RÉPUBLIQUE
FRANÇAISE
*Liberté
Égalité
Fraternité*

Céreq
ÉTUDES
53
2024

# Enquête 2016 auprès de la Génération 2013

## Méthodologie et bilan

Christophe BARRET, Mady CISSÉ,
Christophe DZIKOWSKI, Émilie GAUBERT,
Zora MAZARI, Manon OLARIA, Ines OUJIA,
Alexie ROBERT, Florence RYK, Mélanie VIGNALE
Avec la collaboration d'Anthony SATTLER, ancien stagiaire

*Coord.* : Manon OLARIA

Équipe ingénierie et gestion d'enquête
Céreq > Département Entrées et évolutions dans la vie active

Génération 2013
Enquête 2016

# Enquête 2016 auprès de la Génération 2013

## Méthodologie et bilan

*Coord. :* Manon OLARIA

*Auteurs :* Christophe BARRET, Mady CISSÉ, Christophe DZIKOWSKI, Émilie GAUBERT, Zora MAZARI, Manon OLARIA, Ines OUJIA, Alexie ROBERT, Florence RYK, Mélanie VIGNALE

Avec la collaboration d'Anthony SATTLER, ancien stagiaire

Équipe ingénierie et gestion d'enquête
Céreq > Département Entrées et évolutions dans la vie active

# Synthèse

Ce document de travail présente un panorama complet des travaux liés à l'enquête 2016 auprès de la Génération 2013. Première et unique interrogation de cette génération, trois ans après la sortie du système éducatif. Après une brève présentation du dispositif d'enquêtes Génération et plus spécifiquement de l'enquête 2016, ce document détaille toutes les étapes de sa réalisation dans un ordre chronologique. La première étape consiste en la construction de la base de sondage d'élèves présumés sortants du système éducatif en 2012/2013. Il s'agit d'une base de données d'élèves collectée auprès des établissements de formation initiale situés en France et dans les départements et régions d'outre-mer. Après avoir défini un plan de sondage stratifié et équilibré, le tirage aléatoire de l'échantillon est effectué à l'aide de l'algorithme du Cube[1]. Vient ensuite une phase dite de préparation qui englobe le développement informatisé du questionnaire en CATI (*computer assisted telephone interview*), l'enrichissement des coordonnées téléphoniques et postales et la mise en place du protocole de contact et de diffusion d'informations autour de l'enquête (lettre et mail avis, site internet, page Facebook, etc.). La phase de collecte s'est donc déroulée d'avril à juillet 2016, soit quatre mois d'enquête. Enfin, l'ensemble des traitements post-collecte sont détaillés tels que la création des bases d'exploitation (apurement, codification, redressement des salaires, etc.) et le traitement de la non-réponse (redressement et calage, etc.).

Dans le cadre de la réflexion autour de la rénovation du dispositif d'enquêtes Génération, une nouvelle expérimentation a été menée en parallèle de l'enquête principale téléphonique à partir d'un échantillon disjoint. Pour la première fois, cette expérimentation multimode (collecte par internet et téléphone) est réalisée sur une enquête à trois ans.

[1] La macro SAS CUBE est un algorithme d'échantillonnage qui permet de tirer de manière aléatoire un échantillon équilibré sur un ensemble de totaux connus à partir d'informations auxiliaires disponibles dans la base de sondage. La méthode consiste à choisir un échantillon tel que les estimateurs d'Horvitz-Thompson des totaux des variables servant à l'équilibrage coïncidant avec les vrais totaux.

# Sommaire

# 1. Présentation de l'enquête 2016 auprès de la Génération 2013

## 1.1. Le dispositif des enquêtes Génération

Les enquêtes Génération s'inscrivent dans le cadre du dispositif d'enquêtes de l'Observatoire national des entrées dans la vie active (Oneva). Elles s'intéressent à l'insertion et au cheminement des sortants du système éducatif lors de leurs premières années de vie active. Elles ont deux objectifs principaux : d'une part, elles permettent de produire des indicateurs d'insertion (taux d'emploi, taux de chômage, taux d'emploi à durée indéterminée, etc.), selon les niveaux de formation, les filières, les spécialités, à destination des acteurs publics et sociaux ; d'autre part, elles recueillent des informations qui contribuent à la compréhension des processus d'insertion et des différenciations des parcours en début de carrière.

**Encadré 1 • Le CÉREQ (Centre d'études et de recherches sur les qualifications)**

Créé en 1971, devenu établissement public en 1985, le Centre d'études et de recherches sur les qualifications (Céreq) est placé sous la tutelle des ministères en charge de l'éducation nationale et du travail, de l'emploi, de la formation professionnelle et du dialogue social.

Il a pour missions de développer des études et des recherches, de collecter et d'exploiter des données originales dans le domaine de la relation formation-emploi, et de formuler des avis et propositions destinés à éclairer les choix en matière de politiques de formation.

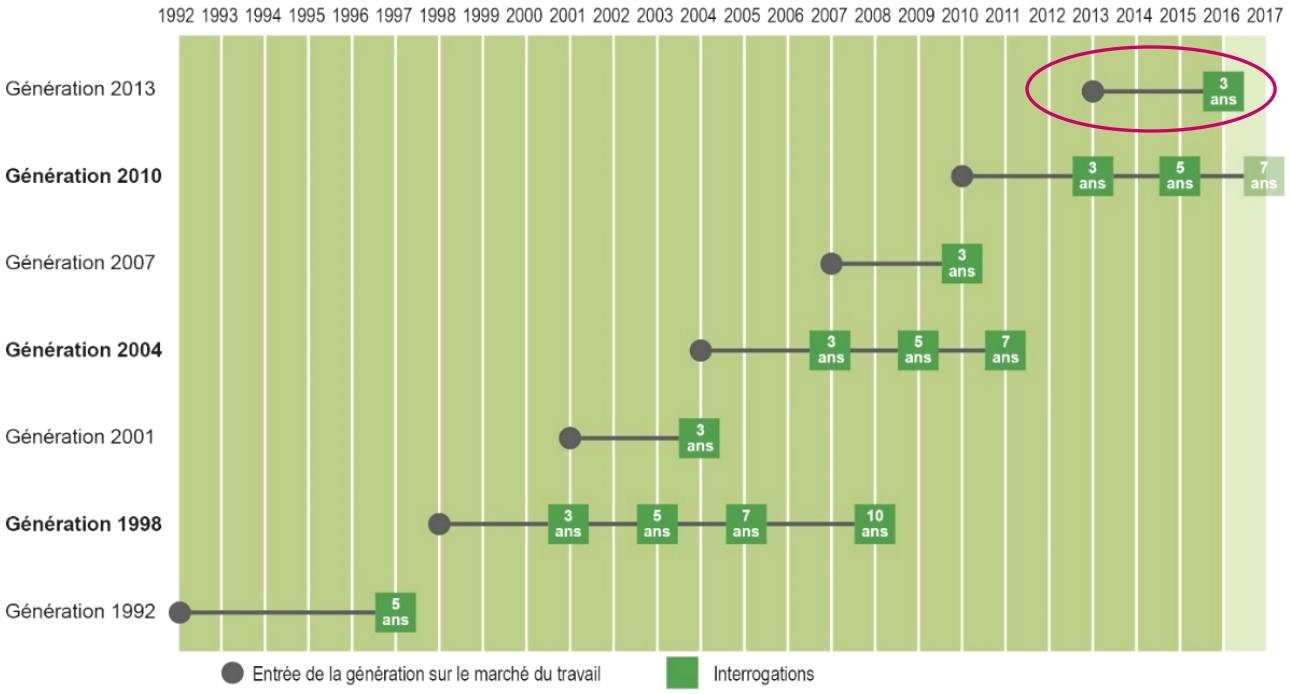
La première enquête Génération a été effectuée en 1997 auprès de jeunes sortis du système éducatif en 1992 et issus de tous les niveaux de formation. Un dispositif régulier d'interrogation a été mis en place à partir de l'enquête de 2001 effectuée auprès de sortants du système éducatif en 1998 : les enquêtes s'enchaînent au rythme d'une génération nouvelle de sortants interrogée tous les trois ans et avec une alternance entre une enquête « génération pleine » et une « génération légère ».

L'enquête dite « génération pleine » comprend plusieurs interrogations. La première interrogation, trois ans après la sortie du système éducatif, est principalement tournée vers la production d'indicateurs et des usages institutionnels.

Les interrogations suivantes, à cinq et sept ans, sont plutôt centrées sur les usages analytiques (notamment sur la question des parcours, des mobilités sur moyen terme). Ces réinterrogations permettent aussi d'approfondir certains constats issus de l'exploitation de la première interrogation (par exemple sur les insertions les plus problématiques).

L'enquête « génération légère » se limite à la première interrogation à trois ans, avec une taille d'échantillon plus réduite, et un questionnement allégé. Cette enquête a vocation à réactualiser sur une nouvelle cohorte les indicateurs d'insertion sur les trois premières années d'insertion selon une grille d'analyse moins fine que pour les « générations pleines ».

**Figure 1 • Calendrier du dispositif des enquêtes Génération**




Le cœur de l'enquête, pour la première interrogation comme pour les autres quand il y a lieu, est constitué par le calendrier d'activité qui permet de suivre mois par mois la situation des jeunes à l'issue de leur formation initiale, et de décrire les situations successives d'emploi et de non-emploi.

Les enquêtes Génération permettent aussi de répondre à des demandes d'extensions nationales ou régionales portant sur les jeunes issus de certains niveaux ou spécialités de formations ou sur les jeunes ayant bénéficié de certaines mesures pour la formation. Les échantillons peuvent donc être élargis en conséquence.

**Tableau 1 • Détail des enquêtes Génération et effectifs de répondants**

| Génération | Année de la première interrogation | Nombre de répondants | Dont liés aux extensions d'échantillon |
|---|---|---|---|
| 1992 | 1997 | 26 000 | - |
| 1998 | 2001 | 55 000 | 19 000 |
| 2001 | 2004 | 25 000 | 15 000 |
| 2004 | 2007 | 65 000 | 37 000 |
| 2007 | 2010 | 29 000 | 26 000 |
| 2010 | 2013 | 38 500 | 13 500 |
| **2013** | 2016 | 22 700 | 13 800 |

## 1.2. Le champ de l'enquête

La Génération 2013 concerne les primo-sortants de formation initiale en 2012-2013 (année scolaire). Les sortants de formation qui avaient déjà interrompu leurs études au moins un an avant l'année scolaire considérée sont hors-champ. Tous les niveaux et domaines de formations sont concernés. De façon plus précise, les critères d'éligibilité pour être dans le champ retenu, nommé ensuite « champ Céreq », sont les suivants :

- avoir été inscrit dans un établissement de formation en France (métropolitaine + DOM) durant l'année scolaire 2012-2013 ;
- avoir quitté le système éducatif entre octobre 2012 et octobre 2013[2] ;
- ne pas avoir interrompu ses études durant une année ou plus avant l'année scolaire 2012-2013 (sauf pour raison de santé) ;
- ne pas avoir repris ses études pendant l'année qui a suivi l'entrée sur le marché du travail ;
- avoir 35 ans ou moins en 2013 ;
- être localisé en France (métropolitaine + DOM) au moment de l'enquête (ce qui exclut notamment les personnes résidant à l'étranger à la date d'enquête).

Toutes ces conditions sont cumulatives.

Quelques points particuliers concernent l'application de ces critères :

- **Cas du contrat d'apprentissage.** Bien qu'il s'agisse de contrats de travail au même titre que les contrats de professionnalisation (ou de qualification), les contrats d'apprentissage sont considérés comme relevant de la formation initiale. Une personne repérée comme sortant de formation en 2012-2013 qui poursuit par un contrat d'apprentissage en 2013-2014 est considérée en poursuite d'études. À la différence du contrat de professionnalisation, cette dernière est classée hors-champ Céreq.
- **Cours par correspondance.** Une personne sortie d'un établissement de formation en 2012-2013 qui poursuit des cours par correspondance ou des cours du soir en 2013-2014 est considérée en poursuite d'études. Elle est classée hors-champ Céreq dans le cas où elle n'occupe pas un emploi en parallèle. Dans le cas contraire, elle est considérée dans le champ Céreq.
- **Statut d'élève fonctionnaire.** Contrairement aux cohortes précédentes, une personne sortie d'un établissement de formation en 2012-2013, mais poursuivant ses études en 2013-2014 comme élève fonctionnaire n'est pas considérée en poursuite d'études ; elle est classée dans le champ Céreq.

Les différences de champ Céreq entre les six cohortes enquêtées (G98 à G13) sont mineures :

- les sortants de classes de 6e et 5e, inclus dans le champ de la Génération 1998 et de la Génération 2001 sont exclus du champ depuis la Génération 2004 ;
- les étudiants étrangers sortants de l'Université (repérables dans le système d'information sur le suivi de l'étudiant, SISE) étaient hors-champ pour la Génération 1998 et la Génération 2001 ; ils sont inclus dans le champ depuis la Génération 2004 ;

[2] Pour certaines formations (école fonction publique, sport/animation, santé/social, formations en alternance et thèse) la date de fin de formation peut s'étendre jusque décembre 2013.

- la période de référence pour la date de sortie était l'année civile pour la Génération 1998 et la Génération 2001 ; le choix s'est porté sur l'année scolaire depuis la Génération 2004 (octobre n à octobre n+1) ;

- la réforme de la formation des métiers de l'enseignement (mastérisation des concours d'enseignement) mise en place sous régime transitoire l'année 2010, et le rattachement administratif des IUFM aux universités n'ont pas permis d'intégrer dans l'enquête Génération 2010 les sortants de première année de cette formation.

Pour l'enquête 2016, le champ a légèrement évolué par rapport aux précédentes enquêtes. Auparavant, les élèves fonctionnaires étaient considérés entrant sur le marché du travail dès l'obtention d'une rémunération au cours de leur formation. Autrement dit, ils étaient considérés sortants du système de formation initiale avant leur entrée dans leur école de fonctionnaire. Désormais, ils sont considérés sortants du système éducatif au moment de leur sortie de leur formation de fonctionnaire.

Ce changement est motivé par deux raisons principales :

- **Pour le questionnement.** Le jeune primo-sortant considère (le plus souvent) son école de fonctionnaire comme la poursuite naturelle de ses études plutôt qu'une première situation d'emploi (même si celle-ci est rémunérée). Dès lors, le questionnement s'en trouve plus adapté ; la description se porte sur la situation d'emploi réelle plutôt que sur celle à cheval avec les études.

- **Pour l'analyse.** La définition du niveau de formation est simplifiée et permet de classer les sortants à leur « vrai » niveau (et non au niveau déclaré au moment de leur entrée dans l'école).

L'évolution du champ impacte également la création des bases comparables avec les anciennes générations de sortants. À cet effet, des questions supplémentaires ont été posées dans le questionnaire filtre afin d'identifier précisément les jeunes sortants de formation d'enseignement.

**Encadré 2 • Une enquête de la statistique publique**

L'enquête 2016 auprès de la Génération 2013, extensions comprises, est une enquête de la statistique publique et relève à ce titre de la loi n° 51-711 du 7 juin 1951 sur l'obligation, la coordination et le secret en matière de statistiques. Celle-ci définit le cadre de procédures destinées à garantir l'intérêt, la pertinence et la qualité des enquêtes publiques, ainsi que la confidentialité des informations collectées.

L'opportunité de l'enquête 2016 auprès de la Génération 2013 a ainsi été discutée au Comité national de l'information statistique (CNIS), qui a émis un avis favorable en mars 2015 puis lui a attribué un label d'intérêt général et de qualité statistique en février 2016 (label n° 2016X713AU).

Les réponses apportées à l'enquête sont confidentielles. Les fichiers des réponses détaillées à l'enquête qui sont mis à disposition des chercheurs sont anonymes et ne comportent pas d'informations susceptibles de permettre une identification directe ou indirecte des personnes enquêtées. La réalisation de l'enquête et sa diffusion a fait l'objet d'une déclaration auprès de la Commission nationale informatique et libertés (CNIL).

## 1.3. Les extensions

Les acteurs intervenant dans le domaine de la formation sont multiples, avec leurs questionnements propres, mais aussi avec un questionnement commun sur le devenir des bénéficiaires. À la demande de certains d'entre eux, ministères, conseils régionaux ou observatoires de branches, des extensions sont régulièrement adossées aux enquêtes Génération. Elles permettent de situer les analyses ciblées sur des publics, des filières ou des territoires particuliers dans un contexte plus large, par rapprochement avec des indicateurs de référence nationaux.

Plusieurs types d'extensions existent, parfois combinées :

- **Extension d'échantillon,** de façon à disposer d'un nombre de répondants suffisant pour permettre des analyses statistiquement pertinentes sur les catégories de sortants ciblées.

- **Extension de champ,** pour interroger également sur les catégories de sortants ciblées les personnes non retenues dans le champ Céreq (par exemple en levant la restriction de première sortie de formation initiale).

- **Extension de questionnement,** pour poser quelques questions complémentaires sur des catégories de sortants ciblées.

Pour l'enquête 2016, neuf partenaires ont signé une convention. À ces neuf partenariats, le Céreq a décidé pour ses besoins propres d'ajouter une extension d'échantillon sur les sortants de baccalauréat professionnel. L'objectif est de contribuer à évaluer l'effet de la réforme de la voie professionnelle avec la généralisation du baccalauréat professionnel en 3 ans en comparant la Génération 2010 (avant réforme) et la Génération 2013 (après réforme). L'extension d'échantillon par rapport à l'échantillon national du Céreq permet notamment de faire des analyses plus fines par spécialité détaillée.

Le tableau suivant présente pour chaque partenariat mené dans le cadre de l'enquête 2016 auprès de la Génération 2013, le nombre de questionnaires conventionnés et exploitables ainsi que le type d'extension.

**Tableau 2a • Liste des partenaires d'extensions de l'enquête Génération 2013**

| Institution ou organisme | Population cible | Champ | Nombre de questionnaires conventionnés | Nombre de questionnaires exploitables | Type d'extension |
|---|---|---|---|---|---|
| CGDD | Sortants de formation initiale en environnement | Céreq | 2 900 | 3 470 | Échantillon + Questionnement |
| DREES | Sortants des formations sanitaires et sociales | Céreq + post-initiaux | 5 200<br>Répartis en :<br>– Santé 3 600<br>– Social 1 600 | 5 590<br>*(Dont 2 135 post-initiaux)*<br>Répartis en :<br>– Santé 3 840<br>– Social 1 750 | Échantillon + Questionnement |
| DGESIP | Sortants de doctorats<br><br>Sortants des autres formations du supérieur | Céreq | 1500<br><br>3 000 | 1 906<br>*Dont 1 659 diplômés*<br><br>9 525 | Échantillon + Questionnement<br><br>Échantillon |
| INJEP | Ensemble des sortants de la Génération | Céreq | 20 000 | 18 679 | Questionnement |
| DGAFP | Ensemble des sortants de la Génération ayant passé au moins un concours de la fonction publique | Céreq | 20 000 | 19 498 dont :<br>*– 13 146 ont décrit la partie situation d'emploi actuelle*<br><br>*– 2 235 ont déclaré avoir créé leur propre activité ou avoir entamé des démarches pour le faire. Parmi eux,*<br>*1 441 ont déclaré l'avoir fait parce qu'ils en avaient envie et sont donc entrés dans le module.*<br><br>*– 19 498 ont répondu au module rapport au travail* | Questionnement |
| CGET | Sortants résidant en QPV au moment de la sortie de formation initiale | Céreq | 1 250 | 1 645 en QPV<br>*Part 1 : 19 498*<br>*Part 2 : 22 737 (dont 3 239 post-initiaux)*<br>*Part 3 : 22 737 (dont 3 239 post-initiaux)* | Échantillon + Questionnement |
| AGEFIPH | Sortants en situation de handicap | Céreq | 20 000 | 19 498<br>*Dont 2 922 entrants dans le module* | Questionnement |
| DJEPVA | Sortants de formations sport | Céreq + post-initiaux | 1 500 | 1 690<br>*Dont 1 104 post-initiaux* | Échantillon + Questionnement |
| ERASMUS + | Ensemble des sortants de la Génération | Céreq | 20 000 | 19 498<br>*Dont 11 033 entrants dans le module* | Questionnement |
| CÉREQ | Sortants de baccalauréat professionnel | Céreq | 1500 | 2 054 | Échantillon |

## 1.4. Le questionnaire

### 1.4.1. Phase de concertation

- **En interne :** une première phase a consisté en la réunion de groupes de travail en interne au Céreq sur les différentes parties du questionnaire afin de recueillir les attentes et besoins des chargés d'études en lien avec leurs thématiques d'études. Il s'agit d'identifier les évolutions des politiques d'emploi et de formation pour tenter de répondre aux questionnements d'institutionnels. Pour exemple, identifier les sortants du nouveau baccalauréat professionnel en 3 ans pour évaluer l'impact de la réforme, quantifier les bénéficiaires du contrat d'accompagnement « garantie jeunes », du service civique ou du CIVIS (Contrat d'insertion dans la vie sociale), et ce afin de mesurer l'impact sur leur insertion professionnelle, mais aussi observer l'évolution de l'accès des femmes aux postes de cadres et aux fonctions d'encadrement en France (pour des perspectives de comparaisons européennes), etc.

- **Avec les partenaires d'extensions :** de nombreux échanges avec les partenaires d'extensions ont également lieu. En effet, la rédaction des questions conventionnées est le fruit d'une collaboration entre le Céreq et le partenaire d'extensions. Le but étant de répondre au mieux à leurs besoins d'exploitation.

- **Avec les instances de concertation :** avant le passage au comité du label, plusieurs réunions de travail sont programmées avec les membres du comité de concertation.

Le comité de concertation du dispositif d'enquêtes Génération est composé de représentants des principales directions des deux ministères de tutelle du Céreq, des autres ministères, de représentants des partenaires sociaux et de représentants d'observatoires, d'organismes d'études et de recherche.

À la demande du comité du label, le comité de concertation a été élargi en 2015 au Commissariat général à l'Égalité des territoires (CGET), à l'Institut national de la jeunesse et de l'éducation populaire (INJEP) et à une association d'étudiants. En revanche, les propositions faites à des associations de familles n'ont pas abouti à ce jour.

Les organismes membres du comité de concertation sont les suivants :

**Tableau 2b • Membres du comité de concertation du dispositif d'enquêtes Génération**

| | |
|---|---|
| **Le Céreq** | – Département Entrées et évolutions dans la vie active<br>– Trois centres associés régionaux (CAR) |
| **Les deux ministères de tutelle** | – DGESCO – Direction générale de l'enseignement scolaire<br>– DEPP – Direction de l'évaluation, de la prospective et de la performance<br>– DGESIP – Direction générale de l'enseignement supérieur et de l'insertion professionnelle<br>– DARES – Direction de l'animation de la recherche, des études et des statistiques<br>– DGEFP – Direction générale de l'emploi et de la formation professionnelle |
| **Observatoires, organismes d'études et de recherche** | – INSEE<br>– RESOSUP – Réseau des observatoires de l'enseignement supérieur<br>– OREFQ – Observatoire régional de l'emploi, de la formation et des qualifications de Lorraine<br>– Universitaires |
| **Autres ministères ou opérateurs** | – DGAFP – Direction générale de l'administration et de la fonction publique<br>– CGET – Commissariat général à l'égalité des territoires<br>– CNEFOP – Conseil national de l'emploi, de la formation et de l'orientation professionnelle<br>– Pôle emploi<br>– INJEP – Institut national de la jeunesse et de l'éducation populaire<br>– ONISEP – Office national d'information sur les enseignements et les professions |
| **Partenaires sociaux ou associations** | – Medef<br>– CGT<br>– CFDT<br>– FO<br>– UNSA<br>– FAGE |

Ce comité se réunit habituellement deux fois par an pour examiner la préparation des enquêtes (questionnaire en particulier), le bilan des enquêtes administrées, et les premiers résultats d'exploitation des données. Une première réunion du comité s'est déroulée en juillet 2015 et la seconde en janvier 2016 à Paris.

À partir d'une version provisoire du questionnaire, des avis et propositions sont formulés pour l'améliorer. Ces réunions permettent également de recueillir l'expression des attentes des institutionnels, des partenaires sociaux et des utilisateurs experts des enquêtes Génération et de bénéficier des avis et expertises des principaux acteurs mobilisés sur la question de la relation formation-emploi.

### 1.4.2. Description du questionnaire

Cette enquête est la première interrogation trois ans après la sortie du système éducatif des sortants de formation initiale de l'année scolaire 2012-2013.

Le questionnaire débute par une partie filtre qui permet l'identification de l'individu et la vérification des critères d'éligibilité :

- **Validation de la cible.** L'interlocuteur est-il le « bon individu » ? Une vérification de son identité ainsi que le mois et année de naissance, le nom de l'établissement de formation fréquenté en 2012-2013.

- **Validation du champ.** L'interlocuteur fait-il partie du champ (champ Céreq, champ d'une extension ou hors-champ) ?

Le questionnaire aborde ensuite successivement les thèmes suivants : le parcours scolaire, l'identification du plus haut diplôme, le calendrier mensuel d'activité sur les trois années suivant la sortie du système éducatif avec une description détaillée de l'ensemble des situations professionnelles, l'opinion sur l'emploi actuel, les perspectives professionnelles et les caractéristiques individuelles et l'environnement familial.

À ce questionnaire de base s'ajoutent des modules de questionnement développés par les chargés d'études du Céreq pour répondre à des besoins de recherche ou des modules proposés par les partenaires d'extensions :

- Module « Thèse ».

- Module « Post-initiaux des formations du sport et de l'animation ».

- Module « Post-initiaux des formations du domaine de la santé et du social ».

- Module « Concours et attractivité de la fonction publique » (nouveau module par rapport au questionnaire de la Génération 2010).

- Module « Séjours à l'étranger ».

- Module « Développement durable ».

- Module « Entrepreunariat – création d'entreprise » (nouveau module).

- Module « Dispositifs d'accompagnement » (nouveau module).

- Module « Rapport des jeunes au travail » (nouveau module).

- Module « Handicap et problème de santé durable »

La plupart de ces modules sont posés à l'ensemble de la Génération dans le champ Céreq, au sens « primo-sortant », à l'exception des modules des formations du sport et de l'animation et celui des formations du domaine de la santé et du social. Demandés par les partenaires des formations respectives, ces deux derniers modules sont proposés en alternative au module parcours scolaire aux seuls individus en formation post-initiale (individus ayant déjà arrêté leurs études plus d'un an avant leur sortie des études en 2013).

De plus, cinq extensions d'échantillon adossées à cette enquête permettent des focus sur une partie du champ de l'enquête. Deux d'entre elles permettent en outre un élargissement du champ Céreq à des sortants post-initiaux. Les cinq extensions portent sur les sortants :

- des formations « Sport-Animation » (y compris élargissement aux post-initiaux) ;
- des formations « Santé-Social » » (y compris élargissement aux post-initiaux) ;
- des formations « Thèse » ;
- des formations environnementales ;
- issus des quartiers prioritaires de la politique de la ville.

### 1.4.3. Zoom sur le calendrier d'activité

Le questionnaire de l'enquête Génération s'articule autour d'un calendrier d'activité. Il retrace mois par mois le parcours professionnel de l'individu entre la date de fin d'études et la date de l'enquête. Ce module constitue le cœur de l'information recherchée et commande l'ouverture des modules de description détaillée des situations professionnelles déclarées.

Ce calendrier permet de distinguer six situations :

- **situation d'emploi en entreprise :** toute activité rémunérée y compris : emplois non salariés, congés payés, congés maladie courts, congés maternité/paternité, stage postdoctoral ;
- **situation d'emploi en intérim :** toute activité rémunérée par le biais d'une agence d'intérim ;
- **situation de recherche d'emploi :** personne inscrite ou non inscrite, rémunérée ou non rémunérée, par pôle emploi ;
- **situation de reprise d'études :** à temps plein dans un établissement scolaire ;
- **situation de formation[3], autre qu'une reprise d'études :** toute formation (hors contrat de travail) ;
- **autres situations :** congé parental, homme/femme au foyer (sans démarche de recherche d'emploi), congés de maladie longue durée et invalidante (sans activité professionnelle), le bénévolat, etc.

Le calendrier d'activité se présente sous forme de tableau croisant les mois et les situations possibles suivi d'un récapitulatif de saisie modifiable. Il est grisé jusqu'au dernier mois de fin d'études (information collectée en début de questionnaire). Il recueille des informations complémentaires (notamment le nom et la commune de l'entreprise pour les situations d'emploi).

Par construction, une seule situation peut être déclarée pour un mois donné et la durée minimum pour la description d'une situation est d'un mois. La déclaration successive d'une même situation est impossible (pour exemple, deux situations de recherche d'emploi ou d'emploi en intérim ne peuvent être contigües[4]). Exception faite pour la situation d'emploi en entreprise, suite à un changement d'établissement employeur (nom d'entreprise et commune différents). Ce dernier cas ne tient pas compte du changement de profession ou du contrat de travail au sein d'un même établissement.

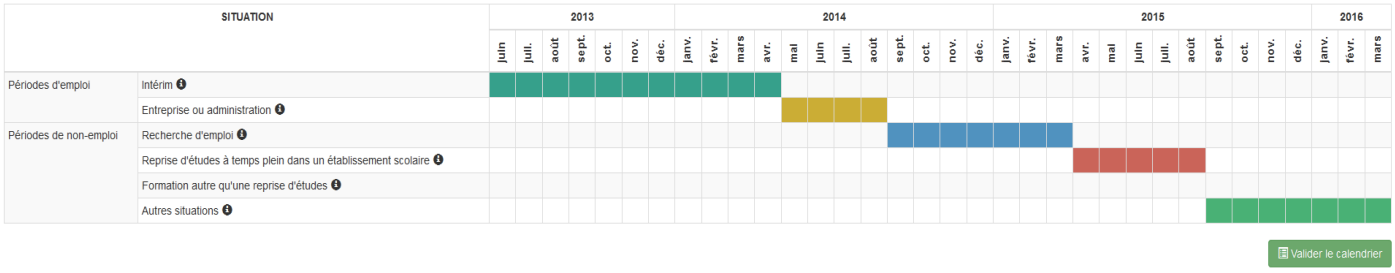
[3] La déclaration d'une formation dans le calendrier d'activité en lien avec un contrat d'apprentissage (plus d'un an après la sortie), un contrat de professionnalisation ou un statut d'élève fonctionnaire est considérée comme une situation d'emploi « en entreprise ». L'ouverture du module de description de la situation est raccordée à l'emploi correspondant.

[4] Cas d'une situation d'emploi en intérim : plusieurs missions dans plusieurs entreprises différentes et en fonction de critères de sélection qui dépendent de la position de la situation dans le calendrier ou de la durée de cette dernière, plusieurs cas possibles : description de la première mission si la situation d'intérim correspond au premier emploi de l'individu après la fin des études ; description de la dernière mission dans le cas où celle-ci correspond à la situation actuelle ; dans tous les autres cas, description de la mission la plus longue.

Au moment de la collecte, le remplissage de l'ensemble des situations d'emploi et/ou de non-emploi est réalisé de façon successive sur toute la période d'observation. Le calendrier d'activité, entièrement renseigné, fournit une liste de vingt-deux codes, appelés « CAL » (voir tableau 3 ci-après), dans un ordre chronologique. Pour chacune des situations déclarées, un code est affecté en fonction de sa nature, sa durée, du moment où elle apparait dans la description du parcours. Cette liste pilote la suite du questionnement avec ouverture ou non de modules associés à chacune des situations décrites.

Dans le cadre de l'enquête 2016 auprès de la Génération 2013, seules la situation de premier emploi et celle à la date de l'enquête (emploi ou non-emploi) sont décrites *a posteriori* dans un module.

**Figure 2 • Le calendrier mensuel d'activité**




Récapitulatif de la saisie

| | Début | Fin | Type de période | Nom | Commune |
|---|---|---|---|---|---|
| | juin 2013 | avril 2014 | Emploi - Intérim | v | MARSEILLE 4E ARRONDISSEMENT |
| | mai 2014 | août 2014 | Emploi - Entreprise | v | MARSEILLE 4E ARRONDISSEMENT |
| | septembre 2014 | mars 2015 | Recherche d'emploi | | |
| | avril 2015 | août 2015 | Reprise d'études | | |
| | septembre 2015 | mars 2016 | Autres situations | | |

Note de lecture : Exemple d'une interrogation en mars 2016 décrivant un parcours professionnel comprenant cinq situations distinctes depuis la fin des études en juin 2013.

Le tableau ci-après présente l'ensemble des vingt-deux « CAL » possibles et les règles à appliquer. Pour chaque code ; le nom du module à ouvrir (le cas échéant), les données à collecter *via* les « pop-up » de description et les variables à créer et/ou imputer.

**Tableau 3 • Définition des « CAL » issus du calendrier d'activité activant le pilotage des modules de description**

<table>
<tr><th>CAL</th><th>Descriptif du CAL</th><th>Durée</th><th>Module à ouvrir</th><th>Description du « pop-up » avec données à recueillir et à stocker pour informer et/ou filtrer dans le module</th></tr>
<tr><td>01</td><td>Emploi court du passé</td><td>≤ 12 mois</td><td rowspan="2">EP<br>Uniquement sur la situation de premier emploi passé (différente de la situation d'emploi actuelle)</td><td rowspan="4">Dans tous les cas :<br>- Dates de début et de fin de séquence<br>- Nom de l'établissement employeur<br>- Commune de l'établissement employeur (code postal, code INSEE, libellé)<br>- Historique des établissements déjà saisis<br><br>En plus, dans le cas d'une situation d'emploi en intérim :<br>- Travaille dans une ou plusieurs entreprises<br><br>Les variables suivantes doivent être calculées pour moduler la suite du questionnement :<br>- Variable INTER (pour les intérimaires)<br>- Variable RETOU (si le couple commune + nom de l'établissement employeur déjà présent dans l'historique)</td></tr>
<tr><td>02</td><td>Emploi long du passé</td><td>> 12 mois</td></tr>
<tr><td>03</td><td>Emploi court actuel</td><td>≤ 6 mois</td><td rowspan="2">EA<br><br>Cas 1 : premier emploi ≠ emploi actuel<br><br>Cas 2 : premier emploi = emploi actuel</td></tr>
<tr><td>04</td><td>Emploi long actuel</td><td>> 6 mois</td></tr>
<tr><td>05</td><td>Recherche d'emploi courte du passé</td><td rowspan="3">≤ 3 mois</td><td rowspan="6">Pas de module</td><td rowspan="10">Dans tous les cas :<br>- Dates de début et de fin de situation<br><br>En plus, dans le cas d'une situation de formation :<br>- Type de formation<br><br>En plus, dans le cas d'une autre situation :<br>- Définition de la situation</td></tr>
<tr><td>07</td><td>Inactivité courte du passé</td></tr>
<tr><td>09</td><td>Formation courte du passé</td></tr>
<tr><td>06</td><td>Recherche d'emploi longue du passé</td><td rowspan="3">> 3 mois</td></tr>
<tr><td>08</td><td>Inactivité longue du passé</td></tr>
<tr><td>10</td><td>Formation longue du passé</td></tr>
<tr><td>11</td><td>Recherche d'emploi courte actuelle</td><td>≤ 3 mois</td><td rowspan="2">NEA</td></tr>
<tr><td>12</td><td>Recherche d'emploi longue actuelle</td><td>> 3 mois</td></tr>
<tr><td>13</td><td>Inactivité courte actuelle</td><td>≤ 3 mois</td><td rowspan="2">Pas de module</td></tr>
<tr><td>14</td><td>Inactivité longue actuelle</td><td>> 3 mois</td></tr>
<tr><td>15</td><td>Formation courte actuelle</td><td>≤ 3 mois</td><td rowspan="2">NEA</td></tr>
<tr><td>16</td><td>Formation longue actuelle</td><td>> 3 mois</td></tr>
</table>

**Suite tableau 3 :**

| CAL | Descriptif du CAL | Durée | Module à ouvrir | Description du « pop-up » avec données à recueillir et à stocker pour informer et/ou filtrer dans le module |
|---|---|---|---|---|
| **17** | Reprise d'études du passé | *Pas de notion de durée* | *Pas de module* | Dans tous les cas :<br>- *Dates de début et de fin de situation* |
| **18** | Reprise d'études actuelle | | ETUA | |
| **20** | Job de vacances | *Situation d'emploi entre juin et sept 2013 ET ≤ 4 mois* | *Pas de module* | |
| **21** | Vacances | *1re situation en autre situation ET ≤ 3 mois* | | |
| **22** | Reprise d'études à temps plein dans un établissement scolaire ou universitaire dans l'année qui suit la fin de formation | ***Hors-champ*** | | |

### 1.4.4. Les extensions de questionnement

Les extensions de questionnement concernent des modules proposés dans le cadre de conventionnement avec des partenaires institutionnels. Certains partenariats sont historiques et les modules produits (parfois rafraîchis) ont été intégrés dans les questionnaires de plusieurs cohortes de sortants.

#### *Module « Post-Initiaux des formations du sport et de l'animation »*

Ce module concerne uniquement les diplômés en post-initial inclus dans l'extension d'échantillon sur les formations du sport et de l'animation. Pour cette population, le module se substitue aux questions sur le parcours scolaire.

Les questions abordent la situation d'activité avant la dernière formation terminée en 2013 (type d'employeur, profession, etc.), les objectifs et origine du financement de cette formation, l'obtention d'un baccalauréat ou d'un diplôme de niveau supérieur et/ou la détention d'autres diplômes ou brevets dans le domaine du sport et de l'animation.

#### *Module « Post-initiaux des formations du domaine de la santé et du social »*

Ce module concerne uniquement les diplômés en post-initial inclus dans l'extension d'échantillon sur les formations de la santé et du social. Pour cette population, le module se substitue aux questions sur le parcours scolaire.

Les questions abordent le temps passé dans différentes situations professionnelles (emploi, recherche d'emploi, etc.) entre la première interruption d'études et le démarrage de la formation en lien avec le diplôme obtenu en 2013, la situation d'activité (type d'employeur, profession, etc.), les objectifs et origine du financement de cette formation, l'obtention d'un baccalauréat ou d'un diplôme de niveau supérieur.

### Module « Développement durable »

Ce module concerne les diplômés de formation initiale dont la spécialité est liée à l'environnement et au développement durable. Ce court module pose la question du degré de correspondance avec une formation environnementale, des opportunités d'insertion professionnelle, des opinions sur la dégradation de l'environnement et sur l'avenir des métiers verts.

Ce module, posé dans les précédentes éditions des enquêtes Génération, a évolué quelque peu suite aux nouvelles orientations d'exploitation dans le cadre de cette enquête.

### Module « Thèse »

Ce module concerne les diplômés de doctorat (excepté ceux du domaine de la santé). Les questions du module abordent le projet professionnel au moment de la soutenance de la thèse, le cadre institutionnel de la thèse (nature de la rémunération, nature du laboratoire), les démarches de valorisation des travaux (publications, participation à des colloques, séminaires, etc.), la situation postdoctorale (stages postdoctorat, qualification au CNU).

### Module « Handicap et problème de santé durable »

Ce module concerne l'ensemble des sortants de la Génération (individus issus d'extension compris).

Plusieurs questions abordent successivement la présence d'une maladie ou un problème de santé (chronique ou de caractère durable), l'existence d'un handicap et sa nature, l'identification du moment d'apparition et plus précisément les démarches liées à la reconnaissance administrative.

Ce module, posé dans les précédentes éditions des enquêtes Génération, a également évolué notamment dans la construction du questionnement.

### Module « Séjours à l'étranger en cours d'études »

Ce module concerne l'ensemble des sortants de la génération (individus issus d'extension compris).

Ce module aborde le cadre de réalisation du/des stage(s) en cours d'études (stage conventionné, période d'études, période de travail, séjour linguistique, etc.). Si le stage est en lien avec les études ; le diplôme préparé, la durée, le pays d'accueil, mais aussi les moyens financiers (bourse, aide financière, etc.) et les compétences acquises pour améliorer les chances d'insertion professionnelle.

**Pour cette édition, quatre nouveaux modules et une question ont été intégrés :**

### Module « Concours et attractivité de la fonction publique »

Ce module concerne l'ensemble des sortants de la génération (individus issus d'extension compris).

Ce module aborde en introduction l'intérêt ou non porté à la fonction publique, le nombre de tentatives de passage de concours, l'admissibilité et/ou l'admission. Plus précisément, une description des cinq derniers concours succincts ; intitulé du concours, catégorie hiérarchique, versant de la fonction publique, date de passage des premières épreuves, admissibilité. De manière aléatoire, une description plus détaillée d'un seul concours ; les étapes de sélection du concours, les épreuves (écrites et/ou orales), l'admissibilité et/ou l'admission, les raisons de non-admissibilité/non-admission, le recrutement et les motivations pour l'inscription au concours.

Dans le cas où l'individu ne s'est inscrit à aucun concours pour devenir fonctionnaire, une question lui est posée afin d'en connaître les motifs.

### *Module « Entreprenariat – création d'entreprise »*

Ce module concerne l'ensemble des sortants de la génération (individu issu d'extension compris).

Ces questions, intercalées dans le module « Perspectives professionnelles », abordent la notion de création d'activité. À savoir, les démarches entreprises ou envisagées à court et/ou long terme depuis la fin des études et les motivations.

### *Module « Dispositifs d'accompagnement »*

Ce module concerne uniquement les individus ayant connu une situation de non-emploi (recherche d'emploi ou formation) depuis la fin des études.

Ce module fait référence au bénéfice de mesures d'accompagnement dans le cadre de l'insertion professionnelle. Trois questions sont posées autour de dispositifs spécifiques ; le contrat garantie jeunes, le contrat d'insertion dans la vie sociale (CIVIS) et la participation à une mission de service civique.

### *Module « Rapport des jeunes au travail »*

Pour aborder le rapport au travail, deux axes de réflexion ont été définis (deux modules) : le premier s'oriente vers les relations professionnelles, notamment avec l'employeur, le supérieur hiérarchique ; le second porte sur les conditions d'emploi (rémunération, possibilités d'évolutions professionnelles, etc.) et l'environnement de travail (sécurité de l'emploi, reconnaissance, etc.).

Dans l'articulation du questionnaire, le premier module s'adresse aux individus dans une situation d'emploi salariée à la date de l'enquête et le second module est proposé à l'ensemble des sortants de la génération (individus issus d'extension compris).

### *Adresse de résidence l'année du baccalauréat*

Pour des besoins d'études spécifiques, l'adresse de résidence l'année du baccalauréat est collectée – notamment pour identifier la proportion de jeunes issus des quartiers prioritaires de la politique de la ville (QPV) (seule la commune figure dans les précédentes interrogations).

## 1.4.5. Ajout de questions complémentaires au questionnaire (hors extensions)

### *Dans le questionnaire filtre*

Suite au changement de champ sur la question des élèves fonctionnaires (cf. partie1.2.), des questions ont été ajoutées.

Par ailleurs, une question a été ajoutée afin de connaître la spécialité des individus en thèse hors santé.

Nous avons également rajouté des questions afin de détecter les masters MEEF (Métiers de l'enseignement, de l'éducation et de la formation).

La question concernant la situation de l'individu (arrêt ou poursuite d'études) lors de l'année 2013-2014 a été modifiée de manière à s'adapter au cas où l'individu n'avait pas obtenu son diplôme.

Par ailleurs, des questions ont été ajoutées pour caractériser les individus hors-champ. Deux cas d'hors-champ sont explorés plus en détail : les individus qui déclarent être en poursuite d'études l'année scolaire 2013-2014 et les individus qui ont interrompu leur scolarité plus d'un an pour des raisons autres que des problèmes de santé avant 2013. Pour les individus en poursuite d'études, il leur est demandé leur situation en 2016, la date d'arrêt des études. Pour ceux qui ont interrompu, nous avons ajouté des questions afin de connaître la durée d'interruption, de savoir s'il s'agissait d'une année de césure ou d'une année à l'étranger et la raison de cette interruption.

### *Dans la partie parcours scolaire et plus haut diplôme*

La typologie des diplômes a été mise à jour.

L'adresse du baccalauréat a été modifiée. En effet, cette question a fait l'objet d'un approfondissement par le biais d'une relance pour les individus n'ayant pas obtenu le baccalauréat, mais ayant suivi une année de terminale et qui sont concernés par cette question.

#### *Emploi passé et actuel*

Les questions sur l'activité de l'entreprise ont été mises à jour afin de passer d'une nomenclature NES à une nomenclature NAF Rév2 (choix déjà effectué pour la deuxième interrogation en 2015 de la Génération 2010).

Les contrats de travail ont également été mis à jour.

La dernière question du module sur l'opinion de l'emploi a également été modifiée suite à l'expérimentation de la deuxième interrogation de la Génération 2010.

#### *Non-emploi actuel ou reprise d'études actuelle*

Deux questions ont été ajoutées afin de connaître les actions réalisées dans le cadre de la recherche d'emploi.

Dans le cas d'une reprise d'étude, deux questions ont été ajoutées : l'une afin de savoir si l'individu est élève fonctionnaire et l'autre afin de connaître le diplôme préparé.

Des questions ont été reformulées de manière à être plus claires et d'autres ont été mises à jour.

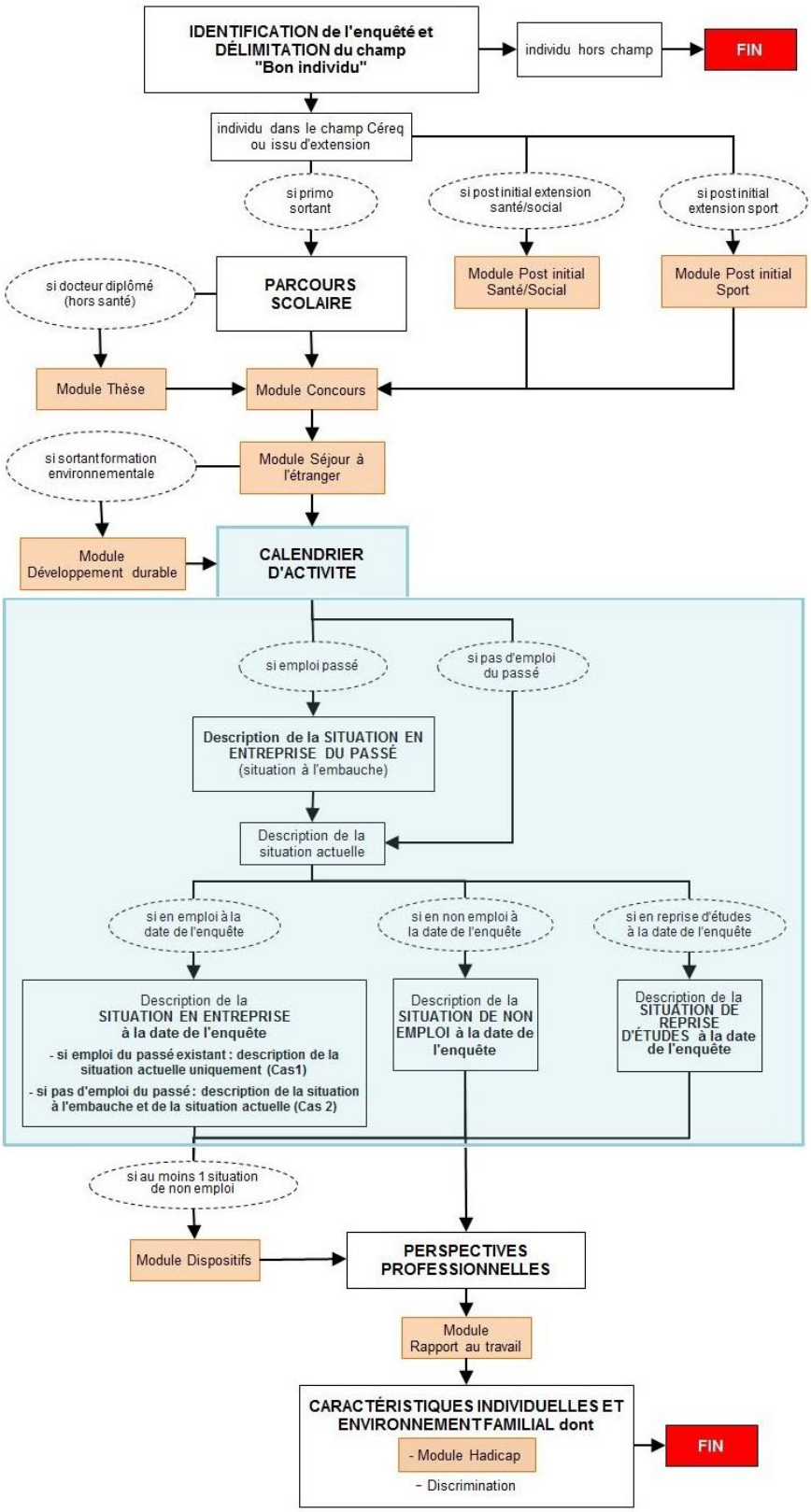
#### *Caractéristiques individuelles et environnement familial*

Étant dans une enquête dite « légère », le calendrier habitat a été substitué par une question pour connaître la situation résidentielle actuelle de l'individu.

Pour terminer, lors de la confirmation des coordonnées, nous avons laissé la possibilité à l'individu de changer uniquement son nom (nom d'usage).

### 1.4.6. Schéma détaillé de l'ossature du questionnaire

**Figure 3 ● Schéma du questionnaire de l'enquête 2016 auprès de la Génération 2013**



## 1.5. L'intérêt et les atouts du dispositif

### *Un cadre d'analyse homogène et cohérent*

Contrairement à d'autres enquêtes d'insertion qui visent des publics segmentés (apprentis, lycéens, sortants de grandes écoles ou d'université…), seul le dispositif d'enquêtes Génération propose un questionnement, une méthodologie et un cadre d'analyse homogène pour tous, quels que soient le parcours scolaire, les diplômes obtenus, les domaines et voies de formation. Il est donc possible de comparer et d'évaluer l'impact de ces différentes caractéristiques sur les variations observées au cours des premières années de vie active : qui accède rapidement à un emploi ? Qui reste durablement au chômage ? À quel type d'emploi accède-t-on ? À quel niveau de rémunération ? Telles sont les questions auxquelles le dispositif permet de répondre. Plus généralement, il met en évidence les phénomènes de concurrence ou de complémentarité entre niveaux, domaines et voies de formation.

### *Des informations riches et diversifiées*

Grâce à un questionnaire détaillé et un échantillon important, les enquêtes permettent, au-delà des caractéristiques du parcours scolaire et des diplômes obtenus, de prendre en compte d'autres critères. Le genre, l'origine sociale, l'origine nationale, le lieu de résidence, les mobilités géographiques, le statut familial, les réseaux sociaux, mais aussi la place et le rôle des dispositifs publics sont autant de dimensions que le dispositif permet d'intégrer pour analyser les différences observées au cours des premières années de vie active.

