Direction de la recherche, des études,
de l'évaluation et des statistiques
DREES

SERIE
SOURCES ET METHODES

DOCUMENT DE TRAVAIL

Échantillonnage, apurements et redressement de la non réponse dans l'enquête IAD

Enquête auprès des intervenants au domicile de personnes fragilisées (2008)

Rémy MARQUIER

n° 11 – juin 2010

MINISTERE DU TRAVAIL, DE LA SOLIDARITE ET DE LA FONCTION PUBLIQUE
MINISTERE DE LA SANTE, ET DES SPORTS
MINISTERE DU BUDGET, DES COMPTES PUBLICS ET DE LA REFORME DE L'ÉTAT

Sommaire

Introduction 5

1. Unités enquêtées, champ de l'enquête, protocole de l'enquête 7

2. Description du plan de sondage 9
2.1.Tirage des intervenants 9
2.1.1.Premier degré : tirage de départements 9
2.1.2.Tirage des intervenants : tenir compte des deux bases de sondage 10
2.1.3.Tirage des intervenants issus de la base IRCEM (indépendants et mandataires) 12
2.1.4.Tirage des intervenants issus de la base des organismes (prestataires et mandataires) 16
2.2.Les tirages lors des face à face 18
2.2.1.Tirage de la semaine de référence 18
2.2.2.Tirage du jour de référence 19
2.2.3.Tirage de la personne/prestation de référence 20

3. Apurement des données 23
3.1.Corrections des données relatives aux organismes 23
3.2.Correction des données des interviews téléphoniques auprès des intervenants 24
3.3.Correction des données des interviews en face à face auprès des intervenants 24
3.3.1.Nombres de personnes visitées et des heures travaillées 24
3.3.2.Emploi du temps 26
3.3.3.L'intervention de référence 27
3.3.4.Mode d'emploi et employeurs associés 27
3.3.5.Les salaires et revenus du foyer 28
3.3.6.Les congés 29
3.3.7.Les autres variables 29

4. Redressement des poids des organismes et des interviews téléphoniques 31
4.1.Les intervenants des organismes (base ANSP) 32
4.1.1.Redressement des poids des organismes 32
4.1.2.Redressement des poids des intervenants des organismes 34
4.2.Les intervenants des particuliers employeurs (base IRCEM) 37

5. Redressement des poids des interviews en face à face 39
5.1.Les intervenants sélectionnés à partir des organismes 39
5.2.Les intervenants sélectionnés à partir de la base IRCEM 40
5.3.Le partage des poids 41

6. Redressement de la journée et de l'intervention types 45
6.1. La journée type 45
6.2. L'intervention type 46

Bibliographie indicative 49

Annexe 1 : exemple de tirage d'échantillon à probabilités proportionnelles à la taille – cas du tirage des départements 51
Annexe 2 : exemple de tirage d'un échantillon de réserve – cas des OASP 55
Annexe 3 : exemple d'imputation par hot-deck – cas de l'âge du premier arrêt des études 62
Annexe 4 : exemple de calage sur marges – cas des interviews en face à face 64
Annexe 5 : macro de troncature des poids 67
Annexe 6 : macro de calcul des erreurs quadratiques moyennes sur les poids tronqués 70

Introduction

Dans un contexte général de vieillissement de la population et d'allongement de la durée de vie, les besoins d'intervenants au domicile des personnes âgées sont en constante progression. Plusieurs dispositifs récents visent à contribuer au développement et à la structuration progressive du secteur de l'aide à domicile, ainsi qu'à la professionnalisation de ses intervenants.

La mise en place en 2002 de l'allocation personnalisée pour l'autonomie (APA) a ainsi permis de renforcer la solvabilisation de la demande, pour les personnes âgées de plus de 60 ans et dépendantes (GIR 1 à 4), en complément de l'exonération de charges patronales dont bénéficient les employeurs âgés de 70 ans ou plus depuis 1987, et des aides (aide ménagère, aide sociale) attribuées par les caisses de retraite et les conseils généraux aux personnes âgées disposant de faibles ressources.

Les évolutions législatives récentes cherchent ainsi à structurer l'offre de services d'aide à domicile, la loi du 2 janvier 2002 a intégré les services d'aide au domicile des personnes âgées et handicapées dans le champ de l'action sociale et médico-sociale, au même titre que les établissements. À ce titre, cette activité est soumise à des procédures d'autorisation, de contrôle, de tarification et aux dispositions garantissant les droits des usagers. Une ordonnance de 2004 a étendu le bénéfice de l'exonération de charges sociales aux entreprises et leur a ouvert l'activité de mandataire. L'APA a également un rôle dans la structuration progressive du secteur de l'aide à domicile, le recours à un service prestataire étant recommandé pour les personnes classées en GIR 1 ou 2.

Les pouvoirs publics promeuvent plus globalement la professionnalisation des intervenants et l'attractivité du secteur : mise en place de diplômes spécifiques, lancement du Plan de développement des services à la personne et création de l'Agence nationale des services à la personne. Ils visent également à renforcer la coordination des divers intervenants auprès des personnes dépendantes (centres locaux d'information et de coordination – CLIC – depuis 2001).

Le manque d'informations, tant qualitatives que quantitatives sur les aides à domicile, a conduit la DREES à mener une enquête auprès d'environ 2 600 intervenants au domicile de personnes fragilisées (personnes âgées, handicapées ou toute autre personne nécessitant de l'aide dans l'accomplissement des tâches de la vie quotidienne). Cette enquête s'est déroulée d'avril à juillet 2008, les intervenants ayant été interrogés d'abord par téléphone, afin de vérifier qu'ils appartenaient bien au champ étudié (c'est-à-dire s'occupaient bien de personnes fragiles), puis en face-à-face. L'enquête a notamment pour objectifs :

- de connaître le profil socio-démographique, d'étudier les trajectoires professionnelles et de formation des intervenants ;
- de connaître les conditions d'exercice de leur métier et la nature précise de leurs interventions, en fonction notamment de la situation de leurs employeurs (niveau de perte d'autonomie, isolement, …) et du cadre dans lequel ils interviennent (service prestataire, emploi par des particuliers avec ou sans service mandataire) ; en particulier de recueillir l'opinion des intervenants sur leur métier, leurs conditions de travail et les

difficultés qu'ils peuvent rencontrer dans leur exercice professionnel ;

- de dénombrer les aides à domicile du champ.

Ce document de travail décrit dans le détail le plan de sondage qui a été choisi pour la réalisation de l'enquête, ainsi que les procédures d'apurements et de redressements nécessaires pour produire les données finales.

1. Unités enquêtées, champ de l'enquête, protocole de l'enquête

L'enquête s'intéresse aux intervenants au domicile de personnes fragilisées, c'est-à-dire âgées en perte d'autonomie, handicapées ou toute autre personne nécessitant de l'aide dans l'accomplissement des actes essentiels de la vie quotidienne. Le champ se fixe sur les aides à domicile au sens large, à l'exclusion des professionnels de santé tels que les aides-soignants. Plus précisément, une fois une première sélection des intervenants effectuée, on retient les intervenants suivants *via* une enquête filtre téléphonique :

- travaillant actuellement (pas d'arrêt prolongé d'activité) auprès de personnes fragilisées telles que définies précédemment ;
- se définissant
  - soit comme aide à domicile, aide ménagère, aide médico-psychologique, assistante de vie, auxiliaire de vie sociale, travailleuse familiale ou technicienne de l'intervention sociale et familiale ;
  - soit comme femme de ménage effectuant en sus les courses ou la préparation des repas, ou l'assistance aux personnes ou l'aide aux démarches administratives ;
  - soit comme aide soignante sans disposer du diplôme correspondant ;
- enfin ne s'occupant pas d'un de leur proche en permanence.

Les intervenants retenus dans le champ sont ensuite interrogés en face à face, pour un questionnaire d'une heure environ.

Les unités statistiques de l'enquête sont :

- les intervenants au domicile de personnes fragilisées ;
- une semaine travaillée par chaque intervenant ;
- un jour de référence tiré dans la semaine travaillée des intervenants ;
- une prestation de référence tirée dans le jour dit.

Le tirage du premier type d'unité statistique (intervenants) s'opère en amont des interrogations, et les suivants lors des interviews en face à face, par l'enquêteur lui-même.

Les intervenants au domicile de personnes fragiles peuvent être employés de plusieurs façons :

- par voie directe : l'intervenant est directement salarié de la personne aidée ;
- par voie prestataire : l'intervenant est salarié d'une structure de services à la personne ;
- par voie mandataire : l'intervenant est salarié de la personne aidée, mais l'organisme mandataire s'occupe de la partie administrative et de la mise en relation des deux parties.

Les intervenants sont tirés dans 30 départements (unités primaires) et, le cas échéant, dans les organismes agréés de services à la personne – OASP – dont ils dépendent (unités secondaires, que ces organismes soient prestataires ou mandataires). Les intervenants, salariés des

personnes fragilisées directement ou en passant par un organisme mandataire sont tirés directement dans les 30 unités primaires, *via* une autre base de données.

Le protocole de l'enquête est donc scindé en deux parties distinctes :

1) l'interrogation des intervenants dépendant d'un organisme (salariés ou par voie mandataire) : l'enquêteur se rend dans l'organisme tiré en première phase pour recueillir la liste de ses intervenants, et sélectionne par tirage systématique les intervenants à enquêter ;
2) l'interrogation des intervenants salariés de particuliers employeurs (en emploi direct ou par l'intermédiaire d'un organisme mandataire) : l'enquêteur interroge directement les intervenants tirés.

L'analyse des résultats de l'enquête se fait distinctement entre les types d'unités statistiques enquêtées : intervenant, semaine et jour de référence, intervention de référence.

2. Description du plan de sondage

2.1. Tirage des intervenants

2.1.1. Premier degré : tirage de départements

Pour des raisons de coûts de déplacement des enquêteurs (l'enquête se déroulant en majeure partie en face à face), le protocole d'enquête a prévu de se baser sur 30 départements de France métropolitaine. Dans l'idéal, le tirage de ces départements devrait se faire proportionnellement aux nombres d'intervenants à domicile. Néanmoins, cette information est difficile à obtenir, car il existe des doubles comptes entre la base des organismes agréés de service à la personne et la base des intervenants auprès de particuliers employeurs (la multi-activité étant possible chez les intervenants à domicile, et les intervenants faisant partie d'un service mandataire pouvant être dans les deux bases de sondage utilisées, puisqu'ils ne sont pas salariés à proprement parler d'une structure – voir *infra*), et qu'il n'est, à l'heure actuelle, pas possible de quantifier ces doubles comptes.

À défaut de pouvoir compter les intervenants, un autre critère de taille a été retenu, que l'on a supposé bien corrélé avec ce nombre : le nombre de bénéficiaires des aides à domicile des départements[1], ces aides étant allouées aux personnes handicapées et aux personnes âgées de plus de 60 ans en perte d'autonomie.

La probabilité de sélection du département d et disposant d'un nombre de bénéficiaires de l'aide sociale X_d est donnée par :

$$\pi_d = 30.\frac{X_d}{\sum_{d=1}^{96} X_d}$$

Pour des départements disposant de nombreux bénéficiaires, il peut arriver que la grandeur π_d ainsi calculée dépasse 1, ce qui pose problème dans la mesure où il s'agit d'une probabilité de tirage. Dans le tirage présent, 3 départements étaient dans ce cas. Ce sont des départements de grande taille, et donc des départements que l'on souhaite tirer. Pour ces départements on pose donc $\pi_d = 1$, ce qui équivaut à les tirer d'office. Ensuite, on recalcule les probabilités π_d sur les départements restants, en excluant du dénominateur les départements sélectionnés d'office.

Une fois le tirage des départements effectué, on vérifie que les 30 départements tirés au sort présentent une certaine diversité en regard de quelques indicateurs reflétant les caractéristiques des départements pouvant influer sur les conditions d'exercice du métier d'aide à domicile auprès de personnes fragilisées : nombre de bénéficiaires de l'allocation personnalisée d'autonomie et de bénéficiaires des prestations de compensation du handicap, d'allocations compensatrices tierce personne ou d'autres aides départementales à destination

[1] Données disponibles dans l'enquête annuelle sur l'aide sociale départementale de la DREES.

des personnes âgées ou handicapées vivant à leur domicile, pourcentage de personnes âgées de plus de 60 ans et de plus de 75 ans dans la population du département, répartition globale de la population. Pour s'assurer une bonne représentativité des départements, on effectue, tout en tenant compte des probabilités précédemment calculées, un tirage systématique[2], en contrôlant à la fois le nombre de bénéficiaires de l'aide sociale départementale à domicile et les dépenses associées.

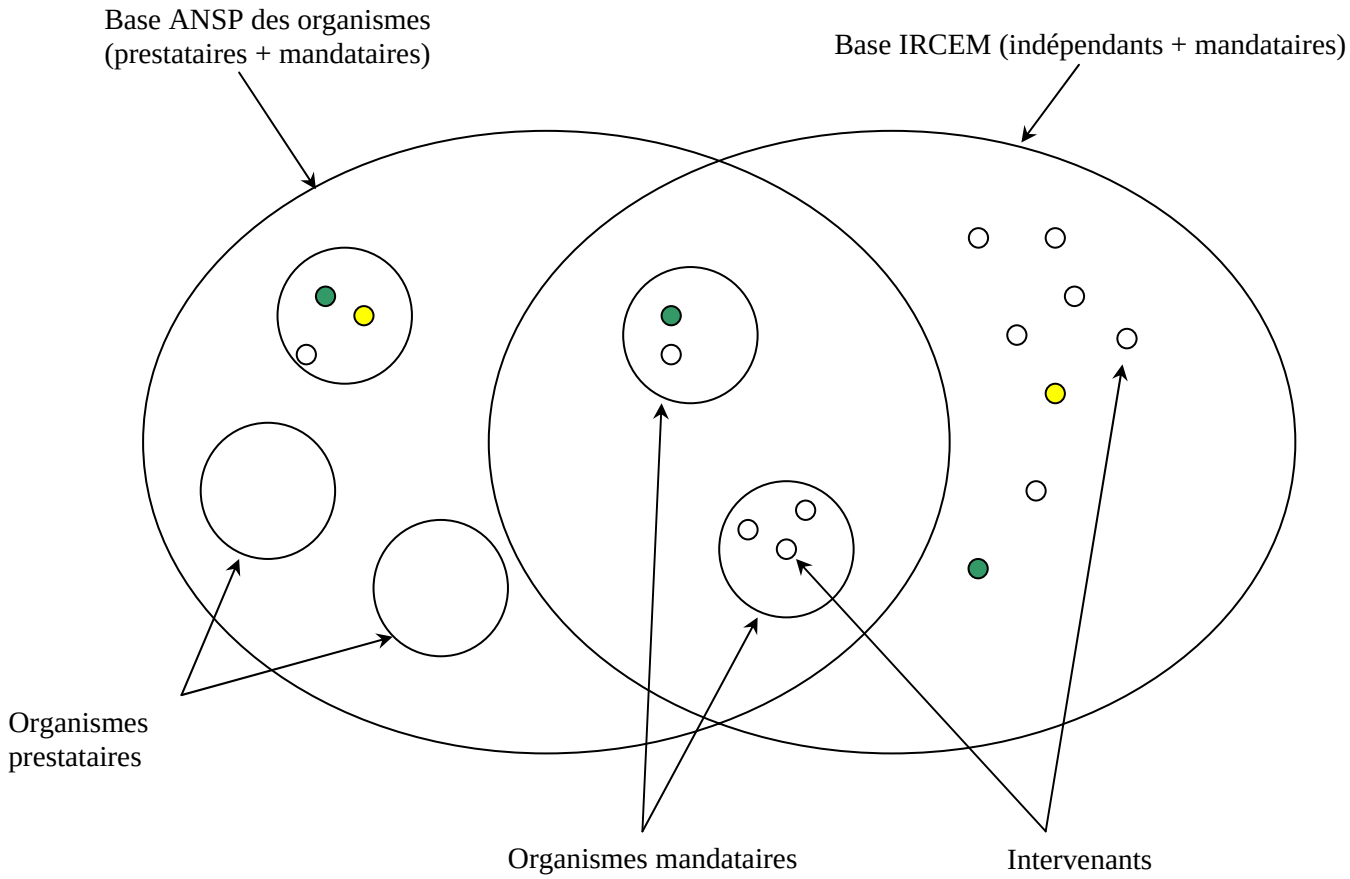
2.1.2. Tirage des intervenants : tenir compte des deux bases de sondage

Les intervenants et organismes contactés sont tous issus des trente départements sélectionnés au premier degré.

