pour la base des individus aux origines connues, 3 pour celle des individus aux origines incertaines.

Tableau 7 : Allocations par strate de tirage

Base de sondage	Strate de tirage	Allocation
Base de sondage des individus à l'origine certaine	Algérie	5 173
	Maroc Tunisie	3 077
	Afrique sahélienne	1 651
	Afrique centrale et guinéenne	1 733(strate exhaustive)
	Italie Espagne	1 183
	Portugal	1 113
	Autres pays de l'UE 28	1 366
	Turquie	1 213
	Vietnam Laos Cambodge	1 648
	Autres pays du monde	2 240
	DOM	1 207
Base de sondage des individus à l'origine incertaine	Origines incertaines Afrique centrale et guinéenne	680 (strate exhaustive)
	Origines incertaines Origines surreprésentées	1 154
	Origines incertaines Origines sous-représentées	479

Au sein de chaque strate, le tirage s'est effectué à probabilités inégales. Ces probabilités ont été calculées de manière à :

- respecter l'allocation de chaque strate de tirage ;
- sous-représenter l'Île-de-France afin que 20 % des FA des sous-échantillons 03 et 04 soient collectées dans la région de gestion de l'Île-de-France³⁹ ;
- minimiser la dispersion des poids des individus à l'issue du tirage des G2 au sein de chaque strate⁴⁰.

³⁹ Sans vigilance particulière, ce nombre se serait élevé à environ 25 %.

⁴⁰ Selon les strates, les poids des individus avant cette étape de tirage peuvent être assez dispersés. Cela provient du fait que :

- au moment du tirage des relevés mairie, les individus qui avaient une probabilité non-nulle d'être d'une origine rare ont été surreprésentés. Mais ils peuvent s'avérer être nés de parents étrangers qui ont une origine relativement fréquente. Cela conduit à avoir, avant ce tirage des G2, des individus avec des poids très faibles parmi les origines suivantes : Algérie, Maroc – Tunisie, Italie – Espagne, Portugal, Autres pays de l'UE 28, Autres pays du monde.

Le calcul des probabilités de tirage $\pi_{tirageG2}$ est réalisé de manière itérative, afin que les probabilités de tirage n'excèdent pas 1. À l'issue de cette itération, la formule retenue pour les probabilités de tirage d'un individu est la suivante pour les unités qui ont une probabilité inférieure à 1 d'être tirées :

$$\pi_{i,tirageG2} = (n_{str} - n_{str}^{exh}) \frac{C_{idf} w_{i,apresreleves}}{\sum_{i \notin exh} C_{idf} w_{i,apresreleves}}$$

i désigne un individu. n_{str} correspond à l'allocation de la strate, n_{str}^{exh} au nombre d'individus tirés exhaustivement pour éviter que leur probabilité de tirage ne dépasse 1 lorsqu'elle est calculée avec la formule ci-dessus. C_{idf} désigne le coefficient de sous-représentation de l'Île-de-France, fixé à 0,57 pour garantir que seuls 20 % des individus tirés vivent en région de gestion de l'Île-de-France. $w_{i,apresreleves}$ est le poids de l'individu i après le tirage des relevés mairie. Il se calcule comme $w_{apresreleves} = W_{EAR} W_{ZAE} W_{jour} W_{releves}$.

avec $w_{releves}$ désignant le poids lié au tirage de l'individu pour l'étape des relevés mairie.

Le fait de calculer cette probabilité proportionnellement au poids des individus après tirage des relevés mairie garantit qu'on limite la dispersion des pondérations finales.

Ainsi le tirage des G2 découle du processus suivant :

- tirage du logement de l'individu dans l'enquête annuelle de recensement
- tirage de la ZAE de l'individu dans l'échantillon-maître
- choix d'un nombre de jours de tirage pour un individu donné⁴¹
- tirage de l'individu pour l'étape des relevés mairie
- tirage de l'individu à l'étape du tirage des G2 : $w_{tirageG2} = \frac{1}{\pi_{tirageG2}}$

Le poids de l'individu à l'issue du tirage des G2 est donc : $w_{G2} = W_{EAR} W_{ZAE} W_{jour} W_{releves} w_{tirageG2}$

Le tirage a conduit à retirer 38 individus en doublon.

-
- les individus nés entre 1960 et 1967 n'ont pas pu être appariés avec STATEC. Donc les individus nés durant cette période et originaires d'Afrique sahélienne, d'Afrique centrale et guinéenne, de Turquie, d'Asie du Sud-Est ou des DOM n'ont pas pu être surreprésentés à l'étape du tirage des relevés mairie. Cela conduit à avoir des individus avec des poids très élevés dans les strates correspondantes

Cette dispersion des poids avant tirage des G2 explique la dispersion des poids par strate après tirage

⁴¹Le 1er janvier est considéré comme un jour comme un autre, car il n'y a pas plus de naissances le 1er janvier qu'un autre jour pour les G2. Ce n'était pas le cas pour le tirage des G1 où le 1er janvier était considéré comme un jour exhaustif compte-tenu de la forte proportion d'individus nés à l'étranger un 1er janvier.

Au final, près de 24 000 individus sont enquêtés en vague 2, répartis de la façon suivante :

- 22 675 au titre du sous-échantillon des descendants d'immigrés (03)
- 1 204 au titre du sous-échantillon des descendants de Domiens (04)

Tableau 8 : Allocation finale des G2 par sous-échantillon (groupes d'origines)

Sous-échantillon (origines)	Effectifs répondants souhaités	Nombre de fiches-adresses sélectionnées
03 – Descendants d'immigrés	9 000	22 675
1. Algérie	1 600	5 161
2. Maroc Tunisie	1 200	3 075
3. Afrique sahélienne	800	1 650
4. Afrique centrale et du golfe de Guinée	800	1 730
5. Espagne-Italie	800	1 180
6. Portugal	800	1 113
7. Autres pays de l'UE 28	800	1 365
7. Turquie	800	1 213
8. Asie du Sud-Est	800	1 647
9. Autres pays	600	2 235
10. Origines incertaines - Afrique centrale et guinéenne		677
11. Origines incertaines - Origines surreprésentées		1 151
12. Origines incertaines - Origines sous-représentées		478
04 – Descendants de Domiens	800	1 204

2.4 Les 3^{èmes} générations non européennes : sondage indirect, collecte par lots

Comme évoqué dans les parties 1.2.2.1 et 2.1.2, cette partie présente la méthodologie pour enquêter les 3^{èmes} générations non européennes, qui font l'objet d'un volet complémentaire à l'enquête TeO2 et à visée expérimentale.

Pour les 1^{ères} et 2^{èmes} générations, les individus sont directement sélectionnés dans des bases de sondage constituées et dont la méthodologie a été présentée plus haut. Le sondage des 3^{èmes} générations est un sondage particulier, car il s'agit d'un sondage indirect. Il est nécessaire de constituer un sous-échantillon spécifique pour atteindre un effectif suffisant de petits-enfants d'immigrés d'origine non européenne. En effet, en l'absence d'une base de sondage nationale permettant de les tirer aléatoirement, leur sélection s'effectue à partir d'une base d'individus dont les coordonnées ont été collectées auprès de leurs parents (individus de 2^{ème} génération non européenne interrogés dans l'enquête TeO2). Ainsi, pendant leur entretien, un module du questionnaire était prévu pour recueillir auprès des parents de 2^{ème} génération les coordonnées (téléphone, adresses postale et électronique) de leurs enfants de plus de 18 ans, cohabitant ou non avec eux.

Pour rappel, en termes de protocole, il s'agissait de collecter les FA éligibles et de les traiter comme une enquête identique dite « enquête clone » avec livraison de 4 lots de FA successifs. Une complexité de

l'interrogation des 3^{èmes} générations non européennes est qu'on ne connaît pas à l'avance le nombre d'individus de 3^{èmes} générations non européennes dont on parvient à récupérer des coordonnées pour les interroger. Des estimations sur les effectifs attendus de personnes de la 3e génération d'origine non européenne ont été réalisées à partir des données de TeO1 et de l'EEC 2017 (cf. annexe 6). À cette première incertitude vient s'ajouter une deuxième, qui est celle du taux de réponse des G3, difficilement prévisible. Cette dernière crée un risque de biais dans la constitution de l'échantillon : si l'objectif de 500 questionnaires est rapidement atteint et la collecte interrompue, ceux qui répondent vite seraient surreprésentés, alors même qu'ils ont probablement des caractéristiques (régionales, professionnelles, etc.) spécifiques. Pour limiter ce risque, il a été envisagé de créer des sous-vagues représentatives au sein de l'échantillon des enfants d'immigré(s) avec la possibilité de rééchantillonner au sein des lots collectés des 3^{èmes} générations non européennes. Enfin, une ultime contrainte de collecte s'ajoute : 1 000 FA seraient mises en collecte au maximum.

Mais au final, il n'y pas eu de sélection à faire pour les 3^{èmes} générations mises en collecte par manque d'individus dont on a récupéré les coordonnées. Au vu de la prévision du nombre total de 3^{èmes} générations non européennes qui pourraient être enquêtées, il a également été décidé :

- de mettre en collecte systématiquement tous les individus de 3^{èmes} générations non européennes pour chaque lot
- d'autoriser à enquêter un 3^{ème} génération non européen qui habite le même logement qu'un ou plusieurs 2^{ème(s)} génération(s) déjà interrogé(s)
- d'autoriser à enquêter un 1^{er}, 2^{ème} génération, 3^{ème} génération non européen qui habite le même logement.

3. Déroulement et bilan de la collecte de TeO2

3.1 Vague 1 de TeO2 : immigrés, Domiens et population générale

3.1.1 Bilan de collecte de la vague 1 : juillet-décembre 2019

La collecte de la vague 1 a eu lieu du 1^{er} juillet 2019 au 31 décembre 2019, sur 26 semaines ininterrompues, par le réseau des enquêteurs de l'Insee. Pour rappel, cette collecte concernait celle des sous-échantillons des immigrés, des natifs des DOM et de la population générale. Ainsi, 26 851 fiches-adresses (FA) ont été allouées dans le but d'obtenir 13 200 questionnaires remplis. Le bilan de la collecte de la vague 1 (cf. tableau ci-après) est très satisfaisant puisque :

- avec 15 605 entretiens réalisés, le taux de collecte national atteint 58 %, au-delà de la cible fixée en début de collecte (54 %)
- les objectifs de nombre de FA attendues par sous-échantillon ont été dépassés. Pour rappel, les objectifs étaient d'atteindre 9 000 questionnaires d'immigrés, 800 de Domiens et 3 400 de population générale. Au final, 10 287 questionnaires d'immigrés, 854 de Domiens et 4 464 de population générale (c'est-à-dire avant réallocation des questionnaires immigrés, natifs d'un DOM et descendants d'immigrés dans les groupes correspondants) ont été collectés.

Tableau 9 : Synthèse de la collecte de la vague 1 par sous-échantillon

	Nb FA attendus	Nb FA confiées	Nb FA réussies	Taux de collecte
Vague 1 dont,	13 200	26 851	15 605	58 %
Immigrés	9 000	18 277	10 287	56 %
Natifs d'un DOM	800	1 646	854	52 %
Population générale	3 400	6 928	4 464	64 %

Pour rappel, au sein du sous-échantillon des immigrés, des cibles par groupe d'origines avaient également été fixées. Ces cibles ont toutes été atteintes voire dépassées dans certains cas sauf pour le groupe « Chine » qui est nouveau parmi les origines surreprésentées du sous-échantillon des immigrés par rapport à TeO1. En effet, le nombre de FA collectées des immigrés chinois atteint 570, en-deçà du seuil des 700 FA attendues. Les enquêteurs ont par ailleurs fait remonter pendant la collecte, que la population originaire de Chine était très complexe à interroger. Ces difficultés concernaient à la fois la prise de contact mais également la passation du questionnaire en langue française. C'est pourquoi, d'une part, des approches de sensibilisation par des associations avaient été mises en place dans certaines régions ; d'autre part, le dispositif de repérage des non francophones a permis de détecter une part plus importante de non francophones d'origine chinoise⁴², soit environ 21% des 436⁴³ FA non francophones repérés (cf. tableau ci-après) et également de non francophones d'origine turque (environ 21 %). Ces personnes ainsi que d'autres personnes immigrées repérées comme non francophones ont été recontactées lors de l'enquête de rattrapage des non francophones en janvier 2020.

⁴² Les immigrés d'origine chinoise font partie de la population immigrée les plus récemment installés en France.

⁴³ 436 FA non francophones repérées : 425 FA du sous-échantillon des immigrés et 11 FA issues de celui de la population générale

Tableau 10 : Synthèse de la collecte du sous échantillon 1 (immigrés)

	Nombre FA attendus	Nombre FA confiées	Nombre FA réussies	Taux de collecte	Nombre FA NF identifiées
Immigrés, dont	9000	18277	10287	56,3%	425
Algérie	1000	2043	1240	60,7%	9
Maroc Tunisie	1000	2195	1329	60,5%	41
Afrique sahélienne	800	1719	887	51,6%	23
Afrique centrale et du golfe de Guinée	800	1907	1063	55,7%	4
Espagne, Italie, Portugal	1000	1731	1061	61,3%	15
Autres pays de l'UE 28	800	1459	831	57,0%	32
Turquie	800	1510	859	56,9%	91
Asie du Sud-Est	800	1554	950	61,1%	31
Chine	700	1430	570	39,9%	93
Pays avec de nombreux réfugiés	700	1414	728	51,5%	60
Autres pays du monde	600	1315	769	58,5%	26

Si le résultat de la collecte est positif, quelques difficultés dans le déroulement de la collecte ont nécessité dans certains cas, des adaptations dans le protocole :

1. la collecte a démarré progressivement du fait de la période de congés d'été. Toutefois, le rythme de collecte s'est ensuite intensifié jusqu'à la fin de collecte.
2. le temps de passation du questionnaire observé a été plus long qu'anticipé : 1h16 en moyenne au lieu de 1h. En effet, la structure de l'échantillon de la vague 1, qui concerne en majorité des immigrants, induit un temps moyen plus long. Les analyses de temps ont montré que la durée moyenne de passation était sensiblement différente selon les sous-échantillons. Pour la vague 2, la durée moyenne a été courte (1h04).
3. Les équipes régionales chargées de superviser la collecte ont transmis systématiquement la date et le lieu de naissance en cas de déménagement (ainsi que toutes les informations annexes collectées par l'enquêteur). Cette adaptation du protocole a été envisagée pour pallier les difficultés

de repérage uniquement par les informations nominatives des individus (noms et prénoms) et dont la qualité diffère selon leur mode de saisie (manuelle ou lecture optique)⁴⁴.

Encadré 6 : Les consignes de collecte pour la vague 1 de TeO2

En cours de collecte, des consignes ont été adressées aux Divisions enquêtes ménages (Dem) en Directions Régionales dont la liste est présentée ci-dessous :

Au cours de la 17^e semaine, les Dems ont reçu comme consigne :

- de maintenir le rythme de collecte au même niveau que celui atteint depuis début septembre 2019
- d'apporter une vigilance particulière, lorsque c'est possible, sur les FA pour lesquelles les personnes sont originaires de Chine, mais aussi de Turquie et des pays avec de nombreux réfugiés (Angola, Sri Lanka, Thaïlande, Pakistan, Haïti, Russie et ex-Yougoslavie hors UE)

Plus spécifiquement, en Île-de-France, il a été demandé d'être également vigilant sur les groupes "Algérie", "Maroc, Tunisie" et "Autres pays du monde", ainsi que sur l'échantillon population générale, pour lesquels l'allocation régionale a été divisée par deux en IDF par rapport à l'allocation proportionnelle (un bon taux de collecte sur ces groupes sous-représentés dans l'échantillon étant nécessaire pour éviter d'obtenir des poids corrigés de la non-réponse trop grands et donc perturbateurs des analyses).

Le comité de suivi du 08/11/2019 s'est accordé sur le fait que les consignes déjà diffusées aux Dem restent valables.

Le comité de suivi du 02/12/2019 s'est accordé sur le fait que les moyens de pilotage en cours de collecte étant relativement limités, il n'y avait pas de message supplémentaire à communiquer aux Dem pour le suivi de la fin de la vague 1, malgré les possibles perturbations du mouvement social de fin 2019.

3.1.2 Bilan de l'enquête de rattrapage des non francophones : janvier-août 2020

L'enquête de rattrapage des non francophones a été réalisée par l'Ined du 13 janvier au 31 août 2020. L'Insee avait préalablement transmis 436 FA repérées comme non francophones à l'Ined. L'Ined a mis en collecte 416 FA⁴⁵ et avait recruté 9 enquêteurs traducteurs pour réaliser l'enquête en 8 langues (portugais, russe, tamoul, vietnamien, anglais, arabe, chinois, turc). Pour 315 FA, l'enquêté avait renseigné ses coordonnées afin d'être recontacté. En termes de repérage et de prise de contact, de multiples modes de contact ont été mobilisés pour atteindre l'enquêté (envoi de courrier, mail, SMS, pages blanches, réseaux sociaux).

Le bilan de collecte de l'enquête de rattrapage des non francophones est également très satisfaisant puisque 229 FA ont été collectées (cf. tableau ci-après), ce qui porte le taux de collecte à 55 %, et à 72 % des FA joignables. 90 % des entretiens se sont déroulés par téléphone. La durée moyenne de passation est de 1h45⁴⁶.

⁴⁴ Ceci est une conséquence du tirage de l'échantillon hors saisie Jouve qui a été nécessaire pour disposer de suffisamment de personnes pour certaines origines.

⁴⁵ 20 FA non attribuées correspondant à des langues d'Afrique francophone

⁴⁶ Temps mesurés par les enquêteurs Ined

Tableau 11 : Synthèse de la collecte de l'enquête de rattrapage des immigrés non francophones

	Nombre de FA confiées	Nombre de FA réussies
NF dont,	416	229
Portugais	17	8
Russe	24	9
Tamoul	25	8
Vietnamien	18	10
Anglais	81	29
Arabe	58	36
Chinois	94	53
Turc	99	76

Au final sur les 26 851 FA mises en collecte, 15 834 FA⁴⁷ ont pu être collectées ce qui porte le taux de collecte de la vague 1 à 59 %, soit 5 points au-dessus du taux fixé au début de la collecte.

3.2 Vague 2 TeO2 : descendants d'immigrés et de domiens

3.2.1 Bilan de l'opération mairie : juin - août 2019

Pour rappel, l'échantillonnage des descendants a nécessité une opération spécifique de collecte dans les mairies afin de rechercher des informations sur l'origine des parents dans les bulletins de naissance des individus échantillonnés. Cette dernière opération a concerné 100 014 relevés et a mobilisé 658 agents entre juin et août 2019. Parmi eux, 13 agents de la direction générale de l'Insee et 10 de l'Ined sont intervenus en renfort pour effectuer une partie des relevés de la région Île-de-France⁴⁸.

Globalement, l'opération s'est déroulée sans difficulté dans chacune des régions. Les données ont été collectées et transmises dans les temps. Dans quelques mairies, il a fallu rappeler la procédure et les raisons pour lesquelles les agents de l'Insee étaient autorisés à réaliser ces relevés. Au final, les 100 014 relevés mairie ont pu être réalisés.

⁴⁷ 15834 FA = 15 605 FA de la vague 1 + 229 FA de l'enquête NF

⁴⁸ Le renfort a porté sur les départements de la Seine-Saint-Denis (93) et du Val-de-Marne (94). Les agents de la direction générale de l'Insee qui se sont portés volontaires ont recueilli les informations sur le lieu de naissance des parents des individus de l'échantillon TeO2 Mairie via un poste nomade et directement sur l'application poste de collecte CAPI. Les volontaires de l'Ined ont relevé les informations sur des formulaires papier fournis par l'Insee à compléter en mairie. Ces formulaires, une fois remplis, ont été envoyés à la Division enquêtes ménage de la Direction régionale Insee Île-de-France qui s'est chargée de la saisie sur le poste de gestion Capi afin d'alimenter l'échantillon TeO2 Mairie. La procédure est détaillée dans les instructions de collecte pour l'opération mairie.

Encadré 7 : Cadre juridique de l'accès aux registres des actes de naissance dans les mairies

La demande d'accès à l'état civil des personnes, qui a permis de constituer l'échantillon des descendants d'immigrés et de Domiens de l'enquête TeO2, est fondée sur l'article L. 213-3 du code du patrimoine qui permet à l'administration des archives de délivrer les autorisations de consultation de documents d'archives publiques avant l'expiration des délais de communicabilité fixés à l'article L. 213-2 du même code, en l'occurrence 75 ans en ce qui concerne les registres des actes de naissance et de mariage. En pratique, l'autorisation a été prise après avis favorable du tribunal de grande instance de Paris, qui a procédé aux vérifications nécessaires (identité des agents participant à l'opération mairie).

Pour rappel, à l'époque de TeO1 (2008), il était nécessaire d'obtenir de chaque procureur de la République l'autorisation nominative d'accès aux registres d'état civil pour les agents participant à l'opération mairie. Avec l'accord de la Chancellerie, les demandes qui sont adressées depuis 2010 au Service interministériel des Archives de France (SIAF)⁴⁹ pour l'état civil de l'ensemble du territoire national sont transmises, pour avis préalable, au tribunal de grande instance de Paris, ceci afin d'éviter l'encombrement des greffes, qui seraient sans cela sollicités chaque année par plusieurs centaines d'utilisateurs (généalogistes professionnels, chercheurs, etc.).

3.2.2 Le déroulé de la collecte de la vague 2 : janvier-novembre 2020

3.2.2.1 Un bon début de collecte, puis une collecte suspendue du fait de la crise sanitaire

La vague 2 de l'enquête TeO2 a débuté le 1^{er} janvier 2020. Pour les descendants d'immigrés et de Domiens, 23 879 fiches-adresses (FA) ont été allouées dans le but d'obtenir 12 800 questionnaires remplis⁵⁰. Le rythme en début de collecte était satisfaisant puisqu'au bout de 2 mois de collecte, le taux de collecte de la vague 2 atteignait 14 %, ce qui était supérieur à celui de la vague 1 (11 %) à l'issue de 2 mois de collecte. Au mois de mars 2020, la collecte a été interrompue en raison de la crise sanitaire. La suspension de la collecte a duré 3 mois du 16/03/2020 au 15/06/2020.

3.2.2.2 Expérimentation de la reprise en face-à-face

Suite au déconfinement, une expérimentation a été menée avec une soixantaine d'enquêteurs volontaires de 9 établissements régionaux de l'Insee⁵¹, du 15 au 30 juin 2020 afin de tester la reprise de la collecte en face-à-face. Pour cette expérimentation :

- seules les enquêtes qui n'avaient fait l'objet d'aucune prise de contact avant le confinement, étaient concernées ;
- un protocole sanitaire spécifique a été mis en place (port du masque, cartes codes placées dans un porte-vues ...) et en cas de vulnérabilité des personnes, la possibilité de passer l'enquête par téléphone a été laissée.

Le bilan étant positif, la reprise généralisée des collectes en face-à-face a été décidée par le comité de direction de l'Insee le 6 juillet 2020, pour une reprise effective le 15 juillet 2020.

⁴⁹ Pour TeO2, la procédure est simplifiée (par rapport à TeO1) : la demande est centralisée par la direction générale de l'Insee qui s'adresse à un interlocuteur unique du SIAF. Ce dernier transmet la demande au TGI de Paris qui donne l'autorisation. Dans certains cas, le SIAF était intervenu directement auprès des procureurs (Belfort, Besançon en particulier) pour les informer de la procédure.

⁵⁰ 12 800 est la somme de 800 descendants de domiens, 9 000 descendants d'immigrés et 3 000 descendants de Français nés à l'étranger.