### *Un recul temporel nécessaire*

Certaines enquêtes d'insertion sont réalisées quelques mois seulement après la sortie du système scolaire. L'option retenue est alors de disposer d'indicateurs qui peuvent être mis rapidement à disposition des décideurs, des familles et des étudiants. Avec le dispositif d'enquêtes Génération, la première interrogation est réalisée trois ans après la sortie du système scolaire. Les résultats des premières enquêtes ont mis en évidence l'importance de ce recul temporel. En effet, il faut attendre plusieurs années pour que la stabilisation professionnelle soit établie pour le plus grand nombre. Enquêter tôt après la sortie de formation donne une photographie datée et imparfaite des situations par rapport à l'emploi, qui accentue fortement les différences, alors que les enquêtes Génération montrent que celles-ci tendent à se réduire avec le temps.

### *Un suivi longitudinal*

Le questionnaire permet aux jeunes débutants de décrire systématiquement, mois par mois, les différentes situations qu'ils ont connues depuis leur sortie du système éducatif. Ce mode d'interrogation permet de construire différents indicateurs comme le taux de chômage ou le taux d'emploi, et d'aborder la qualité de l'emploi (niveau de rémunération, type de contrat). Il permet aussi de construire des typologies de parcours à partir de la description des situations mois par mois. Ces typologies offrent une vision synthétique des premières années sur le marché du travail : trajectoire d'accès rapide à l'emploi, trajectoire d'accès différé à l'emploi, trajectoire de décrochage, etc. L'insertion est une réalité multidimensionnelle qui ne peut se réduire à un ou deux indicateurs.

### *La même conjoncture pour tous*

Les générations sont construites en fonction de la date de sortie de formation et non de l'année de naissance. Quel que soit leur niveau de formation, les jeunes arrivent donc dans un contexte de marché du travail plus ou moins favorable, mais identique pour tous. Il est donc plus facile *a priori* de comparer les trajectoires d'accès à l'emploi. Mais cette conjoncture a-t-elle les mêmes effets pour tous ? À qui profitent les embellies ? Qui souffre le plus des retournements ? Quels effets sur les taux de chômage, l'importance des CDD ou de l'intérim, et pour qui ? Telles sont les questions auxquelles le caractère récurrent des enquêtes Génération permet de répondre.

## 1.6. Le calendrier de l'enquête

### 1.6.1. Calendrier de collecte

La réalisation de l'enquête s'étend sur quatre années. Les bases de sortants de formation initiale sont collectées dès leur disponibilité (au cours de l'année 2014).

Une fois l'échantillon tiré dans la base de sondage constituée, pour chaque enquête, trois phases sont distinguées :

- Une première phase de restructuration, normalisation/validation postale est effectuée par un sous-traitant pour normaliser les adresses postales. Cette phase intègre le rachat des adresses pour les individus ayant déménagé.

- Une deuxième phase consiste à rechercher les coordonnées téléphoniques des individus. Si l'individu n'est pas retrouvé à l'adresse de la base de sondage, deux types de recherche seront effectués : l'un basé sur le nom et le prénom, étendu à la France métropolitaine (homonyme nom/prénom) ; l'autre basé sur le nom uniquement (potentiellement membres de la même famille et ainsi personnes contact). Ces recherches sont réalisées *via* les annuaires France Télécom (pour les fixes et mobiles). Ces deux types de recherche permettent de retrouver une partie des individus ayant connu une mobilité géographique entre le lieu de leur formation initiale et leur résidence à la date de l'enquête.

- Un autre type de recherche en complément du premier est réalisé à partir d'une autre base de données qui permet d'enrichir les coordonnées téléphoniques issues des fournisseurs d'accès à internet et des opérateurs de téléphonie mobile.

- Ces opérations sont prévues en janvier-février 2016.

- Dans une troisième phase, le questionnaire est administré. Cette phase est prévue d'avril à juin 2016. La phase de préqualification transparente pour tous est intégrée cette année dans le questionnaire général. Elle concerne les numéros de téléphone des homonymes. Ils sont dits « qualifiés » par un questionnement préalable. Ces qualifications servent à trouver le bon individu parmi les numéros de téléphone d'homonymes. Elles permettent également de contrôler si l'individu appartient au champ de l'enquête. L'opération de préqualification n'étant plus distinguée, l'opération est donc lancée en même temps que le plateau d'enquête en avril 2016.

**Tableau 4 • Chronologie des principales étapes de constitution de l'enquête 2016 auprès de la Génération 2013**

| Nº | Étape | Début | Fin |
|---|---|---|---|
| 1 | Collecte des bases d'élèves auprès des établissements* | Janvier 2015 | Mai 2015 |
| 2 | Conception du questionnaire et modules spécifiques | Avril 2015 | Novembre 2015 |
| 3 | Élaboration des conventions | Novembre 2015 | Février 2016 |
| 4 | Tirage de l'échantillon | Octobre 2015 | Janvier 2016 |
| 5 | Conception du CATI<br>– version provisoire pour premier test<br>– derniers ajustements de la version définitive | Novembre 2015<br>Décembre 2015 | Novembre 2015<br>Mars 2016 |
| 6 | Tests du questionnaire au format CATI (x3) | 2 décembre 2015<br>21 mars 2016<br>29 février 2016 | 4 décembre 2015<br>26 mars 2016<br>5 mars 2016 |
| 7 | Recherches de coordonnées téléphoniques et RNVP | 27 janvier 2016 | 25 février 2016 |
| 8 | Terrain d'enquête | 4 avril 2016 | 30 juillet 2016 |
| 9 | Constitution des bases, apurements, codifications, pondération, documentation | Septembre 2016 | Juin 2016 |
| 10 | Livraison des bases aux extensions | Juillet 2017 | Juillet 2017 |
| 11 | Bilan méthodologique | Septembre 2016 | Septembre 2017 |

** cette période concerne la collecte des établissements hors rectorats (collecte réalisée au premier semestre 2015)*

### 1.6.2. Calendrier de diffusion

Le fichier provisoire d'enquête, à vocation interne, est mis à disposition en janvier 2017. Le fichier définitif est livré aux partenaires d'extension à l'été 2017.

Les études et chiffres clés sont publiés dans les collections du Céreq (Bref, Notes Emploi Formation, RELIEF) à partir du premier trimestre 2017.

Pour la mise à disposition, les fichiers d'enquêtes sont rendus anonymes, de manière à ne permettre aucune identification directe ou indirecte.

Les destinataires de ces fichiers sont :

- Les partenaires des extensions à l'enquête (DGESIP, DREES, DGAFP, Régions, etc.).
- Les deux ministères de tutelle du Céreq (Éducation nationale, Travail – *via* la Depp et la Dares)
- Le centre Maurice-Halbwachs pour diffusion auprès des chercheurs (à l'été 2019)

# 2. La constitution de la base de sondage

Construire une base d'élèves nécessite une collecte de fichiers auprès des établissements de formations initiales dispensées en France entière (y compris les collectivités d'outre-mer). Il n'existe pas de base d'élèves nominative centralisée couvrant l'ensemble des formations. De ce fait, la base de sondage de Génération est constituée spécifiquement pour cette enquête à partir de différentes sources et en opérant un certain nombre de traitements.

La base ainsi constituée présente un défaut de couverture dû aux établissements qui ne fournissent pas de bases d'élèves lors de la constitution de la base de sondage.

Par ailleurs, les données récupérées sont souvent des listes de sortants d'un établissement : il s'agit de personnes inscrites en 2012-2013 et non réinscrites en 2013-2014 dans le même établissement. Elles peuvent néanmoins s'être réinscrites ensuite ailleurs et sont alors hors du champ de l'enquête. La base comporte aussi des doublons : les jeunes inscrits dans plusieurs établissements ou dans plusieurs filières qui ne seraient pas repérés.

Il n'est possible de traiter que partiellement ces deux aspects lors de la construction de la base de sondage. C'est à partir de la déclaration de l'individu au moment de l'enquête que sera déterminée son éligibilité.

## 2.1. La collecte auprès des établissements

Pour chaque individu scolarisé pendant l'année scolaire 2012-2013 (présumé sortant), les informations suivantes sont récupérées : noms et prénoms, adresse et numéros de téléphone, mail, date et lieu de naissance, sexe, diplôme et spécialité de formation, obtention ou non du diplôme préparé.

Des informations relatives à l'établissement de formation sont également collectées. Lorsque le numéro d'identifiant national d'étudiant (INE) est disponible (informations issues des rectorats, des universités et des CFA), il est récupéré pour les traitements d'apurements de la base (repérage des individus en poursuite d'études et des doublons). Dans ce cadre et pour les CFA, les fichiers d'inscrits en 2013-2014 ont également été récupérés.

### 2.1.1. Collecte des bases rectorales

Les rectorats disposent des listes d'inscrits dans les établissements de leur ressort : collèges, lycées et BTS dépendant du ministère en charge de l'éducation nationale (y compris les établissements privés sous contrat), grâce au système informatique *Scolarité*.

Le Céreq réalise lui-même la collecte des informations souhaitées auprès des rectorats, initialisée par un mail de la direction de l'évaluation, de la prospective et de la performance (Depp), et selon un dessin d'extraction également défini avec la Depp.

La collecte a eu lieu au second trimestre 2014. Elle a permis de récupérer pour chaque académie l'ensemble des inscrits au cours de l'année scolaire 2012-2013, ainsi que celui des inscrits en 2013-2014.

La couverture des établissements est ainsi exhaustive sur le champ rectoral.

### 2.1.2. Collecte des bases de sortants de formation du sport et de l'animation

Dans le cadre de la Génération 2013, cette collecte a été réalisée en collaboration avec le ministère de la Jeunesse et des Sports pour lequel le Céreq réalise une extension d'échantillon et de questionnement.

Certaines directions régionales de la jeunesse, des sports et de la cohésion sociale (DRJSCS) ont rencontré des difficultés pour constituer la base d'élèves pour raison de nouvelles méthodes de gestion de leurs inscrits (accès logiciel et réorganisations).

Par conséquent, le faible taux de réponse a suscité la mise en place d'une collecte supplémentaire réalisée en octobre 2015. Elle a permis de récupérer les bases d'élèves de quatre grandes DRJSCS.

### 2.1.3. Collecte des bases de sortants d'universités

Contrairement à la Génération précédente, la collecte des universités a été réalisée en collaboration avec l'Agence de mutualisation des universités et établissements (Amue). En effet, le Céreq a été déclaré comme destinataire des données gérées par le logiciel *Apogée* dans la délibération RU13 du 10 décembre 2009, délibération CNIL. Pour chacune des universités, un patch d'extraction des sortants universitaires fonctionnant sous le logiciel *Apogée* leur a été fourni. De plus, un appariement avec la base *Examen* a été effectué afin de récupérer des informations sur l'obtention du diplôme.

### 2.1.4. Collecte des bases de sortants des écoles de la santé et du social

Les écoles de la santé et du social ne sont pas extraites de la Banque centrale des établissements (BCE). La collecte est effectuée auprès de la Direction de la recherche, des études, de l'évaluation et des statistiques (Drees) afin de récupérer les bases d'établissements issues du répertoire *FINESS*. Une sélection des formations est alors réalisée de façon précise dans les bases réceptionnées. Les nomenclatures associées sont livrées au même moment.

La collecte a été étendue aux sortants sans diplôme pour se mettre en conformité avec le champ des enquêtes Génération. Les individus non diplômés ne semblent pas être intégrés dans les fichiers.

### 2.1.5. Collecte des bases de sortants bénéficiaires du dispositif CIFRE

La collecte des fichiers CIFRE a été réalisée en interne. Cette collecte a été faite courant 2015.

### 2.1.6. Collecte des bases de sortants de CFA

Les données centralisées au niveau des académies *via* la base *SIFA* ont été collectées pour la première fois. Un arbitrage entre le gain de temps grâce à l'exhaustivité de la collecte face au nombre de variables disponibles a été nécessaire. Par exemple, avec la base centralisée, la collecte de l'adresse mail des individus n'a pas été possible.

En parallèle de cette collecte, les fichiers *SIFA* nationaux ont été récupérés comme sources de référence pour les deux années consécutives 2012-2013 et 2013-2014.

Ces fichiers ont permis notamment l'acquisition d'informations plus précises sur les diplômes et la suppression des individus réinscrits l'année suivante.

### 2.1.7. Collecte des bases de sortants d'écoles de la fonction publique

Les établissements de la fonction publique ne sont renseignés que pour moitié dans la BCE. Systématiquement, une liste des établissements est fournie par la Direction générale de l'administration et de la fonction publique (DGAFP). En vue de les intégrer dans le système de collecte, une phase de validation de cette liste est réalisée avec le Céreq. De plus, une lettre de soutien rédigée par la DGAFP a été envoyée à ces établissements.

### 2.1.8. Collecte des bases de sortants de lycées agricoles

Une base centralisée des sortants de lycées agricoles a été collectée au niveau ministériel. Cette base a permis d'éviter une collecte indépendante auprès de plus de 800 établissements. Néanmoins, les coordonnées téléphoniques ne sont pas disponibles pour tous les élèves. Une expertise précise des fichiers permettra d'évaluer le gain de la collecte de la base centralisée par rapport à une collecte classique auprès de chaque établissement de formation (bases contenant plus d'informations individuelles).

### 2.1.9. Collecte des bases de sortants d'« autres » établissements

Selon la BCE, en dehors des établissements dépendants des rectorats, des CFA, des lycées agricoles, environ 3 500 établissements dispenseraient de formations initiales.

**Encadré 3 ● La Base centrale des établissements (BCE)**

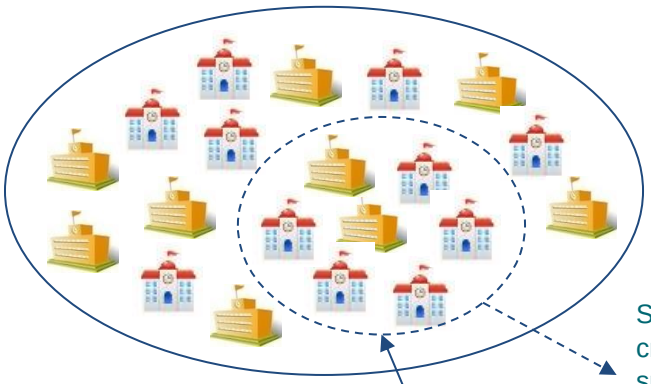
La BCE, gérée par le ministère en charge de l'éducation nationale, est le répertoire national des établissements assurant une activité de formation initiale générale, technique ou professionnelle, de la maternelle à l'enseignement supérieur, qu'ils soient publics ou privés, sous tutelle ou non du ministère de l'Éducation nationale et de la Recherche. Elle concerne également les structures d'administration du système éducatif public ainsi que certaines parties d'établissements qui ont besoin d'être identifiées pour la gestion du système éducatif. La BCE a pour rôle d'alimenter toutes les applications informatiques nationales de gestion administrative, financière, statistique ou documentaire du ministère de l'Éducation nationale et de la Recherche. Elle est mise à jour à partir des bases rectorales d'établissements, en temps réel.
Source : https://data.education.gouv.fr

Ces établissements divers entrent directement dans le champ de l'enquête Génération. Parmi eux, des universités, des écoles de commerce, des écoles d'ingénieurs, etc.

La liste de ces établissements avec les coordonnées de contact est extraite de la BCE à la demande de la Depp pour le Céreq.

La collecte des bases d'élèves s'est déroulée de février à mai 2015 selon le schéma organisationnel de collecte suivant.

**Figure 4 • Schéma organisationnel de la collecte auprès des « autres » établissements**



BCE
Pré-sélection des établissements
(y compris ceux issus des Dom)

Sélection des établissements selon divers critères (formation initiale, extension spécifique, ministères, etc.)
→ 3 500 établissements envoyés en production

Envoi d'un courrier à chaque directeur (pour accord de principe) et désignation du correspondant technique en charge de la constitution de la base d'élèves

Céreq
Formation - Travail - Emploi
www.cereq.fr

*Collecte des coordonnées du correspondant technique via différents modes de contact (avec l'aide d'un prestataire)*

Formulaire web à remplir par l'établissement

Contact téléphonique avec appel sur numéro Azur

Contact téléphonique suite à appel du prestataire

Interlocuteur fichier défini

Relances mails automatiques

Relances téléphoniques

Réception, contrôle et traitement de la base d'élèves par N vacataires au Céreq

***Le déroulement de la collecte auprès des « autres » établissements :***

- **Phase 1 :** envoi d'un courrier signé par le directeur du Céreq sollicitant la mise à disposition des données individuelles nécessaires à la constitution de la base de sondage et d'un courrier technique à l'attention de la personne qui réalisera l'extraction du fichier. Une plaquette d'information décrivant le dispositif d'enquêtes Génération (notamment la description du *champ*) est également fournie.

  Les objectifs de ces courriers sont : l'obtention de l'accord de principe du responsable de l'établissement, la désignation d'un correspondant technique et la transmission des instructions à la constitution de la base d'élèves (format et dessin d'enregistrement du fichier, liste des informations souhaitées). La signification du correspondant technique (nom et coordonnées) est rendue possible *via* un questionnaire web ou par téléphone (n° Azur).

- **Phase 2 :** échange(s) téléphonique(s) avec le correspondant technique de chaque établissement pour fixer une date de transmission de la base d'élèves.

- **Phase 3 :** relance(s) par téléphone et/ou par mail du correspondant technique dès lors que la base d'élèves n'est pas parvenue à la date convenue.

L'ensemble des opérations sus-présentées ont été réalisées par un prestataire.

Hormis les contacts initiaux par courrier, les contacts avec les établissements ont été principalement téléphoniques assistés par un système CATI, puis par mail. Afin de diminuer les taux de refus des établissements, plusieurs ministères de tutelle des établissements, préalablement sensibilisés à l'enquête, ont transmis un courrier de soutien.

***La gestion des bases d'élèves post-collecte :***

Les bases d'élèves sont ensuite réceptionnées directement par le Céreq qui vérifie systématiquement la lisibilité et la conformité des données (année de référence correcte, disponibilité de l'ensemble des informations attendues, etc.).

Le protocole de suivi de la collecte intègre :

- La saisie manuelle, *via* une base de données des informations sur les fichiers reçus (identification de l'établissement, détection des anomalies, etc.), de manière journalière. Cette base permet notamment d'automatiser l'envoi de mails aux correspondants techniques pour confirmer la bonne réception du fichier ou pour demander, *le cas échéant,* des informations manquantes attendues dans la base.

  Les fichiers « révisés » par les correspondants techniques sont directement réceptionnés au Céreq (sans solliciter de nouveau le prestataire).

- Un traitement automatisé des fichiers reçus pour vérifier les informations contenues dans les bases importées et la présence des variables obligatoires demandées.

Cette opération est réalisée par des vacataires présents au Céreq. Ils ont aussi une mission de codification des diplômes en parallèle de la gestion des fichiers reçus.

## 2.2. Compilation et apurement de la base de sondage

### 2.2.1. Les fichiers reçus

**Tableau 5 • Détail de la réception des fichiers par type d'établissement (hors bases collectées *via* les rectorats)**

| Type d'établissement de formation | | Génération 2010 | | | Génération 2013 | | |
|---|---|---|---|---|---|---|---|
| | | Nombre total d'établissements | Nombre de fichiers d'établissements reçus | % | Nombre total d'établissements | Nombre de fichiers d'établissements reçus | % |
| A | Écoles ingénieurs | 225 | 126 | 64 | 259 | 180 | 69 |
| AR | Rajout – écoles ingénieurs | | | | 30 | 13 | 43 |
| B | Écoles de commerce et de gestion | 312 | 125 | 44 | 311 | 139 | 45 |
| BR | Rajout – écoles de commerce et de gestion | | | | 5 | 3 | 60 |
| C | Universités et établissements rattachés | 78 | 78 | 100 | 72 | 71 | 99 |
| C2 | Universités de technologie | | | | 3 | 1 | 33 |
| C3 | Autres établissements publics d'enseignement universitaire | | | | 11 | 7 | 64 |
| C4 | Pôles recherche et enseignement supérieur | | | | 19 | 5 | 26 |
| C5 | Communautés d'établissement | | | | 8 | | 0 |
| CR | Rajout – autres établissements publics d'enseignement universitaire | | | | 238 | 46 | 19 |
| D | Écoles de formation sociales | 171 | 75 | 74 | 195 | 105 | 54 |
| E | Écoles de formation de la santé | 657 | | 75 | 638 | 449 | 70 |
| ED | Écoles de formation santé et social | | | | 24 | 13 | 54 |
| F | Drjscs | | | | 36 | 22 | 53 |
| GR | Rajout – collèges spécialisés | | | | 1 | 91 | 27 |
| GR | Rajout – écoles professionnelles spécialisées | | | | 5 | | |
| GR | Rajout – écoles secondaires spécialisées (2nd cycle) | | | | 1 | | |
| GR | Rajout – écoles composées uniquement de Sts et ou Cpge | | | | 301 | | |
| GR | Rajout – établissements régionaux d'enseignement adapté | | | | 2 | | |
| GR | Rajout – lycées polyvalents | | | | 28 | | |

**Suite tableau 5 :**

| Type d'établissement de formation | | Génération 2010 | | | Génération 2013 | | |
|---|---|---|---|---|---|---|---|
| | | Nombre total d'établissements | Nombre de fichiers d'établissements reçus | % | Nombre total d'établissements | Nombre de fichiers d'établissements reçus | % |
| JR | Rajout – écoles dans le secteur du service | | | | 19 | 6 | 32 |
| K | Écoles dans le secteur industriel | 27 | | 52 | 31 | 12 | 39 |
| KR | Rajout – écoles dans le secteur industriel | | | | 7 | 2 | 29 |
| L | Écoles de formation agricole | 15 | | 73 | 15 | 10 | 67 |
| LR | Rajout – écoles de formation agricole | | | | 1 | | 0 |
| M | Centres ou facultés privés | 12 | | 58 | 55 | 15 | 27 |
| NR | Rajout – lycées d'enseignement général et/ou technologique | | | | 86 | 120 | 31 |
| NR | Rajout – lycées professionnels | | | | 303 | | |
| R | Centres de formation pédagogique privés | 36 | | 69 | 35 | 20 | 57 |
| S | Écoles administration publique | 6 | | 100 | 6 | 4 | 67 |
| T | Écoles normales supérieures | 3 | | 67 | 4 | 2 | 50 |
| U | Écoles de la Dga | 3 | | 33 | 2 | 1 | 50 |
| V | Instituts études politiques | 10 | | 90 | 11 | 8 | 73 |
| W | Écoles architectures et artistiques | 350 | | 50 | 380 | 165 | 43 |
| WR | Rajout – écoles architectures et artistiques | | | | 1 | 1 | 100 |
| Z | Écoles de la Dgafp – supplément | 40 | | 75 | 64 | 24 | 38 |
| **ENSEMBLE** | | **2 137** | | | **3 458** | **1 621** | **47** |

La base élèves académique (BEA) a été collectée à 100 %.

Une fois les fichiers uniformisés et vérifiés, l'ensemble des fichiers a été compilé pour constituer une base globale. Une fois constituée, cette base est apurée pour supprimer, quand cela était possible, les individus hors-champ (détectés comme poursuivants en 2013-2014) et les doublons.

### 2.2.2. Les couplages et apurements des fichiers

Selon les établissements, le Céreq a collecté des individus inscrits ou sortants.

**Tableau 6 • Nature des fichiers collectés**

| Type d'établissement de collecte | Collectés auprès des établissements, bases nominatives | | Collectés auprès de la Depp et de la Dgesip, bases individuelles mais non nominatives | |
|---|---|---|---|---|
| | 2012-2013 | 2013-2014 | 2012-2013 | 2013-2014 |
| RECTORATS | Inscrits | Inscrits | | |
| SIFA | Sortants | | *Sifa* inscrits | *Sifa* inscrits |
| UNIVERSITÉS | Sortants | | *Sise* inscrits et *sise* couplage : sortants et inscrits | *Sise* inscrits |
| DRJS – SPORTS | Sortants | | | |
| ÉCOLES FONCTION PUBLIQUE<br>+ ENS<br>+ DGA | Sortants<br>(fonctionnaire ou non) | | | |
| SANTÉ SOCIAL | Sortants | | | |
| ÉCOLES DE COMMERCE | Sortants | | | |
| ÉCOLES D'INGÉNIEURS | Sortants | | | |
| ÉCOLES DE LA FORMATION AGRICOLE | Sortants | | | |
| IEP | Sortants | | | |
| ÉCOLES DANS LES SECTEURS SERVICES / INDUSTRIEL | Sortants | | | |
| LYCÉES AGRICOLES, CULTURE, CENTRES ET FACULTÉS PRIVÉS + CENTRES DE FORMATION PEDAGOGIQUE PRIVÉS | Sortants | | | |

Pour les bases transmises par les rectorats, pour les bases SIFA, et pour les fichiers SISE des inscrits en 2012-2013 et en 2013-2014, la disponibilité d'un identifiant commun de gestion (l'identifiant national étudiant – INE) a permis de repérer de manière assez précise les poursuites d'études internes à ces trois champs en comparant les fichiers 2012-2013 et les fichiers disponibles pour 2013-2014.

Les poursuites d'études, comme le passage du champ des rectorats ou des centres de formation des apprentis vers l'université, ne sont pas toujours repérables à partir de l'INE (car certains d'entre eux sont manquants ou provisoires). Une partie des poursuites d'études ont toutefois été repérées à partir des noms, prénoms, date de naissance (mois, année) entre les fichiers des inscrits en 2012-2013 et les fichiers des inscrits en 2013-2014.

Il reste malgré ces étapes de nombreux individus hors-champ dans la base de sondage finale. Les individus poursuivants qui n'ont pas pu être repérés lors de cette étape sont détectés au moment de l'enquête au niveau du questionnaire filtre. Pour l'enquête 2016 auprès de la Génération 2013, le repérage des individus hors-champ lors de la collecte représente 50 % des répondants.

Outre les poursuites d'études, des doublons parmi les inscrits de 2012-2013 sont également repérés, détectés à partir des noms, prénoms et date de naissance (mois, année). Le dédoublonnage est réalisé sur la base complète contenant l'ensemble des établissements collectés (tout type d'établissement confondu).

Pour les individus inscrits dans deux cursus simultanément, un ordre de priorité a été défini afin de sélectionner la formation réellement suivie durant l'année 2012-2013.

Par ailleurs, les individus sortants de classes préparatoires, ne rentrant pas dans le champ de l'enquête, ont été supprimés.

Les individus âgés de plus de 35 ans en 2013 ont également été supprimés, sauf ceux appartenant à des extensions de champ. Il s'agit des individus en post-initial issus des formations du sport et de l'animation, du domaine de la santé et du social ainsi que des écoles de la fonction publique.

La base complète regroupe tous les individus ayant effectué une formation initiale en France (métropolitaine et DOM), qu'ils soient étudiants étrangers ou non.

En définitive, la base de sondage se compose de **1 282 273** individus.

### 2.2.3. Bilan général de la collecte auprès de l'ensemble des établissements

À l'issue de la collecte des bases d'élèves, la structure des bases de données qui servent à constituer la base de sondage est la suivante :

**Tableau 7 • Bilan comparatif de la collecte des bases d'élèves auprès des établissements**

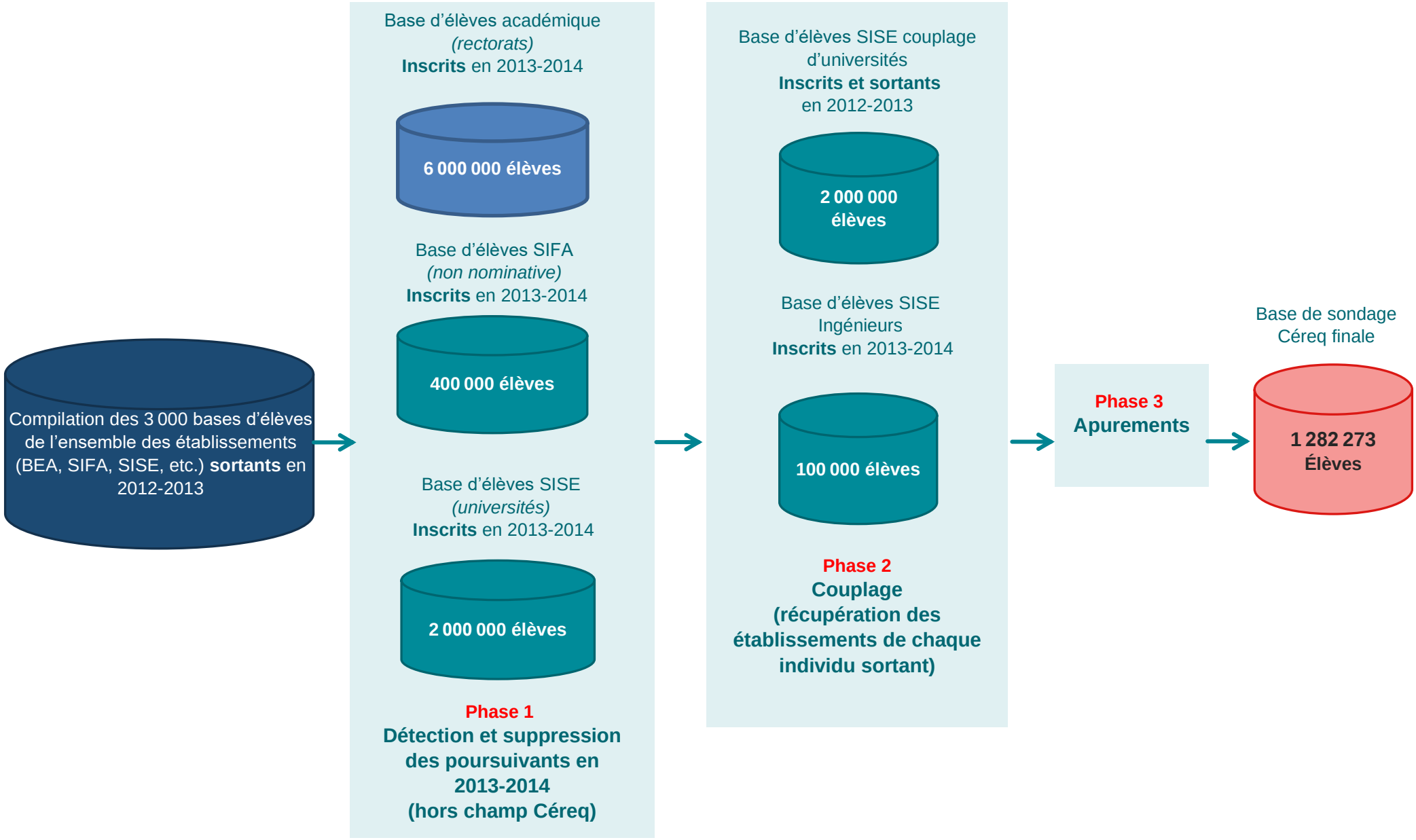
| | Génération 2010 (Source : base de sondage) | | Génération 2013 (Source : base de sondage) | |
|---|---|---|---|---|
| **Type d'établissement** | **Effectif** | **%** | **Effectif** | **%** |
| **Rectorat** | 511 219 | 44,5 | 545 561 | 42,5 |
| **Université** | 319 617 | 27,8 | 332 107 | 25,9 |
| ***Sifa*** | 146 429 | 12,8 | 178 230 | 13,9 |
| **Établissements agricoles** | 55 862 | 4,9 | 55 906 | 4,4 |
| **Autres établissements** | 114 513 | 10 | 171 469 | 13,4 |
| **Total** | **1 147 640** | **100** | **1 282 273** | **100** |

Certains établissements supplémentaires qui n'étaient pas retenus auparavant ont été intégrés dans la collecte de la base de sondage (quelques établissements privés, catholiques, classes supplémentaires de STS). Ces établissements qui offrent en partie de la formation initiale ont donc été collectés, cependant il est présupposé que globalement ces établissements présentent des taux d'hors-champ importants.

Malgré l'effort de collecte auprès de tous les établissements de formation sur le territoire métropolitain et ultramarin (trois collectes distinctes, dont une sous-traitée), un certain nombre d'établissements ne transmettent pas leurs fichiers d'élèves inscrits. Sur la base du fichier des établissements collectés individuellement, le taux de réponse est de 47 % (environ 1 600 établissements).

Ce taux, assez faible, peut en partie s'expliquer par l'absorption d'établissements tels que les CFA, les lycées agricoles dans les bases d'élèves centralisées.

**Figure 5 • Processus de constitution de la base de sondage**



### 2.2.4. Estimation du taux de couverture des individus de la base de sondage

L'estimation de la couverture de la base de sondage, qui renvoie au degré d'exhaustivité de la collecte des fichiers d'élèves provenant des établissements de formation, a été effectuée en s'appuyant sur diverses sources administratives. Sauf mention contraire, les taux de couverture présentés portent sur les effectifs couverts et non sur le nombre d'établissements couverts.

Chaque individu de la base de sondage provient soit de bases centralisées fournies par les ministères, soit de fichiers collectés directement auprès des établissements de formation.

Par ailleurs, chaque établissement de formation est classé dans une nomenclature correspondant au type de formation (école d'ingénieur, école de commerce…).

L'estimation du taux de couverture se fait selon la source (base centralisée ou non) et le type d'établissement :

**Pour les bases centralisées** * (*Sysca* : établissements du secondaire ; *Sifa* : CFA, Formations agricoles), on considère que la couverture est de 100 %, car les bases reçues sont censées être exhaustives.

**Pour les établissements de la collecte**, on estime le taux de couverture selon les informations annexes détenues :

- Dans certains cas nous disposons d'effectifs de référence fournis par une source annexe (transmis par les ministères, par exemple pour les universités et établissements rattachés, ou bien récupérés dans diverses sources comme le RERS[5]). Nous comparons alors nos effectifs de sortants à ces effectifs de référence[6]. Nous estimons ainsi le ratio d'inscrits collectés pour estimer un taux de couverture pour le type d'établissement concerné.

- Lorsqu'aucun effectif de référence n'est disponible, le taux de couverture est estimé selon le ratio d'établissements collectés parmi les établissements repérés dans notre champ (méthode moins précise, car les effectifs d'élèves sont variables d'un établissement à l'autre, ce qui fait qu'un faible nombre d'établissements collectés peut correspondre à un pourcentage élevé d'élèves collectés, et inversement).

Dans l'étape finale, pour estimer le taux de couverture global de la base de sondage, on utilise les taux de couverture précédemment calculés pour estimer les effectifs de sortants attendus dans la base de sondage. Le ratio entre effectif de la base de sondage et effectif attendu donne le taux de couverture : 85 % pour la base de sondage de Génération 2013.

[5] Repères et références statistiques, publication annuelle de la Depp et du Sies.
[6] À noter que l'on compare parfois des effectifs de diplômés à des effectifs sortants, qui peuvent être non diplômés, mais cela est supposé rare pour les types de formations concernés comme les écoles de commerce ou d'ingénieurs.

**Tableau 8 • Taux de couverture global et par type d'établissement de formation**

| Type d'établissement | Effectif collecté dans la base de sondage | Taux de couverture moyen en % |
|---|---|---|
| **Lycées et collèges MEN *** | 568 432 | 100 |
| **Universités** | 352 386 | 90 |
| **Centre de formation des apprentis *** | 182 608 | 100 |
| **Lycées agricoles *** | 56 107 | 100 |
| **Écoles santé social** | 47 248 | 49 |
| **Écoles de commerce** | 30 142 | 91 |
| **Universités autres** | 23 317 | 26 |
| **Écoles ingénieurs** | 20 060 | 83 |
| **Écoles ministère de la Culture** | 12 087 | 37 |
| **DRJSCS** | 11 921 | 70 |
| **Écoles secteur service** | 4 719 | 28 |
| **DGAFP** | 3 955 | 46 |
| **Écoles secteur industriel** | 2 222 | 62 |
| **IEP** | 1 881 | 85 |
| **CIFRE** | 960 | 95 |
| **Centres privés d'enseignement** | 836 | 52 |
| **Écoles formations agricoles** | 809 | 50 |
| **Écoles administrations publiques** | 489 | 39 |
| **Écoles normales supérieures** | 480 | 40 |
| **Ensemble** | **1 320 659** | **85** |

### 2.2.5. Amélioration de la qualité de la base de sondage : les numéros de téléphone, les mails

Pour la base de sondage de l'enquête 2016 auprès de la Génération 2013, tout a été mis en œuvre pour obtenir un nombre maximum de numéros de téléphone par individu.

Lors de la collecte des fichiers individus auprès des établissements, la structure du fichier individus demandée intégrait la nécessité de disposer de l'ensemble des numéros de téléphone disponibles, notamment les numéros de portable.

Au total, 80 % des individus présents dans la base de sondage ont au moins un numéro de téléphone. Pour ces derniers, le nombre moyen de numéros de téléphone est estimé à 1,3 par individu. Cet indicateur a peu évolué depuis la Génération 2007.

**Tableau 9 • Évolution du nombre de numéros de téléphone disponibles**

| Génération | Individus ayant au moins un numéro dans la base de sondage |
|---|---|
| 2013 | 80 % |
| 2010 | 82 % |
| 2007 | 75 % |
| 2004 | 27 % |

**Tableau 10 • Fréquence des numéros de téléphone disponibles par individu**

| Nombre total de numéros de téléphone | Effectif | % | Effectif cumulé | % cumulé |
|---|---|---|---|---|
| *0* | *262 950* | *20,51* | *262 950* | *20,51* |
| *1* | *702 508* | *54,79* | *965 458* | *75,29* |
| *2* | *309 196* | *24,11* | *1 274 654* | *99,41* |
| *3* | *7 619* | *0,59* | *1 282 273* | *100,00* |

Pour les 20 % restants, une procédure de recherche téléphonique est réalisée (cf. partie 4.3.) pour tenter de récupérer un ou plusieurs numéros de téléphone pour chaque individu ne disposant d'aucun numéro dans la base de sondage. En plus des numéros fournis par les établissements, les individus pour lesquels au moins un numéro est disponible sont également inclus dans cette phase de recherche afin d'augmenter les chances de contact.

Concernant les mails, la base de sondage les fournit pour seulement 33 % des individus.

### 2.2.6. Le géocodage de la base de sondage

Pour le tirage de l'échantillon, la base de sondage de l'enquête 2016 auprès de la Génération 2013 a été géocodée afin d'identifier les sortants de quartier prioritaire de la politique de la ville (QPV). Le géocodage a été réalisé à partir des adresses des parents/jeunes à la sortie de leur formation en 2012-2013.

Ainsi, nous identifions 8,4 % des jeunes de la base de sondage qui résidaient en QPV l'année de la sortie du système éducatif (ou dont les parents résidaient en QPV).

Cette variable est utilisée dans le calcul des probabilités de tirage dans l'échantillon.

# 3. Le plan de sondage et la constitution de l'échantillon

## 3.1. Objectifs du plan de sondage

Le plan de sondage de la Génération 2013 est construit pour répondre aux objectifs suivants :

- réaliser un nombre minimum de questionnaires équilibrés par rapport aux formations des jeunes sortants ;
- tenir compte de certains types de formations sur lesquelles le Céreq souhaite réaliser une analyse fine ou sur lesquelles l'attrition lors des réinterrogations est élevée ;
- satisfaire les demandes des partenaires d'extension ;
- disposer d'une réserve, utilisable totalement ou partiellement, en fonction des taux de réponse observés en cours d'enquête.

L'échantillon est construit de façon à atteindre environ 24 000 questionnaires permettant de répondre aux besoins du Céreq et de ses partenaires. Pour les travaux du Céreq, 10 000 questionnaires sont prévus, incluant une extension d'échantillon interne de 1 500 questionnaires pour étudier les effets de la réforme du baccalauréat professionnel. Pour cela, environ 1 000 questionnaires supplémentaires sont prévus pour les sortants de cette formation et des questionnaires additionnels pour les formations « entourant » ce cursus : 500 pour les CAP. Les BTS, et les licences professionnelles étant suréchantillonnés par ailleurs. L'échantillon devra également assurer l'obtention d'environ 14 000 questionnaires pour répondre aux demandes des partenaires d'extension.

À titre indicatif, l'ordre de grandeur de l'échantillon, réserve comprise, devrait se situer autour de 180 000 unités statistiques.

Cette partie présente les différentes étapes de l'échantillonnage de l'enquête 2016 auprès de la Génération 2013. La méthodologie retenue permet de prendre en compte trois difficultés importantes pour la constitution de l'échantillon.

En effet, le Céreq construit lui-même la base de sondage des élèves sortants du système éducatif en 2012-2013. Cette base de sondage présente un défaut de couverture lié à la non-réponse de certains établissements et un défaut de surcouverture lié à la présence dans la base d'individus hors-champ de l'enquête.

Par ailleurs, de nombreux acteurs publics partenaires financent des extensions d'échantillon à l'enquête sur leur population d'intérêt. Le plan de sondage tient compte de ces demandes d'extension. La difficulté principale provient du fait que certaines extensions se croisent. Une solution innovante, par calage sur marges sur les cibles d'extension, est proposée pour éviter d'aboutir à des échantillons dans lesquels le poids des intersections entre extensions soit trop important.

La méthodologie peut se décomposer en deux phases (parties 3.2. et 3.3.), qui elles-mêmes se décomposent en différentes étapes :

- **Phase A : Calcul des probabilités individuelles de tirage.**
  - Étape A1 : taux de couverture.
  - Étape A2 : détermination des probabilités de tirage en l'absence d'extension.
  - Étape A3 : prise en compte des cibles d'extensions dans le calcul des probabilités de tirage.

– **Phase B : Tirage équilibré de l'échantillon.**

  - Étape B1 : tirage de l'échantillon global (principal + réserve).
  - Étape B2 : tirage de l'échantillon principal.

La partie 3.4. présente le bilan de l'échantillon tiré.

## 3.2. Phase A : Calcul des probabilités individuelles de tirage

### 3.2.1. Étape A1 : taux de couverture

L'échantillon Génération est tiré dans une base de sondage construite par le Céreq. Le tableau suivant donne par grands types d'établissement :

- Les effectifs présents dans la base de sondage (après opération de dédoublonnage).
- Le taux de couverture moyen (en %).
- Le poids de couverture moyen.
- L'effectif estimé de la base de sondage en l'absence de défaut de couverture.

Pour un type d'établissement donné, le taux de couverture est estimé en utilisant la meilleure information externe disponible (par ordre de priorité, nombre de sortants nationaux, nombre de diplômés, nombre d'élèves inscrits). En l'absence d'informations externes, le taux de couverture est estimé par le taux de réponse des établissements pour un type donné.

Pour certains types d'établissement, plusieurs sources ou données détaillées (notamment sur des effectifs par formation) peuvent être mobilisées. Cela explique, qu'au sein d'un même type d'établissement, il y ait plusieurs taux de couverture.

Par exemple, pour le type d'établissement « universités », le taux de couverture est calculé en rapportant les effectifs de sortants présents dans la base de sondage aux effectifs de sortants présents dans la base ministérielle SISE (système d'information sur le suivi de l'étudiant). Ce calcul s'effectue au niveau du croisement des variables *région*, *niveau de formation* et *discipline de formation*. Ainsi, pour la région « Midi-Pyrénées », le niveau de formation « M2 » et la discipline « sciences humaines et sociales », la base de sondage issue de la collecte auprès des universités contient 765 individus présumés sortants. Dans la source SISE, il y a un effectif de 772 sortants. Ainsi le taux de couverture, à ce niveau de croisement, est de 765/772 = 99 %.

Ensuite, pour chaque individu présent dans la base de sondage, lui est attribué le poids de couverture qui le concerne. En reprenant l'exemple, le poids de couverture des individus concernés (universités de Midi-Pyrénées, M2, SHS) est donc de 1/0,99 = 1,01.

Pour les notations suivantes, le poids de couverture individuel sera noté $pcouv_i$.

**Tableau 11 • Sous-couverture de la base de sondage par grands types d'établissements**

| Type d'établissement | Effectif base de sondage | Taux de couverture moyen (en %) ** | Poids de couverture moyen | Effectif théorique corrigé du défaut de couverture* |
|---|---|---|---|---|
| LYCÉES ET COLLÈGES (MEN) | 548 922 | 100 | 1,00 | 548 922 |
| UNIVERSITÉS | 332 514 | 90 | 1,11 | 367 745 |
| CFA | 177 490 | 100 | 1,00 | 177 490 |
| LYCÉES AGRICOLES | 55 742 | 100 | 1,00 | 55 742 |
| ÉCOLES PROFESSIONS SOCIALES | 47 106 | 49 | 2,04 | 96 171 |
| ÉCOLES DE COMMERCE | 28 716 | 39 | 2,55 | 73 357 |
| AUTRES UNIVERSITÉS | 18 939 | 33 | 3,17 | 56 777 |
| ÉCOLES D'INGÉNIEURS | 18 338 | 48 | 2,10 | 38 578 |
| ÉCOLES – MINISTÈRE DE LA CULTURE | 12 036 | 37 | 2,68 | 32 197 |
| DRJSCS | 11 697 | 70 | 1,43 | 16 785 |
| CLASSES PRÉPARATOIRES ET AUTRES ÉTABLISSEMENTS | 4 902 | 25 | 4,08 | 19 949 |
| ÉCOLES SECTEUR SERVICE | 4 664 | 27 | 3,64 | 16 962 |
| DGAFP | 3 644 | 41 | 2,44 | 8 903 |
| ÉCOLES SECTEURS INDUSTRIEL | 2 154 | 69 | 1,44 | 3 111 |
| IEP | 1 847 | 85 | 1,18 | 2 177 |
| CIFRE | 943 | 95 | 1,05 | 994 |
| ÉCOLES FORMATION AGRICOLES | 756 | 46 | 2,19 | 1 646 |
| CENTRES PRIVÉS D'ENSEIGNEMENT | 693 | 55 | 1,82 | 1 263 |
| ÉCOLES ADMINISTRATIONS PUBLIQUES | 467 | 39 | 2,56 | 1 197 |
| ÉCOLES NORMALES SUPÉRIEURES | 437 | 36 | 2,78 | 1 214 |
| **ENSEMBLE** | **1 272 007** | **84** | **1,21** | **1 521 180** |

* l'effectif théorique corrigé du défaut de couverture est égal à la somme des poids de couverture individuels.
** le taux de couverture moyen est obtenu en faisant le ratio « effectif base de sondage » sur « effectif théorique corrigé du poids de couverture ». Cet indicateur a été préféré à la moyenne des taux de couverture individuels. Les écarts entre les deux indicateurs sont faibles.

Le taux de couverture de l'ensemble de la base de sondage est ainsi estimé à 84 %. La qualité de la couverture n'est pas uniforme selon l'origine des données : la couverture est quasiment exhaustive lorsque les fichiers sont centralisés[7]. À l'inverse, les taux de couverture sont moins bons pour les types d'établissements concernés par la collecte auprès des établissements. Dans l'ensemble, près de 90 % des individus de la base de sondage sont issus d'une réception de fichiers centralisés, ce qui représente environ les trois quarts des individus de notre champ.

[7] BEA pour les bases rectorales lorsqu'il s'agit des collèges et des lycées. Fichier Agri pour les lycées agricole. SISE pour les Universités. SIFA pour les apprentis. Pour ces différentes sources, la couverture est quasi-exhaustive.

### 3.2.2. Étape A2 : détermination des probabilités de tirage en l'absence d'extension

Cette étape consiste à déterminer les probabilités de tirage qui simulent un échantillon national donnant lieu à 8 500 répondants dans le champ Céreq[8]. Pour ce calcul, il est nécessaire de faire intervenir le poids de couverture, la probabilité de réponse anticipée ainsi que la probabilité d'appartenir au champ de l'enquête.

En effet, la deuxième difficulté liée à la base de sondage provient de la présence (importante) d'individus hors-champ. Il s'agit essentiellement de poursuivants l'année scolaire suivante (non repérés au préalable) mais également d'individus qui ont déjà interrompu leur scolarité dans le passé (post-initiaux). Sur l'ensemble de la base de sondage, le taux d'hors-champ est proche de 50 %. L'effectif théorique de la base de sondage corrigé du défaut de couverture étant d'environ 1 500 000 individus, la population théorique du champ Céreq est approximativement de l'ordre de 750 000 individus.

#### *Définitions – Propriétés*

***Probabilité de réponse anticipée***

Les probabilités de réponses anticipées individuelles sont calculées à l'aide de modèles logistiques sur les données de l'enquête 2013 auprès de la Génération 2010. Les modèles font intervenir les variables explicatives : âge, appartenance à une zone urbaine sensible (Zus), région de l'établissement de formation, indicatrice valant 1 s'il s'agit d'un niveau terminal (classe permettant d'obtenir un diplôme à la fin de l'année scolaire), présence d'au moins un numéro de téléphone, et strate de sortie. Ici, est utilisé (historiquement) de manière inappropriée le terme de « strate » pour désigner une variable agrégeant des classes de sortie sans qu'elle constitue une variable de stratification au sens propre pour le tirage de l'échantillon. Cette variable a été calculée à partir d'une classification des classes de sortie concernant leur homogénéité sur les taux de réponse et d'hors-champ. Cette variable « strate » est donc utilisée comme variable explicative du taux de réponse avec les autres variables suscitées. Par la suite, la probabilité de réponse d'un individu i, est notée :

$$pr_i = P(i\ accepte\ de\ répondre \mid i \in s)$$

***Probabilité de réponse dans le champ anticipé***

La présence d'individus hors-champ dans la base de sondage impose d'introduire, en plus de la probabilité de réponse, la notion de probabilité de répondre dans le champ. En effet, globalement, un individu sur deux issu de la base de sondage ne fait pas partie du champ de l'enquête. Ce taux d'appartenance au champ Céreq varie sensiblement selon les classes de sortie. Cela implique donc de définir une probabilité de répondre dans le champ qui est l'évènement « a répondu et est dans le champ ». Cet évènement est modélisé à partir de l'enquête précédente et des mêmes variables explicatives (issues uniquement de la base de sondage) utilisées précédemment.

La probabilité de réponse dans le champ d'un individu i, est notée :

$$pr_i^C = P(\{i\ accepte\ de\ répondre\ \cap i\ dans\ le\ Champ\} \mid i \in s)$$

***Estimation de la taille de l'échantillon à partir des probabilités de tirage***

Soit :

- $U$ la base de sondage ;
- $s$ un échantillon ;
- $\pi_i = P(i \in s)$ la probabilité que l'échantillon $s$ contienne l'individu $i$ (probabilité de tirage de i).

[8] Le Céreq finance pour ses besoins propres 10 000 questionnaires afin d'obtenir un échantillon national « représentatif » de 8 500 répondants et une extension d'échantillon sur les sortants de baccalauréat professionnel de 1 500 répondants.

L'estimation de la taille de l'échantillon $n$ est donnée par :

$$\sum_{i \in U} \pi_i = E(n)$$

La somme des probabilités de tirage sur toute la base de sondage $U$ est égale à l'espérance de la taille de l'échantillon.

