Méthodologie statistique

N°M0501

CORRECTION DE LA NON REPONSE ETCALAGE DE L'ENQUETE SANTE 2002

Nathalie Caron et Sylvie Rousseau

Document de travail

Institut National de la Statistique et des Etudes Economiques

INSTITUT NATIONAL DE LA STATISTIQUE ET DES ÉTUDES ÉCONOMIQUES
Série des Documents de Travail
de la

DIRECTION DES STATISTIQUES DEMOGRAPHIQUES ET SOCIALES

Unité « Méthodes Statistiques »

Série des Documents de Travail
Méthodologie Statistique

N° M0501

CORRECTION DE LA NON REPONSE ET CALAGE DE L'ENQUETE SANTE 2002

Nathalie CARON et Sylvie ROUSSEAU
INSEE, UMS

Novembre 2005

Correction de la non-réponse et calage de l'enquête Santé 2002

Nathalie Caron et Sylvie Rousseau
Insee, UMS

Résumé:

Réalisée régulièrement tous les 10 ans par l'Insee depuis 1960, l'enquête « Santé » a pour objectif principal de décrire les consommations médicales et l'état de santé de la population métropolitaine. Lors de la dernière édition en 2002, près de 40 000 individus ont été interrogés. Chacun devait répondre à trois questionnaires, un au cours de chacune des trois visites que leur rendaient les enquêteurs, d'un mois sur l'autre.

L'objectif de ce document est de décrire la démarche méthodologique adoptée pour établir les pondérations finales qui sont utilisées pour exploiter l'enquête Santé. Ces poids tiennent compte de la correction de la non-réponse totale et du calage des données d'enquête.

Deux jeux de pondérations sont proposés au niveau des individus :

- le premier jeu s'applique aux individus ayant répondu à la 1ère visite et permet d'exploiter les variables disponibles dans le premier questionnaire (remis lors de la visite 1)
- le second jeu s'applique aux individus ayant répondu aux trois visites à la fois (en nombre naturellement plus faible que ceux n'ayant répondu qu'à la première visite). Cette pondération permet d'exploiter l'ensemble des informations disponibles pour ces individus, qu'elles aient été recueillies aux cours des visites 1, 2 ou 3.

Pour chacun de ces deux jeux de poids, les différentes étapes de correction de la non-réponse totale et de calage sont détaillées dans ce document. Un soin particulier a été apporté pour justifier les orientations retenues.

Mots clés : correction de la non-réponse, calage sur marges, macro CALMAR, précision, enquête santé.

Sommaire

I. Caractéristiques générales

I.1. Présentation de l'enquête

L'enquête « Santé » réalisée tous les 10 ans a pour **principal objectif** de fournir une information détaillée sur l'état de santé de la population, sur la consommation de soins et sur la prévention. Cette enquête constitue une référence sur la mesure de la santé et de la consommation médicale car elle permet de faire le lien entre les dépenses de santé qui sont déjà connues au niveau macroéconomique, les consommations médicales effectives et les caractéristiques socio-démographiques des individus effectuant ces consommations. La dernière enquête a eu lieu sur la période octobre 2002-septembre 2003. **L'unité statistique interrogée** sur la consommation médicale **est l'individu**.

L'enquête est réalisée **en face à face au cours de 3 visites** d'une heure environ et qui sont espacées d'un mois. Sa saisie est administrée sous CAPI (collecte assistée par informatique).

Lors de la première visite, deux types de questionnaires sont remplis : un questionnaire ménage et autant de questionnaires « individu » que de personnes du ménage (y compris pour les enfants). Le **questionnaire "ménage"** comporte un « tronc commun » fournissant des données biographiques ainsi que des informations sur le logement, les aides, le revenu, la protection sociale des individus composant le ménage, leurs gênes et difficultés dans la vie quotidienne. **Le questionnaire "individuel"** comporte des questions générales sur la santé, une liste des maladies en cours, des questions relatives à la grossesse, les traitements, la vue, les dents, l'audition et le recours au médecin pendant les 12 derniers mois.