⁵¹ Ces enquêteurs font partie du réseau de 9 divisions enquêtes ménages (Dem) : Aquitaine, Bourgogne, Provence-Alpes-Côte d'Azur, Pays de la Loire, Centre, Île-de-France, Champagne-Ardenne, Picardie, Bretagne

3.2.2.3 Reprise de la collecte et fin de collecte de la vague 2

À la reprise de la collecte en face-à-face le 15 juillet, le protocole a été adapté :

- de nouvelles lettres avis reprenant le protocole sanitaire mis en œuvre sont mises à disposition
- l'argumentaire ainsi que la fiche de contact pour le contact par téléphone ont été adaptés
- les cartes-codes en français sont insérées dans un porte-vues, les cartes-codes en langues étrangères ont été montrées sur l'écran de l'ordinateur, en version numérique⁵²
- en cas de vulnérabilité à la Covid-19, l'entretien en face-à-face devait se tenir en septembre 2020 si possible, sinon, il devait se réaliser par téléphone

Le calendrier de collecte initialement prévu a été décalé. La fin de la collecte de la vague 2 a été repoussée de 2 mois, ce qui a amené à clôturer cette collecte à la fin novembre 2020. Le détail des consignes est présenté dans l'encadré ci-après.

Encadré 8 : Chronologie et consignes de collecte pour la vague 2 de TeO2

En début de collecte, les Dems ont reçu comme consigne :

- de traiter en priorité les descendants d'immigrés,
- de suspendre les collectes dans l'Oise et dans une commune du Morbihan à compter du 02/03/2020, en raison de l'épidémie de Covid-19.
- concernant l'opération 3G, il est convenu qu'en vague 2 les enquêteurs débutent par les FA d'origine non européenne (tout sauf : Espagne, Italie, Portugal, autres pays de l'UE 28 et Dom).

Le 16 mars 2020, suite aux consignes relatives à la crise sanitaire l'enquête TeO2 a été suspendue jusqu'au 15 juin 2020, une expérimentation de la reprise de TeO2 a débuté pour 15 jours.

Le 15 juillet 2020 : reprise de la collecte avec comme consignes de :

- réaliser en priorité les entretiens en face-à-face avec la possibilité de faire les entretiens par téléphone
- bien collecter les numéros de téléphone des enquêtés,
- commencer par traiter les FA non européennes (pour repérer les G3 assez tôt)
- alerter en cas de difficultés pour obtenir les coordonnées des G3

Le 1^{er} septembre 2020, de nouvelles consignes sont données :

- réaliser en priorité les entretiens en face-à-face en suivant le protocole
- les FA mises en attente depuis mi-juillet et toutes les FA confiées à des enquêteurs vulnérables peuvent être réalisées par téléphone.

Le 7 octobre 2020, suite à la décision du Comité de pilotage TeO2 du 30/09, il a été décidé :

- pour l'ensemble des établissements :
 - d'élargir la possibilité de recourir au téléphone pour TeO2 pour l'ensemble des FA non affectées, ou pour lesquelles l'enquêteur n'est pas en état de réaliser l'entretien.
 - de faire l'enquête par téléphone si l'enquêteur constate un déménagement hors zone de collecte.
- pour la DR d'Île-de-France, les directions régionales (DR) délégataires et la DR de PACA : il est autorisé de réaliser la collecte par téléphone pour les FA déjà affectées situées en Ile de France ou sur Provence-Alpes-Côte d'Azur. Pas de bascule de la collecte par téléphone avec re-répartition des FA entre DR. Tout au plus cela est possible au cas par cas au sein d'une même DR.
- pour faciliter la passation des enquêtes TeO par téléphone, les cartes codes ont été mises en ligne sur le site de l'Insee afin que l'enquêteur puisse proposer à l'enquêté de s'y référer s'il en est d'accord, et s'il dispose d'internet.

Le 26 octobre 2020, une consigne a été passée pour prioriser les FA G3 (lots 1, 2 et 3) avant d'ensuite reprendre la collecte des G2.

Le 02 novembre 2020, suite à l'annonce du 2^{ème} confinement débutant le 30/10,

- le repérage de toutes les enquêtes, dont TeO2, est suspendu jusqu'à nouvel ordre en raison de ce second confinement.
- la collecte de l'enquête TeO2 se fera exclusivement par téléphone.

⁵² Dans la pratique cette consigne a été peu ou pas appliquée car les individus enquêtés en vague 2 sont des descendants. Le problème de compréhension de la langue ne s'est pas posé pour ces individus.

- priorisation de l'enquête TeO2 devant l'enquête Histoire de Vie et Patrimoine pour la fin de collecte

Le 04 novembre 2020, pour l'ensemble des enquêtes habituellement en face à face, la consigne est donnée aux enquêteurs d'envoyer des SMS et mails avis aux personnes enquêtées en l'absence de repérage. Des modèles de SMS et mails avis sont transmis à cet effet.

Le 10 novembre 2020, le repérage est désormais possible pour les enquêteurs volontaires. La collecte de l'enquête TeO2 reste exclusivement par téléphone.

30 novembre 2020 : clôture de la collecte de la vague 2 de TeO2. Il a été décidé de prolonger la collecte des lots 1, 2, 3 des G3 jusqu'au 15 décembre 2020. Un 4^{ème} lot est constitué et destiné à être collecté en janvier 2021.

3.2.3 Bilan de la participation à la vague 2

Le bilan de collecte de la vague 2 (cf. tableau ci-après) est satisfaisant compte-tenu des difficultés de terrain et des incertitudes liées à la crise sanitaire :

- avec 11 721 entretiens réalisés, le taux de collecte national atteint 49 %. Ce taux est inférieur à celui de la vague 1 (58 %) et en deçà de la cible fixée en début de collecte (54 %)
- les objectifs par sous-échantillon n'ont pas été dépassés mais s'en approchent. Pour rappel, les objectifs étaient d'atteindre 9 000 questionnaires de descendants d'immigrés, 3 000 d'enfants de français nés à l'étranger, 800 de descendants de Domiens. Au final, 11 143 questionnaires de descendants de parents nés à l'étranger, 578 de descendants de Domiens ont été collectés.

Tableau 12 : Synthèse de la collecte de la vague 2 par sous-échantillon

	Nombre FA attendus	Nombre FA confiées	Nombre FA réussies	Taux de collecte
Vague 2 dont,	12 800	23 879	11 721	49 %
Descendants de personnes nées à l'étranger	12 000	22 675	11 143	49 %
Descendants de Domiens	800	1 204	578	48 %

Le taux de collecte du sous-échantillon des descendants de personnes nées à l'étranger par groupe d'origines est donné par le tableau ci-après.

Tableau 13 : Synthèse de la collecte du sous échantillon 3
(descendants de personnes nées à l'étranger)

	Nombre FA attendus	Nombre FA confiées	Nombre FA réussies	Taux de collecte
Descendants de personnes nées à l'étranger, dont	9000	22675	11149	49,2%
Algérie	1600	5161	2620	50,8%
Maroc Tunisie	1200	3075	1529	49,7%
Afrique sahélienne	800	1650	688	41,7%
Afrique centrale et du golfe de Guinée	800	1730	665	38,4%
Espagne, Italie	800	1180	686	58,1%
Portugal	800	1113	623	56%
Autres pays de l'UE 28	800	1365	768	56,3%
Turquie	800	1213	612	50,5%
Asie du Sud-Est	800	1647	847	51,4%
Autres pays du monde	600	2235	1155	51,7%
Origines incertaines - Afrique centrale et du golfe de Guinée		677	241	35,6%
Origines incertaines - Origines surreprésentées		1151	497	43,2%
Origines incertaines - Origines sous-représentées		478	218	45,6%

Rappelons que parmi les FA collectées, une partie sont des FA d'individus descendants d'immigrés ; l'autre partie étant des FA d'individus de descendants de français nés à l'étranger. Afin d'aider au pilotage de la collecte, le nombre d'enfants d'immigrés attendus par groupe d'origines a été estimé à partir du nombre de FA réussies et du taux d'enfants d'immigrés parmi les enfants de parents nés à l'étranger, repris de l'enquête TeO1 (cf. tableau ci-après).

Tableau 14 : Nombre de descendants d'immigrés attendus par groupe d'origines

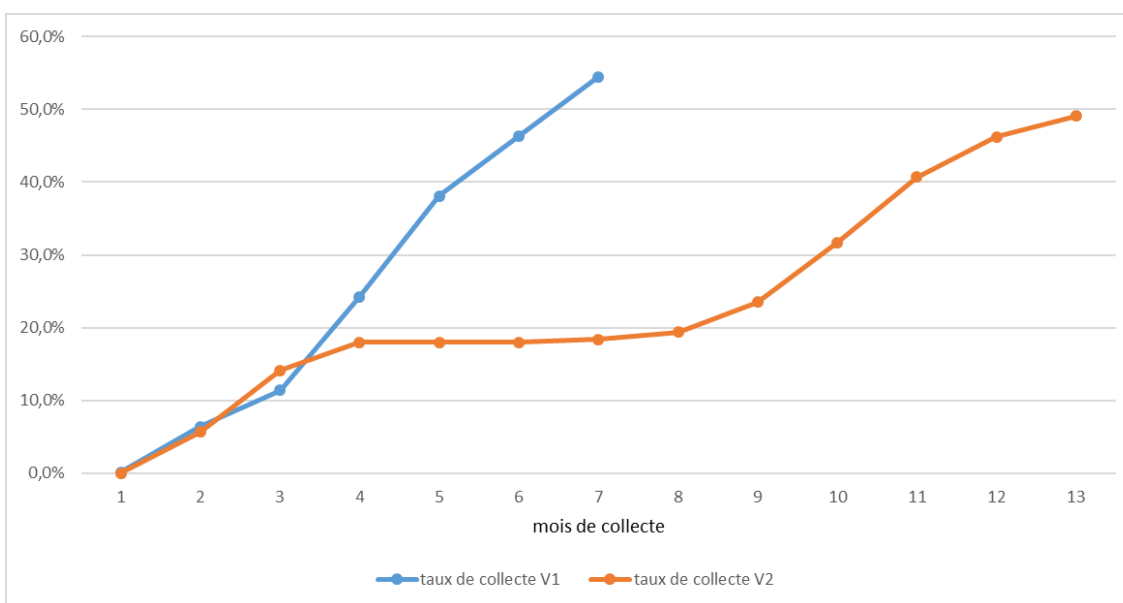
Groupe d'origines	Nombre FA attendus	Nombre FA confiées	Taux de collecte	Taux d'enfants d'immigrés anticipé	Nombre de descendants d'immigrés attendus	Taux nombre de descendants d'immigrés attendus / nombre FA
Algérie	1600	5161	50,8%	49%	1284	80,2%
Maroc Tunisie	1200	3075	49,7%	59%	902	75,2%
Afrique sahélienne	800	1650	41,7%	75%	516 ⁵³	64,5%
Afrique centrale et du golfe de Guinée	800	1730	38,4%	69%	459	78,1%
Espagne, Italie	800	1180	58,1%	97%	665	83,1%
Portugal	800	1113	56%	99%	617	77,1%
Autres pays de l'UE 28	800	1365	56,3%	87%	668	83,5%
Turquie	800	1213	50,5%	99%	606	75,8%
Asie du Sud-Est	800	1647	51,4%	73%	618	77,2%
Autres pays du monde	600	2235	51,7%	37%	427	71,2%

Le graphique 1 compare l'évolution du taux de collecte des 2 vagues selon les mois de collecte. Il met en évidence :

- un meilleur rythme de début de collecte de la vague 2 comparativement à celui de la vague 1 mais une progression quasi continue de la collecte de la vague 1 jusqu'à la clôture
- l'arrêt de la collecte de la vague 2 au bout du 3^{ème} mois de collecte et pendant 3 mois
- la reprise lente de la vague 2 suite au déconfinement et qui a eu lieu durant les congés d'été
- la progression plus forte du rythme de collecte de la vague 2 jusqu'à la clôture, liée à un agenda de collecte plus léger et à une priorisation de la collecte de TeO2.

⁵³ Ce chiffre cumule le nombre de descendants d'immigrés attendus du groupe « Afrique centrale et golfe de Guinée » et « Origines incertaines - Afrique centrale et du golfe de Guinée »

Graphique 1 : Comparaison de l'évolution du taux de collecte TeO2 – vague 1 vs vague 2



3.2.4 Bilan de la collecte par origine et statut migratoire

Le tableau suivant présente le bilan de la collecte par origine et statut migratoire, en comparant les objectifs de collecte initiaux aux effectifs de répondants réalisés. Contrairement aux tableaux précédents, tous les répondants ne sont pas nécessairement issus des FA confiées dans chaque sous-échantillon. Par exemple, les répondants immigrés originaires d'Algérie peuvent avoir été collectés dans le sous-échantillon de la population générale (05) ou même à partir d'un autre sous-échantillon dans lequel ils ne devaient à priori pas apparaître (par exemple, celui des Domiens, dans le cas d'une erreur de remplissage du bulletin du recensement). De même, il s'agit des répondants immigrés au sens de TeO, c'est-à-dire en incluant les réintégré, alors que les sous-échantillons ont été constitués à partir de la définition usuelle.

La collecte globale dépasse l'objectif initial (27 181 répondants contre 26 000 visés, soit 105%), notamment en raison du bon taux de collecte en vague 1 (58% contre 54% attendu). Si la collecte a été supérieure aux attentes pour les immigrés (116% de réussite), sauf ceux originaires de Chine (85%), et pour la population majoritaire (109%), cette dernière est en revanche inférieure aux cibles attendues pour les descendants de 2^e génération immigrée (92%) et Domiens (87%) en raison de la situation sanitaire. Comme pour TeO1, le déficit de répondants par rapport aux cibles est le plus fort pour les descendants d'immigrés originaires d'Afrique subsaharienne (<70%).

Tableau : Effectifs cibles et réalisés par origine et statut migratoire

Origine géographique	Effectif cible	Nombre de répondants	% de réussite	Effectif cible	Nombre de répondants	% de réussite
	G1			G2		
Immigrés G1 ou Descendants G2	9 000	10 396	116	9 000	8 294	92
Algérie	1 000	1 231	123	1 600	1 500	94
Maroc, Tunisie	1 000	1 382	138	1 200	1 142	95
Afrique sahélienne	800	836	105	800	554	69
Afrique centrale et du golfe de Guinée	800	1 066	133	800	533	67
Espagne, Italie	1 000	1 099	110	800	913	114
Portugal				800	789	99
Autres pays de l'UE 28	800	856	107	800	681	85
Turquie	800	923	115	800	674	84
Asie du Sud-Est	800	901	113	800	745	93
Chine	700	593	85			
Pays avec de nombreux réfugiés	700	721	103			
Autres pays du monde	600	788	131	600	763	127
Dominiens (G1) et descendants (G2)	800	849	106	800	692	87
Population majoritaire	6 400	6 950	109			
dont enfants de Français nés à l'étranger	3 000	3 123	104			
dont G3		906				
Total	26 000	27 181	105			

3.3 L'enquête expérimentale TeO2-G3

3.3.1 Déroulement de l'opération TeO2-G3 : mars 2020 - janvier 2021

Pour rappel, la constitution de l'échantillon des 3^{èmes} générations non européennes découlant directement de la collecte des descendants d'immigrés, il était nécessaire de démarrer la collecte de l'enquête TeO2-G3 après celle de la vague 2 avec quelques mois de décalage afin d'avoir le temps de recueillir suffisamment de coordonnées des 3G non européens. La constitution de 4 lots de collecte successifs était prévue.

La collecte de l'enquête TeO2-G3 a débuté le 10 mars 2020 juste avant la suspension de la collecte le 16 mars en raison de la crise sanitaire. La collecte a repris le 15 juillet 2020. Suite à la reprise, le taux de collecte des FA est resté faible : au mois de septembre, le nombre de FA collectées ne dépassait pas 50. Le nombre de lots a d'abord été revu à 3 compte-tenu du faible volume de G3 éligibles. Toutefois un 4^{ème} lot pouvait être constitué (entre mi-octobre et fin novembre) sans pouvoir être enquêté puisque la fin de collecte était prévue à la fin novembre. En effet, comme pour la vague 2, la collecte a été décalée à la fin novembre 2020. Le 26 octobre 2020, une consigne a permis de prioriser les FA G3 (lots 1, 2 et 3) avant de reprendre la collecte des descendants. Avant la clôture de la collecte de la vague 2 (30 novembre 2020), il a été décidé de :

1. prolonger la collecte des lots 1, 2, 3 des G3 jusqu'au 15 décembre 2020 et
2. d'interroger le 4^{ème} lot au mois de janvier 2021.

Encadré 9 : Chronologie et consignes de collecte pour l'opération TeO2-G3

En début de collecte, les Dems ont reçu comme consigne :

- qu'en vague 2 les enquêteurs débutent par les FA d'origine non-européenne (tout sauf : Espagne, Italie, Portugal, autres pays de l'UE 28 et Dom)

Le 16 mars 2020, suite aux consignes relatives à la crise sanitaire, l'enquête TeO2 a été suspendue

Le 15 juillet 2020 : reprise de la collecte avec comme consignes :

- de commencer par traiter les FA non-européennes (pour repérer les G3 assez tôt)
- d'alerter en cas de difficultés pour obtenir les coordonnées des G3

Le 26 octobre 2020, une consigne était de prioriser les FA G3 (lots 1, 2 et 3) avant de reprendre la collecte des descendants.

Mi-novembre 2020 : il a été décidé

- d'interroger le lot 4 des G3 au mois de janvier 2021.
- de prolonger la collecte des lots 1, 2, 3 des G3 jusqu'au 15 décembre 2020.

3.3.2 Bilan de la participation à l'enquête TeO2-G3

Le bilan de collecte de l'enquête TeO2-G3 est mitigé. 242 FA ont été collectées sur un objectif initial de 500. Le volume d'individus G3 non européens repérés par le module G3 du questionnaire a atteint presque 1 000 FA, chiffre conforme aux estimations réalisées avant la collecte, à partir des données de TeO1 (environ 1 400) et de l'Enquête emploi en continu - EEC 2017 (environ 1 200). L'objectif de 500 FA a été établi à partir d'hypothèses fortes et particulièrement optimistes, à savoir un taux de collecte proche de l'enquête principale (50 %) et l'absence de déperdition lors de la phase d'obtention des coordonnées des G3 par leur parent enquêté.

En pratique, cette déperdition a été d'environ 2/3 du stock initial de FA éligibles. Elle s'explique principalement par le refus des parents de communiquer les coordonnées de leurs enfants⁵⁴. Le nombre de FA « enquêtées » et mises en collecte, c'est-à-dire celles pour lesquelles les coordonnées étaient disponibles, a ainsi été de 362 FA (cf. tableau ci-après).

Le taux de FA enquêtées, défini comme le rapport du nombre de FA confiées au nombre de FA de G3 éligibles, est relativement proche entre chaque lot de collecte (43% pour le lot 1, puis 1/3 pour les lots suivants). Sur les 362 FA confiées, 242 ont été réalisées, ce qui porte le taux de collecte à 67%. Ce taux est supérieur à celui de l'enquête principale et il découle principalement de la consigne de priorisation. Les taux de collecte entre chaque lot sont relativement homogènes. Si ces taux sont au final satisfaisants, une grosse incertitude a pesé lors de la collecte. Cela est dû :

- aux difficultés de terrain liées à la crise sanitaire
- au fait de cumuler les collectes de la vague 2 et celle de l'enquête TeO2-G3

⁵⁴ Question COORDENF du questionnaire - Pour pouvoir le/les contacter, je dois noter leur numéro de téléphone (portable, si possible), leur adresse email et leur adresse postale. Les coordonnées de vos enfants seront conservées dans le seul but de réaliser l'enquête. Elles seront ensuite détruites. Acceptez-vous de me communiquer les coordonnées de votre/vos enfants ?

À un mois de la fin de collecte initialement prévue, le nombre de FA réussies atteignait seulement 42 FA. Le rattrapage a été possible grâce aux consignes de priorisation du mois de novembre 2020.

Tableau 15 : Synthèse de la collecte de l'enquête TeO2-G3

	LOT1	LOT 2	LOT 3	LOT 4	TOTAL
Période constitution des lots	01/01 - 24/02	24/02 - 07/09	07/09 - 19/10	19/10 - 30/11	
Durée constitution des lots (semaines)	8	13	6	6	
Date de collecte	10/03 - 30/11	28/09 - 30/11	29/10 - 30/11	02/01 - 31/01	Du 10/03/2020 au 31/01/2021
Durée de collecte (semaines)	40	9	5	4	58
Nombre G3 éligibles	377	213	206	195	991
Nombre G3 confiés	163	68	66	65	362
Taux enquêtables	43 %	32 %	32 %	33 %	37 %
Nombre G3 réussies	103	49	41	49	242
Taux de collecte	63 %	72 %	62 %	75 %	67 %

3.4 Retour sur certaines spécificités de la collecte de l'enquête TeO2

Cette partie aborde le bilan de la collecte sous trois aspects pour lesquels une vigilance et une surveillance ont été nécessaires tout au long de la collecte. Le point 3.4.1 porte sur la durée de passation du questionnaire. Le bilan sur le suivi des déménagements est décrit dans le point 3.4.2. Enfin, le point 3.4.3 présente brièvement la répartition géographique de la collecte.

3.4.1 Des temps de passation du questionnaire variables

La durée moyenne de passation de l'enquête a été de 1h10. Cette estimation ainsi que les suivantes ont été réalisées après retrait des observations non fiables et des temps extrêmes⁵⁵.

Plusieurs analyses des temps ont été menées lors de la vague 1 suite à des retours terrain signalant que la durée d'entretien était supérieure à 1h. La première estimation a été réalisée début septembre et a mis en évidence une durée moyenne de passation de 1h15. Cela a amené à réviser la charge de collecte allouée à la passation du questionnaire. Les autres estimations réalisées pendant la vague 1 ont confirmé cette durée moyenne.

⁵⁵ Calcul réalisé sur l'ensemble des questionnaires de la vague 1 et de la vague 2. Retrait des observations non fiables : les valeurs négatives sont retirées, de même que les FA réalisées sur plusieurs jours car la mesure du temps n'est pas fiable. Retrait des temps extrêmes : pour limiter l'effet des points aberrants, on écarte de notre analyse les 5% de temps extrêmes, c'est-à-dire les durées inférieures au 2,5e percentile ou supérieures au 97,5e percentile.

En fin de vague 1, une dernière estimation⁵⁶ a été réalisée sur l'ensemble des entretiens collectés durant cette vague. La durée moyenne de passation de la vague 1 s'établit à 1h16 avec des temps variables selon les échantillons (cf. tableau 18):

- 1h21 en moyenne pour le sous-échantillon des immigrés,
- 1h14 en moyenne pour le sous-échantillon des Domiens,
- 1h06 en moyenne pour le sous-échantillon de la population générale

Sans surprise le temps moyen de passation du sous échantillon des immigrés est supérieur à celui des autres sous-échantillons et peut s'expliquer pour plusieurs raisons : entretiens en langues étrangères, différences de parcours, trajectoires pour certaines populations, etc.

La durée moyenne de passation de la vague 2 est de 1h04⁵⁷, inférieure à celle de la vague 1 avec des temps homogènes selon les échantillons (cf. tableau ci-après):

- 1h04 en moyenne pour le sous-échantillon des descendants d'immigrés,
- 1h02 en moyenne pour le sous-échantillon des descendants de Domiens

Tableau 16 : Temps de passation du questionnaire par vague de collecte et sous échantillon

	Temps moyens	Temps médian	Percentile 2,5	Percentile 97,5
Vague 1, dont	1h16	1h07	0h34	4h52
sous-échantillon immigrés	1h21	1h12	0h40	3h25
sous-échantillon Domiens	1h14	1h04	0h38	2h56
sous-échantillon population générale	1h06	0h57	0h37	2h47
Vague 2, dont	1h04	0h57	0h34	2h27
sous-échantillon descendants	1h04	0h57	0h34	2h28
sous-échantillon descendants de	1h02	0h57	0h34	2h13

3.4.2 Bilan sur le suivi des personnes ayant déménagé

Pour rappel, l'enquête TeO est une enquête auprès d'individus identifiés. Elle présente donc une difficulté supplémentaire par rapport aux enquêtes auprès des ménages car il ne suffit pas de repérer un logement ni même un ménage mais d'enquêter un individu en particulier. Le protocole mis en place pour suivre les personnes ayant déménagé est décrit dans la partie 1.4.1. Cette partie présente le bilan de collecte sur les déménagements.