Dans le cadre d'un sondage équilibré sur les probabilités d'inclusion, une égalité entre la taille de l'échantillon et la somme des probabilités de tirage sur *U* est vraie si le tirage est parfaitement équilibré (phase de vol uniquement).

$$\sum_{i \in U} \pi_i = n$$

***Estimation du nombre de répondants dans le champ***

L'estimation du nombre de répondants dans le champ $\mathrm{nr}^C$ est donnée par :

$$\sum_{i \epsilon U} P(\{i \; répond \cap i \; dans \; le \; Champ\} \mid i \in s) * \pi_i \; = \; \mathrm{E} \; (nombre \; de \; répondants \; dans \; le \; champ)$$

Soit,

$$\sum_{i \epsilon U} pr_i^C * \pi_i \; = \; \mathrm{E}(\mathrm{nr}^C)$$

Ainsi l'espérance du nombre de répondants dans le champ est calculée en faisant cette somme sur l'intégralité de la base de sondage $U$.

### *Calcul des probabilités de tirage pour simuler un échantillon socle permettant d'obtenir 8 500 répondants*

Une fois ces notations introduites, les probabilités de tirage $\pi_{i1}$ sont déterminées pour simuler un échantillon socle permettant d'obtenir 8 500 répondants dans le champ Céreq.

Les probabilités de tirage sont construites pour surreprésenter les individus dont les taux de réponse attendus sont plus faibles ainsi que ceux appartenant à un type d'établissement mal couvert dans la base de sondage. La structure des répondants sera proche de la structure théorique de la population cible (du moins sur les variables utilisées pour le calcul des poids de couverture et des probabilités de réponse). Ainsi, les probabilités de tirage $\pi_{i1}$sont de la forme :

$$\pi_{i1} = \frac{poids \; de \; couverture_i}{probabilité \; de \; répondre_i} * coeff1 = \frac{pcouv_i}{pr_i} * coeff1$$

Où $coeff1$ est un coefficient de dilatation identique pour tous les individus de la base de sondage. Ce coefficient est calculé pour simuler 8 500 répondants dans le champ à partir de la relation suivante :

$$\sum_{i \epsilon U} pr_i^C * \pi_{i\,1} = \; \mathrm{E}(\mathrm{nr}^C) \text{ Soit } coeff1 = \; 8500 / \sum_{i \in U} \left( pr_i^C * \frac{pcouv_i}{pr_i} \right)$$

La valeur du coefficient $coeff1$ est environ 0,011.

### 3.2.3. Étape A3 : prise en compte des cibles d'extension dans le calcul des probabilités de tirage

Pour cette enquête, cinq partenaires publics financent des extensions d'échantillon sur une ou plusieurs populations d'intérêt. À ces demandes externes, certaines populations font également l'objet d'extension pour les besoins propres du Céreq (collège, baccalauréats généraux, professionnels et technologiques, CAP, docteurs en santé).

Pour chacune des populations d'intérêt, une cible de questionnaires (*i.e*. d'individus répondants dans le champ) a été déterminée. Le tableau, ci-après, précise pour chaque sous-population d'intérêt, la cible souhaitée ainsi que le nombre de questionnaires espéré à partir de l'échantillon socle (*i.e.* après l'étape 2 de calcul des probabilités de tirage pour obtenir 8 500 questionnaires).

L'objectif de cette troisième étape est de déterminer, à partir du calcul des probabilités de tirage de l'étape 2, les coefficients de suppléments de tirage nécessaires pour atteindre les cibles sur chaque population d'intérêt.

La principale difficulté est liée au recoupement des extensions. Le risque est d'aboutir à une structure de l'échantillon qui soit atypique du point de vue des extensions (surreprésentation des intersections notamment dans le cas particulier d'une inclusion d'une extension dans une autre).

**Tableau 12 • Effectifs cibles par sous-populations d'extensions**

| Sous population d'extension | Codage des sous-populations | Cible (effectif extension +réserve) | Nombre de questionnaires à partir de l'échantillon socle |
|---|---|---|---|
| COLLEGE | c1 | 903 | 404 |
| BACCALAURÉAT GÉNÉRAL ET TECHNOLOGIQUE | c2 | 1 068 | 570 |
| BACCALAURÉAT PROFESSIONNNEL | c3 | 2 391 | 1 374 |
| CAP | c4 | 1 373 | 860 |
| CGDD | c5 | 3 349 | 287 |
| *CGDD niveau 1* | c6 | 842 | 83 |
| *CGDD niveau 2* | c7 | 817 | 33 |
| *CGDD niveau 3* | c8 | 567 | 31 |
| *CGDD niveau 4* | c9 | 748 | 102 |
| *CGDD niveau 5* | c10 | 416 | 38 |
| DGESIP | c11 | 7 026 | 4 503 |
| *DGESIP autre* | c12 | 1 959 | 1 692 |
| *DGESIP BTS* | c13 | 972 | 697 |
| *DGESIP Grandes écoles* | c14 | 771 | 705 |
| *DGESIP IUT* | c15 | 350 | 122 |
| *DGESIP Licence* | c16 | 1 464 | 607 |
| *DGESIP Master* | c17 | 1 509 | 681 |
| QUARTIER PRIORITAIRE DE LA POLITIQUE DE LA VILLE | c18 | 2 287 | 752 |
| SANTÉ SOCIAL | c19 | 5 981 | 455 |
| SANTÉ SOCIAL *aide-soignante* | c20 | 1 180 | 144 |
| SANTÉ SOCIAL *assistant social* | c21 | 349 | 17 |
| SANTÉ SOCIAL *auxiliaire puéricultrice* | c22 | 454 | 26 |
| SANTÉ SOCIAL *conseiller économie familiale* | c23 | 208 | 10 |
| SANTÉ SOCIAL *éducateur jeunes* | c24 | 324 | 10 |
| SANTÉ SOCIAL *éducateur spécialisé* | c25 | 474 | 31 |
| SANTÉ SOCIAL *ergothérapeute* | c26 | 62 | 3 |
| SANTÉ SOCIAL *infirmier* | c27 | 1 191 | 164 |
| SANTÉ SOCIAL *masseur kinésithérapeute* | c28 | 427 | 13 |
| SANTÉ SOCIAL *moniteur éducateur* | c29 | 461 | 19 |
| SANTÉ SOCIAL *orthophoniste* | c30 | 78 | 3 |
| SANTÉ SOCIAL *orthoptiste* | c31 | 31 | 1 |
| SANTÉ SOCIAL *podologue pédicure* | c32 | 140 | 3 |
| SANTÉ SOCIAL *psychomotricien* | c33 | 84 | 3 |
| SANTÉ SOCIAL *sage-femme* | c34 | 306 | 7 |
| SPORT | c35 | 1 678 | 50 |
| THESE | c36 | 2 446 | 128 |
| DOCTEUR EN SANTÉ | c37 | 112 | 98 |
| **ENSEMBLE** | **c38** | **24 700** | **8 490** |

## *Explication des méthodes par un exemple*

Supposons de deux sous-populations d'intérêt A et B. Les cibles de questionnaires fixées sont respectivement $C_A$ et $C_B$.

Sur l'échantillon socle (visant 8 500 questionnaires), le nombre de questionnaires estimé de A, noté $nsoc_A$, et de B, noté $nsoc_B$.

***1er cas : supposons $A \cap B = \emptyset$***

Dans ce cas, les coefficients de dilatation, notés $dil_A$ et $dil_B$, sont immédiats et donnés par :

$$dil_A = \frac{C_A}{nsoc_A} \; et \; dil_B = \frac{C_B}{nsoc_B}$$

Et les probabilités de tirage tenant compte des extensions sont données par :

$$\text{pour } i \epsilon A, \pi_{i2} = \pi_{i1} \times dil_A = \pi_{i1} \times \frac{C_A}{nsoc_A}$$
$$\text{pour } i \epsilon B, \pi_{i2} = \pi_{i1} \times dil_B = \pi_{i1} \times \frac{C_B}{nsoc_B}$$

Prenons les données fictives suivantes :

Les cibles sont fixées à $C_A = 1000$ et $C_B = 1300$, dans toute la suite.

Supposons que $nsoc_A = 100$ et $nsoc_B = 200$ (sans oublier que $nsoc_{A \cap B} = 0$)

Par conséquent :

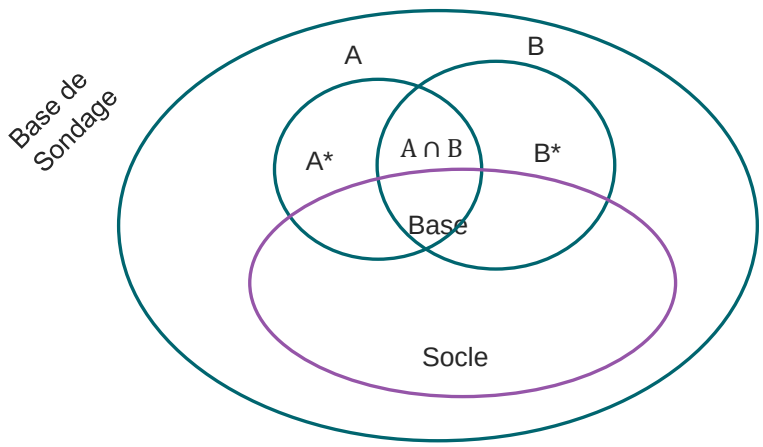
$$\text{dil}_A = \frac{1000}{100} = 10 \text{ et dil}_B = \frac{1300}{200} = 6{,}5$$

Autrement dit, la probabilité de tirage des individus de A sera multipliée uniformément par 10, pour ceux de B par 6,5.

***2e cas : supposons $A \cap B \neq \emptyset$***

La principale difficulté de la prise en compte des cibles d'extensions est liée au fait que certaines sous-populations d'intérêt se croisent.
Introduisons les notations supplémentaires suivantes :

$A^* = A \backslash (A \cap B)$
$B^* = B \backslash (A \cap B)$
$Z = SOCLE \backslash (A \cup B)$



Supposons, comme dans le cas 1, que $nsoc_A = 100$ et $nsoc_B = 200$

Dans l'hypothèse où $nsoc_{A\cap B} = 50$ alors $nsoc_{A^*} = 50$ et $nsoc_{B^*} = 150$

La taille de l'échantillon socle est de $8\,500$, par conséquent :

$$nsoc_Z = 8500 - 50 - 50 - 150 = 8250$$

Une première méthode naturelle consiste à traiter séquentiellement chaque extension. Les probabilités de tirage des individus de A sont dilatées (avec $dil_A = 1000/100$). Comme l'intersection est non vide avec B, cette étape va conduire à gonfler le nombre d'individus de B présents dans l'échantillon (et répondants dans le champ). Le coefficient de dilation $dil_B$ est recalculé avec les nouveaux effectifs ($dil_B = 1300/650 = 2$), ce qui revient à doubler les poids de tirage de B. Le tableau ci-dessous détaille les calculs effectués pour cette première méthode et sa variante qui consiste à réaliser le même processus en commençant par B.

**Tableau 13 • Méthodes séquentielles de calcul des suppléments de tirage**

| | | MÉTHODE 1 (M1) : D'abord A puis B | | MÉTHODE 1bis (M1bis) : D'abord B puis A | |
|---|---|---|---|---|---|
| | | D'abord A (dila = 1000/100 appliqué à la population A) | Puis B (dilb = 1300/650 appliqué à la population B) | D'abord B (dilb = 1300/200 appliqué à la population A) | Puis A (dila = 1000/375 appliqué à la population B) |
| | Nombre de répondants estimé dans l'échantillon socle | Nombre de répondants estimé dans l'échantillon intermédiaire | Nombre de répondants estimé dans l'échantillon final méthode 1 (F1) | Nombre de répondants estimé dans l'échantillon intermédiaire | Nombre de répondants estimé dans l'échantillon final méthode 1bis (F1bis) |
| **A*** | 50 | 500 | 500 | 50 | 133 |
| **B*** | 150 | 150 | 300 | 975 | 975 |
| **A∩B** | 50 | 500 | 1 000 | 325 | 867 |
| **Z** | 8 250 | 8 250 | 8 250 | 8 250 | 8 250 |
| **Total** | 8 500 | 9 400 | 10 050 | 9600 | 10 225 |
| **A** | 100 | 1 000 | 1 500 | 375 | 1 000 |
| **B** | 200 | 650 | 1 300 | 1 300 | 1 842 |

Il y a un double inconvénient à cette méthode « intuitive » :

- Les résultats dépendent de l'ordre dans lequel sont traitées les cibles d'extensions.

- Dans chacune des variantes de cette première méthode, le bon nombre de questionnaires (par rapport à la cible fixée) est obtenu uniquement pour l'une des deux sous-populations (celle traitée en dernier). En revanche, pour l'autre population d'intérêt, le nombre de questionnaires obtenus sera sensiblement supérieur à la cible fixée. Dans l'échantillon F1, 1 500 répondants de A (au lieu des 1 000 cibles) et dans F1bis, 1 842 répondants de B (au lieu des 1 300 cibles).

Une deuxième méthode, basée sur la précédente, permet de pallier le problème de dépendance de l'ordre de traitement des cibles d'extensions. Il suffit de cumuler les coefficients multiplicatifs initiaux. De fait, les coefficients de dilatation sont définis tels que :

$$dil_A = 1 + \alpha_A \; et \; dil_B = 1 + \alpha_B$$

Et les probabilités de tirage sont dilatées comme suit :

$$pour \; i \epsilon A^*, \; \pi_{i2} = \pi_{i1} \times (1 + \alpha_A)$$

$$pour \; i \epsilon B^*, \; \pi_{i2} = \pi_{i1} \times (1 + \alpha_B)$$

$$pour \; i \epsilon A \cap B, \; \pi_{i2} = \pi_{i1} \times (1 + \alpha_A + \alpha_B)$$

**Tableau 14 • Méthode simultanée de calcul des suppléments de tirage**

| | | MÉTHODE 2 (M2) : A et B simultanément | |
|---|---|---|---|
| | Nombre de répondants estimé dans l'échantillon socle | Coefficient de dilatation | Nombre de répondants estimé dans l'échantillon final méthode 2 (F2) |
| A* | 50 | 10 | 500 |
| B* | 150 | 6,5 | 975 |
| A∩B | 50 | 15,5 | 775 |
| Z | 8 250 | 1 | 8 250 |
| Total | 8 500 | | 10 500 |
| A | 100 | | 1 275 |
| B | 200 | | 1 750 |

Cette méthode n'est plus dépendante de l'ordre des cibles d'extensions. En revanche, les deux cibles sur les sous-populations A et B sont dépassées.

Au-delà des inconvénients déjà cités, le principal problème des deux méthodes présentées résulte en des échantillons qui surreprésentent exagérément les individus de l'intersection des deux extensions. En voici la démonstration dans le tableau suivant :

**Tableau 15 • Poids des intersections**

| Poids des répondants de A∩B | Socle | F1 | F1bis | F2 |
|---|---|---|---|---|
| **Parmi ceux de A** | 50 % | 67 % | 87 % | 61 % |
| **Parmi ceux de B** | 25 % | 77 % | 47 % | 44 % |

L'échantillon socle représente le poids « naturel » de l'intersection dans chacune des sous-populations d'intérêt. Les deux méthodes conduisent à très largement surreprésenter les individus de l'intersection.

### *Émergence d'une 3e méthode : calcul de coefficients multiplicatifs par une méthode de calage*

Pour pallier les inconvénients précédents (dépendance de l'ordre des extensions, dépassement des cibles, surreprésentation exagérée des intersections), le choix s'est porté sur le calcul de coefficients multiplicatifs de tirage par une méthode de calage sur un tableau de données particulier.

L'idée de base de cette méthode est d'itérer les méthodes M1 et M1bis jusqu'à convergence. C'est la raison pour laquelle cette méthode de calage décrite ci-après est la mieux adaptée.

### *Les marges de calage*

Les marges de calage sont issues du tableau 12 (effectifs cibles par sous-populations d'extensions). Pour que l'algorithme de calage se déroule normalement, les niveaux généraux CGDD (c5), santé social (c19) et ensemble (c38) ont été retirés pour éviter les colinéarités entre les variables de calages. D'autre part, des contraintes facilement atteintes en pratique ; DGESIP licence (c16) et quartier prioritaire de la politique de la ville (c18) et une population tirée exhaustivement dans l'échantillon global (thèse [c36]) ont été également retirés. Il y a donc au final 32 marges de calage. La table SAS des marges pour l'application de la macro *Calmar* est la suivante :

**Tableau 16 • Les 32 marges de calage**

| Var | N | MAR1 | MAR2 |
|---|---|---|---|
| **c1** | 2 | 903 | 23 797 |
| **c2** | 2 | 1 068 | 23 632 |
| **c3** | 2 | 2 391 | 22 309 |
| **c4** | 2 | 1 373 | 23 327 |
| **c6** | 2 | 842 | 23 858 |
| **c7** | 2 | 817 | 23 883 |
| **c8** | 2 | 567 | 24 133 |
| **c9** | 2 | 748 | 23 952 |
| **c10** | 2 | 416 | 24 284 |
| **c11** | 2 | 7 026 | 17 674 |
| **c12** | 2 | 1 959 | 22 741 |
| **c13** | 2 | 972 | 23 728 |
| **c14** | 2 | 771 | 23 929 |
| **c15** | 2 | 350 | 24 350 |
| **c17** | 2 | 1 509 | 23 191 |
| **c20** | 2 | 1 180 | 23 520 |
| **c21** | 2 | 349 | 24 351 |
| **c22** | 2 | 454 | 24 246 |
| **c23** | 2 | 208 | 24 492 |
| **c24** | 2 | 324 | 24 376 |
| **c25** | 2 | 474 | 24 226 |
| **c26** | 2 | 62 | 24 638 |
| **c27** | 2 | 1 191 | 23 509 |
| **c28** | 2 | 427 | 24 273 |
| **c29** | 2 | 461 | 24 239 |
| **c30** | 2 | 78 | 24 622 |
| **c31** | 2 | 31 | 24 669 |
| **c32** | 2 | 140 | 24 560 |
| **c33** | 2 | 84 | 24 616 |
| **c34** | 2 | 306 | 24 394 |
| **c35** | 2 | 1 678 | 23 022 |
| **c37** | 2 | 112 | 24 588 |

Dans Mar1 sont enregistrées les cibles à atteindre pour chaque sous-population. Dans Mar2 se trouvent le complémentaire de Mar1 pour atteindre la cible des 24 700 observations.

### *Données auxquelles est appliquée la méthode de calage*

Le calage est appliqué sur la table obtenue en croisant les 32 indicatrices d'extension retenues. Ce croisement d'indicatrices d'extension crée une partition de la base de sondage. Le croisement des 32 indicatrices donne une table avec 94 observations, autrement dit, il s'agit d'une partition en 94 parties. Pour chacune de ces parties, le nombre d'enquêtes réalisées est estimé à partir de la première simulation (correspondant à l'échantillon socle de 8 500). Cette table est appelée « Partition ».

Le tableau suivant est un extrait de la table « Partition » sur laquelle est appliqué l'algorithme de calage. La colonne *Effectif Agrégé* donne, pour chaque partie de la partition, les effectifs anticipés de répondants dans le champ Céreq élargi[9]. De fait, cette variable *Effectif Agrégé* est calée sur les marges des totaux souhaités.

**Tableau 17 • Extrait de la table Partition**

| Obs | c1 | c2 | c3 | c4 | c6 | c7 | c8 | c9 | c10 | c11 | c12 | c13 | c14 | c15 | c17 | c20 | c21 | c22 | c23 | c24 | c25 | c26 | c27 | c28 | c29 | c30 | c31 | c32 | c33 | c34 | c35 | c37 | Effectif agrégé |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 0,38 |
| 2 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 1,95 |
| 3 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 38,1 |
| 4 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 363,41 |
| 5 | 2 | 1 | 2 | 2 | 2 | 2 | 2 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 0,68 |
| 6 | 2 | 1 | 2 | 2 | 2 | 2 | 2 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 15,39 |
| 7 | 2 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 0,05 |
| ... | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 94 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 90,38 |

* les indicatrices c5, c16, c18, c19, c36 et c38 ont été retirées pour la création de cette table.
Note de lecture : l'observation 6 de ce tableau correspond à la partie nº6 qui est composée des individus appartenant aux sous-populations c2 et c9 sans appartenir à aucune autre sous-population. À partir de l'échantillon socle simulé, le nombre de répondants dans le champ est estimé à 15,39 pour la partie 6.

À l'issue de l'étape de calage, pour chaque observation p de la table « Partition », le coefficient $coeff_p$ est calculé. Celui-ci correspond au rapport entre les valeurs finales et initiales de la variable *Effectif Agrégé*. Pour préciser les variables mises en jeu, le système sur lequel est appliqué l'algorithme est : pour chacune des 32 sous-populations j, (les 38 indicatrices de sous-populations moins les 6 retraits et l'ensemble) l'algorithme cherche à résoudre le problème suivant :

$$min_{somme\ calée_p} \sum_{p \in Partition} eff_p * G\left(\frac{eff\ calé_p}{eff_p}\right)$$

Sous la contrainte :

$$\sum_{p \in Partition} eff\ calé_p * \binom{1_{pj1}}{1_{pj2}} = t_{Xj} \text{ Où } t_{Xj} = \binom{Mar1_j}{Mar2_j}$$

- Le système de poids initial considéré est le nombre de répondants obtenus sur chacune des parties p, à savoir $effectif\ agrégé_p$ $(eff_p)$. Les poids finaux des observations *p* sont enregistrés dans la variable $eff\ calé_p$. C'est le nombre de répondants souhaité dans la partie p.

- L'indice p parcourt les données du tableau « Partition ». Ce dernier contient 94 observations, une pour chaque partie de la partition.

[9] En considérant le champ plus vaste « champ Céreq + champ spécifique (post-initiaux) des extensions sport et santé », la simulation conduit à 8 900 questionnaires et non à 8 500 comme anticipé avec le champ Céreq uniquement. Et c'est sur la base de ce nombre de questionnaires réalisés dans le cadre de ce champ élargi que sont faites les étapes suivantes d'ajustement des probabilités de tirage.

- Dans le vecteur $\begin{pmatrix} 1_{pj1} \\ 1_{pj2} \end{pmatrix}$,

- $1_{pj1}$ est l'indicatrice valant 1 si la case de la $p^{ième}$ ligne et de la $j^{ième}$ colonne de Partition vaut 1. Et 0 sinon.

- De même $1_{pj2}$ est l'indicatrice valant 1 si la case de la $p^{ième}$ ligne et de la $j^{ième}$ colonne de « Partition » vaut 2. Et 0 sinon.

- Le vecteur $t_{Xj}$ a pour composantes les marges définies par la $j^{ième}$ ligne du tableau 16 sur les marges de calage. La première composante étant Mar1, la deuxième Mar2.

- Le coefficient $coeff_p$ est alors le ratio $eff\ calé_p / eff_p$ .

- Avec G une fonction de distance. Dans le cas présent, il s'agit de la distance de la méthode linéaire dans la macro *Calmar*.

Après la réalisation du calage, les probabilités de tirage $\pi_{i2}$ sont de la forme :

$$\pi_{i2} = \pi_{i1} * coeff_p$$

L'hypothèse faite ici est qu'il existe une relation approximativement proportionnelle entre les probabilités de tirage et le nombre de répondants. L'approximation vient du fait que certaines probabilités dépassent 1 et que celles-ci sont ramenées à ce seuil maximal.

$$\pi_{i2} = \min\left(\pi_{i1} * coeff_p, 1\right)$$

Cette étape de calage ne fait pas tout, mais réalise tout de même l'essentiel du travail de détermination des coefficients de suppléments de tirage $coeff_p$. En effet, en plus du contrôle des valeurs des probabilités de tirage (inférieures ou égales à 1), les valeurs des probabilités de tirage $\pi_{i2}$ sont testées par rapport à leur valeur initiale $\pi_{i1}$. En effet, dans la démarche retenue, les probabilités de tirage $\pi_{i2}$ doivent être supérieures aux premières probabilités de tirage $\pi_{i1}$. Or, les contraintes spécifiques sur les cibles d'extensions ne prennent pas en compte les contraintes sur les probabilités de tirage $\pi_{i1}$. Il est donc nécessaire de corriger les probabilités d'inclusion $\pi_{i2}$ en forçant à ce qu'elles soient supérieures ou égales à $\pi_{i1}$.

Le calage permet d'aboutir au résultat souhaité. Cependant, les corrections effectuées en aval provoquent un écart entre le résultat obtenu et le résultat souhaité sans que cela soit dommageable quant aux objectifs de cet échantillon.

## 3.3. Phase B : Tirage équilibré de l'échantillon

### 3.3.1. Étape B1 : tirage de l'échantillon global (principal + réserve)

Trente contraintes d'équilibrage, toutes construites de la même manière, ont été utilisées pour contrôler les effectifs envoyés en production des sous-populations les plus sensibles (il s'agit essentiellement de contraintes d'équilibrage pour garantir les objectifs de questionnaires pour les partenaires d'extensions). Les sous-populations sur lesquelles des contraintes d'équilibrage ont porté sont les suivantes :

1. Baccalauréat professionnel
2. CAP
3. CGDD niveau 1
4. CGDD niveau 2
5. CGDD niveau 3
6. CGDD niveau 4
7. CGDD niveau 5
8. DGESIP BTS
9. DGESIP Grandes écoles
10. DGESIP IUT
11. DGESIP Licence
12. DGESIP Master
13. Sport
14. Thèse
15. Santé social conseiller économie familiale
16. Santé social éducateur jeunes
17. Santé social éducateur spécialisé
18. Santé social ergothérapeute
19. Santé social infirmier
20. Santé social masseur kinésithérapeute
21. Santé social moniteur éducateur
22. Santé social orthophoniste
23. Santé social orthoptiste
24. Santé social podologue pédicure
25. Santé social psychologue
26. Santé social sage femmes
27. Santé social aide-soignante
28. Santé social assistant social
29. Santé social auxiliaire puéricultrice

L'équilibrage est réalisé sur la variable probabilité de tirage pour chaque sous-population J. Les contraintes d'équilibrage sont ainsi données par :

$$\sum_{i\in S} \frac{1_{i\in J} * \pi_{i2}}{\pi_{i2}} = \sum_{i\in U} 1_{i\in J} * \pi_{i2}$$

Soit,

$$\sum_{i\in S} 1_{i\in J} = \sum_{i\in U} 1_{i\in J} * \pi_{i2}$$

Soit,

$$n_{j=} \sum_{i\in U} 1_{i\in J} * \pi_{i2}$$

Où : – $n_j$ est la taille de la sous-population *J* dans l'échantillon
– $1_{i\in J}$ est l'indicatrice d'appartenance à l'extension *J*

En équilibrant sur les probabilités de tirage correspondant à une extension particulière (délimitée par une indicatrice), le sous-échantillon est de taille fixe pour la sous-population *J*. Cela garantie un nombre d'individus à enquêter adapter au nombre de répondants attendu sur cette sous-population.

Ici est utilisée la loi des grands nombres : bien que les probabilités d'obtenir un questionnaire varient selon les individus, en moyenne le nombre de questionnaires espéré est assez bien connu. Les probabilités de tirage ont été déterminées pour définir une taille d'échantillon adaptée pour atteindre une cible donnée de

questionnaires pour la sous-population *J*. Ces probabilités d'inclusion sont respectées lors de la phase de vol du tirage équilibré[10]. En équilibrant sur les probabilités d'inclusion, la taille du sous-échantillon adéquate est déterminée de façon à atteindre les objectifs sur la sous-population *J*.

Un échantillon global, contenant la réserve et l'échantillon principal, a été tiré de manière équilibrée avec le package Sampling du logiciel *R*. L'échantillon global contient 171 000 observations.

### 3.3.2. Étape B2 : tirage de l'échantillon principal

Une contrainte d'équilibrage a été utilisée pour affecter les individus, soit à l'échantillon principal, soit à la réserve à partir de l'échantillon global. Le coefficient $coeffprin_i$ correspond à la probabilité que l'individu $i$ échantillonné se retrouve dans l'échantillon principal. Cette probabilité est de l'ordre de 0,9 pour les sous-populations disposant d'une réserve, et elle est égale à 1 si la sous-population est envoyée intégralement en production dans l'échantillon principal (c'est-à-dire que l'individu est automatiquement affecté à l'échantillon principal et en l'absence d'échantillon de réserve pour cette sous-population). Cette variation du coefficient $coeffprin_i$ est due aux fluctuations des nombres de questionnaires anticipés suite aux corrections manuelles nécessaires après la phase de calage.

La contrainte d'équilibrage s'écrit de la manière suivante :

$$\sum_{i \in S\ prin} \frac{coeffprin_i}{coeffprin_i} = \sum_{i \in S} coeffprin_i$$

Soit,

$$\sum_{i \in S\ prin} 1 = \sum_{i \in S} coeffprin_i$$

Soit,

$$n_{principal} = \sum_{i \in S} coeffprin_i$$

Où $S\ prin$ désigne l'échantillon principal et $S$ l'échantillon global.

Cette contrainte assure que le nombre d'observations comprises dans l'échantillon principal respecte globalement les probabilités d'affectation. Les probabilités d'inclusion finales dans l'échantillon principal sont donc :

$$\pi_{i3} = \pi_{i2} * coeffprin_i$$

Les individus non sélectionnés sont dans la réserve.

L'échantillon principal contient 159 432 observations et la réserve 11 654.

Pour l'enquête 2013 auprès de la Génération 2010, la réserve n'a pas été mobilisée. L'objectif est le même pour cette enquête. Si toutefois celle-ci devait être mobilisée, cette mesure serait appliquée : en cours de production, si les projections du nombre de répondants d'une sous-population montrent que la cible ne sera pas atteinte avant la fin du terrain d'enquêtes, alors toute la réserve de cette sous-population particulière sera débloquée.

[10] Le tirage de l'échantillon est quasiment parfaitement équilibré. À l'issue de la phase de vol, l'algorithme a statué sur 22 observations (sur 170 000 au regard de l'échantillon global). Le calcul d'un échantillon de 22 observations qui s'éloigne le moins possible des contraintes d'équilibrage est trop gourmand en temps de calcul. L'option choisie pour la phase d'atterrissage est donc de supprimer les contraintes (method = 2 dans la fonction *samplecube* du *package sampling*).

## 3.4. Bilan de l'échantillonnage

***Simulation du nombre de répondants à partir de l'échantillon principal***

Le tableau suivant montre les effectifs de répondants anticipés à partir de l'échantillon global $S$. Pour chaque sous-population J, le nombre de répondants attendu à partir de l'échantillon global est estimé par :

$$\sum_{i\epsilon\, S\cap J} p(i\ repond\ dans\ le\ champ\ ) = \mathrm{E}(nombre\ de\ répondants\ dans\ le\ champ\ pour\ J)$$

Il s'agit de l'espérance du nombre de questionnaires calculée sur les individus échantillonnés et faisant partie de la sous-population $J$. Le même calcul est utilisé pour les effectifs estimés pour l'échantillon principal en restreignant cette fois à $S\ prin\ \cap J$. Les estimations des effectifs répondants dans le champ à partir de l'échantillon principal sont fournies dans le tableau 19.

**Tableau 18 • Écart entre échantillon global et cible**

| Sous-population d'extension | Effectif répondant à partir de l'échantillon global | Cible (effectif extension + réserve) | Différence entre l'échantillon global et les cibles |
|---|---|---|---|
| COLLEGE | 895 | 903 | -8 |
| BACCALAURÉAT GÉNÉRAL ET TECHNOLOGIQUE | 1 059 | 1 068 | -9 |
| BACCALAURÉAT PROFESSIONNEL | 2 403 | 2 391 | 12 |
| CAP | 1 372 | 1 373 | -1 |
| CGDD | 3 372 | 3 349 | 23 |
| *CGDD niveau 1* | 831 | 842 | -11 |
| *CGDD niveau 2* | 826 | 817 | 9 |
| *CGDD niveau 3* | 564 | 567 | -3 |
| *CGDD niveau 4* | 744 | 748 | -4 |
| *CGDD niveau 5* | 407 | 416 | -9 |
| DGESIP | 7 350 | 7 026 | 324 |
| *DGESIP autre* | 2 044 | 1 959 | 85 |
| *DGESIP BTS* | 974 | 972 | 2 |
| *DGESIP Grandes écoles* | 802 | 771 | 31 |
| *DGESIP IUT* | 348 | 350 | -2 |
| *DGESIP Licence* | 1 666 | 1 464 | 202 |
| *DGESIP Master* | 1 516 | 1 509 | 7 |
| QUARTIER PRIORITAIRE DE LA POLITIQUE DE LA VILLE (QPV) | 2 145 | 2 287 | -142 |
| SANTÉ SOCIAL | 5 971 | 5 981 | -10 |
| *SANTÉ SOCIAL aide-soignante* | 1 171 | 1 180 | -9 |
| *SANTÉ SOCIAL assistant social* | 395 | 349 | 46 |
| *SANTÉ SOCIAL auxiliaire puéricultrice* | 510 | 454 | 56 |
| *SANTÉ SOCIAL conseiller économie familiale* | 207 | 208 | -1 |
| *SANTÉ SOCIAL éducateur jeunes* | 323 | 324 | -1 |
| *SANTÉ SOCIAL éducateur spécialisé* | 531 | 474 | 57 |
| *SANTÉ SOCIAL ergothérapeute* | 62 | 62 | 0 |
| *SANTÉ SOCIAL infirmier* | 1 190 | 1 191 | -1 |
| *SANTÉ SOCIAL masseur kinésithérapeute* | 425 | 427 | -2 |
| *SANTÉ SOCIAL moniteur éducateur* | 521 | 461 | 60 |
| *SANTÉ SOCIAL orthophoniste* | 77 | 78 | -1 |
| *SANTÉ SOCIAL orthoptiste* | 31 | 31 | 0 |
| *SANTÉ SOCIAL podologue pédicure* | 140 | 140 | 0 |
| *SANTÉ SOCIAL psychomotricien* | 84 | 84 | 0 |
| *SANTÉ SOCIAL sage-femmes* | 304 | 306 | -2 |
| SPORT | 2 605 | 1 678 | 927 |
| THÈSE | 2 486 | 2 446 | 40 |
| DOCTEUR EN SANTÉ | 110 | 112 | -2 |
| **ENSEMBLE** | **26 734** | **24 700** | **2 034** |

Les cibles visées sont respectées à l'issue du tirage de l'échantillon. Le nombre d'observations échantillonnées garantit systématiquement les cibles des conventions. En effet, les faibles écarts négatifs (un nombre de répondants inférieur aux effectifs cibles) ne constituent pas un risque, les cibles de l'échantillon principal ayant été définies avec une légère sécurité.

D'autre part, un léger écart existe sur la taille de la population totale suite aux corrections manuelles apportées ; notamment suite aux ajustements manuels des probabilités de tirage (effectif calé de 24 700 observations passe à 26 734).

Cet écart s'explique en bonne partie par la décision d'envoyer en production l'intégralité des diplômés DRJSCS (par expérience, il s'agit d'une population difficile à enquêter). Les fichiers collectés étant en outre de qualité moindre (absence de numéro de téléphone plus fréquente, moins d'informations identifiantes), la probabilité de réponse supposée de cette population parait inférieure à celle observée sur l'enquête précédente. Cette contrainte aurait pu être intégrée initialement dans les marges de calage, mais l'ajustement sur les effectifs de l'extension sport a été réalisé à l'issue du calage.

Le reste de la différence sur le total de la population se fait sur le complémentaire des indicatrices d'extension. Il s'agit ici en l'occurrence des mentions complémentaires, les formations à l'issue du CAP ou du baccalauréat professionnel.

Par ailleurs plusieurs cibles sont ici indicatives. Il s'agit des cibles pour les quartiers prioritaires de la politique de la ville (QPV) et celles sur les sous-populations des formations de la santé et du social. En effet, la cible de la convention pour les QPV était assez naturellement dépassée. Les cibles sur les sous-populations de la santé et du social se rapprochent des cibles que les partenaires auraient souhaitées idéalement. Cependant une étude de faisabilité a montré que certains effectifs de sous-populations ne pouvaient pas être atteints. La convention mentionne donc deux objectifs pour l'ensemble des formations de la santé d'une part et l'ensemble des formations du social d'autre part. Les populations qui ne pouvaient être atteintes ont été envoyées exhaustivement. L'échantillon devrait par ailleurs permettre d'atteindre les deux cibles générales.

Dans la dernière étape, les individus sont affectés à l'échantillon principal. Les estimations du nombre de répondants sont données dans le tableau suivant :

**Tableau 19 • Écart entre échantillon principal et cible**

| Sous-population d'extension | Effectif répondant à partir de l'échantillon principal | Cibles échantillon principal | Différence entre échantillon principal et les cibles |
|---|---|---|---|
| COLLÈGE | 895 | 903 | -8 |
| BACCALAURÉAT GÉNÉRAL ET TECHNOLOGIQUE | 1 059 | 1 068 | -9 |
| BACCALAURÉAT PROFESSIONNEL | 2 403 | 2 391 | 12 |
| CAP | 1 372 | 1 373 | -1 |
| CGDD | 3 163 | 3 080 | 83 |
| *CGDD niveau 1* | 753 | 765 | -12 |
| *CGDD niveau 2* | 750 | 742 | 8 |
| *CGDD niveau 3* | 509 | 515 | -6 |
| *CGDD niveau 4* | 744 | 680 | 64 |
| *CGDD niveau 5* | 407 | 378 | 29 |
| DGESIP | 6 665 | 6 387 | 278 |
| *DGESIP autre* | 1 840 | 1 781 | 59 |
| *DGESIP BTS* | 887 | 884 | 3 |
| *DGESIP Grandes écoles* | 730 | 701 | 29 |
| *DGESIP IUT* | 318 | 318 | -1 |
| *DGESIP Licence* | 1 518 | 1 331 | 186 |
| *DGESIP Master* | 1 373 | 1 372 | 1 |
| QUARTIER PRIORITAIRE DE LA POLITIQUE DE LA VILLE (QPV) | 1 980 | 2 079 | -99 |
| SANTÉ SOCIAL | 5 595 | 5 438 | 158 |
| *SANTÉ SOCIAL aide-soignante* | 1 071 | 1 073 | -1 |
| *SANTÉ SOCIAL assistant social* | 359 | 318 | 41 |
| *SANTÉ SOCIAL auxiliaire puéricultrice* | 461 | 413 | 48 |
| *SANTÉ SOCIAL conseiller économie familiale* | 207 | 208 | -1 |
| *SANTÉ SOCIAL éducateur jeunes* | 323 | 324 | -1 |
| *SANTÉ SOCIAL éducateur spécialisé* | 488 | 431 | 57 |
| *SANTÉ SOCIAL ergothérapeute* | 62 | 62 | 0 |
| *SANTÉ SOCIAL infirmier* | 1 085 | 1 083 | 2 |
| *SANTÉ SOCIAL masseur kinésithérapeute* | 425 | 427 | -2 |
| *SANTÉ SOCIAL moniteur éducateur* | 477 | 419 | 58 |
| *SANTÉ SOCIAL orthophoniste* | 77 | 78 | 0 |
| *SANTÉ SOCIAL orthoptiste* | 31 | 31 | 0 |
| *SANTÉ SOCIAL podologue pédicure* | 140 | 140 | 0 |
| *SANTÉ SOCIAL psychomotricien* | 84 | 84 | 0 |
| *SANTÉ SOCIAL sage-femmes* | 304 | 306 | -2 |
| SPORT | 2 360 | 1 525 | 835 |
| THÈSE | 1 863 | 1 500 | 363 |
| DOCTEUR EN SANTÉ | 100 | 100 | 0 |
| **ENSEMBLE** | **24 570** | **23 000** | **1 570** |

La différence sur les docteurs hors santé s'explique par le fait que la convention mentionne une cible de diplômés et non de sortants. Une variable indiquant l'obtention du diplôme existe dans la base de sondage pour certaines formations, sans certitude quant à la qualité de cette variable. De ce fait, le nombre d'observations tirées, qui concernent ici tous les sortants, est légèrement supérieur à la cible sachant que l'estimation de la part des diplômés se situe entre 70 % et 80 %.

## 3.5. Phase d'apurement

Lors de la création du fichier d'import, un apurement de l'échantillon a dû être réalisé afin de réaliser une détection plus fine des hors-champ. Ainsi, 1 031 individus ont été supprimés, réduisant l'échantillon à 158 401 individus.

# 4. Préparation de la collecte

## 4.1. Développement informatisé du questionnaire

### 4.1.1. Technique de développement du calendrier

L'enquête 2016 auprès de la Génération 2013 est une enquête réalisée par téléphone uniquement. La technique de recueil de l'information utilisée est le système CATI (*computer assited telephone interviewing*). Cet outil permet d'assister l'enquêteur grâce à l'utilisation de l'informatique.

Le développement informatisé du questionnaire a nécessité l'utilisation de deux logiciels. Un premier, *Quancept,* pour la gestion des appels téléphoniques intégrant le module de questions sur l'identification du bon individu. Le second, *Confirmit*, utilisé est la version web du logiciel pour le développement du questionnaire dans sa majorité. Enfin, du fait de la spécificité de l'enquête Génération autour de deux calendriers interactifs, la programmation spécifique d'une application intégrée a été développée en *Javascript*. L'intégration de cette application au reste du questionnaire a été transparente pour l'enquêteur en phase de collecte.

La complexité de ce développement n'a pas eu d'impact direct sur la passation de l'enquête. En revanche, le manque de fluidité dans le passage de la phase d'identification au questionnaire a eu un effet négatif sur la durée du questionnaire et le travail des enquêteurs notamment par l'intégration de manipulations multiples.

### 4.1.2. Test du CATI et calendrier associé

Trois tests en réel du CATI ont été programmés au lieu de 2 habituellement.

#### *Test 1*

Pour la première fois, le marché de l'enquête de cheminement à tous niveaux Génération 2013 a prévu la réalisation du premier test du questionnaire, pour alimenter le label, avec le prestataire. En effet, ce test s'effectuait auparavant en interne au Céreq. Les délais ont été redéfinis tout en respectant les contraintes légales liées au marché public.

L'opération s'est déroulée du 02 au 05 décembre 2015 (soit 4 jours de tests). 7 enquêteurs, 2 chefs d'équipe et 1 écouteur ont été formés pour la passation de l'enquête téléphonique (ET). Les plages horaires choisies sont 12-17 heures et 15-21 heures afin de pouvoir interroger tout type de jeunes actifs ou inactifs.

Plusieurs objectifs pour ce test :

- réaliser une estimation du temps d'enquête par module ;
- repérer les questions à reformuler et/ou à supprimer notamment pour les nouveaux modules de questionnement ;
- repérer les problèmes de filtres.

Un échantillon principal a été tiré de la base de sondage constituée des sortants du système éducatif au cours de l'année scolaire 2012-2013. Le tirage a tenu compte des strates agrégées de formation. La taille de l'échantillon est de 2 624 individus de tous niveaux.

Un échantillon de réserve constitué de 800 individus a aussi été tiré pour assurer la fin du plateau d'enquête. En effet, le nombre de numéros de téléphone par individu était de 2 au maximum alors qu'en enquête réelle il varie de 4 à 8 numéros disponibles.

L'échantillon principal a été brassé lors des 2 premiers jours de tests avec 2 appels sur chaque numéro. L'échantillon de réserve a ainsi été injecté à partir du 3e jour.

**Tableau 20 • Bilan du premier test**

| STATUT DU DERNIER APPEL | ÉCHANTILLON PRINCIPAL | | ÉCHANTILLON DE RÉSERVE | | TOTAL | |
|---|---|---|---|---|---|---|
| | Eff. | % | Eff. | % | Eff. | % |
| Interviews réalisées | 94 | 5,6 (1) | 20 | 4,3 (1) | 114 | 5,3 (1) |
| Répondeur – Individu identifié sur nom prénom | 192 | 11,4 (1) | 41 | 8,9 (1) | 233 | 10,9 (1) |
| Répondeur – Individu non identifié sur nom prénom | 875 | 52,0 (1) | 245 | 53,3 (1) | 1 120 | 52,3 (1) |
| NRP/OCCUPE | 146 | 8,7 (1) | 48 | 10,4 (1) | 194 | 9,1 (1) |
| RDV | 10 | 0,6 (1) | 2 | 0,4 (1) | 12 | 0,6 (1) |
| FAUX NUMÉRO | 322 | 15,6 (2) | 129 | 20,8 (2) | 451 | 16,8 (2) |
| HORS CHAMP | 285 | 16,9 (1) | 83 | 18,0 (1) | 368 | 17,2 (1) |
| REFUS | 78 | 4,6 (1) | 20 | 4,3 (1) | 98 | 4,6 (1) |
| Abandon en cours de questionnaire | 3 | 0,2 (1) | 1 | 0,2 (1) | 4 | 0,2 (1) |
| Non appelés | 62 | 3,0 (2) | 30 | 4,8 (2) | 92 | 3,4 (2) |
| **Total fichier d'appel** | **2 067** | **-** | **619** | **-** | **2 686** | **-** |
| **Total fichier d'appel sans les faux numéros et strates vides** | **1 683** | **100** | **460** | **100** | **2 143** | **100** |

Lecture du tableau : (1) : Pourcentage calculé sur le total du fichier d'appel sans les faux numéros et les strates vides ; (2) : Pourcentage calculé sur le total du fichier d'appel

En fin de terrain la durée moyenne enregistrée pour les 114 interviews réalisées est de 33 minutes sans la phase de contact. La cible de ce questionnaire était de 22 minutes en moyenne.

Plusieurs raisons peuvent expliquer la longueur du questionnaire :

Le questionnement était assez long avec certains modules de partenaires qui ont été intégrés peu de temps avant le début de ce test. Le module « Thèse » sera notamment raccourci ainsi que la partie sur les caractéristiques individuelles et l'environnement familial de l'individu.

Des problèmes techniques et une ergonomie à faire évoluer comme par exemple : menus déroulants non adaptés à une recherche en autocomplétion à venir, normalisation des adresses à réaliser, le calendrier professionnel qui doit être optimisé…

Une consigne avait été donnée aux enquêteurs de relever d'éventuelles anomalies dans le questionnaire (défauts d'affichage, erreurs ou incohérences de filtres) et de les noter pour un traitement post-enquête ou en direct. Cette consigne donnée sur les deux premiers jours a ralenti le temps de passation.

### Test 2

Un échantillon a été tiré de la base de sondage constituée des sortants du système éducatif au cours de l'année scolaire 2012-2013. Le tirage a tenu compte des strates agrégées de formation. La taille de l'échantillon est de 3 875 individus de tous niveaux.

**Tableau 21 • Bilan du second test**

| STATUT DU DERNIER APPEL | Échantillon principal | |
|---|---|---|
| | Eff. | % |
| Interviews réalisées | 202 | 9,2 (1) |
| NRP/OCCUPE | 930 | 42,2 (1) |
| RDV | 55 | 2,5 (1) |
| FAUX NUMÉRO | 1 215 | 31 (2) |
| HORS CHAMP | 442 | 20,1 (1) |
| REFUS | 119 | 5,4 (1) |
| Abandon en cours de questionnaire | 455 | 20,6 (1) |
| Non appelés | 457 | 13 (2) |
| **Total fichier d'appel** | **3 875** | **-** |
| **Total fichier d'appel sans les faux numéros et les non appelés** | **2 203** | **100** |

Lecture du tableau : (1) : Pourcentage calculé sur le total du fichier d'appel sans les non appelés ; (2) : Pourcentage calculé sur le total du fichier d'appel

202 questionnaires ont été réalisés. En fin de terrain, la durée moyenne du questionnaire était de 30 minutes sans la phase de contact.

### Test 3

Un échantillon principal a été tiré de la base de sondage constituée des sortants du système éducatif au cours de l'année scolaire 2012-2013. Le tirage a tenu compte des strates agrégées de formation. La taille de l'échantillon est de 3 387 individus de tous niveaux.

**Tableau 22 • Bilan du troisième test**

| STATUT DU DERNIER APPEL | Échantillon principal | |
|---|---|---|
| | Eff. | % |
| Interviews réalisées | 161 | 7,7 (1) |
| Répondeur – Individu identifié sur nom prénom | 304 | 14,5 (1) |
| Répondeur – Individu non identifié sur nom prénom | 881 | 42,1 (1) |
| NRP/OCCUPE | 313 | 14,9 (1) |
| RDV | 22 | 1 (1) |
| FAUX NUMÉRO | 357 | 10,5 (2) |
| HORS CHAMP | 347 | 16,5 (1) |
| REFUS | 56 | 2,7 (1) |
| Abandon en cours de questionnaire | 13 | 0,6 (1) |
| Non appelés | 933 | 27,5 (2) |
| **Total fichier d'appel** | **3 387** | **-** |
| **Total fichier d'appel sans les faux numéros et les non appelés** | **2 097** | **100** |

Lecture du tableau : (1) : Pourcentage calculé sur le total du fichier d'appel sans les non appelés ; (2) : Pourcentage calculé sur le total du fichier d'appel

161 questionnaires ont été réalisés. La durée moyenne du questionnaire était de 32 minutes sans la phase de contact.

## 4.2. Restructuration, normalisation, validation postale des adresses et rachat des déménagés

Ici est présenté le détail de la procédure appliquée à l'échantillon principal envoyé en production. L'échantillon de réserve a également été traité dans les mêmes conditions. N'ayant pas été utilisés au moment de l'enquête, les résultats présentés n'intègrent pas les données de l'échantillon de réserve.

### 4.2.1. Restructuration Normalisation Validation Postale (RNVP)

Suite au tirage de l'échantillon, une phase de validation ou recherche des coordonnées postales, téléphoniques et mails est réalisée pour chacun des individus. Une première étape consiste en une mise à jour des adresses postales avec un double objectif ; favoriser l'enrichissement des coordonnées téléphoniques basé sur l'adresse complète de l'individu et faciliter la transmission de la lettre avis dont le but est de contacter le plus grand nombre.

Cette phase de validation des adresses postales a été effectuée par un prestataire externe à partir des fichiers d'adresses de La Poste. Le fichier transmis regroupe les individus de l'échantillon avec l'ensemble des informations disponibles dans la base de sondage.

Les recherches ont donné les résultats suivants :

**Tableau 23 • Résultats du traitement RNVP**

| Classement des adresses | Échantillon principal | % |
|---|---|---|
| Adresses validées | 122 497 | 77,33 |
| Adresses litigieuses | 6 510 | 4,11 |
| Adresses rejetées | 29 394 | 18,56 |
| **Total** | **158 401** | **100** |

Un classement des adresses permet de les définir en tant que :

- *Validées* : adresses confirmées ou corrigées par le prestataire. Ces dernières sont mises à jour directement dans l'échantillon.

- *Litigieuses* : adresses corrigées (avec réserve) et non validées. Ces adresses ne sont pas prises en considération.

- *Rejetées* : adresses non corrigées, car présentant des incohérences ; voies inconnues dans la localité, déménagées dont l'adresse n'est pas commercialisable (fichier CNIL), etc.

S'agissant d'une simple normalisation des adresses postales et non d'une recherche de nouvelles adresses par individu, la lettre avis est transmise aux seules adresses validées par le prestataire.

### 4.2.2. Recherche des adresses des individus ayant déménagé

Les adresses postales présentes dans la base de sondage datent de 2013. Une actualisation est nécessaire pour une partie des individus ayant déménagé dans la période 2013-2016.

Parmi les 158 401 adresses envoyées en traitement, 15 366 ont un statut de déménagé (soit environ 7 %) dont 9 978 exploitables.