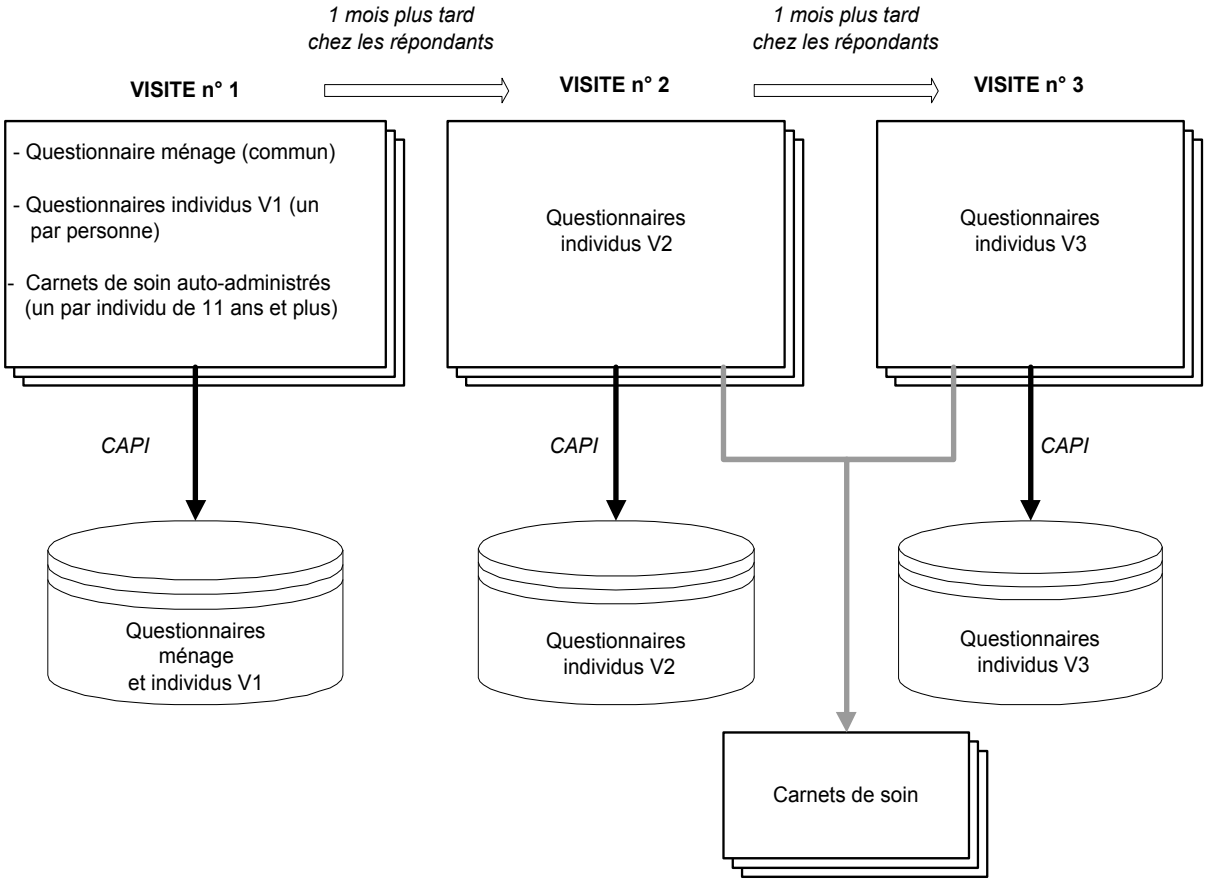
De plus, l'enquêteur laisse à la première visite **un carnet de soin auto-administré** pour toutes les personnes de 11 ans et plus, aptes à la tenue d'un tel document (en version allégée pour les deux tranches d'âge suivantes : 11-14 ans et 15-17 ans). Celui-ci aborde certains thèmes personnels tels que la dépression, l'asthme et les problèmes respiratoires, la migraine, la lombalgie, les événements au cours de la vie et les conditions de travail. Il sert d'aide-mémoire pour les 2ème et 3ème visites : chaque enquêté y reporte au quotidien les dates de soins, le contenu des ordonnances, le nom des médicaments… Ce carnet est repris par l'enquêteur lors d'une prochaine visite.

Lors de la seconde visite, tous les individus répondants lors de la première visite remplissent un questionnaire qui porte sur les nouvelles maladies depuis la dernière visite, les hospitalisations des 12 derniers mois, les alitements et les diverses visites à des médecins ou dentistes, les examens médicaux, les soins infirmiers ou de kinésithérapie, les médicaments. Il renseigne également sur la taille et poids.

Le **questionnaire de la troisième visite** est une reprise allégée des questions de la deuxième visite, portant cette fois sur le dernier mois. En outre, il mesure la consommation de médicaments de la veille de l'enquête et renseigne sur la prévention les comportements et les facteurs de risque. Il s'adresse à tous les répondants de la seconde visite.

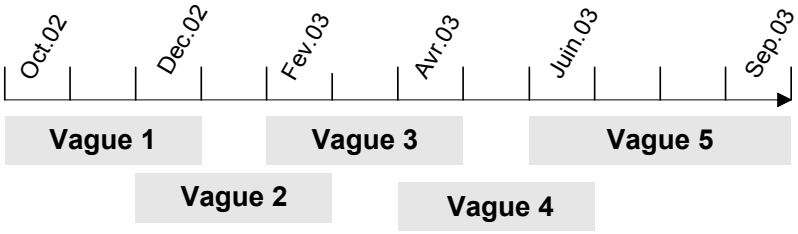
Pour chaque ménage sélectionné, le dispositif prévoit donc un questionnaire ménage, autant de questionnaires individus que de personnes du ménage et de visites et d'autant de carnets de soin que de personnes de plus de 11 ans dans le ménage (voir schéma 1).

Schéma 1 : le dispositif d'interrogation



Enfin, **pour appréhender l'effet saisonnier lié à la santé**, **la collecte** de l'enquête **s'est déroulée en 5 vagues**, de septembre 2002 à octobre 2003, chacune sur une période de 3 mois sauf la vague d'été qui a duré 4 mois. La présence de la vague estivale constitue d'ailleurs une nouveauté par rapport à la précédente enquête « Santé » de 1990-91 (de 4 vagues de 3 mois sans l'été). Le calendrier des vagues se chevauchent comme précisé dans le schéma 2.

Schéma 2 : le calendrier de collecte



I.2. Les principaux objectifs d'exploitation

L'enquête « Santé » a pour **principaux objectifs, au niveau national** :

- de dresser un panorama de l'état de santé de la population,
- d'analyser les liens entre l'état de santé (ressenti ou objectif) et le recours aux soins en fonction des caractéristiques socio-démographiques des individus,
- d'analyser les trajectoires de prise en charge médicale en fonction des caractéristiques socio-démographiques des individus,
- et de fournir des éléments nécessaires en volume et en valeur à l'estimation de la consommation de soins et de prévention.

Ces objectifs se déclinent également **au niveau régional** dans les cinq régions où des extensions ont été réalisées : Nord-Pas-de-Calais, Provence-Alpes-Côte-d'Azur, Picardie, Ile-de-France et Champagne-Ardennes. Grâce à un échantillon de taille plus importante, ces extensions permettent de disposer de davantage de données et autorisent les comparaisons entre les régions concernées et/ou avec le niveau national.

De plus**, à titre expérimental**, il est prévu d'**apparier**[1] les données de l'enquête **avec la base de données inter-régime d'assurance maladie** qui contient l'ensemble des dépenses portées au remboursement des assurés sociaux et de leurs ayant droit. Cette approche permet d'une part de pouvoir comparer la consommation déclarée dans l'enquête avec celle effective sur la période étudiée et d'autre part de connaître le montant annuel de consommation remboursée.