Comme pour TeO1, la mobilité des personnes échantillonnées s'est avérée relativement forte. Entre l'enquête annuelle de recensement 2018, qui a servi de base de sondage et la collecte l'enquête TeO2 (juillet 2019 – novembre 2020), 27 % des individus de l'échantillon avaient changé d'adresse. Ce pourcentage est légèrement plus élevé au sein de la vague 2 (29 %) par rapport à celle de la vague 1 (25 %) en raison du décalage dans le temps des 2 vagues (cf. tableau ci-après).

⁵⁶ Après retrait des observations non fiables et des temps extrêmes repérés en vague 1

⁵⁷ Après retrait des observations non fiables et des temps extrêmes repérés en vague 2

Tableau 17 : Synthèse déménagements collecte TeO2

	FA confiées	déménagements	Taux de déménagement
Ensemble collecte dont,	50 730	13 700	27 %
Vague 1, dont	26 851	6 758	25,2 %
sous-échantillon immigrés	18277	4577	25%
sous-échantillon Domiens	1646	515	31,3
sous-échantillon population générale	6928	1666	24
Vague 2, dont	23 879	6 942	29,1 %
sous-échantillon descendants d'immigrés	22675	6570	29%
sous-échantillon descendants de Domiens	1204	372	30,9%

Au final, le protocole prévu pour le suivi des adresses n'a permis de retrouver l'adresse que dans environ un cas sur quatre, ce qui est inférieur au taux de récupération d'adresses de TeO1 qui était de 50 %. Lorsque l'adresse a été retrouvée, le taux de collecte atteint environ 80 % (cf. tableau ci-après).

Tableau 18 : Synthèse suivi des déménagements collecte TeO2

	déménagements	adresses retrouvées	retrouvées/déménagements	FA réalisées	réalisées/retrouvées
Ensemble collecte dont,	13 700	3 791	27,7 %	3 067	81 %
Vague 1, dont	6 758	1 743	25,8 %	1 422	82 %
sous-échantillon immigrés	4577	1102	24,1%	893	81%
sous-échantillon Domiens	515	119	23,1%	92	77,3%
sous-échantillon population générale	1666	522	31,3%	837	83,7%
Vague 2, dont	6 942	2 048	29,5 %	1 645	80 %
sous-échantillon descendants d'immigrés	6570	1942	29,6%	1552	77,9%
sous-échantillon descendants de Domiens	372	106	28,5%	93	87,7%

Comme pour beaucoup d'enquêtes individuelles, la mobilité des personnes a constitué un facteur important de déperdition d'entretiens. Parmi les personnes n'ayant pas changé d'adresse, le taux de collecte dans le champ (part du nombre de FA réalisées rapporté au nombre de FA dans le champ) atteint 66 % (cf. tableau ci-après). Il atteint seulement 24 % pour les personnes ayant déménagé.

Tableau 19 : Taux de collecte selon la mobilité de l'enquêté

	FA confiées	FA hors champ	FA dans le champ	FA réalisées	Taux de collecte dans le champ
Ensemble collecte dont,	50 730	1 205	49 525	27 211	55 %
aucun déménagement	37 030	488	36 542	24 144	66 %
présence d'un déménagement	13 700	717	12 983	3 067	24 %

3.4.3 Répartition géographique de l'échantillon et des répondants

La région Île-de-France⁵⁸ concentre environ un tiers de l'échantillon de collecte. L'échantillon des trois régions Auvergne-Rhône-Alpes, Occitanie et PACA représente le deuxième tiers. Le dernier tiers des FA est réparti sur les autres régions (cf. tableau ci-après).

Tableau 20 : Répartition du nombre de FA mise en collecte en pourcentage

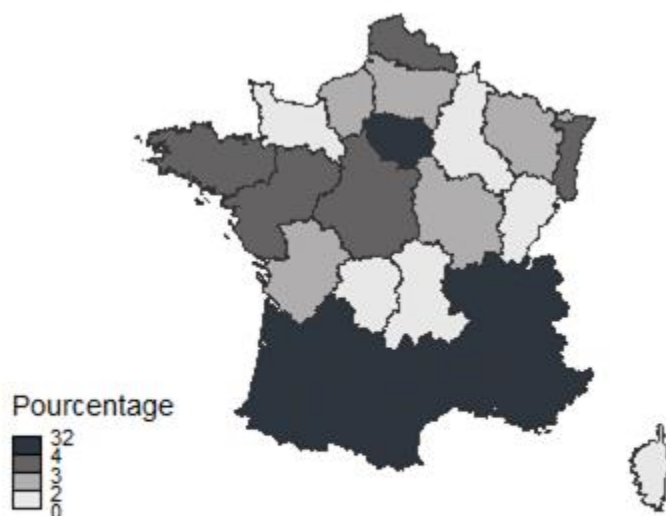
Nom Région	Effectif (%)
Île-de-France	31,77
Auvergne-Rhône-Alpes	12,26
Occitanie	10,17
PACA	9,09
Grand Est	8,16
Nouvelle-Aquitaine	6,98
Hauts-de-France	5,96
Centre-Val de Loire	3,52
Pays de la Loire	3,28
Bourgogne-Franche-Comté	2,97
Bretagne	2,92
Normandie	2,60
Corse	0,33
Ensemble	100,00

Pour la suite, nous commenterons les résultats à travers le découpage géographique des anciennes régions plus proches du découpage géographique de la collecte⁵⁹,

⁵⁸ Région Île-de-France au sens administratif i.e. composée des 8 départements : Paris, la Seine-et-Marne, les Yvelines, l'Essonne, les Hauts-de-Seine, la Seine-Saint-Denis, le Val-de-Marne et le Val-d'Oise

⁵⁹ Le réseau enquêteurs est basé sur les régions de gestion (RGES) qui est un découpage proche des anciennes régions modulo quelques ajustements pour certains régions délégataires : Centre, Champagne Ardennes, Haute Normandie. Ces régions ont en plus à leur charge un département de l'Île de France: 91 pour Centre, 77 Champagne Ardennes, 95 Haute Normandie. Le but est de répartir la charge de collecte de l'Île de France.

Carte 1 : nombre de FA confiées par région



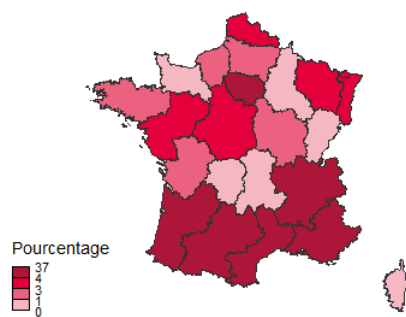
Ensemble

La répartition du nombre de FA mises en collecte s'explique par la structure particulière de l'échantillon de l'enquête. Plus de 80 % des FA ont été allouées au sous-échantillon des immigrés et des descendants d'immigrés. La population des immigrés et des descendants est fortement concentrée dans les zones urbaines et particulièrement dans la région Île-de-France. L'Île-de-France représente 37 % de l'ensemble des FA du sous-échantillon des immigrés et 32 % de celui des descendants d'immigrés contre 14 % de celui du sous échantillon de la population générale⁶⁰. Pour rappel, et à titre de comparaison, 19 % de la population de France métropolitaine vit en Île-de-France en 2018⁶¹. La carte 2 présente la répartition géographique du nombre de FA confiées pour chaque sous-échantillon. Pour l'ensemble des sous-échantillons, les zones de forte concentration sont les régions Île-de-France, Rhône-Alpes, PACA. La répartition par région du sous-échantillon des descendants d'immigrés est proche de celle des immigrés mais légèrement moins concentrée. Les zones de concentration sont plus dispersées pour le sous-échantillon de la population générale.

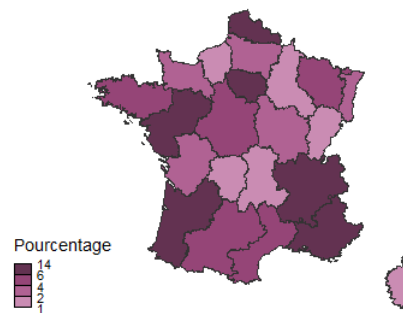
⁶⁰ A noter que le sous-échantillon de la population générale comprend également des individus immigrés, domiens, descendants d'immigrés et de domiens. En excluant ces individus du champ, le taux de l'Ile de France serait inférieur à 14%.

⁶¹ Selon les populations légales au 1^{er} janvier 2018 (cf. Rannou-Heim Ch., « *Population légale de l'Île-de-France - 12 213 447 Habitants au 1er janvier 2018* », Insee Flash Île-de-France n° 52, décembre 2020.

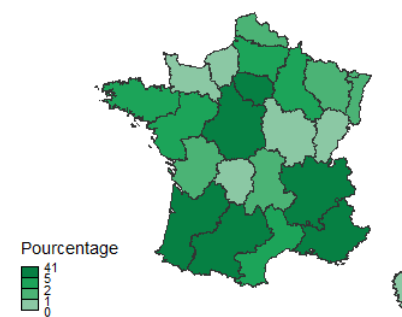
Carte 2 : pourcentage de FA confiées par région



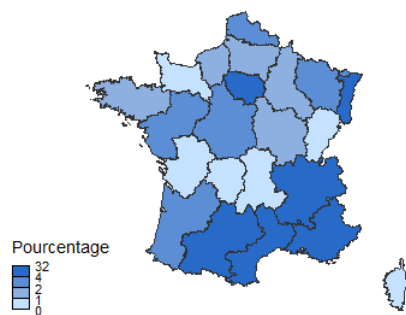
sous échantillon immigrés



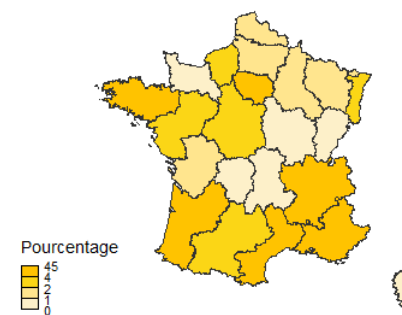
sous échantillon population générale



sous échantillon domiens



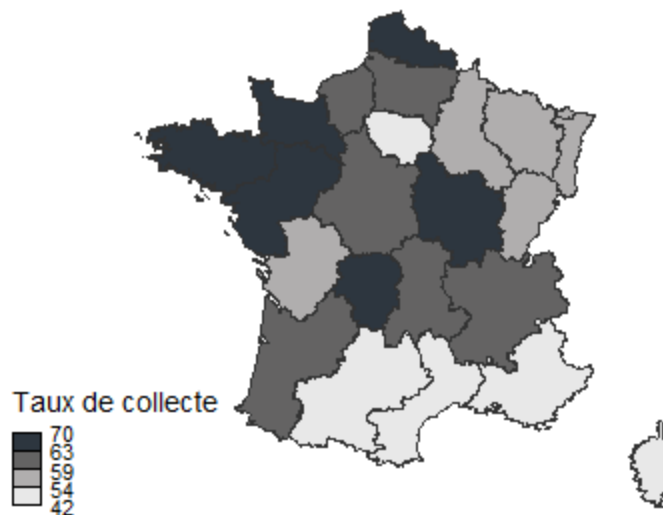
sous échantillon descendants d'immigrés



sous échantillon descendants de domiens

Le taux de collecte⁶² varie entre 40 % et 70 % selon les régions (cf. carte 3). D'une part, la Bretagne, les Pays-de-la-Loire, la Bourgogne, le Nord-Pas-de-Calais, le Limousin font partie des régions dont le taux de collecte est le plus élevé, D'autre part, l'Île-de-France, Midi-Pyrénées, PACA, le Languedoc-Roussillon, la Corse font partie des régions dont le taux de collecte est le plus faible. Le taux de collecte a été bien anticipé car les régions à faible taux de collecte sont celles pour lesquelles la part de FA⁶³ dans l'ensemble des régions est le plus élevé.

Carte 3 : Taux de collecte par région



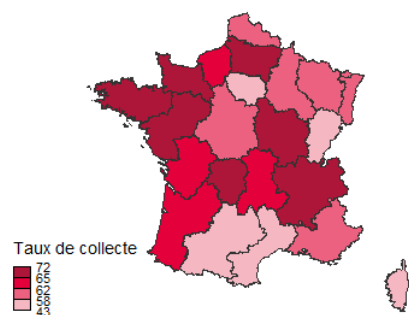
Ensemble

Cette répartition est plus ou moins conservée selon les sous-échantillons (cf. carte 4). La Bretagne, les Pays-de-la-Loire, la région Centre figurent parmi les régions dont les taux de collecte sont les meilleurs, et ce pour quasiment tous les sous-échantillons. À l'inverse, l'Île-de-France, Midi-Pyrénées, le Languedoc-Roussillon disposent des taux de collecte les plus bas et ce pour pratiquement tous les sous-échantillons. Le sous-échantillon des descendants présente de bons taux de collecte sur la diagonale « nord-ouest - sud-est ».

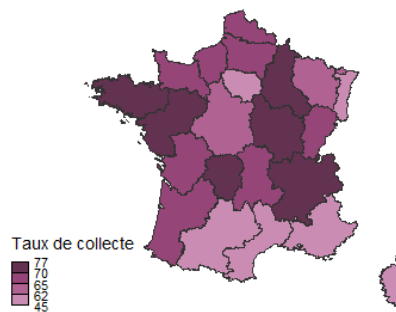
⁶² Pour rappel, le taux de collecte est défini comme le pourcentage des enquêtes réalisées par rapport à l'ensemble des fiches-adresses confiées. Ici il n'intègre pas les enquêtes partielles dans les enquêtes réalisées et recouvre uniquement les questionnaires qui peuvent être utilisés pour les études.

⁶³ Il s'agit du rapport entre le nombre de FA de la région et le nombre total de FA sur l'ensemble des régions de France métropolitaine.

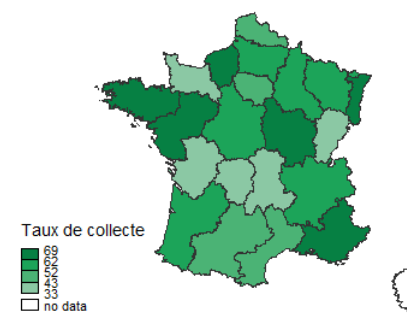
Carte 4 : Taux de collecte par région et sous-échantillon



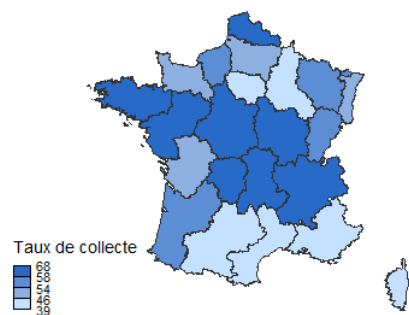
sous échantillon immigrés



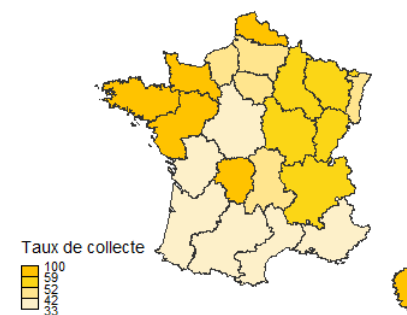
sous échantillon population générale



sous échantillon domiens



sous échantillon descendants d'immigrés



sous échantillon descendants de domiens

4. Pondération et redressement de l'échantillon TeO2⁶⁴

La collecte de l'enquête TeO2 s'est traduite par une déperdition d'environ 44% de l'échantillon mis en collecte initialement. Afin de maintenir la représentativité de l'échantillon final obtenu (27 181 observations), des travaux méthodologiques de redressements de l'échantillon ont été menés. L'objectif de cette partie est de présenter la méthodologie générale de ces redressements.

4.1 Définition de la non-réponse totale

Parmi les FA effectivement mises en collecte pour l'enquête TeO2⁶⁵, 2% étaient hors champ et 44% étaient classées non-répondantes. Cela comprend pour l'essentiel des personnes impossibles à joindre, mais aussi des refus de réponses complets, des FA non traitées ou dont l'adresse n'est pas connue et certains cas où l'entretien a été accepté mais l'enquêté l'a stoppé avant la fin.

Au final, ce sont :

- 21 781 questionnaires individus pour l'ensemble de la collecte (soit 54% des FA mises en collecte), qui sont exploitables et pour lesquels une pondération individu a été calculée soit :
 - 15 469 questionnaires individus pour la collecte de la vague 1 ;
 - 11 712 questionnaires individus pour la collecte de la vague 2

Tableau 21 : Répartition des FA selon le statut de réponse

Statut de réponse	Vague 1	Vague 2	Ensemble
Hors champ	758	447	1205
Non-réponse	10 624	11 609	22 233
Réponse	15 469	11 712	27 181

Par la suite, les FA hors champ ne participent pas au modèle de correction de la non-réponse.

4.2 Chaîne de redressements de TeO2 (hors calage)

Cette partie synthétise les redressements de l'enquête TeO2 en explicitant les traitements mis en œuvre pour chaque étape avant le calage qui sera abordé dans la partie 4.5.

Rappelons ici que l'enquête TeO2 est réalisée auprès d'individus et concerne cinq sous-échantillons distincts :

1. un sous-échantillon (01) de personnes immigrées ;
2. un sous-échantillon (02) de personnes originaires des DOM ;

⁶⁴ Cette partie aborde le redressement de l'enquête TeO2. La description du redressement de l'enquête TeO2-G3 fait l'objet d'un autre document de travail à paraître.

⁶⁵ Parmi les FA mises en collecte, 111 ont été retirées en cours de collecte de la vague 2 en raison d'un recrutement infructueux. L'échantillon a fait l'objet d'une sous allocation avec transfert de poids sur les autres FA. Par conséquent, le statut de réponse a été calculé sur l'ensemble de la collecte moins ces 111 FA. Par la suite, ces FA ne participent donc pas au modèle de correction de la non-réponse.

3. un sous-échantillon (03) de descendants d'immigrés ;
4. un sous-échantillon (04) de descendants de Domiens ;
5. un sous-échantillon (05) en population générale.

Par la suite, par souci de simplification, les individus des sous-échantillons (01) et (02) seront appelés individus de première génération (G1). De même, les individus des sous-échantillons (03) et (04) seront appelés individus de deuxième génération (G2). Tous les autres i.e. les individus qui ne sont ni des G1 ni des G2 seront appelés G0.

Pour rappel, les poids bruts de l'échantillon initial sont le produit de plusieurs poids construits dans plusieurs étapes de tirage et décrites dans la partie 2.2.3 pour le tirage des G1 et la population générale et dans la partie 2.3.8 pour le tirage des G2.

L'enquête pose 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. Les redressements de l'enquête (hors calage) TeO2 sont appliqués sur 3 groupes distincts dont la constitution est relative aux enjeux de diffusion de l'enquête :

1. le groupe des immigrés et des domiens (G1)
2. le groupe des descendants d'immigrés et de Domiens(G2)
3. le groupe des « autres individus »(G0)

4.2.1 Chaîne de redressements des immigrés et des Domiens (G1)

La chaîne de redressements des immigrés et des Domiens est constituée de 2 étapes :

1. une première étape de correction de la non-réponse au questionnaire pour les individus des sous-échantillons (01) et (02) pour corriger du biais de sélection lié à la non-réponse de ces sous-échantillons
2. une deuxième étape de partage des poids car une fiche adresse des sous-échantillons (01) ou (02) aurait pu être sélectionnée dans le sous-échantillon (05) représentatif de l'ensemble de la population

Concernant l'étape de correction de la non-réponse, des modèles différents ont été appliqués suivant les origines des individus afin de mieux adhérer aux enjeux de diffusion de l'enquête, ce point est développé dans la partie 4.3.1.

Concernant l'étape de partage des poids, les individus échantillonnés dits de première génération, qu'ils soient immigrés ou Domiens, peuvent être issus :

1. soit du sous-échantillon spécifique aux G1 c'est à dire le sous-échantillon (01) ou le sous-échantillon (02) ;
2. soit du sous-échantillon de la population générale (05).

Par conséquent, un individu immigré ou domien appartient simultanément aux champs de ces deux sous-échantillons, et ceci même si ces sous-échantillons sont par construction exclusifs⁶⁶ l'un de l'autre. Il importe donc d'en tenir compte dans le calcul de son poids. Une opération de partage des poids a été effectuée pour en tenir compte (cf. partie 4.4.2). Au préalable de l'opération du partage de poids, il a été

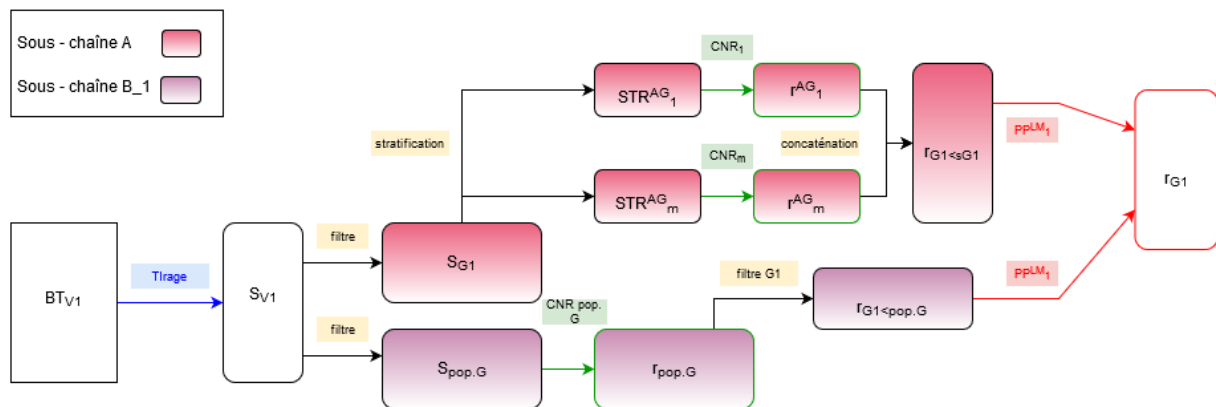
⁶⁶ En effet, le caractère exclusif de ces sous-échantillons est assurée par des tirages sur les dates de naissance essentiellement incompatible entre les deux. Toutefois, pour certaines origines rares, les individus nés du 14 au 19 de chaque mois sont susceptible d'être sélectionnés dans chacun des sous-échantillons (01) et (05). Une disjonction manuelle est opérée pour éviter cela. Cette dernière induit une repondération afin de corriger le biais qu'elle génère.

nécessaire de réaliser une étape de correction de la non-réponse pour le sous-échantillon de la population (05).

Au total, la chaîne de repondérations (hors calage) des immigrés et Domiens s'obtient en assemblant :

1. la sous-chaîne de repondérations des individus des sous-échantillons (01) et (02), présentée dans le schéma ci-après (sous-chaîne A) ;
2. la sous-chaîne de repondérations des individus du sous-échantillon (05), présentée dans le schéma, restreinte aux seuls individus G1 (sous-chaîne B_1).

Schéma 3 : Chaîne de redressements des immigrés et domiens



Où BT_{V1} est la base de tirage pour la vague 1,

S_{V1} l'échantillon tiré dans BT_{V1} . C'est la concaténation des sous-échantillon (01), (02), (05)

S_{G1} le sous-échantillon spécifique aux G1 c'est à dire la concaténation des sous-échantillons (01) et (02)

$S_{pop,G}$ le sous-échantillon de la population générale c'est à dire le sous-échantillon (05) $r_{pop,G}$
l'ensemble des individus répondants du sous-échantillon $S_{pop,G}$

$r_{G1<pop,G}$ désigne les répondants G1 issus de $r_{pop,G}$

Pour i compris entre 1 et m , STR_i^{AG} désigne le sous-échantillon de S_{G1} obtenu par restriction au groupe agrégé d'origine d'indice i

Pour i compris entre 1 et m , r_i^{AG} désigne l'ensemble des individus répondants de STR_i^{AG}

$r_{G1<S_{G1}}$ est la concaténation des sous-échantillons r_i^{AG} : il s'agit de l'ensemble des individus G1 répondants atteints par le sous-échantillon spécifique aux G1.

r_{G1} désigne l'ensemble de tous les individus G1 répondants. Ils sont de 2 types : d'une part les G1 répondants issus du sous-échantillon spécifique aux G1 et d'autre part les G1 répondants issus du sous-échantillon de la population générale. Autrement dit, r_{G1} est la concaténation du sous-échantillon $r_{G1<pop,G}$ et du sous-échantillon $r_{G1<S_{G1}}$

Le terme CNR désigne la correction de la non-réponse. Le terme $PPLM$ désigne le partage des poids dans un contexte de liens multiples.