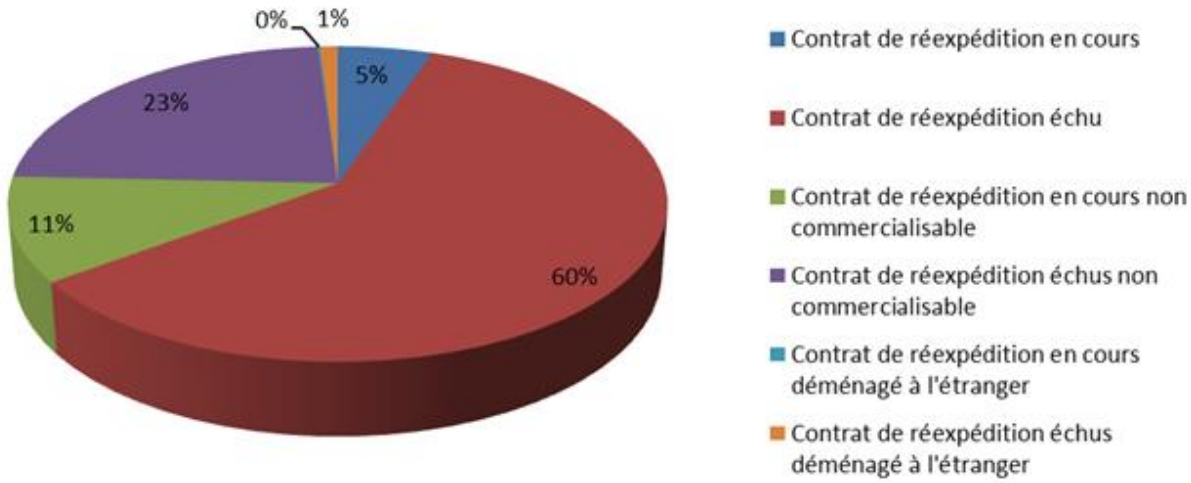
**Tableau 24 • Résultats des recherches des individus ayant déménagé**

| Types d'adresse | Effectif | % |
|---|---|---|
| **Total commercialisable (ayant pu être achetées)** | **9 978** | **6,30** |
| - Contrat de réexpédition en cours | 809 | 0,51 |
| - Contrat de réexpédition échu | 9 169 | 5,79 |
| **Total non commercialisable (n'ayant pas pu être achetées)** | **5 388** | **3,41** |
| - Contrat de réexpédition en cours non commercialisable | 1 642 | 1,04 |
| - Contrat de réexpédition échu non commercialisable | 3 574 | 2,26 |
| - Contrat de réexpédition en cours déménagé à l'étranger | 19 | 0,01 |
| - Contrat de réexpédition échu déménagé à l'étranger | 153 | 0,10 |

La recherche, réalisée à partir des fichiers de La Poste, distingue les adresses commercialisables des non commercialisables (à la demande du ménage) et avec une précision sur le statut du contrat de réexpédition (en cours ou échu). Les individus ayant déménagé dans un pays étranger sont également non commercialisables.

En conséquence, pour les 5 388 individus de l'échantillon, la seule information disponible est « individu ayant déménagé » sans récupération d'une nouvelle adresse. Toutefois, l'adresse postale initiale est supprimée et aucune lettre avis papier n'est transmise.

**Figure 6 • Répartition des individus ayant déménagé**




Quelques précisions sur le contrat de réexpédition qui permet le transfert du courrier vers une autre adresse en France :

- *En cours* : nouvelle adresse de déménagé dont le contrat de réexpédition souscrit auprès de La Poste est en cours (pendant 6 mois, renouvelable 1 fois).
- *Échu* : nouvelle adresse de déménagé dont le contrat de réexpédition souscrit auprès de La Poste est arrivé à son terme.
- *Non commercialisable* : ancienne adresse de déménagé avec mention de non-autorisation de commercialiser la nouvelle adresse.
- *Déménagé à l'étranger* : ancienne adresse d'individu ayant déménagé à l'étranger.

## 4.3. Enrichissement des coordonnées téléphoniques

### 4.3.1. Le protocole de recherche

La validation des numéros de téléphone disponibles dans l'échantillon est une étape primordiale pour la réussite de l'enquête. En parallèle, une recherche de numéros supplémentaires permet de favoriser un taux de réponse maximum. Suite à la mise à jour des adresses postales, un protocole spécifique est mis en place pour valider et enrichir les numéros disponibles.

72 % des individus de l'échantillon disposent d'au moins un numéro de téléphone et peuvent avoir déclaré jusqu'à trois numéros à leur établissement de formation. L'enrichissement des coordonnées téléphoniques permet d'une part de collecter des numéros de téléphone pour les individus n'en disposant pas ou ayant déménagé et d'autre part de mettre à jour les coordonnées téléphoniques déjà présentes dans la base.

La recherche peut aboutir à la collecte d'un ou plusieurs numéros supplémentaires qui peuvent être liés soit à l'individu, soit à un membre de sa famille.

Le prestataire choisi pour réaliser cette phase de recherche dispose de deux sources d'informations ; l'annuaire de France Télécom et une base de données contenant des numéros de téléphones mobiles et des numéros spéciaux de fournisseurs d'accès internet (nommée base partenaires).

Un protocole de recherche par étapes successives est pratiqué par le prestataire pour récupérer un maximum de numéros de téléphone avec un taux de fiabilité acceptable.

- *Phase A* : recherche, dans l'annuaire France Télécom, des individus pour lesquels les critères *nom, prénom et adresse* complète aboutissent à un seul écho. Pour pallier les éventuelles imprécisions des données individuelles, une entorse aux critères est faite en intégrant de légers écarts d'orthographe sur le nom, le prénom ou l'adresse, voire en l'absence du prénom ou de l'adresse.

- *Phase D* : recherche, dans la base partenaires, des individus avec les mêmes critères de sélection que la phase A.

- *Phase B* : recherche dite « élargie », dans les deux sources de données, des individus sur les critères suivants :

  - Nom, prénom et commune.
  - Nom, prénom et département.
  - Nom, prénom et région.
  - Nom, prénom et Île-de-France.
  - Nom, prénom et reste de la France.

  Le critère de la localisation géographique est levé pour augmenter le nombre d'échos et par conséquent les possibilités de recueil de nouveaux numéros. Seuls les individus exclus de la phase A et D sont considérés.

- *Phase C* : recherche dite « élargie », dans les deux sources de données, des individus sur les critères suivants :

  - Nom et commune.
  - Nom et département.

  Le critère du prénom est levé. Parmi les individus sélectionnés, la présence de membres de la famille ou d'homonyme est possible. Comme pour la phase B, seuls les individus exclus de la phase A et D sont considérés.

Pour les recherches A et D, le prestataire fournit au Céreq un seul écho unique. Le procédé de recherche se fait les critères du nom, prénom et adresse complète. Le rapprochement calcule un score et une classe de déduplication. Des méthodes sont appliquées pour prendre en charge l'orthographe approchée, la phonétique, etc.

Seules certaines classes de déduplication sont suffisamment fiables pour valider le rapprochement. Parmi celles-ci, le choix de la classe est celle qui présente le meilleur score ; idéalement nom, prénom et adresse identique. Par conséquent, si plusieurs échos possibles, le choix est porté sur celui qui concorde le mieux avec les données en entrée. Par exemple, pour un couple marié à la même adresse, seul le prénom diffère, le prénom le plus proche de celui en entrée présente un meilleur score.

L'ordonnancement de la recherche par étapes successives appliqué à l'échantillon défini par le Céreq résulte d'une priorisation quant à la qualité des échos reçus.

**Tableau 25 • Synthèse des recherches**

| Ordre | Phase | Définition des critères | Annuaire France Télécom | Base partenaires |
|---|---|---|---|---|
| 1 | A | Nom, prénom et adresse | ✓ | |
| 2 | D | Nom, prénom et adresse | | ✓ |
| 3 | C + B | Nom et commune + nom, prénom et département | ✓ | |
| 4 | C | Nom et commune | ✓ | |
| 5 | B | Nom, prénom et département | ✓ | |
| 6 | C + B | Nom et commune + nom, prénom et département | ✓ | ✓ |
| 7 | C | Nom et commune | | ✓ |
| 8 | B | Nom, prénom et département | | ✓ |
| 9 | C | Nom et département | ✓ | |
| 10 | C | Nom et département | | ✓ |
| 11 | B | Nom, prénom et région | ✓ | |
| 12 | B | Nom, prénom et région | | ✓ |
| 13 | B | Nom, prénom et Île-de-France | ✓ | |
| 14 | B | Nom, prénom et reste de la France | | ✓ |

La recherche d'un individu s'appuie sur 14 requêtes susceptibles de répondre à l'objectif d'une collecte de numéros de téléphone de qualité. La première requête fournit un numéro de meilleure qualité que la dernière. Chaque requête peut retourner une multiplicité d'échos. Un seuil de qualité du numéro de téléphone a été établi à cinq échos maximum pour les requêtes de 1 à 12 et d'un unique écho pour les deux dernières.

### 4.3.2. Bilan de l'enrichissement

Le prestataire réalise les recherches en deux temps. Sur la base du premier fichier reçu avec les résultats des phases A et D, le Céreq détermine les individus faisant l'objet de recherches approfondies suivantes (B puis C).

#### *Recherche en phase A :*

Parmi les 158 401 individus envoyés en phase de recherche des coordonnées téléphoniques, 42 676 ont été trouvés suite à la recherche selon les critères de la phase A (26,9 %). Pour chaque individu, un écho unique a été retourné à partir de la recherche dans l'annuaire France Télécom sur les critères *nom* (éventuellement *prénom*) et *adresse.*

**Tableau 26 • Bilan de la phase A (annuaire France Télécom)**

| Classe | Définition des critères | Écho possible | Effectif | % de l'échantillon |
|---|---|---|---|---|
| U11 | Nom, prénom et adresse identiques | 1 | 5 921 | 3,74 |
| U12 | Nom et prénom identiques, adresse approchée | 1 | 311 | 0,20 |
| U21 | Nom ou prénom approché, adresse identique | 1 | 1 144 | 0,72 |
| U22 | Nom ou prénom approché, adresse approchée | 1 | 64 | 0,04 |
| U13 | Nom et prénom identiques, adresse absente | 1 | 74 | 0,05 |
| O111 | Nom identique, adresse identique, prénom différent | 1 | 22 016 | 13,90 |
| O131 | Nom identique, adresse absente | 1 | 518 | 0,33 |
| O121 | Nom identique, adresse approchée | 1 | 1 985 | 1,25 |
| O11* | Nom identique, adresse identique, prénom différent ou approché | n | 8 416 | 5,31 |
| O12* | Nom identique, adresse approchée | n | 861 | 0,54 |
| O13* | Nom identique, adresse absente | n | 36 | 0,02 |
| O211 | Nom approché, adresse identique | 1 | 306 | 0,19 |
| O101 | Nom identique, prénom identique, adresse différente | 1 | 1 024 | 0,65 |
| **TOTAL** | | | **42 676** | **26,9** |

*pour les individus présents dans les classes O11, O12 et O13, la recherche a abouti à plusieurs échos possibles. Après une sélection manuelle par le prestataire, un seul écho est livré par individu.

#### *Recherche en phase D :*

Parmi les 158 401 individus envoyés en phase de recherche des coordonnées téléphoniques, 44 572 ont été trouvés à la suite d'une recherche selon les critères de la phase D (28,1 %). Pour chaque individu, un écho unique a été retourné à partir de la recherche dans la base partenaire sur les critères *nom* (éventuellement *prénom*) et *adresse.* Ces individus sont à additionner avec ceux de la phase A.

**Tableau 27 • Bilan de la phase D (base partenaires)**

| Classe | Définition des critères | Écho possible | Effectif | % de l'échantillon |
|---|---|---|---|---|
| U11 | Nom, prénom et adresse identiques | | 15 344 | 9,69 |
| U12 | Nom et prénom identiques, adresse approchée | 1 | 698 | 0,44 |
| U21 | Nom ou prénom approché, adresse identique | 1 | 2 031 | 1,28 |
| U22 | Nom ou prénom approché, adresse approchée | 1 | 105 | 0,07 |
| U13 | Nom et prénom identiques, adresse absente | 1 | 263 | 0,17 |
| O111 | Nom identique, adresse identique, prénom différent | 1 | 16 731 | 10,56 |
| O131 | Nom identique, adresse absente | 1 | 379 | 0,24 |
| O121 | Nom identique, adresse approchée | 1 | 1 097 | 0,69 |
| O11* | Nom identique, adresse identique, prénom différent ou approché | n | 4 162 | 2,63 |
| O12* | Nom identique, adresse approchée | n | 434 | 0,27 |
| O13* | Nom identique, adresse absente | n | 17 | 0,01 |
| O211 | Nom approché, adresse identique | 1 | 365 | 0,23 |
| O101 | Nom identique, prénom identique, adresse différente | 1 | 2 946 | 1,856 |
| **TOTAL** | | | **44 572** | **28,1** |

*pour les individus présents dans les classes O11, O12 et O13, la recherche a abouti à plusieurs échos possibles. Après une sélection manuelle par le prestataire, un seul écho est livré par individu.

### *Recherche en phase B et C :*

En fonction de la source de données utilisée, les résultats de la recherche élargie des phases B et C sont différents. Un individu a potentiellement plusieurs échos possibles et peut se retrouver à la fois dans les échos de phase B et C et également parmi les échos des requêtes d'une même phase. Toutefois, quelle que soit la source de données, la recherche selon les critères de la phase C a fourni plus d'échos potentiels que celle de la phase B.

**Tableau 28 • Bilan des phases B et C (base France Télécom)**

| Phase | Définition des critères | Effectif |
|---|---|---|
| B | Nom, prénom et commune | 70 |
| B | Nom, prénom et département | 1 646 |
| B | Nom, prénom et reste de la France | 2 015 |
| B | Nom, prénom et Île-de-France | 917 |
| B | Nom, prénom et région | 583 |
| C | Nom et commune | 25 282 |
| C | Nom et département | 28 179 |
| **Total** | | **58 692** |

**Tableau 29 • Bilan des phases B et C (base partenaires)**

| Phase | Définition des critères | Effectif |
|---|---|---|
| B | Nom, prénom et commune | 4 856 |
| B | Nom, prénom et département | 5 211 |
| B | Nom, prénom et reste de la France | - |
| B | Nom, prénom et Île-de-France | - |
| B | Nom, prénom et région | - |
| C | Nom et commune | 19 450 |
| C | Nom et département | 16 260 |
| **Total** | | **45 777** |

### 4.3.3. Mise à jour et classement des coordonnées téléphoniques

La phase de validation et de recherche de nouveaux numéros de téléphone terminée, une mise à jour de la base est réalisée.

Il s'agit maintenant de les ordonner par ordre de fiabilité en fonction des critères de qualité pour augmenter la probabilité de contact de chaque individu. La logique de classement des numéros est de prioriser le numéro de téléphone portable et a été déterminée tel que :

1. Sélection du numéro de téléphone portable issu de la base de sondage.
2. Sinon sélection du numéro de téléphone portable issu de la phase D.
3. Sinon sélection du numéro de téléphone fixe issu de la phase A.
4. Sinon sélection du numéro téléphone fixe issu de la base de sondage.
5. Sinon sélection du numéro de téléphone (fixe ou portable) issu de la recherche élargie (B, C).

Chaque individu dispose alors d'un ou plusieurs numéros de téléphone avec un nombre maximum possible de 8. Au total, pour 10 % de l'échantillon, aucun numéro de téléphone n'a été trouvé.

**Tableau 30 • Nombre de numéros de téléphone disponibles dans la base**

| NBR TEL INITIAL | Fréquence | Pourcentage | Fréquence cumulée | Pourcentage cumulé |
|---|---|---|---|---|
| 0 | 16 638 | 10,50 | 16 638 | 10,50 |
| 1 | 43 552 | 27,49 | 60 190 | 38,00 |
| 2 | 56 248 | 35,51 | 116 438 | 73,51 |
| 3 | 26 648 | 16,82 | 143 086 | 90,33 |
| 4 | 8 148 | 5,14 | 151 234 | 95,48 |
| 5 | 4 126 | 2,60 | 155 360 | 98,08 |
| 6 | 2 276 | 1,44 | 157 636 | 99,52 |
| 7 | 728 | 0,46 | 158 364 | 99,98 |
| 8 | 37 | 0,02 | 158 401 | 100,00 |

Au final, sur les 158 401 individus envoyés en production, 141 763 disposent d'au moins un numéro de téléphone (soit 89,5 %). Au total 308 262 numéros de téléphone sont disponibles dans la base :

**Tableau 31 • Source des numéros de téléphone**

| Origine | Fréquence | Pourcentage | Fréquence Cumulée | Pourcentage Cumulé |
|---|---|---|---|---|
| Base de sondage | 150 892 | 48,95 | 150 892 | 48,95 |
| PHASE A | 22 267 | 7,22 | 173 159 | 56,17 |
| PHASE D | 35 322 | 11,46 | 208 481 | 67,63 |
| PHASE B ou C | 99 781 | 32,37 | 308 262 | 100,00 |

## 4.4. Lettre et mail avis

Avant chaque enquête, l'ensemble des individus de l'échantillon est contacté par voie postale et/ou numérique une semaine avant le démarrage de la production. La transmission de la lettre avis est une recommandation du comité du label.

L'objectif principal est d'informer les individus qu'ils seront contactés par téléphone dans les semaines suivantes afin de participer à une enquête sur l'insertion professionnelle. Au-delà du rôle informatif sur les objectifs, le contenu et la période d'enquête, une mise à jour des coordonnées téléphoniques est proposée *via* un site web, mis en place par le prestataire de collecte, ou à partir d'un numéro vert. La prise de rendez-vous pour répondre à l'enquête est également possible par le biais de ces outils de contact.

Le cœur du questionnaire est un calendrier d'activité qui décrit mois par mois la situation professionnelle de l'individu sur les trois dernières années. Pour faciliter son remplissage lors de l'entretien téléphonique, un modèle de calendrier accompagne la lettre d'informations. Chacun a la possibilité de le préparer en amont de l'enquête.

Pour une question de coûts et de rapidité, le mail a été privilégié pour eux pour l'envoi des lettres avis par rapport à l'envoi de courrier.

Pour tous les individus disposant d'un mail, la lettre avis a donc été envoyée par mail. Pour les individus restants et dont les adresses paraissaient suffisamment fiables, une lettre-avis a été envoyée par courrier postal dès lors qu'ils n'appartenaient pas à une strate de formation à fort taux d'hors-champ.

Cependant, afin d'augmenter les chances de contacts, certains individus ayant un mail ont également reçu un courrier. Les individus de l'extension « Sport » ayant un mail et une adresse fiable ont reçu en plus du mail un courrier papier. Ensuite, parmi les individus des autres extensions ayant un mail et une adresse postale fiable, nous en avons aléatoirement choisi la moitié afin de leur envoyer un courrier papier en plus du mail.

**Tableau 32 • Répartition des lettres avis envoyées**

| Nature de la lettre avis | Effectif | % |
|---|---|---|
| Papier | 56 900 | 35,9 |
| Mail | 52 587 | 33,2 |
| Papier et mail | 13 092 | 8,3 |
| Aucun envoi | 35 822 | 22,6 |
| **Total** | **158 401** | **100,0** |

Pour la première fois, la prestation *Alliage* a été utilisée. Cette dernière consiste à récupérer l'information des plis non distribués grâce à un code apposé sur le courrier. En effet, lorsque les plis ne peuvent être distribués, le facteur le scanne, ce qui permet de stocker les informations des individus dans une base de données. Cette dernière nous est fournie environ un mois après.

Au total, 9 664 courriers (soit 13,8 %) n'ont pas pu être distribués. Le tableau ci-dessous prend en compte cette information.

**Tableau 33 • Répartition du nombre de lettres avis envoyées selon le mode avec prise en compte des plis non distribués**

| Nature de la lettre avis | Effectif | % |
|---|---|---|
| Papier | 49 356 | 31,2 |
| Mail | 54 707 | 34,5 |
| Papier et mail | 10 972 | 6,9 |
| Aucun envoi | 43 366 | 27,4 |
| **Total** | **158 401** | **100,0** |

## 4.5. Site internet

Un site internet dédié aux personnes échantillonnées a été mis en place pour informer sur l'enquête en cours. Accessible à partir de la page d'accueil du site du Céreq, il permettait aux jeunes d'avoir des informations sur l'enquête ainsi que de consulter les résultats des enquêtes précédentes, mais surtout d'accéder à un espace internet (appelé *CAWI de préfidélisation ou d'enrichissement*) afin d'y enrichir ses coordonnées téléphoniques et/ou de prendre rendez-vous.

De plus, pour la première fois, les individus avaient la possibilité de répondre au questionnaire filtre sur le site internet.

## 4.6. Hotline

Un numéro vert a été mis en place pour permettre aux jeunes d'avoir des réponses à des questions qu'ils se poseraient sur l'enquête, de donner de nouvelles coordonnées, de réaliser l'enquête ou de faire part de leur refus d'y répondre. Le numéro était présent sur la lettre-avis. Il était aussi communiqué lors des contacts téléphoniques à des tiers qui refusaient de transmettre les coordonnées de l'individu ou aux proches d'individus difficiles à joindre.

Ce numéro vert a été ouvert du lundi au samedi, entre le 4 avril et le 31 juillet 2013. L'appel était gratuit, que ce soit d'un poste fixe ou portable. En dehors des horaires d'ouverture, les jeunes pouvaient laisser un message sur un répondeur.

Les individus avaient également la possibilité de contacter la hotline par e-mail.

## 4.7. Facebook

Le plateau d'enquête de l'enquête dure trois mois durant lesquels les jeunes tentent de s'informer sur l'enquête avant de répondre à nos appels. Au-delà du site internet du Céreq, une page Facebook Céreq Génération 2013 (www.facebook.com/cereqG13) a été créée pour pallier ce manque d'interactions et d'échanges avec l'équipe en charge de l'enquête. De fait, l'alimentation de ce compte est réalisée par l'équipe.

La création de la page Facebook répond à plusieurs objectifs :

– créer un lien entre le Céreq et les jeunes ;

– recueillir les commentaires des jeunes sur l'enquête ;

– donner plus de visibilité au Céreq et à l'enquête Génération en particulier.

Au-delà de l'information donnée sur le Céreq et ses missions, la page Facebook était principalement destinée à informer les jeunes répondants sur l'enquête Génération elle-même. La page contenait les objectifs de l'enquête, les modalités, le déroulement, la périodicité et les dates d'interrogation, la présentation de l'équipe en charge de l'enquête, les actualités, les évènements en lien avec l'enquête, les informations post-enquêtes, des vidéos et les premiers résultats. Toutes les informations qui permettent de favoriser la réponse positive à l'enquête.

Par exemple, répondre au questionnaire permet d'enrichir les informations sur leur orientation, sur le choix des filières, la situation économique à l'entrée sur le marché du travail, le type de contrat proposé, le salaire, etc.

Il existe également un compte Twitter du Céreq (https://twitter.com/PRESSECEREQ), orienté presse, destiné plus particulièrement aux journalistes et au réseau institutionnel du Céreq. Le compte Facebook est synchronisé avec le compte Twitter. Ainsi, sur ce dernier apparaissent toutes les publications postées sur la page Facebook.

## 5. La collecte par téléphone

La collecte de l'enquête s'est déroulée par téléphone d'avril à juillet 2016 chez le prestataire IPSOS dont le plateau est basé à Bordeaux (33000). 158 401 individus ont été mis en production, la réserve tirée n'a pas été utilisée. Le plateau d'enquête a été exceptionnellement prolongé en juillet au vu du nombre insuffisant de répondants et au regard de l'avancée du protocole mis en place (environ 15 000 questionnaires réalisés fin juin).

L'enquête téléphonique est assistée par ordinateur, donc l'enquêteur interroge les individus par téléphone tout en suivant sur un écran d'ordinateur un script préétabli qui affiche les questions qu'il doit poser et les éventuelles modalités de réponse (système CATI), avec la consigne de lire strictement les questions et de lister les modalités, sauf mention contraire. Les réponses sont saisies directement sur informatique. Le CATI est capable de gérer des filtres et d'orienter l'individu vers des questions différentes en fonction des réponses précédemment données.

L'enquête a été développée à partir de deux logiciels qui communiquent entre eux : le logiciel *Quancept* pour la phase de contact et le logiciel *Confirmit* pour le questionnaire. Le calendrier d'activité a quant à lui été développé sous application externe, mais connecté au logiciel *Confirmit*.

Les téléenquêteurs étaient spécialisés sur des cibles particulières selon le niveau de diplôme. L'échantillon était subdivisé en trois terrains d'enquêtes : « niveau bac+2 et plus », « niveau bac et CAP/BEP diplômés » et « niveau CAP/BEP non diplômés et sans qualification ».

## 5.1. Calendrier et organisation générale de la collecte

**Tableau 34 • Calendrier de la collecte**

| Étapes | Délais |
|---|---|
| Transmission du questionnaire | 21 octobre 2015 |
| Réunion de lancement | 2 novembre 2015 |
| Programmation et tests du questionnaire CATI<br>Transmission au CÉREQ d'un lien test | Novembre 2015 |
| Transmission fichier pilote 1<br>*1er pilote CATI – 114 questionnaires – 35 minutes en moyenne* | 2 – 4 décembre 2015 |
| Transmission fichier pilote 2<br>*2e pilote CATI – 202 questionnaires – 33 minutes en moyenne* | 29 février – 5 mars 2016 |
| Transmission fichier pilote 3<br>*3e pilote CATI – 162 questionnaires – 32 minutes en moyenne* | 22 – 26 mars 2016 |
| Envoi des lettres avis courrier | 29 mars 2016 |
| Ouverture numéro vert | 29 mars 2016 |
| Envoi des lettres avis mail | 1er avril 2016 |
| Ouverture du site internet | 1er avril 2016 |
| Formations enquêteurs | 4, 5, 27 avril, 3 mai, 30 juin 2016 |
| Début du plateau d'enquêtes | 4 avril 2016 |
| Relance mail | 28 juin 2016 |
| Relance SMS | 22 juillet 2016 |
| Fin du plateau | 31 juillet 2016 |

## 5.2. Les nouveautés dans la collecte

Pour cette enquête il y a eu deux nouveautés par rapport aux enquêtes précédentes.
Premièrement, les appels étaient effectués en alternant un numéro de téléphone masqué et un numéro de téléphone non masqué. De plus, le numéro non masqué était pour la première fois de type régionalisé c'est-à-dire que le numéro changeait selon la région de l'individu. Il y avait également un numéro de mobile commençant par 06. En cas de rappel de l'individu, l'ensemble des numéros utilisés aboutissaient sur la plateforme des appels du numéro vert. Si un individu, d'office, n'accepte pas les numéros masqués, un démasquage était effectué sur la totalité des tentatives restantes.
Deuxièmement, pour la première fois, les individus avaient la possibilité de répondre au questionnaire filtre par internet. Dans le cas où ils appartenaient au champ, une prise de rendez-vous leur était proposée afin de terminer le questionnaire.

## 5.3. Le suivi de la collecte en chiffres

### 5.3.1. Nombre d'enquêteurs

Initialement, 93 enquêteurs et 6 superviseurs ont été formés. Divisés en trois groupes, ils ont chacun reçu une préformation d'une demi-journée par le chef de plateau puis une seconde demi-journée de formation menée conjointement par le Céreq et le prestataire.

La préformation a permis de former les enquêteurs à l'utilisation des outils *Quancept* et *Confirmit* pour compléter le questionnaire et notamment le calendrier. La formation CÉREQ/prestataire a, quant à elle, consisté à présenter l'enquête Génération, comprenant la méthodologie, le contenu du questionnaire, la phase de contact ; et à réaliser des exercices pratiques permettant aux enquêteurs de se familiariser avec le questionnaire et les consignes.

Après trois semaines d'enquêtes, il s'est avéré que le nombre d'enquêteurs était insuffisant. En effet, deux facteurs se sont combinés : le questionnaire s'est révélé plus long que prévu et il a fallu pallier l'absence récurrente d'enquêteurs, notamment pour cause de grève. Ainsi, en l'espace d'une semaine, 57 nouveaux enquêteurs ont été formés.

À cela s'ajoutent 16 enquêteurs formés fin juin afin d'avoir un nombre suffisant d'enquêteurs pour la prolongation du plateau au mois de juillet.

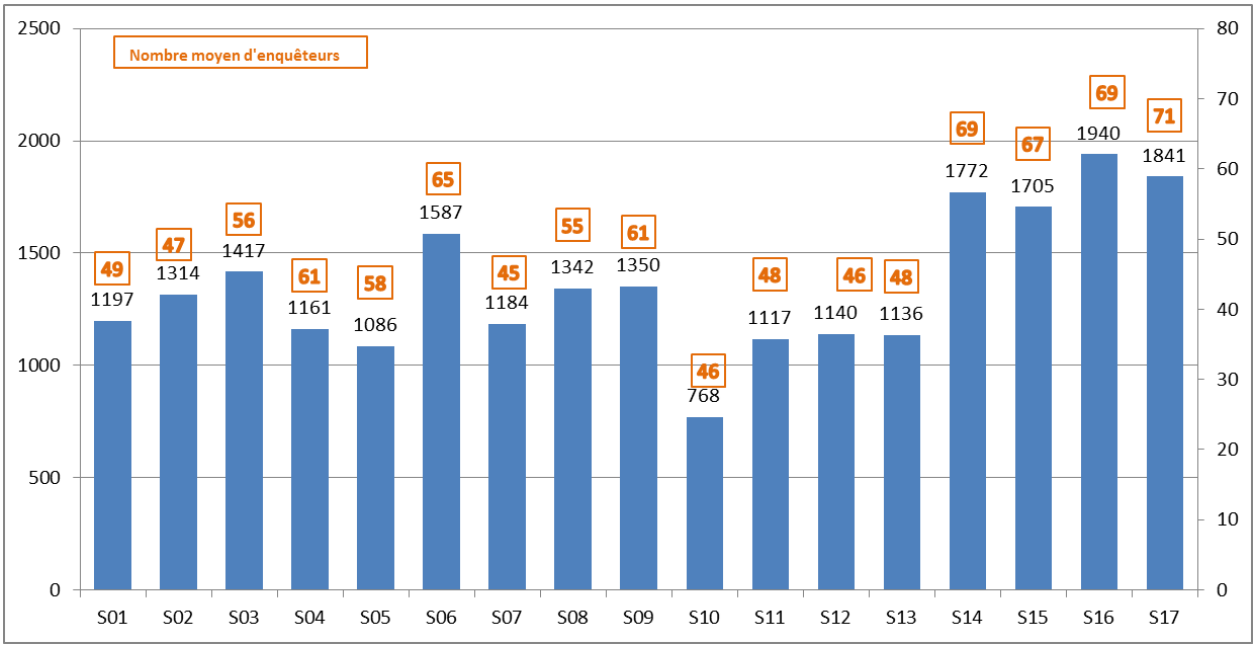
Au total, 6 formations ont été réalisées pour former 166 enquêteurs et 6 superviseurs.

Le tableau ci-dessous détaille chaque formation réalisée.

**Tableau 35 • Formation des enquêteurs**

| | Équipe A | Équipe B | Équipe C | Équipe D | Équipe E | Équipe F | Global |
|---|---|---|---|---|---|---|---|
| **DATE BRIEFING** | **04/04/2016** | **05/04/2016** | **05/04/2016** | **27/04/2016** | **03/05/2016** | **30/06/2016** | **-** |
| **Durée préformation** | 1/2 journée | 1/2 journée | 1/2 journée | - | 1/2 journée | - | **2 jours** |
| **Durée formation** | 1/2 journée | 1/2 journée | 1/2 journée | 1 journée | 1/2 journée | 1 journée | **4 jours** |
| **Enquêteurs** | 31 | 32 | 30 | 32 | 25 | 16 | **166** |
| **Superviseurs** | 4 | 0 | 2 | 0 | - | 0 | **6** |
| **Début terrain** | 04/04/2016 | 05/04/2016 | 05/04/2016 | 28/04/2016 | 04/05/2016 | 01/07/2016 | **-** |

**Figure 7 • Nombre d'enquêtes réalisées selon la période et le nombre de téléenquêteurs**



### 5.3.2. Statistique de la hotline

Au total, la hotline a été sollicitée 25 048 fois par les individus. Ces demandes s'étalent entre le 1er avril et le 27 juillet 2016.

Les demandes concernent plus précisément :

- 72 % par téléphone avec un contact direct.
- 27 % par téléphone sans contact direct. Parmi eux seul 1 individu sur 3 a laissé un message sur le répondeur.
- 1 % par mail.

Le tableau suivant présente les motifs pour lesquels la hotline a été contactée.

**Tableau 36 • Motifs de contacts de la hotline**

| Motifs | Effectif | % |
|---|---|---|
| Message répondeur / appel vide | 7 018 | 28,0 |
| Prise de rendez-vous | 5 316 | 21,2 |
| Autres (coordonnées, disponibilités, cas particulier…) | 4 586 | 18,3 |
| Demande d'informations… | 4 080 | 16,3 |
| Suite à un appel enquêteur, attend le rappel… | 3 152 | 12,6 |
| Refus…ne pas appeler… | 487 | 1,9 |
| Hors-champs | 257 | 1,0 |
| Passation questionnaire filtre | 151 | 0,6 |
| Correction par hotline (admin) du nom ou prénom | 1 | 0,0 |
| **Total** | **25 048** | **100,0** |

Parmi les rendez-vous pris depuis la hotline, 56 % se transforment en interviews (court ou long).

### 5.3.3. Statistique du site internet

24 708 individus se sont connectés au site internet, soit 15,6 % de l'échantillon, et ils se sont connectés en moyenne 2,3 connexions par individus. Plus précisément :

- 6 070 (24,6 %) ont mis à jour leurs coordonnées.
- 938 (3,8 %) ont pris un rendez-vous. Parmi eux, 735 ont réalisé une interview, soit un taux de transformation de 78 %.
- 1 326 individus ont été définis éligibles à la fin du questionnaire filtre dont 1 048 ont été recontactés et ont pu terminer leur interview (soit un taux de transformation de 79 %).
- 976 ont été détectés hors-champ

### 5.3.4. Bilan enrichissement en cours d'enquête

Pour 18 632 individus, les coordonnées téléphoniques ont été enrichies. Ainsi, 19 432 nouveaux numéros de téléphone ont été intégrés.

79 numéros de téléphone issus de la base de sondage ont été supprimés suite à un premier nettoyage de la base (numéro débutant par 00, chaîne de caractère dans le champ…).

### 5.3.5. Statistique des appels

76 % des enquêtes ont été réalisées à partir d'un portable

### 5.3.6. Relance mail

La relance mail a été réalisée le 28 juin 2016.

Deux types de mails ont été envoyés :

- un mail de relance pour les individus n'ayant jamais été contactés ;
- un mail de relance pour les individus qui ont déjà été contactés, mais aucune interview n'a été réalisée.

Au global, 44 901 mails ont été envoyés.

### 5.3.7. Relance SMS

Initialement, la relance SMS devait concerner tout ou partie des individus disposant d'un numéro de téléphone portable et n'ayant toujours pas complété le questionnaire. Compte tenu de l'avancée de la production en date de mi-juillet 2016 et tout particulièrement pour les extensions d'échantillon, il a été convenu que la relance SMS serait plus ciblée sur ces publics. Ainsi, le 22 juillet 2016, une relance SMS a été réalisée pour les individus des extensions DGESIP, CGDD, Santé/Social, ainsi que sur certains diplômes, ayant au moins un numéro de téléphone portable sur les deux premiers numéros et n'ayant toujours pas complété le questionnaire (questionnaires non commencés et incomplets. De plus, sont exclus, parmi ces individus, ceux qui avaient pris un rendez-vous.

Parmi les 32 570 individus concernés, seuls 27 465 ont reçu un SMS de relance.

### 5.3.8. Durée de passation du questionnaire

La durée de questionnaire a été plus longue que prévu. En effet, initialement évaluée à 19 minutes en moyenne, elle a finalement été de 30 minutes.

**Figure 8 • Évolution de la durée de passation du questionnaire**

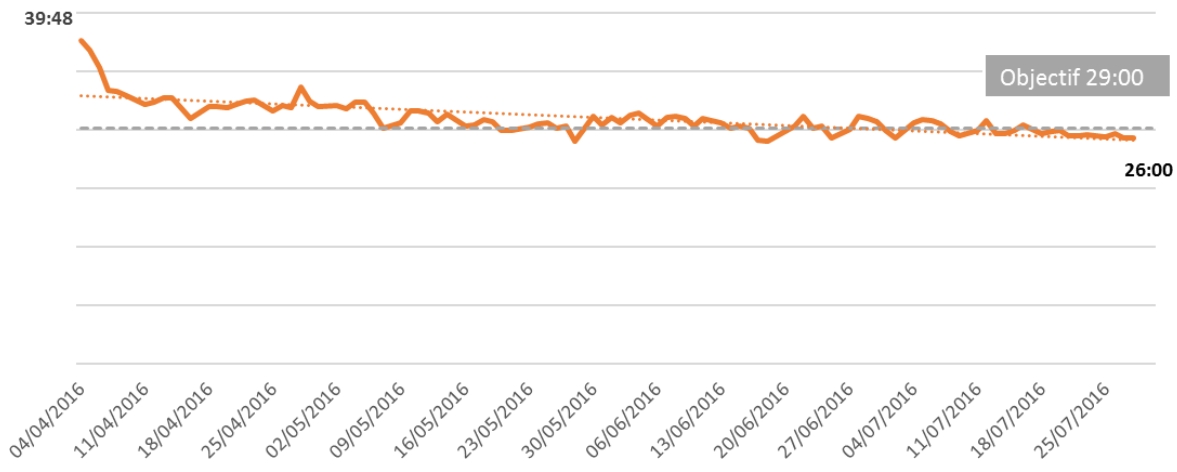
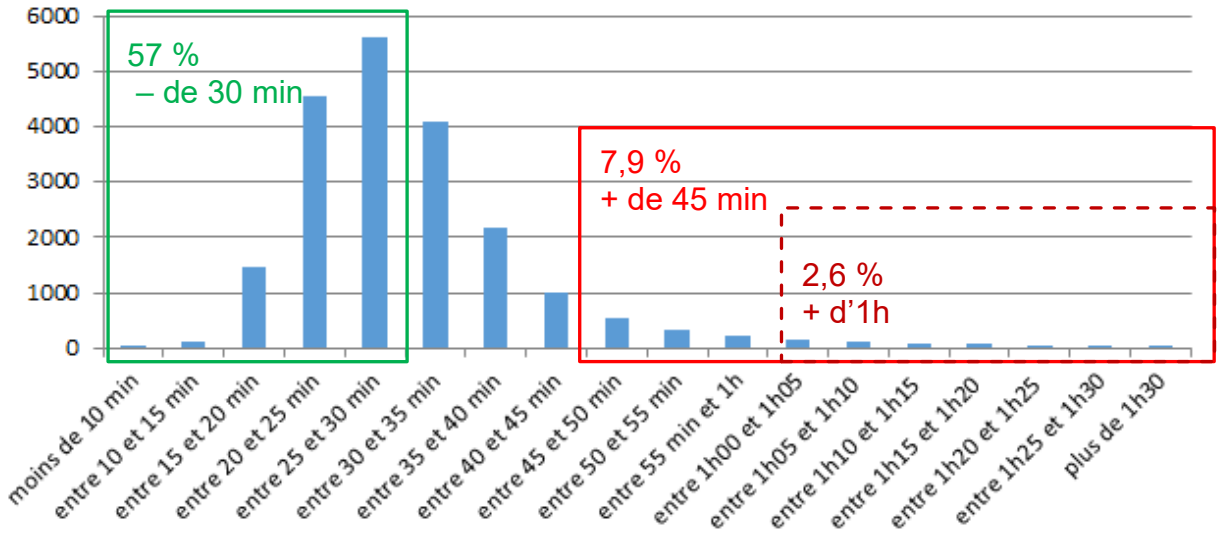


**Figure 9 • Distribution de la durée de passation du questionnaire**




Parmi les répondants ayant terminé le questionnaire, 57 % des entretiens ont duré moins de 30 minutes, 7,9 % ont duré plus de 45 minutes et 2,6 % plus d'une heure.

## 5.4. Les règles de rappel

**Tableau 37 • Règles de rappel détaillées**

| Statut de l'appel | Passage au numéro suivant | Compteur +1 sur le tél. en cours | Enquêteur | Délai de rappel | Règle |
|---|---|---|---|---|---|
| **Ne répond pas** | OUI | OUI | Libre | 80 minutes | Appel sur le 1er et le 2e numéro en alternance avec jusqu'à 20 appels sur chaque numéro (sauf si classement hors cible sur l'un des numéros, dans ce cas continuer les 20 appels sur le numéro restant puis passage au TEL+1) |
| **Occupé** | OUI | OUI | Libre | 10 minutes | |
| **Rendez-vous** | | | | À la date déterminée par l'individu | Honorer tous les rendez-vous |
| **Change le numéro de téléphone** | | NON | Même enquêteur que l'appel précédent | Immédiat | Appel sur le numéro enrichi |
| **Répondeur** | NON | OUI | Libre | 110 minutes | |
| **Abandon questionnaire** | | | | | Appliquer les règles précédentes selon le cas |

**Suite tableau 37 :**

| Statut de l'appel | Passage au numéro suivant | Compteur + 1 sur le tél. en cours | Enquêteur | Délai de rappel | Règle |
|---|---|---|---|---|---|
| **Hors cible** | | | | | |
| **Individu non francophone/pb physique ou mental** | NON / OUI | NON | | | S'il s'agit de la bonne personne, il s'agit bien d'un hors cible mais il ne faut pas passer au TÉL+1. Sinon on passe au TÉL+1 |
| **Injoignable durée d'étude (avant fin juin 2016)** | NON | NON | | | |
| **Il/elle habite à l'étranger** | NON | NON | | | |
| **Il/elle est décédé(e)** | NON | NON | | | |
| **Faux numéro nominatif** | OUI | NON | | | |
| **Homonyme parfait** | OUI | NON | | | |
| **Ne connait pas les coordonnées** | OUI | NON | | | |
| **Autres Hors cibles (contact)** | OUI | NON | | | |
| **Nés avant 78 qui ne sont pas en formation « santé social », ni « sport », ni « thèse »** | OUI | NON | | | |
| **Non-inscrits dans établissement du fichier et date de naissance différente du fichier** | OUI | NON | | | |
| **Hors-champ (il s'agit bien de la cible – le bon individu)** | | | | | |
| **Non-inscrits dans un établissement scolaire** | NON | NON | | | |
| **Primo sortants santé social ou sport de plus de 35 ans** | NON | NON | | | |
| **Ont terminé formation avant octobre 2012 ou après décembre 2013** | NON | NON | | | |
| **Primo sortants docteur de plus de 35 ans** | NON | NON | | | |
| **Post-initiaux docteurs n'ayant pas d'adresse mail pour bifurquer en CAWI** | NON | NON | | | |
| **Poursuite études ou interrompu plus d'un an** | NON | NON | | | |
| **Reprise d'étude dans le calendrier** | NON | NON | | | |

**Suite tableau 37 :**

| Statut de l'appel | Passage au numéro suivant | Compteur + 1 sur le tel en cours | Enquêteur | Délai de rappel | Règle |
|---|---|---|---|---|---|
| **Refus** | | | | | |
| **Raccroche au nez après 2e tentative** | NON | OUI | Même enquêteur que l'appel précédent | | |
| **De la personne concernée – ne plus jamais appeler** | NON | NON | | | |
| **D'un tiers de passer la personne concernée** | OUI | NON | | | |
| **D'un tiers de donner les coordonnées** | OUI | NON | | | |
| **Refus définitif CNIL** | NON | NON | | | |
| **Autre refus** | OUI | NON | | | |

## 5.5. Messages sur répondeurs

À partir du mois de mai 2016, soit un mois après le début de l'enquête, le processus de dépôt de message sur répondeur a été lancé. Cette démarche avait pour objectif de sensibiliser les individus qui ne répondaient pas à la vue du numéro qui s'affichait.

## 5.6. Suivi technique et personnes « qualité »

En raison de la distance entre le plateau d'enquête, basé à Bordeaux, et le Céreq, basé à Marseille, deux personnes « qualité » ont été recrutées pour suivre le plateau téléphonique en continu et effectuer le relais avec les membres de l'équipe ingénierie et gestion d'enquêtes du Céreq. Cela a notamment permis au Céreq d'avoir de l'information en temps réel. Ces personnes « qualité » ont réalisé des écoutes quotidiennes, assisté les superviseurs et assuré la formation des enquêteurs lorsque les agents du Céreq ne pouvaient être présents. Elles ont fait part des difficultés rencontrées sur place et permis au Céreq de réagir efficacement face aux problèmes humains, techniques ou organisationnels.

## 5.7. Les résidents étrangers à la date de l'enquête

Les individus identifiés par un tiers lors de la phase de contact comme résidents étrangers ou ceux le déclarant directement sur le site internet sont considérés hors-champ de l'enquête téléphonique. 2 530 individus ont été topés comme résidents à l'étranger. Leurs questionnaires ne sont donc pas disponibles dans les bases de l'enquête téléphonique. Par ailleurs, il leur est proposé, lorsque l'adresse mail de l'individu est disponible (donné par le tiers ou renseigné par l'individu), de répondre à une enquête expérimentale multimode internet/téléphone développée en simultané. Cette expérimentation fera l'objet d'une expertise afin de déterminer la possibilité d'une future intégration dans le champ de ces jeunes dans les enquêtes. Le bilan de cette enquête multimode devrait paraître prochainement.

## 5.8. Les post-initiaux docteurs

De la même manière que pour les individus résidants à l'étranger, les post-initiaux docteurs ont été interrogés dans l'enquête multimode. C'est-à-dire, que ces 280 questionnaires ne sont pas disponibles dans les bases, car ils sont considérés comme hors-champs de l'enquête principale téléphonique.
L'enquête expérimentale étant en multimode c'est-à-dire internet et téléphone, les post-initiaux docteurs ont eu la possibilité de répondre soit par internet soit par téléphone.

# 6. Taux de réponse

Parmi les 158 401 individus de l'échantillon, 141 775 ont fait l'objet de tentatives d'appels et 16 626, soit 10 %, ne l'ont pas été parce qu'aucune coordonnée téléphonique n'était disponible pour eux. Ce taux était de 9 % pour l'enquête 2013 auprès de la Génération 2010, de 14 % pour la Génération 2007 et de 22 % pour la Génération 2004.

Pour cette enquête, 45 835 individus ont accepté de répondre à l'enquête. On compte :

- 22 737 questionnaires complétés et valides (dont 3 239 appartenant au champ d'extensions spécifiques) ;
- 22 999 questionnaires hors-champ (dont 258 détectés en post-collecte) ;
- 99 questionnaires non exploitables, supprimés *a posteriori*.

Ainsi, le taux de succès global ou taux de réponse (*i.e.* le nombre de répondants rapportés au nombre d'individus dans l'échantillon) est de 28,9 %. Ce taux était de 26,9 % pour l'enquête 2013 auprès de la Génération 2010 et de 25,2 % pour l'enquête 2010 auprès de la Génération 2007.

La pondération fournie dans les fichiers de diffusion permettra de corriger le phénomène de non-réponse observé sur l'échantillon.

Le tableau ci-après présente le classement final des individus échantillonnés.

**Tableau 38 • Classement des individus échantillonnés**

| Résultats par individu | Total | | |
|---|---|---|---|
| | Effectif | % | |
| A – Individus joints ayant complété le questionnaire (Champ Céreq + extensions spécifiques) | 22 737 | 14,4 | Taux de réponse = 28,9 % |
| B – Individus joints classés hors du champ Céreq | 22 999 | 14,5 | |
| C – Questionnaires inexploitables supprimés post-enquête | 99 | 0,1 | |
| D – Individus hors cible | 15 120 | 9,5 | |
| E – Refus de répondre | 11 074 | 7,0 | |
| F – Abandon de questionnaire | 1 440 | 0,9 | |
| G – Ne réponds pas / Occupé / Répondeur | 54 854 | 34,6 | |
| H – Rendez-vous non abouti | 1 574 | 1,0 | |
| I – Aucun contact (Absence de numéro ou faux numéro) | 28 504 | 18,0 | |
| **Total** | **158 401** | **100,0** | |

Note de lecture : 9,5 % des individus échantillonnés ont été classés hors cible (classement D). Cela signifie que l'on n'a jamais réussi à les joindre en personne, mais qu'un contact téléphonique (au moins) a abouti sur un autre interlocuteur.

# 7. Les traitements en aval

L'enquête 2016 auprès de la Génération 2013 est une enquête dite « légère ». Les individus de cette enquête sont interrogés une seule et unique fois. Il n'est pas prévu de réinterrogation. Le questionnaire est construit dans un objectif d'actualisation des indicateurs d'insertion. La structure du questionnaire est donc allégée avec une description du parcours professionnel autour de deux points dans le temps : le premier emploi et la situation actuelle (en emploi ou en non-emploi). Ainsi, la constitution des bases est quelque peu différente par rapport à une enquête Génération dite « pleine ».

Cette partie présente l'ensemble des traitements réalisés pour construire des fichiers livrables à partir des fichiers bruts, à l'exception du calcul de la pondération finale présentée dans la partie suivante qui aborde le traitement de la non-réponse globale et le calage sur marge.

*Remarque* : Les observations entrant dans le champ Céreq et celles relevant d'extensions spécifiques (les post-initiaux des formations du sport et des formations de la santé et du social) n'ont pas été séparées pour les phases d'apurement, de codification et de création de variables. Elles n'ont été dissociées qu'à la fin des traitements pour la constitution des bases finales et les calculs de pondération.

*Remarque 2* : une enquête multimode expérimentale (téléphone et/ou internet) a été menée en parallèle de l'enquête principale. En raison de différences très marquées, le traitement des deux jeux de données a été réalisé distinctement.

## 7.1. Processus général d'apurement des données collectées

À l'issue de la collecte, le prestataire a livré quatre bases brutes :

- Une table *calendrier d'activité* regroupant le parcours professionnel des individus interrogés. Elle contient, pour chaque individu, une ou plusieurs situations décrites dans le cadre d'un calendrier d'activité interactif croisant les mois et situations rencontrées pendant la période d'observation.

- Une table *calendrier du mode de cohabitation* regroupant les situations d'habitat des individus interrogés. Elle contient, pour chaque individu, une ou plusieurs situations d'habitat (même principe que le calendrier d'activité).

- Une table *concours* regroupant les concours tentés décrits spécifiquement dans un module de questions incluant une boucle permettant de collecter jusque cinq concours maximums par individu (cf. partie 1.4.4., module « Concours et attractivité de la fonction publique »).

- Une table *individus* contenant les données individuelles (informations collectées en début et en fin du questionnaire), les données en lien avec le parcours scolaire, la liste des diplômes obtenus ainsi que les informations relatives à la situation à date de l'enquête.