I.3. Rappels sur le plan de sondage[2]

L'unité statistique interrogée dans l'enquête est **l'individu**. Cependant, en l'absence de base de sondage au niveau individu, c'est par l'intermédiaire des logements et des ménages que les individus sont approchés. Plus précisément, **le champ de l'enquête** est défini par l'ensemble des personnes qui, au moment du passage de l'enquêteur, occupent les logements échantillonnés au titre de leur résidence principale.

Les bases de sondage utilisées sont au nombre de quatre : il s'agit d'une part, pour les résidences recensées en mars 1999, de l'échantillon maître (**EM**) complété dans les régions à extension par l'échantillon maître pour les extensions régionales (**EMEX**) ; d'autre part, pour les logements construits après le recensement, de la base de sondage des logements neufs (**BSLN**) enrichie dans les régions à extension par des logements issus de permis de construire déclarés achevés par le ministère de l'équipement **(SITADEL-EMEX**)[3].

Plus précisément,

- L**'échantillon maître** se définit comme une **"réserve" localisée de logements construite à partir du dernier recensement**. Il permet d'alimenter la plupart des échantillons des enquêtes ménages nationales[4] jusqu'au prochain recensement Son existence et son mode de constitution reposent principalement sur le fait que les enquêtes ménages sont réalisées en face à face. De fait, pour limiter au maximum les frais de déplacements des enquêteurs (surtout dans la partie rurale), il convient de ne pas trop disperser les lieux d'enquêtes. De plus, l'EM a été conçu pour concilier cette contrainte pratique avec l'objectif de fournir des données d'enquêtes suffisamment précises au niveau national.

[1] Cet appariement sera pratiqué avec l'aide de trois partenaires avec une technique de « double-aveugle » afin qu'aucun des partenaires ne dispose de l'ensemble des informations identifiées de l'enquête et celles issues de la base de données inter-régime.
[2] Pour en savoir plus, voir la fiche récapitulative du plan de sondage (note n°048/F410 du 22/07/2002) disponible en annexe 2.
[3] Système d'Information et de Traitement Automatisé des Données Elémentaires sur le Logement et les locaux.
[4] Il existe quelques exceptions comme l'enquête Handicap-Incapacités et Dépendances, l'enquête auprès des Sans-domicile ou encore l'enquête mensuelle CAMME de conjoncture auprès des ménages mais celles-ci sont en faible nombre.

La base actuelle -mise en service pour la première fois en 2001- est composée d'environ deux millions de logements provenant du recensement de la population de 1999 (**RP99**), soit près de 7% des logements recensés. Elle résulte d'un **plan de sondage à plusieurs degrés stratifié par catégories de communes,** le nombre de degrés étant variable selon la strate. **La stratification** permet de distinguer les communes rurales des communes urbaines, celles-ci étant classées selon la taille de l'unité urbaine à laquelle elles appartiennent. **Les unités primaires** sont des cantons, des regroupements ou, parfois, des fractions de cantons dans les zones rurales et des unités urbaines ou des regroupements d'unités urbaines ailleurs. Dans le rural, le petit urbain et le moyen urbain, seules certaines unités primaires sont retenues dans l'échantillon-maître (selon un tirage à probabilités inégales proportionnellement au nombre de résidences principales[5] de chaque région). Toutes les "grandes" unités urbaines (de plus de 100 000 habitants) sont retenues au niveau de ce premier degré. Cependant, **un tirage supplémentaire de districts** avec un taux de sondage 1/20 y est réalisé afin de concentrer encore plus les enquêtes. Il en est de même pour les unités urbaines de moyen urbain (de 20 000 à 100 000 habitants). L'échantillon-maître est finalement constitué de l'ensemble des logements appartenant aux unités ainsi obtenues.

- Pour tenir compte des logements construits après le recensement (à défaut, la représentativité de l'échantillon-maître diminuerait avec le temps), l'échantillon-maître est complété par **la base de sondage des logements neufs**. Cette base est constituée à partir des permis de construire autorisés que gére le ministère de l'équipement. Le principe consiste à extraire un échantillon de logements[6] (environ 5% des logements) dans les permis de construire localisés dans les unités primaires de l'échantillon-maître. Les logements sélectionnés font l'objet d'un suivi dans les directions régionales jusqu'à leur achèvement. Ceux qui sont déclarés achevés alimentent ensuite la BSLN et deviennent susceptibles d'être sélectionnés lors des tirages d'enquêtes. La taille de cette base augmente donc chaque année en fonction du volume de la construction neuve (près de 45 000 logements fin 2002).

Comme précisé précédemment, l'enquête « Santé » de 2002-03 a fait l'objet de **cinq extensions régionales.** Les financements locaux ont pratiquement permis, dans les cinq régions concernées, de doubler la taille de l'échantillon par rapport à celle qui aurait été obtenue en absence d'extension[7]. Le tirage de ces extensions a appelé deux bases de sondage complémentaires : l'échantillon maître pour les extensions régionales (**EMEX**) et la base **SIDATEL-EMEX** :

- **L'EMEX** est une réserve supplémentaire qui a été mise en place en 2001 pour pouvoir établir des résultats régionaux avec une précision acceptable[8]. En effet, l'échantillon-maître n'a pas été construit dans l'optique d'obtenir des résultats régionaux à partir des parties régionales des échantillons nationaux, souvent de taille insuffisante[9]. Il est par conséquent nécessaire de réaliser des extensions régionales d'enquêtes, en augmentant la taille de l'échantillon. L'EMEX est ainsi une **réserve localisée de logements recensés.** Sa constitution repose sur les mêmes principes que ceux de l'EM. Notons que l'EMEX permet d'homogénéiser le tirage et les traitements des extensions régionales associées à des enquêtes nationales. Il évite également les problèmes de réserve insuffisante par ponction excessive dans l'EM. De surcroît, ce dispositif permet au niveau national de tirer parti des extensions en exploitant simultanément les données nationales et régionales. Ces dernières alimentent, quant à elles, dans les régions à extension, les exploitations régionales et les comparaisons inter-régionales.