La difficulté vient ici du fait que deux bases de sondage sont disponibles, se recouvrant partiellement (schéma) :

- la base de l'IRCEM, qui recense les intervenants employés directement par des particuliers, mais également les intervenants dépendant d'organismes mandataires (auquel cas ils sont payés les particuliers) ;
- la base de l'ANSP, qui recense les organismes agréés de service à la personne (OASP), tant prestataires (auquel cas les intervenants sont salariés de ces organismes) que mandataires.

[2] Ce qui est possible même dans le cas d'un tirage à probabilités inégales.



Le même intervenant travaille comme indépendant pour certaines personnes et est salarié d'un organisme prestataire pour d'autres interventions.

Le même intervenant travaille à la fois comme indépendant pour certaines personnes, par l'intermédiaire d'un organisme mandataire pour d'autres, et est salarié d'un organisme prestataire.

On ne connaît pas la proportion de recouvrement entre les deux bases, c'est-à-dire qu'*a priori*, il est impossible de savoir, dans la base IRCEM, si les individus sont indépendants et/ou dépendent d'un organisme mandataire ou encore sont multi-actifs entre salariat chez des particuliers employeurs et salariat pour un prestataire, et dans la base des organismes, si des intervenants qui en dépendent travaillent également directement auprès de particuliers.

Il existe donc un risque pour que des individus soient tirés deux fois (ce risque étant cependant très faible compte tenu des tailles des bases de données – on compte en tout environ 1,3 millions d'intervenants toutes catégories confondues[3], aide aux personnes fragiles ou autres –), et même si un intervenant n'est tiré qu'une seule fois, il faut toujours considérer qu'il aurait pu être tiré deux fois, voire plus. Ne pas tenir compte de cette information reviendrait à ne pas tenir compte des doubles comptes, et donc à sur-représenter les intervenants multi-actifs.

[3] Données DARES et BIPE.

Une façon de résoudre ce problème est d'appliquer la méthode du partage des poids, en posant des questions adaptées lors des interviews. Le protocole est alors le suivant :

- tirage par la DREES des intervenants dans la base IRCEM[4] ;
- tirage par la DREES des organismes dans la base ANSP ;
- tirage par le prestataire d'enquête des intervenants dans chaque organisme ;
- repondération par partage des poids en tenant compte des doublons (la méthode est détaillée en partie 5.3).

Au total, on cherche à interroger de 2500 à 3000 professionnels intervenant à domicile. Le test de l'enquête a révélé les éléments suivants :

- concernant la base fournie par l'IRCEM, les numéros de téléphone des intervenants ne sont pas disponibles. 41 % environ ont pu être récupérés (les 59 % restants sont alors considérés comme non répondants). *A posteriori*, la DREES ayant effectué des recherches complémentaires de numéros de téléphone, 63 % des intervenants du fichier ont eu un numéro. Parmi ceux-ci, seuls 18 % ont été joints, ont été qualifiés comme faisant partie du champ de l'enquête et ont accepté le face à face, alors que tous étaient des intervenants travaillant chez des personnes âgées ou dépendantes. Au total, 11,3 % de l'échantillon de départ a pu être interrogé pour toutes les étapes de l'enquête. En extrapolant avec le taux de numéros de téléphone uniquement retrouvés par le prestataire d'enquête, cela revient à estimer que 7,3 % de l'échantillon seulement aurait pu être interrogé tout en faisant partie du champ pour le face à face ;
- concernant la base fournie par l'ANSP, 60 % des organismes ont donné leur accord de participation et ont été interrogés, et 58 % des intervenants sélectionnés ont répondu entièrement à l'enquête.

On se base donc sur ces taux pour anticiper des taux globaux de réponse et d'individus dans le champ pessimistes. Il est par ailleurs impossible à ce stade de savoir combien d'intervenants directs auprès de particuliers travaillent effectivement pour des personnes fragilisées et combien d'intervenants d'organismes ont une partie de leur activité dévouée à ce type de personne. À défaut de propratiser, on tente d'interroger autant d'intervenants issus de la base IRCEM que de la base ANSP (la représentativité sera toutefois assurée par les poids de sondage différents et par les redressements *a posteriori*). Pour obtenir 2 500 à 3 000 interviews complètes, ceci conduit au final à sélectionner entre 17 100 et 20 500 intervenants dans la base IRCEM, entre 720 et 860 organismes dans la base ANSP et 5 intervenants dans chacun des organismes.

2.1.3. Tirage des intervenants issus de la base IRCEM (indépendants et mandataires)

Il s'agit ici de tirer directement les intervenants, le tirage est donc à un seul degré au sein de chaque département.

[4] Les intervenants de la base IRCEM n'étant recensés qu'une seule fois (évitant ainsi les doubles comptes au sein de cette base, ce qui n'est pas forcément le cas dans la base ANSP), il n'est pas nécessaire à ce niveau de connaître le nombre de personnes auprès desquelles ils interviennent sans passer par un organisme.

On tire donc tout d'abord un ensemble de 20 545 intervenants qui comprendra à la fois l'échantillon principal et deux échantillons de réserve, ceux-ci étant destinés à pallier un éventuel taux de non réponse trop important par rapport à celui anticipé (encadré 1). Le lot principal comprend 17 090 intervenants. Les lots complémentaires comprennent chacun environ : (20 545-17 090)/2=1 728 intervenants[5].

Encadré 1 : tirage d'un échantillon principal et d'échantillons de réserve

Le tirage d'un échantillon principal S_p et d'un échantillon supplémentaire de réserve S_r s'effectue comme suit :

1) tirage d'un échantillon global S_G tel que $card(S_G) = card(S_p) + card(S_r)$

2) tirage au sein de S_G de l'échantillon principal S_p

3) S_r est obtenu par déduction de S_p dans S_G

On a donc $S_G = S_p \cup S_r$ et $S_p \cap S_r = \varnothing$.

Dans la pratique, S_r importe peu puisque si l'on active au final cet échantillon, on utilise en fait l'échantillon global S_G.

On en revient à tirer plusieurs échantillons imbriqués, ce qui n'est pas sans soulever certaines difficultés dans le calcul des probabilités d'inclusion. En effet, il faut respecter les propriétés suivantes :

1) $k \in S_p \Rightarrow k \in S_G$, où k est l'individu tiré ;

2) les probabilités d'inclusion de l'individu k sont calculées sur la population entière, de taille N, quel que soit l'échantillon considéré ;

3) le plan de sondage menant *in fine* aux deux échantillons est le même, c'est-à-dire que les probabilités d'inclusion finales, peuvent être retrouvées par le même mode de calcul.

Par exemple, dans le cas d'un sondage à probabilités proportionnelles à une grandeur X, les probabilités de sélection de l'individu k dans les échantillons S_G (de taille $n_G = n_p + n_r$) et S_p (de taille n_p) sont au final :

$$P(k \in S_G) = n_G \frac{X_k}{\sum_{k=1}^{N} X_k}$$

$$P(k \in S_p) = n_p \frac{X_k}{\sum_{k=1}^{N} X_k}$$

Dans la pratique, le théorème de Bayes apporte une solution bien commode pour respecter les trois conditions : partant de la propriété (1), on peut écrire : $P(k \in S_p \cap k \in S_G) = P(k \in S_p)$, autrement dit tirer un individu dans S_p revient à le tirer à la fois dans S_p et dans S_G.

Le théorème de Bayes indique que : $P(k \in S_p / k \in S_G) = \frac{P(k \in S_p \cap k \in S_G)}{P(k \in S_G)}$

[5] Ces nombres diffèrent légèrement de l'objectif théorique, du fait des contraintes de calculs de tailles d'échantillons dans les différentes strates (arrondies à des nombres entiers).

Soit dans notre cas : $\boxed{P(k \in S_p / k \in S_G) = \frac{P(k \in S_p)}{P(k \in S_G)}}$

Dès lors, pour tirer l'individu k dans S_p sachant qu'il a été tiré dans S_G, il suffit de se servir de cette dernière formule : on effectue le tirage de l'échantillon principal, à partir de l'échantillon global, en injectant $P(k \in S_p / k \in S_G)$ dans un tirage à probabilités inégales[6] (voir annexe 2 pour un exemple de tirage avec échantillon de réserve).

Cette méthode se généralise aisément dans le cas de plusieurs échantillons de réserve : on tire d'abord les individus de l'échantillon global, puis ceux des n échantillons de réserve et de l'échantillon principal dans l'échantillon global à l'aide de la formule vue *supra*, puis ceux des n-1 échantillons de réserve et de l'échantillon principal dans les n précédents échantillons (plus le principal), etc.

L'idée d'un tirage proportionnel à l'activité des intervenants est écartée, compte tenu des nombreux domaines d'intérêt de l'enquête. Si un tel tirage était effectué, la précision des estimations jouerait dans les deux sens :

- elle serait améliorée pour tous les indicateurs corrélés à l'activité (emploi du temps, etc…) ;
- elle serait au contraire dégradée pour les indicateurs peu ou pas corrélés avec l'activité (caractéristiques des personnes aidées, des intervenants, etc…).

Choisir un tirage uniforme est plus pertinent, assurant ainsi une précision « moyenne » relativement bonne. On peut en revanche améliorer le tirage, en utilisant deux variables de stratification : l'ancienneté dans le métier (approximée par l'année d'adhésion à l'IRCEM), et le nombre d'employeurs de l'intervenant[7], ces deux variables étant scindées en 4 tranches chacune.

Ces facteurs seront utilisés dans l'analyse, c'est pourquoi ils sont pris en compte au moment du tirage des intervenants.

Il faut par ailleurs que la répartition des intervenants dans chaque strate de l'échantillon soit la même que dans les strates de la population entière issue de la base IRCEM. Il s'agit ici de faire un tirage par allocation proportionnelle selon les nombres d'intervenants auprès de particuliers employeurs par strate. On aura donc le même taux de sondage par strate h : $\frac{n_h}{N_h} = \frac{n}{N}$. En revanche, compte tenu des complexités liées à la gestion des réseaux d'enquêteur et pour des raisons de réduction de la variance, le mieux est de sélectionner le même nombre d'intervenants par département, c'est-à-dire environ $20\,500/30 \approx 685$ en incluant les deux échantillons de réserve. Ce qui revient à appliquer un critère de stratification supplémentaire.

[6] Attention cependant à penser à recalculer les probabilités de sélection après tirage (qui interviennent de fait dans le calcul des pondérations) : SAS donne les probabilités injectées dans le tirage $\frac{P(k \in S_p)}{P(k \in S_G)} \neq P(k \in S_p)$.

[7] Ce qui revient en partie à tenir compte de l'activité, mais on n'est pas dans le cadre d'un tirage à probabilités proportionnelles à celle-ci.

Indépendamment des départements, dans chaque strate h et département d ayant un nombre d'intervenants employés par des particuliers $N_{1,dh}$, on sélectionne ainsi le nombre d'intervenants suivants :

$$n_{1,dh} = 685.\frac{N_{1,dh}}{\sum_{h=1}^{H} N_{1,dh}}$$

Ce calcul peut toutefois poser problème, si certaines strates ont moins de $n_{1,dh}$ intervenants disponibles. Pour ces strates, on force $n_{1,dh}$ au nombre d'intervenants disponibles et on recalcule $n_{1,dh}$ pour les autres strates.

La probabilité d'inclusion de chaque intervenant k de la strate h du département d est donc donnée par :

$$\pi_{1,dhk} = \frac{n_{1,dh}}{N_{1,dh}} = \frac{685.\frac{N_{1,dh}}{\sum_{h=1}^{H} N_{1,dh}}}{N_{1,dh}} = \frac{685}{\sum_{h=1}^{H} N_{1,dh}}$$

et donc :

$$\boxed{\pi_{1,hdk} = \frac{685}{N_{1,d}}, \quad \forall h}$$

… ce qui est assez sympathique, la probabilité d'inclusion ne dépendant pas de la strate, mais uniquement de la taille du département en terme de nombre d'intervenants. Ceci assure donc une répartition des poids des intervenants homogène au sein de chaque département.

Au final dans l'enquête, on n'a pas eu besoin d' « activer » les échantillons de réserve. 17 090 intervenants directs ont donc été sélectionnés (il faut donc remplacer le 685 par 570 dans les formules précédentes), pour obtenir 1 842 interviews en face à face exploitables (intervenants sélectionnés, ayant passé le filtre téléphonique, et ayant répondu au face à face).

La pondération initiale de l'intervenant sélectionné par strate et par département est le produit de la sélection du département puis de l'intervenant employé par un ou des particuliers au sein des départements tirés au premier degré :

$$w_{1,dhk} = \frac{1}{\pi_{1,hdk}.\pi_d}$$

Cette pondération sera ensuite modifiée, au moment du redressement de la non-réponse, pour tenir compte du recouvrement des deux bases de données (IRCEM+ANSP, voir partie 5.3 sur

le partage des poids).

2.1.4. Tirage des intervenants issus de la base des organismes (prestataires et mandataires)

Les coordonnées des intervenants n'étant pas disponibles dans cette base de sondage, le tirage doit ici s'effectuer sur deux degrés :

a. tirage des organismes ;
b. tirage des intervenants dépendant de ces organismes.

Pour minimiser l'effet grappe, on cherche à tirer un maximum d'organismes au premier degré, et moins d'intervenants au deuxième degré. Cependant, sélectionner trop d'organismes supposerait de gérer de nombreuses listes d'intervenants, travail particulièrement fastidieux pour le prestataire d'enquête. Le protocole d'enquête prévoit d'interroger 5 intervenants par organisme, ce qui implique de tirer 520 organismes pour sélectionner les 2 600 salariés ou mandatés par les OASP souhaités (ce chiffre tient compte de la non réponse des intervenants). Néanmoins, en plus de la non réponse des intervenants, il faut également tenir compte de la non réponse des organismes. Si on estime celle-ci à un taux de 40 %, cela revient à sélectionner au total 870 organismes.