### 7.1.1. Stabilisation du nombre de questionnaires exploitables

La version brute des bases de données ne peut être exploitée en l'état. Ces bases contiennent l'ensemble des individus envoyés en production, soit 158 401 observations. Pour la phase de redressement, il est important d'identifier les individus dans le champ Céreq, dans le champ des « post-initiaux » et ceux hors du champ de l'enquête et de déterminer précisément le statut de chaque individu contacté.

Dans le processus d'apurement, il s'agit d'identifier les individus dans le champ et ayant répondu intégralement au questionnaire. À ce stade, 23 094 questionnaires réalisés sont traités et validés pour être livrés. Ce chiffre est provisoire. En effet, une étape importante de vérification de la cohérence du questionnaire est réalisée. Il s'agit de vérifier et de limiter la suppression de questionnaire. Plusieurs éléments permettent de les détecter et le travail consiste à « corriger » lorsque cela est possible ces incohérences. L'objectif étant de disposer d'un maximum de questionnaires exploitables.

Parmi les incohérences rencontrées, la plupart concernent :

- L'ouverture de modules en lien avec le calendrier d'activité. Pour exemple, si la situation actuelle dans le calendrier est de l'emploi, le module emploi actuel doit s'ouvrir. Mais dans quelques cas, ce module de question est vide. Le questionnaire est donc inexploitable, car il n'y a aucune possibilité de correction (l'information est manquante) (41 observations).

- Le calcul du « CAL » (tableau 3) après saisie des situations dans le calendrier d'activité. Pour chaque situation décrite d'emploi ou de non-emploi, une valeur est attribuée à la variable CAL qui pilote l'ouverture des modules. Dans certains questionnaires, la valeur du « CAL » est fausse. Pour exemple, une valeur de « CAL » d'une situation du passé a été attribuée à la situation à la date de l'enquête. Il manque donc la description du module actuel, car il ne s'est pas ouvert. Dans ce cas, les questionnaires ont été supprimés (13 observations).

- Le cas du « CAL 22 », c'est-à-dire les reprises d'études à temps plein dans un établissement scolaire ou universitaire dans l'année qui suit la fin de formation. La particularité de ce « CAL » est qu'il n'a pas la fonction « classique » de définition des situations professionnelles. Il a été créé pour « repêcher » les individus qui ne sont pas dans le champ de l'enquête, mais qui sont parvenus à passer le questionnaire filtre. Son objectif : détecter ces individus et stopper leur questionnement après la validation du calendrier d'activité. Dans le cadre de cette enquête, la restriction de champ a été plus importante, néanmoins beaucoup d'individus ont poursuivi à tort et ont dû être supprimés (hors-champ) (258 observations).

- Le cas de l'intérim. La description dans le calendrier d'activité est particulière (cf. partie 1.4.3). Pour gérer la multitude de situations possibles, il existe des règles de saisie spécifiques. Il existe trois types de situations d'intérim à décrire ; première mission, mission la plus longue et mission actuelle. Des incohérences ont été détectées dans le remplissage des situations, notamment sur les dates de début et de fin de mission. Pour la majorité de ces calendriers, les dates ont été corrigées, dans tous les autres cas, le questionnaire a été supprimé (45 observations).

Au total, moins de 2 % des questionnaires, considérés comme complets à l'issue de la collecte, sont inexploitables.

*Remarque* : concernant les incohérences en lien avec le calcul des « CAL », une partie est imputable à une erreur de programmation du calendrier d'activité. En effet, la possibilité de saisir le calendrier de façon non chronologique était prévue. Cependant, la numérotation des différentes situations était fausse. A été développée une numérotation chronologique en fonction de la saisie dans le calendrier et non en lien avec les dates des différentes situations décrites. Par conséquent :

- Des modules de situation d'emploi ou de non-emploi à la date de l'enquête n'ont pas été ouverts (description manquante).

- Le module de premier emploi a été posé à la mauvaise situation d'emploi décrite dans le calendrier.

### 7.1.2. Vérifications de cohérence générale

Après avoir vérifié la cohérence des questionnaires dans leur globalité en réalisant les correctifs nécessaires. Quelques tests de contrôle sont réalisés sur le calendrier d'activité en particulier.

- Test pour vérifier la présence de chevauchement de situations

- Test pour vérifier que le calendrier est bien complet

- Test de cohérence dans l'ouverture des modules post-calendrier

- Test de cohérence sur la variable RETOU (cf. partie 1.4.3). Il s'agit de vérifier si pour chaque individu la présence multiple d'une même entreprise (couple = commune + code postal) existe et que la variable est bien codée.

L'ensemble de ces tests permettent de garantir l'exploitation de chaque questionnaire mis à disposition.

Concernant la vérification des filtres (inter-question) pour le reste du questionnaire, l'outil d'enquête utilisé permet la suppression automatique des réponses sélectionnées lorsqu'un chemin de questionnement est emprunté par erreur. Cette phase de vérification est apparentée à celle réalisée avant la mise en production du questionnaire (cf. partie 4.1.2.).

## 7.2. Création des premières bases exploitables

### 7.2.1. Création de la base individus

Il s'agit premièrement de supprimer toutes les variables en lien avec le calendrier d'activité, des ouverts de question, une partie des variables d'identification de l'individu (questionnaire filtre) et les variables de filtres non exploitables. Ensuite, des variables individuelles de synthèse sont construites à partir des calendriers d'activité et du mode de cohabitation afin d'exploiter rapidement la base des individus.

#### *Variables construites à partir du calendrier d'activité*

Les variables MOIS1 à MOIS45 permettent de décrire mois par mois la situation professionnelle de l'individu et à quel moment elle se produit dans son parcours professionnel. Ces variables MOISX s'échelonnent de MOIS1 qui correspond au mois de novembre 2013 (premier mois possible observé) à MOIS45 pour juillet 2016 (dernier mois possible observé).

Chaque MOISX est défini sur quatre positions :

- Les deux premières positions correspondent à la variable « CAL » qui définit le type de situation rencontrée (présente dans les bases de situations d'emploi et de non-emploi).

- Les deux dernières positions renvoient au numéro de la situation professionnelle correspondante, décrite dans le parcours

Pour chaque individu, les variables MOISX sont à valeur manquante jusqu'à la date de fin de leur formation (Q20) incluse (autrement dit, MOIS1 correspond au mois de fin de formation+1).

Les variables ID, IF et DUROBS correspondent respectivement aux numéros du premier et du dernier mois observé et au nombre total de mois renseignés dans le calendrier d'activité (DUROBS = IF-ID+1).

Les variables NMEMP, NMJVAC, NMCHO, NMINA, NMFOR, NMETU et NMVAC renseignent sur le nombre total de mois passés dans chaque situation professionnelle, respectivement en emploi, dans un job de vacances, au chômage, en inactivité, en formation, en études et en vacances.

Les variables NSEMP, NSJVAC, NSCHO, NSINA, NSFOR, NSETU et NSVAC renseignent sur le nombre total de situations décrites dans chacune des situations professionnelles possibles. La variable NSTOT donne le nombre total de situations décrites de l'individu.

La variable TAPE renseigne sur le temps d'accès au premier emploi (si aucun emploi occupé, TAPE est à valeur manquante), TRAJPRO sur l'enchaînement des situations et SITDE à la situation professionnelle à la date d'enquête.

### *Variables construites à partir du calendrier du mode de cohabitation*

Le calendrier du mode de cohabitation permet d'observer, mois après mois, la situation d'habitat selon trois statuts ; vit chez ses parents, vit en couple ou vit seul (y compris foyer, colocation pour cette dernière situation).

Comme pour le calendrier d'activité, l'ouverture de ce calendrier s'opère à partir du mois suivant la date de fin de formation (Q20). Les variables s'échelonnent également de HMOIS1 à HMOIS45 avec la même périodicité. Chaque HMOISX est défini sur quatre positions :

- Les deux premières positions renseignent sur le mode d'habitat qui définit le type de situation rencontrée.

- Les deux dernières positions renvoient au numéro de la situation professionnelle correspondante, décrite dans le parcours (mettant ainsi en parallèle mode de cohabitation et situation professionnelle).

Les situations sont exclusives et hiérarchisées. Par exemple, un individu déclarant vivre en couple chez ses parents est considéré comme vivant chez ses parents (mode de cohabitation principal).

Pour compléter les données individuelles collectées, une sélection de variables, issues de la base de sondage, est réalisée pour vérifier la cohérence de l'historique de formation de chaque individu. Ce sont des variables de diplômes et de spécialités, des variables autour du statut de l'individu (apprenti, localisation en Zus, etc.), également des variables pour caractériser l'établissement de formation (région, etc.)… La table Individus est enrichie par des informations issues de la base de sondage et relatives à la scolarité suivie en 2012-2013 : formation suivie, niveau, spécialité…

Ces informations complètent les réponses au questionnaire sur les informations concernant le parcours scolaire (série du bac, diplômes, etc.).

### 7.2.2. Création des bases d'emploi et de non-emploi

La table *calendrier d'activité* est stabilisée suite aux ajustements en lien avec la cohérence de la description de chaque parcours des répondants définitifs. Pour faciliter l'utilisation de ce dernier, plusieurs variables annexes sont fournies :

- Les variables de dates sont recalculées sous forme d'indice.
- Un nouveau numéro de séquence (situation) NSEQ a été créé pour définir le parcours professionnel de chaque individu (suite à la problématique de la numérotation selon la saisie et non selon des dates des situations).
- Les situations d'intérim de plusieurs missions dans plusieurs entreprises sont retraitées en lien avec les règles spécifiques de remplissage du calendrier. En effet, la règle est de sélectionner une situation pour l'ensemble des missions réalisées (lorsqu'elles sont multiples). Dans le cas d'une description de la première mission et de la mission actuelle, est affectée la date du milieu de la période pour définir la date de fin de la première mission et la date de début de celle actuelle.
- La création de la variable SITDE qui synthétise la description de la situation professionnelle à la date de l'enquête : emploi, chômage, formation, reprise d'études ou inactivité

Il est maintenant possible de scinder en deux tables distinctes ce calendrier, une table des situations d'emploi SEQENTR et une table des situations de non-emploi NONEMPL. L'information est recomposée dans ces tables avec respectivement la description des situations d'emploi du passé et celles de non-emploi du passé (le cas échéant) ainsi que la situation actuelle qui peut être soit de l'emploi soit du non-emploi.

#### *Phase dite de « Recopie » dans la base des emplois*

Dans le cadre d'une génération « légère », les informations collectées sont principalement la description de la situation de premier emploi et celle à la date de l'enquête. Autrement dit, pour la situation d'emploi du passé, la collecte se focalise sur la situation à l'embauche de l'individu et pour l'emploi à la date de l'enquête, sa situation actuelle (selon les cas). Pour alléger le questionnement et apporter de la fluidité dans la passation, quelques questions ne sont posées qu'une seule fois au moment de la description d'un emploi. Les questions concernées sont :

- Le libellé de la profession.
- Le contrat de travail.
- Le temps de travail.
- Le salaire et les primes.
- La catégorie socio professionnelle (PCS).

Le tableau 39 présente les différentes définitions d'une situation d'emploi en fonction de sa temporalité et de sa durée. La mention « changement » fait référence, dans le cas d'un emploi dont la durée est supérieure à 6 mois, a un changement de situation dans l'emploi (changement de profession, contrat de travail, etc.). Pour exemple, si l'individu déclare ne pas avoir changé de profession dans l'emploi alors la question sur le libellé de la profession n'est posée qu'une fois, dans le cas contraire, elle est posée deux fois à l'embauche et à la date de l'enquête.

**Tableau 39 • Résumé des informations collectées sur les situations d'emploi**

| Situation | Cal | Durée | Changement | Embauche | Actuel |
|---|---|---|---|---|---|
| **Emploi du passé** | | | | | |
| • Emploi court | 01 | ≤ 12 mois | | ✓ | |
| • Emploi long | 02 | > 12 mois | | ✓ | |
| **Emploi actuel** | | | | | |
| Cas 1 (Premier emploi ≠ emploi actuel) | | | | | |
| - Emploi court | 03 | ≤ 6 mois | | | ✓ |
| - Emploi long | 04 | > 6 mois | | | ✓ |
| Cas 2 (Premier emploi = emploi actuel) | | | | | |
| - Emploi court | 03 | ≤ 6 mois | | | ✓ |
| - Emploi long | 04 | > 6 mois | Non<br>Oui | ✓<br>✓ | <br>✓ |

Pour revenir à la recopie, dans le cadre de cette enquête, un seul cas se produit (cas 2 sans changement dans le tableau 39). Le principe d'une recopie est une duplication de l'information collectée. Ici, si l'individu décrit une situation d'emploi longue et qu'il n'a pas changé de statut dans l'emploi (profession, contrat de travail, etc.) alors la recopie de l'information à l'embauche est faite sur l'information « manquante » à la date de l'enquête.

Dans tous les autres cas, aucune recopie n'est effectuée à la différence d'une enquête Génération « pleine » où tous les points dans le temps sont importants.

### *Normalisation des noms de variables*

Pour compléter, la logique de livraison des bases consiste aussi en une normalisation des noms de variables. Les informations autour de l'emploi actuel sont stockées dans des variables EAXX et celles du non-emploi essentiellement en NEAXX. Par souci de simplification et dans la logique de la construction des bases Génération, toutes les variables d'emploi EA sont renommées EP (sauf dans le cas où une question de la situation actuelle n'existe pas dans le passé alors la variable garde son nom d'origine). La même logique est appliquée aux variables de non-emploi.

Pour déterminer quelles sont les situations d'emploi ou de non-emploi actuelles lors de l'exploitation, l'utilisation de la variable ACTU est préconisée (variable construite, vaut un si c'est une situation actuelle).

## 7.2.3. Création de la base « concours »

La table concours est directement liée à une demande de partenaire d'extension et concerne l'ensemble des individus dans le champ Céreq. Cette base résulte d'un module de questionnement spécifique sur l'attractivité de la fonction publique. Construite sur une boucle, la création d'une table indépendante est apparue comme le format le plus adapté pour l'exploitation des données.

## 7.3. Création des variables synthétiques selon les nomenclatures officielles

### 7.3.1. La codification des diplômes et des spécialités

Le questionnaire de l'enquête 2016 auprès de la Génération 2013 débute par une validation du diplôme de sortie, disponible dans la base de sondage et un approfondissement de l'ensemble de son parcours scolaire avec un focus sur l'obtention de diplôme(s) autre que celui de sortie de formation initiale.
L'ensemble de ces diplômes sont codés *a posteriori* selon la nomenclature INSEE. Les diplômes concernés par cette codification concernent le diplôme de sortie, le baccalauréat et les autres diplômes éventuellement obtenus, permettant *in fine* de déterminer le plus haut diplôme obtenu par l'individu.

Ci-dessous est détaillé l'ensemble des étapes de codification des diplômes.

#### *Apurement des variables ANTER et SUPER*

Les variables ANTER et SUPER (variables transmises par l'établissement de formation et disponibles dans le fichier d'import) déterminent respectivement si l'individu est issu d'une classe terminale et s'il est sortant du supérieur (dont IV+).

Dans le cadre de cette enquête, elles sont recalculées « en direct » en cas d'invalidation d'informations de la base de sondage par l'individu (diplôme de sortie, année terminale). Ces variables sont déterminantes dans la codification du diplôme de sortie et la construction du plus haut diplôme (exemple : la question Q7B – Obtention du diplôme – est filtrée en fonction de la variable ANTER). Il est donc essentiel de mettre à jour, des données collectées, l'information contenue dans ces variables. Ainsi les variables ANTER_NEW et SUPER_NEW ont été créées, potentiellement différentes des variables initiales si invalidées.

#### *Construction des libellés des diplômes*

En fonction des données disponibles pour la constitution des libellés, le traitement est réalisé différemment dans la définition du diplôme de sortie, du baccalauréat ou des autres diplômes.

Le libellé du diplôme de sortie est construit, soit à partir des informations de la base de sondage, soit à partir des informations recueillies en enquête en cas d'invalidation du diplôme ou lorsque l'information initiale est incomplète.

L'information du baccalauréat peut être collectée auprès de l'établissement de sortie par le biais des bases SISE lorsque l'individu est sortant d'une formation dispensée par un établissement de l'enseignement supérieur. L'individu peut également déclarer avoir obtenu un baccalauréat en enquête lors d'un focus sur les diplômes (partie parcours scolaire). Il peut enfin s'agir de son diplôme de sortie s'il est un bac et que l'individu déclare l'avoir obtenu.

Parmi les autres diplômes obtenus par l'individu, le diplôme le plus élevé (autre que son diplôme de sortie), information essentiellement déduite de l'enquête, est sélectionné pour être pris en compte dans la codification du plus haut diplôme.

#### *Codification des diplômes*

La codification des diplômes consiste en l'attribution pour un libellé d'un code permettant d'identifier le diplôme et la spécialité de formation. Le logiciel *Sicore* est utilisé pour la codification automatique des diplômes. Il renvoie, pour chaque libellé, un code sur 8 positions, composé d'un code diplôme dans la nomenclature INSEE et d'un code dans la nomenclature NSF. Il renvoie également un code de qualité de codage, qui renseigne sur la nécessité ou non de vérifier le code automatique attribué ou de coder manuellement les codifications qualifiées d'incertaines.

En entrée du logiciel, il est possible, en plus du libellé de diplôme-spécialité, d'ajouter des variables annexes telles que la date d'obtention du diplôme et le niveau, comme c'est le cas pour le diplôme de sortie (pour lequel l'information est disponible).

Les libellés doivent pouvoir être « exploitables » par *Sicore*. Cela sous-entend un prétraitement des libellés : suppression des caractères spéciaux, signes de ponctuation et lettres avec accent ; mise en majuscule du libellé ; complétion des mots abrégés lorsque cela est possible.

**Encadré 4 • SICORE Environnement diplôme et spécialité (millésime 2013)**

Le logiciel *Sicore* (système informatique de codage des réponses aux enquêtes) est un système de chiffrement automatique développé par le Département des projets de l'INSEE, qui a fait l'objet de tests sur de nombreuses variables et a déjà été appliqué en production avec succès. Son usage est appelé à se généraliser à l'INSEE, et éventuellement à d'autres services statistiques, en France ou dans d'autres pays. À partir de l'enquête auprès de la Génération 1998, le Céreq a utilisé *Sicore* pour coder la profession. Depuis la Génération 2010, son utilisation a été généralisée à la codification des autres grandes variables de l'enquête : activité, diplôme.

Le code renvoyé par *Sicore* est sur 8 positions :

- Positions 1 à 3 : renvoie au niveau du diplôme (ex : 320 = BTS, 500 = CA, etc.).
- Position 4 : zéro.
- Positions 5 à 7 : spécialité selon le code NSF.
- Position 8 : lettre.

En définitive, la codification des diplômes s'est faite de manière distincte et successive pour le diplôme de sortie, le baccalauréat ou les autres diplômes. La procédure intègre une part de codification totale par *Sicore*, une part de codification partielle et le restant concerne la codification purement manuelle.

- Phase 1 : une *codification totale* qui repose sur une codification du libellé complet (diplôme et spécialité) pour lequel *Sicore* renvoie un code qualifié de bonne qualité. Dans ce cas, il n'y a pas de reprise manuelle (sauf cas particuliers nécessitant une révision à la marge du code).
- Phase 2 : une *codification partielle* qui renvoie à une codification en deux temps, pour les libellés non qualifiés lors de la première phase de codification sur le libellé complet. Dans ce cas, le libellé est allégé et ne comprend que le diplôme. La spécialité est donc à coder manuellement *a posteriori*.
- Phase 3 : *codification manuelle* des libellés rejetés par *Sicore* lors des deux phases précédentes.

La codification manuelle de la spécialité en phase 2 et du diplôme et spécialité en phase 3 sont faites dans la nomenclature NSF et INSEE, soit par appariement avec les libellés déjà codés pour la recherche de correspondance, soit manuellement avec chaque libellé qui est associé à un code *Sicore* choisi.

À titre indicatif, la codification des diplômes (diplôme de sortie et partiellement le baccalauréat) intègre dans un premier temps les répondants en post-initial qui sont exclus lors de la définition du plus haut diplôme. À raison, le module parcours scolaire ne leur a pas été posé.

### *Codification du diplôme de sortie*

Le diplôme de sortie est celui transmis par l'établissement de sortie ou celui renseigné et/ou complété par l'individu. Sont codés ici l'ensemble des diplômes de sortie, qu'ils soient réalisés dans le cadre d'une année terminale ou non, qu'il soit obtenu par l'individu ou non.

55 % des diplômes de sortie ont été codés par *Sicore* sur la base du libellé compilant le diplôme et la spécialité (phase 1). Un second passage dans *Sicore* a été effectué pour les diplômes restants avec uniquement le libellé de diplôme (phase 2). Ainsi, parmi les 45 % restants, 38 % ont été codés par *Sicore*. Les autres diplômes ont donc été codés manuellement (phase 3).

**Tableau 40 • Rapport de codification *Sicore* – diplôme de sortie**

| État de codification | % |
|---|---|
| Codification totale | 55 |
| Codification partielle | 17 |
| Codification manuelle | 28 |

### *Codification du baccalauréat*

Les données autour du baccalauréat sont collectées en enquête (questions Q35 et/ou BB39). Elles peuvent également être issues des fichiers SISE. En cas d'informations en doublon, c'est l'information déclarée par l'individu qui est conservée.

En post-codage, l'information du diplôme de sortie, lorsqu'il s'agit d'un baccalauréat, est intégré. La procédure de codification est la même que pour le diplôme de sortie : passage dans *Sicore* avec diplôme et spécialité, puis pour les diplômes en échec, nouveau passage dans *Sicore* avec le diplôme uniquement et enfin codification manuelle totale pour les diplômes restants et partielle pour ceux de la phase 2.

**Tableau 41 • Rapport de codification *Sicore* – diplôme du baccalauréat**

| État de codification | % |
|---|---|
| Codification totale | 80,17 |
| Codification partielle | 3,41 |
| Codification manuelle | 16,42 |

### *Codification des autres diplômes*

Pour rappel, l'individu déclare les diplômes qu'il a obtenus en dehors de son diplôme de sortie. Dans le cas où plusieurs diplômes sont déclarés, c'est le plus élevé qui sera plus précisément décrit (complément sur le libellé du diplôme si besoin de précision et la spécialité). Après vérification de la procédure de détermination du plus haut diplôme, des erreurs ont été décelées :

- La modalité 4 de la variable Q39CA est absente du CATI, 159 individus sont concernés par cette erreur et ont donc des valeurs manquantes.

- Une erreur de détermination du plus haut diplôme a été relevée : 465 individus pour les diplômes du secondaire (Q39CA) et 525 individus pour les diplômes du supérieur (Q39HA). Pour l'essentiel de ces individus, les spécialités renseignées ne sont pas celles du diplôme qui aurait dû être décrit. Sont concernés 366 individus en Q39C2 (79 %) et 525 pour Q39I2 (100 %).

- Pour 224 individus, la précision du diplôme lorsque ce dernier est encore approximatif (exemple : Q39C = 05 « Un autre diplôme de ce niveau ») est manquante. La codification de ces diplômes est donc rendue compliquée.

Ainsi, une nouvelle variable diplôme a été créée qui récupère la bonne information. Concernant la spécialité, le choix a été fait de la conserver, qu'il y ait ou non erreur sur le diplôme. En effet, l'hypothèse sous-jacente est celle d'une proximité des spécialités de formation entre diplômes pour un même parcours scolaire pour l'essentiel des individus.

Compte tenu de la qualité des libellés disponible, la codification s'est faite en seulement deux étapes : codification totale et codification manuelle.

**Tableau 42 • Rapport de codification *Sicore* – autres diplômes**

| État de codification | % |
|---|---|
| Codification totale | 59,64 |
| Codification partielle | 40,36 |

### *Attribution d'un code pour chaque diplôme afin de les hiérarchiser*

L'attribution d'un code *Sicore* ou sa recomposition manuelle se heurte à la difficulté de ne pouvoir coder certains libellés, car trop vague pour la plupart (couple libellé de diplôme + libellé de la spécialité). Lorsque c'est le cas pour le diplôme, le libellé ne sera pas codé. Dans le cas où un code diplôme est attribué et en cas d'échec de la codification de la spécialité, il est conservé car intervient dans la détermination du plus haut diplôme. S'il s'agit au final du plus haut diplôme détenu par l'individu, des variables annexes seront utilisées pour tenter d'identifier une spécialité. Lorsqu'en toute fin de codification, la spécialité du plus haut diplôme n'est pas reconnue, le code de la spécialité du diplôme de sortie est affecté, que ce dernier soit obtenu ou pas.

Le code *Sicore* est décomposé pour distinguer le diplôme de la spécialité. À partir de ces deux informations, le niveau de diplôme est construit sur deux positions (utilisation de la table Passagedip). Pour la spécialité, est également attribuée une lettre L/M (Lettre/Mathématique) ou I/T (Industriel/Tertiaire) selon le code NSF et le diplôme. Ainsi, un premier codage est déterminé pour chaque type de diplôme. La nomenclature choisie, en 21 positions, est la suivante :

00 = NON-DIPLÔMÉ
02I = CAP-BEP-MC INDUSTRIEL
02T = CAP-BEP-MC TERTIAIRE
03I = BAC PRO TECHNO INDUSTRIEL
03T = BAC PRO TECHNO TERTIAIRE
04 = BAC GÉNÉRAL
05 = BAC+2 SANTÉ SOCIAL
06I = BTS-DUT INDUSTRIEL
06T = BTS-DUT TERTIAIRE
07I = AUTRE BAC+2 INDUSTRIEL
07T = AUTRE BAC+2 TERTIAIRE
08 = BAC+3 SANTÉ SOCIAL
09 = LICENCE PRO
10L = AUTRE BAC+3 LSH GESTION DROIT
10M = AUTRE BAC+3 MATHS SCIENCE TECHNIQUE
11 = BAC+4
12L = BAC+5 LSH GESTION DROIT
12M = BAC+5 MATHS SCIENCE TECHNIQUE
13 = ÉCOLE DE COMMERCE
14 = INGÉNIEUR
15 = DOCTORAT

Pour chaque diplôme codé, un niveau agrégé est également déterminé à partir de la nomenclature précédente (sans diplôme, niveau CAP/BEP/MC, niveau bac, niveau bac+2 à bac+5 et doctorat).

### *Détermination du plus haut diplôme*

(Ne sont conservés ici que les individus dans le champ Céreq)

La détermination du plus haut diplôme résulte d'une hiérarchisation de l'ensemble des diplômes obtenus par l'individu. Si plusieurs diplômes de niveau équivalent sont identifiés, parmi lesquels est inclus le diplôme de sortie, c'est ce dernier qui est défini comme étant le plus haut diplôme de l'individu. La priorité établie est la suivante : le diplôme de sortie, puis les autres diplômes et enfin le baccalauréat. La nomenclature du plus haut diplôme utilisée est en 30 positions :

01 = NON-DIPLÔMÉ
02I = CAP-BEP-MC INDUSTRIEL
02T = CAP-BEP-MC TERTIAIRE
03I = BAC PRO-BT-BP INDUSTRIEL
03T = BAC PRO-BT-BP TERTIAIRE
04I = BAC TECHNO INDUSTRIEL
04T = BAC TECHNO TERTIAIRE
05 = BAC GÉNÉRAL
06I = BTS-DUT INDUSTRIEL
06T = BTS-DUT TERTIAIRE
07I = AUTRE BAC+2 INDUSTRIEL
07T = AUTRE BAC+2 TERTIAIRE
08 = BAC+2/3 SANTÉ SOCIAL
09L = LICENCE PRO LSH GESTION DROIT
09M = LICENCE PRO MATHS SCIENCE TECHNIQUE
10L = L3 LSH GESTION DROIT
10M = L3 MATHS SCIENCE TECHNIQUE
11L = AUTRE BAC+3 LSH GESTION DROIT
11M = AUTRE BAC+3 MATHS SCIENCE TECHNIQUE
12L = BAC+4 LSH GESTION DROIT
12M = BAC+4 MATHS SCIENCE TECHNIQUE
13L = M2 LSH GESTION DROIT
13M = M2 MATHS SCIENCE TECHNIQUE
14L = AUTRE BAC+5 LSH GESTION DROIT
14M = AUTRE BAC+5 MATHS SCIENCE TECHNIQUE
15 = BAC+5 ÉCOLE DE COMMERCE
16 = INGÉNIEUR
17 = DOCTORAT SANTÉ
18L = DOCTORAT HORS SANTÉ LSH GESTION DROIT
18M = DOCTORAT HORS SANTÉ MATHS SCIENCE TECHNIQUE

Cela revient à réaliser des ajustements par rapport à la nomenclature en 21 positions vue précédemment.

## 7.3.2. La codification du secteur d'activité de l'établissement employeur

L'activité principale de l'établissement employeur est codée selon la nomenclature NAF révision 2 (nomenclature des activités économiques en vigueur en France depuis le 1er janvier 2008) en 88 divisions. La collecte de cette information est possible *via* une recherche de l'établissement employeur dans un menu dont le code NAF est connu. Dans le cas contraire, un questionnement est prévu, construit autour de la nomenclature officielle selon l'arborescence par niveau. À défaut, l'enquêté a la possibilité de déclarer en clair l'activité principale de son employeur.

Le processus de codification appliqué est multiple et séquentiel. En effet, une première opération consiste en une imputation directe, la seconde *via* une imputation par codage automatique avec l'outil *Sicore* activité et enfin pour les rejets, une codification manuelle.

### *Imputation directe du code NAF*

Cette codification est possible à l'aide de deux modèles d'imputation.

#### *Imputation à partir d'une table de passage*

L'articulation du module de questionnement sur l'activité intègre une base de noms d'établissements contenant l'ensemble des grandes entreprises du *CAC 40* mais aussi des noms génériques d'administrations dont l'activité principale est facilement identifiable. Cette base est mise à jour au lancement de chaque nouvelle enquête, alimentée également des noms d'établissements les plus cités par les individus lors de la dernière

collecte. Ce menu a un double objectif ; d'une part de détecter les établissements dont le code NAF est connu et d'autre part de réduire le questionnement donc le temps de passation de l'enquête.

Ici, il s'agit simplement de récupérer l'information dans notre base de données des établissements. Une vérification a tout de même été réalisée pour valider le code NAF imputé. Cette étape a permis la codification de 42 % de l'ensemble des situations d'emploi.

***Imputation résultant des réponses aux questions fermées***

La refonte du questionnement autour de l'activité permet d'identifier l'activité principale de l'établissement employeur et facilite la détermination du code NAF en fonction des réponses de l'individu. À partir de l'arborescence par niveau de la nomenclature en 88 divisions (Annexe 6), six grandes familles d'activité ont été créées. Pour chacune d'entre elles, le positionnement de l'enquêté a plus ou moins été difficile.

- *Agriculture* : L'ensemble des répondants dont l'activité principale de l'employeur est en lien avec l'agriculture se positionne plutôt facilement. 100 % de codification à cette étape lorsque l'individu s'est bien positionné à la première question. Dans cette catégorie, la modalité « autre » n'a pas été proposée.

- *Industrie* : Pour cette classe, trois possibilités possibles d'arriver à la question ouverte avec tout au long du questionnement la collecte d'informations complémentaires sur l'activité. À savoir ; si l'activité est plutôt en lien avec les industries extractives ou manufacturières et si cela concerne le domaine de la fabrication.

- *Énergie et eau* : Peu d'individus se sont retrouvés dans la modalité « autre ». Le taux de codification est plutôt satisfaisant à ce niveau.

- *Constructions, BTP* : Cette catégorie a généré un certain nombre de déclarations en clair. Les termes « génie civil » plutôt incompris et le manque de précision dans la modalité « travaux de construction spécialisés » ont dû conduire l'individu à préciser.

- *Commerce* : Les individus se positionnent plutôt facilement dans cette classe, mais une partie s'est positionnée dans la modalité « autre » pour définir davantage l'activité.

- *Services* : Cette famille est la plus dispersée et est celle où l'information collectée en clair est la plus importante. En effet, lorsque l'individu indique une activité en lien avec des services du type « hébergement et restauration », « transports, activités spécialisées », « scientifiques et techniques », « santé humaine et action sociale » ou « autres services » alors une précision supplémentaire lui ait demandé.

L'opération d'imputation directe *via* les questions fermées représente 10 % de la codification des situations d'emploi.

*Remarque* : conscients de la difficulté à collecter cette information et du manque de clarté de certaines questions du module. Une réflexion est menée autour des fichiers sirène avec leur utilisation embarquée *via* un menu en autocomplétion.

### *Imputation par codage automatique avec l'utilisation de Sicore activité*

Chaque individu se voit proposer le module de questionnement sur l'activité principale de l'établissement employeur dès lors que le nom de son établissement est absent dans le menu des entreprises. Si l'individu ne parvient pas à se positionner précisément, différents chemins mènent à une question ouverte qui demande explicitement l'activité en clair. Selon le parcours de chacun dans le module, voici les formulations possibles de la question :

L'établissement employeur <nom1> fabriquait principalement quel produit .....1

L'établissement employeur <nom1> vendait principalement quel service .....2

L'établissement employeur <nom1> vendait principalement quel produit .....3

Que faisait principalement l'établissement employeur <nom1> .....4

Quelle était l'activité de votre entreprise.....5

L'énonciation de cette question est très importante, car elle fournit déjà des éléments sur la nature de l'activité. Plusieurs notions sont utilisées par *Sicore* lors de la codification automatique : fabrication/vente, produit/service.

L'outil *Sicore*, combiné à l'environnement d'activité, tente de mettre en correspondance le libellé de l'activité en clair déclaré par l'individu et les informations présentes dans la nomenclature à l'aide d'un algorithme (aucune variable annexe à fournir). Le nom et les données collectées autour de l'établissement employeur ne sont pas utilisés lors de cette étape de codification.

Avant de lancer *Sicore*, le fichier brut de données est corrigé des erreurs d'orthographe. Une procédure de correction manuelle est lancée (avec l'utilisation d'Excel).

**Encadré 5 • SICORE Environnement activité (millésime 2012)**

Le logiciel *Sicore* (système informatique de codage par reconnaissance) permet de coder l'activité en NAF rév. 2. Cet environnement de codage est communément appelé *Sicape* (*i.e. Sicore* APE). Le codage n'utilise pas de variables annexes ; il s'appuie uniquement sur le libellé d'activité déclaré. Cet environnement *Sicore* a la particularité de retourner plusieurs codes APE, avec les probabilités associées. Pour cette raison, *Sicape* est utilisé sur le poste des enquêteurs dans les enquêtes INSEE en face-à-face, car l'enquêté peut choisir le meilleur des codes proposés par *Sicape*. Mais en tenant compte des probabilités, on peut aussi utiliser *Sicape* en aval (traitement automatique).

Pour l'enquête Génération 2013, des règles de décisions ont été définies afin de repérer le meilleur écho.

- 1re règle : si *Sicore* fournit 1 SEUL écho dont la probabilité est supérieure à 40, l'écho en 5 positions est choisi.
- 2e règle : si *Sicore* fournit des échos multiples :
  - Sélection des 2 premières positions des différents échos
  - Calcul de la somme des probabilités des échos identiques sur 2 positions successives (premier écho inclus).

Dans ce deuxième cas, si la somme des probabilités est supérieure à 40, l'écho sur 2 positions est choisi.

Le codage en automatique comptabilise 35 % des situations d'emploi codées à partir du choix du bon écho. *Sicore* peut proposer jusque cinq codes pour une activité à définir. Le code NAF (ou APE) sélectionné est déterminé à partir d'un seuil de probabilité maximum qui garantit sa fiabilité (Annexe 7).

Malgré le degré de précision du codage automatique par *Sicore* qui fournit un code NAF en cinq positions et du fait de la difficulté à coder les rejets, la livraison finale de la codification de l'activité est définie en deux positions selon la nomenclature choisie par le Céreq sur 88 divisions.

### *Codification manuelle*

Pour les situations d'emploi dont le secteur d'activité n'a pas été codé, soit les rejets de la codification automatique, un codage manuel est prévu :

- Une codification réalisée par un prestataire à partir de la raison sociale déclarée, de l'information décrite en clair de l'activité et des réponses aux questions fermées (EP2A à EP10). L'idée est de fournir toutes les données disponibles en lien avec l'établissement employeur pour permettre au prestataire de le qualifier (13 % des situations d'emploi).

- Une phase de recopie du code NAF est effectuée si l'individu déclare dans le calendrier d'activité avoir travaillé plusieurs fois dans la même entreprise dans la période d'observation (variable utilisée : RETOU).

**Tableau 43 • Rapport de codification de l'activité principale NAF**

| Mode de codification du code NAF | % |
|---|---|
| Imputation directe | 52,35 |
| Imputation par codage automatique | 34,77 |
| Codification manuelle | 12,88 |

## 7.3.3. La codification des professions et des catégories socioprofessionnelles

La profession est codée selon la nomenclature des professions et catégories socioprofessionnelles (PCS) dans sa version datant de 2003. La PCS comporte trois niveaux d'agrégation emboîtés : 8 groupes socioprofessionnels, 24 ou 42 catégories socioprofessionnelles et 486 professions.

Le codage de la PCS est basé sur le libellé de la profession demandé à l'individu, complété par d'autres informations relatives à la position professionnelle et aux caractéristiques de l'établissement dans lequel il travaille (appartenance au secteur public, taille de l'entreprise, secteur d'activité).

La codification de la profession selon la nomenclature officielle, à l'aide de l'outil *Sicore*, nécessite la création de variables annexes. Certaines sont disponibles dans l'enquête et d'autres sont à recomposer. Plusieurs étapes de préparation des données sont nécessaires avant l'utilisation de ce logiciel.

**Encadré 6 • SICORE Environnement PCS 2013**

Le logiciel *Sicore* (système informatique de codage par reconnaissance) permet d'effectuer un codage automatique des professions à partir d'un questionnement en face à face ou bien à partir d'un fichier en entrée (fonctionnement en « *batch* »). C'est le fonctionnement qui a été utilisé pour l'enquête Génération 2013.

*Sicore* en *batch* nécessite en entrée un fichier en format texte qui comporte les variables suivantes (Annexe 8 : Définition des variables et des modalités)

- Identifiant.
- Libellé de la profession.
- STATUT : statut dans l'emploi (3 modalités).
- PUB : statut de l'établissement employeur (5 modalités).
- CPF : classification professionnelle ou qualification (10 modalités).
- FN : fonction principale (9 modalités).
- NBS : nombre de salariés employés (4 modalités).
- NAF4 : activité principale de l'établissement sur 4 caractères.
- NAF2 : activité principale de l'établissement sur 2 caractères.
- T : taille de l'entreprise (4 modalités).
- OPA : orientation des productions agricoles (8 modalités).
- DEP : département.
- SAU: surface agricole utilisée (88 modalités).
- S : sexe.
- SP : statut précaire.
- STRE : emploi actuellement (oui ou non).

Il procède en trois étapes successives :

- Étape 1 : analyse le libellé de la profession.
- Étape 2 : introduit des variables annexes, le cas échéant (cf. la liste ci-dessus).
- Étape 3 : livre un code PCS avec un indicateur de fiabilité.

Dans le cadre de l'enquête Génération 2013 l'utilisation de la NAF2 a été privilégiée (la NAF4 n'étant que très partiellement présente) et les variables OPA, DEP, SAU et STRE non utilisées, car informations non disponibles dans l'enquête. Toutes les autres variables ont été intégrées en variables annexes.

Parmi les variables annexes collectées dont un traitement a été réalisé à la fois pour le codage et pour les bases d'exploitation :

### Recomposition du libellé de profession (EMPL)

La collecte de l'intitulé de la profession est faite à l'aide d'un menu qui répertorie un ensemble de professions connues. Afin de capter une information précise, certaines professions sont « topées » lorsque celles-ci sont définies comme floues. Une question subsidiaire est alors posée afin de collecter le domaine de l'emploi. À l'inverse, pour les professions manifestes, les questions de caractérisation sont masquées. Pour exemple, la question sur la fonction professionnelle n'est pas posée aux professions du bâtiment.

Une étape de correction orthographique et de normalisation est réalisée pour créer le fichier de données pour *Sicore*. Il convient lors de la recomposition de l'information disponible sur la profession de veiller à :

- Laisser un seul espace entre les mots
- Supprimer les accents ou trémas et les caractères tels que (, – /. * "…
- Supprimer également les mots redondants en double ou en triple

Pour compléter, des recopies sont faites lorsque l'individu indique ne pas avoir changé de profession dans son emploi (situation d'emploi d'une durée de plus de 6 mois). La variable livrée pour l'exploitation de l'enquête est dénommée EMPL.

### Création des variables de statut de l'établissement employeur (PUB/NATENTR)

La variable PUB détermine si une entreprise appartient au secteur public ou privé. Pour coder cette variable, la table de passage des entreprises est une nouvelle fois utilisée. En plus de l'information sur l'activité (NAF), existe une variable qui détermine si l'entreprise relève du domaine public ou privé. Si l'établissement employeur est absent du menu, la question est alors directement posée à l'enquêté.

PUB est une variable annexe utilisée par *Sicore*. En revanche, la variable NATENTR, est une variable d'intérêt fournie dans les bases d'exploitations. Elle contient une information plus fine sur le statut de l'établissement : Éducation nationale, armée, autre État, collectivités territoriales, hôpitaux, divers secteur public, sécurité sociale, entreprises publiques nationalisées, secteur privé ou indéterminé.

### Création de la variable de classification professionnelle (CPF/POSPRO)

La position professionnelle permet de qualifier le niveau de l'emploi occupé. Le libellé de profession ne suffit pas à le déterminer. Par conséquent, un questionnement spécifique est posé pour obtenir une information précise sur la qualification de la profession. Ce questionnement est double selon le statut de l'emploi et permet :

Dans le cas d'un emploi salarié :

- de préciser le niveau (ingénieur, cadre, employé, etc.) ;
- de déterminer son degré de qualification (technicité).

Dans le cas d'un emploi de fonctionnaire :

- de préciser le niveau à partir de la catégorie hiérarchique (A, B, C ou autre) ;
- de déterminer son degré de qualification (technicité).

La variable CPF est utilisée pour le codage automatique dans *Sicore* et POSPRO est livrée dans les bases finales.

### Création de la variable de fonction (FN/FONCT)

La fonction principale est utilisée pour représenter au mieux l'emploi occupé notamment, car la description de certaines professions n'est pas assez claire et précise. Cette information est définie comme suit :

Production, fabrication, chantiers .....01

Installation, réparation, maintenance.....02

Nettoyage, gardiennage, entretien ménager .....03

Manutention, magasinage, logistique .....04

Secrétariat, saisie, accueil .....05

Gestion, comptabilité .....06

Commerce, technico-commercial .....07

Études, recherche et développement, méthodes .....08

Enseignement, soins aux personnes.....09

Autres fonctions .....10

La variable FN est utilisée pour le codage automatique dans *Sicore* et FONCT est livrée dans les bases finales.

### Codage de la profession selon la nomenclature (PCS)

À l'issue de la création des variables annexes, il s'agit de lancer la codification de la PCS. *Sicore* livre un fichier de résultats contenant les professions codées, les professions non codées (principalement en raison d'un libellé non reconnu) et celles codées avec réserve. Pour cette dernière catégorie, *Sicore* fournit un code PCS mais signale qu'il lui manque une information annexe pour valider le code. À cela s'ajoute, pour chaque libellé de profession, un indice de confiance qui donne la précision sur l'utilisation d'une ou plusieurs variables annexes.

Dans le cadre de cette enquête, le choix s'est porté sur la sélection des PCS codés sans réserve et avec un indice de confiance défini tel qu'au moins une variable annexe demandée n'est pas renseignée.

Les professions qui n'ont pas été codées (selon les critères choisis) par cette procédure ont été transmises à un prestataire externe qui a procédé à une codification manuelle.

**Tableau 44 • Rapport de codification de la profession PCS**

| Mode de codification du code PCS | % |
|---|---|
| Imputation par codage automatique | 69,79 |
| Codification manuelle | 30,21 |

## 7.4. Post-codification manuelle

### 7.4.1. Traitement des questions ouvertes ou semi-ouvertes

Certaines variables présentent une modalité « Autre, précisez » qui ouvre la possibilité d'une déclaration en clair. Dans cette enquête, une cinquantaine de variables font l'objet de recodifications. Toute réponse en clair donnée par un individu a été analysée et reclassée dans une des modalités proposées (si possible), soit a donné lieu à la création d'une nouvelle modalité lorsque la réponse est citée de manière récurrente (avec un minimum de 50 observations). Les réponses n'ayant pas pu être reclassées selon les précédents critères sont restées dans la modalité « Autre ».

Les variables concernées par la création de modalités post-collecte sont signalées dans le dictionnaire des variables. Une indication « modalité ajoutée » est notée à la suite de la modalité.

#### *Le cas des questions à choix multiples*

Dans les enquêtes Génération, les questions à choix multiples sont dichotomisées. Pour une question donnée, chacune des modalités est déclinée en variable de type booléen.

Certaines de ces questions à choix multiples offrent également la possibilité de collecter une réponse en clair. Après la phase de codification, si de nouvelles modalités sont proposées alors sont créées autant de variables dichotomiques. Pour pouvoir les identifier, la mention « modalité ajoutée » apparait cette fois-ci dans le libellé de la variable et également dans le dictionnaire.

#### *L'exemple du contrat de travail*

Variable essentielle de l'enquête, elle est issue de la combinaison de plusieurs questions semi-ouvertes. Ce questionnement est mis à jour à chaque enquête des nouveaux contrats de travail ou dispositifs spécifiques crées dans la période d'observation. Cependant, les individus ont parfois des difficultés à se positionner dans les modalités proposées.

Pour la situation de premier emploi, une variable de synthèse est construite pour définir le contrat de travail à l'embauche. Pour la situation d'emploi à la date de l'enquête, une voire deux variables de synthèse sont construites : contrat de travail à l'embauche (lorsque la durée de l'emploi est supérieure à 6 mois) et celle contenant le contrat de travail à la date d'interrogation (si changement).

Pour définir le type de contrat, une proportion importante des répondants ont recours à la modalité « Autre, précisez » afin de déclarer une information plus précise. La plupart de ces déclarations sont reclassées dans les modalités proposées au moment de l'enquête. Quelques verbatim pour exemple, « *CDD de remplacement* », « *contrat à durée indéterminée intermittent* », « *contrat de stagiairisation* »…

### 7.4.2. Géolocalisation des données

L'enquête Génération prévoit la collecte de la commune de résidence à différents points du parcours de l'individu. Notamment, la commune de résidence en sixième, au baccalauréat (le cas échéant) et à la date de l'enquête en 2016. La commune du lieu de travail à la date d'interrogation est également demandée (si l'individu est en emploi).

Les communes sont repérées à partir de leur code officiel géographique, ce qui permet de mobiliser les différents zonages administratifs (région, département) et d'études développés par l'INSEE (unités urbaines, aires urbaines, zone d'emploi). Les nomenclatures actualisées (en 2013) ont été intégrées pour cette nouvelle enquête Génération.

Pour le zonage en aires urbaines (ZAU), les informations retenues sont la taille de l'aire urbaine et le type d'espace auquel la commune appartient (pôles urbains, couronne périurbaine, commune multipolarisée, communes isolées).

Chaque commune est caractérisée au plus précis pour sa géolocalisation ainsi les informations d'intérêt pour l'exploitation sont déclinées dans plusieurs variables. Le libellé de chacune d'elles est formé par le couple : « temporalité du lieu de résidence + type du code géographique ».

Pour exemple, la codification de la variable de commune de résidence en sixième a donné lieu à neuf variables d'intérêt :

- SIXIEMEDEP : Département.
- SIXIEMEREG : Code de la région.
- SIXIEMEREGIONF : Libellé de la région.
- SIXIEMECATAEU : Catégorie d'espace du ZAU 2013.
- SIXIEMETAU : Tranche d'aire urbaine 2013.
- SIXIEMESTATUTUU : Statut de communes selon la définition des unités urbaines 2013 (ville centre, banlieue, isolé, rural).
- SIXIEMETYPEUU : Type de communes selon la définition des unités urbaines 2013 (rural/urbain).
- SIXIEMEZE : Code de la zone d'emploi.
- SIXIEMELIBZE : Libellé de la zone d'emploi.

À noter que seule la géolocalisation obtenue ne permet pas l'identification de l'individu. De plus la commune reste une variable confidentielle est n'est pas livrée en tant que telle sans passage au comité du secret.

### 7.4.3. Redressement des salaires (primes incluses)

Dans le cadre de cette enquête, pour chaque situation d'emploi, le salaire net mensuel est collecté à l'embauche lorsqu'il s'agit d'un premier emploi. S'il s'agit d'une situation d'emploi à la date de l'enquête ; le salaire net mensuel perçu à l'embauche (lorsque la durée de l'emploi est supérieure à 6 mois) et celui à la date d'interrogation (si changement).

Cette question sur les salaires est « sensible », l'enquêté a le choix soit de déclarer le montant directement en clair, soit de donner un montant par le biais de tranches de salaire. Toutefois, il a la possibilité de ne pas répondre s'il refuse ou s'il ne connaît pas le montant même approximatif de la rémunération mensuelle.

Seuls les salaires des individus avec un statut d'aide familial ne sont pas collectés.

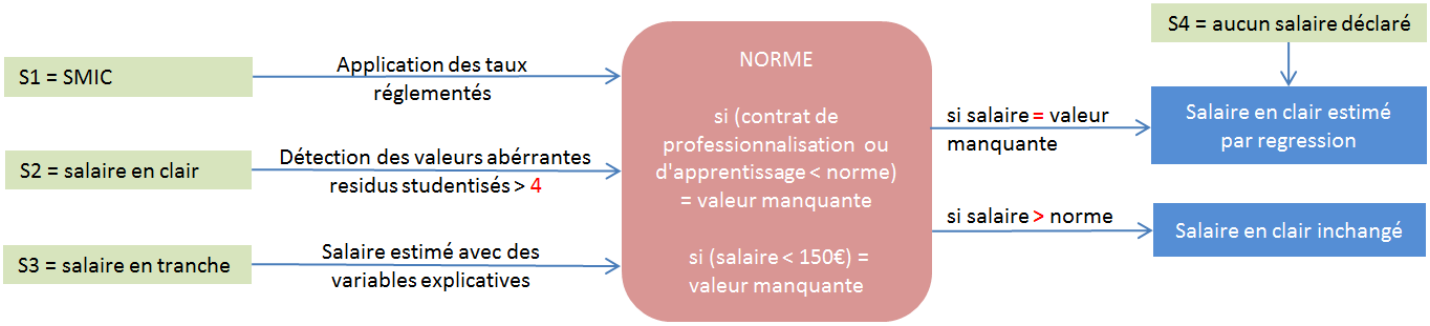
**Tableau 45 • Mode de déclaration du montant du salaire**

| Salaire déclaré | À l'embauche | À la date de l'enquête |
|---|---|---|
| Réponse en clair | 73,24 | 81,61 |
| SMIC | 9,64 | 2,21 |
| Réponse en tranche | 15,72 | 7,17 |
| Non-réponse | 1,40 | 9,01 |

Les salaires des situations d'emploi liées aux individus à leur compte sont différenciés dans leur traitement, car les niveaux de rémunérations ne sont pas comparables. Pour ces derniers, lorsque le montant du salaire est déclaré en clair, l'information brute est directement livrée dans les bases d'exploitation. Dans le cas le contraire, le salaire est recalculé. Pour les post-initiaux dans le champ Céreq, la même procédure est appliquée.