4.2.2 Chaîne de redressements des descendants d'immigrés et de Domiens

La chaîne de redressements des G2 est plus compliquée qu'elle ne l'est pour les G1. En effet, en plus d'une étape classique de correction de la non-réponse, un traitement supplémentaire est réalisé, relatif à l'opération de relevés mairie, laquelle est propre à l'enquête.

La chaîne de redressements des G2 est donc constituée :

1. d'une première étape de correction de la non-réponse pour les relevés mairie ;
2. d'une deuxième étape de correction de la non-réponse au questionnaire pour les individus des sous-échantillons (03) et (04) ;
3. d'une troisième étape de partage des poids.

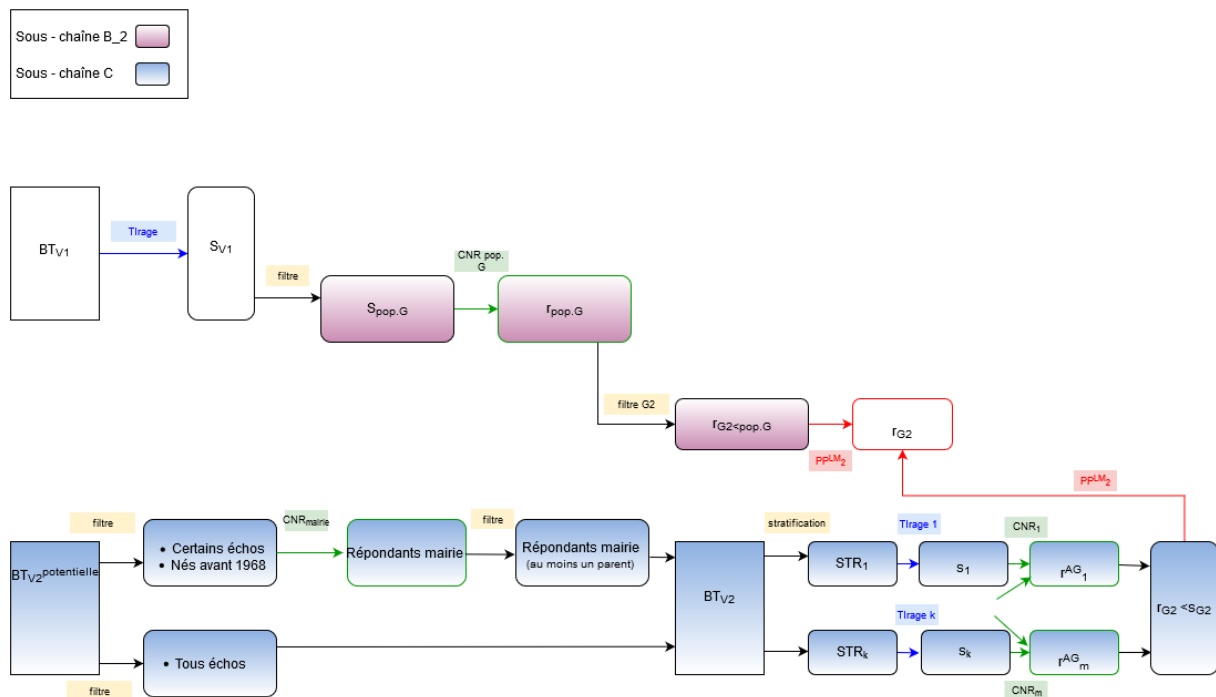
Concernant la troisième étape (partage des poids, de façon semblable à ce qui a été dit pour les individus G1), les individus G2 appartiennent à la fois :

1. au champ du sous-échantillon des descendants d'immigrés (03) et de Domiens (04) ;
2. au champ du sous-échantillon de la population générale (05).

Là encore, une opération de partage des poids a donc été effectuée pour affecter à ces individus des poids composites construits à partir des pondérations de ces deux sous-échantillons. La chaîne de repondérations (hors calage) des descendants d'immigrés et de domiens est présentée dans le 5. Elle est obtenue par concaténation de :

1. la sous-chaîne de repondérations des individus du sous-échantillon (05), présentée dans le schéma ci-après, restreinte aux seuls individus G2 (sous chaîne B_2);
2. la sous-chaîne de repondérations des descendants d'immigrés et de domiens présentée au schéma (sous chaîne C).

Schéma 4 : Chaîne de redressements des descendants d'immigrés et de Domiens



Où BT_{V_1} est la base de tirage pour la vague 1,

S_{V_1} l'échantillon tiré dans BT_{V_1} . C'est la concaténation des sous-échantillons (01), (02), (05)

$S_{pop.G}$ le sous-échantillon de la population générale c'est à dire le sous-échantillon (05)

$r_{pop.G}$ l'ensemble des individus répondants du sous-échantillon $S_{pop.G}$

$r_{G2 < pop.G}$ désigne les répondants G2 issus de $r_{pop.G}$

Certains individus vont faire l'objet d'une recherche en mairie (consultation de leur bulletin individuel de naissance, sur lequel figurent les pays de naissance de leurs parents). Pour des raisons de coût de collecte, il a été décidé de limiter les relevés en mairie à un nombre de 100 000 environ ce qui a conduit au tirage d'une base $BT_{V_2}^{potentielle}$.

$BT_{V_2}^{potentielle}$ désigne la base obtenue après tirage dans la base de sondage (les répondants à l'EAR 2018 qui ne sont ni immigrés ni nés dans un DOM) restreinte à l'échantillon maître (EM⁶⁷). Cette base est filtrée suivant les résultats après appariement avec les fichiers d'état civil anonyme (STATEC). Les individus pour lesquels certains échos seulement (individus de même sexe nés le même jour dans la même commune) ont leurs parents nés à l'étranger ou dans les DOM ainsi que les individus nés avant 1968, font l'objet d'une opération de relevés en mairie. Parmi eux, ceux pour lesquels l'opération a abouti sont appelés « répondants mairie ». Un filtre est alors appliqué pour conserver uniquement ceux qui ont au moins un parent né à l'étranger ou dans les DOM : on les appelle « répondants mairie (au moins un parent) ». Dans le schéma ci-dessus, ceux-là vont alimenter la base de tirage de la vague 2, noté BT_{V_2} . Il en est de même des individus pour lesquels tous les échos ont des parents nés à l'étranger ou dans les DOM : on les appelle « tous échos ».

Pour i compris entre 1 et k , STR_i désigne la restriction de cette base de tirage au groupe agrégé d'origines d'indice i . On désigne par S_i l'échantillon tiré dans cette dernière.

Pour j compris entre 1 et m , r_j^{AG} désigne un regroupement de l'ensemble des répondants au sein de l'agrégation des groupes d'origines d'indice j .

$r_{G2 < s_{G2}}$ est l'ensemble des répondants G2 qui ont été atteints par l'échantillon spécifique aux G2.

r_{G2} désigne l'ensemble de tous les individus G2 répondants. Ils sont de 2 types : d'une part les G2 répondants issus du sous-échantillon spécifique aux G2 et d'autre part les G2 répondants issus du sous-échantillon de la population générale. Autrement dit, r_{G2} est la concaténation du sous-échantillon $r_{G2 < pop.G}$ et du sous-échantillon $r_{G2 < s_{G2}}$

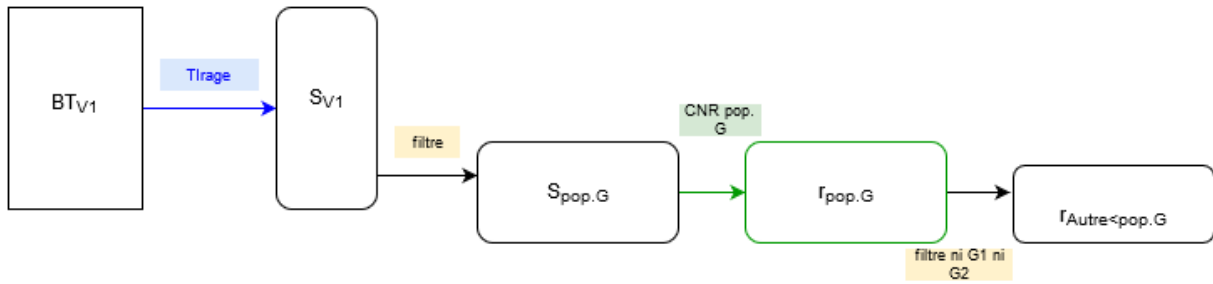
Le terme CNR désigne la correction de la non-réponse. Le terme PP^{LM} désigne le partage des poids dans un contexte de liens multiples.

4.2.3 Chaîne de redressements des « autres individus »

Les individus de l'échantillon qui ne sont ni des G1 ni des G2 (ou individus de la « population majoritaire ») n'appartiennent qu'au champ du sous-échantillon (05) de la population générale. Une étape de correction de la non-réponse au questionnaire pour les individus du sous-échantillon (05) est réalisée. Contrairement aux G1 et aux G2, ces individus ne sont pas concernés par l'opération de partage des poids. Leur chaîne de redressements (hors calage) est donc obtenue par restriction à la « population majoritaire ».

⁶⁷ Afin de limiter le coût des déplacements des enquêteurs

Schéma 5 : Chaîne de redressements des autres individus



Où S_{V1} l'échantillon tiré dans BT_{V1} . C'est la concaténation des sous-échantillons (01), (02), (05)

$S_{pop.G}$ le sous-échantillon de la population générale c'est à dire le sous-échantillon (05)

$r_{pop.G}$ l'ensemble des individus répondants du sous-échantillon $S_{pop.G}$

$r_{Autre<pop.G}$ résulte de l'application d'un filtre aux individus de $r_{pop.G}$ qui ne sont ni des G1 ni des G2.

4.2.4 Chaîne de redressements globale de l'enquête TeO2 (hors calage)

La chaîne de redressements globale (calage exclu) de l'enquête TeO2 est obtenue par branchement :

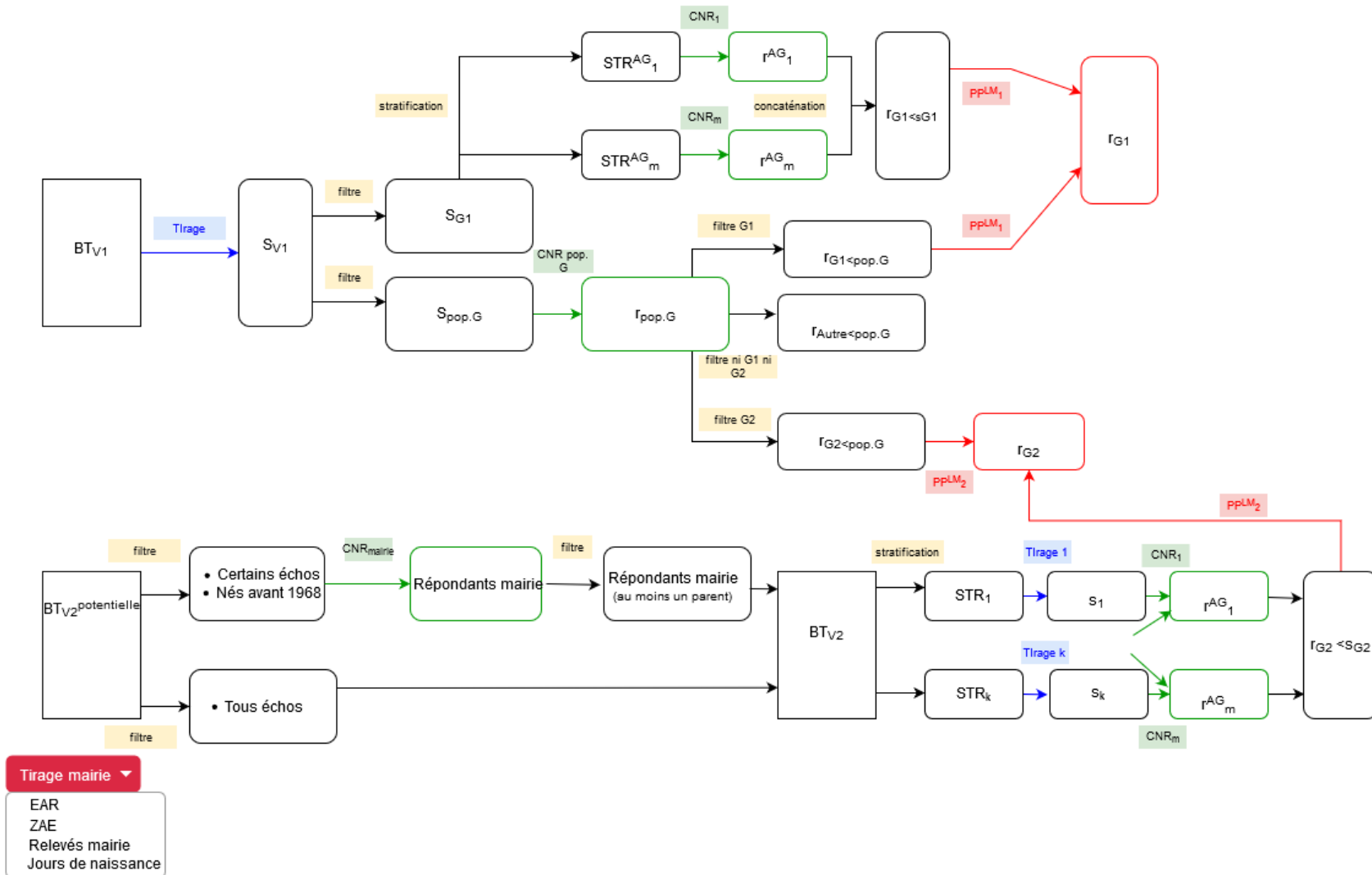
1. de la chaîne de redressements des immigrés et Domiens (voir schéma 3) ;
2. de la chaîne de redressements des descendants d'immigrés et de Domiens (voir schéma 4) ;
3. de la chaîne de redressements des individus qui ne sont ni immigrés, ni Domiens, ni descendants d'immigrés, ni descendants de Domiens, i.e. la « population majoritaire » (voir schéma 5).

Cette chaîne est présentée dans le schéma ci-après.

Les redressements de l'enquête complémentaire TeO 2 - G3, qui prennent en compte les particularités de son mode de collecte, sont décrits plus en détails dans Guin et Thao Khamsing et Tanneau (2022)⁶⁸.

⁶⁸Guin, O., Tanneau, P., Thao Khamsing, W. (2022). « Redressements de l'enquête complémentaire sur les descendants d'immigrés de troisième génération d'origine non-européenne (TeO 2 - G3) », Journées de méthodologie statistique de l'Insee.

Schéma 6 : Chaîne de redressements globale de l'enquête TeO2 (hors calage)



4.3 Correction de la non-réponse de TeO2

L'étape de correction de la non-réponse totale consiste à estimer dans le champ des répondants et des non-répondants, la probabilité de répondre en fonction des caractéristiques disponibles sur l'ensemble de ce champ. Le poids initial des répondants (inverse de la probabilité d'inclusion dans l'échantillon tiré dans la base de sondage) est ensuite corrigé en le multipliant par l'inverse de la probabilité de réponse estimée. Cette opération permet de corriger de la non-réponse totale et de rendre ainsi l'échantillon des répondants représentatif de l'ensemble du champ de l'enquête.

4.3.1 Correction de la non-réponse de TeO2 – vague 1

La correction de la non-réponse des individus procède en deux sous-étapes distinctes :

1. modélisation de la non-réponse des individus par le biais d'une régression logistique dichotomique. Pour chaque individu, cette régression produit une probabilité de réponse estimée.

2. construction automatique de groupes de réponse homogènes (GRH) à partir de la probabilité de réponse estimée par la méthode de Haziza-Beaumont⁶⁹.

La réunion des sous-échantillons (01) et (02) est l'échantillon noté S_{G1} , qui regroupe tous les individus G1 sélectionnés en vague 1. Cet échantillon est obtenu par application d'un filtre à S_{V1} (cf. schéma 3).

L'approche retenue est celle des multi-modèles. En effet, afin de mieux adhérer aux enjeux de diffusion de l'enquête, il a été décidé d'appliquer des modèles de correction de la non-réponse spécifiques aux groupes d'origines des individus, telles qu'elles sont connues dans la base de sondage. Si l'on suspecte des effets différenciés suivant le groupe d'origine des facteurs explicatifs de la non-réponse sur cette dernière, alors, appliquer des modèles spécialement construits pour des groupes d'origines permet de mieux appréhender le comportement de réponse des individus, et in fine de mieux corriger le biais de sélection induit par la non-réponse.

Différents scénarios de non-réponse ont alors été envisagés, consistant à :

- choisir une stratification des origines des individus ; l'enquête TeO2 est déjà stratifiée en groupes d'origines, il s'agit ici d'envisager des agrégations de ces groupes d'origines ;
- pour chaque strate ainsi définie, proposer un ensemble de variables explicatives de la non-réponse. Ces ensembles de variables peuvent ainsi être théoriquement différents d'une strate à l'autre, car l'on peut suspecter que certaines variables soient significatives pour un groupe d'origines mais pas pour un autre.

La stratification des origines doit répondre à certaines contraintes :

- tout d'abord, des contraintes d'effectifs : des strates de tailles trop petites affaibliraient la puissance statistique des modèles, qui risqueraient de manquer de stabilité et de produire en conséquence des estimateurs manquant de précision ;
- ensuite, des contraintes d'homogénéité : il s'agit ici de regrouper autant que possible dans une même strate des origines pour lesquelles les connaissances métier laissent à penser que les effets des variables explicatives de la non-réponse sont sensiblement les mêmes d'une origine à l'autre.

⁶⁹ Haziza, D., Beaumont, J.-F. (2007), On the construction of imputation classes in surveys, International Statistical Review. »

Modèles de correction de la non-réponse du sous-échantillon (01) immigrés

Après plusieurs essais de modélisation, les groupes d'origines et les variables retenus⁷⁰ dans la modélisation de la non-réponse des individus du sous-échantillon (01) sont les suivants (cf. tableau ci-après) :

Tableau 22 : Groupes d'origines et variables retenus pour la CNR du sous-échantillon (01)

Modèle	Origines	Variables
Modèle 1	Algérie, Maroc Tunisie	Sexe Tranche d'âge Situation d'emploi de l'individu Niveau de diplôme Appartenance du logement de l'individu à un organisme HLM Présence d'un déménagement de l'individu Région de gestion de l'enquêteur Tranche d'unité urbaine ⁷¹ Quartier prioritaire de la politique de la ville ⁷²
Modèle 2	Afrique sahélienne, Afrique centrale et du golfe de Guinée	Tranche d'âge Statut d'occupation du logement Situation d'emploi de l'individu Appartenance du logement de l'individu à un organisme HLM Présence d'un déménagement de l'individu Région de gestion Quartier prioritaire de la politique de la ville
Modèle 3	Espagne, Italie, Portugal, Autres pays de l'UE 28, Autres pays du monde	Sexe Tranche d'âge Statut d'occupation du logement Type de logement Situation d'emploi de l'individu Niveau de diplôme Appartenance du logement de l'individu à un organisme HLM Présence d'un déménagement de l'individu Région de gestion
Modèle 4	Turquie, Asie du Sud-Est	Nombre de personnes dans le logement Statut d'occupation du logement Niveau de diplôme Appartenance du logement de l'individu à un organisme HLM Présence d'un déménagement de l'individu Région de gestion Tranche d'unité urbaine Quartier prioritaire de la politique de la ville
Modèle 5	Chine, Pays avec de nombreux réfugiés	Statut d'occupation du logement Situation d'emploi de l'individu Niveau de diplôme Appartenance du logement de l'individu à un organisme HLM Présence d'un déménagement de l'individu Région de gestion Quartier prioritaire de la politique de la ville

Les résultats de la modélisation figurent en annexe 7.

⁷⁰ Issus de la base de sondage

⁷¹ Appartenance du lieu de résidence à la tranche d'unité urbaine

⁷² Appartenance du lieu de résidence de l'individu à un quartier prioritaire de la politique de la ville

À l'issue de la correction de la non-réponse, l'étape de construction des groupes de réponses homogènes (GRH) a permis de constituer un certain nombre de groupes dont les informations sont résumées dans le tableau ci-après.

Tableau 23 : GRH – modèles CNR sous-échantillon (01)

Modèle	Nombre de GRH	Taille ⁷³ du plus petit GRH	Probabilité de réponse la plus faible	Probabilité de réponse la plus forte
Modèle 1	9	181	0,18	0,80
Modèle 2	9	198	0,14	0,78
Modèle 3	9	159	0,13	0,84
Modèle 4	10	139	0,18	0,83
Modèle 5	9	126	0,14	0,77

Modèles de correction de la non-réponse du sous-échantillon (02) Domiens

Les variables qui interviennent dans la modélisation de la non-réponse des individus du sous-échantillon (02) sont présentées dans le tableau ci-après :

Tableau 24 : Variables retenues pour la CNR du sous-échantillon (02)

Sous-échantillon	Variables
(02) Domiens	Sexe Tranche d'âge Statut d'occupation du logement Situation d'emploi de l'individu Présence d'un déménagement de l'individu Région de gestion

Après correction de la non-réponse, l'étape de construction des groupes de réponses homogènes a permis de constituer 9 groupes (cf. tableau ci-après). Le plus petit groupe contient 143 observations.

Tableau 25 : GRH – modèle CNR sous-échantillon (02)

Modèle	Nombre de GRH	Taille du plus petit GRH	Probabilité de réponse la plus faible	Probabilité de réponse la plus forte
Modèle sous-échantillon (02)	9	143	0,11	0,82

⁷³ En nombre d'observations

Modèles de correction de la non-réponse du sous-échantillon (05) population générale

Les variables qui interviennent dans la modélisation de la non-réponse des individus du sous-échantillon (05) sont présentées dans le tableau ci-après :

Tableau 26 : Variables retenues pour la CNR du sous-échantillon (05)

Sous-échantillon	Variables
(05) population générale	Sexe Statut d'occupation du logement Type de logement Situation d'emploi de l'individu Niveau de diplôme Présence d'un déménagement de l'individu Région de gestion Tranche d'unité urbaine

Après correction de la non-réponse, l'étape de construction des groupes de réponses homogènes a permis de constituer 9 groupes (cf. tableau ci-après). Le plus petit groupe contient 176 observations.

Tableau 27 : GRH – modèle CNR sous-échantillon (05)

Modèle	Nombre de GRH	Taille du plus petit GRH	Probabilité de réponse la plus faible	Probabilité de réponse la plus forte
Modèle sous-échantillon (05)	9	176	0,11	0,83

À l'issue de cette étape, on dispose donc de jeux de pondérations associés à ces différents groupes.

4.3.2 Correction de la non-réponse de TeO2 – vague 2

Correction de la non-réponse mairie

Par définition, un individu descendant d'immigré ou de Domiens (G2), est un individu dont l'un au moins des parents est immigré⁷⁴ (resp. domien) mais n'étant pas lui-même immigré (resp. domien). Pour rappel, il n'existe pas de base de sondage des descendants d'immigrés ou de Domiens, en particulier les informations disponibles dans l'Enquête Annuelle de Recensement 2018 (EAR 2018) ne sont pas suffisantes pour identifier clairement ces derniers. La méthodologie retenue pour pallier cette difficulté a été décrite en détail précédemment (cf. partie 2.3). Elle a nécessité une opération de recherche en mairie afin d'établir quels individus d'un échantillon d'individus non-immigrés et non-Domiens tiré dans la base de sondage (l'EAR 2018) sont effectivement descendants de parents nés à l'étranger ou dans les DOM.

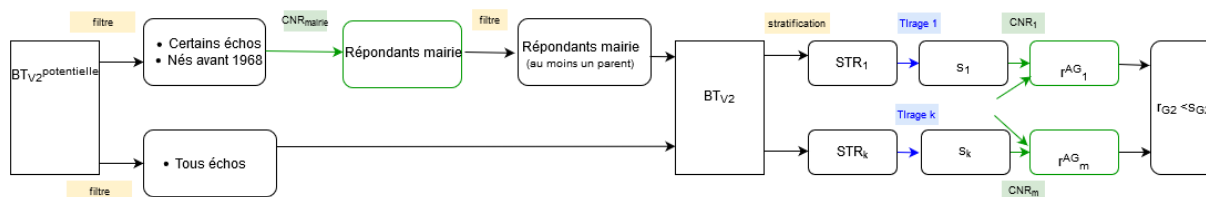
Cette opération fait elle-même fait l'objet d'une étape de correction de la non-réponse.