  L'EMEX a été utilisé pour la première fois en septembre 2002 pour le tirage des extensions régionales de l'enquête Santé : dans les régions à extension, la partie recensée de l'échantillon a été sélectionnée dans la base de sondage formée de la

[5] et qui est équilibré selon le revenu et l'âge des individus en trois tranches d'âge au niveau de regroupement de régions.
[6] repérés uniquement par un numéro séquentiel au sein des permis sélectionnés (faute d'informations complémentaires dans Sitadel).
[7] La répartition régionale de la charge d'enquête est décrite dans la fiche récapitulative du plan de sondage (disponible en annexe 2).
[8] Les premières estimations de précision sont disponibles dans la note interne Insee n° 072/F410 d'août 2005. Elles ont été calculées avec le logiciel POULPE, conçu à l'Insee.
[9] L'EM n'a pas, a priori, une représentativité régionale suffisante, notamment parce que le nombre d'unités primaires tirées par région est trop faible. De plus, il n'y a pas de condition explicite de couverture du territoire de la région par l'échantillon-maître : ainsi, celui-ci peut impacter la région de manière plus ou moins bien répartie et il peut arriver que certains départements ne contiennent pas d'unités primaires rurales.

réunion « EM + EMEX » « représentative»[10] de chaque région au regard de critères d'âge et de revenu. Dans les autres régions, seul l'EM a été mobilisé.

- De la même façon, pour le tirage des logements neufs dans une enquête à extension régionale, la BSLN est complétée par un **extrait de logements dans la liste des permis déclarés achevés** au sens du ministère de l'équipement **et localisés dans les zones « EMEX** ». Pour l'enquête Santé, dans les régions à extension, les logements construits depuis mars 1999 ont ainsi été échantillonnés dans l'ensemble « BSLN - SIDATEL (EMEX) » (dans la BSLN seule, pour les autres régions).

Pour finir, l'échantillon de l'enquête Santé résulte d'un **plan de sondage stratifié par catégorie de communes**. Le tirage s'effectue de telle sorte que **tous les ménages aient la même chance d'être interrogés** quelle que soit leur localisation sur le territoire. Ici, un **distinguo** s'établit naturellement dans **les régions à extension** où cette probabilité s'accroît avec la taille de l'échantillon, à hauteur des financements locaux. Partout ailleurs, le taux de sondage de tous les logements principaux est identique. Ce taux s'applique aussi aux logements neufs que l'on considère destinés à l'habitat principal. Par ailleurs, pour tenir des changements de catégorie de logement intervenus depuis le recensement de mars 1999, l'échantillon comprend également des logements déclarés occasionnels, secondaires ou vacants en mars 1999. Leur taux de sondage est cependant moindre que ceux des résidences principales compte-tenu du champ d'interrogation (cf. tableau 1 et annexe 2 pour de plus amples précisions).

Au final, **l'échantillon comporte 25 021 logements**, répartis sur l'ensemble du territoire métropolitain. **Chacune des 5 vagues** de collecte **porte** sur un échantillon de **5 000 logements** environ.

Tableau 1 : nombre de logements sélectionnés par catégorie de logement

Catégorie de logement	*Effectif échantillonné*
Résidences principales au RP99	21 511
Résidences secondaires au RP99	1 123
Résidences occasionnelles au RP99	118
Résidences vacantes au RP99	1 619
Logements achevés après le RP99	650
Total	25 021

[10] au sens du tirage équilibré. Pour plus de détail, consulter par exemple L. Wilms (2000) ou S. Rousseau, F. Tardieu (2004).

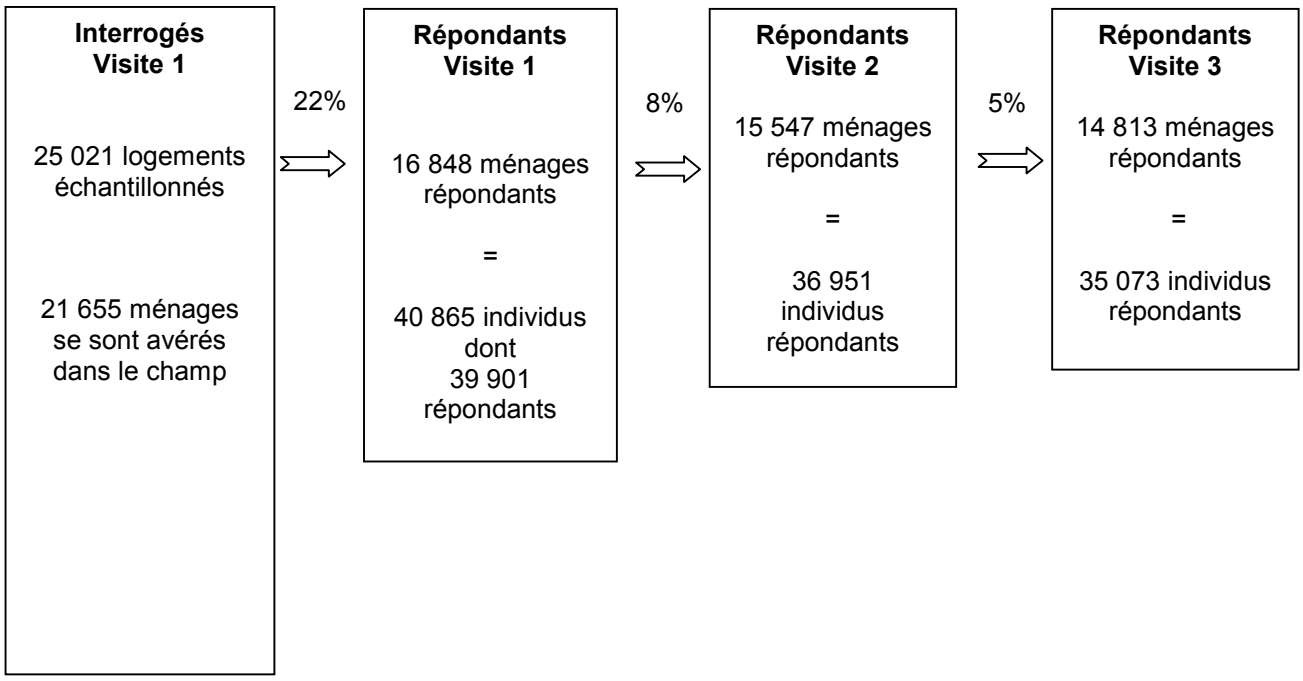
II. Description de la méthodologie adoptée