Le traitement des salaires est distinct selon le mode de réponse. Le processus de redressement suit la logique suivante :

**Figure 10 • Gestion des salaires**




### *Procédure de redressement*

À l'aide de modèles de régression, une estimation du salaire est faite pour les individus ayant déclaré un salaire en tranche et pour ceux dont le salaire a été invalidé (seuil de la norme) ou manquant.

L'option *stepwise* a été utilisée pour l'ensemble des modèles estimés. La régression pas à pas est une méthode d'ajustement des modèles de régression dans laquelle le choix des variables prédictives est effectué par une procédure automatique. L'objectif est d'obtenir une meilleure variabilité du salaire. Comme pour l'enquête Génération 2010, deux modèles distincts sont construits pour estimer les salaires manquants en lien avec la quotité de travail ; un pour le temps plein et l'autre pour le temps partiel. En revanche, pour les salaires déclarés en tranche, l'estimation est réalisée à partir d'un seul modèle.

Les variables utilisées dans les modèles sont :

- Sexe.
- Âge de l'enquêté.
- Niveau de sortie : non-diplômé, secondaire, bac+2, bac+3/4, bac+5.
- Plus haut diplôme obtenu en 15 positions.
- Spécialité de formation : général, industriel ou tertiaire.
- Type de contrat de travail : indépendant, fonctionnaire, cdi, cdd, contrats aidés.
- Catégorie socioprofessionnelle : ouvrier, profession intermédiaire, cadre, employé, autre.
- Ancienneté.
- Expérience : *non sélectionnée par l'option stepwise.*
- Origine des parents : *non sélectionnée par l'option stepwise.*
- Région de l'entreprise : Île-de-France, autre région, étranger.
- Taille de l'entreprise : moins de 20 salariés.
- Activité de l'entreprise : NAF en 8 postes.

Plusieurs étapes mises en œuvre pour l'imputation d'environ 23,6 % de l'ensemble des salaires déclarés :

### *Étape 1 : Détermination des salaires aberrants*

Pour détecter les salaires aberrants, un premier modèle de régression est construit incluant les variables sexe, âge, niveau de sortie, plus haut diplôme obtenu, spécialité de formation, type de contrat de travail, catégorie socio-professionnelle, région et taille de l'entreprise, activité de l'entreprise, ancienneté.

Pour tester l'adéquation du modèle statistique et obtenir une information sur les données dites « aberrantes », une analyse des résidus studentisés est réalisée. Il faut s'attendre à ce que 95 % des valeurs soient comprises entre −2 et +2. Les valeurs, en petit nombre, très éloignées de 0, tels que supérieures à 3 ou 4, font l'objet d'un examen particulier. Elles sont considérées comme des données suspectes.

Le résultat du modèle montre que les résidus studentisés supérieurs à 4 sont à éliminer et par conséquent les salaires correspondants sont exclus du modèle. Cette opération concerne moins de 5 % des situations d'emploi et représente soit le salaire à l'embauche, soit celui à la date de l'enquête.

### *Étape 2 : Imputation des salaires déclarés en tranche*

Lorsque le salaire a été déclaré par le biais des montants en tranche, un modèle de régression mobilisant les variables suivantes a été spécifié : sexe, âge, niveau de sortie, plus haut diplôme obtenu, type de contrat de travail, catégorie socio-professionnelle, région et activité de l'entreprise, ancienneté ainsi qu'une variable qui détermine *la classe de la tranche de salaire*. Ce modèle utilise les observations des individus ayant déclaré un salaire en clair et sont mis en correspondance avec les tranches de salaire. Le salaire est alors redressé dans la classe d'imputation concernée, constitué par le croisement des variables explicatives du modèle et du type d'emploi.

### *Étape 3 : Comparaison des salaires déclarés aux normes en vigueur*

L'ensemble des salaires est rapporté à une *norme* construite à partir du taux horaire du SMIC, selon la période observée, mais aussi en fonction du contrat et de la quotité de travail. Pour les cas particuliers des contrats de professionnalisation et d'apprentissage, des taux réglementés sont introduits dans le calcul de la norme en adéquation avec l'âge et l'avancement dans la formation. Lorsque le montant du salaire s'avère inférieur à cette norme ou inférieur à 150 euros, une imputation est réalisée à l'aide d'une équation de salaire. Ce dernier traitement est aussi appliqué aux individus n'ayant déclaré aucun salaire.

– **Le salaire minimum de croissance.**

Pour faciliter le traitement des salaires, la modalité SMIC est proposée pour un positionnement rapide de l'enquêté. Cela permet notamment d'éviter les déclarations approximatives. Le montant du SMIC est donc calculé post-collecte. La valeur du SMIC a été réévaluée 5 fois au cours de la période d'observation entre novembre 2012 et juillet 2016.

**Tableau 46 • Revalorisation du SMIC entre novembre 2012 et juillet 2016**

| Date de revalorisation | Montant net du SMIC* |
|---|---|
| Novembre 2012 | 1 116,87 |
| Janvier 2013 | 1 120,43 |
| Janvier 2014 | 1 128,70 |
| Janvier 2015 | 1 135,99 |
| Janvier 2016 | 1 141,61 |

* source INSEE

– **Le contrat de professionnalisation (nouveau contrat de qualification).**

Ce contrat s'adresse aux jeunes de 16 à 25 ans, aux demandeurs d'emploi de 26 ans et plus et aux bénéficiaires de certaines allocations ou contrats. Les bénéficiaires de 16 à 25 ans révolus sont rémunérés en pourcentage du SMIC selon leur âge et leur niveau de formation. Les salariés âgés de 26 ans et plus perçoivent une rémunération qui ne peut être ni inférieure au SMIC ni à 85 % du salaire minimum conventionnel.

Pour chaque individu, le calcul du montant du salaire prend en compte l'âge et le niveau de formation. La variable « nisor » (niveau de sortie sur 15 positions) est utilisée comme référence pour le niveau de formation.

**Tableau 47 • Barème de rémunération du contrat de professionnalisation**

| Formation initiale | Moins de 21 ans | 21 ans et plus |
|---|---|---|
| Niveau inférieur au bac | 55 %* | 70 %* |
| Niveau bac et plus | 65 %* | 80 %* |

* pourcentage du SMIC

Ce salaire ne peut être inférieur à 55 % du SMIC pour les bénéficiaires âgés de moins de 21 ans et à 70 % du SMIC pour les bénéficiaires de 21 ans et plus. Ces rémunérations ne peuvent être inférieures, respectivement à 65 % et 80 % du SMIC, dès lors que le bénéficiaire est titulaire d'une qualification au moins égale à celle d'un baccalauréat professionnel ou d'un titre ou diplôme à finalité professionnelle de même niveau.

– **Le contrat d'apprentissage.**

Ce contrat s'adresse aux jeunes de 16 à 25 ans, mais des dérogations à ces limites d'âge sont possibles. L'apprenti perçoit un salaire minimum légal déterminé en pourcentage du SMIC et dont le montant varie en fonction de l'âge du bénéficiaire et de la progression dans le cycle de formation faisant l'objet de l'apprentissage.

**Tableau 48 • Barème de rémunération du contrat d'apprentissage**

| Année d'exécution du contrat | Moins de 18 ans | De 18 ans à moins de 21 ans | 21 ans et plus |
|---|---|---|---|
| 1re année | 25 %* | 41 %* | 53 %* |
| 2e année | 37 %* | 49 %* | 61 %* |
| 3e année | 53 %* | 65 %* | 78 %* |

* pourcentage du SMIC

Au titre de la progression dans un cycle de formation, l'apprenti bénéficie d'une rémunération variant en fonction de l'année d'exécution du contrat. Ainsi, le salaire minimum perçu par l'apprenti correspond au pourcentage du SMIC (ou, dans certains cas, du salaire minimum conventionnel), allant de 25 % à 78 % déterminé en fonction de l'âge et de l'évolution dans le cycle.

Dans l'enquête, la situation d'emploi dans laquelle le répondant déclare être apprenti est supposée correspondre à son lieu d'apprentissage. Par conséquent, pour déterminer l'année d'exécution du contrat, la variable de durée est utilisée par défaut comme paramètre pour définir l'ancienneté dans le contrat. Après imputation, un test de cohérence est réalisé, comme pour l'ensemble des contrats, pour vérifier si le salaire déclaré en enquête est proche du salaire redressé.

***Étape 4 : Imputation des salaires manquants***

Les salaires inférieurs à 150 euros et les salaires manquants sont imputés par un modèle faisant intervenir les variables : sexe, âge, niveau de sortie, plus haut diplôme obtenu, spécialité de formation, type de contrat de travail, catégorie socio – professionnelle, région et taille de l'entreprise, activité de l'entreprise, ancienneté. Ce modèle a été dupliqué, distinguant les salaires à temps partiel et à temps complet. Le modèle de régression explique environ la moitié de la variance totale.

***Étape 5 : Le traitement des primes***

Dans l'enquête, la notion de salaire ne se distingue pas directement des primes. En effet, la question du salaire ne mentionne aucune précision sur les primes. L'enquêté a la possibilité de déclarer un montant de salaire hors primes ou primes incluses. Les primes sont abordées par la suite, notamment la prime de 13e mois. Comme pour les salaires, la déclaration peut être en clair, en tranche ou un refus de répondre.

Le traitement nécessite la gestion de différents types de primes. Pour la prime dite de « 13e mois », le montant annuel collecté est transformé en une prime mensuelle et ajouté au salarié corrigé (le cas échéant).

Pour les « autres primes » mensuelles collectées en clair ou en tranche. La procédure de redressement appliquée est identique que celle du salaire. Un montant moyen par tranche et par statut d'emploi est calculé et imputé à partir des autres primes déclarées en clair.

L'objectif étant de déterminer un montant de salaire net mensuel « primes incluses » pour toutes les situations d'emploi (hormis les jobs de vacances et les emplois de statut aide familiale).

*Remarque* : les variables contenant les informations sur les primes ne sont pas livrées dans les bases d'exploitation.

## 7.5. Anonymisation des données et finalisation

Afin d'assurer la confidentialité des réponses des personnes enquêtées et conformément au principe du secret statistique garanti par la loi de n° 51-711 du 7 juin 1951 sur l'obligation, la coordination et le secret en matière de statistiques, les fichiers de résultats ont été rendus anonymes.

L'anonymisation consiste d'abord à supprimer les informations directement nominatives (nom et prénom des personnes), ainsi que les coordonnées téléphoniques et postales, nécessaires à la réalisation de la collecte de l'enquête. Les localisations géographiques trop fines (en l'occurrence la commune), les noms des entreprises sont également supprimés.

Elle consiste aussi à s'assurer qu'une identification indirecte par recoupement des différentes informations disponibles est impossible. Pour cela, certaines variables ont été supprimées, d'autres ont vu leurs modalités regroupées. En particulier, les informations de localisations géographiques fines ne sont pas diffusées (code commune etc.). Les questions ouvertes sont supprimées, seule est conservée la codification de ces variables.

## 7.6. Les bases exploitables

### 7.6.1. Les bases finales de la Génération 2013 (champ Céreq)

À l'issue des différents traitements post-collecte, les fichiers de résultats de l'enquête 2016 auprès de la Génération 2013 sont livrés en quatre tables.

Ces tables contiennent uniquement les individus dans le champ Céreq. Les individus considérés « hors-champ Céreq » ne sont pas disponibles dans les bases. À l'instar des post-initiaux, également « hors-champ Céreq », qui ont été interrogés dans le cadre de conventionnement avec des partenaires d'extensions et qui figurent dans des tables distinctes.

Pour l'exploitation des bases, un identifiant unique (IDENT) par individu est disponible dans chacune des tables. Également un numéro de situation (NSEQ) est présent, dans les tables des situations d'emploi et de

non-emploi, et permet de retracer chronologiquement le parcours professionnel de chaque répondant. Enfin, dans la dernière table sur les concours existe un numéro (NBC) qui détermine le nombre et l'ordonnancement des concours passés.

**Tableau 49 • Base finale**

| Table | Observations | Données disponibles |
|---|---|---|
| **G13individusvf** | 19 498 répondants | Ensemble des données individuelles de l'enquêté, y compris les variables de synthèse des calendriers d'activité et du mode de cohabitation.<br><br>Pondération.<br><br>731 variables. |
| **G13seqentrvf** | 41 773 situations d'emploi | Ensemble des situations d'emploi décrites dans le calendrier d'activité.<br><br>88 variables. |
| **G13nonempvf** | 26 972 situations de non-emploi | Ensemble des situations de non-emploi décrites dans le calendrier d'activité.<br><br>61 variables. |
| **G13concoursvf** | 6 206 concours décrits | Ensemble des concours décrits.<br><br>84 variables. |

Chaque partenaire d'extension reçoit la version des tables livrées à l'ensemble des chercheurs. Cependant pour les demandes ayant un impact sur le champ Céreq, l'interrogation d'individus en post-initial par exemple, la production de nouvelles tables intégrant les individus de leurs champs respectifs est réalisée. Un jeu de pondération spécifique est également fourni.

### 7.6.2. Les bases comparables 2010 et 2013 (champ Céreq)

Réservée à un usage interne, la création de bases de données à champ comparable a été réalisée à partir des enquêtes à 3 ans des Générations 2010 et 2013. Souvent utilisée dans la publication de Bref de premiers résultats, la comparaison dans le temps permet d'observer l'impact de la conjoncture sur l'insertion professionnelle des jeunes.

Pour ce faire, une restriction de champ est nécessaire ainsi que la mise à jour des jeux de pondération de chaque enquête. En effet, les sortants d'une école supérieure du professorat ou de l'éducation ont été supprimés, car ce champ n'était pas présent dans la Génération 2010.

## 7.7. Format, labels et dictionnaire des variables

Les labels des questions de l'enquête ont été récupérés pour la majeure partie des tables SAS de l'enquête à trois ans, mais aussi du questionnaire.

Nous avons créé un fichier SAS contenant tous les formats des variables de la base dont l'ordre de création est basé sur l'enchaînement des questions du questionnaire d'origine de Génération 2013 à 3 ans. Pour la grande majorité, les formats ont été créés en tenant compte des modifications des modalités de certaines variables du questionnaire, en particulier des questions ouvertes (traitées en amont avec leurs nouvelles modalités), mais aussi de certains modules du questionnaire ajoutés.

En parallèle, un dictionnaire des variables a été créé. Le dictionnaire contient un index papier et un index créé automatiquement par ordre alphabétique avec les numéros de pages en face des noms de variables, qui permet à l'utilisateur grâce au lien hypertexte affecté aux variables de le rediriger directement sur l'information.

# 8. La pondération finale

## 8.1. Le principe général

Pour Génération 2013, un total de 19 498 questionnaires a été collecté auprès d'individus appartenant au champ du Céreq. Pour ces individus, la pondération finale est obtenue en tenant compte des paramètres suivants en amont de la collecte :

- couverture de la base de sondage ;
- probabilité individuelle d'appartenance à l'échantillon principal.

En aval de la collecte :

- probabilité d'entrer en contact téléphonique avec l'individu ou un de ses proches sachant que l'individu appartient à l'échantillon principal ;
- probabilité que l'individu accepte de répondre, sachant que lui-même ou un proche a été contacté ;
- probabilité de terminer le questionnaire sachant que l'individu a accepté de répondre au questionnaire ;
- coefficient de calage sur marges.

La correction du taux de couverture vise à pallier les défauts d'exhaustivité de la base de sondage, qui résultent de la non-réponse de certains établissements à la phase de collecte des fichiers d'inscrits.

La probabilité d'appartenance à l'échantillon est égale, pour les individus de l'échantillon utilisé, à leur probabilité d'être sélectionné dans la phase de tirage de l'échantillon. Pour Génération 2013, l'échantillon de réserve n'a pas été mobilisé et seul l'échantillon principal a été utilisé. La probabilité d'appartenance à l'échantillon est donc égale à la probabilité d'appartenance à l'échantillon principal.

Pour déterminer la probabilité de répondre tout en étant dans le champ et en ayant terminé le questionnaire, trois modèles successifs sont mobilisés.

La probabilité d'entrer en contact téléphonique avec l'individu ou un de ses proches a été calculée pour l'échantillon principal et corrige de la probabilité d'être réellement envoyé en enquête en fonction des coordonnées.

La probabilité d'accepter de répondre est calculée pour les individus de l'échantillon principal avec lesquels le contact a pu être établi.

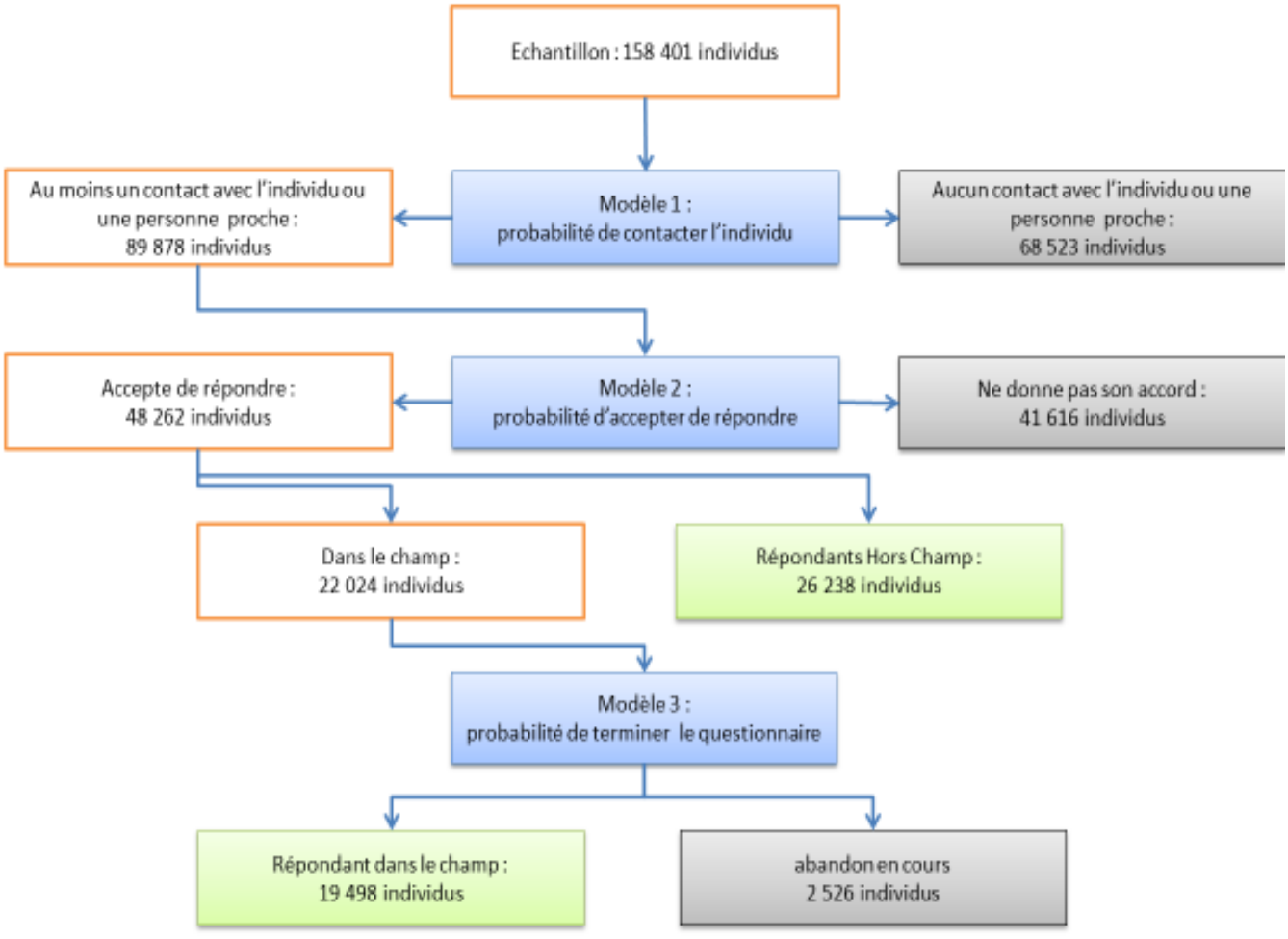
La probabilité de terminer le questionnaire est calculée pour les individus ayant accepté de répondre et qui appartiennent au champ du Céreq.

À partir des probabilités de réponses estimées, des groupes homogènes de non-réponse sont constitués pour estimer une probabilité de répondre au sein de chaque groupe homogène de non-réponse. Cette étape est une étape de robustesse qui permet de se prémunir contre une mauvaise spécification du modèle de correction de la non-réponse.

Après la phase de repondération liée à la modélisation de la non-réponse, un calage des données est effectué sur les effectifs selon le plus haut diplôme croisé avec le sexe. Pour le champ Céreq, le calage est fait à partir de résultats publiés par la DEPP[11] qui mobilisent l'enquête emploi de l'INSEE.

Finalement, le poids d'un individu répondant dans le champ du Céreq est déterminé de la manière suivante :

$$\forall i, pondef(i) = p_{couverture}(i) * p_{échantillonnage}(i) * p_{contact}(i) * p_{accepte\ \ répondre}(i) * p_{ter\,mine\ \ questionnaire}(i) * p_{calage}(i)$$



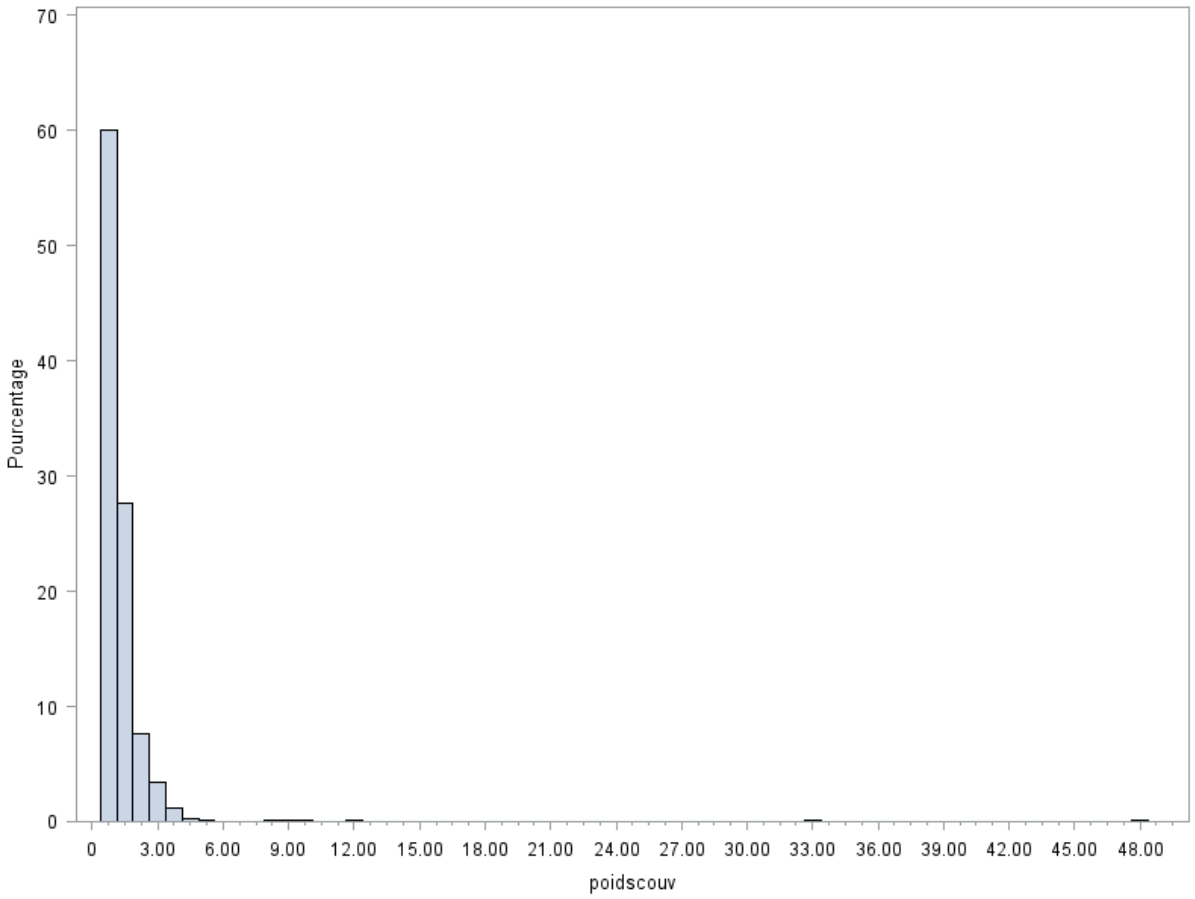

[11] Ouvrage Repères et références statistiques sur les enseignements, la formation et la recherche – RERS 2016 (fiche *Le niveau d'étude à la sortie du système éducatif* en donnée provisoire).

## 8.2. Le poids de couverture

Le calcul du poids de couverture est présenté dans la partie 3.2.1. Ce calcul est réalisé sur l'ensemble des sortants avec l'hypothèse sous-jacente d'une homogénéité des taux de couverture entre les individus dans le champ Céreq et les individus hors-champ.

Pour les 19 498 individus dans le champ de l'enquête et répondant au questionnaire, il est compris entre 1 (dans le cas où la base de sondage est considérée comme exhaustive) et 48,2.

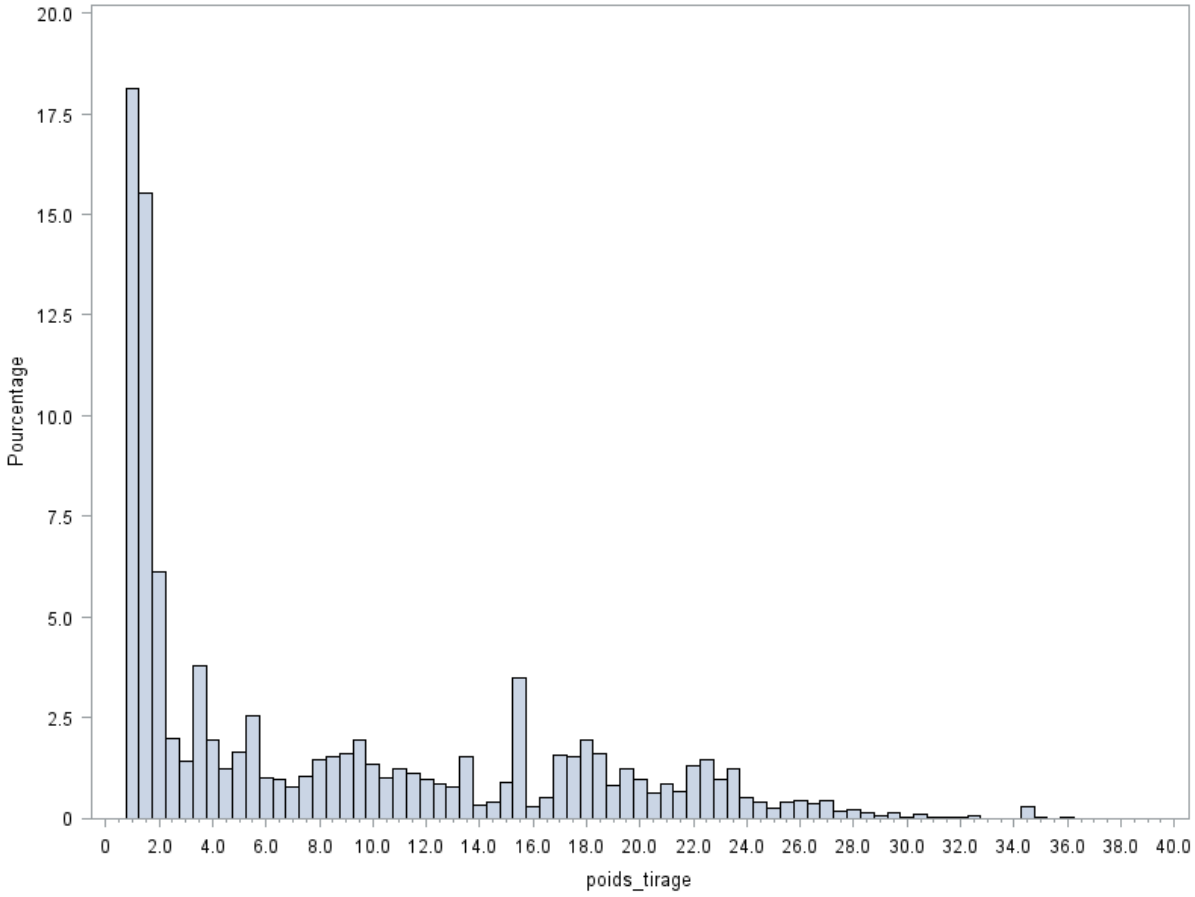
**Figure 11 • Distribution des poids de correction liés au défaut de couverture de la base de sondage**



## 8.3. Le poids d'échantillonnage

Pour les individus dans le champ de l'enquête et répondant au questionnaire, il est compris entre 1 (l'individu est échantillonné d'office) et 35,9.

**Figure 12 • Distribution des poids d'échantillonnage**



## 8.4. Le poids relatif au fait de contacter l'individu ou un proche

### 8.4.1. Modélisation de la probabilité de contact

L'échantillon est constitué de 158 401 individus. Un contact téléphonique a pu être établi pour 89 878 d'entre eux (56,74 %). La probabilité de contact n'est pas indépendante de caractéristiques de l'individu connues dans l'échantillon. On modélise la probabilité de contact grâce à une régression logistique sur l'échantillon.
Les variables retenues dans le modèle sont :

- sexe ;
- le mode d'envoi de la lettre avis ;
- la présence du nom de commune dans l'adresse de l'individu ;
- le niveau d'études ;
- formation effectuée par apprentissage en 2013 ;
- type d'établissement ;
- information sur l'adresse de l'individu en 2013 (QPV ou non) ;
- déménagement entre 2013 et 2016 ;
- indicateur de qualité du numéro de téléphone ;
- présence du mail dans les informations de l'individu ;
- indicatrice de la fiabilité du numéro de téléphone (si donnée par l'établissement de formation et est dans l'annuaire) ;
- nombre d'appels.

Les individus dont les coordonnées téléphoniques disponibles étaient les plus fiables ont été les plus faciles à contacter. Ceux dont l'adresse postale était située en QPV ou était mal renseignée (code postal absent…) ont été moins souvent contactés que les autres. Le type de formation suivie est également lié à la probabilité de joindre l'individu ou un de ses proches.

### 8.4.2. Cohérence de la modélisation

Le modèle présente un $R^2$ ajusté de 0,29 et 76,2 % de paires concordantes.

Sur l'ensemble de l'échantillon, la probabilité moyenne de contact estimée est égale à 0,57 (min = 0,03 ; max = 0,98 ; ET = 0,24).

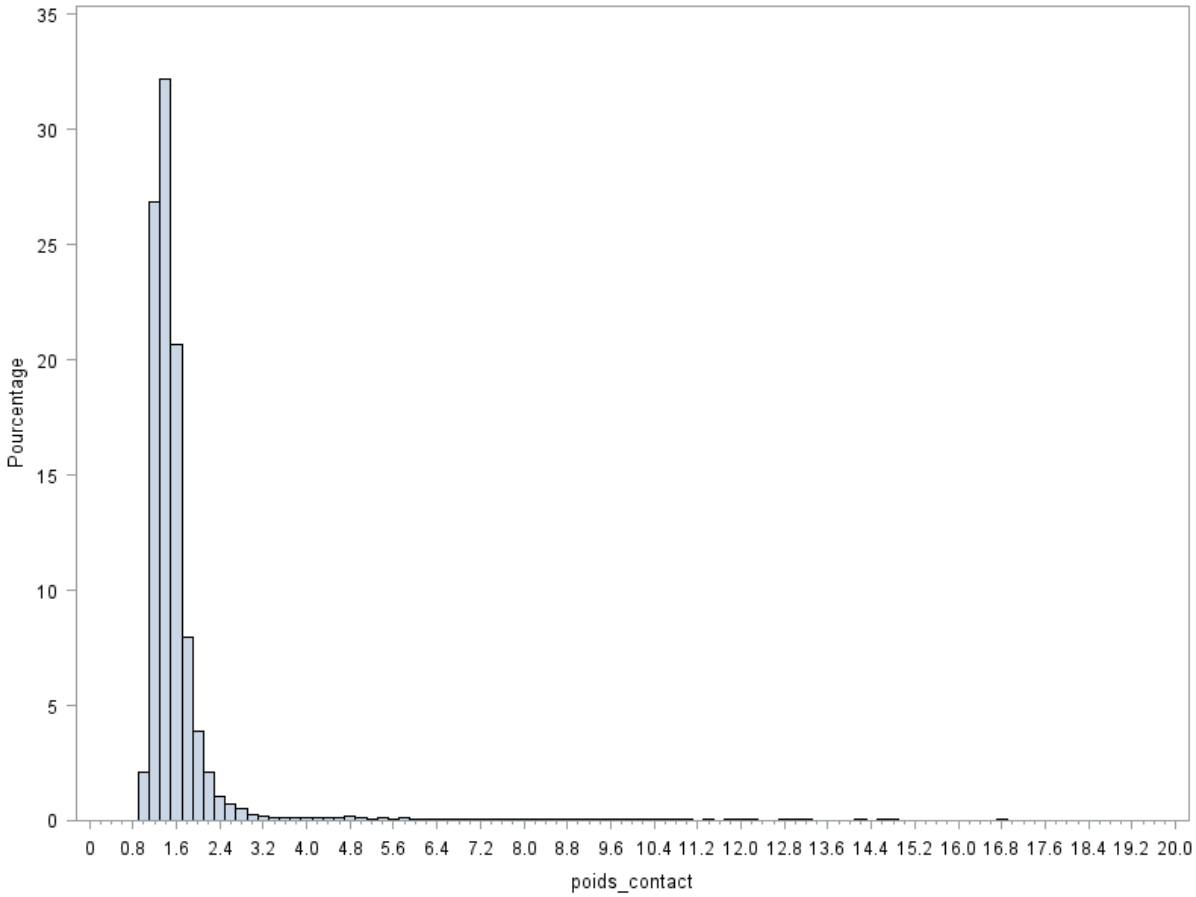
Pour les 89 878 individus ayant effectivement été contactés, le poids relatif au contact est égal à l'inverse de la probabilité de contact estimée par la procédure logistique.

La somme des poids relatifs au contact pour les individus effectivement contactés est égale à 155 392. On note une légère contraction des poids par rapport à l'échantillon (158 401 individus, soit un écart de 1,9 %).

### 8.4.3. Poids de contact des individus ayant complété un questionnaire dans le champ du Céreq

Pour les 19 498 individus dans le champ de l'enquête et répondant au questionnaire, il est compris entre 1,02 et 16,8.

**Figure 13 • Distribution des poids de contact**



## 8.5. Le poids relatif au fait d'accepter de répondre

### 8.5.1. Modélisation de la probabilité d'accepter de répondre

À l'issue de la phase de contact, il reste 89 878 individus potentiellement dans le champ du Céreq.

Parmi eux, 48 262 (53,7 %) ont accepté de répondre au questionnaire. On modélise à nouveau la probabilité d'accepter de répondre en fonction des caractéristiques individuelles grâce à une régression logistique.

Les variables retenues dans le modèle sont :

- mode d'envoi de la lettre avis ;
- présence du nom de commune dans l'adresse de l'individu ;
- type d'établissement ;
- niveau d'études ;
- information sur l'adresse de l'individu en 2013 (QPV ou non) ;
- région de l'établissement ;
- indicateur de qualité du numéro de téléphone ;
- présence du mail dans les informations de l'individu.

Les individus dont les coordonnées téléphoniques disponibles étaient renseignées et les plus fiables ont plus souvent accepté de répondre. Ceux qui ont suivi une formation par apprentissage acceptent moins fréquemment de répondre. La strate de formation et la région de l'établissement de formation ont également leur importance.

### 8.5.2. Cohérence de la modélisation

Le modèle présente 57,8 % de paires concordantes. Le pouvoir prédictif du modèle est donc faible, nettement inférieur à celui du modèle de contact précédent. Si les caractéristiques individuelles jouent un rôle majeur sur la probabilité de joindre un individu ou un proche, le fait d'accepter de répondre sachant qu'on a été contacté en dépend moins.

Sur l'ensemble des 89 878 individus ayant fait l'objet d'un contact, la probabilité moyenne d'acceptation estimée est égale à 0,54 (min = 0,12 ; max = 0,80 ; ET = 0,08).

Pour les 48 262 individus ayant effectivement accepté de répondre, le poids relatif au fait d'accepter de répondre est égal à l'inverse de la probabilité d'accepter de répondre estimée par la procédure logistique.

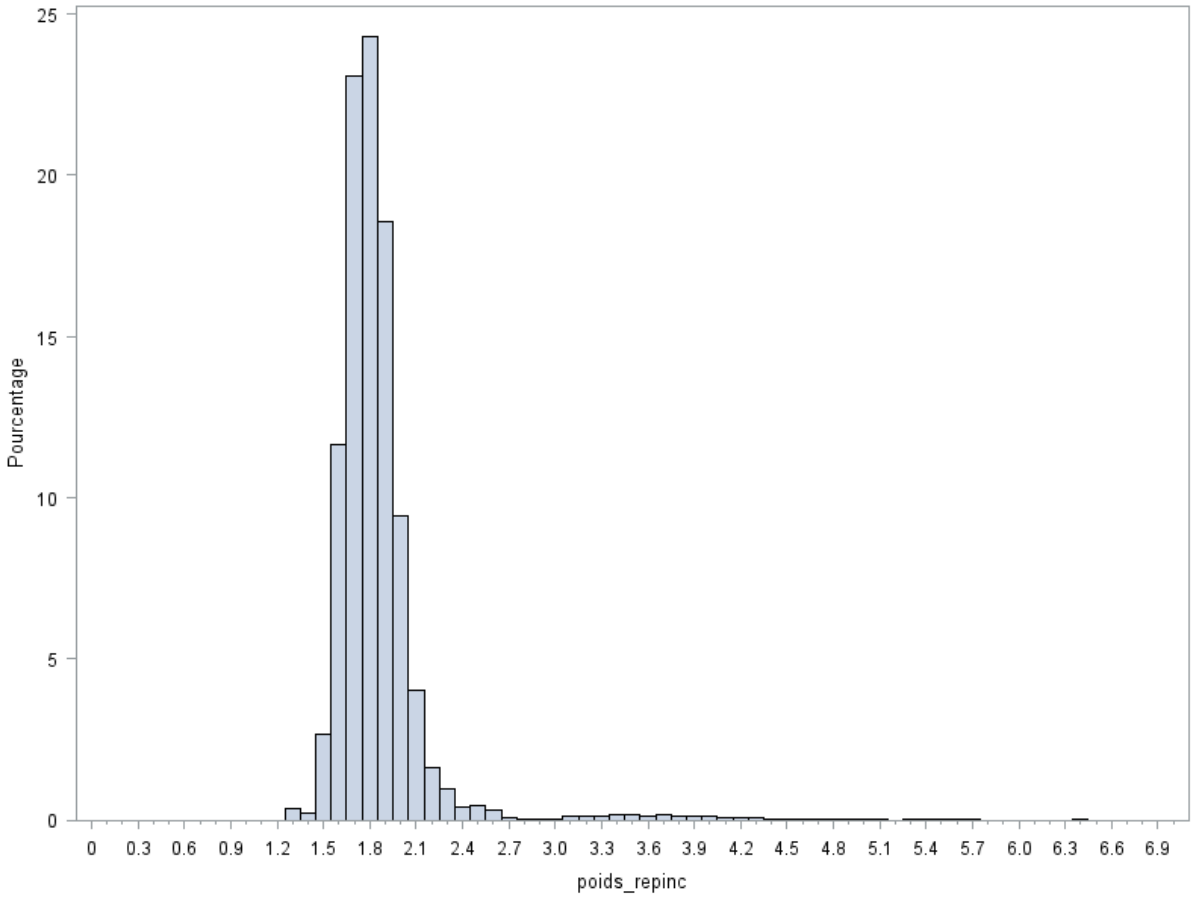
### 8.5.3. Mise hors-champ des individus à l'issue du questionnaire filtre

Un total de 28 238 individus ayant accepté de répondre s'est révélé hors du champ de l'enquête. Il s'agit principalement d'individus ayant poursuivi leurs études à la rentrée scolaire 2013-2014, ou d'individus post-initiaux (ayant interrompu leurs études une première fois puis les ayant reprises avant de les arrêter à nouveau en 2012-2013). Ces individus sont donc considérés comme répondants hors-champ.

### 8.5.4. Poids relatif au fait d'accepter de répondre pour les individus ayant complété un questionnaire dans le champ du Céreq

Pour les 19 498 individus dans le champ de l'enquête et répondant au questionnaire, il est compris entre 1,25 et 6,43.

**Figure 14 • Distribution des poids relatifs au fait d'accepter de répondre**




## 8.6. Probabilité de répondre à l'intégralité du questionnaire sachant que l'on appartient au champ de l'enquête

### 8.6.1. Modélisation de la probabilité de terminer le questionnaire dans le champ

À l'issue du questionnaire filtre et après exclusion des individus hors du champ, 22 024 personnes ont commencé un questionnaire.

Parmi elles, 19 498 (89 %) ont terminé le questionnaire. On modélise la probabilité de terminer un questionnaire sachant que l'on est dans le champ du Céreq en fonction des caractéristiques individuelles grâce à une régression logistique.

Les variables retenues sont :

- obtention du diplôme si individu en classe terminale à la sortie du système éducatif ;
- mois de l'interrogation – avril à juillet 2016 ;
- nombre de connexion au questionnaire ;
- mode d'envoi de la lettre avis ;
- présence du nom de commune dans l'adresse de l'individu ;
- niveau d'études ;
- type d'établissement ;
- information sur l'adresse de l'individu en 2013 (QPV ou non) ;
- indicateur de qualité du numéro de téléphone.

### 8.6.2. Cohérence de la modélisation

Le modèle permet de prédire le comportement réel des individus dans 68,3 % des cas.

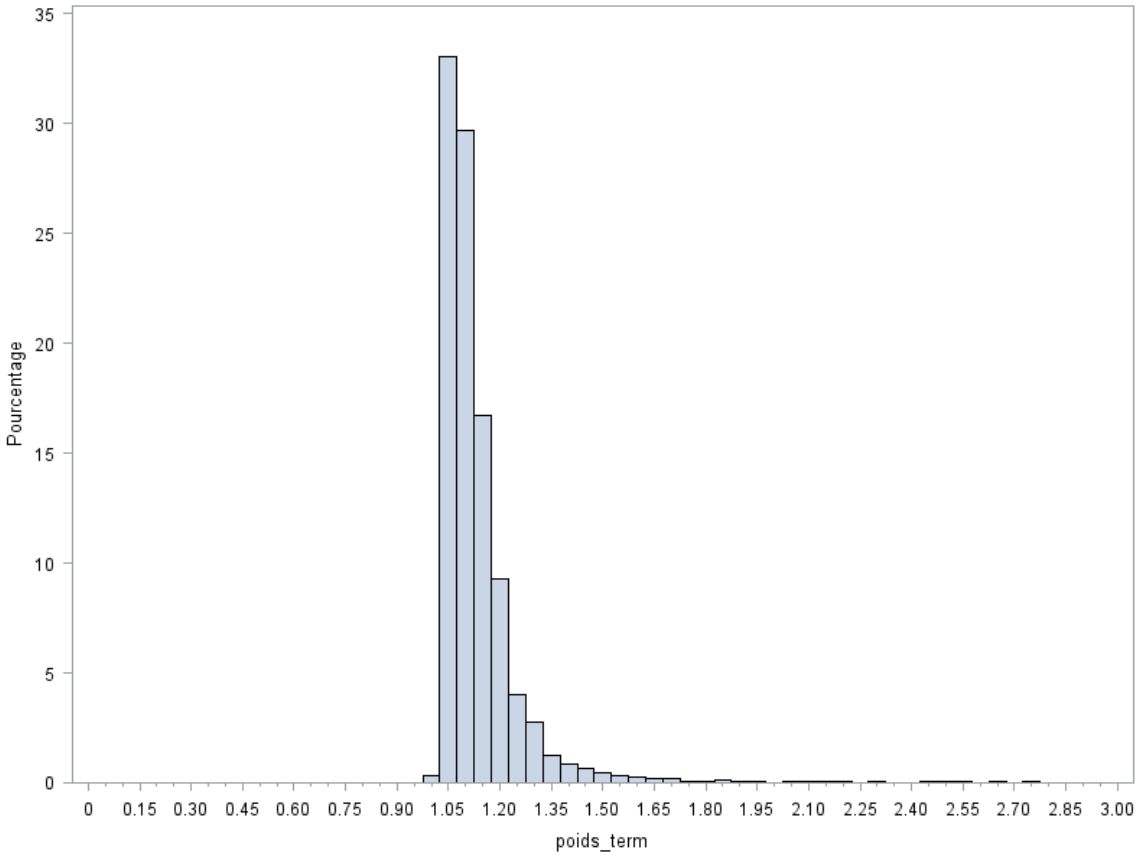
Sur l'ensemble des 22 024 individus dans le champ ayant commencé un questionnaire, l'estimation de la probabilité moyenne de le terminer est égale à 0,89 (min = 0,29 ; max = 0,99 ; ET = 0, 08).

Pour les individus ayant effectivement terminé un questionnaire, le poids relatif est égal à l'inverse de la probabilité de terminer le questionnaire estimé par la procédure logistique.

### 8.6.3. Poids relatif au fait de terminer un questionnaire exploitable pour les individus mis à disposition dans la table d'exploitation

Pour les 19 498 individus dans le champ de l'enquête et présents dans la base mise à disposition, il est compris entre 1 et 2,75.

**Figure 15 • Distribution des poids relatifs au fait de terminer un questionnaire exploitable**



## 8.7. Lissage des poids par groupes homogènes de non-réponse

À partir des probabilités de réponses estimées, des groupes homogènes de non-réponse sont constitués pour estimer une probabilité de répondre au sein de chaque groupe homogène de non-réponse. Cette étape est une étape de robustesse qui permet de se prémunir contre une mauvaise spécification du modèle de correction de la non-réponse.

La population des 19 498 répondants est divisée en sous-populations supposées homogènes au sens de la non-réponse et on suppose que la probabilité de non-réponse est constante au sein de sous-groupes.
Les individus sont ordonnés selon les poids de repondération :

$$p_{couverture}(i) * p_{échantillonnage}(i) * p_{contact}(i) * p_{accepte\ \ répondre}(i) * p_{ter\ min\ e\ \ questionnaire}(i)$$

Puis on divise les répondants en 25 groupes de tailles approximativement égales (méthode des quantiles). Au sein de chaque groupe, le poids est égal à la moyenne du poids de repondération du groupe.

Pour les 19 498 individus dans le champ de l'enquête et présents dans la base mise à disposition, il est compris entre 2,35 et 165,63.

## 8.8. Le calage sur marges

L'enquête Emploi de l'INSEE sert de référence dans les publications du ministère de l'Éducation nationale. Afin d'obtenir des résultats cohérents avec cette source, les données de Génération sont également calées dans ce sens. On utilise la répartition par sexe et plus haut diplôme du tableau « *Le niveau d'étude à la sortie du système éducatif* » de la publication « Repères et références statistiques » (RERS 2016) pour réaliser le calage.

On choisit d'utiliser 16 marges de calage, issues du croisement du genre et du niveau du plus haut diplôme obtenu.
Les données du RERS concernent uniquement la France métropolitaine. Le calage est donc réalisé sur les 19 498 individus formés en France métropolitaine. On estime le nombre de sortants du système éducatif au cours ou à l'issue de l'année scolaire 2012-2013 avant calage à 667 292 et après calage à 672 000.

Le calage a été effectué à l'aide de la macro-procédure CALMAR, de l'INSEE. Le tableau 50 présente les marges utilisées, estimées en compilant les enquêtes trimestrielles de trois années successives, 2012 à 2014.