On tire tout d'abord un ensemble de 870 organismes qui comprend à la fois l'échantillon principal et deux échantillons de réserve. Le lot principal comprend 720 organismes. Les lots complémentaires comprennent chacun environ : (870-720)/2=75 organismes[8].

Pour les mêmes raisons que précédemment, il faudrait au final interroger le même nombre d'intervenants dans chaque département. De ce fait, le nombre d'organismes sélectionnés doit également être identique par département, soit 870/30=29.

Dans l'idéal, pour chaque département (pris ici comme une strate), le nombre d'organismes sélectionnés devrait être proportionnel au nombre d'intervenants dépendant de ces organismes. La difficulté vient ici du fait que le nombre d'intervenants (et même de salariés) n'est pas connu pour tous les organismes. On a alors fait le choix de séparer l'échantillon en deux strates :

- la première avec les effectifs connus, en effectuant un tirage proportionnel au nombre de salariés de l'organisme ;
- la deuxième avec les effectifs inconnus, en effectuant un tirage uniforme des organismes.

Il faut d'abord déterminer, pour chaque département, le nombre d'organismes à tirer dans chaque strate effectifs connus/inconnus. Ce nombre sera calculé par allocation proportionnelle au nombre d'organismes par strate et département $M_{2,dh}$:

[8] Ce qui diffère là aussi légèrement de l'objectif théorique, du fait des contraintes de calculs de tailles d'échantillons dans les différentes strates (arrondies à des nombres entiers).

$$m_{2,dh} = 29.\frac{M_{2,dh}}{\sum_{h=1}^{H} M_{2,dh}}$$

Devant la faiblesse de l'échantillon par département, on peut se retrouver avec des $m_{2,hd}$ nuls. Il faut donc veiller à ce que l'on ait bien $\sum_{d=1}^{30} m_{2,dh} = m_{2,h}$ tel que $\frac{m_{2,h}}{M_{2,h}} = \frac{m_2}{M_2}$, totaux issus de l'ensemble des départements sélectionnés.

1) Tirage dans la strate des effectifs connus

Dans le cas où les effectifs salariés des organismes sont connus, le tirage au sein de chaque département se fait à probabilités inégales, proportionnellement aux nombres de salariés dépendant des organismes. La probabilité d'inclusion de chaque organisme j du département *d* est donnée par :

$$\boxed{\pi_{21,dj} = m_{21,d} \frac{N_{21,dj}}{\sum_{j=1}^{J_d} N_{21,dj}}}$$

On doit par ailleurs encore s'assurer que toutes les probabilités de tirage du second degré soient inférieures à 1 pour l'ensemble des organismes. Pour s'assurer une bonne représentativité des organismes, on effectue, tout en tenant compte des probabilités précédemment calculées, à un tirage systématique, en contrôlant à la fois les effectifs salariés et la catégorie juridique.

2) Tirage dans la strate des effectifs inconnus[9]

Pour les organismes dont l'effectif salarié est inconnu, le tirage au sein de chaque département se fait par tirage uniforme systématique, en contrôlant la catégorie juridique. La probabilité d'inclusion de chaque organisme j est donnée par :

$$\boxed{\pi_{22,dj} = \frac{m_{22,d}}{M_{22,d}}}$$

Au final dans l'enquête, les deux échantillons de réserve ont été utilisés, soit 870 organismes (860+10 pour l'arrondi dans le nombre sélectionné par département), et même ceci fait, seuls 263 ont répondu à l'enquête. Le taux de non réponse a donc été largement sous-estimé lors du test de l'enquête.

[9] Les organismes dont l'effectif de salariés était nul ou supérieur à 300 (peu plausible *a priori*) ont été mis dans la strate des effectifs inconnus. Las, l'histoire nous apprendra après l'enquête que les organismes avec plus de 300 intervenants existent bel et bien (allant même jusqu'à plus de 2 000 intervenants !). Une troncature des poids a dû être opérée par la suite, pour limiter la trop grande dispersion de ceux-ci (encadré 3).

Au sein des organismes, on sélectionne les intervenants par tirage uniforme systématique, sur la liste des intervenants triée par ordre alphabétique, cette liste étant recueillie directement par l'enquêteur au moment de sa visite dans l'organisme. Au plus 5 intervenants sont tirés dans chaque organisme sélectionné, moins si l'organisme n'a que très peu d'intervenants.

Si la liste d'intervenants d'un organisme comprend T_j individus, on calcule d'abord le pas qui vaut la partie entière du rapport $\frac{T_j}{Min(5,T_j)}$. On sélectionne ensuite les intervenants de la façon suivante :

- tirage aléatoire du premier intervenant, en le sélectionnant au hasard entre le 1er et le $\frac{T_j}{5}$ème de la liste (pour les organismes disposant de moins de 5 salariés, tous les intervenants sont sélectionnés) ;
- tirage des intervenants suivants en rajoutant $\frac{T_j}{5}$ au numéro d'ordre après chaque sélection.

La pondération initiale de l'intervenant ainsi sélectionné par strate, département et organisme est le produit de la sélection du département, puis de l'organisme et de l'intervenant qui en dépend :

$$w_{2,dhjk} = \frac{1}{\pi_{2,dhj}.\pi_d}.\frac{T_j}{Min(5,T_j)}$$

Cette pondération est toutefois beaucoup plus variable que dans le cas des intervenants directs, du fait des probabilités de sélection d'intervenants inconnues avant l'enquête : cette pondération est très forte pour les très gros organismes, et à l'inverse très faible pour les plus petits.

Cette pondération sera ensuite modifiée, au moment du redressement de la non-réponse, pour tenir compte du recouvrement des deux bases de données (IRCEM+ANSP, voir partie 5.3 sur le partage des poids).

2.2. Les tirages lors des face à face

Le questionnaire d'enquête prévoit plusieurs modules spécifiques : l'intervenant doit ainsi décrire une semaine et un jour de référence, ainsi qu'une personne chez qui il intervient. Ces situations ont elles-mêmes fait l'objet d'échantillonnages, mais au cours de l'interview, et non pas en amont.

2.2.1. Tirage de la semaine de référence

Au cours de l'entretien, l'intervenant est amené à décrire une semaine type de son activité, dite « semaine de référence ». La sélection de cette semaine ne doit pas être trop compliquée pour l'enquêté comme pour l'enquêteur, et surtout doit rester proche de la date d'entretien, car

dans ce cas l'effet mémoire joue à plein. En effet, il est plus facile de décrire la semaine passée que celle ayant eu lieu deux mois plus tôt.

Le protocole d'enquête prévoit d'interroger l'intervenant sur la dernière semaine travaillée au cours du mois précédent, en partant de la date d'enquête. On entend par semaine travaillée une activité minimum, à savoir l'intervention chez au moins une personne fragilisée.

Ce protocole de sélection semble correct, même s'il ne tient pas compte d'une possible saisonnalité, et notamment de la présence des jours fériés du mois de mai, au milieu de la période d'enquête. Ceci étant, cela permet également de mesurer l'activité des intervenants les jours fériés, et ceci d'autant plus que les personnes aidées, étant par définition de l'enquête des personnes âgées ou handicapées, restent probablement chez elles ces jours particuliers, et ont donc également besoin d'aide.

La pondération de la semaine de référence est égale à celle de l'intervenant sélectionné, puisqu'on ne cherche pas à décrire la situation d'un mois ou d'une année type de référence (auquel cas il faudrait multiplier les poids respectivement par 4,5 et 52).

2.2.2. Tirage du jour de référence

Au même titre que pour la semaine de référence, les intervenants sont amenés à décrire une journée type. Cette journée est sélectionnée dans la semaine de référence. Là encore, le protocole de sélection ne doit pas être trop fastidieux, et surtout ne doit pas être laissé à la discrétion de l'enquêteur ou de l'enquêté.

La solution retenue est de tirer un jour au hasard parmi les jours travaillés, par tirage uniforme, et interroger l'intervenant dessus. Ceci suppose de demander au préalable à l'intervenant quels jours il a travaillé pendant la semaine de référence (plus précisément quels jours il s'est déplacé au moins une fois chez une personne fragilisée). Ainsi, tous les intervenants devront décrire un jour de référence, et d'autre part il n'y a pas de risque de mauvaise représentativité de chaque jour, puisque chaque jour tiré est par définition un jour actif.

La probabilité de tirage du jour j de l'intervenant k est donc :

$$\boxed{\pi_{kj} = \frac{1}{J_k}}$$

où J_k correspond au nombre total de jours travaillés de l'intervenant k lors de la semaine de référence, variable selon l'intervenant considéré.

La pondération des jours de référence tirés pour chaque intervenant est égale à l'inverse de la probabilité de sélection de ce jour, multiplié par la pondération de l'intervenant soit :

$$w_{kj} = \frac{1}{\pi_{kj}}.w_k = J_k.w_k$$

2.2.3. Tirage de la personne/prestation de référence

Remarque préliminaire : si le tirage correspond bien à une personne, l'exploitation finale est plus à appréhender en termes de prestation, une même personne pouvant être vue plusieurs fois par l'intervenant dans la journée. On ne pourra pas, *a priori*, avoir de résultats du type « XXXX personnes sont aidées ». On tire donc plus une prestation qu'une personne de référence. Si l'on voulait un résultat de ce type, il eût fallu gérer le fait qu'une personne puisse être sélectionnée au travers de différentes prestations ou intervenants, et poser des questions à ce sujet, ce qui aurait alourdi et pour le moins complexifié le questionnement.

Même si les intervenants à domicile travaillent pour la plupart à temps partiel, il n'en reste pas moins que les personnes auprès desquelles ils viennent en aide peuvent être nombreuses. Aussi est-il exclu d'interroger l'intervenant pour chaque personne chez qui il se déplace (pour cause de coût de l'enquête et de pénibilité dans l'interrogation). On a donc choisi de n'interroger l'intervenant que sur une seule personne aidée.

La sélection de la personne de référence est délicate : en effet, il paraît difficile de demander à un intervenant de parler de la personne n° 2 qu'il est allée voir dans la semaine, ou de la personne n° 5, etc., sachant que dans ce cas, l'effort de mémoire est sans doute trop important.

Le protocole initial proposait d'interroger chaque intervenant sur deux personnes aidées : la personne chez qui il intervient depuis le plus longtemps, et celle chez qui il intervient depuis le moins longtemps. Ceci aurait pour avantage au moins de tenir compte des différences de liens relationnels qui peuvent exister entre l'intervenant et les personnes aidées. En revanche, on ne peut pas vraiment étudier un « individu moyen » (ou plutôt une « prestation moyenne ») par cette méthode.

Une solution alternative est de tirer une personne au hasard dans le jour de référence. Dans le questionnaire, l'intervenant doit décrire très précisément le jour de référence (horaires des visites, ...). Il ne doit donc pas avoir de difficultés pour visionner telle ou telle intervention de la journée. Il suffit alors à l'enquêteur de tirer au hasard un numéro d'ordre, et d'annoncer la partie du questionnaire correspondante de la façon suivante : « Nous allons maintenant parler de la personne que vous avez vue à 14H » par exemple. Cette méthode permet par ailleurs de limiter la description à une seule personne, ce qui réduit le temps de passation du questionnaire. C'est cette solution qui a été choisie au final.

La probabilité de tirage de la prestation i du jour j de l'intervenant k est alors :

$$\boxed{\pi_{kji} = \frac{1}{I_{kj}}}$$

où I_{kj} correspond au nombre total d'interventions du jour j de l'intervenant k, variable selon l'intervenant et le jour considéré.

La pondération des prestations de référence tirées pour chaque intervenant et jour sélectionnés est égale à l'inverse de la probabilité de sélection de l'intervention, multiplié par la pondération finale du jour soit :

$$w_{kji} = \frac{1}{\pi_{kji}}.w_{kj} = I_{kj}.J_k.w_k$$

3. Apurement des données

A réception des données, il convient de repérer les situations aberrantes et dans la mesure du possible, de les corriger. La correction de ces aberrations s'est faite pour chaque questionnaire d'enquête : au niveau des organismes, des interviews téléphoniques des intervenants, et des interviews en face à face.

Dans tous les cas, on s'est basé sur une stratégie « prudentielle » : en effet, la réalité peut être très complexe et une situation en apparence impossible peut en fait exister. Seules les données extrêmes ont donc été redressées lors de cette étape.

Par ailleurs, certaines données indispensables à l'analyse n'ont pas toujours été renseignées lors de l'enquête. Un redressement de la non réponse partielle a donc été effectué en même temps que l'apurement.

3.1. Corrections des données relatives aux organismes

Avant l'arrivée d'un enquêteur dans l'organisme, celui-ci devait tout d'abord remplir un coupon-réponse indiquant sa forme juridique, son type d'agrément (simple, qualité, ou simple et qualité), le nombre de personnes fragiles bénéficiaires des intervenants du service et le nombre d'intervenants associé, ces deux derniers indicateurs étant demandés pour le mode prestataire uniquement, pour le mode mandataire uniquement, et pour les deux modes à la fois.

Le statut juridique ainsi que le type d'agrément n'étant pas systématiquement renseignés, un redressement a été opéré directement à partir de la base de sondage de l'ANSP.

Certains organismes ont visiblement confondu les bénéficiaires et intervenants sous les deux statuts à la fois avec la somme de ceux issus du service prestataire et ceux issus du service mandataire. Pour ces organismes, les nombres de bénéficiaires du service et d'intervenants sous les deux statuts à la fois ont été mis à 0[10]. Cette opération concerne 3 organismes sur 263 pour ce qui est des intervenants, et 7 pour ce qui est des bénéficiaires.

Par ailleurs, si tous les organismes déclarent disposer d'intervenants auprès de personnes fragiles, certains déclarent n'avoir aucune personne fragile parmi leur clientèle. On suppose dans ce cas qu'il s'agit d'une non réponse. La correction de celle-ci se fait par hot-deck métrique, en imputant une valeur d'un organisme « proche » :

a. on trie les données selon le type d'agrément, la forme juridique et le nombre d'intervenants du statut donné ;
b. on impute le nombre de bénéficiaires du service sous ce statut en prenant la valeur de l'organisme répondant précédent dans la liste.

[10] Ce choix n'est d'ailleurs pas neutre, au vu du peu d'organismes répondants, et donc des poids finals élevés. Enlever quelques dizaines d'intervenants à l'un d'eux peut baisser considérablement l'estimation du total d'intervenants dépendant d'organismes.

Cette opération se répète pour chaque mode d'intervention. Au final, 5 valeurs sur 263 ont été imputées pour les bénéficiaires d'un service prestataire uniquement, 2 pour les bénéficiaires d'un service mandataire uniquement, et une seule pour les bénéficiaires sous les deux statuts, prestataire et mandataire.

3.2. Correction des données des interviews téléphoniques auprès des intervenants

Trois variables ont été corrigées dans la base des interviews téléphoniques : l'âge, le nombre de personnes composant le foyer et le nombre d'enfants à charge des intervenants.

- Les âges déclarés de moins de 20 ans ou plus de 70 ans (queues de la distribution des âges) ont été considérés comme suspects. De ce fait, on a comparé la valeur de l'âge déclaré avec la base de sondage de l'IRCEM – dans le cas des intervenants en emploi direct – et avec les réponses données par les organismes sur les intervenants tirés au sort. Cette comparaison a également permis de redresser de la non réponse partielle. Après correction, il apparaît tout de même que des intervenants très jeunes ou particulièrement âgés exercent le métier : leur âge s'étend en effet de 17 à 80 ans.