⁷⁴ i.e. né étranger à l'étranger

Pour rappel, l'échantillon d'individus qui a fait l'objet d'une opération de relevés mairie est composé de 100 014 individus. Cet échantillon se décompose de la façon suivante (cf. schéma ci-après) :

1. 75 032 individus nés entre 1968 et 2001 pour lesquels seulement certains échos ont un parent né à l'étranger ou dans un DOM ;
2. 24 982 individus nés avant 1968, pour lesquels aucun appariement avec les fichiers anonymisés d'état-civil n'a pu être fait.

Schéma 7 : CNR du sous-échantillon (03) des descendants d'immigrés et (04) des descendants de domiens



Pour certains individus (5 133), l'opération relevés mairie n'a pas abouti, de sorte qu'il est impossible de déterminer si ces individus sont ou ne sont pas dans le champ de l'enquête. Par conséquent, et afin de corriger un éventuel biais de sélection, une étape de correction de la non-réponse - appelée correction de la non-réponse mairie - visant à transférer les poids des individus pour lesquels l'opération mairie n'a pas abouti sur ceux pour lesquels elle a abouti, a été réalisée pour les individus ayant fait l'objet d'une recherche en mairie.

Les variables qui interviennent dans la modélisation de la non-réponse mairie sont présentées dans le tableau ci-après.

Tableau 28 : Variables retenus pour la CNR mairie

Sous-échantillon	Variables
Mairie	Sexe Statut d'occupation du logement Type de logement Nombre de personnes dans le logement Niveau de diplôme Appartenance du logement de l'individu à un organisme HLM Région de gestion Tranche d'unité urbaine Mode de recueil des informations (ordinateur/papier) Type d'échos ⁷⁵ Nombre d'origines ⁷⁶

Les résultats de la modélisation figurent en annexe 8.

⁷⁵ Variable construite suite à l'appariement et l'exploitation des fichiers anonymisés d'état civil (cf. partie 2.3.2). Elle comprend 4 modalités : né avant 1968, l'appariement donne un seul écho, l'appariement donne plusieurs échos avec une seule origine, l'appariement donne plusieurs échos avec plusieurs origines.

⁷⁶ L'exploitation des fichiers anonymisés d'état civil a permis d'identifier le nombre d'origines différentes pour tous les échos.

À l'issue de la correction de non-réponse, l'étape de construction des groupes de réponses homogènes a permis de constituer 6 groupes (cf. tableau ci-après). Le plus petit groupe contient 2 230 observations.

Tableau 29 : GRH – modèle CNR mairie

Modèle	Nombre de GRH	Taille du plus petit GRH	Probabilité de réponse la plus faible	Probabilité de réponse la plus forte
Modèle mairie	6	2230	0,81	0,98

Pour rappel, l'opération mairie aboutit à la construction de la base de sondage des descendants de personnes nées à l'étranger ou dans les DOMS. Elle est une étape permettant l'identification des enfants d'immigrés ou de nés dans les DOMS potentiels. La base de sondage (BT_{V2}) ainsi construite et contenant 38 036 individus est constituée de :

- 29 784 individus pour lesquels les relevés en mairie ont conclu que ces individus appartenaient au champ de ce sous-échantillon,
- ainsi que de 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.

Près de 24 000 FA ont été sélectionnées pour la collecte de la vague 2 de l'enquête TeO2.

Correction de la non-réponse questionnaire

Une étape de repondération a été entreprise pour corriger la non-réponse au questionnaire de l'enquête. Comme pour les G1, et pour les mêmes raisons, cette correction de la non-réponse s'est faite par groupe d'origines.

Modèles de correction de la non-réponse du sous-échantillon (03) descendants d'immigrés

Après plusieurs essais de modélisation, les groupes d'origines et les variables retenus dans la modélisation de la non-réponse des individus du sous-échantillons (03) sont les suivants (cf. tableau ci-après) :

Tableau 30 : Groupes d'origines et variables retenus pour la CNR du sous-échantillon (03)

Modèle	Origines	Variables
Modèle 1	Algérie Maroc Tunisie	Sexe Tranche d'âge Statut d'occupation du logement Type de logement Situation d'emploi de l'individu Niveau de diplôme Appartenance du logement de l'individu à un organisme HLM Présence d'un déménagement de l'individu Région de gestion Tranche d'unité urbaine Quartier prioritaire de la politique de la ville
Modèle 2	Afrique sahélienne Afrique centrale et du golfe de Guinée	Sexe Tranche d'âge Statut d'occupation du logement Type de logement Situation d'emploi de l'individu Appartenance du logement de l'individu à un organisme HLM Présence d'un déménagement de l'individu Tranche d'unité urbaine G2 vivant dans le même logement qu'un individu enquêté en vague 1 ⁷⁷ Quartier prioritaire de la politique de la ville
Modèle 3	Espagne Italie, Portugal Autres pays de l'UE 28	Statut d'occupation du logement Type de logement Situation d'emploi de l'individu Niveau de diplôme Appartenance du logement de l'individu à un organisme HLM Présence d'un déménagement de l'individu Tranche d'unité urbaine G2 vivant dans le même logement qu'un individu enquêté en vague 1
Modèle 4	Turquie Asie du Sud-Est Autre pays du monde	Tranche d'âge Statut d'occupation du logement Type de logement Niveau de diplôme Appartenance du logement de l'individu à un organisme HLM Présence d'un déménagement de l'individu Tranche d'unité urbaine Quartier prioritaire de la politique de la ville
Modèle 5	Origines incertaines - Afrique centrale et du golfe de Guinée Origines incertaines - Origines surreprésentées Origines incertaines - Origines sous-représentées	Tranche d'âge Statut d'occupation du logement Situation d'emploi de l'individu Niveau de diplôme Appartenance du logement de l'individu à un organisme HLM Présence d'un déménagement de l'individu Tranche d'unité urbaine

⁷⁷ Pour des raisons de contraintes d'échantillonnage, il a été autorisé d'enquêter en vague 2 des individus G2 vivant dans le même logement qu'un individu interrogé en vague 1 (environ 900 individus). Lors de la correction de la non-réponse, nous avons vérifié si le fait de vivre dans un même logement qu'un individu enquêté en vague 1 est une variable explicative de la non-réponse. Cette variable est corrélée à la thématique de l'enquête puisque des individus qui vivent chez leurs parents ou sont en couple avec une personne immigrée ont une trajectoire sans doute différente de celle d'autres descendants d'immigrés. De plus, cette variable peut être corrélée à la réponse de l'enquêté car la charge de collecte pesant sur ce ménage est plus importante.

Les résultats de la modélisation figurent en annexe 8.

À l'issue de la correction de la non-réponse, l'étape de construction des groupes de réponses homogènes (GRH) a permis de constituer un certain nombre de groupes dont les informations sont résumées dans le tableau ci-après.

Tableau 31 : GRH – modèle CNR sous-échantillon (03)

Modèle	Nombre de GRH	Taille ⁷⁸ du plus petit GRH	Probabilité de réponse la plus faible	Probabilité de réponse la plus forte
Modèle 1	11	510	0,14	0,82
Modèle 2	10	151	0,07	0,68
Modèle 3	8	212	0,23	0,77
Modèle 4	9	312	0,15	0,79
Modèle 5	9	108	0,15	0,77

Modèle de correction de la non-réponse du sous-échantillon (04) descendants de Domiens

Après plusieurs essais de modélisation, les variables retenues dans le modèle de correction de la non-réponse des individus du sous-échantillon (04) sont les suivantes (cf. tableau ci-après) :

Tableau 32 : Variables retenues pour la CNR du sous-échantillon (04)

Sous-échantillon	Variables
(04) descendants de Domiens	Tranche d'âge Nombre de personnes dans le logement Statut d'occupation du logement Présence d'un déménagement de l'individu Quartier prioritaire de la politique de la ville

Les résultats de la modélisation sont présentés en annexe 8.

À la suite de la correction de la non-réponse, l'étape de construction des groupes de réponses homogènes a permis de constituer 6 groupes (cf. tableau ci-après). Le plus petit groupe contient 134 observations.

Tableau 33 : Probabilités de réponse des GRH – modèle CNR sous-échantillon (04)

Modèle	Nombre de GRH	Taille du plus petit GRH	Probabilité de réponse la plus faible	Probabilité de réponse la plus forte
(04) descendants de Domiens	6	134	0,22	0,71

À l'issue de cette étape, on dispose donc de jeux de pondérations associés à ces différents groupes.

⁷⁸ En nombre d'observations

4.4 Partage des poids de TeO2

4.4.1 Enjeux du partage des poids de TeO2

Les modèles de correction de la non-réponse (CNR) sont spécifiques à chaque croisement sous-échantillon x regroupement de strate d'origines. À l'issue de cette étape, on obtient donc des jeux de pondération assurant que chaque croisement est bien représentatif de son champ. Cependant, la concaténation de ces jeux de pondérations ne permet pas d'assurer à ce stade la représentativité de l'échantillon global des répondants. L'opération de partage des poids vise à intégrer le fait que certains des répondants puissent théoriquement être captés via différents sous-échantillons dans la base de sondage (liens multiples). Cette opération permet ainsi de produire des poids composites assurant la représentativité de l'échantillon global vis-à-vis de la base de sondage. Le calcul d'un tel poids pour un individu donné repose sur le principe général suivant :

- détermination des liens entre cet individu et les unités de la base de sondage ; il s'agit de repérer a posteriori (avec les informations de la collecte sur les seuls répondants) les individus qui potentiellement pouvaient être tirés dans deux sous-échantillons (voir ci-dessous)
- puis utilisation de ces liens et des poids de ces unités pour affecter un poids à l'individu.

4.4.2 Partage des poids de TeO2 – G1

Les individus échantillonnés dits de première génération, qu'ils soient immigrés ou Domiens, peuvent être issus :

1. soit du sous-échantillon spécifique aux G1 c'est à dire le sous-échantillon (01) ou le sous-échantillon (02) ;
2. soit du sous-échantillon de la population générale (05).

Par conséquent, un individu immigré ou domien appartient simultanément aux champs de ces deux sous-échantillons, et ceci même si, ces sous-échantillons sont par construction exclusifs l'un de l'autre (cf. annexe 9). Il importe donc d'en tenir compte dans le calcul de son poids. Une opération de partage des poids a donc été effectuée pour en tenir compte (i.e. pour tenir compte du nombre de liens). Ainsi, pour un individu immigré qui potentiellement peut être sélectionné via le sous-échantillon (05) « population générale » mais également via le sous-échantillon (01) spécifique aux « immigrés », ce nombre de liens vaut 2. De la même façon, un individu domien qui potentiellement peut être sélectionné via le sous-échantillon (05) « population générale » mais également via le sous-échantillon (02) spécifique aux « Domiens », ce nombre de liens vaut 2 également. Il faut tenir compte de nombre de liens dans l'élaboration des pondérations (méthode généralisée du partage des poids). L'application brute des liens (nombre de fois qu'un individu aurait pu être échantillonné) entraîne une dispersion des poids considérable.

En effet :

1. une première source de dispersion est liée au fait qu'un individu aura un poids très différent s'il est atteint par le sous-échantillon (05) sans surreprésentation ou par un des sous-échantillons (01) ou (02) avec des surreprésentations élevées ;
2. une deuxième source de dispersion est liée à des surreprésentations différentes d'une origine à l'autre.

Par conséquent, il s'avère nécessaire de pondérer les liens, et de faire dépendre ces pondérations des origines des individus. Nous retenons l'approche des liens développée par Lavallée⁷⁹ (2009). Des liens

⁷⁹ Lavallée, P. (2009). Indirect sampling. Springer Science & Business Media.

pondérés sont ainsi calculés, en reprenant la méthodologie proposée par Davezies, Landre, Murat & Rousseau⁸⁰ (2005) (cf. annexe 9).

La date de naissance du 1er janvier fait l'objet d'un traitement particulier pour les immigrés de certaines origines (Maroc, Tunisie, Afrique sahélienne, Turquie). Ainsi, on considère que ces individus ont été échantillonnés exhaustivement au titre du sous-échantillon (01) et n'ont, de ce fait, qu'un lien avec la base de sondage.

4.4.3 Partage des poids de TeO2 – G2

Par construction, les individus échantillonnés dits de deuxième génération, qu'ils soient descendants d'immigrés ou de Domiens, peuvent être issus :

1. soit du sous-échantillon spécifique aux G2 c'est à dire le sous-échantillon (03) ou le sous-échantillon (04) ;
2. soit du sous-échantillon de la population générale (05).

Comme pour les G1, une étape de partage des poids a été réalisée pour tenir compte du nombre de liens entre les individus répondants descendants d'immigrés et de Domiens et la base de sondage. Ce nombre de liens est systématiquement égal à 2. La formule classique du partage des poids par simple décompte mène à une forte dispersion des poids, et donc à un niveau élevé de variance, ce qui est peu souhaitable. Encore une fois, et pour les mêmes raisons, le partage des poids a été réalisé en tenant compte de l'origine (notion de liens pondérés).

4.4.4 Partage des poids de TeO2 – G0

Si l'individu n'est ni immigré ni domien, ni descendant de personne née à l'étranger ou dans les DOM, alors il n'a pu être sélectionné qu'à partir du sous échantillon (05). Dans ce cas, l'individu a un seul lien avec la base de sondage. Autrement dit, son poids reste invariant après partage des poids.

4.5 Calage sur marge de l'enquête TeO2⁸¹

4.5.1 Principe du calage et application à l'enquête TeO2

À l'issue de l'opération du partage des poids, on obtient donc un jeu de pondération unique sur l'échantillon global des répondants.

L'objectif du calage sur marges est double :

1. assurer une cohérence entre les totaux estimés de certaines variables (dites variables de calage) et les vrais totaux de ces variables connus, par une information externe, sur la population
2. réduire la variance des estimateurs des variables d'intérêt de l'enquête

L'estimateur calé possède de bonnes propriétés statistiques, et ce d'autant plus que les variables de calage sont corrélées aux variables d'intérêt de l'enquête. En effet :

- le calage vise à ne modifier que légèrement les poids, de sorte que le biais qu'il introduit reste maîtrisé ;

⁸⁰ Davezies, L., Landre, C., Murat, F., & Rousseau, S. (2005) Problèmes théoriques et pratiques de la mise en œuvre d'une sur-représentation des ZUS dans les échantillons d'enquête : le cas de l'enquête IVQ.

⁸¹ On notera que les tableaux et figures présentés ci-après correspondent à une version non définitive des travaux de calage: ces résultats se rapportent notamment à un schéma de redressement privilégiant une troncature uniforme au 99ème percentile (contre une troncature par origine au 95ème percentile finalement retenue), et ne prennent pas en compte certains retraitements des origines inconnues au moment de l'échantillonnage. En revanche, le texte présente bien la méthode finalement adoptée.

- par ailleurs, les contraintes de calage assurent que les estimateurs par expansion des totaux des variables de calage sont de variances nulles, et la bonne corrélation de ces dernières avec les variables d'intérêt assure que les estimateurs des variables d'intérêt sont eux-mêmes de variances presque nulles.

Le calage nécessite de disposer d'une source externe qui fournisse les totaux connus de variables auxiliaires (les marges).

Pour les descendants d'immigrés et de Domiens, il n'existe aucune source fiable pour donner des marges utilisables. L'enquête Emploi comporte des informations mais son échantillon de descendants est trop faible. En effet, de manière générale on peut utiliser une enquête ou une source pour calculer des marges pour une autre enquête dès lors que l'échantillon de la première est dix fois plus gros que l'échantillon de la seconde. Et cela, de façon à permettre d'obtenir des estimateurs beaucoup plus précis que l'enquête à laquelle est appliquée le calage. Cette source a donc été écartée, elle a néanmoins servi comme source de validation externe en particulier sur les descendants d'immigrés, les enfants de français nés à l'étranger, les niveaux de diplômes, le statut d'emploi etc.

On retient comme source de référence pour le calcul des marges, l'enquête annuelle de recensement de l'année 2019 couplée avec celle de 2020 (qui donne une photographie de la population en moyenne sur l'année 2019). La taille de cette source - près de 5 millions d'observations sur le champ de TeO2 - est son principal avantage. En particulier, elle permet de mesurer des caractéristiques sur les populations rares comme les Domiens (moins de 1% de la population).

La procédure du calage porte sur 3 populations :

1. les immigrés
2. les Domiens
3. les autres (ni immigrés, ni Domiens)

L'EAR est pour l'essentiel auto-administrée alors que l'enquête TeO2 interroge des individus principalement en face-à-face. Ainsi de nombreuses caractéristiques ne peuvent pas être calées en raison d'un mode de questionnement trop différent. Par exemple, le diplôme de l'individu fait l'objet d'une question dans l'EAR et d'un questionnement détaillé dans TeO2 (codification automatique). Concernant les immigrés, la question des équivalences des diplômes est nécessairement traitée différemment. Il en est de même pour la situation d'emploi. Les questions posées sont les mêmes mais la nature de l'information collectée diffère trop pour permettre un calage sur l'activité de l'individu enquêté par l'enquête TeO2 (PCS, statut...).

Enfin, une analyse préalable des poids après partage des poids montre que ceux-ci sont relativement dispersés. La solution appliquée a été de tronquer les poids au sein de chaque origine avant même le calage. Ainsi, les poids les plus élevés ont été tronqués au 95^{ème} percentile.

4.5.2 Calage sur marges des immigrés

Les variables intégrées dans le calage des immigrés sont les suivantes :

- sexe croisé avec la variable tranche d'âge
- couple
- type de logement
- nationalité
- pays de naissance d'origine
- région de résidence
- tranche d'unité urbaine

Les marges et leur répartition dans l'échantillon et la population figurent dans le tableau ci-après.

Après plusieurs essais de méthode de calage, le calage est réalisé avec la méthode logistique, en prenant comme bornes des rapports de poids 0,4 et 2,9. Cette méthode évite d'avoir des poids négatifs et des valeurs de rapports de poids trop élevés ce qui est peu souhaitable. Cette méthode converge au bout de 5 itérations. L'histogramme du rapport des poids conserve une allure relativement gaussienne (cf. figure ci-après). Les distributions du rapport des poids et des poids après calage sont présentées dans les tableaux ci-après.

Tableau 34 : Marges et répartition avant calage des immigrés

Variable	Modalité	Marge échantillon	Marge population	Pourcentage échantillon	Pourcentage population
Croisement sexe-âge	Homme – moins de 25 ans	132678	218573	3,7	5,2
	Homme – de 26 à 35 ans	330678	442683	9,1	10,5
	Homme – de 36 à 49 ans	731008	819412	20,2	19,4
	Homme – plus de 50 ans	538131	528488	14,8	12,5
	Femme – moins de 25 ans	156065	223705	4,3	5,3
	Femme – de 26 à 35 ans	449303	540177	12,4	12,8
	Femme – de 36 à 49 ans	790751	907083	21,8	21,5
Couple	Oui	2683728	2839206	74,0	67,2
	Non	942590	1388335	26,0	32,8
Type de logement	Maison	1233619	1217941	34,0	28,8
	Autre	2392699	3009600	66,0	71,2
Nationalité	Française	1381482	1500952	38,1	35,5
	Etrangère	2244836	2726589	61,9	64,5
Pays de naissance	01 – Algérie	75292	89401	2,1	2,1
	02 – Maroc	502302	564898	13,9	13,4
	03 – Tunisie	574012	568331	15,8	13,4
	04 – Afrique guinéenne ou centrale	142465	204564	3,9	4,8
	05 – Afrique sahélienne	361087	375282	10,0	8,9
	06 – Espagne, Italie	149895	230438	4,1	5,5
	07 – Portugal	116550	158865	3,2	3,8
	08 – Autres pays de l'UE 15	299729	328207	8,3	7,8
	09 – Autres pays de l'UE 28	172908	228220	4,8	5,4
	10 – Vietnam, Laos, Cambodge	182376	206602	5,0	4,9
	11 – Turquie	222233	194113	6,1	4,6
	12 – Chine	68278	89148	1,9	2,1
	13 – Autre Afrique	190084	214794	5,2	5,1
	14 – Autre Asie	125611	212482	3,5	5,0
	15 – Autre Europe	170883	238902	4,7	5,7
	16 – Autres pays et TOM, Suisse et AELE	272609	323294	7,5	7,7
Région de résidence	11 – Ile de France	1323442	1715348	36,5	40,6
	24 – Centre-Val de Loire	127532	121536	3,5	2,9
	27 – Bourgogne-Franche-Comté	121591	119398	3,4	2,8
	28 – Normandie	96653	104557	2,7	2,5
	32 – Hauts-de-France	230058	214067	6,3	5,1
	44 – Grand Est	252295	317014	7,0	7,5
	52 – Pays de la Loire	114306	117358	3,2	2,8
	53 – Bretagne	110168	89274	3,0	2,1
	75 – Nouvelle-Aquitaine	227812	239595	6,3	5,7
	76 – Occitanie	314524	326151	8,7	7,7
	84 – Auvergne-Rhône-Alpes	443949	501469	12,2	11,9
	93 – PACA	255511	338392	7,1	8,0
	94 – Corse	8477	23382	0,2	0,6
Tranche d'unité urbaine	Rural (< 2000 habitants)	215105	285954	5,9	6,8
	Autre	3411212	3941587	94,1	93,2

Figure 1 : Histogramme du rapport des poids pour le calage des immigrés

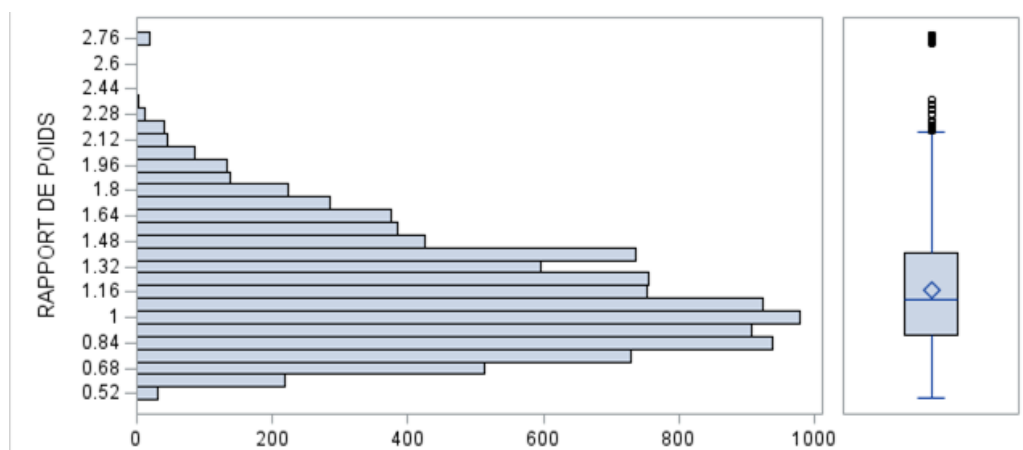


Tableau 35 : Distribution du rapport des poids pour le calage des immigrés

Moyenne	1,17 ⁸²
Maximum	2,78
99%	2,10
95%	1,83
90%	1,67
75% Q3	1,40
50% Médiane	1,11
25% Q1	0,89
10%	0,75
5%	0,68
1%	0,60
Minimum	0,49

Tableau 36 : Distribution des poids calés des immigrés

Moyenne	413
Maximum	12477
99%	3902
95%	1230
90%	794
75% Q3	412
50% Médiane	214
25% Q1	121
10%	71
5%	53
1%	32
Minimum	7

⁸² Le calage permet de corriger un défaut de couverture ce qui amène à avoir une moyenne du rapport des poids supérieurs à 1. Cette remarque est valable pour les calages à suivre.

4.5.3 Calage sur marge des Domiens

Les variables intégrées dans le calage des Domiens sont les suivantes :

- sexe croisé avec la variable tranche d'âge
- couple
- type de logement
- région de résidence
- tranche d'unité urbaine

Les marges et leur répartition dans l'échantillon et la population figurent le tableau ci-après.