**Tableau 50 • Les marges de calage pour la France métropolitaine**

| Plus haut diplôme | Homme (en %) | Femme (en %) |
|---|---|---|
| Aucun diplôme ou diplôme national du brevet | 7,88 | 5,51 |
| CAP, BEP ou équivalent | 7,33 | 5,47 |
| Bac. professionnel, bac. technologique | 11,45 | 10,13 |
| Baccalauréat général | 3,44 | 4,00 |
| Bac+2 et santé social | 7,30 | 8,33 |
| Bac+3/4 | 4,47 | 5,05 |
| Grandes écoles | 3,34 | 2,91 |
| Bac+5 – Docteurs | 4,58 | 8,81 |
| **Total** | **49,78** | **50,22** |

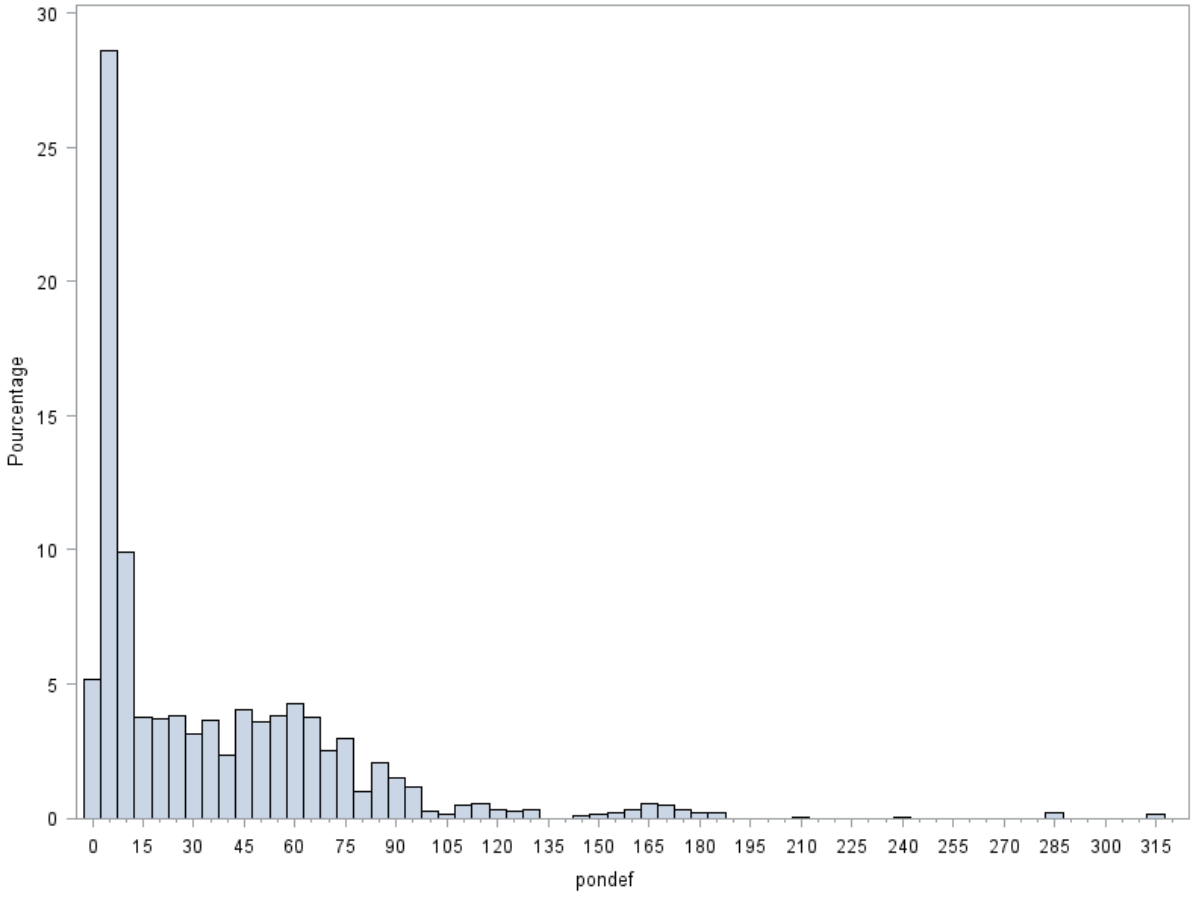
**Tableau 51 • La ventilation selon le plus haut diplôme pour les enquêtes Génération 2004 à 2013**

| Données France métropolitaine | Génération 2004 | Génération 2007 | Génération 2010 | Génération 2013 |
|---|---|---|---|---|
| Non-diplômés hommes | 10,53 | 11,17 | 10,6 | 7,88 |
| Non-diplômées femmes | 6,17 | 6,83 | 5,74 | 5,51 |
| CAP – BEP et équivalent hommes | 10,29 | 10,22 | 7,94 | 7,33 |
| CAP – BEP et équivalent femmes | 7,00 | 6,78 | 6,35 | 5,47 |
| Bac professionnel et technologique hommes | 9,72 | 9,52 | 9,78 | 11,45 |
| Bac professionnel et technologique femmes | 8,64 | 8,00 | 9,62 | 10,13 |
| Bac général hommes | 2,24 | 2,20 | 4,11 | 3,44 |
| Bac général femmes | 3,44 | 3,29 | 5,03 | 4,00 |
| Bac+2 (ensemble), bac+3 de la santé et du social hommes | 7,89 | 7,05 | 6,91 | 7,30 |
| Bac+2 (ensemble), bac+3 de la santé et du social femmes | 10,69 | 8,95 | 8,35 | 8,33 |
| Bac+3 (hors santé social), bac+4 (ensemble) hommes | 4,59 | 4,50 | 3,65 | 4,47 |
| Bac+3 (hors santé social), bac+4 (ensemble) femmes | 7,42 | 7,26 | 4,70 | 5,05 |
| Bac+5 et plus hommes | 6,18 | 7,45 | 7,95 | 7,92 |
| Bac+5 et plus femmes | 5,19 | 6,79 | 9,27 | 11,72 |
| **Ensemble** | **100** | **100** | **100** | **100** |
| **Effectifs totaux** | **737 000** | **739 000** | **685 500** | **672 000** |
| Source | Céreq, enquête Génération 2004, pondération calée | Céreq, enquête Génération 2007, pondération calée | Céreq, enquête Génération 2010, pondération calée hors sortants IUFM | Céreq, enquête Génération 2013, pondération calée |

## 8.9. Le récapitulatif

Pour les 19 498 individus dans le champ de l'enquête et répondant au questionnaire, le poids final est compris entre 1,25 et 315,4.

**Figure 16 • Distribution des poids définitifs**



# 9. Effet des phases de pré-enquête

## 9.1. Effet de la lettre avis

**Tableau 52 • Type d'adresse de l'individu selon sa réponse ou non à l'enquête**

| Type adresse | Non-répondants | | Répondants | | Total | |
|---|---|---|---|---|---|---|
| | N | % | N | N | N | % |
| Conforme | 77 949 | 68,9 | 35 175 | 31,1 | **113 124** | **100** |
| Déménagé rachetée | 6975 | 69,9 | 3 003 | 30,1 | **9 978** | **100** |
| Aucune adresse | 27 642 | 78,3 | 7 657 | 21,7 | **35 298** | **100** |
| **Total** | **112 566** | **71,1** | **45 835** | **28,9** | **158 401** | **100** |

Au total, plus de 120 000 individus disposent d'une adresse valide (adresse conforme ou déménagée rachetée), ce qui correspond à 78 % des individus, tandis que 22 % n'ont pas d'adresse renseignée.

Chez les individus ayant une adresse confirmée, il y a 31,1 % de répondants contre 68,9 % de non-répondants. Cette répartition est assez similaire chez les adresses de déménagés rachetées, avec 30,1 % de répondants. Toutefois, un faible taux de réponse est constaté pour ceux n'ayant pas d'adresse, avec seulement 21,7 %.

**Tableau 53 • Répartition des déménagés rachetés selon la réponse ou non à l'enquête**

| Déménagé racheté | Non-répondants | | Répondants | | Total | |
|---|---|---|---|---|---|---|
| | N | % | N | % | N | % |
| Oui | 6 975 | 69,9 | 3 003 | 30,1 | **9 978** | **100** |
| Non | 3 891 | 72,2 | 1 497 | 27,8 | **5 388** | **100** |
| **Total** | **10 866** | **71,1** | 4 500 | **28,9** | **15 366** | **100** |

En ce qui concerne les adresses de déménagés, il y a deux cas de figure :

- L'adresse est disponible et peut être rachetée au prestataire, ce qui est le cas de la ligne supérieure du tableau

- L'adresse de l'individu n'est pas rachetable pour diverses raisons, c'est donc la deuxième ligne du tableau

Ainsi, le rachat de l'adresse des déménagés (et donc l'envoi de courrier) permet d'avoir un taux de réponse de 30,1 %, tandis que chez les déménagés qui n'ont pas reçu de courrier étant donné que leur adresse n'a pas pu être obtenue, celui-ci n'est que de 27,8 %.

**Tableau 54 • Format de lettre avis envoyée selon le type d'adresse de l'individu**

| Lettre avis | Conforme | | Déménagé rachetée | | Aucune adresse | | Total | |
|---|---|---|---|---|---|---|---|---|
| | N | % | N | % | N | % | N | % |
| Aucune | 23 107 | 53,3 | 1 371 | 3,2 | 18 888 | 43,5 | **43 366** | **100** |
| Mail | 34 801 | 63,6 | 3 495 | 6,4 | 16 411 | 30 | **54 707** | **100** |
| Papier | 45 692 | 92,6 | 3 664 | 7,4 | 0 | 0 | **49 356** | **100** |
| Papier et mail | 9 524 | 86,8 | 1 448 | 13,2 | 0 | 0 | **10 972** | **100** |
| **Total** | **113 124** | **71,4** | **9978** | **6,3** | **35 299** | **22,3** | **158 401** | **100** |

Au total, 115 035 individus en ont reçu une (ou deux dans le cas de l'envoi par papier et mail).

Pour les individus ayant reçu un mail, 63,6 % ont une adresse confirmée, 30 % n'ont aucune adresse et 6,4 % ont déménagé et ont été rachetées. Cette répartition est différente pour les individus ayant reçu un courrier papier, étant donné que ceux ne disposant d'aucune adresse ne sont pas concernés. 92,6 % des courriers ont ainsi été envoyés aux adresses confirmées, et 7,4 % aux adresses de déménagés. La répartition des envois communs de lettre papier et mail concerne un peu plus les déménagés rachetés, étant donné qu'ils ont été relativement ciblés pour cet envoi. Il y a ainsi 86,8 % de ces envois adressés à des adresses confirmées, et 13,2 % à des déménagés rachetés.

**Tableau 55 • Format de lettre avis envoyée selon la réponse ou non à l'enquête**

| Lettre avis | Non-répondants | | Répondants | | Total | |
|---|---|---|---|---|---|---|
| | N | % | N | % | N | % |
| Aucune | 34 850 | 80,4 | 8 516 | 19,6 | **43 366** | **100** |
| Mail | 36 784 | 67,2 | 17 923 | 32,8 | **54 707** | **100** |
| Papier | 34 514 | 69,9 | 14 842 | 30,1 | **49 356** | **100** |
| Papier et mail | 6 418 | 58,5 | 4 554 | 41,5 | **10 972** | **100** |
| **Total** | **112 566** | **71,1** | **45 835** | **28,9** | **158 401** | **100** |

Chez les individus pour lesquels aucune lettre d'avis n'a été envoyée, il y a une part de non-répondants très élevée. En effet, plus de 80 % des individus qui n'en reçoivent pas n'ont pas répondu, et seuls 20 % ont répondu à l'enquête.

Toutefois, chez les individus ayant reçu une lettre d'avis, le taux de réponse est plus élevé et dépend du support. Il monte à 30 % lorsque l'individu reçoit un courrier papier, et 32,8 % lors de l'envoi d'un mail. Il atteint même plus de 41 % pour ceux qui ont reçu la lettre papier et mail, soit le double par rapport à ceux qui n'ont pas eu de lettre d'avis. Ainsi, les personnes averties de l'enquête, qui représentent 72,6 % du total des individus, sont à l'origine de plus de 81 % des questionnaires récoltés.

Il y a un effet non négligeable des lettres avis. En effet, le rôle informatif qu'elles jouent permet aux enquêtés d'être plus réceptifs aux sollicitations de l'enquêteur et de se préparer. Le coût engendré par cette phase est donc nécessaire, car il permet de toucher un plus grand nombre d'individus et de recueillir plus de questionnaires. De plus, le rachat des adresses des individus ayant déménagé est également utile, dans l'optique où celles-ci sont utilisées pour l'envoi de courrier.

## 9.2. Effet de l'enrichissement des coordonnées

### 9.2.1. Enrichissement pendant l'enquête

Pour 18 632 individus (dont 37 n'ayant pas de numéros initialement dans la base), les coordonnées téléphoniques ont été enrichies de 1 à 4 numéros. Ainsi, 19 432 nouveaux numéros de téléphone ont été intégrés.

79 numéros issus de la base de sondage ont été supprimés lors d'un premier nettoyage de la base (numéro débutant par 00, chaine de caractère dans le champ…).

Au total, 327 615 numéros sont disponibles dans la base :

**Tableau 56 • Origine du numéro de téléphone**

| Origine | Fréquence | Pourcentage | Fréquence cumulée | Pourcentage cumulé |
|---|---|---|---|---|
| **Base de sondage** | 150 813 | 46,03 | 150 813 | 46,03 |
| **Phase A** | 22 267 | 6,80 | 173 080 | 52,83 |
| **Phase D** | 35 322 | 10,78 | 208 402 | 63,61 |
| **Phases B et C** | 99 781 | 30,45 | 308 183 | 94,07 |
| **Enrichis pendant l'enquête** | 19 432 | 5,93 | 327 615 | 100,00 |

16 626 individus ne possèdent aucun numéro de téléphone à la fin de l'enquête.
Parmi eux, 94 se sont connectés au site internet : 43 ont répondu au questionnaire filtre *via* le site internet et ont été classés hors-champs.

### 9.2.2. Coordonnées téléphoniques

Une étude a été réalisée afin de savoir si l'enrichissement téléphonique est pertinent.

**Tableau 57 • Classement final des numéros de téléphone appelés selon la source**

| Source | Répondants* | | Contactés | | Hors Cibles | | Faux Numéros | | Sans Réponse | | Total | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | N | % | N | % | N | % | N | % | N | % | N | % |
| *Base de sondage* | 27 768 | 22,24 | 11 011 | 8,82 | 5 679 | 4,55 | 14 493 | 11,61 | 65 883 | 52,78 | 124 834 | 100 |
| Phase A | 1 470 | 8,82 | 1 385 | 8,31 | 995 | 5,97 | 2 479 | 14,88 | 10 333 | 62,02 | 16 662 | 100 |
| Phase D | 4 430 | 15,76 | 2 975 | 10,58 | 1 640 | 5,83 | 3 025 | 10,76 | 16 045 | 57,07 | 28 115 | 100 |
| Phases B et C | 1 593 | 2,70 | 3 517 | 5,97 | 6 298 | 10,69 | 9 028 | 15,32 | 38 499 | 65,32 | 58 935 | 100 |
| *Enrichis pendant l'enquête* | 9 594 | 51,09 | 1 911 | 10,18 | 478 | 2,55 | 365 | 1,94 | 6 430 | 34,24 | 18 778 | 100 |
| **Total** | **44 855** | **18,14** | **20 799** | **8,41** | **15 090** | **6,10** | **29 390** | **11,88** | **137 190** | **55,47** | **247 324** | **100** |

* Ne prend pas en compte les 976 individus classés hors-champs qui ont répondu au questionnaire Filtre par internet.

Ce tableau présente le classement final des numéros selon les sources (phases A à D) mais également de ceux issus de la base de sondage et pendant l'enquête. Le premier cas concerne les numéros présents dans les fichiers d'individus envoyés par les établissements, tandis que le deuxième concerne les numéros recueillis au cours de l'enquête soit par téléphone (par exemple si l'enquêteur tombe sur une tierce personne qui accepte de communiquer un numéro pour joindre l'individu souhaité), soit par internet *via* l'espace personnel des individus.

### *Focalisation sur les phases d'enrichissement ayant fait l'objet de recherches en phase de pré-enquête.*

#### *Les enrichissements issus de la phase D*

79 % des numéros issus de cette phase ont été contactés et ce sont ceux qui donnent les meilleurs taux de réponse. En effet, 15,8 % des numéros appelés ont permis de réaliser une enquête (courte ou longue), 10,6 % ont permis un contact avec l'individu ou un tiers. Cela peut s'expliquer par le fait que la phase D renvoie uniquement des numéros de portable, qui sont beaucoup plus utilisés par les individus que les fixes.

#### *Les enrichissements issus de la phase A*

75 % des numéros issus de la phase A ont été appelés. Ils donnent des résultats assez satisfaisants, dans la lignée de ceux de la phase D, avec néanmoins un taux de répondants plus faible de 8,8 %, 8,3 % ont permis un contact.

#### *Les enrichissements issus des phases B et C*

Ces numéros, quant à eux, semblent également peu pertinents. En effet, seuls 2,7 % ont permis de réaliser une enquête et seuls 6 % ont abouti à un contact avec l'individu. Ce sont ceux qui présentent le plus haut taux de hors cible, presque 11 % et de faux numéros, 15,3 %. En effet, il s'agit d'une recherche géographique élargie qui reste assez approximative. De plus, 41 % des numéros n'ont pas été appelés, car ils sont considérés comme peu fiables et sont donc en dernière position de l'ordre d'appel.

**Tableau 58 • Source de l'appel ayant permis de faire un questionnaire long dans le champ Céreq**

| Source | Non pondérés | | Pondérés | |
|---|---|---|---|---|
| | N | % | N | % |
| Établissement de formation (BS) | 12 456 | 63,9 | 371 757,9 | 53,68 |
| Annuaire téléphonique (A) | 478 | 2,5 | 13 316,99 | 1,92 |
| Annuaire téléphonique (D) | 1 895 | 9,7 | 78 723,39 | 11,37 |
| Annuaire téléphonique (BC) | 497 | 2,5 | 39 910,36 | 5,76 |
| Enrichissement pendant l'enquête | 4 172 | 21,4 | 188 815,1 | 27,26 |
| **Total** | **19 498** | **100** | **692 523,7** | **100** |

Il est intéressant de comparer les sources des numéros qui ont permis d'avoir une réponse à l'enquête complète, en fonction des effectifs non pondérés et pondérés. En effet, certaines sources de numéros peuvent avoir une plus grande importance lorsque les individus ont été pondérés, c'est-à-dire que ces numéros portent sur des individus dont la présence est nécessaire.

Le tableau ci-dessus montre que la pondération modifie l'importance des différentes phases d'enrichissements, ainsi que des numéros obtenus autrement. En effet, la part des numéros ayant permis l'enquête et provenant de la base de sondage passe de 63,9 % à 53,68 % avec la pondération. Il y a donc dans cette catégorie un bon nombre d'individus qui ne sont pas « rares » dans la base des répondants. À contrario, le faible nombre de répondants en sources B et C voit sa part doubler lors de la prise en compte des pondérations, passant de 2,5 % à 5,8 %. Ces individus sont donc plus difficiles à capter, mais importants

pour l'enquête. Pour les sources A et D, la différence est plus faible (-0,6 et +1,6), tandis que les numéros enrichis pendant l'enquête prennent 6 points lors de l'ajout des pondérations.

### 9.2.3. Statistiques sur les appels

**Tableau 59 • Nombre de numéros de téléphone appelés pour les répondants**

| Nombre de numéros de tél. appelés | Nombre de répondants* | Pourcentage par rapport au total |
|---|---|---|
| **1** | 22 375 | 49,8 |
| **2** | 17 097 | 37,3 |
| **3** | 4 512 | 9,8 |
| **4** | 640 | 1,4 |
| **5** | 171 | 0,4 |
| **6** | 68 | 0,2 |
| **7** | 23 | 0,1 |
| **8** | 5 | 0,0 |

* Le nombre de répondants est différent, car 945 personnes ont été détectées hors-champ en répondant au questionnaire filtre par internet avant d'être appelées (dont 43 pour lesquels aucun de numéro de téléphone n'a été obtenu)

Cela correspond au nombre de numéros de téléphone différents qui ont été appelés pour les individus répondants. Ainsi, la moitié des répondants ont été appelés sur un seul numéro uniquement, 37,3 % sur deux numéros et 9,8 % ont répondu sur le troisième numéro appelé. La part de répondant lors des appels au-delà du quatrième numéro est quasi nulle, ce qui laisse penser que les numéros classés dans les trois premiers pour un individu jouent un rôle clé. Le nombre moyen de numéros de téléphone appelés par individu répondant est de 1,7.

**Tableau 60 • Nombre d'appels sur le dernier numéro appelé des répondants**

| Nombre d'appels sur le dernier numéro appelé | Nombre de répondants* | Pourcentage par rapport au total |
|---|---|---|
| **0** | 976** | 2,13 |
| **Entre 1 et 3** | 22 846 | 49,85 |
| **Entre 4 et 6** | 11 985 | 26,15 |
| **Entre 7 et 9** | 5 690 | 12,42 |
| **Plus de 10** | 4 334 | 9,46 |

* 2 personnes n'ont pas l'information du numéro qui a permis de faire l'enquête
** 976 n'ont pas été contactés avant de faire l'enquête : 945 n'ont jamais été appelés (dont 43 pour lesquels aucun de numéro de téléphone n'a été obtenu), 31 qui ont été appelés au moins une fois sur un ou plusieurs numéros qui ont été classés et qui ont répondu au questionnaire en ligne avant qu'on puisse les appeler sur la nouvelle fiche.

Il s'agit ici du nombre d'appels pour le numéro ayant permis l'enquête (incluant les hors-champs). Il y a la moitié des individus qui ont répondu positivement aux sollicitations de l'enquêteur avec moins de trois appels sur le numéro ayant permis l'enquête. Un quart des individus répondants ont nécessité entre quatre et six appels, et 12,6 % entre sept et neuf appels. Près de 10 % des répondants ont dû être appelés plus de dix fois sur le même numéro, il est donc important de persévérer afin de relancer les individus. Le nombre moyen d'appels sur le dernier numéro des répondants est de 4,5.

.

# 10. Publications récentes

Barret, C., Cissé, M. & Dzikowski, C. (2018). Plan de sondage des enquêtes Génération : utilisation d'un calage pour suréchantillonner les extensions. *13es Journées de méthodologie statistique de l'INSEE (JMS), 12-14 juin, Paris.*

Barret, C., Cissé, M., Gaubert, É., Mazari, Z. & Olaria, M. (2018). Efficacité d'un protocole multimode (téléphone et internet). *13es Journées de méthodologie statistique de l'INSEE (JMS), 12-14 juin, Paris.*

Calmand, J. & Robert, A. (2019). Séjours des jeunes à l'étranger : des objectifs européens partiellement atteints, mais un accès encore inégal à la mobilité. *Céreq Bref, 371.*

Cissé, M. & Barret, C. (2018). Agrégation de données multimode : impact sur la modélisation des variables présentant un effet de mesure. *13es Journées de méthodologie statistique de l'INSEE (JMS), 12-14 juin, Paris.*

Couppié, T. & Vignale, M. (2020). Que deviennent les jeunes des quartiers prioritaires de la ville après leur bac ? *Céreq Bref, 391.*

Gaubert, E., Henrard, V., Robert, A. & Rouaud, P. (2017). Enquête 2016 auprès de la Génération 2013 – Pas d'amélioration de l'insertion professionnelle pour les non-diplômés**.** *Céreq Bref, 356.*

Henrard, V. & Ilardi, V. (coord.) (2017). *Quand l'école est finie. Premiers pas dans la vie active de la Génération 2013. Résultats de l'enquête 2016.* Marseille : Céreq coll « Enquêtes » (nº1).

Ilardi, V., Joseph, O. & Sulzer, E. (2018). L'entrée sur le marché du travail des jeunes de la voie professionnelle rénovée. *Céreq Bref, 365.*

Joseph, O., Sulzer, E., Toutin, M.-H. (2020). Construire les compétences de demain dans le BTP. *Céreq Bref, 389.*

Kergoat, P. (dir.), Sulzer, E. (coord.), Cart, B., Capdevielle-Mougnibas, V., Ilardi, V., Saccomanno, B. & Toutin, M.-H. (2017). *Mesure et analyse des discriminations d'accès à l'apprentissage*. Rapport d'évaluation. Paris : INJEP.

Merlin, F. (2020). Une insertion plus difficile pour les jeunes « recalés » à l'entrée du supérieur. *Céreq Bref, 399*.

# Annexes

## Annexe 1. Table des illustrations

### Les encadrés



### Les figures (graphiques, schémas)

## Les tableaux

# Annexe 2. Lettres avis de contact avec les jeunes

## Lettre envoyée par courrier POSTAL

Céreq
Formation - Travail - Emploi
www.cereq.fr

Génération 2013
Enquête 2016

<Nom> <Prenom>
<Adr1>
<Adr2>
<Adr3>
<CP> <Commune>

Marseille, le 22 avril 2016

**Objet :** Enquête statistique nationale « Génération 2013 » sur l'insertion professionnelle des sortants du système éducatif en 2013

Bonjour <Prénom> <Nom>,

Le Centre d'études et de recherches sur les qualifications (Céreq) mène une enquête statistique nationale pour analyser les parcours professionnels des jeunes au début de leur vie active, en fonction des formations suivies.

Cette enquête cherche à décrire les parcours et les situations professionnelles des sortants du système éducatif en 2012-2013 quel que soit le diplôme préparé ou la formation suivie. Les résultats permettront d'améliorer l'information des jeunes et leur famille au moment de l'orientation scolaire.

Vous êtes sortis du système éducatif en 2012-2013. A ce titre, **vous avez été sélectionné(e)** au hasard dans un échantillon à partir des listes de sortants fournies par les établissements scolaires. **Vous serez interrogé(e) à partir du mois d'avril 2016 par téléphone par un de nos enquêteurs.** Afin d'observer la diversité des parcours et d'assurer ainsi la qualité statistique des résultats, il est très important que vous participiez à cette enquête.

L'objectif de cet appel sera de décrire votre parcours professionnel depuis la fin de vos études. Il durera vingt minutes environ et pourra varier selon les différentes situations professionnelles que vous avez connues. Vous trouverez ci-joint des documents qui vous aideront à préparer cette interrogation.

Conformément à la loi, vos réponses resteront confidentielles et serviront uniquement à la réalisation de statistiques qui alimenteront des études sur l'insertion professionnelle.

En vous remerciant par avance de votre participation,

Le directeur du Céreq
Alberto Lopez

*Le numéro de téléphone dont nous disposons pour vous joindre est le suivant : **<TEL1 >***
***Pour actualiser ou confirmer votre numéro de téléphone, et pour faciliter l'entretien téléphonique,***
*vous pouvez accéder à votre espace en flashant le QR-code ci-après*
*ou en vous connectant au lien suivant :*
*https://generation2013.cereq.fr*

*Vos identifiants sont : Identifiant : **<IDENT>***
*Mot de passe : **<MDP>***

*Vous pouvez également nous contacter au numéro vert **0 800 710 459** (appel gratuit depuis un téléphone fixe, en semaine entre 10h et 21h, le samedi entre 10h et 16h) en mentionnant votre identifiant : **<IDENT>***

Vu l'avis favorable du Conseil National de l'Information Statistique, cette enquête est reconnue d'intérêt général et de qualité statistique sans avoir de caractère obligatoire.
Visa n°2016X713AU du Ministre de l'Éducation nationale, de l'Enseignement supérieur et de la recherche, du Ministre du Travail, de l'Emploi, de la Formation professionnelle et du Dialogue social, du Ministre des finances et des comptes publics et du Ministre de l'économie, de l'industrie et du numérique valable pour l'année 2016.
En application de la loi n°51-711 du 7 juin 1951, les réponses à ce questionnaire sont protégées par le secret statistique et destinées à la production de statistiques publiques.
La loi n°78-17 du 6 janvier 1978 modifiée relative à l'informatique, aux fichiers et aux libertés, s'applique aux réponses faites à la présente enquête. Elle garantit aux personnes concernées un droit d'accès et de rectification pour les données les concernant. Ce droit peut être exercé auprès du Céreq, 10 place de la Joliette, 13 567 Marseille Cedex 02, ou bien par un mail adressé au correspondant informatique et libertés du Céreq à l'adresse suivante : cil-generation@cereq.fr en mentionnant dans l'objet « Génération 2013 » suivi de votre identifiant apparaissant sur la première page de ce courrier.

## *DOCUMENT A CONSERVER – NE PAS ENVOYER*

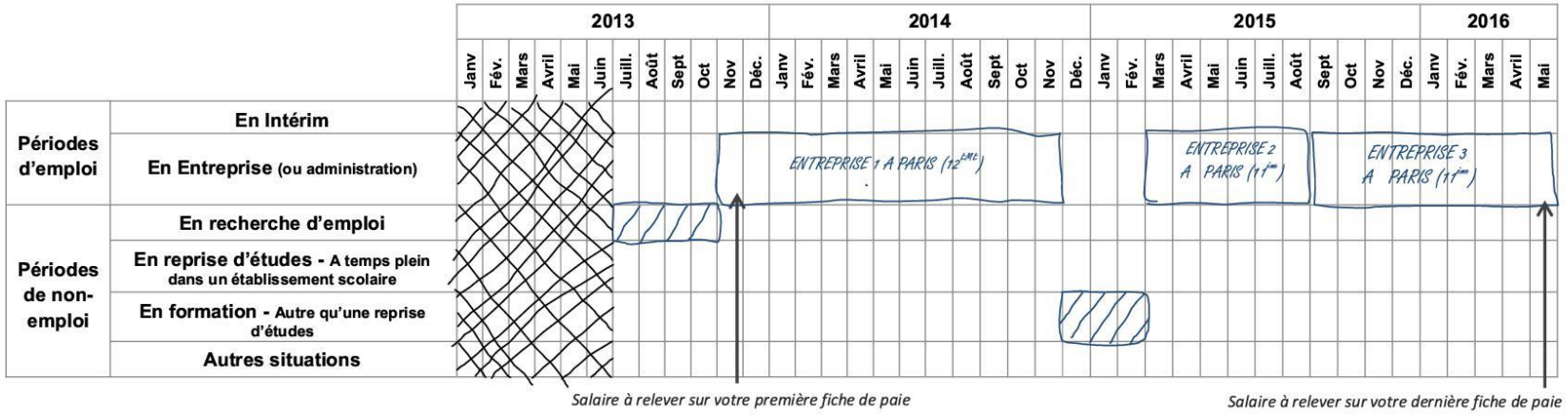
### Votre parcours depuis 2013

**Nous vous incitons à remplir ce calendrier afin de faciliter l'entretien téléphonique. L'enquêteur disposera sur écran d'un calendrier vierge qu'il remplira avec vous oralement. Vous trouverez dans la plaquette jointe un guide vous permettant de compléter le calendrier ci-dessous.**

| | | 2013 | | | | | | | | | | | | 2014 | | | | | | | | | | | | 2015 | | | | | | | | | | | | 2016 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Janv | Fév. | Mars | Avril | Mai | Juin | Juill. | Août | Sept | Oct | Nov | Déc. | Janv | Fév. | Mars | Avril | Mai | Juin | Juill. | Août | Sept | Oct | Nov | Déc. | Janv | Fév. | Mars | Avril | Mai | Juin | Juill. | Août | Sept | Oct | Nov | Déc. | Janv | Fév. | Mars | Avril | Mai |
| **Périodes d'emploi** | **En Intérim** | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | **En Entreprise** (ou administration) | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| **Périodes de non-emploi** | **En recherche d'emploi** | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | **En reprise d'études** - A temps plein dans un établissement scolaire | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | **En formation** - Autre qu'une reprise d'études | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | **Autres situations** | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |

***Exemple de parcours :***

**Dans cet exemple, sorti(e) en juin 2013 du système éducatif, vous avez recherché un emploi de juillet à octobre 2013. De novembre 2013 à novembre 2014 vous avez travaillé dans une première entreprise à Paris. Vous avez ensuite effectué une formation de trois mois. De mars à aout 2015, vous avez été en emploi dans une seconde entreprise à Paris. Depuis septembre 2015, vous travaillez dans une troisième entreprise à Paris.**



## Lettre envoyée par courrier ÉLECTRONIQUE

De : Le Céreq (Centre d'Etudes et de Recherches sur les Qualifications)

**Objet :** Enquête statistique nationale « Génération 2013 » sur l'insertion professionnelle des sortants du système éducatif en 2013


Bonjour <Prénom> <Nom>,

Le Centre d'études et de recherches sur les qualifications (Céreq) mène une enquête statistique nationale pour analyser les parcours professionnels des jeunes au début de leur vie active, en fonction des formations suivies.

Cette enquête cherche à décrire les parcours et les situations professionnelles des sortants du système éducatif en 2012-2013 quel que soit le diplôme préparé ou la formation suivie. Les résultats permettront d'améliorer l'information des jeunes et leur famille au moment de l'orientation scolaire.

Vous êtes sortis du système éducatif en 2012-2013. A ce titre, **vous avez été sélectionné(e)** au hasard dans un échantillon à partir des listes de sortants fournies par les établissements scolaires. **Vous serez interrogé(e) à partir du mois d'avril 2016 par téléphone par un de nos enquêteurs**. Afin d'observer la diversité des parcours et d'assurer ainsi la qualité statistique des résultats, il est très important que vous participiez à cette enquête.

*Le numéro de téléphone dont nous disposons pour vous joindre est le suivant :* ***<TEL1 >***
***Pour actualiser ou confirmer votre numéro de téléphone, et pour faciliter l'entretien téléphonique,***
*vous pouvez accéder à votre espace en cliquant sur le lien suivant :*

***Lien personnalisé => cawi enrichissement***

*Vous pouvez également nous contacter au numéro vert* ***0 800 710 459*** *(appel gratuit depuis un téléphone fixe, en semaine entre 10h et 21h, le samedi entre 10h et 16h) en mentionnant la référence :* ***<IDENT>.***

L'objectif de cet appel sera de décrire votre parcours professionnel depuis la fin de vos études. Il durera vingt minutes environ et pourra varier selon les différentes situations professionnelles que vous avez connues. Vous trouverez ci-joint des documents qui vous aideront à préparer cette interrogation.

Conformément à la loi, vos réponses resteront confidentielles et serviront uniquement à la réalisation de statistiques qui alimenteront des études sur l'insertion professionnelle.

Des informations complémentaires sur notre enquête sont disponibles ici.

En vous remerciant par avance de votre participation,

Le directeur du Céreq
Alberto Lopez

Vous pouvez aussi nous retrouver sur notre page Facebook spéciale Génération 2013 en cliquant ici

Vu l'avis favorable du Conseil National de l'Information Statistique, cette enquête est reconnue d'intérêt général et de qualité statistique sans avoir de caractère obligatoire.
Visa n°2016X713AU du Ministre de l'Éducation nationale, de l'Enseignement supérieur et de la recherche, du Ministre du Travail, de l'Emploi, de la Formation professionnelle et du Dialogue social, du Ministre des finances et des comptes publics et du Ministre de l'économie, de l'industrie et du numérique valable pour l'année 2016.
En application de la loi n°51-711 du 7 juin 1951, les réponses à ce questionnaire sont protégées par le secret statistique et destinées à la production de statistiques publiques.
La loi n°78-17 du 6 janvier 1978 modifiée relative à l'informatique, aux fichiers et aux libertés, s'applique aux réponses faites à la présente enquête. Elle garantit aux personnes concernées un droit d'accès et de rectification pour les données les concernant. Ce droit peut être exercé auprès du Céreq, 10 place de la Joliette, 13 567 Marseille Cedex 02, ou bien par un mail adressé au correspondant informatique et libertés du Céreq à l'adresse suivante : cil-generation@cereq.fr en mentionnant dans l'objet « Génération 2013 » suivi de votre identifiant apparaissant sur la première page de ce courrier.

Pour ne plus recevoir de courrier électronique de la part du Céreq cliquer ici.

# Annexe 3. Nomenclature des diplômes

Nomenclature des diplômes pour la codification de l'enquête emploi (version au 01/01/2013) de niveau, détaillé en **104 postes,** utilisée pour la codification des diplômes dans l'enquête emploi. Les diplômes qui ne sont plus délivrés actuellement sont en italique souligné.

## Diplômes de niveau bac+5 ou plus

1000 DU NIVEAU BAC+5 (DIPLÔME UNIVERSITAIRE DE NIVEAU BAC+5) 1100 MAGISTÈRE

1110 MASTÈRE SPÉCIALISÉ

1140 DRT (DIPLÔME DE RECHERCHE ET TECHNOLOGIE)

1200 *DEA (DIPLÔME D'ÉTUDES APPROFONDIES)*, *DES (DIPLÔME D'ÉTUDES SPÉCIALISÉES)*

1210 MASTER RECHERCHE (LMD)

1300 *DESS (DIPLÔME D'ÉTUDES SUPÉRIEURES SPÉCIALISÉES)*

1310 MASTER PROFESSIONNEL (LMD)

1420 DNSEP NIVEAU BAC+5 (DIPLÔME NATIONAL SUPÉRIEUR D'EXPRESSION PLASTIQUE)

1450 BEES 3E DEGRÉ (BREVET D'ÉTAT D'ÉDUCATEUR SPORTIF 3E DEGRÉ)

1500 CAPME (DIPLÔME DE CAPACITÉ DE MÉDECINE)

1600 ÉCOLE SUPÉRIEURE DE COMMERCE NIVEAU BAC+5

1640 AUTRE TITRE OU CERTIFICAT NIVEAU BAC+5 (ARCHITECTE, EXPERT- COMPTABLE, DSCG…)

1700 INGÉNIEUR (ÉCOLE D'INGÉNIEUR)

1800 PROFESSEUR D'ENSEIGNEMENT SECONDAIRE (CAPES, CAPET, CAPLP, CAFEP, PROFESSORAT DE SPORT…) NIVEAU BAC+5

1880 PROFESSEUR DES ÉCOLES (CAPE) NIVEAU BAC+5

1900 AGRÉGATION

1960 DOCTORATS PROFESSIONS DE SANTÉ (MÉDECINE, PHARMACIE, DENTAIRE, VÉTÉRINAIRE)

1970 DOCTORATS DE RECHERCHE (HORS SANTÉ)

1980 AUTRE DIPLÔME NIVEAU BAC+5 OU PLUS (AVOCAT, NOTAIRE, MAGISTRAT, EXPERT GÉOMÈTRE, SCIENCES PO…)

## Diplômes de niveau bac+3 ou 4 (licence, maîtrise)

2000 LICENCE, LICENCE GÉNÉRALE LMD (L3)

2010 MAÎTRISE, MAÎTRISE INTERMÉDIAIRE (M1)

2200 *MST MAÎTRISE DE SCIENCES ET TECHNIQUES*

2300 DIPLÔME D'INGÉNIEUR MAÎTRE (MAÎTRISE D'IUP)

2400 DSAA (DIPLÔME SUPÉRIEUR DES ARTS APPLIQUÉS)

2410 DNAT (DIPLÔME NATIONAL D'ART ET DE TECHNOLOGIE), DNAP (DIPLÔME NATIONAL D'ARTS PLASTIQUES) NIVEAU BAC+3

2420 *DNSEP (DIPLÔME NATIONAL SUPÉRIEUR D'EXPRESSION PLASTIQUE) NIVEAU BAC+4*

2450 BEES NIVEAU BAC+4 (BREVET D'ÉTAT D'ÉDUCATEUR SPORTIF 2E DEGRÉ)

2460 DESE, DEST (DIPLÔME D'ÉTUDES SUPÉRIEURES DU CNAM)

2500 LICENCE PRO (LICENCE PROFESSIONNELLE)

2560 DU NIVEAU BAC+3/4 (DIPLÔME UNIVERSITAIRE DE NIVEAU BAC+3/4)

2580 DREA (DIPLÔME DE RECHERCHE ET D'ÉTUDES APPLIQUÉES)

2600 *ÉCOLE SUPÉRIEURE DE COMMERCE NIVEAU BAC+4*

2640 AUTRE TITRE OU CERTIFICAT NIVEAU BAC+3/4 (DCG, *DESCF, DECF*, BEAUX-ARTS…)

2800 *IUFM, CAPES, CAPET, AUTRES CONCOURS D'ENSEIGNEMENT SECONDAIRE (PLP2, PROFS DE SPORT…) NIVEAU BAC+3/411*

2880 *IUFM, CAPE (PROFESSEUR DES ÉCOLES), ENSEIGNEMENT DU 1ER DEGRÉ NIVEAU BAC+3/4*

2900 *AGRÉGATION NIVEAU BAC+4*

2960 DIPLÔMES DE LA SANTÉ ET DU TRAVAIL SOCIAL DE NIVEAU BAC+3/4 (SAGE- FEMME, INFIRMIÈRE, OSTÉOPATHE, *DSTS DIPLÔME SUPÉRIEUR DE TRAVAIL SOCIAL*…)

2980 AUTRE DIPLÔME NIVEAU BAC+3/4 (JOURNALISTE, ART, ÉTUDES JUDICIAIRES…)

## Diplômes de niveau bac+2

3200 BTS (BREVET DE TECHNICIEN SUPÉRIEUR)

3210 DMA (DIPLÔME DES MÉTIERS D'ART)

3220 DTS (DIPLÔME DE TECHNICIEN SUPÉRIEUR), DNTS, DPECF

3230 BTSA (BREVET DE TECHNICIEN SUPÉRIEUR AGRICOLE)

3350 BM NIVEAU BAC+2 (BREVET DE MAÎTRISE SUPÉRIEUR)

3400 *DNAT (DIPLÔME NATIONAL D'ART ET DE TECHNOLOGIE) NIVEAU BAC+2*

3420 *DNAP (DIPLÔME NATIONAL D'ARTS PLASTIQUES) NIVEAU BAC+2*

3460 DPCT, DPCE (DIPLÔMES DU CNAM NIVEAU BAC+2)

3500 DUT (DIPLÔME UNIVERSITAIRE DE TECHNOLOGIE)

3510 *PROPÉDEUTIQUE*

3520 DEUG (DIPLÔME D'ÉTUDES UNIVERSITAIRES GÉNÉRALES), AUTRES DIPLÔMES UNIVERSITAIRES NIVEAU BAC+2 (PCEM, *DUEL, DUES*…)

3550 DEUST (DIPLÔME D'ÉTUDES UNIVERSITAIRES SCIENTIFIQUES ET TECHNIQUES)

3560 DU NIVEAU BAC+2 (DIPLÔME UNIVERSITAIRE DE NIVEAU BAC+2)

3630 CSA (CERTIFICAT DE SPÉCIALISATION AGRICOLE) NIVEAU BAC+2

3640 AUTRE TITRE OU CERTIFICAT NIVEAU BAC+2 (ÉCOLE DE VENTE, *DEFA D'ARCHITECTE*, CLERC DE NOTAIRE…)

3880 *ÉCOLE NORMALE D'INSTITUTEUR, PEGC*

3960 DIPLÔMES DE LA SANTÉ ET DU TRAVAIL SOCIAL DE NIVEAU BAC+2 (KINÉ, LABORANTIN, PUÉRICULTRICE, ÉDUCATEUR, *INFIRMIÈRE AVANT 2010*…)

3980 AUTRE DIPLÔME NIVEAU BAC+2 (ÉCOLE D'ART…)

## Diplômes de niveau bac

4000 BAC PRO (BACCALAURÉAT PROFESSIONNEL)
4010 BMA (BREVET DES MÉTIERS D'ART)
4020 BTM (BREVET TECHNIQUE DES MÉTIERS)
4030 BAC PRO AGRICOLE (BACCALAURÉAT PROFESSIONNEL AGRICOLE)
4060 *BMS BREVET DES MÉTIERS DU SPECTACLE*
4100 *BEI BREVET D 'ENSEIGNEMENT INDUSTRIEL*
4110 *BEC BREVET D 'ENSEIGNEMENT COMMERCIAL*
4120 *BEH BREVET D'ENSEIGNEMENT HÔTELIER*
4130 *BEA BREVET D'ENSEIGNEMENT AGRICOLE*
4140 *BES BREVET D'ENSEIGNEMENT SOCIAL*
4200 BT (BREVET DE TECHNICIEN)
4230 BTA (BREVET DE TECHNICIEN AGRICOLE)
4300 BAC TECHNO (BACCALAURÉAT TECHNOLOGIQUE) : STG, STI, STL, ST2S, *SMS, STT, F, G, H*
4330 BAC TECHNO AGRICOLE : STAV, *STAE, STPA*
4350 BM (BREVET DE MAÎTRISE)
4370 MC POST BAC (MENTION COMPLÉMENTAIRE NIVEAU BAC)
4450 BEES NIVEAU BAC (BREVET D'ÉTAT D'ÉDUCATEUR SPORTIF 1ER DEGRÉ), BPJEPS
4500 BP (BREVET PROFESSIONNEL)
4530 BPA (BREVET PROFESSIONNEL AGRICOLE)
4600 *BSEC (BREVET SUPÉRIEUR D'ENSEIGNEMENT COMMERCIAL)*
4630 CSA NIVEAU BAC (CERTIFICAT DE SPÉCIALISATION AGRICOLE NIVEAU BAC)
4640 AUTRE TITRE OU CERTIFICAT NIVEAU BAC (BEPECASER, SECRÉTAIRE MÉDICALE…)
4700 BAC GENERAL : L, ES, S, *A, B, C, D, E, PHILO, MATH-ELEM, SCIENCES EX, MATHEMATIQUES ET TECHNIQUES*
4880 CAPACITE EN DROIT, DAEU (DIPLÔME D'ACCES AUX ÉTUDES UNIVERSITAIRES), *ESEU*
4900 *BREVET SUPÉRIEUR*
4960 DIPLÔMES DE LA SANTÉ ET DU TRAVAIL SOCIAL NIVEAU BAC (ASSISTANTE DENTAIRE, MONITEUR ÉDUCATEUR…)
4980 AUTRE DIPLÔME NIVEAU BAC (PERSONNEL NAVIGANT COMMERCIAL, DIPLÔME D'ÉTUDES MUSICALES…)

## Diplômes de niveau CAP-BEP

5000 CAP (CERTIFICAT D'APTITUDE PROFESSIONNELLE)

5030 CAPA (CERTIFICAT D'APTITUDE PROFESSIONNELLE AGRICOLE)

5100 BEP (BREVET D'ÉTUDES PROFESSIONNELLES)

5130 BEPA (BREVET D'ÉTUDES PROFESSIONNELLES AGRICOLES)

5230 *BAA (BREVET D'APPRENTISSAGE AGRICOLE)*

5320 BC (BREVET DE COMPAGNON)

5370 MC AU CAP-BEP (MENTION COMPLÉMENTAIRE NIVEAU CAP-BEP)

5400 *PREMIÈRE PARTIE BACCALAURÉAT*, CFES (CERTIFICAT DE FIN D'ÉTUDES SECONDAIRES)

5450 *BREVET ÉLÉMENTAIRE, BEPS BREVET ENSEIGNEMENT PRIMAIRE SUPÉRIEUR*

5530 BPA (BREVET PROFESSIONNEL AGRICOLE)

5560 *EFAA (EXAMEN DE FIN D'APPRENTISSAGE ARTISANAL)*

5630 CSA NIVEAU CAP-BEP (CERTIFICAT DE SPÉCIALISATION AGRICOLE NIVEAU CAP/BEP)

5640 AUTRE TITRE OU CERTIFICAT NIVEAU CAP-BEP (TITRE AFPA NIVEAU CAP…)

5960 DIPLÔME SANTÉ ET TRAVAIL SOCIAL NIVEAU CAP-BEP (AIDE SOIGNANTE, AUXILIAIRE PUÉRICULTRICE…)

5980 AUTRE DIPLÔME NIVEAU CAP-BEP (*MONITEUR AUTO-ÉCOLE, ENSEIGNEMENT MÉNAGER*…)

## Diplômes de niveau Brevet

6400 DNB (DIPLÔME NATIONAL BREVET), *BREVET DES COLLÈGES, BEPC*

6410 *CEPRO (CERTIFICAT D'ÉDUCATION PROFESSIONNELLE)*

## Diplômes de niveau CEP ou aucun diplôme

7400 *CEP (CERTIFICAT D 'ÉTUDES PRIMAIRES)*, DFEO (DIPLÔME FIN ÉTUDES OBLIGATOIRES)

7510 CFG (CERTIFICAT DE FORMATION GÉNÉRALE)

7990 AUCUN DIPLÔME RECONNU

# Annexe 4. Nomenclature des niveaux d'études

Cette nomenclature de niveaux en **137 postes** n'a pas de caractère officiel, elle est utilisée par l'INSEE pour la codification des enquêtes ménages. Les niveaux des diplômes qui ne sont plus délivrés actuellement sont en italique souligné.

Les années terminales de préparation des diplômes ont un code dont les 4 premiers caractères sont identiques à ceux du code diplôme, et le dernier est égal à 0.

Les années non terminales de préparation des diplômes ont un code dont le 5e caractère est différent de 0. Elles sont surlignées en gras.