- Les intervenants ayant déclaré avoir plus de 10 personnes dans leur foyer (y compris eux-mêmes) ont également fait l'objet d'une correction : on a imputé à cette variable la valeur suivante :

  1 (pour l'intervenant lui-même) ;
  + 1 si l'intervenant vit en couple ;
  + le nombre d'enfants à charge ;
  + le nombre d'autres personnes à charge vivant au domicile de l'intervenant.

- De même, on a plus finement regardé les intervenants déclarant avoir plus de 10 enfants. Tous ont indiqué des nombres par tranches d'âge beaucoup trop élevés (par exemple 15 enfants entre 12 et 18 ans). Néanmoins, n'ayant pas d'information auxiliaire pertinente pour appliquer des valeurs par tranches d'âge aux nombres d'enfants, celles-ci ont été mises en non réponse.

3.3. Correction des données des interviews en face à face auprès des intervenants

3.3.1. Nombres de personnes visitées et des heures travaillées

Par rapport aux nombres de personnes visitées la semaine de référence et au nombre d'heures travaillées, plusieurs problèmes apparaissent :

- 2 intervenants indiquent n'avoir travaillé auprès d'aucune personne la semaine de référence, alors même qu'ils parlent de la dernière semaine travaillée en tant qu'aide à domicile. Ces deux intervenants indiquent de plus n'avoir travaillé qu'un seul jour dans la semaine. Par ailleurs :

a. Un de ces intervenants a indiqué n'avoir vu aucune personne au cours de la journée de référence mais avoir tout de même fait une intervention ; celui-ci n'a par ailleurs quasiment pas répondu aux questions sur l'intervention de référence (hormis le lieu du déjeuner). Pour cet intervenant, on a finalement indiqué qu'il n'avait pas travaillé en tant qu'aide à domicile au cours du mois qui venait de s'écouler (partie B du questionnaire mise à blanc) ;

b. Le deuxième intervenant a indiqué avoir effectué une visite à domicile au cours de la journée de référence, auprès d'une personne fragile. On a donc choisi le type de personne visitée lors de la semaine de référence (ce qui revient au même que lors de la journée de référence, étant donné qu'il n'a travaillé qu'un jour) et le nombre d'heures travaillées auprès de cette personne en prenant les valeurs d'un individu proche, c'est-à-dire une aide à domicile étant intervenue un seul jour la semaine de référence et auprès d'une seule personne fragilisée. Les heures travaillées auprès de cette personne ont été calculées à partir de l'emploi du temps de la journée.

- À l'inverse, d'autres intervenants indiquent s'être occupé d'un nombre anormalement important de personnes lors de la semaine de référence (jusqu'à 150). La valeur seuil de 30 personnes visitées en une semaine (correspondant au 99ème centile) a été utilisée pour analyser les valeurs extrêmes. Plusieurs cas ont été repérés :
-

a. Des erreurs de CAPI possibles : dans plusieurs cas, une valeur non nulle relativement élevée a été affectée au nombre d'interventions auprès de tous les types de personne (âgée, handicapée, malade,…), alors même que l'intervenant a indiqué n'avoir effectué aucune heure de travail dans quatre catégories sur cinq[11]. Dans ce cas, la valeur affectée au nombre de personnes visitées dans la semaine a été fixée à 0 pour les catégories à heures nulles, et on a gardé la valeur initiale pour la catégorie avec un nombre d'heures non nul ;

b. Des erreurs de saisie : certaines réponses sont clairement extravagantes (96 personnes âgées visitées en une semaine pour 0 heure travaillée auprès de ce type de personne par exemple) et ont été corrigées par hot-deck métrique en prenant comme critère de classification le nombre de jours travaillés dans la semaine de référence et le nombre d'heures auprès de chaque catégorie de personne. Les cas de plus faible suspicion d'erreur n'ont pas été redressés.

- Enfin, des intervenants indiquent un nombre d'heures travaillées lors de la semaine de référence trop important. 5 intervenants indiquent ainsi travailler plus de 24 heures par jour auprès de personnes fragiles (sans compter une autre activité éventuelle). On peut assimiler ces valeurs aberrantes à des erreurs de saisie de la part de l'enquêteur. Ces valeurs extrêmes ont été corrigées selon deux méthodes :

[11] Par exemple l'intervenant indique avoir travaillé auprès de 7 personnes âgées, 7 enfants handicapés, 7 adultes handicapés, 7 autres personnes fragiles, et 7 personnes non fragiles alors que les heures de travail associées sont de 21, 0, 0, 0 et 0.

a. Pour ceux qui n'ont travaillé qu'un seul jour dans la semaine en tant qu'aide à domicile, en se basant sur l'emploi du temps de la journée (décrit plus loin dans le questionnaire) ;

b. Pour les autres, par hot-deck métrique, en prenant comme critères de classification le nombre de jours travaillés lors de la semaine de référence et le nombre de personnes visitées par catégorie (âgée, handicapée, etc.).

En prenant en compte les heures travaillées pour une autre activité que l'aide à domicile lors de la semaine de référence, il peut également arriver que l'on dépasse les 24 heures par jour. Ce cas de figure peut cependant se présenter dans le cas où l'intervenant s'occupe 24h/24 d'une personne et exerce une autre activité par ailleurs. Aucune correction n'a été effectuée pour la question de l'activité complémentaire.

Certains intervenants ont indiqué n'avoir rendu visite à aucune personne le jour de référence[12], alors même qu'elles déclarent un nombre d'interventions non nul ce même jour. La correction appliquée est la suivante :

a. Pour les intervenants n'ayant fait qu'une visite, on cale le nombre de personnes visitées à 1 ;

b. Pour les intervenants ayant fait plus d'une visite dans la journée, on se sert de l'emploi du temps de cette journée pour essayer de voir s'il a effectué plusieurs interventions auprès de la même personne. Un seul cas est ambigu, l'intervenant ayant fait deux interventions, l'une le matin et l'autre l'après-midi. Il a cependant indiqué dans les deux cas qu'il ne s'agissait pas d'une personne fragile. On peut donc légitimement penser que ce type de personne a moins besoin qu'une personne fragile de recevoir plusieurs visites par jour. On a donc choisi de compter une personne différente pour chaque visite dans ce cas.

3.3.2. Emploi du temps

Pour l'emploi du temps détaillé de la journée de référence, il y a vraisemblablement eu quelques fautes de saisie :

- certains horaires sont en non-réponse, aucune correction n'a été effectuée dans ce cas ;
- surtout, le repérage des horaires aberrants a été effectué comme suit :

Si la condition suivante n'est pas vérifiée, alors on peut suspecter une erreur de saisie :
'Heure de départ du domicile' < 'Heure d'arrivée chez le premier client'
et 'Heure d'arrivée chez le 1er client' < 'Heure de départ de chez le 1er client'
et 'Heure de départ de chez le 1er client' < 'Heure d'arrivée chez le 2ème client'
…
et 'Heure de départ de chez le dernier client' < 'Heure de retour au domicile'

[12] On parle bien ici du jour de référence, et non plus de la semaine de référence.

Suite à ce repérage, plusieurs cas ont pu être identifiés :

a. les erreurs de saisie « évidentes » : par exemple 00H11 au lieu de 11H00, qui ont été corrigées dans la foulée ;
b. les horaires de nuit : l'intervenant n'a pas décrit sa journée de 00H00 à 23H59, mais de 20H00 à 9H30 par exemple. Aucune correction n'a été apportée pour eux, mais ce cas de figure méritera réflexion lors de l'étude des emplois du temps, les réponses ne correspondant pas en quelque sorte à la même question suivant que l'intervenant a bien répondu dans la tranche imposée ou non ;
c. les horaires juxtaposés : l'intervenant a par exemple déclaré 10H00-11H00 pour deux de ses interventions. Aucune correction n'a été effectuée pour l'instant, mais là encore une étude approfondie devra être menée lors de l'exploitation de cette partie ;
d. des cas plus complexes : par exemple lorsque l'intervenant a vraisemblablement confondu l'heure du retour au domicile avec l'heure du déjeuner, ou quand l'enquêteur a probablement saisi les heures de l'après midi en douzaine plutôt qu'en base 24 (06H00 à la place de 18H00). Les corrections ont été effectuées au cas par cas ici ; une étude approfondie sera menée plus tard lors de l'exploitation de cette partie.

3.3.3. L'intervention de référence

Trois variables de l'intervention de référence ont été corrigées, notamment dans le but d'effectuer le calage des poids de celle-ci :

- 9 aides à domicile ne connaissent pas l'âge de la personne visitée lors de l'intervention. Celui-ci a été redressé par hot-deck métrique, sur le fait que cette personne dispose ou non de l'APA et sur l'ancienneté des interventions auprès de cette personne ;
- 2 aides à domicile ne savent pas si la personne est handicapée. Elles déclarent cependant qu'elle est malade ou momentanément invalide. On suppose que la personne n'est pas handicapée ;
- 2 aides à domicile déclarent le symétrique : une personne dont elles ne savent pas si elle est malade ou temporairement invalide. Elles déclarent en revanche dans les deux cas qu'elle est handicapée. On suppose alors que la personne n'est pas malade.

3.3.4. Mode d'emploi et employeurs associés

En ce qui concerne le mode d'emploi, deux cas de figure se présentent :

- 3 intervenants indiquent ne pas savoir s'ils sont salariés d'un organisme de services à la personne ou de particuliers employeurs. Le redressement s'est effectué en analysant les réponses de la suite du questionnaire : celui qui indique que sa hiérarchie lui donne des consignes est normalement salarié d'un organisme, et ceux qui ont répondu explicitement qu'ils n'étaient pas salariés d'un organisme le sont forcément par au moins un particulier employeur ;

- d'autre part, les intervenants semblent confondre les statuts mandataires et prestataires. En effet, des problèmes de cohérence ont été repérés entre les réponses des organismes et celles des intervenants. Une variable supplémentaire a été introduite (MODE), construite comme suit :

a. Si l'organisme a indiqué que l'intervenant ne travaillait qu'en mode prestataire, et que celui-ci a déclaré de même, on le classe en prestataire exclusif ;
b. Si l'intervenant déclare ne dépendre d'aucun organisme, on le classe en emploi direct exclusif ;
c. Les autres cas correspondent à de l'emploi mandataire ou mixte (c'est-à-dire sous plusieurs modes).

Les nombres d'employeurs associés à chacun des modes d'emploi posent également problème dans certains cas :

- 12 intervenants déclarent plus de structures mandataires que de particuliers employeurs. Dans ces cas, on a calé le nombre d'organismes mandataires sur le nombre de ces particuliers pour ne pas trop distordre la déclaration initiale ;
- en ce qui concerne les valeurs manquantes, une imputation par hot-deck métrique a été effectuée, sur les critères de nombre de personnes visitées la journée de référence et de nombre de particuliers employeurs (lorsque la non réponse concerne le nombre d'organismes prestataires) ou d'organismes prestataires (lorsque la non réponse concerne les nombres de particuliers employeurs et de structures mandataires).

3.3.5. Les salaires et revenus du foyer

Dans l'enquête, les aides à domicile ont dû déclarer les montants des deux derniers mois de salaire. Même si certains salaires semblent particulièrement élevés (plus de 3 000 euros), une analyse fine montre que les intervenants concernés ont la plupart du temps une forte activité (approximée sur la semaine de référence) ou alors les salaires des deux derniers mois sont proches, ce qui ne plaide pas en faveur d'une erreur de frappe.

Au final, un seul salaire a été corrigé (9 350 euros tout de même !) par hot-deck métrique sur la déclaration de variations de revenus importantes d'un mois sur l'autre, d'activité la semaine de référence et sur le 2ème salaire déclaré.

Certains revenus du foyer de l'intervenant sont également particulièrement élevés. Ceux dont la valeur est très éloignée du salaire de l'aide à domicile (au moins 6 fois plus) et supérieurs à 4 500 euros (correspondant au 99ème centile de la distribution) sont considérés comme suspects. Les cas sont cependant complexes à analyser. En effet, les hauts revenus peuvent trouver plusieurs explications : conjoint particulièrement bien rémunéré, vie avec la fratrie, avec les parents, etc. Quatre cas semblent cependant problématiques et ont été redressés par hot-deck métrique sur le nombre de personnes du foyer non à charge, la moyenne des deux derniers salaires de l'aide à domicile, et l'aide sociale éventuelle perçue par le foyer.

3.3.6. Les congés

Des intervenants déclarent avoir pris plus de congés que ce à quoi ils indiquent avoir droit, alors même qu'ils n'ont pas pris de congés sans solde. Il s'agit à l'évidence d'erreurs de saisie : l'intervenant doit s'exprimer soit en jour, soit en semaines, et ensuite indiquer le nombre de jours/semaines. Certains indiquent qu'ils ont droit à 15 jours de congés mais qu'ils ont pris 15 semaines. On rétablit en corrigeant les semaines en jours.

D'autres valeurs semblent aberrantes pour les congés : par exemple 32 semaines de congés payés. On corrige les semaines en jours au regard de ce que l'intervenant a pris comme congés réels (et vice-versa) : par exemple sur les 32 semaines, l'intervenant sus-cité aurait pris 24 jours. On cale alors le droit à congés à 32 jours.

3.3.7. Les autres variables

- Pour l'âge de fin des études initiales, on a considéré que les âges inférieurs à 10 ans et supérieurs à 29 ans étaient douteux. Une imputation par hot-deck a donc été effectuée, en se servant de la tranche d'âge de l'intervenant et de son niveau d'études maximal atteint comme variables de classes. Un exemple d'imputation par hot-deck est présenté en annexe 3 ;

- 6 aides à domicile ont déclaré plus d'arrêts de travail pour accident du travail ou maladie professionnelle que le nombre d'arrêts totaux. Les nombres d'arrêts totaux ont été recalés sur le total des arrêts pour accident ou maladie ;

- 6 aides à domicile ont indiqué qu'elles étaient arrivées en France avant leur naissance. Dans ces cas, on recale la date d'arrivée sur celle de l'année de naissance.

4. Redressement des poids des organismes et des interviews téléphoniques

Comme dans toutes les enquêtes par sondage, on n'échappe pas à la non-réponse des différents protagonistes, qu'il s'agisse d'organismes ou des intervenants eux-mêmes. Cette non-réponse peut résulter de plusieurs facteurs :

- des erreurs d'adresses dans les bases de sondage ;
- des impossibilités à joindre les personnes lors de l'enquête (congés, mauvaises plages horaires…) ;
- des refus de réponse à l'enquête, refus qui peuvent s'exprimer au cours du recueil de données (manque de temps, questionnaire trop long…).

De ce fait, les poids de sondage ne reflètent plus la réalité. Au mieux, si la non-réponse n'est pas sélective, c'est-à-dire qu'elle ne dépend pas des caractéristiques des organismes ou individus interrogés, seuls les totaux ne sont plus bons. Mais dans la majorité des cas, les indicateurs sont biaisés si l'on ne redresse pas les poids initiaux.

Pour l'enquête IAD, les poids ont tous été redressés selon la méthode dite du calage sur marges, dont on trouvera un exemple d'utilisation en annexe 4 (encadré 2).