Tableau 37 : Marges et répartition avant calage des domiens

Variable	Modalité	Marge échantillon	Marge population	Pourcentage échantillon	Pourcentage population
Croisement sexe-âge	Homme – moins de 25 ans	16081	31466	5,7	9,6
	Homme – de 26 à 35 ans	35478	43695	12,6	13,3
	Homme – de 36 à 49 ans	49612	51111	17,6	15,5
	Homme – plus de 50 ans	34766	33296	12,3	10,1
	Femme – moins de 25 ans	23818	36838	8,5	11,2
	Femme – de 26 à 35 ans	37972	44942	13,5	13,7
	Femme – de 36 à 49 ans	50037	49282	17,7	15,0
Couple	Oui	163883	168786	58,1	51,3
	Non	118096	160451	41,9	48,7
Type de logement	Maison	115795	97508	41,1	29,6
	Autre	166184	231729	58,9	70,4
Région de résidence	11 – Ile de France	94482	135433	33,5	41,1
	24 – Centre-Val de Loire	8425	13376	3,0	4,1
	27 – Bourgogne-Franche-Comté	6329	7362	2,2	2,2
	28 – Normandie	17163	8823	6,1	2,7
	32 – Hauts-de-France	11747	11987	4,2	3,6
	44 – Grand Est	23648	15541	8,4	4,7
	52 – Pays de la Loire	21011	14982	7,5	4,6
	53 – Bretagne	19220	13828	6,8	4,2
	75 – Nouvelle-Aquitaine	18861	27199	6,7	8,3
	76 – Occitanie	27726	32762	9,8	10,0
	84 – Auvergne-Rhône-Alpes	22599	30965	8,0	9,4
93-94 – PACA Corse	10767	16979	3,8	5,2	
Tranche d'unité urbaine	0 – Commune rurale (< 2000 habitants)	31293	31106	11,1	9,5
	Autre	250686	298131	88,9	90,6

Après plusieurs essais de méthode de calage, le calage est réalisé avec la méthode logistique, en prenant comme bornes des rapports de poids 0,4 et 2,4. Cette méthode converge au bout de 6 itérations. L'histogramme des rapports de poids est présenté dans la figure ci-après. Les distributions du rapport des poids et des poids après calage sont présentées dans les tableaux ci-après.

Figure 2 : histogramme du rapport des poids pour le calage des Domiens

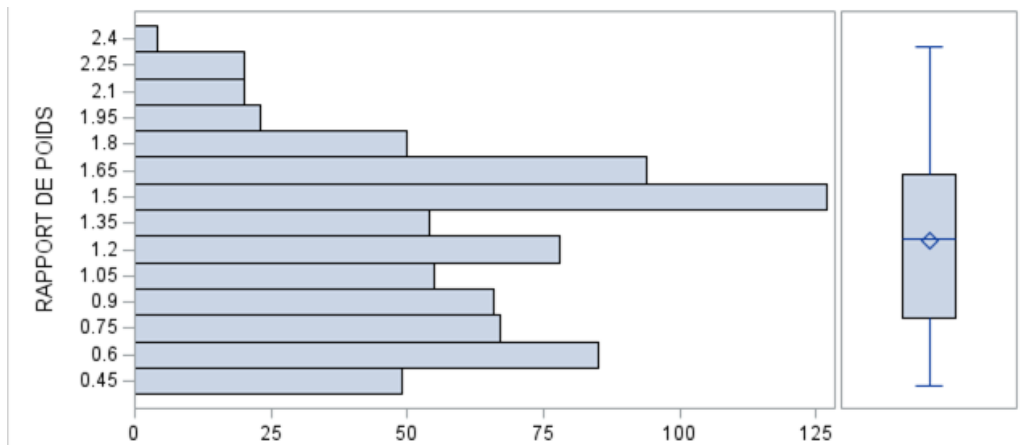


Tableau 38 : distribution du rapport des poids pour le calage des Domiens

Moyenne	1,25
Maximum	2,35
99%	2,30
95%	2,05
90%	1,82
75% Q3	1,62
50% Médiane	1,26
25% Q1	0,81
10%	0,59
5%	0,51
1%	0,45
Minimum	0,43

Tableau 39 : distribution des poids calés des Domiens

Moyenne	416
Maximum	5128
99%	2931
95%	1110
90%	725
75% Q3	469
50% Médiane	287
25% Q1	185
10%	123
5%	98
1%	68
Minimum	22

4.5.4 Calage sur marge des « autres »

Les variables intégrées dans le calage des « autres » populations sont les suivantes :

- sexe croisé avec la variable tranche d'âge
- couple
- type de logement
- nationalité
- région de résidence
- tranche d'unité urbaine

Les marges et leur répartition dans l'échantillon et la population figurent le tableau ci-après.

Tableau 40 : Marges et répartition avant calage des « autres »

Variable	Modalité	Marge échantillon	Marge population	Pourcentage échantillon	Pourcentage population
Croisement sexe-âge	Homme – moins de 25 ans	2555796	2665131	9,4	9,3
	Homme – de 26 à 35 ans	3002404	3171722	11,1	11,1
	Homme – de 36 à 49 ans	4672418	4783394	17,2	16,7
	Homme – plus de 50 ans	3202310	3581442	11,8	12,5
	Femme – moins de 25 ans	2612087	2596565	9,6	9,0
	Femme – de 26 à 35 ans	3192331	3251661	11,8	11,3
	Femme – de 36 à 49 ans	4446159	4906656	16,4	17,1
	Femme – plus de 50 ans	3468056	3751982	12,8	13,1
Couple	Oui	17850740	16890567	65,7	58,8
	Non	9300820	11817986	34,3	41,2
Type de logement	Maison	17710191	17508111	65,2	61,0
	Autre	9441369	11200442	34,8	39,0
Nationalité	Française de naissance	26762967	28273021	98,6	98,5
	Française par acquisition	310773	356205	1,1	1,2
	Etrangère	77821	79327	0,3	0,3
Région de résidence	11 – Ile de France	4120966	5004933	15,2	17,4
	24 – Centre-Val de Loire	963373	1106557	3,6	3,9
	27 – Bourgogne-Franche-Comté	1277431	1228416	4,7	4,3
	28 – Normandie	1431368	1524876	5,3	5,3
	32 – Hauts-de-France	2630412	2878291	9,7	10,0
	44 – Grand Est	2463564	2507402	9,1	8,7
	52 – Pays de la Loire	1774917	1766067	6,5	6,2
	53 – Bretagne	1874805	1539950	6,9	5,4
	75 – Nouvelle-Aquitaine	2551146	2666416	9,4	9,3
	76 – Occitanie	2271243	2592049	8,4	9,0
	84 – Auvergne-Rhône-Alpes	3868967	3604213	14,3	12,6
Tranche d'unité urbaine	1 – Commune rurale (< 2000 habitants)	6420548	6764716	23,7	23,6
	2 – UU de 2 000 à moins de 100 000 habitants	8570625	8657971	31,6	30,2
	3 – UU de 100 000 à moins de 2 000 000 habitants	9413731	8949154	34,7	31,2
	4 – Agglomération de Paris	2746656	4336712	10,1	15,1

Après plusieurs essais de méthode de calage, le calage est réalisé avec la méthode logistique, en prenant comme bornes des rapports de poids 0,4 et 2,5. Cette méthode converge au bout de 6 itérations. L'histogramme du rapport des poids conserve une allure relativement gaussienne (cf. figure ci-après). Les distributions du rapport des poids et des poids après calage sont présentées dans les tableaux ci-après.

Figure 3 : histogramme du rapport des poids pour le calage des « autres »

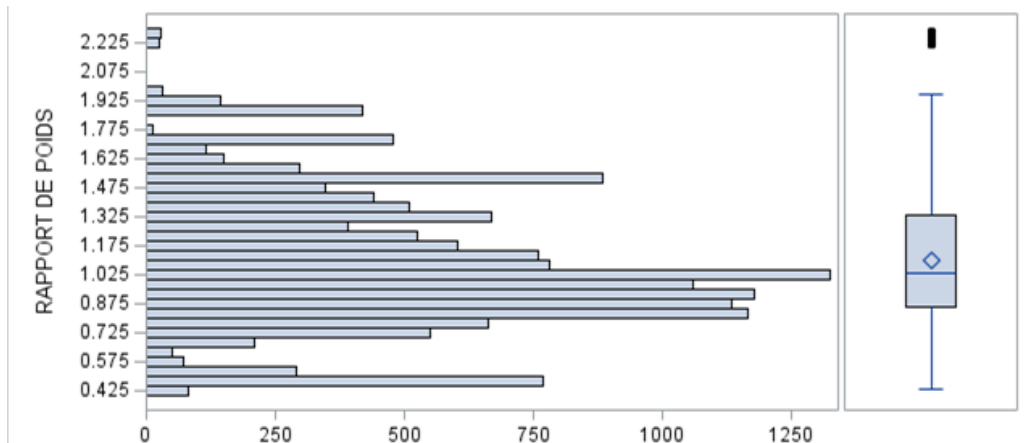


Tableau 41 : distribution du rapport des poids pour le calage des « autres »

Moyenne	1,10
Maximum	2,28
99%	1,91
95%	1,71
90%	1,56
75% Q3	1,34
50% Médiane	1,03
25% Q1	0,86
10%	0,71
5%	0,48
1%	0,45
Minimum	0,43

Tableau 42 : distribution des poids calés des « autres »

Moyenne	1778
Maximum	56407
99%	16616
95%	7322
90%	5538
75% Q3	1159
50% Médiane	394
25% Q1	196
10%	94
5%	60
1%	28
Minimum	5

Pour tester la cohérence et la qualité des pondérations, des validations croisées sont effectuées en annexe 10. Elles comparent les résultats obtenus sur des variables annexes, c'est-à-dire n'ayant pas servi aux calages, à partir de TeO2 et d'une source externe, ici l'enquête Emploi.

Annexes

Annexe 1 – Calendrier détaillé des tests et collecte

TeO - FOCUS GROUPE		
Étapes	Dates	
Livraison du paramétrage Opale	10 avril 2017	
Intégration de l'échantillon dans Opale	10 avril 2017	19 avril 2017
Finalisation du questionnaire	début septembre 2017	
Rédaction des instructions de collecte et de la grille d'évaluation	début septembre 2017	fin septembre 2017
Élaboration du module de formation	début septembre 2017	fin septembre 2017
Formation	3 octobre 2017	
Collecte terrain	9 octobre 2017	17 novembre 2017
Bilan	17 novembre 2017	
Diffusion de la synthèse du focus group	1 décembre 2017	8 décembre 2017
Prise en compte des remarques du focus group	mi-novembre 2017	fin janvier 2018
TeO - TEST Consultation Mairies		
Étapes	Dates	
sous-échantillons descendants		
Autorisation des procureurs	été 2017	
1 ^{er} tirage dans le RP	fin octobre 2017	
Remise des spécifications du questionnaire	14 septembre 2017	
Remise des spécifications postes de gestion	5 octobre 2017	
Livraison du DM à CAPI	19 octobre 2017	
Livraison de l'échantillon à Opale	4 décembre 2017	
Remise des instructions aux DR (y compris listes des personnes à rechercher, courriers aux petites communes)	22 décembre 2017	
Livraison du poste de collecte	8 janvier 2018	
Livraison des postes de gestion	8 janvier 2018	
Intégration de l'échantillon dans Capi	début décembre 2017	mi-décembre 2017
Relevés en mairie	15 janvier 2018	16 mars 2018
Réunions de Bilan	22 mars 2018	
Spécifications aval Capi	fin novembre 2017	
Aval Capi	mi-mars 2018	
Livraison du fichier	fin mars 2018	
Diffusion du bilan de test	fin mars 2018	
Étapes	Dates	
Mise à disposition échantillon Opale aux DR	janvier 2018	
Remise des spécifications du questionnaire - lot 1	31 janvier 2018	
Remise des spécifications du questionnaire - lot 2	19 février 2018	28 février 2018
1 ^{ère} livraison DM	1 mars 2018	
Remise des spécifications pour le contrôle des dates et le questionnaire qualité	19 février 2018	semaine du 20 mars
Finalisation des documents à traduire	Fin mars 2018	
Priorisation des corrections sur le DM	22 mars 2018	13 avril 2018
Tirage sous-échantillons domiens, immigrés, témoins dans RP	3 avril 2018	13 avril 2018
Tirage sous-échantillons descendants suite opération mairies	3 avril 2018	13 avril 2018
Livraison échantillon par la division sondages	16 avril 2018	

Répartition de l'échantillon par les Dem	28 mars 2018	11 avril 2018
Livraison des spécifications aval	début avril 2018	
Envoi des fichiers au CNIN pour impression des FA	16 avril 2018	
Remise des spécifs Capi	16 avril 2018	
Préparation de l'échantillon de l'opération terrain	16 avril 2018	20 avril 2018
Finalisation des documents de collecte (hors traduction)	13 avril 2018	
Validation DM	20 avril 2018	
Livraison du DM à l'Ined	23 avril 2018	
Livraison du DM à CAPI	23 avril 2018	
Saisie des exercices de formation sur le DM final et livraison à Capi	23 avril 2018	
Impression FA (envoi réel aux DRs)	fin avril 2018	
Test du poste de collecte	26 avril 2018	4 mai 2018
Test du poste de gestion	7 mai 2018	16 mai 2018
Livraison des postes de collecte	11 mai 2018	
Livraison des postes de gestion	18 mai 2018	
Finalisation de la formation (hors exercices)	début mai 2018	
Formation des enquêteurs DR IdF	14 et 15 mai 2018	
Formation des enquêteurs DR Marseille	14 et 15 mai 2018	
Formation des enquêteurs DR Reims	17 et 18 mai 2018	
Opération terrain	22 mai 2018	13 juillet 2018
Bilan DR IdF	12 juillet 2018	
Bilan DR Reims	11 juillet 2018	
Bilan DR Marseille	13 juillet 2018	
Synthèse des réunions de bilan du test Capi	16 juillet 2018	20 juillet 2018
Livraison des fichiers ADT	16 juillet 2018	
Analyse des temps ADT	16 juillet 2018	20 juillet 2018
Arbitrage des remarques du test Capi	23 juillet 2018	7 septembre 2018
Prise en compte des évolutions dans les spécifications et les documents	10 septembre 2018	4 octobre 2018
Dépôt du dossier Label	11 octobre 2018	
Pré-Label	30 octobre 2018	
Passage au Label	14 novembre 2018	
Prise en compte des remarques du Label	fin octobre 2018	mi-novembre 2018
Aval CAPI	mi-juillet 2018	octobre 2018
Test CAPI Non francophones	3 septembre 2018	17 septembre 2018
Formation des enquêteurs-traducteurs de l'Ined	3 septembre 2018	4 septembre 2018
Bilan test CAPI-NF	17 septembre 2018	
Étapes	Dates	
Remise des modifications des spécifications du questionnaire	mi-octobre 2018	
1 ^{ère} livraison DM	début novembre 2018	
Validation DM	fin décembre 2018	
Livraison du DM à CAPI	fin décembre 2018	
Remise des spécifications des postes de collecte et de gestion	fin décembre 2018	
Livraison des spécifications aval	début janvier 2019	mi-janvier 2019
Livraison du poste de collecte	21 janvier 2019	25 janvier 2018
Livraison des postes de gestion	21 janvier 2019	25 janvier 2018
Mise à disposition de l'échantillon	21 janvier 2019	25 janvier 2018
Opération en DR	28 janvier 2019	8 février 2019
Bilan	mi-février 2019	

Aval CAPI	mi-février 2019	fin mars 2019
Test des reprises	avril 2019	
TeO - Échantillon		
Étapes	Dates	
Déclaration Cnil pour l'échantillon	novembre 2018	décembre 2018
sous-échantillons domiens / immigrés / témoins - vague 1		
Calcul des allocations provisoires	mi-octobre 2018	
Livraison de l'échantillon provisoire à Opale	mi-octobre 2018	
Tirage de l'échantillon dans RP		
Envoi au CNIN	début mars 2019	
Impression et envoi des Fiches-adresses aux DR	mi-mars 2019	fin mars 2019
Livraison de l'échantillon définitif à Opale	1 avril 2019	1 mai 2019
Répartition dans Opale	1 avril 2019	mi-mai 2019
Export Opale-Capi	début mai 2019	mi-mai 2019
sous-échantillons descendants - vague 2		
Spécification du tirage de l'échantillon mairie	fin novembre 2018	début décembre 2018
Préparation de la base de sondage pour l'échantillon mairie	novembre 2018	décembre 2018
1 ^{er} tirage dans le RP (descendants à rechercher)	décembre 2018	janvier 2019
Contrôles et mises aux normes du fichier	décembre 2018	janvier 2019
Communication du nombre de relevés à effectuer par mairie aux SES	janvier 2019	
Transmission des noms et prénoms des personnes responsables des relevés	début février 2019	fin février 2019
Envoi au TGI de Paris des noms des personnes responsables des relevés	février 2019	
Livraison de l'échantillon à Opale	mi-mars 2019	mi-avril 2019
Répartition dans Opale	mi-avril 2019	mi-mai 2019
Remise des spécifications du questionnaire mairie	fin janvier 2019	fin mars 2019
Remise des spécifications des postes de gestion pour l'opération mairie	début avril 2019	
Livraison du DM mairie à CAPI	mi-avril 2019	
Livraison du poste de collecte mairie	mi-mai 2019	
Livraison postes de gestion mairie	mi-mai 2019	
Remise des instructions aux DR (y compris listes des personnes à rechercher, courriers aux petites communes)	début mai 2019	mi-mai 2019
Production du fichier de données antérieures	fin mars 2019	début mai 2019
Export Opale-Capi	mi-mai 2019	
Remise de l'échantillon aux DR	mi-mai 2019	mi-mai 2019
Formation		
Relevés en mairie	1 juin 2019	août 2019
Réunions de Bilan	1 septembre 2019	
Spécifications aval Capi	début mai 2019	
Aval Capi	septembre 2019	
Livraison du fichier des descendants avec origine	septembre 2019	fin septembre 2019
Impression et envoi des FA vague 2 aux DR	début octobre 2019	fin octobre 2019
Livraison des échantillons vague 2 à Opale	début octobre 2019	fin octobre 2019
Répartition vague 2 dans Opale	début octobre 2019	mi-décembre 2019
Export vague 2 Opale-Capi	début décembre 2019	mi-décembre 2019
TeO - ENQUETE		

Étapes	Dates	
Lancement du marché pour les cadeaux	octobre 2018	
GT sur le vademecum suivi des personnes et mail type	novembre 2018	décembre 2018
Choix des langues pour les traductions	20 novembre 2018	
Validation du contenu des documents à traduire (ou rétro traduire)	20 novembre 2018	
Lancement du marché de traduction	début janvier 2019	
Envoi du DCPOD et étude d'impact au DPD	20 décembre 2018	
Traduction des documents	novembre 2018	janvier 2019
Production d'un 4 pages de résultats	décembre 2018	janvier 2019
Envoi du dossier Cnil (échantillon + enquête)	fin janvier 2019	
Remise des modifications des spécifications du questionnaire	mi-février 2019	
Retour sur les rétro-traductions	11 mars 2019	
Validation des documents traduits	20 mars 2019	
Finalisation des documents de collecte pour impression	25 mars 2019	
1 ^{ère} livraison DM	5 mars 2019	
Tournage des vidéos de formation	1 ^{ère} semaine d'avril 2019	
Montage des séquences vidéo	avril 2019	
Validation des séquences vidéos "formation"	mi-mai 2019	
Validation des séquences vidéos "grand public" : présentation générale	mi-mai 2019	
Validation des séquences vidéos "grand public" : données sensibles	1 septembre 2019	
Livraison des documents de collecte en Dem	début mai 2019	
Validation DM	26 avril 2019	
Livraison du DM à CAPI	26 avril 2019	
Remise des spécifications des postes gestion et concepteur	26 avril 2019	
Livraison Poste de collecte vague 1	9 mai 2019	
Export Opale-Capi	17 juin 2019	
Livraison Poste de gestion vague 1	20 juin 2019	
Livraison Poste concepteur vague 1	25 juin 2019	
Réception des cadeaux en DR	fin mai 2019	
Formation des gestionnaires	13 mai 2019	22 mai 2019
Formation des enquêteurs	juin 2019	juin 2019
Ouverture du site TeO2	10 mai 2019	
Opération terrain - Vague 1	1 juillet 2019	31 décembre 2019
Analyse des temps ADT	2 septembre 2019	
Tirage de l'échantillon vague 2	1 septembre 2019	25 octobre 2019
Livraison de l'échantillon vague 2 à Opale	6 novembre 2019	
Analyse des temps ADT	mi-novembre 2019	
Production des DLM pour les FA papier	12 novembre 2019	21 novembre 2019
Validation des BAT des FA	22 novembre 2019	
Impression et livraison en Dem	28 novembre 2019	6 décembre 2019
Fin de répartition Opale	11 décembre 2019	
Export Opale - Capi vague 2	12 décembre 2019	
Production des échantillons Capi	12 novembre 2019	13 décembre 2019
Livraison des postes de collecte, gestion et concepteur vague 2	16 décembre 2019	17 décembre 2019
Transmission des dernières FA vague 1	31 décembre 2019	

Apurement vague 1	1 juillet 2019	7 janvier 2020
Contrôles d'exhaustivité vague 1	8 janvier 2020	10 janvier 2020
Production du retour du confié	13 janvier 2020	14 janvier 2020
Intégration du retour du confié dans Opale	15 janvier 2020	
Opération Non francophones	13 février 2020	30 août 2020
Formation enquêteurs-traducteurs Ined	10 février 2020	12 février 2020
Collecte groupe 1	13 février 2020	9 juillet 2020
Bilan groupe 1	10 juillet 2020	
Livraison à l'Insee des données collectées : groupe 1	16 juillet 2020	
Collecte groupe 2	13 février 2020	30 août 2020
Bilan groupe 2	31 août 2020	
Livraison à l'Insee des données collectées : groupe 2	2 septembre 2020	
Opération terrain - Vague 2	1 janvier 2020	30 novembre 2020
Livraison du poste de collecte G3	début février 2020	mi-février 2020
Extraction aval vague 2	24-févr	
Sélection des identifiants à conserver dans le lot	24-févr.	27-févr.
Ajout des clés et production du csv de publipostage pour les FA papier	27-févr.	
Mise au format Opale	27-févr.	
Livraison de l'échantillon Opale à Opale	27-févr.	
Livraison du csv de publipostage au PENM (table éligible + clé + indicatrice de sélection dans le lot)	27-févr.	
Intégration Opale	28-févr.	
Production et envoi des pdf en Dem	4-mars	
Répartition par les Dem	28-févr.	4-mars
Export Opale-Capi (fichier .csv)	5-mars	
Livraison du .tot à Capi	5-mars	
Mise à disposition des postes Capi et échantillons en Dem	5-mars	9-mars
Opération terrain - G3 lot 1	10 mars 2020	15 décembre 2020
Extraction aval vague 2	7 septembre 2020	
Sélection des identifiants à conserver dans le lot	7-sept.	10-sept.
Ajout des clés et production du CSV de publipostage pour les FA papier	15-sept.	
Mise au format Opale	15-sept.	
Livraison de l'échantillon Opale à Opale	16-sept.	
Livraison du csv de publipostage au PENM (table éligible + clé + indicatrice de sélection dans le lot)	18-sept.	
Intégration Opale	18-sept.	
Production et envoi des pdf en Dem	21-sept.	
Répartition par les Dem	18-sept.	24-sept.
Export Opale-Capi (fichier .csv)	25-sept.	
Livraison du .TOT à Capi	25-sept.	
Mise à disposition des postes Capi et échantillons en Dem	25-sept.	27-sept.
Opération terrain - G3 lot 2	28 septembre 2020	15 décembre 2020
Extraction aval vague 2	19 octobre 2020	
Sélection des identifiants à conserver dans le lot	19-oct.	21-oct.
Ajout des clés et production du csv de publipostage pour les FA papier	21-oct.	
Mise au format Opale	21-oct.	
Livraison de l'échantillon Opale à Opale	21-oct.	