II.1. Les choix adoptés

Comme nous l'avons signalé dans la partie I.1., pour chaque individu nous devrions disposer des trois questionnaires recueillis au cours des différentes visites de l'enquêteur, de l'autoquestionnaire si cet individu a plus de 11 ans et du questionnaire relatif au ménage auquel il appartient. Cependant, comme toute enquête, l'enquête « Santé » est confrontée à la **non-réponse** ; celle-ci est susceptible de se produire pour le remplissage de chacun des questionnaires.

D'après les consignes de collecte, **seuls les individus répondants à la première (resp. à la seconde) visite ont été contactés pour une seconde visite (resp. la dernière visite).** Par construction, le taux de non-réponse ne peut que stagner ou augmenter avec le nombre de visites en fonction de « l'effet lassitude » des personnes enquêtées. Ainsi, parmi celles qui ont répondu à la première visite, 13 % n'ont pas souhaité répondre lors de la troisième visite (voir schéma 3 ci-dessous). Il s'avère que la non-réponse intervient principalement au moment de la prise de contact avec le ménage pour la première visite et entre les visites 1 et 2 ; par contre peu de non-répondants supplémentaires surviennent entre les visites 2 et 3.

Schéma 3 : Comportements de non-réponse observés et taux de non-réponse des ménages d'une visite à l'autre



Au maximum, 5 jeux de pondérations sont possibles :

i. Une pondération ménage,
ii. Une pondération individu pour ceux qui ont répondu au premier questionnaire,
iii. Une pondération individu pour ceux qui ont répondu aux deux premiers questionnaires,
iv. Une pondération individu pour ceux qui ont répondu aux troix questionnaires,
v. Enfin une pondération individu pour ceux qui ont répondu à l'autoquestionnaire.

Afin de **simplifier l'exploitation** de cette enquête pour ses utilisateurs, nous avons décidé de **limiter le nombre de jeux de pondération distincts**. **Trois systèmes de poids** sont proposés au niveau des individus :

- Le premier jeu s'applique **aux individus ayant répondu à la 1ère visite** et **permet d'exploiter les variables disponibles dans le premier questionnaire** (remis lors de la visite 1).
- Le second jeu s'applique **aux individus ayant répondu aux trois visites à la fois** (en nombre plus faible que ceux n'ayant répondu qu'à la première visite). Cette pondération **permet d'exploiter l'ensemble des informations disponibles pour ces individus**, qu'elles aient été recueillies aux cours des visites 1, 2 ou 3.
- Un dernier jeu de pondérations est prévu pour **exploiter les autoquestionnaires**. Son calcul a été réalisé ultérieurement car il a d'abord fallu définir la notion de questionnaires ''exploitables''. Ceci nécessitait que les différents partenaires de l'enquête déterminent ensemble la liste minimale des questions auxquelles les individus devaient avoir répondu. Si on obtenait une réponse à toutes les questions de cette liste, alors le questionnaire était déclaré exploitable ; si ce n'était pas le cas, le questionnaire était considéré comme manquant, c'est-à-dire comme une non-réponse totale. La détermination de la liste a été d'autant plus délicate à réaliser que les partenaires avaient des objectifs d'exploitation différents. De plus, elle devait être relativement courte, sinon un trop grand nombre d'autoquestionnaires risquaient d'être éliminés, ce qui aurait par conséquent gonflé le taux de non-réponse. Ce n'est qu'à l'issue de cette étape qu'une pondération relative à l'exploitation des réponses recueillies dans les autoquestionnaires a été élaborée. Ce système de pondération n'est pas présenté dans ce document du fait de la similitude des traitements effectués avec les précédents.

Le choix des jeux de pondérations à utiliser a été fixé au vu des éléments suivants :

- **L'unité statistique interrogée étant l'individu**, il semble naturel de privilégier le redressement à ce niveau. Le redressement comprend deux opérations distinctes[11] : la correction de la non-réponse totale et le calage sur des données exogènes;

- La perte d'information entre les visites 1 et 3 est significative, tant au niveau des ménages que des individus. Cet élément participe au choix de **ne pas s'en tenir à une unique pondération, en fin de visite 3** ;

- A contrario, les non-réponses les plus nombreuses s'observent dès la 1ère visite et concernent essentiellement les ménages (il y a peu de de non-réponse supplémentaire des individus dans les ménages répondants lors de la visite 1). En outre, les questionnaires ménages et individus de la visite 1 sont suffisamment complets pour répondre à certains objectifs d'exploitations de l'enquête. C'est pourquoi nous avons construit **un jeu de pondérations pour exploiter** les données recueillies au cours de **la 1ère visite** ;

- Le faible nombre d'individus non-répondants supplémentaires entre la visite 2 et la visite 3 contribue **à écarter le choix d'une pondération spécifique à la 2ème visite** (la perte d'information qui s'en déduit est acceptable) ;

- **quant à l'autoquestionnaire**, **la différence de champ** entre les personnes concernées et celui des questionnaires justifie à elle seule une nouvelle pondération.

D'autre part, il nous semble important **de tenir compte des deux éléments suivants** dans le redressement :

- le fait que le plan de sondage **prenne explicitement en considération les 5 extensions régionales financées,**
- et le fait que **la collecte de l'enquête a été étalée sur 5 vagues** (de taille égale en nombre de logements enquêtés) pour mieux appréhender les effets saisonniers.