Tout redressement de la non-réponse se fait suite à une analyse de celle-ci. Dans l'enquête IAD, la probabilité de non réponse d'une unité enquêtée a été estimée à l'aide de modèles logistiques, sur les variables disponibles dans les bases de sondage (pour les organismes et les intervenants issus de la base de l'IRCEM), ou du degré d'interrogation précédant lui-même redressé (par exemple pour les intervenants sélectionnés dans les organismes, et pour lesquels la seule information auxiliaire disponible était dans les réponses des gestionnaires d'organismes).

Encadré 2 - Principes du calage sur marges

On part pour exemple de l'estimateur standard d'un total d'une variable *Y*, donné par : $\hat{Y}_w = \sum_{k \in S} w_k y_k$ (dans le cas de non réponse sélective, cet estimateur est baisé).

On dispose par ailleurs de variables auxiliaires $X_j \, (j = 1,..., J)$, dont on connaît les totaux sur l'ensemble de la population : $X_j = \sum_{k \in U} x_{jk}$.

Pour tenir compte de cette information, on cherche à transformer les poids w_k de façon à ce que les totaux des variables X_j restent exacts sur l'échantillon, et plus seulement sur la population d'origine. En d'autres termes, l'estimateur du total de *Y* devient :

$$\hat{Y}_p = \sum_{k \in S} p_k y_k \ ,$$

avec p_k les nouveaux poids respectant les équations de calage :

$$\sum_{k \in S} p_k x_{jk} = X_j , \quad \forall j = 1,..., J$$

Tout l'art d'un bon calage est de trouver les nouveaux poids p_k satisfaisant les équations de calage et les plus proches possibles des poids initiaux w_k . On introduit alors une fonction G mesurant la distance entre les poids p_k et w_k . Réaliser le calage sur marges consiste à résoudre le programme suivant :

$$\begin{cases} \underset{p_k}{Min} \sum_{k \in S} w_k G(p_k / w_k) \\ sc \quad \sum_{k \in S} p_k x_{jk} = X_j \quad \forall j \end{cases}$$

Plusieurs formes sont utilisables pour la fonction G, correspondant chacune à une méthode. Dans la macro SAS CALMAR2, développée par l'INSEE, cinq méthodes sont actuellement disponibles pour réaliser le calage : linéaire, exponentielle (également appelée raking ratio), logistique, linéaire tronquée, sinus hyperbolique.

Asymptotiquement toutes ces méthodes sont équivalentes, l'estimateur final calé correspond à l'estimateur par régression. Dans les faits, on se base sur quelques critères pour déterminer celle qui amène aux « meilleurs » poids : la plus faible dispersion des poids, la plus faible étendue, l'allure générale de la distribution…

La méthode du calage sur marges est initialement prévue pour recalibrer l'échantillon final sur certaines variables représentant la population entière (on cale sur des marges, tirées de la base de sondage), après correction de la non réponse totale. Néanmoins, sous la condition que les variables de calages incluent l'ensemble des variables explicatives de la non réponse, on peut effectuer à la fois la correction de la non réponse et le calibrage en une seule étape de calage.

4.1. Les intervenants des organismes (base ANSP)

Le redressement de la non réponse pour les intervenants des organismes se fait en deux étapes : d'abord le redressement des organismes eux-mêmes, puis celui des intervenants sélectionnés dans ces mêmes organismes.

4.1.1. Redressement des poids des organismes

263 organismes agréés de services à la personne ont répondu à l'enquête IAD sur un total de 870, soit un taux de réponse global de 30,2 %.

L'analyse préalable montre que les variables suivantes influent sur la probabilité de réponse des organismes :

- la catégorie juridique ;
- les effectifs salariés (en tranches, dont une spécifique aux organismes dont on ne connaissait pas *a priori* l'effectif) ;
- si l'organisme délivre des services de type ménage ;
- si l'organisme délivre des services de type entretien du linge ;
- si l'organisme délivre des services de type assistance aux personnes âgées ;
- si l'organisme délivre des services de type assistance aux personnes handicapées ;
- si l'organisme délivre des services de type assistance administrative.

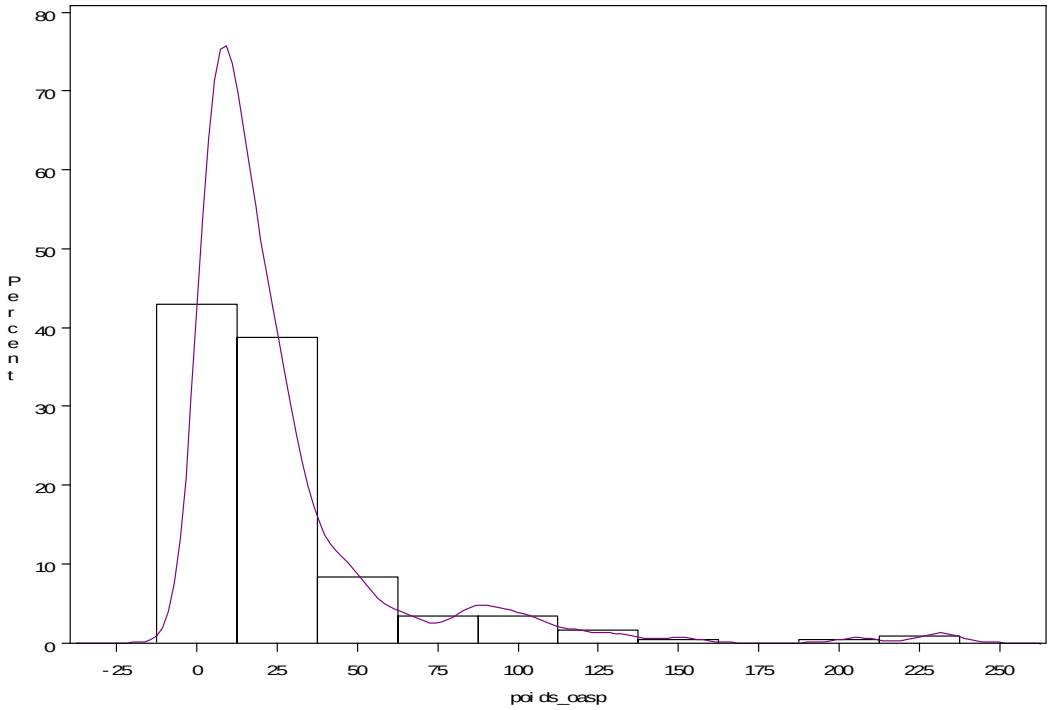
Les totaux de toutes ces variables, calculés à partir de la base de sondage de l'ANSP, ont été introduites dans le calage sur marges.

Comme vu *supra* (partie 2), les poids initiaux des répondants sont calculés comme suit, produit de la sélection des départements et des organismes :

$$w_{2,dhj} = \frac{1}{\pi_{2,dhj} . \pi_d}$$

h se référant à la variable de stratification effectifs connus/effectifs inconnus.

La fonction linéaire tronquée a été choisie pour le redressement par calage. La distribution finale des poids des organismes prend la forme suivante :



Statistiques sur les poids redressés des organismes selon la méthode de calage

	Linéaire	Exponentielle	Logistique	*Linéaire tronquée*	Sinus hyperbolique
Moyenne	26,3	26,3	26,3	***26,3***	26,3
Médiane	16,4	16,3	13,8	***15,1***	17,3
Écart-type	33,3	32,9	34,8	***34,0***	34,2
Rapport Q3/Q1	3,9	3,8	3,8	***4,3***	6,3
Rapport C95/C5	32,7	28,6	31,1	***35,3***	77,7
Rapport Max/Min	-297,6	144,2	128,8	***122,3***	549,8

Les écarts de poids peuvent paraître trop élevés. Cependant, l'échantillon ayant été constitué pour partie par tirage proportionnel à la taille, il est normal de se retrouver au final avec de tels écarts. Les écarts de poids se sont même réduits, puisque sur l'échantillon initial, comprenant les non répondants, le rapport Max/Min était de 301,9. En tout état de cause, les organismes ne sont pas étudiés lors des exploitations de l'enquête, et en tout cas pas sur des domaines[13]. Les écarts de pondération ne seront vraiment importants à analyser que lorsqu'on s'intéressera aux poids des intervenants eux-mêmes.

[13] Domaine au sens statistique du terme : variable de croisement qui n'est pas une variable de stratification.

4.1.2. Redressement des poids des intervenants des organismes

Les poids des intervenants tirés des organismes précédemment sélectionnés sont redressés de la même façon que les organismes, en se basant sur la pondération initiale :

$$w_{2,dhjk} = \frac{1}{\pi_{2,dhj}.\pi_d}.\frac{T_j}{Min(5,T_j)}$$

Où T_j est le nombre d'intervenants de l'organisme *j*.

Une analyse préalable des poids initiaux montre que ceux-ci sont beaucoup trop dispersés (voir tableau ci-dessous). Pour rappel, lors d'un tirage à plusieurs degrés où les degrés amont sont tirés à probabilité proportionnelle à la taille, les poids finals des individus du dernier degré sont sensiblement identiques, ou au moins très faiblement dispersés. Une trop forte dispersion introduit une variance trop importante dans les estimateurs sur domaines, rendant délicates les analyses futures, et ceci même si les estimateurs sont théoriquement sans biais.

Statistiques sur les poids initiaux des intervenants des organismes

Moyenne	357,8
Médiane	121,3
Écart-type	1 180,0
Rapport Q3/Q1	4,1
Rapport C95/C5	76,7
Rapport Max/Min	2 699,0

La solution appliquée ici est de tronquer les poids, avant même le calage (encadré 3). Les variables d'intérêt dans l'analyse de la troncature des poids sont les suivantes :

- l'âge de l'intervenant ;
- son ancienneté dans l'organisme ;
- son nombre d'heures travaillées au cours du dernier mois.

Pour ces trois variables, l'erreur quadratique moyenne a tendance à décroître ou à rester stable au fur et à mesure que l'on tronque les poids. Elle augmente même aux faibles troncatures pour les variables d'ancienneté et d'heures travaillées (poids tronqués aux 99ème et 98ème centiles des poids initiaux). Le biais relatif des trois variables, s'il est assez élevé (jusqu'à 9% pour les heures travaillées), a toutefois tendance à diminuer à mesure de la troncature, ou au moins à rester constant.

Au final, on se sert dans le calage des poids tronqués au 88ème centile des poids initiaux.

Encadré 3 - troncature des poids

Les poids tronqués doivent respecter certaines conditions :
- le taille de la population, estimée à partir de l'échantillon pondéré, doit rester la même ;
- les estimateurs des variables d'intérêt ne doivent pas « trop » s'éloigner des estimateurs sans biais, calculés à l'aide des poids initiaux.

La deuxième condition suppose de choisir une batterie d'estimateurs, pour lesquels on calculera le biais absolu, le biais relatif, la variance et l'erreur quadratique moyenne.

Suivant Potter (1990), en fixant un poids maximal w_0 et en corrigeant quelques fautes de frappe dans le texte original, les poids tronqués sont calculés comme suit :

$$w_{kt} = \tau_k w_0 + (1-\tau_k) w_k \frac{\sum_{k\in S}(w_k - \tau_k w_0)}{\sum_{k\in S}(1-\tau_k w_k)} \quad \textbf{(E)}$$

Avec $\tau_k = 1$ si $w_k \geq w_0$, 0 sinon.

w_k représente les poids initiaux, w_0 le poids plafond pour la troncature, et w_{kt} les poids tronqués.

$A = \frac{\sum_{k\in S}(w_k - \tau_k w_0)}{\sum_{k\in S}(1-\tau_k w_k)}$ est le coefficient multiplicateur des poids inférieurs à w_0. Ce coefficient est constant.

Le lecteur courageux pourra par ailleurs vérifier que la première condition est satisfaite[14] :

$$\sum_{k\in S} w_{kt} = \sum_{k\in S} w_k$$

Ce qu'oublie de mentionner Potter, c'est qu'après cette opération, certains poids, après avoir été multipliés par *A*, peuvent être supérieurs à w_0. La solution consiste alors à calculer de nouveaux poids tronqués, en injectant dans (E) les poids tronqués à la place des poids initiaux (le coefficient *A* diminue). Cette opération doit se répéter jusqu'à ce que le processus converge et qu'on vérifie $Max(w_{kt}) = w_0$. On trouvera la macro SAS permettant de calculer les poids tronqués en annexe 5.

Le calcul de la variance des estimateurs (ici des moyennes ou des proportions) des variables d'intérêt est approximé par la *proc surveymeans* de SAS. Le calcul de l'erreur quadratique moyenne est donné par (Potter, 1993[15]) :

$$EQM = (\hat{\bar{Y}}_t - \hat{\bar{Y}})^2 + V(\hat{\bar{Y}}_t)$$

On trouvera la macro de calcul de l'erreur quadratique moyenne en annexe 6. Les poids tronqués *optima* sont ceux qui permettent de trouver l'EQM minimale. Dans les faits, on s'intéresse également beaucoup au biais relatif, qui ne doit pas être trop élevé.

[14] L'étourdi Potter a quant à lui posé $A = \frac{\sum_{k\in S}(1-\tau_k w_0 \pi_k)}{\sum_{k\in S}(1-\tau_k) w_k}$ dans son texte, avec $w_k = \frac{1}{\pi_k}$, ce qui ne vérifie pas la première condition.

[15] De nombreux estimateurs sont disponibles pour le calcul de l'erreur quadratique moyenne. Voir par exemple Alavi et Beaumont (2004), Potter (1988), Potter (1990) et Potter (1993).

Statistiques sur les poids initiaux tronqués au 88ème centile des intervenants des organismes

Moyenne	357,8
Médiane	469,2
Écart-type	142,1
Rapport Q3/Q1	1,9
Rapport C95/C5	7,0
Rapport Max/Min	26,8

991 intervenants sélectionnés dans les organismes ont répondu à l'interview téléphonique, soit un taux de réponse de 77 %.

La non réponse des intervenants sélectionnés dans les organismes ne dépend que du mode d'exercice (prestataire, mandataire ou sous les deux statuts). Cette variable est donc introduite dans le calage. D'autres variables de calage sont également introduites, correspondant aux variables d'intérêt dans l'analyse de la troncature des poids mises en tranches.

Les marges des variables sont calculées *via* les réponses des organismes sur leurs intervenants sélectionnés, en utilisant les poids initiaux non tronqués (estimateurs sans biais). Ceci a notamment l'avantage de corriger le biais sur les variables d'intérêt introduit par la troncature.

Au final, on cale les poids tronqués par la méthode logistique. Les poids calés sont beaucoup plus dispersés que les poids tronqués, mais aussi beaucoup moins que les poids initiaux. On aurait pu améliorer ces écarts en tronquant un peu plus les poids initiaux, mais les mesures de biais et d'erreurs quadratiques moyennes deviennent alors trop importantes. Il s'agit donc d'un arbitrage entre qualité des estimateurs et variance des poids.

Statistiques sur les poids tronqués et calés des intervenants des organismes (téléphone)

Moyenne	464,6
Médiane	389,3
Écart-type	395,2
Rapport Q3/Q1	1,9
Rapport C95/C5	25,4
Rapport Max/Min	101,8

4.2. Les intervenants des particuliers employeurs (base IRCEM)

4 923 intervenants tirés dans la base de l'IRCEM ont répondu totalement à l'interview téléphonique, soit un taux de réponse global de 28,8%[16].