Livraison du CSV de publipostage au PENM (table éligible + clé + indicatrice de sélection dans le lot)	22-oct.	
Intégration Opale	21-oct.	
Production et envoi des pdf en Dem	22-oct.	
Répartition par les Dem	22-oct.	26-oct.
Export Opale-Capi (fichier .csv)	27-oct.	
Livraison du TOT à Capi	27-oct.	
Mise à disposition des postes Capi et échantillons en Dem	28-oct.	28-oct.
Opération terrain - G3 lot 3	29 octobre 2020	15 décembre 2020
Extraction aval vague 2	11 décembre 2020	
Sélection des identifiants à conserver dans le lot	11-déc.	11-déc.
Ajout des clés et production du csv de publipostage pour les FA papier	11-déc.	
Mise au format Opale	11-déc.	
Livraison de l'échantillon Opale à Opale	14-déc.	
Livraison du csv de publipostage au PENM (table éligible + clé + indicatrice de sélection dans le lot)	14-déc.	
Intégration Opale	14-déc.	
Production et envoi des pdf en Dem	14-déc.	
Répartition par les Dem	14-déc.	16-déc.
Export Opale-Capi (fichier .csv)	16-déc.	
Livraison du.TOT à Capi	17-déc.	
Mise à disposition des postes Capi et échantillons en Dem	17-déc.	21-déc.
Opération terrain - G3 lot 4	1 janvier 2021	31 janvier 2021
Apurement en DR	juillet 2019	8 décembre 2020
Codification et reprise Sicore (CS et diplôme) - lot 1 (FA vague 1)	octobre 2019	avril 2020
Codification et reprise Sicore (CS et diplôme) - lot 2 (reste FA vague 1)	janvier 2020	mai 2020
Codification et reprise Sicore (CS et diplôme) - lot 3 (FA vague 2)	avril 2020	fin décembre 2020
Codification et reprise Sicore (CS et diplôme) - lot 3bis (vague 1 FA NF)	septembre 2020	octobre 2020
Codification et reprise Sicore (CS et diplôme) - lot 4 (FA vague 2)	mi-décembre 2020	fin mars 2021
Codification et reprise Sicore (CS et diplôme) - lot 5 (lots 1,2 et 3 des G3)	mi-janvier 2021	mi-avril 2021
Codification et reprise Sicore (CS et diplôme) - lot 6 (lot 4 des G3)	mi-février 2021	mi-avril 2021
Livraison des tables aval Capi (ou livraison du fichier intermédiaire)	décembre 2020	fin mai 2021
Redressement, calage hors 3G	décembre 2020	août 2021
Livraison du fichier d'études aux 2 MOA hors 3G	début sept 2021	
1ers résultats hors 3G	début mars 2022	
Redressement, calage 3G	mars 2021	décembre 2021
Livraison du fichier d'études yc 3G aux 2 MOA	juin 2022	
Livraison du fichier d'études yc 3G au CASD	juin 2022	
Livraison du FPR à Quételet	fin 2022	

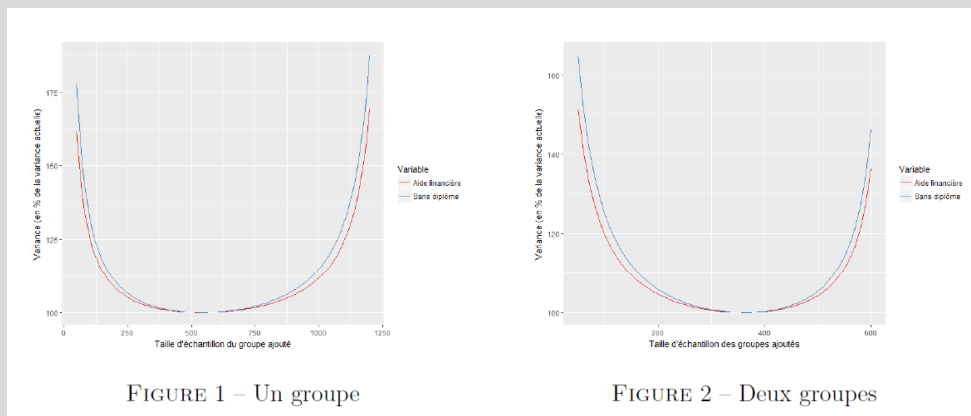
Annexe 2 – Simulation de l'impact de l'intégration de nouveaux groupes d'origines sur la précision globale

Encadré : Simulation groupe d'origines

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 $n_r \approx 1200 - n_1$;
2. deux groupes, avec des tailles d'échantillons égales $n_1 = n_2$, ce qui réduit la taille de l'échantillon pour le groupe résiduel à environ $n_r \approx 1200 - (n_1 + n_2)$

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.



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.

Annexe 3 – Méthodologie échantillon complémentaire vague 1

La logique du tirage a suivi en grande partie celle décrite pour les sous échantillons de la vague 1 (trois sous-échantillons, calcul des allocations, etc.) à quelques exceptions près :

Allocation échantillon complémentaire

Sous-échantillon (origines)	Part des quatre générations (79, 80, 89, 90)	Nombre de FA complémentaires
01 – Immigrés	9,84 %	1 656
Algérie	8,91 %	168
Maroc, Tunisie	10,70 %	214
Afrique Sahélienne	11,21 %	195
Afrique centrale et du golfe de Guinée	11,18 %	175
Espagne, Italie, Portugal	6,63 %	109
Autres pays de l'UE 28	9,38 %	127
Turquie	10,68 %	147
Asie du Sud-Est	6,19 %	91
Chine	11,83 %	153
Pays avec de nombreux réfugiés	11,18 %	144
Autres pays du monde	11,09 %	133
02 – Domiens	10,62 %	159
05 – Population générale	9,19 %	585

Les jours de naissance mobilisés sont étendus aux jours entre le 14 et le 19 de chaque mois pour les origines suivantes :

- « Asie du Sud-Est »
- « Chine »
- « Turquie », ce qui est différent de l'échantillon principal

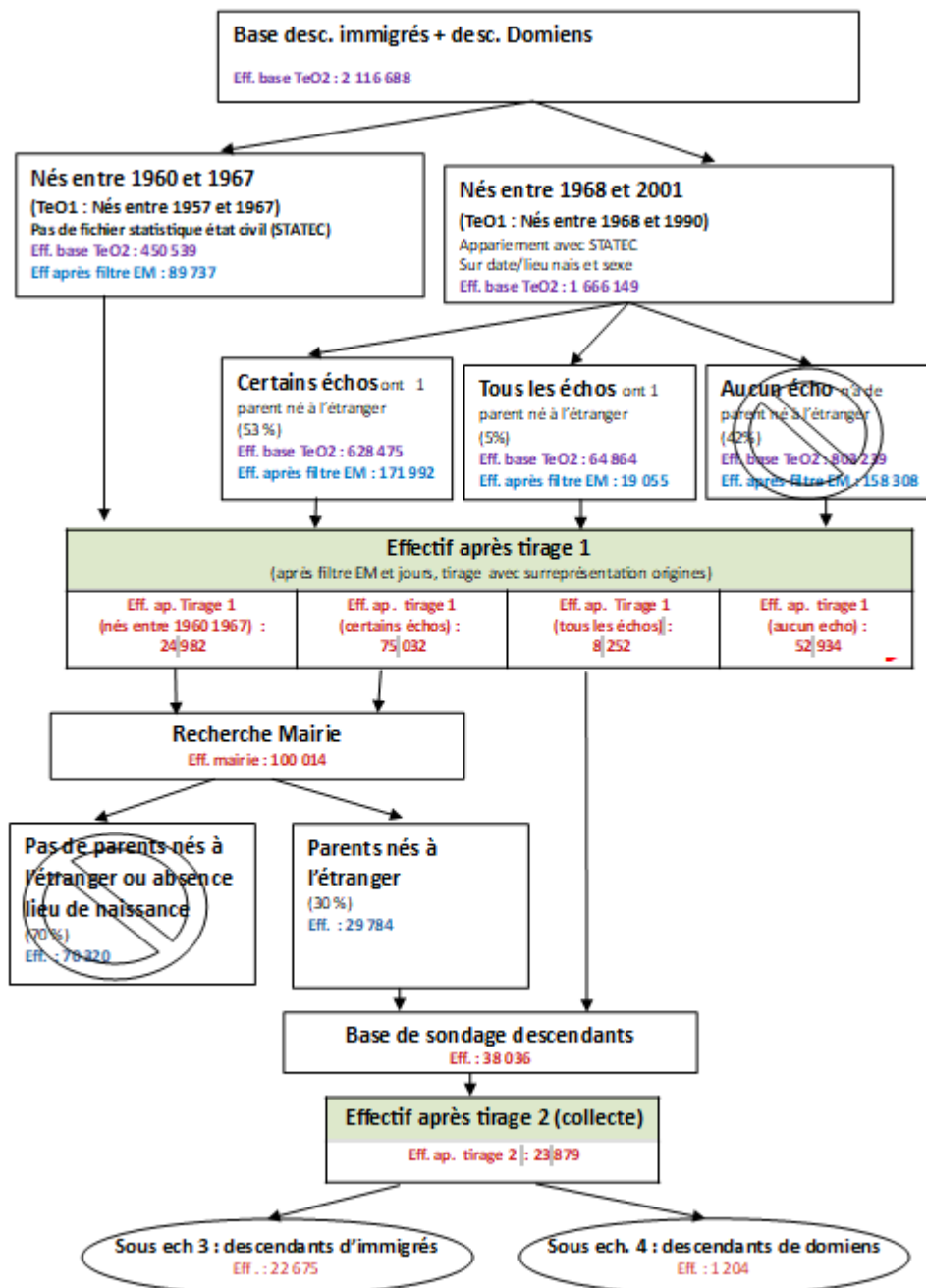
- Le tirage se fait au sein de l'Échantillon-Maître Octopusse, sauf pour les individus de l'origine « Asie du Sud-Est » ; contrairement à l'échantillon principal, les origines « Chine » et « Turquie » ne demandent pas d'extension.

- La gestion de la disjonction entre le sous-échantillon 05 et les autres, ainsi que la gestion de l'interrogation d'un seul individu par logement a été effectuée mais uniquement parmi les individus nés dans les quatre générations nécessitant le complémentaire.

Annexe 4 – Schéma du processus d'échantillonnage des G2

Le schéma ci-après résume le processus d'échantillonnage des G2

Processus d'échantillonnage des G2



Annexe 5 – Estimation de l'allocation par cible de répondants et origines géographiques

Encadré : Estimation de l'allocation par cible de répondants et origines géographiques

Avant le tirage de l'échantillon de la vague 2 (descendants) il est nécessaire de calibrer les allocations de FA par groupe d'origine. Cette partie décrit trois scénarios qui permettent d'estimer le nombre de FA à mettre en collecte pour atteindre un nombre suffisant de répondants par groupe d'origines.

On détermine le nombre de FA à mettre en collecte pour chaque groupe d'origine, en fonction de 2 paramètres :

- le taux de réponse anticipé pour chaque groupe respectif
- le taux d'« enfants d'immigrés » parmi les enfants de personnes nées à l'étranger pour chaque groupe respectif. Ce dernier paramètre est important à prendre en compte dans la mise en collecte des FA car une partie de ces FA sera couverte par les enfants de parents français nés à l'étranger¹

Ces deux paramètres varient selon les sources utilisées et les hypothèses effectuées.

3 scénarios sont testés :

1. Scénario 1 : utiliser TeO1 pour estimer un
 - taux de réponse anticipé pour chaque groupe respectif. On reprend ici le taux de réponse de TeO1
 - taux d'« enfants d'immigrés » parmi les enfants de personnes nées à l'étranger. On se cale sur le ratio (nombre de questionnaires courts/ nombre de questionnaires courts + nombre de questionnaires longs) de l'enquête TeO1 qui donne une approximation du nombre de FA d'enfants de parents français nés à l'étranger.
2. Scénario 2 : utiliser un
 - taux de réponse anticipé des descendants (source TeO2) repris des taux de réponse anticipées de la vague 1 de TeO2 faute d'information nouvelle ou récente. Ces taux sont plus faibles que ceux de la vague 2 de TeO1 (hypothèse pessimiste)
 - taux d'enfants d'immigrés parmi les enfants de parents nés à l'étranger (source EEC)
3. Scénario 3 : utiliser un
 - taux de réponse anticipé des descendants (source TeO2) repris des taux de réponse anticipées de la vague 1 de TeO2 faute d'information nouvelle ou récente. Ces taux sont plus faibles que ceux de la vague 2 de TeO1 (hypothèse pessimiste)
 - taux d'enfants d'immigrés parmi les enfants de parents nés à l'étranger (source TeO1, cf supra).

Les scénarios 2 et 3 utilisent 2 sources pour l'estimation alors que le scénario n'en utilise qu'une seule. Par ailleurs, bien que l'EEC présente un avantage sur TeO1 pour estimer le taux d'enfants d'immigrés car plus récente, les taux bruts issus de l'EEC sont particulièrement bas, car la définition d'un descendant d'immigré(s) n'est pas la même que celle de l'enquête TeO. L'instruction des scénarios met en évidence que quelque soit le scénario envisagé, le stock estimé de FA à mettre en collecte est suffisant pour chaque groupe d'origines.

Si utiliser l'enquête TeO1 pour estimer le nombre de FA à mettre en collecte présente certaines limites :

- les taux de réponse et d'enfants d'immigrés datent de 2009
- l'enquête TeO2 est désormais obligatoire alors que TeO1 ne l'était pas
- la tranche d'âge des descendants a été étendue de 50 ans à 60 ans dans TeO2

Cela rend ces taux difficilement anticipables. Toutefois, on peut faire l'hypothèse que la hiérarchie du taux de réponse selon les groupes d'origines observés dans TeO1 sera conservée dans TeO2. Au final, c'est le scénario 1 qui a été retenu (utiliser TeO1).

Annexe 6 – Estimation du nombre d'individus de 3ème génération maghrébine et non européenne

Un cadrage démographique est réalisé à partir de l'enquête TeO1 et de l'enquête emploi 2017 afin d'estimer le nombre de d'individus de 3ème génération maghrébine et non européenne.

1. Chiffrage TeO1

Estimations fondées sur la base enfants de TeO1, avec clauses d'âge (nés avant 2001), lieu de naissance (nés en France) dont le parent enquêté (ego) est descendant d'immigré(s) et dont le conjoint actuel n'est pas immigré.

Le nombre brut d'enfants G3 dans TeO1 est de 3043, dont 692 ont un parent descendant d'immigré(s) maghrébin(s) et 985 un parent descendant d'immigré(s) non européen(s). **Les G3 d'origine maghrébine représentent donc 70,2% des G3 d'origine non européenne.**

Avec l'échantillon enquêté dans TeO1 où les G2 maghrébins représentent 50,2% des non européens, le nombre de G3 maghrébins enquêtés parmi 500 G3 non européens (objectif fixé à l'enquête G3) **serait de 351.**

Cependant, la structure de l'échantillon de TeO2 diffère pour l'instant de celui de TeO1. En utilisant la répartition annoncée des objectifs par groupes d'origine en G2 où **les G2 maghrébins (2800) représentant 42,4% des non européens (6600)**, on peut recalculer les effectifs attendus. Pour ce faire, on estime un ratio entre le nombre d'individus G2 dans TeO1 et le nombre d'enfants G3 éligibles correspondant dans l'enquête :

Maghrébins : 2433 G2 ont eu 692 enfants éligibles, soit un ratio de 0,28
Non européens : 4849 G2 ont eu 985 enfants éligibles, soit un ratio de 0,20

Les effectifs attendus pour TeO2 sont un peu différents :

Maghrébins : $2800 \times 0,28 = 784$
Non européens : $6600 \times 0,20 = 1360$

Le poids des Maghrébins parmi les non Européens est alors de 57,6%, ce qui signifie que sur 500 G3 non européens on peut s'attendre à trouver **288 G3 d'origine maghrébine.**

2. Chiffrage EEC 2017

Le tableau ci-dessous présente la répartition des descendants d'immigrés non européens et maghrébins en 2017 :

Tableau 1 : Répartition des descendants d'immigrés non européens et maghrébins en 2017 selon la tranche d'âge

	Non européens	Maghrébins	Total descendants
18-29 ans	46%	36%	35%
30-39 ans	28%	29%	25%
40-49 ans	17%	23%	23%
50-59 ans	9%	12%	17%
Total 18-59 ans	100%	100%	100%

Champ : ménages ordinaires, France métropolitaine,
Source enquête emploi 2017

Lecture : en 2017, 46 % des descendants d'immigrés non européens de 18-59 ans étaient âgés de 18 à 29 ans.

Dans un premier temps, nous appliquons cette répartition par tranche d'âge aux objectifs attendus pour les descendants non européens (6600) et maghrébins (2800).

Tableau 2 : Répartition du nombre attendu de descendants d'immigrés non européens et maghrébins selon la tranche d'âge dans TeO2

	Non européens	Maghrébins
18-29 ans	3031	1012
30-39 ans	1865	823
40-49 ans	1113	635
50-59 ans	591	330
Total 18-59 ans	6600	2800

Champ : ménages ordinaires, France métropolitaine,
Source enquête emploi 2017, enquête TeO2

Le nombre de 3G non européens et maghrébins est estimé à partir d'hypothèses sur :

- le nombre moyen d'enfants âgés de 18 à 59 ans pour la population des descendants d'immigrés non européens et maghrébins approximé par celui des immigrés et des descendants (cf. tableau 3). Nous nous restreignons au champ d'âge de TeO2 i.e. 18-59 ans
 - L'hypothèse « mixte » correspond à une fécondité prise comme la moyenne entre le total des descendants et des immigrés. Cette hypothèse tend plutôt à surévaluer la fécondité des descendants non européens car d'après la littérature, cette dernière est plus proche de celle des descendants que de celles des immigrés.

Tableau 3 : Nombre moyen d'enfants selon la tranche d'âge chez les immigrés et leurs descendants

	Total Immigrés	Total descendants	Hypothèse mixte
18-29 ans	0	0	0
30-39 ans	0,04	0,02	0,03
40-49 ans	0,68	0,56	0,62
50-59 ans	1,98	1,61	1,795
Total	0,68	0,41	0,545

Champ : ménages ordinaires, France métropolitaine,
Source enquête Famille logement 2011

- la mixité du couple parentale (en couple avec population majoritaire ou avec un descendant) (cf. tableau 4).
 - Les enfants nés d'un parent immigré et d'un parent descendant n'étant pas considérés comme des personnes de 3ème génération, il est nécessaire de les retirer.

Tableau 4 : répartition des couples descendants selon le lien à la migration

Parmi les couples	Non européens	Maghrébins	Total descendants
En couple avec pop majoritaire	40%	35%	60%
En couple avec un descendant	28%	32%	21%
En couple avec un immigré	32%	33%	19%

Champ : ménages ordinaires, France métropolitaine,
Source enquête emploi 2017

Lecture : en 2017, 40 % des descendants non européens sont en couple avec une personne issue de la population majoritaire

Pour TeO2, nous estimons les effectifs attendus des 3G par tranche d'âge.

Par exemple, le nombre de 3G non européens âgés de 30-39 ans : $1865 \cdot 0.03 \cdot .40 + 1865 \cdot 0.03 \cdot 0.28 = 38$.

À noter qu'on ne divise pas par deux le nombre d'enfants issus d'un couple de descendants. En effet, comme l'enquête n'est passée qu'à une personne par ménage, il n'y a aucune chance qu'un enfant 3G aient ses deux parents dans l'échantillon de descendants.

Pour TeO2, l'effectif attendu est ainsi de :

- 3G Maghrébins : 680
- 3G non européens : 1236

Tableau 5 : Estimation du nombre de 3G non européens et maghrébins attendus selon la tranche d'âge.

	Non européens	Maghrébins
18-29 ans	0	0
30-39 ans	38	17
40-49 ans	472	265
50-59 ans	726	398
Total 18-59 ans	1236	680

Annexe 7 – Modèles de correction de la non-réponse – vague 1

Après plusieurs essais de modélisation, certaines modalités de variables ont été regroupées.

Modèles de correction de la non-réponse du sous-échantillon (01) - immigrés

Les résultats de la modélisation sont présentés dans les tableaux ci-après.

Résultats CNR sous-échantillon (01) modèle 1

Variable	Modalité	Estimate	Std, Error	z value	Pr(> z)
	(Intercept)	0,506	0,149	3,402	0,001
Sexe	Homme	Ref			
	Femme	0,198	0,076	2,596	0,009
Age	36 ans et plus	Ref			
	35 ans et moins	-0,156	0,082	-1,901	0,057
Situation d'emploi	En emploi	Ref			
	Autre	-0,202	0,078	-2,593	0,010
Niveau de diplôme	Bas ⁸³	Ref			
	Haut ⁸⁴	0,233	0,075	3,093	0,002
HLM	Non	Ref			
	Oui	-0,291	0,077	-3,760	0,000
Déménagement	Non	Ref			
	Oui	-2,224	0,092	-24,174	0,000
Région	Ile de France	Ref			
	Nord-Est	0,962	0,125	7,674	0,000
	Nord-Ouest	0,960	0,132	7,267	0,000
	Sud-Ouest	1,022	0,145	7,041	0,000
	Sud-Est	0,857	0,116	7,387	0,000
TUU ⁸⁵	UU de moins de 100 000 habitants	Ref			
	UU de 100 000 habitants et plus	-0,238	0,088	-2,711	0,007
QPV ⁸⁶	Non	Ref			
	Oui	-0,223	0,083	-2,703	0,007

⁸³ Jusqu'au CAP, BEP ou diplôme de niveau équivalent

⁸⁴ Bac et études supérieures

⁸⁵ Tranche d'unité urbaine

⁸⁶ Quartier prioritaire de la politique de la ville

Résultats CNR sous-échantillon (01) modèle 2

Variable	Modalité	Estimate	Std. Error	z value	Pr(> z)
	(Intercept)	0,658	0,150	4,399	0,000
Age	36 ans et plus	ref			
	35 ans et moins	-0,141	0,083	-1,694	0,090
Statut d'occupation du logement	Autre	Ref			
	Propriétaire	-0,349	0,128	-2,720	0,007
Situation d'emploi	En emploi	Ref			
	Autre	-0,281	0,082	-3,446	0,001
HLM	Non	Ref			
	Oui	-0,386	0,092	-4,205	0,000
Déménagement	Non	Ref			
	Oui	-2,045	0,097	-21,056	0,000
Région	Ile de France	Ref			
	Nord-Est	1,017	0,119	8,575	0,000
	Nord-Ouest	1,043	0,098	10,669	0,000
	Sud-Ouest	1,283	0,156	8,232	0,000
	Sud-Est	1,111	0,129	8,631	0,000
QPV	Non	Ref			
	Oui	-0,147	0,087	-1,679	0,093

Résultats CNR sous-échantillon (01) modèle 3

Variable	Modalité	Estimate	Std. Error	z value	Pr(> z)
	(Intercept)	0,658	0,164	4,006	0,000
Sexe	Homme	Ref			
	Femme	0,188	0,075	2,525	0,012
Age	36 ans et plus	Ref			
	35 ans et moins	-0,190	0,084	-2,273	0,023
Statut d'occupation du logement	Autre	Ref			
	Propriétaire	-0,266	0,094	-2,823	0,005
Type de logement	Maison	Ref			
	Autre	-0,316	0,094	-3,378	0,001
Situation d'emploi	En emploi	Ref			
	Autre	-0,170	0,084	-2,031	0,042
Niveau de diplôme	Bas	Ref			
	Haut	0,243	0,077	3,167	0,002
HLM	Non	Ref			
	Oui	-0,273	0,106	-2,575	0,010
Déménagement	Non	Ref			
	Oui	-2,300	0,091	-25,159	0,000
Région	Ile de France	Ref			
	Nord-Est	0,916	0,116	7,915	0,000
	Nord-Ouest	1,101	0,126	8,767	0,000
	Sud-Ouest	1,259	0,130	9,668	0,000
	Sud-Est	1,106	0,108	10,261	0,000