## Formations de niveau bac+5 ou plus

10000 DU NIVEAU BAC+5 (DIPLÔME UNIVERSITAIRE NIV BAC+5)

11000 MAGISTÈRE

11100 MASTÈRE SPÉCIALISE

11400 DRT (DIPLÔME RECHERCHE TECHNOLOGIE)

12000 DEA (DIPLÔME ÉTUDES APPROFONDIES), *DES (DIPLÔME ÉTUDES SPÉCIALISÉES)*

12100 MASTER RECHERCHE (LMD)

13000 DESS DIPLÔME ÉTUDES SUPÉRIEURES SPÉCIALISÉES

13100 MASTER PROFESSIONNEL (LMD)

14200 DNSEP NIVEAU BAC+5 (DIPLÔME NATIONAL SUPÉRIEUR D'EXPRESSION PLASTIQUE)

14500 BEES NIVEAU BAC+5 (BREVET ÉTATÉDUCATEUR SPORTIF 3E DEGRÉ)

15000 CAPME (CAPACITE DE MÉDECINE)

16000 ÉCOLE SUPÉRIEURE DE COMMERCE NIVEAU BAC+5

16400 AUTRE TITRE HOMOLOGUE BAC +5 (ARCHITECTE, EXPERT-COMPTABLE, DSCG…)

17000 INGENIEUR (ÉCOLE D'INGENIEUR)

18000 PROFESSEUR D'ENSEIGNEMENT SECONDAIRE NIVEAU MASTER : CAPES, CAPET, CAPLP, CAFEP, PROFESSORAT DE SPORT…

18800 PROFESSEUR D'ENSEIGNEMENT PRIMAIRE NIVEAU MASTER : CAPE

19000 AGRÉGATION

19600 DOCTORAT DES PROFESSIONS DE SANTÉ (MÉDECINE, PHARMACIE, DENTAIRE, VÉTÉRINAIRE)

19700 DOCTORAT DE RECHERCHE (HORS SANTÉ)

19800 AUTRES DIPLÔMES DE NIVEAU BAC+5 OU PLUS (AVOCAT, NOTAIRE, MAGISTRAT, SCIENCES PO…)

**19991 ANNÉE NON TERMINALE DE DOCTORAT DE RECHERCHE (HORS SANTÉ)**

**19992 ANNÉE NON TERMINALE DE DOCTORAT MÉDECINE, PHARMACIE, DENTAIRE…**

**19999 BAC+5 ET PLUS (SANS AUTRE INDICATION)**

## Formations de niveau bac+3 et bac+4

20000 LICENCE, LICENCE GÉNÉRALE LMD (L3)

20100 MAÎTRISE, 1RE ANNÉE MASTER (M1)

22000 *MST (MAÎTRISE SCIENCES TECHNIQUES)*

23000 DIPLÔME D'INGENIEUR MAITRE (MAÎTRISE D IUP)

24000 DSAA (DIPLÔME SUPÉRIEUR ARTS APPLIQUÉS)

24100 DNAT (DIPLÔME NATIONAL D'ART ET DE TECHNOLOGIE), DNAP (DIPLÔME NATIONAL D'ARTS PLASTIQUES) NIVEAU BAC+3

24200 *DNSEP ( DIPLÔME NATIONAL SUPÉRIEUR EXPRESSION PLASTIQUE) NIVEAU BAC+4*

24500 BEES NIVEAU BAC+4 (BREVET ÉTAT ÉDUCATEUR SPORTIF 2E DEGRÉ)

24600 DESE, DEST (DIPLÔME D'ÉTUDES SUPÉRIEURES DU CNAM)

25000 LICENCE PRO (LICENCE PROFESSIONNELLE)

25600 DU NIVEAU BAC+3/4 (DIPLÔME UNIVERSITAIRE NIV BAC+3/4)

25800 DREA (DIPLÔME DE RECHERCHE ET D'ÉTUDES APPLIQUÉES)

26000 *ÉCOLE SUPÉRIEURE DE COMMERCE NIVEAU BAC+4*

26400 AUTRE TITRE HOMOLOGUE BAC+3/4 (DCG, *DESCF, DECF,* BEAUX-ARTS…)

28000 IUFM, CAPES, CAPET, CAPLP, CAFEP, PROFESSORAT DE SPORT… NIVEAU LICENCE

28800 *IUFM. CAPE (PROFESSEUR DES ÉCOLES),ENSEIGNEMENT PRIMAIRE NIVEAU LICENCE*

29000 *AGRÉGATION*

29600 DIPLÔMES DE LA SANTÉ ET DU TRAVAIL SOCIAL DE NIVEAU BAC+3/4 (SAGE- FEMME, INFIRMIÈRE OSTÉOPATHE, *DSTS ( DIPLÔME SUPÉRIEUR DE TRAVAIL SOCIAL)*

29800 AUTRES DIPLÔMES DE NIVEAU BAC+3/4 (JOURNALISTE, ART, ÉTUDES JUDICIAIRES…)

**29991 1RE ANNÉE IUFM**

**29998 BAC+3 (SANS AUTRE INDICATION)**

**29999 BAC+4 (SANS AUTRE INDICATION)**

## Formations de niveau bac+2

32000 BTS (BREVET DE TECHNICIEN SUPÉRIEUR)

32100 DMA (DIPLÔME DES MÉTIERS D'ART)

32200 DTS (DIPLÔME DE TECHNICIEN SUPÉRIEUR), DNTS, *DPECF*

32300 BTSA (BREVET DE TECHNICIEN SUPÉRIEUR AGRICOLE)

33500 BM NIVEAU BAC+2 (BREVET DE MAÎTRISE SUPÉRIEUR)

34000 *DNAT (DIPLÔME NATIONAL ART ET TECHNOLOGIE)*

34200 *DNAP (DIPLÔME NATIONAL D'ARTS PLASTIOUES)*

34600 *DPCT, DPCE (DIPLÔME PREMIER CYCLE CNAM)*

35000 DUT (DIPLÔME UNIVERSITAIRE TECHNOLOGIE)

35100 *PROPEDEUTIQUE*

35200 DIPLÔMES UNIVERSITAIRES GÉNÉRAUX 1ER CYCLE (DEUG, *DUEL, DUES…)*, 2E ANNÉE LICENCE (L2)

35500 DEUST (DIPLÔME ÉTUDES UNIVERSITAIRES SCIENTIFIQUES TECHNIQUES)

35600 DU NIVEAU BAC+2 (DIPLÔME UNIVERSITAIRE DE NIVEAU BAC+2)

36300 CSA NIVEAU BAC+2 (CERTIFICAT SPÉCIALISATION AGRICOLE NIVEAU BAC+2)

36400 AUTRE TITRE HOMOLOGUE DE NIVEAU BAC+2 (ÉCOLE DE VENTE, *DEFA D'ARCHITECTE,* CLERC DE NOTAIRE…)

38800 *ÉCOLE NORMALE D'INSTITUTEUR PEGC*

**38881 ANNÉE POST BTS POST DUT FCIL NIVEAU BAC+2**

**38883 ANNÉE NON TERMINALE ÉCOLE SUPÉRIEURE DE COMMERCE**

**38884 ANNÉE NON TERMINALE ÉCOLE INGÉNIEUR**

**38885 ANNÉE NON TERMINALE DES AUTRES FORMATIONS DE NIVEAU > A U BAC+2 (ARCHITECTURE, MAGISTRATURE, VÉTÉRINAIRE, SCIENCES PO, JOURNALISTE…)**

39600 DIPLÔMES DE LA SANTÉ ET DU TRAVAIL SOCIAL DE NIVEAU BAC+2 (KINÉ, LABORANTIN, PUÉRICULTRICE, ÉDUCATEUR, *INFIRMIÈRE AVANT 2010…)*

39800 AUTRE FORMATION NIVEAU BAC+2 (ÉCOLE D'ART NIVEAU BAC+2…)

**39991 DCEM 2E CYCLE MÉDECINE ODONTOLOGIE PHARMACIE**

**39999 BAC+2 (SANS AUTRE INDICATION)**

## Formations de niveau bac et bac+1

40000 BAC PRO (BACCALAURÉAT PROFESSIONNEL)
40100 BMA (BREVET DES MÉTIERS D'ART)
40200 BTM (BREVET TECHNIQUE DES MÉTIERS)
40300 BAC PRO AGRICOLE (BACCALAURÉAT PROFESSIONNEL AGRICOLE)
40600 *BMS BREVET DES MÉTIERS DU SPECTACLE*
41000 *BEI BREVET D'ENSEIGNEMENT INDUSTRIEL*
41100 *BEC BREVET D'ENSEIGNEMENT COMMERCIAL*
41200 *BEH BREVET D'ENSEIGNEMENT HOTELIER*
41300 *BEA BREVET D'ENSEIGNEMENT AGRICOLE*
41400 *BES BREVET D'ENSEIGNEMENT SOCIAL*
42000 BT (BREVET DE TECHNICIEN)
42300 BTA (BREVET DE TECHNICIEN AGRICOLE)
43000 BAC TECHNO (BACCALAURÉAT TECHNOLOGIQUE) : STG, STI, STL, ST2S, *SMS, STT, F, G, H*
43300 BAC TECHNO AGRICOLE : STAV, *STAE, STPA*
43500 BM (BREVET DE MAÎTRISE)
43700 MC APRES BAC (MENTION COMPLÉMENTAIRE NIVEAU BAC)
44500 BEES NIVEAU BAC (BREVET ÉTAT ÉDUCATEUR SPORTIF PREMIER DEGRÉ), BPJEPS
45000 BP (BREVET PROFESSIONNEL)
45300 BPA (BREVET PROFESSIONNEL AGRICOLE)
46000 *BSEC (BREVET SUPÉRIEUR ENSEIGNEMENT COMMERCIAL)*
46300 CSA NIVEAU BAC (CERTIFICAT DE SPÉCIALISATION AGRICOLE NIVEAU BAC)
46400 AUTRE TITRE HOMOLOGUÉ NIVEAU BAC (BEPECASER, SECRÉTAIRE MÉDICALE…)
47000 BAC GENERAL : L, ES, S, *A, B, C, D, E, PHILO, MATH-ELEM, SCIENCES EX, BREVET SUPÉRIEUR*
**48201 ANNÉE NON TERMINALE DEUST, DUT**
**48202 ANNÉE NON TERMINALE BTS, BTSA**
**48203 ANNÉE NON TERMINALE DIPLÔMES DU TRAVAIL SOCIAL OU DE LA SANTÉ DE NIVEAU BAC+2 OU BAC+3/4**
**48204 ANNÉE NON TERMINALE AUTRE CYCLE NIVEAU BAC+2 (ART, INSTITUTEUR…)**
48800 CAPACITÉ EN DROIT, DAEU, *ESEU*
49600 DIPLÔMES DE LA SANTÉ ET DU TRAVAIL SOCIAL NIVEAU BAC (ASSISTANTE DENTAIRE, MONITEUR ÉDUCATEUR…)
49800 AUTRE DIPLÔME NIVEAU BAC (PERSONNEL NAVIGANT COMMERCIAL, DIPLÔME D'ÉTUDES MUSICALES…)
**49991 CPGE (CLASSE PRÉPARATOIRE GRANDES ÉCOLES)**
**49992 1RE ANNÉE DEUG, 1RE ANNÉE LICENCE (L1)**
**49993 FORMATION COMPLÉMENTAIRE POST-BAC : CLASSE PRÉPARATOIRE AUX ÉCOLES DE SANTÉ, AUX CONCOURS…**
**49999 BAC+1 (SANS AUTRE INDICATION)**

## Formations de niveau CAP, BEP, seconde et première de lycée

50000 CAP (CERTIFICAT D'APTITUDE PROFESSIONNELLE)

50300 CAPA (CERTIFICAT D'APTITUDE PROFESSIONNELLE AGRICOLE)

51000 BEP (BREVET D'ÉTUDES PROFESSIONNELLES)

51300 BEPA (BREVET D'ÉTUDES PROFESSIONNELLES AGRICOLES)

52300 *BAA (BREVET APPRENTISSAGE AGRICOLE)*

53200 BC (BREVET COMPAGNON)

53700 MC AU CAP, BEP (MENTION COMPLÉMENTAIRE NIVEAU CAP-BEP)

54500 *BE (BREVET ELEMENTAIRE)*, *BEPS (BREVET ENSEIGNEMENT PRIMAIRE SUPÉRIEUR)*

55300 BPA (BREVET PROFESSIONNEL AGRICOLE)

55600 *EFAA (EXAMEN FIN APPRENTISSAGE ARTISANAL)*

56300 CSA NIVEAU CAP-BEP (CERTIFICAT DE SPÉCIALISATION AGRICOLE NIVEAU CAP-BEP)

56400 AUTRE TITRE HOMOLOGUE NIVEAU CAP-BEP (TITRE AFPA NIVEAU CAP…)

**58881 SECONDE TECHNOLOGIQUE**

**58882 SECONDE PRO (SECONDE PROFESSIONNELLE), SECONDE BT, SECONDE SPÉCIALE**

**58883 1RE TECHNO, PREMIERE ADAPTATION**

**58884 PREMIERE BAC PRO, PREMIERE BT**

**58885 ANNÉE NON TERMINALE TITRE PROFESSIONNEL NIVEAU BAC (BP…)**

**58886 FORMATION COMPLÉMENTAIRE POST CAP-BEP**

59600 DIPLÔME SANTÉ ET TRAVAIL SOCIAL NIV CAP-BEP (AIDE SOIGNANTE, AUXILIAIRE PUÉRICULTRICE…)

59800 AUTRE FORMATION NIVEAU CAP-BEP *(MONITEUR AUTO-ÉCOLE, ENSEIGNEMENT MÉNAGER…)*

**59991 SECONDE GÉNÉRALE, SECONDE INDÉTERMINÉE**

**59992 1RE GÉNÉRALE**

## Formation de niveau Brevet des collèges et année non terminale de CAP ou BEP

64000 TROISIÈME, DNB, BEPC, BREVET DES COLLÈGES

64100 CEPRO (CERTIFICAT ÉDUCATION PROFESSIONNELLE)

**68881 ANNÉE NON TERMINALE CAP**

**68882 ANNÉE NON TERMINALE BEP**

**68883 ANNÉE NON TERMINALE AUTRE DIPLÔME NIVEAU CAP-BEP**

## Formation de niveau CEP, 6e, 5e, ou 4e de collège, ou aucun diplôme

74000 CEP (CERTIFICAT D'ÉTUDES PRIMAIRES), DFEO (DIPLÔME FIN ÉTUDES OBLIGATOIRES)

**78882 CPPN, CPA, CLASSE TRANSITION**

79900 AUCUN DIPLÔME RECONNU

**79991 ÉTUDES PRIMAIRES**

**79992 6E, 5E**

**79993 4E**

**79995 IME (INSTITUT MÉDICO EDUCATIF), IMP (INSTITUT MÉDICO PÉDAGOGIQUE), IMPRO (INSTITUT MÉDICO PROFESSIONNEL)**

# Annexe 5. Nomenclature des spécialités

## 1 DOMAINES DISCIPLINAIRES

### 10 Formations générales

100 Formations générales

### 11 Mathématiques et sciences

110 Spécialités pluriscientifiques
111 Physique-Chimie
112 Chimie-biologie, biochimie
113 Sciences naturelles (biologie-géologie)
114 Mathématiques
115 Physique
116 Chimie
117 Sciences de la terre
118 Sciences de la vie

### 12 Sciences humaines et droit

120 Spécialités pluridisciplinaires sciences humaines et droit
121 Géographie
122 Économie
123 Sciences sociales (y compris démographie, anthropologie)
124 Psychologie
125 Linguistique
126 Histoire
127 Philosophie, éthique et théologie
128 Droit, sciences politiques

### 13 Lettres et arts

130 Spécialités littéraires et artistiques plurivalentes
131 Français, littérature et civilisation française
132 Arts plastiques
133 Musique, arts du spectacle
134 Autres disciplines artistiques et spécialités artistiques plurivalentes
135 Langues et civilisations anciennes
136 Langues vivantes, civilisations étrangères et régionales

## 2 DOMAINES TECHNICO-PROFESSIONNELS DE LA PRODUCTION

### 20 Spécialités pluritechnologiques de la production

200 Technologies industrielles fondamentales (génie industriel et procédés de transformation, spécialités à dominante fonctionnelle)
201 Technologies de commandes des transformations industrielles (automatismes et robotique industriels, informatique industrielle)

### 21 Agriculture, pêche, forêts et espaces verts

210 Spécialités plurivalentes de l'agronomie et de l'agriculture
211 Productions végétales, cultures spécialisées et protection des cultures (horticulture, viticulture, arboriculture fruitière…)
212 Production animale, élevage spécialisé, aquaculture, soins aux animaux y compris vétérinaire
213 Forêts, espaces naturels, faunes sauvages, pêche
214 Aménagement paysager (parcs, jardins, espaces verts, terrains de sport)

### 22 Transformations

220 Spécialités pluritechnologiques des transformations
221 Agroalimentaire, alimentation, cuisine.
222 Transformations chimiques et apparentées (y compris industrie pharmaceutique)
223 Métallurgie (y compris sidérurgie, fonderie, non-ferreux…)
224 Matériaux de construction, verre, céramique
225 Plasturgie, matériaux composites
226 Papier, carton
227 Énergie, génie climatique (y compris énergie nucléaire, thermique, hydraulique, utilités : froid, climatisation, chauffage)

### 23 Génie civil, construction et bois

230 Spé. pluritechnologiques, génie civil, construction, bois
231 Mines et carrières, génie civil, topographie
232 Bâtiment : construction et couverture
233 Bâtiment : finitions
234 Travail du bois et de l'ameublement

### 24 Matériaux souples

240 Spécialités pluritechnologiques matériaux souples
241 Textile
242 Habillement (y compris mode, couture)
243 Cuirs et peaux

### 25 Mécanique, électricité, électronique

250 Spé pluri technologiques mécanique-électricité (y compris maintenance mécano-électrique)
251 Mécanique générale et de précision, usinage
252 Moteurs et mécanique auto
253 Mécanique aéronautique et spatiale
254 Structures métalliques (y compris soudure, carrosserie, coque bateau, cellule avion)
255 Électricité, électronique (non compris automatismes, productique)

## 3 DOMAINES TECHNICO-PROFESSIONNELS DES SERVICES

### 30 Spécialités plurivalentes des services

300 Spécialités plurivalentes des services

### 31 Échanges et gestion

310 Spécialités plurivalentes des échanges et de la gestion (y compris administration générale des entreprises et des collectivités)
311 Transport, manutention, magasinage
312 Commerce, vente
313 Finances, banque, assurances
314 Comptabilité, gestion
315 Ressources humaines, gestion du personnel, gestion de l'emploi

### 32 Communication et information

320 Spécialité plurivalente de la communication
321 Journalisme et communication (y compris communication graphique et publicité)
322 Techniques de l'imprimerie et de l'édition
323 Techniques de l'image et du son, métiers connexes du spectacle
324 Secrétariat, bureautique
325 Documentation, bibliothèques, administration des données
326 Informatique, traitement de l'information, réseaux de transmission des données

### 33 Services aux personnes

330 Spécialités plurivalentes sanitaires et sociales
331 Santé
332 Travail social
333 Enseignement, formation
334 Accueil, hôtellerie, tourisme
335 Animation culturelle, sportive et de loisirs
336 Coiffure, esthétique et autres spécialités des services aux personnes

### 34 Services à la collectivité

340 Spécialités plurivalentes des services à la collectivité
341 Aménagement du territoire, développement, urbanisme
342 Protection et développement du patrimoine
343 Nettoyage, assainissement, protection de l'environnement
344 Sécurité des biens et des personnes, police, surveillance (y compris hygiène et sécurité)
345 Application des droits et statuts des personnes
346 Spécialités militaires

## 4 DOMAINES DU DÉVELOPPEMENT PERSONNEL

### 41 Domaines des capacités individuelles

410 Spécialités concernant plusieurs capacités
411 Pratique sportive (y compris arts martiaux)
412 Développement des capacités mentales et apprentissages de base
413 Développement des capacités comportementales et relationnelles
414 Développement des capacités individuelles d'organisation
415 Développement des capacités d'orientation, d'insertion ou de réinsertion sociale et professionnelle

### 42 Domaines des activités quotidiennes et de loisirs

421 Jeux et activités spécifiques de loisirs
422 Économie et activités domestiques
423 Vie familiale, vie sociale et autres formations au développement personnel

# Annexe 6. Nomenclature NAF rev2 en 88 divisions (A88)

| Section | Libellé des sections | Code division | Intitulé |
|---|---|---|---|
| A | AGRICULTURE, SYLVICULTURE ET PÊCHE | 01 | Culture et production animale, chasse et services annexes |
| | | 02 | Sylviculture et exploitation forestière |
| | | 03 | Pêche et aquaculture |
| B | INDUSTRIES EXTRACTIVES | 05 | Extraction de houille et de lignite |
| | | 06 | Extraction d'hydrocarbures |
| | | 07 | Extraction de minerais métalliques |
| | | 08 | Autres industries extractives |
| | | 09 | Services de soutien aux industries extractives |
| C | INDUSTRIE MANUFACTURIÈRE | 10 | Industries alimentaires |
| | | 11 | Fabrication de boissons |
| | | 12 | Fabrication de produits à base de tabac |
| | | 13 | Fabrication de textiles |
| | | 14 | Industrie de l'habillement |
| | | 15 | Industrie du cuir et de la chaussure |
| | | 16 | Travail du bois et fabrication d'articles en bois et en liège, à l'exception des meubles ; fabrication d'articles en vannerie et sparterie |
| | | 17 | Industrie du papier et du carton |
| | | 18 | Imprimerie et reproduction d'enregistrements |
| | | 19 | Cokéfaction et raffinage |
| | | 20 | Industrie chimique |
| | | 21 | Industrie pharmaceutique |
| | | 22 | Fabrication de produits en caoutchouc et en plastique |
| | | 23 | Fabrication d'autres produits minéraux non métalliques |
| | | 24 | Métallurgie |
| | | 25 | Fabrication de produits métalliques, à l'exception des machines et des équipements |
| | | 26 | Fabrication de produits informatiques, électroniques et optiques |
| | | 27 | Fabrication d'équipements électriques |
| | | 28 | Fabrication de machines et équipements n.c.a. |
| | | 29 | Industrie automobile |
| | | 30 | Fabrication d'autres matériels de transport |
| | | 31 | Fabrication de meubles |
| | | 32 | Autres industries manufacturières |
| | | 33 | Réparation et installation de machines et d'équipements |
| D | PRODUCTION ET DISTRIBUTION D'ÉLECTRICITÉ, DE GAZ, DE VAPEUR ET D'AIR CONDITIONNÉ | 35 | Production et distribution d'électricité, de gaz, de vapeur et d'air conditionné |

**Suite table sur la nomenclature NAF**

| Section | Libellé des sections | Code division | Intitulé |
|---|---|---|---|
| E | PRODUCTION ET DISTRIBUTION D'EAU ; ASSAINISSEMENT, GESTION DES DÉCHETS ET DÉPOLLUTION | 36 | Captage, traitement et distribution d'eau |
| | | 37 | Collecte et traitement des eaux usées |
| | | 38 | Collecte, traitement et élimination des déchets ; récupération |
| | | 39 | Dépollution et autres services de gestion des déchets |
| F | CONSTRUCTION | 41 | Construction de bâtiments |
| | | 42 | Génie civil |
| | | 43 | Travaux de construction spécialisés |
| G | COMMERCE ; RÉPARATION D'AUTOMOBILES ET DE MOTOCYCLES | 45 | Commerce et réparation d'automobiles et de motocycles |
| | | 46 | Commerce de gros, à l'exception des automobiles et des motocycles |
| | | 47 | Commerce de détail, à l'exception des automobiles et des motocycles |
| H | TRANSPORTS ET ENTREPOSAGE | 49 | Transports terrestres et transport par conduites |
| | | 50 | Transports par eau |
| | | 51 | Transports aériens |
| | | 52 | Entreposage et services auxiliaires des transports |
| | | 53 | Activités de poste et de courrier |
| I | HÉBERGEMENT ET RESTAURATION | 55 | Hébergement |
| | | 56 | Restauration |
| J | INFORMATION ET COMMUNICATION | 58 | Édition |
| | | 59 | Production de films cinématographiques, de vidéo et de programmes de télévision ; enregistrement sonore et édition musicale |
| | | 60 | Programmation et diffusion |
| | | 61 | Télécommunications |
| | | 62 | Programmation, conseil et autres activités informatiques |
| | | 63 | Services d'information |
| K | ACTIVITÉS FINANCIÈRES ET D'ASSURANCE | 64 | Activités des services financiers, hors assurance et caisses de retraite |
| | | 65 | Assurance |
| | | 66 | Activités auxiliaires de services financiers et d'assurance |
| L | ACTIVITÉS IMMOBILIÈRES | 68 | Activités immobilières |
| M | ACTIVITÉS SPÉCIALISÉES, SCIENTIFIQUES ET TECHNIQUES | 69 | Activités juridiques et comptables |
| | | 70 | Activités des sièges sociaux ; conseil de gestion |
| | | 71 | Activités d'architecture et d'ingénierie ; activités de contrôle et analyses techniques |
| | | 72 | Recherche-développement scientifique |
| | | 73 | Publicité et études de marché |
| | | 74 | Autres activités spécialisées, scientifiques et techniques |
| | | 75 | Activités vétérinaires |

**Suite table sur la nomenclature NAF**

| Section | Libellé des sections | Code division | Intitulé |
|---|---|---|---|
| **N** | ACTIVITÉS DE SERVICES ADMINISTRATIFS ET DE SOUTIEN | **77** | Activités de location et location-bail |
| | | **78** | Activités liées à l'emploi |
| | | **79** | Activités des agences de voyages, voyagistes, services de réservation et activités connexes |
| | | **80** | Enquêtes et sécurité |
| | | **81** | Services relatifs aux bâtiments et aménagement paysager |
| | | **82** | Activités administratives et autres activités de soutien aux entreprises |
| **O** | ADMINISTRATION PUBLIQUE | **84** | Administration publique et défense ; sécurité sociale obligatoire |
| **P** | ENSEIGNEMENT | **85** | Enseignement |
| **Q** | SANTÉ HUMAINE ET ACTION SOCIALE | **86** | Activités pour la santé humaine |
| | | **87** | Hébergement médico-social et social |
| | | **88** | Action sociale sans hébergement |
| **R** | ARTS, SPECTACLES ET ACTIVITÉS RÉCRÉATIVES | **90** | Activités créatives, artistiques et de spectacle |
| | | **91** | Bibliothèques, archives, musées et autres activités culturelles |
| | | **92** | Organisation de jeux de hasard et d'argent |
| | | **93** | Activités sportives, récréatives et de loisirs |
| **S** | AUTRES ACTIVITÉS DE SERVICES | **94** | Activités des organisations associatives |
| | | **95** | Réparation d'ordinateurs et de biens personnels et domestiques |
| | | **96** | Autres services personnels |
| **T** | ACTIVITÉS DES MÉNAGES EN TANT QU'EMPLOYEURS ; ACTIVITÉS INDIFFÉRENCIÉES DES MÉNAGES EN TANT QUE PRODUCTEURS DE BIENS ET SERVICES POUR USAGE PROPRE | **97** | Activités des ménages en tant qu'employeurs de personnel domestique |
| | | **98** | Activités indifférenciées des ménages en tant que producteurs de biens et services pour usage propre |
| **U** | ACTIVITÉS EXTRATERRITORIALES | **99** | Activités des organisations et organismes extraterritoriaux |

# Annexe 7. Procédure de codification de l'activité selon la NAF rev2

## Annexe 7.1. Définition du fichier

Ensemble des séquences en entreprise :

- Entreprise non trouvée dans le menu entreprise.
- Réponses aux questions EP3 à EP7 **avec précision** de l'activité en clair en ep10.

## Annexe 7.2. Traitement du fichier en entrée

*Définition des catégories à partir des réponses aux questions EP3 à EP7*

01 Public : entreprise nationalisée.
02 Public : autre administration.
03 Fabrication pdt.
04 Vente service entreprise.
05 Vente service autre.
06 Vente service particulier.
07 Commerce pdt entreprise.
08 Commerce pdt indéterminé.
09 Commerce pdt non alim particulier.
10 Commerce pdt indéterminé particulier.
11 Commerce pdt alim particulier lieu indéterminé.
12 Commerce pdt alim & non alim particulier lieu indéterminé.
13 Autres activités.

Un regroupement a été nécessaire pour définir des mots clés pour préciser l'activité en clair pour le codage dans *Sicore*. Pour certaines catégories, les mots clés n'ont pas été retenus car non pertinents dans le codage de *Sicore*.

***Mots clés utilisés :***

1- FABRICATION : catégorie 03.
2- COMMERCE GROS : catégorie 07 (commerce interentreprise).
3- COMMERCE : catégorie 08.
4- COMMERCE DÉTAIL : catégories 09 10 11 et 12 (commerce pour des particuliers).

*Sicore* utilise des synonymes de ces termes :

1- PRODUCTION.
2- DISTRIBUTION.
3- COMMERCE.
4- DÉTAIL.

### *Correction orthographique et suppression des accents et caractères spéciaux sur le libellé en clair en ep10*

Procédure effectuée à partir d'Excel avec l'utilisation du correcteur orthographique.

## Annexe 7.3. *Sicore*

Utilisation de l'environnement *Sicore* activité « APE_2012_06.ENV » (passer par environnement->TEST puis charger l'environnement).

Pour coder l'activité, *Sicore* n'utilise qu'une seule variable LIBELLÉ. C'est pourquoi plusieurs fichiers en entrée ont été testés pour coder un maximum de séquences.

- Fichier 1 :

  LIBELLÉ : libellé en clair (brut) de l'activité (sans normalisation de *Sicore*).

Pour les fichiers suivants, la normalisation de *Sicore* a permis de définir les 6 mots déterminants de chaque libellé brut (en supprimant notamment les mots de liaison).
À partir du fichier 1 en sortie, le libellé en clair est construit en concaténant les mots déterminants.

- Fichier 2 :

  LIBELÉ : mot clé de la catégorie + libellé en clair de l'activité (reconstruit) ou libellé en clair de l'activité (pour les catégories sans mots clés).

- Fichier 3 :

  LIBELLÉ : 2 premiers mots du fichier 2.

- Fichier 4 :

  LIBELLÉ : premier mot du fichier 2 à l'exception du mot clé.

## Annexe 7.4. Codification en nomenclature NAF rev2 de l'activité

L'ensemble des séquences ont été codées 4 fois. Dans chaque fichier codé, *Sicore* fournit entre 0 et 5 échos.

### *Choix du bon code APE dans chacun des 4 fichiers*

Au vu des résultats, 2 règles ont été définies et appliquées dans les 4 fichiers pour choisir le meilleur écho :

- 1re règle : si 1 SEUL écho dont la probabilité est supérieure à 40, l'écho en 5 positions est choisi.
- 2e règle : si échos multiples, alors
  - sélection des 2 premières positions des échos ;
  - somme des probabilités des échos identiques SUCCESSIFS (premier écho inclus).

  Si la somme est supérieure à 40, l'écho en 2 positions est choisi.

Au final, pour chaque séquence :

- un code APE en 5 positions ;
- un code APE en 2 positions ;
- pas de code.

### *Reprise de codes APE sur le fichier 2 et 3*

Après analyse des résultats, les règles ont été élargies afin de récupérer des codes du fichier 2 et 3 pour les séquences qui n'ont pas été codées.

- 1re règle : si le premier écho dont la probabilité est supérieure à 40, l'écho en 5 positions est choisi.
- 2e règle : si échos multiples :
  - Sélection des 2 premières positions des échos.
  - Comparaison des différents échos (1 à 5).
  - Somme des probabilités des échos identiques NON SUCCESSIFS (premier écho inclus).

Si la somme est supérieure à 40, l'écho en 2 positions est choisi.

### *Choix du meilleur code APE*

À l'issue des étapes précédentes, un fichier global contenant les 4 résultats de codage a été construit.

- Le code du fichier 2 a été choisi en priorité 1.
- Le code du fichier 3 a été choisi en priorité 2.
- Le code du fichier 4 a été choisi si code absent dans le fichier 1 et inversement.
- Si le code du fichier 4 et celui du fichier 1 sont égaux le code a été choisi (sur 2 ou 5 positions).
- Si le code du fichier 4 et celui du fichier 1 sont différents aucun code n'a été choisi (85 séquences concernées).

Une étape d'imputation s'est avérée nécessaire pour les catégories suivantes (1 seul code sur 2 positions dans la nomenclature NAF rev2) :

- Commerce gros ou détail pour les automobiles et motocycles : 45.
- Commerce gros : 46.

Commerce détail : 47.

Lorsque les séquences des 2 dernières catégories ne sont pas codées alors on impute le code APE correspondant (46 ou 47), excepté celles dont l'activité en clair contient les mots en rapport avec l'automobile (code 45 imputé).

# Annexe 8. Définition des variables annexes utilisées dans *Sicore* PCS

**STATUT :** statut dans l'emploi (clivage salarié / indépendant).

Les modalités de STATUT pour les règles *Sicore* sont les suivantes :

- 1 indépendant.
- 2 salariés de sa propre entreprise, gérant mandataire, PDG.
- 3 salariés.
- (ou blanc) manquant.

**PUB :** statut de l'établissement employeur (distinction public / privé). Cette variable ne concerne que les salariés.
Les modalités de PUB pour les règles *Sicore* sont les suivantes :

- 1 état.
- 2 collectivités territoriales, HLM, hôpitaux.
- 3 sécurité sociale.
- 4 entreprises publiques nationalisées.
- 5 privé.
- (ou blanc) manquant.

**SP :** emploi précaire (distinction apprenti / autre). Cette variable ne concerne que les salariés.
Les modalités de SP pour les règles *Sicore* sont les suivantes :

- 1 apprenti.
- (ou blanc) manquant.

**CPF :** classification professionnelle ou Qualification

Cette variable ne concerne *a priori* que les salariés (cf. PCS 2003 page 43). Les modalités de CPF pour les règles *Sicore* sont les suivantes :

- 1 manœuvre ou ouvrier spécialisé.
- 2 ouvrier qualifié ou ouvrier hautement qualifié ou technicien d'atelier.
- 3 agent de maîtrise.
- 4 directeur général ou adjoint direct au directeur.
- 5 technicien, dessinateur, VRP.
- 6 instituteur, assistante sociale, infirmière et autres personnels de catégorie B de la fonction publique.
- 7 ingénieur ou cadre.
- 8 professeur et personnel de catégorie A de la fonction publique.
- 9 employés de bureau, de commerce, agents de service, aides-soignantes, gardiennes d'enfants, personnels de catégorie C ou D de la fonction publique.
- 0 autres.
- * (ou blanc) manquant.

**FN** : fonction principale

Les modalités de FN pour les règles *Sicore* sont les suivantes :

- 1 production, fabrication, chantiers.
- 2 installation, réparation, maintenance.
- 3 nettoyage, gardiennage, entretien ménager.
- 4 manutention, magasinage, logistique.
- 5 secrétariat, saisie, accueil.
- 6 gestion, comptabilité.
- 7 commerce, vente, technico-commercial.
- 8 études, recherche et développement, méthodes.
- 0 autres.
- (ou blanc) manquant.

**NBS :** nombre de salariés employés

Cette variable ne concerne *a priori* que les chefs d'entreprise ou personnes installées à leur compte.

Les modalités de NBS pour les règles *Sicore* sont les suivantes :

- 1 aucun salarié.
- 2 un ou deux salarié(s).
- 3 trois à neuf salariés.
- 4 dix salariés ou plus.
- * (ou blanc) manquant.

**NAF2** et **NAF** : activité principale de l'établissement

Les modalités de NAF2 et NAF pour les règles *Sicore* sont les suivantes :

- Toutes les modalités possibles de la nomenclature d'activité.
- ** ou **** (ou blanc) manquant.

La variable NAF2 comprend les deux premières positions du code NAF (niveau « division » de la nomenclature). La variable NAF comprend les codes NAF à quatre positions.

Si l'on dispose de la NAF à quatre positions, créer aussi la variable annexe NAF2. Certains codages ne nécessitant que la NAF2.

**S :** sexe

Les modalités de S pour les règles *Sicore* sont les suivantes :

- 1 masculin.
- 2 féminin.
- * (ou blanc) manquant.

**T** : taille de l'entreprise

Les modalités de T pour les règles *Sicore* sont les suivantes :

- 0 de 0 à 9 salariés.
- P de 10 à 49 salariés.
- M de 50 à 499 salariés.
- G 500 salariés et plus.
- * (ou blanc) manquant.

**OPA :** orientation des productions agricoles

Les modalités de OPA pour les règles *Sicore* sont les suivantes :

- 1 polyculture (culture des terres labourables).
- 2 maraîchage ou horticulture.
- 3 vigne ou arbres fruitiers.
- 4 élevage d'herbivores (bovins, ovins…).
- 5 élevage de granivores (volailles, porcs…).
- 6 polyculture-élevage.
- 7 élevage d'herbivores et de granivores.
- 8 autre.
- * (ou blanc) manquant.

**DEP :** département

Les modalités de DEP pour les règles *Sicore* sont les suivantes :

- Tous les numéros de départements métropolitains (pour la Corse, on peut distinguer 2A et 2B ou ne pas distinguer en mettant 20).
- 97 pour les DOM (pas de distinction).
- ** (ou blanc) manquant.

Le département peut être utile pour la codification de certaines professions d'agriculteurs.

# Annexe 9. Pondérations des extensions

La pondération des extensions santé/social et sport diffère de celle de l'enquête Céreq, car le champ n'est pas le même et porte sur l'ensemble des sortants de formation initiale. Des individus post-initiaux sont interrogés en plus des individus du champ Céreq (primo-sortants de formation initiale) pour constituer la base de ces extensions spécifiques. De ce fait la pondération doit en tenir compte.

Le traitement de la non-réponse totale se fait en trois étapes :

- Une modélisation de la probabilité d'obtenir un contact avec l'enquêté ou un de ses proches.
- Une modélisation de la probabilité que l'enquêté réponde au questionnaire en entier sachant qu'il a été contacté (ou un de ses proches).
- Un traitement par groupes homogènes de réponse pour réduire la variance des poids obtenus.

Ensuite, un calage a été réalisé sur les individus diplômés des formations des métiers de la santé et du social, ainsi que pour les diplômés de formation de sport et de l'animation.

Enfin, la mise hors-champ des individus répondants, mais ne remplissant pas de questionnaire est effectuée. Il s'agit d'individus poursuivant leurs études après l'année 2013, n'étant pas sortis de formation en santé/social ou sport durant l'année 2012-2013 ou résidant à l'étranger à la date de l'enquête.

## Annexe 9.1. Extension santé/social pour la DREES

7 238 sortants de formation en santé et social ont été interrogés dans le cadre de l'enquête Génération 2013 à 3 ans. Parmi eux, 3 398 sont des primo-sortants de formation initiale (champ Céreq), 2 113 sont des sortants post-initiaux et 1 727 sont hors du champ de l'enquête. Ils se répartissent de la façon suivante :

**Tableau A1 • Effectifs de sortants de formation en santé et social**

| Formation | Effectif Champ Céreq | Effectif Post-initiaux | Total Champ DREES | Effectif Hors Champ | Total |
|---|---|---|---|---|---|
| Orthophoniste | 167 | 0 | 167 | 18 | 185 |
| Orthoptiste | 48 | 0 | 48 | 17 | 65 |
| Sage-femme | 362 | 5 | 367 | 81 | 448 |
| Aide-soignant | 288 | 679 | 967 | 240 | 1207 |
| Auxiliaire de puériculture | 157 | 281 | 438 | 78 | 516 |
| Ergothérapeute | 86 | 2 | 88 | 8 | 96 |
| Masseur-Kinésithérapeute | 374 | 18 | 392 | 76 | 468 |
| Pédicure-Podologue | 77 | 4 | 81 | 147 | 228 |
| Psychomotricien | 96 | 12 | 108 | 34 | 142 |
| Infirmier | 779 | 338 | 1117 | 275 | 1392 |
| **Total santé** | 2434 | 1339 | 3773 | 974 | 4747 |
| Assistant de service social | 218 | 119 | 337 | 158 | 495 |
| Conseiller en économie sociale et familiale | 149 | 53 | 202 | 118 | 320 |
| Educateur de jeunes enfants | 234 | 131 | 365 | 115 | 480 |
| Educateur spécialisé | 260 | 243 | 503 | 183 | 686 |
| Moniteur-Educateur | 103 | 228 | 331 | 179 | 510 |
| **Total social** | 964 | 774 | 1738 | 753 | 2491 |
| **Total** | 3398 | 2113 | 5511 | 1727 | 7238 |

### *Modélisation de la non-réponse totale*

La probabilité pour un individu d'avoir été contacté, lui-même ou un de ses proches, ainsi que la probabilité de répondre entièrement au questionnaire sont modélisées par des régressions logistiques.

Celles-ci nous indiquent que les hommes, ceux qui ont reçu un mail-avis, les sortants de formations supérieures ont plus de chances d'aboutir à une interrogation complète. Des variables concernant la qualité estimée des coordonnées et l'effort de contact ont également été incluses dans le modèle de contact.

### *Groupes homogènes de réponse*

Pour réduire la variance des poids après correction de la non-réponse totale, des groupes de réponse homogène ont été effectués. Les individus ont été répartis en 50 groupes par poids croissant. Chaque individu s'est vu affecter le poids moyen du groupe auquel il appartient.

### Calage sur marges

Le calage a été effectué sur les effectifs de sortants diplômés des formations en santé et social, à partir des marges fournies par la DREES et disponibles sur leur open data (http://www.data.drees.sante.gouv.fr, Onglet « Professions de santé et du social »).

Le calage a été effectué à l'aide de la macro SAS *Calmar* en utilisant la méthode linéaire.

Il a été réalisé sans inclure les sortants non diplômés par manque d'information provenant de source externe. Cependant, les poids de ces individus ont été modifiés à la suite du calage : pour chaque formation, le même ratio de calage a été appliqué au poids des individus non diplômés qu'à celui des diplômés.

Exception : pas de calage pour les sortants de formation d'orthophoniste et d'orthoptiste, car nous ne disposons pas de marge de calage. Leur poids n'a donc pas été modifié après les groupes homogènes de réponse.

**Tableau A2 • Calage et marges pour l'extension santé et social**

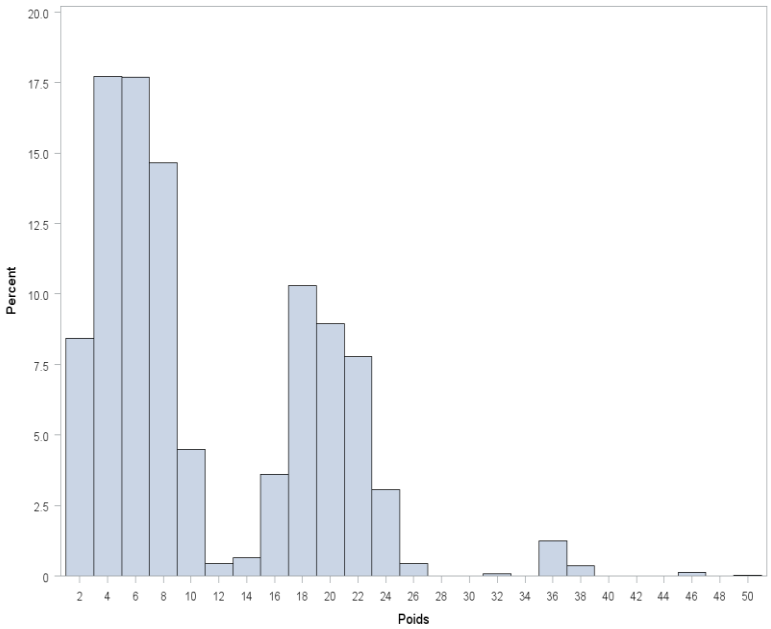
| Formation | Effectif diplômé dans le champ | Marge de calage | Ratio de calage | Effectif diplômé dans le champ pondéré après calage |
|---|---|---|---|---|
| Aide-soignant | 947 | 21 495 | 1.036 | 18 314 |
| Auxiliaire de puériculture | 426 | 4 601 | 1.223 | 4 166 |
| Ergothérapeute | 88 | 496 | 0.805 | 451 |
| Sage-femme | 366 | 914 | 0.841 | 851 |
| Infirmier | 1 087 | 25 323 | 0.965 | 21 998 |
| Masseur-Kinésithérapeute | 390 | 2 220 | 1.301 | 1 997 |
| Pédicure-Podologue | 80 | 533 | 1.941 | 412 |
| Psychomotricien | 108 | 689 | 1.252 | 541 |
| Orthophoniste | 166 | x | 1 | 513 |
| Orthoptiste | 43 | x | 1 | 141 |
| **Total santé** | 3 701 | 56 271 | | 49 384 |
| Assistant de service social | 321 | 2 351 | 1.029 | 1 987 |
| Conseiller en économie sociale et familiale | 182 | 1 177 | 0.917 | 925 |
| Educateur de jeunes enfants | 359 | 1 516 | 1.024 | 1 347 |
| Educateur spécialisé | 494 | 4 408 | 0.997 | 3 890 |
| Moniteur-Educateur | 312 | 2 734 | 1.345 | 2 163 |
| **Total social** | 1 668 | 12 186 | | 10 311 |
| **Total** | 5 369 | 75 742 | | 59 695 |

Note : le calage des diplômés de formation de sage-femme s'est fait sur les individus sortant d'université ainsi que ceux sortant d'école de sage-femmes. Ne sachant pas si les sortants d'université sont inclus dans le tableau de la DREES, il peut y avoir sous-estimation du nombre de sage-femmes.

## Poids finaux

Les poids finaux sont définis par le produit de l'inverse du taux de couverture, du poids d'échantillonnage, du poids de non-réponse et du ratio de calage. Ils sont distribués comme ceci :

**Figure A1 • Histogramme des poids des sortants de formation santé et social**




Par formation :

**Tableau A3 • Poids des sortants de formation santé et social par diplôme**

| Formation | Effectif non-pondéré | Moyenne | Ecart-type | Minimum | Médiane | Maximum |
|---|---|---|---|---|---|---|
| AIDE-SOIGNANT | 967 | 19.35 | 4.56 | 5.02 | 19.34 | 38.87 |
| AUXILIAIRE DE PUERICULTURE | 438 | 9.84 | 4.89 | 6.67 | 9.16 | 45.90 |
| ERGOTHERAPEUTE | 88 | 5.12 | 0.68 | 3.90 | 4.91 | 7.05 |
| INFIRMIER | 1117 | 20.22 | 5.23 | 5.46 | 19.57 | 36.19 |
| MASSEUR-KINESITHERAPEUTE | 392 | 5.12 | 0.82 | 3.70 | 4.98 | 8.85 |
| PEDICURE-PODOLOGUE | 81 | 5.16 | 0.88 | 3.87 | 5.04 | 7.79 |
| PSYCHOMOTRICIENS | 108 | 5.01 | 1.64 | 3.56 | 4.70 | 19.64 |
| SAGE-FEMME | 367 | 2.33 | 3.10 | 1.68 | 1.88 | 31.53 |
| ORTHOPHONIE | 167 | 3.09 | 0.63 | 1.99 | 3.07 | 5.45 |
| ORTHOPTISTE | 48 | 3.28 | 0.63 | 2.43 | 3.25 | 5.45 |
| ASSISTANT DE SERVICE SOCIAL | 337 | 6.19 | 0.76 | 4.35 | 6.04 | 9.00 |
| CONSEILLER EN ECONOMIE SOCIALE FAMILIALE | 202 | 5.08 | 0.80 | 3.87 | 5.00 | 8.02 |
| EDUCATEUR DE JEUNES ENFANTS | 365 | 3.75 | 0.53 | 2.91 | 3.61 | 6.01 |
| EDUCATEUR SPECIALISE | 503 | 7.88 | 1.79 | 4.21 | 7.68 | 37.38 |
| MONITEUR-EDUCATEUR | 331 | 6.92 | 2.71 | 4.55 | 6.51 | 50.46 |

## Annexe 9.2. Extension sport

1690 sortants de formation des métiers du sport et de l'animation ont été interrogés dans le cadre de l'enquête Génération 2013 à 3 ans. Parmi eux, 586 sont des primo-sortants de formation initiale (champ Céreq) et 1104 sont des sortants post-initiaux. Ils se répartissent de la façon suivante :

**Tableau A4 • Effectifs de sortants de formation des métiers du sport et de l'animation**

| Diplôme | Effectif Champ Céreq | Effectif Post-initiaux | Total |
|---|---|---|---|
| BAFA | 6 | 11 | 17 |
| BAPAAT | 18 | 19 | 37 |
| BEES | 31 | 79 | 110 |
| BPJEPS | 457 | 783 | 1240 |
| DEDPAD | 0 | 3 | 3 |
| DEJEPS | 69 | 177 | 246 |
| DESJEPS | 5 | 32 | 37 |
| DEFA | 0 | 0 | 0 |
| BEATEP | 0 | 0 | 0 |
| **Total** | 586 | 1104 | 1690 |

### *Modélisation de la non-réponse totale*

La probabilité pour un individu d'avoir été contacté, lui-même ou un de ses proches, ainsi que la probabilité de répondre entièrement au questionnaire sont modélisées par des régressions logistiques.

Celles-ci nous indiquent que les hommes, ceux qui ont reçu un mail-avis, les sortants de formations supérieures ont plus de chances d'aboutir à une interrogation complète. Des variables concernant la qualité estimée des coordonnées et l'effort de contact ont également été incluses dans le modèle de contact.

### *Groupes homogènes de réponse*

Pour réduire la variance des poids après correction de la non-réponse totale, des groupes de réponse homogène ont été effectués. Les individus ont été répartis en 25 groupes par poids croissant. Chaque individu s'est vu affecter le poids moyen du groupe auquel il appartient.

### *Calage sur marges*

Le calage a été effectué sur les effectifs de sortants diplômés des formations en santé et social, à partir des marges fournies par la MEOS et présentes dans la publication de l'INJEP *« Les chiffres clés du sport »*. Puisqu'il n'y a qu'une variable, le calage (peu importe la méthode) est équivalent à une règle de trois. Le ratio de calage de chaque formation a été rapporté sur les sortants non diplômés.

Par manque d'effectifs dans certains diplômes, des regroupements ont été effectués :

- les sortants de BEES tous degrés confondus sont réunis dans une seule catégorie ;
- les sortants de DESJEPS et DEJEPS sont réunis.

Par manque d'information sur les sortants de DEDPAD, les poids n'ont pas été modifiés par calage.

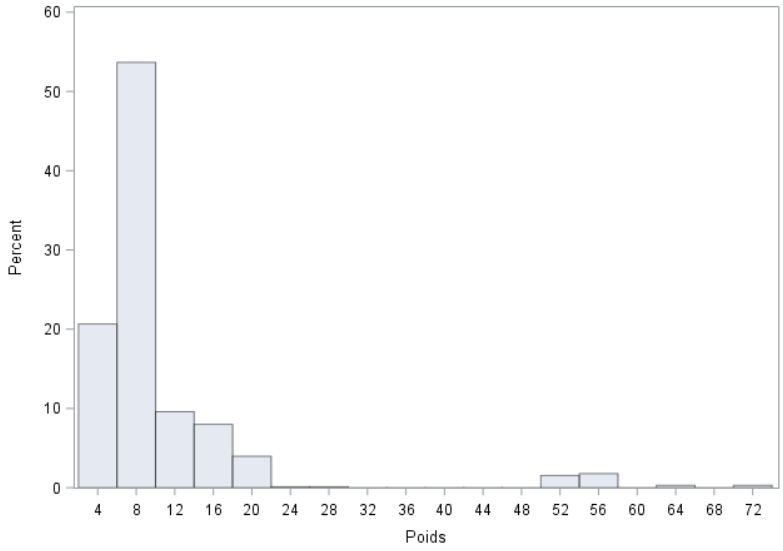
**Tableau A4 • Calage et marges pour l'extension sport**

| Diplôme | Effectif | Effectif pondéré avant calage | Marge de calage | Ratio de calage |
|---|---|---|---|---|
| BAPAAT | 32 | 263 | 663 | 2,52 |
| BEES | 107 | 600 | 1 297 | 2,16 |
| BPJEPS | 1 200 | 5 037 | 9 948 | 1,97 |
| DEJEPS+DESJEPS | 277 | 3 019 | 5 464 | 1,81 |
| **Total** | 1 616 | 8 919 | 17 372 | 1,95 |

## *Poids finaux*

Les poids finaux sont définis par le produit de l'inverse du taux de couverture, du poids d'échantillonnage, du poids de non-réponse et du ratio de calage. Ils sont distribués comme ceci :

**Figure A2 • Histogramme des poids des sortants de formation en sport et animation**



Par niveau de diplôme :

**Tableau A5 • Poids des sortants de formation en sport et animation par diplôme**

| Diplôme | Effectif | Moyenne | Ecart-type | Minimum | Médiane | Maximum |
|---|---|---|---|---|---|---|
| BAFA | 17 | 7,01 | 5,74 | 4,38 | 5,61 | 28,97 |
| BAPAAT | 37 | 19,16 | 21,81 | 7,38 | 10,50 | 72,91 |
| BEES | 110 | 12,13 | 11,23 | 6,34 | 9,47 | 62,61 |
| BPJEPS | 1240 | 8,37 | 7,92 | 5,08 | 6,77 | 57,21 |
| DEDPAD | 3 | 7,44 | 2,13 | 5,61 | 6,94 | 9,78 |
| DEJEPS | 246 | 20,89 | 11,00 | 10,15 | 17,70 | 52,42 |
| DESJEPS | 37 | 11,32 | 7,20 | 7,55 | 10,15 | 52,42 |

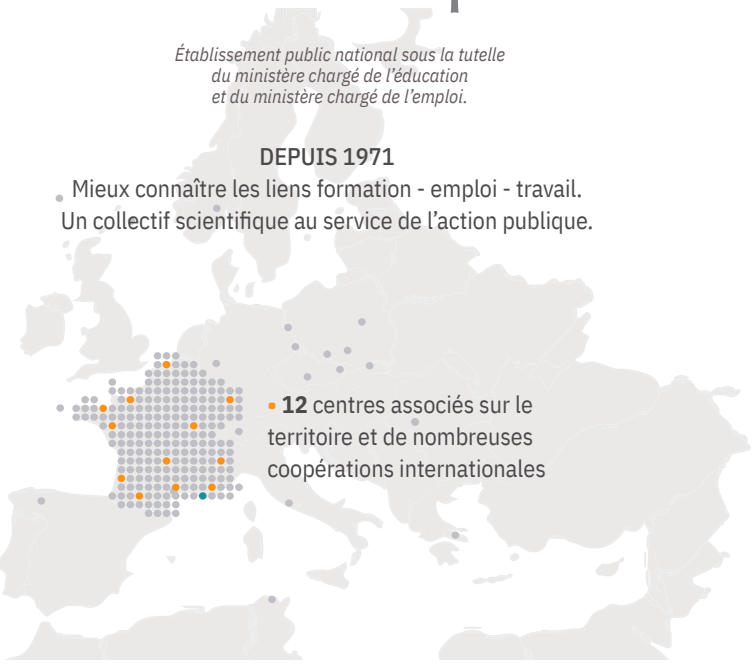
ISSN
2497-6873

Céreq

*Établissement public national sous la tutelle du ministère chargé de l'éducation et du ministère chargé de l'emploi.*

**DEPUIS 1971**

Mieux connaître les liens formation - emploi - travail.
Un collectif scientifique au service de l'action publique.



• **12** centres associés sur le territoire et de nombreuses coopérations internationales

**+ d'infos**
et tous les travaux

**À explorer**
www.cereq.fr

**+ de 600 publications**
Accessibles librement