Les poids initiaux introduits dans le calage sur marges sont donnés directement par :

$$w_{1,dhk} = \frac{1}{\pi_{1,dhk} . \pi_d}$$

h représentant les différentes strates et d les départements.

Ces poids, faiblement dispersés, ne nécessitent pas d'être tronqués préalablement au calage. Les variables expliquant la non réponse sont les suivantes :

- l'ancienneté d'adhésion à l'IRCEM en tranches ;
- l'âge en tranches ;
- le sexe ;
- le nombre d'employeurs en tranches ;
- si l'intervenant a des employeurs utilisant la déclaration nominative simplifiée (DNS) ;
- si l'intervenant a des employeurs bénéficiant de la prestation d'accueil du jeune enfant (PAJE).

À ces variables, utilisées en tranches, on rajoute également le nombre total d'heures travaillées et la masse salariale dans les variables de calage. Les poids calés ne présentent aucun problème de dispersion (tableau). La méthode logistique a été retenue pour le calage.

Statistiques sur les poids calés des intervenants de particuliers employeurs (téléphone)

Moyenne	128,7
Médiane	117,5
Écart-type	46,0
Rapport Q3/Q1	1,6
Rapport C95/C5	3,2
Rapport Max/Min	7,4

[16] La non réponse est aussi due en grande partie au manque de coordonnées téléphoniques pour joindre les intervenants sélectionnés.

5. Redressement des poids des interviews en face à face

Les interviews en face à face se font uniquement auprès des intervenants qui travaillent auprès de personnes fragiles, soit après le passage du questionnaire filtre téléphonique. Sur les 5 914 aides à domicile interrogées téléphoniquement, 2 676 sont déclarées hors champ, soit 45,2 %. L'estimateur de la population des intervenants au domicile des personnes fragilisées correspond à la somme des poids des individus du champ dans l'enquête téléphonique (un peu moins de 710 000). Cet estimateur ne tient pas encore compte des doubles comptes, mais ce sont tout de même ces poids qui serviront à l'initialisation du redressement de la non-réponse totale. En effet, seuls ces poids permettent le calcul de marges sur les variables auxiliaires, que celles-ci soient issues des interviews téléphoniques ou des bases de sondage.

Pour assurer une cohérence entre les interviews téléphoniques et les interviews en face à face, on procède à deux calages :

- l'un redressant de la non réponse, en se basant sur des variables du questionnaire téléphonique ;
- l'autre calant les poids redressés de la non réponse sur les mêmes variables que celles ayant servi au calage des poids des interviews téléphoniques.

5.1. Les intervenants sélectionnés à partir des organismes

745 intervenants sélectionnés à partir des organismes et déclarés dans le champ ont répondu à l'interview en face à face, soit un taux de réponse de 81,9%.

L'analyse de la non réponse fait ressortir les variables explicatives suivantes :

- les tâches habituellement effectuées dans le cadre du travail au domicile de personnes ;
- les types de personnes chez qui l'intervenant travaille ;
- si l'activité d'aide au domicile de personnes fragilisées constitue l'activité principale de l'intervenant.

Ces variables sont utilisées dans le calage sur marges. La méthode du sinus hyperbolique est retenue à ce niveau.

Pour le calage complémentaire sur poids redressés, on utilise les variables précédemment utilisées lors du redressement de l'interview téléphonique (réponses des gestionnaires des organismes) :

- le mode d'exercice (prestataire, mandataire ou sous les deux statuts) ;
- l'âge de l'intervenant en tranches ;
- son ancienneté dans l'organisme en tranches ;
- son nombre d'heures travaillées au cours du dernier mois en tranches.

Les 5 méthodes de calage présentent des résultats très proches. On choisit la méthode du sinus hyperbolique, qui présente la dispersion la plus faible. Cette dernière reste toutefois forte (rapport Max/Min de 123, tableau).

Statistiques sur les poids redressés et calés des intervenants d'organismes (face à face)

Moyenne	556,9
Médiane	458,9
Écart-type	495,2
Rapport Q3/Q1	2,1
Rapport C95/C5	24,5
Rapport Max/Min	122,6

5.2. Les intervenants sélectionnés à partir de la base IRCEM

1 842 intervenants sélectionnés parmi les particuliers employeurs et déclarés dans le champ ont répondu à l'interview en face à face, soit un taux de réponse de 79,1 %.

L'analyse de la non réponse fait ressortir des variables explicatives différentes (sauf une) de celles des intervenants issus d'organismes :

- l'âge de l'intervenant en tranches ;
- le métier qu'il déclare ;
- les tâches habituellement effectuées dans le cadre du travail au domicile de personnes.

Ces variables sont utilisées dans le calage sur marges. Les méthodes de calage donnent des résultats sensiblement équivalents. La méthode logistique est retenue à ce niveau.

Pour le calage complémentaire sur poids redressés, on utilise les variables précédemment utilisées lors du redressement de l'interview téléphonique (issues de la base de sondage de l'IRCEM) :

- l'ancienneté d'adhésion à l'IRCEM en tranches ;
- l'âge en tranches ;
- le sexe ;
- le nombre d'employeurs en tranches ;
- si l'intervenant a des employeurs utilisant la déclaration nominative simplifiée (DNS) ;
- si l'intervenant a des employeurs bénéficiant de la prestation d'accueil du jeune enfant (PAJE) ;
- le nombre d'heures travaillées ;
- la masse salariale.

Les 5 méthodes de calage présentent des résultats identiques. On choisit la méthode linéaire, qui a l'avantage d'être la plus simple et dont les estimateurs associés correspondent parfaitement aux estimateurs par régression.

Statistiques sur les poids redressés et calés des intervenants de particuliers employeurs (face à face)

Moyenne	160,0
Médiane	150,5
Écart-type	57,3
Rapport Q3/Q1	1,6
Rapport C95/C5	3,2
Rapport Max/Min	10,0

5.3. Le partage des poids

Comme on l'a vu *supra* (partie 2), les intervenants peuvent être sélectionnés plusieurs fois. Une dernière correction des pondérations est donc nécessaire, pour ne pas sur-représenter les potentiels doubles comptes. La repondération nécessite de connaître le nombre de liens, pour chaque intervenant sélectionné dans l'une ou l'autre base de sondage, qui existent entre les deux bases. Pour ce, on a posé plusieurs questions aux intervenants interrogés[17] :

- s'ils sont indépendants pour une partie de leur activité ;
- s'ils dépendent d'organismes mandataires pour une partie de leur activité, et si oui de combien ;
- s'ils sont salariés d'organismes prestataires pour une partie de leur activité, et si oui de combien.

Tous ces liens sont fixés à date donnée, c'est-à-dire qu'un intervenant sélectionné dans la base IRCEM par exemple, devra déclarer, à une date donnée, les liens qu'il peut entretenir avec d'autres organismes. L'idéal serait qu'il réponde à ces questions à la date de construction de la base de sondage. À défaut de pouvoir le faire, les intervenants ont été interrogés sur leur situation actuelle, c'est-à-dire à la date d'enquête.

Pour la pondération finale, on applique la méthode du partage des poids : il suffit de diviser la pondération obtenue par intervenant par le nombre de liens relevés lors de l'enquête pour cet intervenant précis, ce qui élimine les doubles comptes[18].

Dans l'enquête en face à face, les questions ont été posées comme suit :

[17] Comme on l'a vu, les réponses des intervenants à ces questions sont assez floues, mais il s'agit des seules informations disponibles pour compter les liens entre les deux bases.

[18] Au cas où un intervenant aurait été sélectionné plusieurs fois, on ne l'aurait interrogé qu'une seule fois, mais aurions gardé trace de ce double tirage pour les pondérations finales. La pondération de celui-ci aurait alors été égale à la somme des deux pondérations issues des deux tirages, divisée par les nombres de liens trouvés lors de l'interview.

C1 Actuellement, êtes-vous salarié(e) d'au moins une entreprise, une association, ou une structure publique d'aide à domicile

1 Oui
2 Non

POSER C1BIS SI C1=1

C1bis Dans combien de ces services d'aide à domicile êtes-vous salarié(e) ?
|___|___|

C2 Toujours actuellement, un particulier au moins est-il votre employeur en tant qu'aide à domicile ?

1 Oui
2 Non

POSER C2BIS SI C2=1

C2bis Combien de ces particuliers sont vos employeurs en tant qu'aide à domicile ?
|___|___|

POSER C3 SI C2=1

C3 Avez-vous été mis en contact avec au moins l'un de ces particuliers employeurs par l'intermédiaire d'une association ou d'un CCAS ?

1 Oui
2 Non

POSER C4 SI C3= 1

C4 Combien d'associations ou de CCAS vous ont servi d'intermédiaire pour entrer en relation avec les personnes qui vous emploient directement actuellement ?
|___|___|

Le nombre de liens entre les bases d'enquête des intervenants se voit alors comme suit :

- les intervenants déclarant X employeurs directs (sans passer par la voie mandataire) ne peuvent être issus que de la base de l'IRCEM ;
- les intervenants déclarant Y organismes mandataires peuvent être issus des Y organismes associés dans la base de l'ANSP et de la base de l'IRCEM. Pour cette dernière, ils ne peuvent y être qu'une seule fois (pas de doublons). Par ailleurs, si l'intervenant fait aussi de l'emploi direct, on ne le compte qu'une fois dans la base IRCEM ;
- les intervenants déclarant Z organismes prestataires ne peuvent être issus que des Z organismes associés dans la base de l'ANSP.

En termes mathématiques, cela revient à écrire le nombre de liens comme suit :

$$L_k = \underbrace{1_{\{k\in\Omega_d\}} + 1_{\{k\notin\Omega_d\}}.1_{\{k\in\Omega_m\}}}_{\text{Liens IRCEM}} + \underbrace{M_k + P_k}_{\text{Liens ANSP}}$$

Ω_d et Ω_m étant les ensembles des intervenants respectivement en emploi direct et dépendant d'organismes mandataires.
M_k et P_k sont les nombres d'organismes respectivement mandataires et prestataires dont l'intervenant k dépend.

Par exemple :

- si un intervenant a été sélectionné uniquement dans la base IRCEM, et qu'il est complètement indépendant (i.e. ne peut être tiré dans la base ANSP), son poids final sera $w_k = w_{1,dhk}$;
- si un intervenant a été sélectionné dans la base ANSP, et qu'il fait partie d'un organisme mandataire, il aurait pu être sélectionné de fait dans la base IRCEM. Son poids final sera $w_k = \frac{w_{2,dh'jk}}{2}$.

Avec partage des poids, c'est-à-dire en tenant compte des doubles comptes, on estime le nombre total d'aides à domicile de personnes fragilisées à un peu plus de 515 000.

Les poids sont cependant à nouveau trop dispersés (tableau). On recourt donc à nouveau à la troncature de ces derniers, en prenant comme variables d'intérêt pour l'analyse des conséquences de cette opération :

- l'ancienneté dans le métier ;
- les nombres de personnes par type auprès desquelles l'intervenant à domicile travaille ;
- les nombres d'heures travaillées par type de personne aidée ;
- le nombre d'employeurs directs dont l'intervenant est salarié ;
- le nombre d'organismes mandataires dont l'intervenant dépend ;
- le nombre d'organismes prestataires dont l'intervenant est salarié.

Statistiques sur les poids partagés des intervenants (face à face)

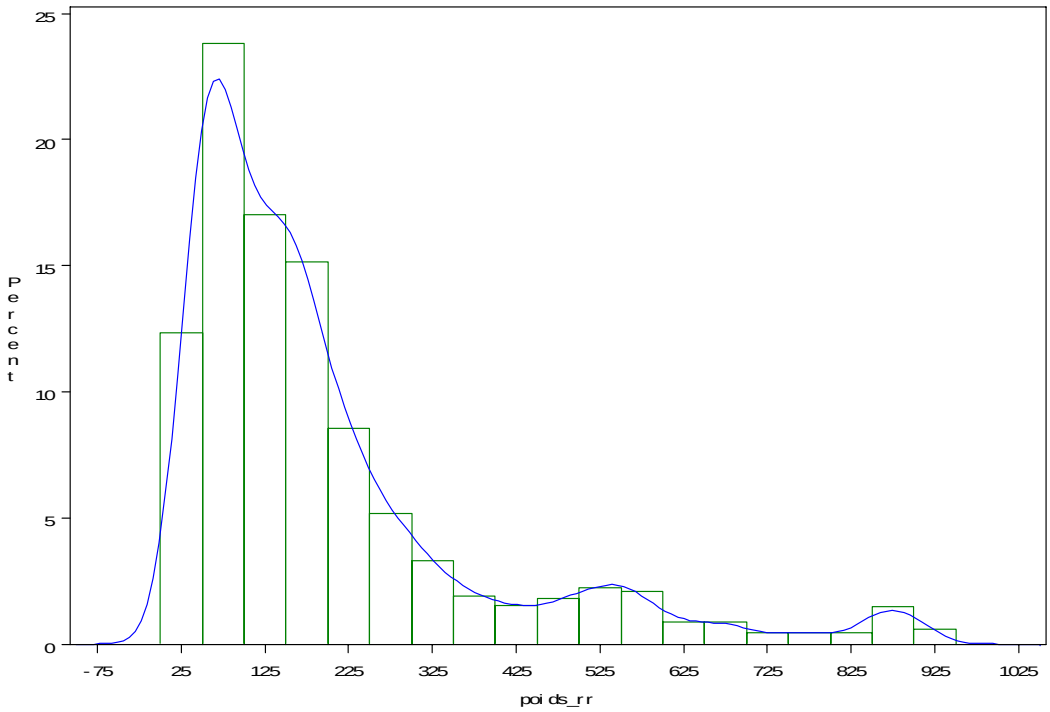
Moyenne	199,2
Médiane	135,8
Écart-type	221,9
Rapport Q3/Q1	3,2
Rapport C95/C5	16,4
Rapport Max/Min	275,6

Les variations – selon le seuil de troncature des poids – de l'erreur quadratique moyenne sont différentes suivant la variable considérée. En revanche, le biais relatif dû à la troncature augmente assez rapidement sur certaines variables, incitant à tronquer *a minima*. En tout état de cause, tronquer les poids au 98ème centile semble suffisant (rapport Max/Min des poids tronqués=82,4). Les variables d'intérêt utilisées dans l'analyse de la troncature sont ré-utilisées pour caler les poids tronqués (en utilisant des tranches pour l'ancienneté), et ainsi corriger le biais introduit. Les marges sont calculées en utilisant la pondération des poids partagés avant troncature sur les variables. Au final, c'est la méthode exponentielle qui est retenue (tableau).

Statistiques sur les poids partagés, tronqués et calés des intervenants (face à face)

Moyenne	199,2
Médiane	141,4
Écart-type	186,3
Rapport Q3/Q1	3,3
Rapport C95/C5	16,2
Rapport Max/Min	69,6

Distribution finale des poids (face à face)



6. Redressement de la journée et de l'intervention types

Le questionnaire face à face prévoit des questions sur une journée et une intervention types. Toutes deux sont sélectionnées aléatoirement par l'enquêteur :

- la journée type est tirée parmi les journées travaillées par l'intervenant la dernière semaine d'activité ;
- l'intervention type est tirée parmi l'ensemble des interventions auprès de personnes fragiles de la journée type.