Résultats CNR sous-échantillon (01) modèle 4

Variable	Modalité	Estimate	Std. Error	z value	Pr(> z)
	(Intercept)	0,677	0,209	3,243	0,001
Nombre de personnes dans le ménage	Moins de 3 personnes	Ref			
	3 personnes ou plus	0,183	0,100	1,833	0,067
Statut d'occupation du logement	Autre	Ref			
	Propriétaire	-0,352	0,107	-3,282	0,001
Niveau de diplôme	Bas	Ref			
	Haut	0,261	0,090	2,895	0,004
HLM	Non	Ref			
	Oui	-0,512	0,129	-3,965	0,000
Déménagement	Non	Ref			
	Oui	-2,018	0,118	-17,065	0,000
Région	Ile de France	Ref			
	Nord-Est	1,260	0,132	9,517	0,000
	Nord-Ouest	1,076	0,130	8,299	0,000
	Sud-Ouest	0,656	0,168	3,901	0,000
	Sud-Est	0,837	0,120	6,951	0,000
TUU	UU de moins de 100 000 habitants	Ref			
	UU de 100 000 habitants et plus	-0,215	0,101	-2,126	0,034
QPV	Non	Ref			
	Oui	-0,473	0,116	-4,082	0,000

Résultats CNR sous-échantillon (01) modèle 5

Variable	Modalité	Estimate	Std. Error	z value	Pr(> z)
	(Intercept)	0,531	0,157	3,373	0,001
Statut d'occupation du logement	Autre	Ref			
	Propriétaire	-0,541	0,105	-5,143	0,000
Niveau de diplôme	Bas	Ref			
	Haut	0,366	0,093	3,946	0,000
HLM	Non	Ref			
	Oui	-0,316	0,122	-2,592	0,010
Déménagement	Non	Ref			
	Oui	-2,283	0,120	-19,042	0,000
Région	Ile de France	Ref			
	Nord-Est	0,946	0,131	7,216	0,000
	Nord-Ouest	1,138	0,134	8,516	0,000
	Sud-Ouest	0,878	0,192	4,567	0,000
	Sud-Est	0,878	0,141	6,209	0,000
QPV	Non	Ref			
	Oui	-0,267	0,117	-2,283	0,022

Modèles de correction de la non-réponse du sous-échantillon (02) - Domiens

Résultats CNR - sous-échantillon (02)

Variable	Modalité	Estimate	Std. Error	z value	Pr(> z)
	(Intercept)	0,278	0,167	1,666	0,096
Sexe	Homme	Ref			
	Femme	0,332	0,117	2,823	0,005
Age	27 ans et plus	Ref			
	26 ans et moins	-0,328	0,162	-2,023	0,043
Situation d'emploi	En emploi	Ref			
	Autre	-0,270	0,139	-1,941	0,052
Statut d'occupation du logement	Autre	Ref			
	Propriétaire	-0,597	0,143	-4,186	0,000
Déménagement	Non	Ref			
	Oui	-2,149	0,143	-14,998	0,000
Région	Ile de France	Ref			
	Nord-Est	1,125	0,188	5,983	0,000
	Nord-Ouest	1,163	0,161	7,214	0,000
	Sud-Ouest	1,242	0,202	6,155	0,000
	Sud-Est	1,200	0,184	6,525	0,000

Modèles de correction de la non-réponse du sous-échantillon (05) – population générale

Résultats CNR - sous-échantillon (05)

Variable	Modalité	Estimate	Std. Error	z value	Pr(> z)
	(Intercept)	0,383	0,130	2,950	0,003
Sexe	Homme	Ref			
	Femme	0,143	0,058	2,464	0,014
Statut d'occupation du logement	Autre	Ref			
	Propriétaire	-0,225	0,071	-3,155	0,002
Type de logement	Maison	Ref			
	Autre	-0,291	0,077	-3,781	0,000
Situation d'emploi	En emploi	Ref			
	Autre	-0,190	0,062	-3,068	0,002
Niveau de diplôme	Bas	Ref			
	Haut	0,290	0,060	4,800	0,000
Déménagement	Non	Ref			
	Oui	-2,057	0,069	-29,753	0,000
Région	Ile de France	Ref			
	Nord-Est	1,006	0,123	8,205	0,000
	Nord-Ouest	1,063	0,124	8,600	0,000
	Sud-Ouest	0,922	0,133	6,939	0,000
	Sud-Est	0,922	0,121	7,649	0,000
TUU	UU de moins de 100 000 habitants	Ref			
	UU de 100 000 habitants et plus	-0,209	0,066	-3,174	0,002

Annexe 8 – Modèles de correction de la non-réponse – vague 2

Modèles de correction de la non-réponse mairie

Résultats CNR mairie

Variable	Modalité	Estimate	Std, Error	z value	Pr(> z)
	(Intercept)	3,139	0,086	36,586	0,000
Sexe	Homme	Ref			
	Femme	-0,153	0,029	-5,242	0,000
Statut d'occupation du logement	Autre	Ref			
	Propriétaire	0,309	0,040	7,783	0,000
Type de logement	Autre	Ref			
	Maison	0,309	0,042	7,371	0,000
Nombre de personnes dans le ménage	Moins de 3 personnes	Ref			
	3 personnes ou plus	-0,084	0,032	-2,597	0,009
Niveau de diplôme	Bas	Ref			
	Haut	0,275	0,032	8,551	0,000
HLM	Non	Ref			
	Oui	0,178	0,039	4,543	0,000
Région	Ile de France	Ref			
	Nord-Est	-0,258	0,046	-5,578	0,000
	Nord-Ouest	-0,312	0,050	-6,195	0,000
	Sud-Ouest	-0,633	0,042	-15,028	0,000
	Sud-Est	-0,087	0,059	-1,479	0,139
TUU	UU de moins de 100 000 habitants	Ref			
	UU de plus de 100 000 habitants	-0,241	0,038	-6,316	0,000
		Ref			
	taille_com02_grande_commune_deplacement	0,253	0,071	3,562	0,000
Type d'échos	né avant 1968	Ref			
	un seul écho	-0,893	0,187	-4,773	0,000
	plusieurs échos, une seule origine	-0,854	0,203	-4,215	0,000
	plusieurs échos et plusieurs origines	-1,187	0,057	-20,838	0,000
Nombres d'origines	né avant 1968	Ref			
	0 origine	1,415	0,245	5,766	0,000
	1 origine	0,751	0,185	4,060	0,000
	2 origines	0,683	0,047	14,605	0,000
	3 origines et plus	0,298	0,048	6,225	0,000

Modèles de correction de la non-réponse du sous-échantillon (03) – descendants d’immigrés

Les résultats des modélisations sont présentés dans les tableaux ci-après.

Résultats CNR sous-échantillon (03) modèle 1

Variable	Modalité	Estimate	Std. Error	z value	Pr(> z)
	(Intercept)	0,481	0,109	4,401	0,000
Sexe	Homme	Ref			
	Femme	0,116	0,049	2,375	0,018
Age	47 ans et plus	Ref			
	De 37 à 46 ans	-0,171	0,064	-2,670	0,008
	De 28 à 36 ans	-0,248	0,070	-3,545	0,000
	27 ans et moins	-0,235	0,077	-3,039	0,002
Statut d'occupation du logement	Autre	Ref			
	Propriétaire	0,478	0,062	7,765	0,000
Situation d'emploi	En emploi	Ref			
	Autre	-0,182	0,058	-3,126	0,002
Niveau de diplôme	Bas	Ref			
	Haut	0,175	0,053	3,296	0,001
HLM	Non	Ref			
	Oui	-0,427	0,070	-6,068	0,000
Déménagement	Non	Ref			
	Oui	-1,548	0,061	-25,530	0,000
Région	Ile de France	Ref			
	Nord-Est	0,477	0,088	5,433	0,000
	Nord-Ouest	0,741	0,094	7,919	0,000
	Sud-Ouest	0,167	0,078	2,153	0,031
	Sud-Est	0,339	0,097	3,499	0,000
TUU	UU de moins de 100 000 habitants	Ref			
	UU de 100 000 habitants et plus	-0,275	0,055	-5,025	0,000
QPV	Non	Ref			
	Oui	-0,193	0,076	-2,558	0,011

Résultats CNR sous-échantillon (03) modèle 2

Variable	Modalité	Estimate	Std. Error	z value	Pr(> z)
	(Intercept)	0,310	0,140	2,208	0,027
Sexe	Homme	Ref			
	Femme	0,316	0,076	4,174	0,000
Age	33 ans et plus	Ref			
	32 ans et moins	-0,138	0,083	-1,676	0,094
Statut d'occupation du logement	Autre	Ref			
	Propriétaire	0,339	0,114	2,976	0,003
Type de logement	Autre	Ref			
	Maison	0,401	0,116	3,467	0,001
Situation d'emploi	En emploi	Ref			
	Autre	-0,261	0,082	-3,192	0,001
HLM	Non	Ref			
	Oui	-0,292	0,098	-2,981	0,003
Déménagement	Non	Ref			
	Oui	-1,260	0,088	-14,299	0,000
TUU	UU de moins de 100 000 habitants	Ref			
	UU de 100 000 habitants et plus	-0,204	0,108	-1,888	0,059
G2 dans le même logement que G1	Non	Ref			
	Oui	-0,287	0,162	-1,772	0,076
QPV	Non	Ref			

	Oui	-0,256	0,101	-2,540	0,011
--	-----	--------	-------	--------	-------

Résultats CNR sous-échantillon (03) modèle 3

Variable	Modalité	Estimate	Std. Error	z value	Pr(> z)
	(Intercept)	0,832	0,133	6,245	0,000
Statut d'occupation du logement	Autre	Ref			
	Propriétaire	0,402	0,096	4,193	0,000
Type de logement	Autre	Ref			
	Maison	0,322	0,093	3,443	0,001
Situation d'emploi	En emploi	Ref			
	Autre	-0,272	0,085	-3,200	0,001
Niveau de diplôme	Bas	Ref			
	Haut	0,286	0,077	3,712	0,000
HLM	Non	Ref			
	Oui	-0,567	0,133	-4,259	0,000
Déménagement	Non	Ref			
	Oui	-1,348	0,087	-15,480	0,000
TUU	UU de moins de 100 000 habitants	Ref			
	UU de 100 000 habitants et plus	-0,376	0,078	-4,808	0,000
G2 dans le même logement que G1	Non	Ref			
	Oui	-0,712	0,289	-2,467	0,014

Résultats CNR sous-échantillon (03) modèle 4

Variable	Modalité	Estimate	Std. Error	z value	Pr(> z)
	(Intercept)	0,875	0,106	8,274	0,000
Age	33 ans et plus	Ref			
	32 ans et moins	-0,326	0,063	-5,194	0,000
Statut d'occupation du logement	Autre	Ref			
	Propriétaire	0,532	0,081	6,580	0,000
Type de logement	Autre	Ref			
	Maison	0,305	0,079	3,846	0,000
Niveau de diplôme	Bas	Ref			
	Haut	0,225	0,068	3,297	0,001
HLM	Non	Ref			
	Oui	-0,626	0,095	-6,609	0,000
Déménagement	Non	Ref			
	Oui	-1,366	0,072	-18,891	0,000
TUU	UU de moins de 100 000 habitants	Ref			
	UU de 100 000 habitants et plus	-0,329	0,072	-4,582	0,000
QPV	Non	Ref			
	Oui	-0,220	0,103	-2,145	0,032

Résultats CNR sous-échantillon (03) modèle 5

Variable	Modalité	Estimate	Std. Error	z value	Pr(> z)
	(Intercept)	0,730	0,162	4,516	0,000
Statut d'occupation du logement	Autre	Ref			
	Propriétaire	0,549	0,120	4,586	0,000
Age	37 ans et plus	Ref			
	De 28 ans à 36 ans	-0,402	0,119	-3,386	0,001
	27 ans et moins	-0,428	0,128	-3,333	0,001
Situation d'emploi	En emploi	Ref			
	Autre	-0,192	0,110	-1,754	0,080
Niveau de diplôme	Bas	Ref			
	Haut	0,244	0,105	2,335	0,020
HLM	Non	Ref			
	Oui	-0,265	0,122	-2,164	0,030
Déménagement	Non	Ref			
	Oui	-1,588	0,114	-13,890	0,000
TUU	UU de moins de 100 000 habitants	Ref			
	UU de 100 000 habitants et plus	-0,550	0,114	-4,814	0,000

Modèle de correction de la non-réponse du sous-échantillon (04) – descendants de Domiens

Résultats CNR - sous-échantillon (04)

Variable	Modalité	Estimate	Std. Error	z value	Pr(> z)
	(Intercept)	0,151	0,125	1,208	0,227
Age	28 ans et plus	Ref			
	27 ans et moins	-0,428	0,144	-2,979	0,003
Nombre de personnes dans le ménage	Moins de 3 personnes	Ref			
	3 personnes ou plus	0,312	0,132	2,362	0,018
Statut d'occupation du logement	Autre	Ref			
	Propriétaire	0,533	0,136	3,931	0,000
Déménagement	Non	Ref			
	Oui	-1,349	0,145	-9,293	0,000
QPV	Non	Ref			
	Oui	-0,498	0,200	-2,493	0,013

Annexe 9 – Partage des poids

La détermination des liens entre les répondants et la base de sondage

Nous reprenons les notations utilisées par Lavallée (2009), U^A , désigne la population décrite par la base de sondage, U^B est la population d'intérêt, $i \in r^B$ est un individu répondant lors de la collecte, $j \in U^A$ est une unité de la base de sondage⁸⁷, $L_{j,i}$ indique la présence d'un lien entre j et i :

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

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

est le nombre de liens entre l'individu i de la population U^B et la base de sondage.

D'après ce qui a été dit en 4.4, le calcul du nombre de liens est immédiat :

$$L_i^B = \begin{cases} 2 & \text{si l'individu est immigré, domien, descendants d'immigré ou descendant de domien} \\ 1 & \text{sinon} \end{cases}$$

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

$$\omega_i^{B,CNR} = \sum_{j \in r^A} \frac{L_{j,i}}{L_i^B} \omega_j^{A,CNR}$$

où r^A désigne l'ensemble des répondants de la population U^A , $\omega_j^{A,CNR}$ le poids corrigé de la non-réponse de l'unité répondante j et $\omega_i^{B,CNR}$ le poids de l'individu i déduit des liens et des poids $\omega_j^{A,CNR}$ par partage des poids.

Cependant, l'application brute de cette formule entraîne une dispersion considérable des poids, En effet :

- une première source de dispersion est liée au fait qu'un individu aura un poids très différent s'il est atteint par le sous-échantillon (05) sans surreprésentation ou par un des sous-échantillons (01) à (04) avec des surreprésentations élevées ;
- une deuxième source de dispersion est liée à des surreprésentations différentes d'une origine à l'autre.

Par conséquent, il s'avère nécessaire de pondérer les liens, et de faire dépendre ces pondérations des origines des individus.

La formule de partage des poids avec les liens pondérés

Afin de pallier le problème qui vient d'être mentionné, il est nécessaire de pondérer les liens des individus tout en tenant compte de leur origine, Faire cela impose d'utiliser une formule un peu plus générale que la formule classique de partage des poids présentée ci-dessus, Cette formule généralisée fait intervenir la notion de **liens pondérés**, présentée par Lavallée (2009), qui utilise et généralise la notion de liens vue précédemment.

Plus précisément, un système de liens pondérés $\widetilde{\theta}_{j,i}$ doit vérifier les conditions suivantes :

- $\forall j, \forall i, \widetilde{\theta}_{j,i} \in [0,1]$
- $\widetilde{\theta}_{j,i} = 0$ s'il n'y a pas de lien entre l'individu i de la population U^B et l'unité j de la population U^A

⁸⁷ Formellement, il s'agit d'un couple (individu, numéro de sous-échantillon).

- iii. $\widetilde{\theta}_{j,i} > 0$ s'il y a au moins un lien entre l'individu i de la population U^B et l'unité j de la population U^A
- iv. $\forall i, \sum_{j \in U^A} \widetilde{\theta}_{j,i} = 1$, autrement dit le lien pondéré total d'un individu i avec la base de sondage est égal à 1

La formule de partage des poids devient alors :

$$\omega_i^{B,CNR} = \sum_{j \in r^A} \widetilde{\theta}_{j,i} \omega_j^{A,CNR} = \sum_{j \in r^A} \widetilde{\theta}_{j,i} \frac{1}{\pi_j \widehat{p}_j} \quad (1)$$

où π_j est la probabilité d'inclusion de l'unité j et \widehat{p}_j est l'estimation de sa probabilité de réponse obtenue par correction de la non-réponse.

Un point important est que, en raison de la disjonction deux à deux des sous-échantillons (01) à (05) mentionnée en partie 2,1, un individu i donné n'admet qu'un seul lien avec l'échantillon r^A des répondants de U^A (même s'il peut admettre deux liens avec la population U^A), de sorte que dans la somme (1) un seul terme est non nul :

$$\omega_i^{B,CNR} = \widetilde{\theta}_{j_i,i} \frac{1}{\pi_{j_i} \widehat{p}_{j_i}} \quad (2)$$

où, par définition, j_i est l'unique unité répondante de r^A correspondant à l'individu répondant i de r^B .

Finalement, pour construire les poids $\omega_i^{B,CNR}$ il reste à définir les liens pondérés $\widetilde{\theta}_{j,i}$.

Le calcul des liens pondérés

Il existe évidemment une infinité de systèmes de liens pondérés possibles. La section 4,6 de Lavallée (2009) présente une méthode de calcul d'un système de liens pondérés minimisant la variance des estimateurs. La solution obtenue est complexe et nous avons procédé un peu différemment, en suivant la méthodologie présentée par Davezies, Landre, Murat et Rousseau (2005), qui a pour avantage de fortement limiter la dispersion des poids, et donc la variance des estimateurs.

Considérons un individu i répondant et j_i l'unique unité répondante de r^A correspondant à i .

On note k_i la strate d'origines de i , où il est entendu que nous faisons les abus de langage suivant :

- les DOM sont considérés comme une strate d'origines ;
- pour un descendant d'immigrés ou de Domiens, on désigne par strate d'origine celle du parent dont il est considéré descendant.

On note enfin $r_1 \dots r_5$ les sous-échantillons de répondants des échantillons (01), ..., (05), n_m^k le nombre de répondants dans r_m dans la strate d'origines k , et m_i le sous-échantillon auquel appartient j_i .

L'expression de $\widetilde{\theta}_{j,i}$ est alors donnée par le tableau ci-après, Elle dépend à la fois du sous-échantillon auquel appartient l'unité j_i tirée, mais aussi de sa strate d'origines.

Liens pondérés associés aux individus en fonction de la strate d'origines et du sous-échantillon

	Immigré	Domien	Descendant d'immigré	Descendant de Domien	Autre
$\widetilde{\theta}_{j,i}$	$\frac{n_{m_i}^{k_i}}{n_1^{k_i} + n_5^{k_i}}$	$\frac{n_{m_i}^{k_i}}{n_2^{k_i} + n_5^{k_i}}$	$\frac{n_{m_i}^{k_i}}{n_3^{k_i} + n_5^{k_i}}$	$\frac{n_{m_i}^{k_i}}{n_4^{k_i} + n_5^{k_i}}$	1

Annexe 10 – Vérifications sur les pondérations

Dans une optique de comparabilité, des vérifications ont été réalisées à partir d'une source externe : l'Enquête emploi en continu⁸⁸ (EEC). Les effectifs et la structure après pondération sont assez proches de ceux mesurés dans l'enquête emploi (cf. tableaux ci-après). Le concept de « descendant d'immigrés » diffère légèrement dans TeO2 et dans l'EEC. Dans TeO2, les enfants dont les parents nés Français dans une ancienne colonie (notamment l'Algérie) et qui auraient perdu leur nationalité française à l'indépendance, sont reclassés comme descendants d'immigrés. De même, TeO2 étant une enquête centrée sur l'origine des personnes, les réponses des enquêtés sur leurs origines et celles de leurs parents peuvent être différentes par rapport à une enquête généraliste (EEC). Dans cette section, les statistiques sur les descendants d'immigrés de TeO2 utilisent le concept TeO2.

Comparaison des effectifs des personnes dont les parents sont nés à l'étranger dans TeO 2 et dans l'EEC

	Effectifs			Proportion parmi les descendants G2		
	TeO2	EEC	Ecart Teo/EEC	TeO2	EEC	Ecart Teo/EEC
Maghreb	1 608 799	1 267 752	27%	39%	36%	3%
Algérie	836 730	629 463	33%	20%	18%	2%
Maroc, Tunisie	772 068	638 288	21%	19%	18%	1%
Autres pays d'Afrique	373 153	304 454	23%	9%	9%	0%
Afrique sahélienne	134 444	96 794	39%	3%	3%	1%
Afrique guinéenne ou centrale	122 155	96 071	27%	3%	3%	0%
Turquie, Moyen-Orient	167 764	166 027	1%	4%	5%	-1%
Turquie	131 372	125 173	5%	3%	4%	0%
Autres pays d'Asie	204 289	158 185	29%	5%	4%	0%
Asie du Sud-Est	139 702	100 997	38%	3%	3%	1%
Europe du Sud	1 335 971	1 186 648	13%	32%	34%	-1%
Portugal	608 762	498 865	22%	15%	14%	1%
Espagne, Italie	727 209	687 783	6%	18%	20%	-2%
Autres pays de l'UE 27	309 733	267 661	16%	7%	8%	0%
Autres pays d'Europe	91 719	94 720	-3%	2%	3%	0%
Amérique, Océanie	43 214	71 021	-39%	1%	2%	-1%
Refus, NSP	143	3 434	-96%	0%	0%	0%
Total descendants G2	4 134 785	3 519 902	17%	100%	100%	0%
Descendants de Domiens	392 868	295 349	33%			
Descendants de Français nés à l'étranger	1 437 104	1 599 522	-10%			

Les résultats obtenus sont globalement proches entre les deux sources, en effectifs et plus encore en proportions. Pour les descendants d'immigrés G2, l'écart en effectifs sur le total (+17 %) provient essentiellement des personnes originaires du Maghreb (+27 %), d'Afrique subsaharienne (+23 %) et d'Asie hors Moyen-Orient (+29 %), c'est-à-dire les anciennes colonies pour lesquelles la nationalité de naissance des parents n'est pas toujours bien connue. Les écarts sont parfois beaucoup plus importants pour des niveaux très fins, car il s'agit de pays n'ayant pas servi de strate de tirage. De ce fait, les exploitations ne peuvent se focaliser sur des sous-populations d'effectifs trop faibles. Enfin, les proportions de chaque origine parmi les descendants d'immigrés sont très similaires entre TeO 2 et l'EEC.

Pour les descendants G2 de Domiens, l'écart en effectifs sur le total est plus important (+ 33 % dans TeO 2 par rapport à l'EEC). Comme pour les descendants d'immigrés G2, il s'agit d'un phénomène déjà rencontré dans TeO1. Les écarts étaient même plus importants. Ces différences pourraient s'expliquer, outre par l'inégalité importance accordée à la question des origines dans les deux enquêtes, par des stratégies d'échantillonnage et de redressements très différentes.

⁸⁸ Années 2019-2020

Des comparaisons sur des caractéristiques annexes (emploi, niveau d'études) ont également été menées. Même si le questionnement sur les diplômes est différent dans deux enquêtes, les résultats, à champ identique, sont très proches. Pour la situation professionnelle, la part de personnes en emploi est relativement proche (75%) mais on observe en revanche de forts écarts dans la répartition entre chômeurs et inactifs. Les personnes sans emploi se déclarent plus fréquemment chômeuses dans TeO2 que dans l'EEC. Dans TeO2, ce statut est auto-déclaré alors que dans l'EEC, le chômage est mesuré plus finement au sens du BIT.

Comparaison de la structure par situation d'emploi, diplôme dans TeO2 et l'EEC en pourcentage

	TeO2	EEC	Ecart Teo/EEC
Situation d'emploi			
En emploi	75	74,2	0,7
Chômage	10	6,5	3,9
Inactivité	15	19,3	-4,5
Diplôme			
Diplôme supérieur à Bac + 2 ans	24,9	24,9	0,0
Bac + 2 ans	15,5	15,3	0,2
Bac ou brevet professionnel	20,0	20,5	-0,5
CAP- BEP	23,2	23,3	-0,1
Brevet des collèges	5,5	4,7	0,8
Aucun diplôme	11,0	11,3	-0,3