[11] Voir partie II.3. ci dessous

II.2. La source externe retenue pour le calage

Le calage de toutes les enquêtes « ménages » réalisées à l'Insee hormis les plus grosses - en pratique Logement et Formation et Qualification Professionnelle - utilisent comme source externe **l'enquête Emploi.**

Cette pratique ancienne de l'institut s'explique tout simplement par le fait que la taille de l'échantillon de l'enquête Emploi est plus importante que celle des autres enquêtes ménages : jusqu'en 2002, l'enquête Emploi, réalisée auprès de 80 000 ménages, était annuelle et les marges utilisées étaient celles issues de l'enquête Emploi réalisée l'année de collecte de l'enquête à caler. L'enquête en continu interroge quant à elle 54 000 logements chaque trimestre, avec un taux de rotation d'1/6 chaque trimestre, ce qui conduit à enquêter un peu plus de 80 000 ménages distincts chaque année.

Avec la mise en place en 2002 de ce nouveau dispositif de collecte pour l'enquête emploi, le contexte a changé et il a été nécessaire de redéfinir la manière d'utiliser les résultats de l'enquête Emploi, publiés désormais à un rythme trimestriel, pour le calage des autres enquêtes ménages.

En vue d'une utilisation comme source de calage, l'enquête Emploi est spécifiquement calée sur la pyramide des âges de la population vivant en ménage ordinaire seulement, à l'exclusion, donc, des communautés (cette pyramide étant issue des estimations de population calculées par l'Insee à partir de l'état civil et du recensement de la population). On obtient les marges annuelles fondées sur l'année civile en effectuant la moyenne simple des 4 structures trimestrielles - processus non théoriquement optimal au sens de la précision mais néanmoins simple et robuste[12].

Ainsi, toutes les enquêtes dont la collecte se déroule durant l'année n sont calées sur une structure moyenne estimée à partir des résultats des 4 trimestres de l'enquête Emploi de l'année n-1. Remarquons que, in fine, les enquêtes « ménages » sont implicitement calées sur la pyramide des âges de la population vivant en ménage ordinaire.

II.3. Les principes généraux

Pour la mise au point de chacun des jeux de pondérations, **le redressement de l'enquête Santé est réalisé en deux temps en séparant l'étape de correction de la non-réponse totale[13] de celle du calage sur des marges exogènes.** Ces deux étapes sont détaillées ci-dessous et résumées dans le schéma 4.

Au préalable, plusieurs traitements statistiques devaient être réalisés avant de passer à la correction de la non-réponse proprement dite. En particulier, la correction de la non-réponse totale ne se fait que sur le champ des résidences principales considérées comme telles à la date de l'enquête (autrement dit les logements **« hors champ** » n'entrent à aucun moment en considération pour le redressement de la non-réponse). Il convient donc de séparer au préalable, parmi les logements pour lesquels on n'a aucun retour, les logements « hors champ » des logements dans le champ de l'enquête. De même, il est important de vérifier si les **questionnaires** remplis sont bien **« exploitables** ».

[12] On démontre (voir J.C. Deville, 1999) que la meilleure information auxiliaire que l'on peut bâtir à partir de deux enquêtes consiste en une combinaison linéaire des estimations issues de ces 2 enquêtes. Cette solution qui conduit donc à l'utilisation de marges différentes pour chaque enquête à redresser risque de poser un problème de diffusion : afin que les totaux les plus stratégiques (au moins taille de la population, nombre de logements, ventilation par âge x sexe, …) gardent une cohérence entre différentes enquêtes auprès des ménages, il a été décidé de bâtir les marges à partir de l'enquête Emploi uniquement.

[13] La non-réponse est dite totale lorsqu'un individu échantillonné ne répond à aucune des questions posées ou est considéré comme tel car ses réponses sont inexploitables. A contrario, la non-réponse est dite partielle lorsque seuls certains renseignements font défaut.

II.3.1. Traitement de la non-réponse totale

a. Justification de ce traitement

Le traitement de la non-réponse totale vise principalement **à diminuer le biais** occasionné par les non-répondants. En effet, la présence de données manquantes influe sur la qualité de l'inférence. Utiliser les formes habituelles d'estimateurs en les restreignant aux seuls répondants est problématique si les non-répondants possèdent, en moyenne, un comportement différent de celui des répondants pour le thème de l'enquête. Or il en est généralement ainsi, car **la non-réponse est rarement le fruit du hasard**. Par exemple, on peut aisément imaginer que des personnes gravement malades répondent moins facilement à l'enquête que celles en bonne santé ; ignorer ces malades faute de réponse conduirait à sur-estimer l'état de santé moyen des français.

D'autre part, les estimateurs construits sur les seuls répondants admettent **une variance d'échantillonnage plus importante** (ils seront donc moins précis), puisque la taille de l'échantillon exploitable est plus faible que celle prévue au moment du tirage.

b. Principe général de la méthode

Le traitement de la non-réponse totale a été réalisé par repondération : le principe consiste à modifier le poids des unités répondantes pour compenser la présence de non-réponse totale pour extrapoler les résultats obtenus à la population de référence. Le poids d'échantillonnage de chaque unité répondante est ainsi augmenté par l'inverse de sa **probabilité de réponse**, quantité inconnue **qu'il faut estimer**.

Dans ce but, nous avons supposé que le **mécanisme de réponse était homogène à l'intérieur de sous-populations**[14]. Cette approche repose sur l'hypothèse qu'à l'intérieur de sous-populations particulières (qui sont à définir) les individus possèdent tous la même probabilité de répondre et que leurs comportements de réponse sont indépendants. La probabilité de réponse au sein d'un groupe donné est estimée en rapportant le nombre d'unités répondantes à l'effectif échantillonné (et appartenant au champ de l'enquête).