6.1. La journée type

2 477 aides à domicile sur 2 587 ont travaillé le mois précédent l'enquête, et ont donc renseigné sur une semaine et une journée types. Les pondérations initiales du jour de référence sont données par :

$$w_{kj} = J_k . w_k$$

où J_k est les nombre de jours travaillés par l'intervenant k pendant la semaine de référence.

Les pondérations ainsi calculées ne donnent pas une distribution des jours tirés identique à celle déclarée par les intervenants lorsqu'on leur demande de décrire leur semaine type (tableau). Le week-end est ainsi légèrement sous représenté dans le tirage des journées type. Une première étape est donc de recadrer cette distribution, par simple post-stratification (règle de trois).

Distribution des jours travaillés selon les déclarations des intervenants sur leur semaine type et selon le tirage aléatoire du jour de référence

	Jours déclarés	Jours tirés
Lundi	17,5	16,9
Mardi	18,1	19,4
Mercredi	16,9	17,1
Jeudi	17,7	18,3
Vendredi	18,0	18,3
Samedi	7,7	6,3
Dimanche	4,2	3,7
Total	100,0	100,0

Suite à ce recadrage, les poids apparaissent à nouveau trop dispersés : rapports Max/Min=259,8 et C95/C5=19,2. Les poids ont donc été tronqués, au 93ème centile des poids initiaux puis recalés par la méthode linéaire tronquée sur les variables d'intérêt suivantes, afin de corriger du biais introduit :

- la distribution des jours travaillés la semaine de référence ;
- le nombre d'interventions total le jour tiré ;
- le nombre d'interventions auprès de personnes fragiles le jour tiré ;
- l'amplitude horaire de travail le jour tiré, c'est-à-dire le temps passé entre le départ et le retour au domicile.

Statistiques sur les poids tronqués et calés de la journée type

Moyenne	991,2
Médiane	699,6
Écart-type	817,7
Rapport Q3/Q1	3,5
Rapport C95/C5	15,4
Rapport Max/Min	99,0

6.2. L'intervention type

Les pondérations initiales de l'intervention de référence, tirées dans le jour de référence pour chaque intervenant, sont données par :

$$w_{kji} = I_{kj}.J_k.w_k$$

Avec I_{kj} le nombre total d'interventions auprès de personnes fragiles de l'intervenant k le jour j.

Suite à une erreur de programmation dans le questionnaire électronique, certaines interventions de référence n'ont pas été prises en compte dans l'enquête. Plus précisément, aucune intervention n'a été tirée dans les cas suivants :

- l'intervenant n'a effectué qu'une seule intervention dans la journée, et elle était auprès de personnes fragiles ;
- l'intervenant a effectué plusieurs interventions dans la journée, dont une seule auprès de personnes fragiles, celle-ci étant la dernière de la journée.

En revanche, les intervenants ayant effectué plusieurs interventions dans la journée, dont une seule auprès de personnes fragiles, celle-ci n'étant pas la dernière de la journée[19], ont bien pu répondre aux questions sur l'intervention de référence.

Ceci implique que l'on n'a pas d'information sur les interventions des aides à domicile à « faible variété d'activité » (une seule intervention dans la journée). Et pour celles qui font une seule intervention auprès de personnes fragiles parmi d'autres, on n'a qu'une information parcellaire. On prend le parti de compter ces absences de tirage comme de la non réponse, que l'on redresse par règle de trois, en recalant les poids des interventions des « répondants » – parmi ceux qui n'ont effectué qu'une intervention auprès de personnes fragiles dans la journée – sur le total des poids des interventions auprès d'une seule personne fragile des aides à domicile[20].

Suite à ce premier redressement, on tronque les poids des interventions, ceux-ci étant trop dispersés. On les recale ensuite sur les variables suivantes :

[19] Il faut suivre…
[20] Ce total étant bien disponible, car les pondérations théoriques des interventions ont été calculées, même en l'absence de tirage effectif.

- l’ancienneté auprès de la personne concernée par l’intervention ;
- la configuration du foyer où a lieu l’intervention (personne seule, couple, autre) ;
- l’âge de la personne aidée ;
- si la personne aidée est handicapée ;
- si la personne aidée est malade ou momentanément invalide.

Les poids ont été tronqués au 87ème centile des poids initiaux, puis calés par la méthode linéaire tronquée.

Statistiques sur les poids tronqués et calés de l’intervention type

Moyenne	4 351,3
Médiane	3 311,7
Écart-type	3 139,0
Rapport Q3/Q1	4,4
Rapport C95/C5	15,0
Rapport Max/Min	96,2

Bibliographie indicative

Ardilly P., 2006 : *Les techniques de sondage – 2ème édition*, Paris, Éditions Technip, 675p.

Alavi A., Beaumont J.-F., 2004 : « Estimation robuste par la régression généralisée », *Techniques d'enquêtes*, Statistique Canada, vol.30, n°2, pp. 217-231, décembre.

Le Guennec J., Sautory O., 2005 : « La macro CALMAR2 – Redressement d'un échantillon par calage sur marges », Insee-Cepe, avril.

Potter F., 1988 : « Survey of procedures to control extreme sampling weights », *Proceedings of the Section on Survey Research*, American Statistical Association, pp.453-458.

Potter F., 1990 : « A study of procedures to identify and trim extreme sampling weights », *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp. 225-230.

Potter F., 1993 : « The effect of weight trimming on nonlinear survey estimates », *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp. 758-763.

Annexe 1 - Exemple de tirage d'échantillon à probabilités proportionnelles à la taille – cas du tirage des départements

```
/* Calcul des probabilités d'inclusion */
/*-------------------------------------*/

data dept;set dept;
      Btot=btotag+btothand;
      Dtot=dtotag+dtothand;
run;

proc means data=dept sum;
      var btot;
      output out=out_dept sum=sumBtot;
run;

data dept;set dept;i=1;
data out_dept;set out_dept;i=1;drop _TYPE_ _FREQ_;run;

data dept;merge dept out_dept;by i;run;

* Calcul des probas ;

data dept;set dept;
      prob=30*btot/sumBtot;
run;

* Probas > 1 ? ;

proc univariate data=dept;
      var prob;
run;

* oui -> 3 depts => recalcul en enlevant ces 3 ;

data t1 dept2;set dept;
      if prob>1 then output t1;
      else output dept2;
run;

proc univariate data=dept2;
      var prob;
run;

proc means data=dept2 sum;
      var btot;
      output out=out_dept2 sum=sumBtot2;
run;

data out_dept2;set out_dept2;i=1;drop _TYPE_ _FREQ_;run;

data dept2;merge dept2 out_dept2;by i;run;


* Calcul des nouvelles probas ;
```

```
data dept2;set dept2;
      prob2=27*btot/sumbtot2;
run;

* proba > 1 ? ;

proc univariate data=dept2;
      var prob2;
run;

* => OK, nettoyage ;

data dept2;set dept2;
      drop i sumBtot prob;
run;

data dept2;set dept2;
      rename prob2=prob sumBtot2=sumBtot;
run;


/* Tirage de l'échantillon */
/*-------------------------*/

data t1;set t1;
      prob=1;
      samplingweight=1;
      expectedhits=prob;
      drop i sumBtot;
run;

* les 27 avec prob<1 ;

data dept2;set dept2;
      drop sumBtot;
run;

proc sort data=dept2;by btot dtot;run;


%macro tir_dept(rep=10);

      /* tirage des 27 */

      proc surveyselect data=dept2
                              stats
                              method=pps_sys
                              sampsize=27
                               seed=1234
                               rep=&rep
                              out=ech27;
            size prob;
            control btot dtot;
      run;

      /* nettoyage */

      data ech27;set ech27;drop numberhits;run;
```

```
      /* constitution */

      %do i=1 %to &rep;
            data ech30_&i;set ech27;
                  if replicate=&i then output;
            run;

            data t1;set t1;replicate=&i;run;

            data ech30_&i;set ech30_&i t1;run;

            * vérif ;

            proc means data=ech30_&i sum;
                  var Btot Dtot;
                  weight samplingweight;
            run;

            proc sort data=ech30_&i;by code_dept;run;

            * export ;

            data a.dept30_&i;set ech30_&i;run;

            proc export data= a.dept30_&i
                  outfile= "D:\...\Dept30_&i..xls"
      dbms=EXCEL2000 replace;
            run;
      %end;

%mend tir_dept;

%tir_dept(rep=10);


/* choix des échantillons */
/*-------------------------*/

proc sort data=dept;by code_dept;run;

%macro tri(nb);

      data final;set dept;run;

      %do i=1 %to &nb;
            proc sort data=ech30_&i;by code_dept;run;

            data ech30_&i;set ech30_&i;
                  ech_&i=1;
            run;

            data final;merge final ech30_&i;
                  by code_dept;
                  if ech_&i=. then ech_&i=0;
            run;
      %end;
```

```
      data final;set final;
            drop replicate expectedhits samplingweight;
      run;
%mend;

%tri(nb=10);

data a.final;set final;run;
```

Annexe 2 - Exemple de tirage d'un échantillon de réserve – cas des OASP

```
/* Principe :
   On part de l'échantillon total et on fait comme si on en tirait
   seulement 27 par départements (donc 810 au total, au lieu de 870) ;
   On fait ensuite le différentiel entre l'échantillon total et
   l'échantillon réduit pour trouver le deuxième échantillon de réserve ;

   On connait le nombre d'employés de certains établissements
=> stratification suivant le critère info connue/inconnue
- pour ceux dont on connait l'info -> tirage PPT
- pour ceux dont on ne connait pas l'info -> tirage SAS */
/*-------------------------------------------------------------------------*/

libname a"D:\...";

data echant_tot;set a.echant_total;run;
data oasp;set a.base_oasp;run;


/* Comparaison info connue/inconnue */
/*----------------------------------*/

data oasp;set oasp;
      connu=(efftot^=.);
run;

proc univariate data=oasp;
      var efftot;
      class jur2;
run;

* les jur2=2 sont plus petits => on essaiera d'en tenir compte dans
l'échantillonnage (dans l'instruction control)             ;


/* Échantillonnage */
/*-----------------*/

/* Calcul des allocations par département x strate */
/*------------------------------------------------*/

proc freq data=dept;
      table d_partement_d_activit_*connu /out=out;
run;

data out0 out1;set out;
      if connu=0 then output out0;
      else output out1;
run;

data out0;set out0;
      rename COUNT=Eff_inconnu connu=connu0;
      drop PERCENT;
run;
```

```
data out1;set out1;
      rename COUNT=Eff_connu connu=connu1;
      drop PERCENT;
run;

data alloc;merge out0 out1;
      by d_partement_d_activit_;
run;

data alloc;set alloc;
      total=eff_inconnu+eff_connu;
      pct_inconnu=eff_inconnu/total;
      Eff_ech=27;
      Eff_ech_inconnu=round(27*pct_inconnu);
      Eff_ech_connu=Eff_ech-Eff_ech_inconnu;
run;

data alloc_fin;set alloc;
      keep d_partement_d_activit_ total eff_inconnu eff_connu Eff_ech
            Eff_ech_inconnu Eff_ech_connu;
run;

proc sort data=dept;by d_partement_d_activit_;run;

data dept2;merge dept alloc_fin;by d_partement_d_activit_;run;

data dept2;set dept2;
      rename d_partement_d_activit_=code_dept;
run;


/* Calcul des probabilités d'inclusion */
/*-------------------------------------*/

/* Pour ceux dont on connait le nombre de salariés */
/*-------------------------------------------------*/

proc means data=dept2 n sum;
      class code_dept;
      var Efftot;
      output out=outprob1 sum=Somme1;
      where connu=1;
run;

data outprob1;set outprob1;
      if _TYPE_=0 then delete;
run;

data outprob1;set outprob1;
      drop _TYPE_ _FREQ_;
run;

data prob1;merge dept2 outprob1;by code_dept;run;

* Calcul des probas ;

data prob1;set prob1;
```

```
      probA=eff_ech_connu*efftot/somme1;
      if connu=1 then output;
run;

* Probas > 1 ? ;

proc univariate data=prob1;
      var probA;
run;

* oui -> on isole ;

data t1 prob2;set prob1;
      if probA>=1 then output t1;
      else output prob2;
run;

* calage des probas >= 1 ;

data t1;set t1;
      probA=1;
run;

* recalcul des probas pour les autres ;

proc means data=prob2 n sum;
      class code_dept;
      var Efftot;
      output out=outprob2 sum=Somme2;
run;
```

Et ainsi de suite…

```
* Nettoyage ;

data t1;set t1;
      rename probA=prob_deb;
      drop somme1;
run;

data t2;set t2;
      rename probB=prob_deb;
      drop somme1--eff_ech_connu2;
run;

data t3;set t3;
      rename probC=prob_deb;
      drop somme1--eff_ech_connu3;
run;

data prob4;set prob4;
      rename probD=prob_deb;
      drop somme1--nb_prob3;
run;
```

```
* Calcul des probas de tirage à partir de l'échantillon total déjà tiré ;

data echant_deb;set prob4 t1 t2 t3;run;

data echant_tot1;set echant_tot;
      if connu=1 then output;
      keep serial connu prob_fin;
run;

proc sort data=echant_deb;by serial;
proc sort data=echant_tot1;by serial;
run;

data prob5;merge echant_deb echant_tot1(in=yy);
      by Serial;
      if yy;
run;

* Calcul des nouvelles probas ;

data prob5;set prob5;
      prob_nouv=prob_deb/prob_fin;
run;

* Vérif ;

proc univariate data=prob5;
      var prob_nouv;
run;

* On ne garde que les probas<1
  Attention, dans certains départements les probas avant et après sont
  identiques
-> dû aux arrondis et à la faiblesse de l'échantillon
=> il faudra recalculer les nombres d'organismes à tirer par département ;

data tirage1 tirage2;set prob5;
      if prob_nouv<1 then output tirage1;
      else output tirage2;
run;

proc univariate data=tirage1;
      var prob_nouv;
run;

* OK ;


/* Tirage des organismes */
/*-----------------------*/

* Créations de tranches de tailles pour les contrôles ;

proc univariate data=tirage1;
      var efftot;
run;

data tirage1;set tirage1;
```

```
      if efftot<=20 then efftr=1;
      if 20<efftot<=33 then efftr=2;
      if 33<efftot<=54 then efftr=3;
      if 54<efftot then efftr=4;
run;

proc freq data=tirage1;
      table efftr;
run;

* Création de la table des strates ;

proc freq data=tirage2;
      table code_dept / out=nb0;
run;

data nb0;set nb0;
      drop PERCENT;
      rename COUNT=nb_prob0;
run;

proc sort data=tirage1;by code_dept;run;

data tirage1;merge tirage1(in=xx) nb0;
      by code_dept;
      if xx;
run;

data tirage1;set tirage1;
      if nb_prob0=. then nb_prob0=0;
run;

data tirage1;set tirage1;
      eff_ech_connu0=eff_ech_connu-nb_prob0;
run;

proc sort data=tirage1 out=strates nodupkey;by code_dept;run;

data strates;set strates;
      keep code_dept eff_ech_connu0;
run;

data strates;set strates;
      rename eff_ech_connu0=_NSIZE_;
      label eff_ech_connu0="_NSIZE_";
run;

* Sélection ;

proc sort data=tirage1;by code_dept;run;

proc surveyselect data=tirage1
                        out=echant
                        seed=1
                         sampsize=strates
                        method=pps_sys
                        stats;
      size prob_nouv;
```

```
      strata code_dept;
      control efftr jur2;
run;

* nettoyage -> attention, les poids ne sont pas les bons (il faut mettre
  ceux qu'on trouve comme si on avait tiré au départ 810 organismes ;

data echant;set echant;
      drop expectedhits numberhits samplingweight prob_fin prob_nouv
      eff_ech_connu4 eff_ech_connu0 nb_prob0;
run;

data echant;set echant;
      samplingweight=1/prob_deb;
run;

data echant;set echant;
      rename prob_deb=prob_fin;
run;

* Concaténation avec les probas=1 ;

data tirage2;set tirage2;
      samplingweight=1/prob_deb;
      drop prob_fin eff_ech_connu4 prob_nouv;
run;

data tirage2;set tirage2;
      rename prob_deb=prob_fin;
run;

data echant_fin1;set echant tirage2;run;


/* Ceux dont l'effectif est inconnu */
/*----------------------------------*/

* Il suffit de faire un SAS dans chaque strate de l'échantillon total ;

data base_inc;set dept2;if connu=0 then output;run;

data echant_tot2;set echant_tot;
      if connu=0 then output;
      keep serial connu;
run;

proc sort data=base_inc;by serial;
proc sort data=echant_tot2;by serial;
run;

data tirage;merge base_inc echant_tot2(in=yy);
      by serial;
      if yy;
run;

* Constitution de la table contenant les tailles de strates ;

proc sort data=base_inc out=strates2 nodupkey;by code_dept;run;
```

```
data strates2;set strates2;
      keep code_dept eff_ech_inconnu;
run;

data strates2;set strates2;
      rename eff_ech_inconnu=_NSIZE_;
run;

* Tirage ;

proc sort data=tirage;by code_dept;run;

proc surveyselect data=tirage
                        out=echant2
                        seed=12345
                         method=sys
                        stats
                        sampsize=strates2;
      strata code_dept;
      control jur2;
run;

* Recalcul des probas d'inclusion ;

data echant_fin2;set echant2;
      drop selectionprob samplingweight;
run;

data echant_fin2;set echant_fin2;
      prob_fin=eff_ech_inconnu/eff_inconnu;
      samplingweight=1/prob_fin;
run;


/* Concaténation finale */
/*----------------------*/

data echant_fin;set echant_fin1 echant_fin2;run;

* Sauvegarde ;

data a.echant_reserve1;set echant_fin;run;
```