La partie la plus délicate pour mettre en œuvre cette méthode consiste à **définir les sous-populations**. Plusieurs techniques sont possibles et nous avons privilégié l'approche suivante :

- **Dans un premier temps**, il s'agit de **modéliser le comportement de réponse des unités échantillonnées et présentes dans le champ de l'enquête par des caractéristiques connues à la fois sur les répondants et les non-répondants**. Le modèle peut résulter d'une régression logistique non pondérée de la variable « répond/ ne répond pas » avec comme variables explicatives[15] des données issues des bases de sondage et la vague de collecte. Une attention particulière est portée aux **effets régionaux** pour tenir compte des extensions régionales. **Pour le traitement de la non-réponse de la 3ème visite, des réponses à l'enquête Santé elle-même** (obtenues au cours de de la 1ère visite) ont également pu être utilisées, ce qui a permis d'introduire l'état de santé des enquêtés comme facteur éventuel de non-réponse. Une fois optimisé (par regroupement de modalités de variables, élimination des variables explicatives non pertinentes, etc.), le modèle obtenu permet de dégager les motifs de non-réponse et de former les différentes sous-populations.

- **Les sous-populations s'obtiennent** par des regroupements effectués selon l'une ou l'autre des méthodes suivantes :
  - En croisant toutes les modalités des variables explicatives retenues dans le modèle,
  - En fractionnant la liste des individus préalablement triée par probabilités de réponse estimées croissantes (ou décroissantes) en groupes de taille égale.

[14] Pour plus de détails, consulter par exemple N. Caron (1996).

[15] Les variables explicatives introduites dans les modèles sont des informations issues du RP99 et des bases de sondage des logements neufs (sauf la vague) :
- géographiques : région administrative, densité d'habitat,…
- socio-démographiques : taille du ménage, sexe, âge de la personne de référence
- relatives au logement : nombre de pièces, habitat collectif ou individuel …
- et la vague pour tenir compte de la saisonnalité de la collecte.

La taille de ces sous-populations ne doivent pas être trop faible afin qu'il y ait suffisamment d'individus répondants qui « parlent » au nom des non-répondants. En effet, comme le poids des répondants augmentent pour compenser les non-répondants, il faut éviter que les non-répondants soient représentés par un trop petit nombre de répondants afin d'assurer une certaine « robustesse » aux résultats obtenus.

c. Mise en œuvre : un traitement de la non-réponse totale « en cascade » pour chaque visite

La non réponse totale au niveau individu peut « s'expliquer » par **les trois raisons** suivantes :

- Un logement échantillonné ne répond pas ; par conséquent, l'enquête ne peut pas se dérouler.
- Un individu dans un logement échantillonné et répondant ne souhaite pas poursuivre l'enquête dès la 1ère visite.
- Un individu répondant aux 2 premières visites ne souhaite plus répondre à la 3ème.

Le traitement de la non-réponse a été effectué en **décomposant le comportement de réponse d'un individu selon ces différents niveaux.**

Pour la 1ère visite, deux corrections successives de la non-réponse totale ont donc été réalisées :

i. la première pour corriger la non-réponse totale des ménages,

ii. et la seconde pour corriger la non-réponse totale des individus au sein de ménages répondants.

Cette modélisation a l'avantage de prendre en compte le fait que les individus sont contactés par l'intermédiaire du ménage et de distinguer la cause de non réponse d'un ménage de celle propre aux caractéristiques de l'individu. De plus, sur le plan opérationnel, il paraît indispensable de modéliser la probabilité de réponse des individus à partir d'informations (a priori socio-démographiques) qui ne sont par définition disponibles que dans le cas où le ménage est répondant. Notons que la modélisation du comportement par une probabilité de réponse ne dépend ici que de l'individu et que les modèles de non-réponse sont construits de manière indépendante entre les deux niveaux d'observations, ménages et individus. Cette hypothèse est simplificatrice car en réalité il est naturel d'imaginer que dans un ménage composé de plusieurs personnes, un phénomène de réponse "en grappe" joue à plein (le comportement d'un individu risque en effet d'impacter celui d'autres personnes au sein du même logement).

De même, **le traitement de la non-réponse pour la troisième visite** a aussi été effectué « en cascade » : pour qu'un individu réponde à la 3ème visite, il faut qu'il ait répondu aux deux entretiens précédents (et que son ménage ait été échantillonné et qu'il ait répondu).

Autrement dit, ***la probabilité pour qu'un individu « i » réponde à la 1ère visite (notée V1) s'obtient en multipliant*** :

- la probabilité initiale de tirage du ménage *« m »* auquel il appartient (notée P_m),
- la probabilité estimée de réponse du ménage sachant qu'il a été sélectionné (P_{rm}),
- la probabilité estimée de réponse de l'individu sachant qu'il appartient à un ménage échantillonné et répondant (P_{i1}).

Ainsi, nous obtenons :

$$P(i \text{ réponde à } \mathrm{V}_1) = P_m \times P_{rm} \times P_{i1}$$

Après correction de la non-réponse en V1, la pondération de l'individu « i » du ménage « m » vaut donc :

$$\omega_{ri1} = \frac{1}{P(i \text{ réponde à } \mathrm{V}_1)} = \frac{1}{P_m \times P_{rm} \times P_{i1}}$$

De même, la probabilité pour que l'individu « i » réponde lors de la 3ème visite (V3) est le produit de :

- sa probabilité estimée de réponse à V1 (calculée ci-dessus*),*
- sa probabilité estimée de réponse à V3 (sachant qu'il appartient à un ménage échantillonné et répondant et que lui-même a déjà répondu à V1 et V2). Notons cette quantité $P_{i3/1}$.

Le poids de l'individu « i » du ménage « m » après correction de la non-réponse en V3 devient donc :

$$\omega_{ri3} = \frac{1}{P(i \text{ réponde à } V_3)} = \omega_{ri1} \times \frac{1}{P_{i3/1}}$$

Par construction, au sein d'un même ménage, les pondérations individuelles sont donc différentes. Par exemple, s'il s'avère que les hommes répondent moins souvent que les femmes, dans les couples où les deux conjoints ont répondu et avaient le même poids d'échantillonnage, la femme aura une pondération plus petite que son conjoint après correction de la non-réponse totale.