Annexe 3 - Exemple d'imputation par hot-deck – cas de l'âge du premier arrêt des études

```
proc univariate data=faf;
      var A13;
      histogram A13 /kernel(color=purple);
run;

proc freq data=faf;
      table A13;
run;

* -> certaines données semblent aberrantes, analyse avec le tel ;

data t1 t2 t3;set faf;
      if a13<10 & a13^=-2 then output t1;
      else if a13>29 then output t2;
      else output t3;
run;

proc sort data=t1;by ident;
proc sort data=t2;by ident;
proc sort data=tel;by ident;
run;

data t1_2;set t1(keep=ident a13);rename a13=fa13;run;
data t1_2;merge t1_2(in=xx) tel;by ident;if xx;run;

data t2_2;set t2(keep=ident a13);rename a13=fa13;run;
data t2_2;merge t2_2(in=xx) tel;by ident;if xx;run;

* Solution : imputation
Stratification des donneurs : Tranche d'âge*Niveau d'étude ;

proc sort data=t3;by ident;run;

data t3_2;set t3(keep=ident a13);rename a13=fa13;run;
data t3_2;merge t3_2(in=xx) tel;by ident;if xx;run;

* Taille des strates ;

proc freq data=t1_2;
      table tage*a12 /out=strate1;
run;

proc freq data=t2_2;
      table tage*a12 /out=strate2;
run;

data strate;set strate1 strate2;run;

data strate;set strate;
      rename COUNT=_NSIZE_;
      drop PERCENT;
run;

proc sort data=strate;by tage          a12;run;
```

```
data strate;set strate;
      if tage="3" & a12="6" then _NSIZE_=2;
run;

proc sort data=strate nodupkey;by tage a12;run;

* Tirage ;

proc sort data=t3_2;by tage a12;run;

data t3_x;merge t3_2 strate(in=xx);by tage a12;if xx;run;
data t3_x;set t3_x;drop _NSIZE_;run;

proc surveyselect data=t3_x
                        method=urs
                        out=ech_imput
                        stats
                        seed=1234
                         outhits
                        sampsize=strate;
      strata tage a12;
run;

data ech_imput;set ech_imput(keep=tage a12 fa13);
      i=_N_;
      rename fa13=a13;
run;

data t12;set t1_2 t2_2;run;
proc sort data=t12;by tage a12;run;

data t12;set t12(keep=ident tage a12);
      i=_N_;
run;

data t12_x;merge t12 ech_imput;by i;run;

data t12_x;set t12_x;
      keep ident a13;
run;

data t12_y;set t1 t2;run;

data t12_y;set t12_y;drop a13;run;

proc sort data=t12_y;by ident;
proc sort data=t12_x;by ident;
run;

data t12_z;merge t12_x t12_y;by ident;run;

data faf_2;set t3 t12_z;run;
```

Annexe 4 - Exemple de calage sur marges – cas des interviews en face à face

```
* Calcul des marges pour le vecteur X de redressement de la non-réponse ;

proc freq data=rep_oasp;
      table a6_3 a5_1_2 a5_2_2 a5_3_2 a5_4_2 a5_5_2;
      weight poids_fin;
run;

proc means data=rep_oasp sum;
      var a4_01 a4_02 a4_03 a4_04 a4_05 a4_09;
      weight poids_fin;
run;

data marges;
      input var $ n mar1 mar2;
      cards;
      a6_3 2 370138 44785
      a5_1_2 2 392626 22297
      a5_2_2 2 180395 234528
      a5_3_2 2 194864 220059
      a5_4_2 2 21969 392954
      a5_5_2 2 116073 298850
      a4_01 0 393768 .
      a4_02 0 353689 .
      a4_03 0 321698 .
      a4_04 0 49981 .
      a4_05 0 8530 .
      a4_09 0 16347 .
      ;
run;

* Redressement ;

libname e"D:\...";

options sasmstore=e mstored;

data rep;set rep_oasp;if REP=1 then output;run;

* Méthode linéaire ;

%calmar2(DATAMEN=rep,POIDS=poids_fin,IDENT=ident,MARMEN=marges,
      M=1,EDITPOI=oui,OBSELI=oui,DATAPOI=lin,
      POIDSFIN=poids_lin,MISAJOUR=non,ECHELLE=0);

proc sort data=rep;by ident;
proc sort data=lin;by ident;

data lin;merge rep lin;
      by ident;
      poids_init=poids_fin*1.2596713848;
      d=poids_lin/poids_init;
run;

proc univariate data=lin;
```

```
      var poids_init poids_lin d;
      histogram poids_init poids_lin d /kernel(color=blue);
run;

* Raking-ratio ;

%calmar2(DATAMEN=rep,POIDS=poids_fin,IDENT=ident,MARMEN=marges,
      M=2,EDITPOI=oui,OBSELI=oui,DATAPOI=rr,POIDSFIN=poids_rr,
      MISAJOUR=non,ECHELLE=0);

proc sort data=rep;by ident;
proc sort data=rr;by ident;

data rr;merge rep rr;
      by ident;
      poids_init=poids_fin*1.2596713848;
      d=poids_rr/poids_init;
run;

proc univariate data=rr;
      var poids_init poids_rr d;
      histogram poids_init poids_rr d /kernel(color=red);
run;

* Logit ;

%calmar2(DATAMEN=rep,POIDS=poids_fin,IDENT=ident,MARMEN=marges,
      M=3,EDITPOI=oui,OBSELI=oui,DATAPOI=log,
      POIDSFIN=poids_log,MISAJOUR=non,ECHELLE=0,LO=.81,
         UP=1.24);

proc sort data=rep;by ident;
proc sort data=log;by ident;

data log;merge rep log;
      by ident;
      poids_init=poids_fin*1.2596713848;
      d=poids_log/poids_init;
run;

proc univariate data=log;
      var poids_init poids_log d;
      histogram poids_init poids_log d /kernel(color=purple);
run;

* Linéaire tronquée ;

%calmar2(DATAMEN=rep,POIDS=poids_fin,IDENT=ident,MARMEN=marges,
      M=4,EDITPOI=oui,OBSELI=oui,DATAPOI=lt,POIDSFIN=poids_lt,
      MISAJOUR=non,ECHELLE=0,LO=.81,UP=1.24);

proc sort data=rep;by ident;
proc sort data=lt;by ident;

data lt;merge rep lt;
      by ident;
      poids_init=poids_fin*1.2596713848;
      d=poids_lt/poids_init;
```

```
run;

proc univariate data=lt;
      var poids_init poids_lt d;
      histogram poids_init poids_lt d /kernel(color=brown);
run;

* Sinus hyperbolique ;

%calmar2(DATAMEN=rep,POIDS=poids_fin,IDENT=ident,MARMEN=marges,
      M=5,EDITPOI=oui,OBSELI=oui,DATAPOI=sh,POIDSFIN=poids_sh,
      MISAJOUR=non,ALPHA=11.72);

proc sort data=rep;by ident;
proc sort data=sh;by ident;

data sh;merge rep sh;
      by ident;
      poids_init=poids_fin*1.2596713848;
      d=poids_sh/poids_init;
run;

proc univariate data=sh;
      var poids_init poids_sh d;
      histogram poids_init poids_sh d /kernel(color=green);
run;
```

Annexe 5 - Macro de troncature des poids

```
data t0;set oasp2;
      keep ident B2 B6 B6b B7 poids_init;
run;

%MACRO tronq(CENTILE=);

      data t&CENTILE;set t0;run;

      * sortie des statistiques ;

      proc univariate data=t&CENTILE noprint;
            var poids_init;
            output out=s&CENTILE pctlpts=&CENTILE pctlpre=p;
      run;

      * Fusion avec la table initiale ;

      data t&CENTILE;set t&CENTILE;i=1;
      data s&CENTILE;set s&CENTILE;i=1;run;

      data t&CENTILE;merge t&CENTILE s&CENTILE;by i;run;

      /* Boucle des troncatures */
      /*------------------------*/

      data t&CENTILE;set t&CENTILE;
            wkt=poids_init;
      run;

      * Extraire le max de wkt ;

      proc means data=t&CENTILE max;
            var wkt;
            output out=maxwkt max=max;
      run;
      proc means data=t&CENTILE max;
            var p&CENTILE;
            output out=maxctl max=max;
      run;

      data _NULL_;set maxwkt;
            call symput('Maxwkt',max);
      run;
      data _NULL_;set maxctl;
            call symput('Cent',max);
      run;

      %PUT Données de départ :
            Max_wkt : &Maxwkt Centile_&CENTILE : &Cent;

      data t&CENTILE._X;set _NULL_;run;

      %DO %UNTIL(%SYSEVALF(&Maxwkt.<=&Cent.));

            * Troncature ;
```

```
data t&CENTILE;set t&CENTILE;
      tau=(wkt>=p&CENTILE);
      numer=wkt-tau*p&CENTILE;
      denom=(1-tau)*wkt;
run;

* Fin des calculs pour la constante a ;

proc means data=t&CENTILE sum noprint;
      var numer;
      output out=s&CENTILE._numer sum=sum_numer;
run;
proc means data=t&CENTILE sum noprint;
      var denom;
      output out=s&CENTILE._denom sum=sum_denom;
run;

data s&CENTILE._numer;set s&CENTILE._numer;
      i=1;
      drop _TYPE_ _FREQ_;
run;
data s&CENTILE._denom;set s&CENTILE._denom;
      i=1;
      drop _TYPE_ _FREQ_;
run;

data t&CENTILE;merge t&CENTILE s&CENTILE._numer
                     &CENTILE._denom;
      by i;
      a&CENTILE.=sum_numer/sum_denom;
run;

* Pondération finale ;

data t&CENTILE;set t&CENTILE;
      wkt1=tau*p&CENTILE.+(1-tau)*wkt*a&CENTILE;
run;

data t&CENTILE._1 t&CENTILE._2;set t&CENTILE;
      wkt=wkt1;
      if wkt=p&CENTILE then output t&CENTILE._1;
* pour éviter de remettre les poids calés dans la boucle ;
      else output t&CENTILE._2;
run;

data t&CENTILE._X;set t&CENTILE._X t&CENTILE._1;run;
data t&CENTILE;set t&CENTILE._2;
      drop wkt1 sum_numer sum_denom a&CENTILE;
run;

* Réinitialisation de Maxwkt ;

proc means data=t&CENTILE max;
      var wkt;
      output out=maxwkt max=max;
run;
data _NULL_;set maxwkt;
      call symput('Maxwkt',max);
```

```
            run;

            %PUT Centile : &Cent , Max_wkt &Maxwkt;
            %PUT ;

      %END;

      data t&CENTILE;set t&CENTILE t&CENTILE._X;

      * Nettoyage ;

      data t&CENTILE;set t&CENTILE;
            drop p&CENTILE tau numer denom sum_numer sum_denom;
      run;

      proc delete data=s&CENTILE s&CENTILE._numer s&CENTILE._denom
                       t&CENTILE._1 t&CENTILE._2 t&CENTILE._X;
      run;

      options notes;

%MEND tronq;


%tronq(CENTILE=88);
```

Annexe 6 - Macro de calcul des erreurs quadratiques moyennes sur les poids tronqués

```
%MACRO EQM(VAR=,ARRET=);

      * Initialisation ;

      data S_&VAR;set _NULL_;run;

      %DO j=1 %TO &ARRET;

            %LET i=%EVAL(100-&j.);

            %tronq(CENTILE=&i);

            proc surveymeans data=t&i;
                  var &VAR;
                  weight poids_init;
                  ods output Statistics=S&i._1;
            run;

            proc surveymeans data=t&i;
                  var &VAR;
                  weight wkt;
                  ods output Statistics=S&i._2;
            run;

            * Calcul des variances ;

            data s&i._1;set s&i._1;
                  varybar=stderr**2;
                  rename mean=ybar;
            run;
            data s&i._2;set s&i._2;
                  varybart=stderr**2;
                  rename mean=ybart;
            run;

            * Nettoyage ;

            data s&i._1;set s&i._1;keep ybar varybar;
            data s&i._2;set s&i._2;keep ybart varybart;
            run;

            * Calcul EQM ;

            data s&i._1;set s&i._1;i=1;
            data s&i._2;set s&i._2;i=1;

            data s&i;merge s&i._1 s&i._2;
                  by i;
                  EQM=(ybart-ybar)**2+varybart;
                  MOYENNE=ybart;
                  BIAIS=ybart-ybar;
                  BIAIS_rel=(ybart-ybar)/ybar;
                  VARIANCE=varybart;
                  STD=sqrt(varybart);
```

```
            run;

            * Dernier nettoyage de la boucle ;

            data s&i;set s&i;
                  CENTILE=&i;
                  drop ybar varybar i ybart varybart;
            run;

            * Concaténation ;

            data S_&VAR;set S_&VAR s&i;run;

            proc delete data=s&i s&i._1 s&i._2 t&i;run;

      %END;

%MEND EQM;
```