II.3.2. Calage au niveau individu

a. Objectif du calage

Le calage permet d'**améliorer la précision de résultats statistiques issus d'enquêtes par sondage** en utilisant de l'information provenant d'autres sources, y compris des enquêtes. En effet, la théorie montre que dès que la taille de l'enquête qui sert de référence est supérieure à celle de l'enquête que l'on cherche à caler, il existe toujours un gain de précision. Ce gain est d'autant plus important que :

- la différence des tailles d'échantillon l'est, c'est à dire que la source externe est considérée comme plus précise que l'enquête que l'on cherche à caler,
- le choix des variables de calage est judicieux, c'est-à-dire que les corrélations entre la variable d'intérêt et les variables auxiliaires sont grandes.

Le calage permet aussi d'**assurer que les résultats obtenus pour l'enquête soient cohérents avec des données connues par ailleurs sur la population** (totaux de variables numériques ou effectifs des différentes modalités de variables catégorielles).

b. Principe général

La théorie du calage consiste à « construire » une forme d'estimateur permettant de faire coïncider le total d'une variable estimé à partir de l'échantillon avec celui provenant d'une autre source. Ce nouvel estimateur s'obtient en modifiant les pondérations de départ[16] des unités sélectionnées dans l'échantillon. Les nouveaux poids sont aussi proches que possible des poids initiaux, au sens d'une distance choisie au préalable, tout en respectant un ensemble de contraintes appelées équations de calage : la somme pondérée des valeurs de chaque variable auxiliaire obtenue sur les individus de l'échantillon correspond au total connu de cette variable auxiliaire.

c. Cadre d'utilisation

Pour que la technique du calage soit valide, il est essentiel d'assurer une cohérence parfaite des variables entre le fichier de l'enquête et la source externe. Autrement dit, ce doivent être **les mêmes variables**, c'est-à-dire les **mêmes concepts**, recueillis à une **même date** et selon **le même protocole**. Dans certains cas, cette règle de cohérence oblige le concepteur d'enquête à redéfinir des variables spécifiquement pour le calage.

[16] Dans notre cas, les pondérations de départ correspondent aux pondérations corrigées de la non-réponse (ϖ_{ri1} et ϖ_{ri3} établies dans la sous partie précédente selon que l'on s'intéresse au calage pour la visite 1 ou pour la visite 3)

d. Mise en œuvre

La source externe est l'enquête Emploi en continu considérée sous forme annuelle (obtenue en moyennant les résultats trimestriels obtenus au cours d'une même année civile - voir partie II.2.). Comme l'enquête Santé s'est déroulée d'octobre 2002 à septembre 2003, nous avons choisi de retenir les résultats de l'enquête Emploi relatifs à l'année 2003 (disponibles à la date du redressement de l'enquête Santé).

Nous avons supposé que les variables utilisées pour le calage (voir tableau 2) peuvent être considérées comme issues du même protocole de collecte dans l'enquête Emploi et dans l'enquête Santé. Nous avons dû cependant redéfinir la **notion d'âge** spécifiquement pour le calage : en effet, pour assurer une cohérence maximale, il faut considérer l'âge à une même date dans l'enquête Santé et dans l'enquête Emploi. Nous avons retenu la notion d'âge au 31 décembre, variable déjà présente dans les fichiers de l'enquête Emploi et non à la date d'anniversaire comme mesuré par l'enquête Santé. Les individus répondants de la 1ère vague (interrogés d'octobre à décembre 2002) ont ainsi été artificiellement « vieillis » au 31/12/02 tandis que ceux des vagues 3 à 5 (interrogés à partir de février 2003) ont vu leur âge porté au 31/12/03. Ce traitement s'avère plus délicat pour les individus de la vague 2 interrogés sur les deux années (de décembre 2002 à février 2003). Au vu des structures d'âge obtenues dans chaque scénario, nous avons finalement décidé de mesurer l'âge au 31/12 de l'année 2002 afin de ne pas « perdre » trop de jeunes enfants, mécaniquement moins nombreux dans l'autre scénario.

Le calage a été réalisé avec la macro **SAS CALMAR**, **au niveau individuel** comme nous l'avons déjà signalé. Les variables retenues pour le calage (disponibles en annexe 4) sont donc principalement constituées d'informations disponibles sur la population française. Cependant, des informations du niveau ménage ont été également introduites dans le calage réalisé au niveau individu afin de restituer certaines données de référence sur la population des ménages. Pour cela, les variables ménages concernées sont individualisées, c'est-à-dire « redescendues » au niveau individuel proportionnellement à la taille du ménage. Plus précisément, pour une caractéristique X mesurée au niveau ménage selon *p* modalités $X_1,...,X_p$, la méthode[17] consiste à utiliser pour le calage les variables individuelles :

$$\frac{\mathrm{I}_{m\in X_1}}{T_m},...,\frac{\mathrm{I}_{m\in X_p}}{T_m}$$

où T_m désigne la taille[18] réelle du ménage *m* des individus et $\mathrm{I}_{m\,\epsilon\,X1}$ vaut 1 si le ménage *m* présente la modalité X_1 et 0 sinon.

Par exemple, si on souhaite se caler sur la répartition des ménages par taille (considérée en trois modalités), on « redescend » l'information ménage au niveau individu de la manière suivante :

Ménage	Taille	Taille1	Taille2	Taille3
A	1	1	0	0
B	2	0	1	0
C	2	0	1	0
D	3	0	0	1

Individu	Ménage	Taille1	Taille2	Taille3
A1	A	1	0	0
B1	B	0	1/2	0
B2	B	0	1/2	0
C1	C	0	1/2	0
C2	C	0	1/2	0
D1	D	0	0	1/3
D2	D	0	0	1/3
D3	D	0	0	1/3

[17] dont la justification théorique est disponible en annexe 1.

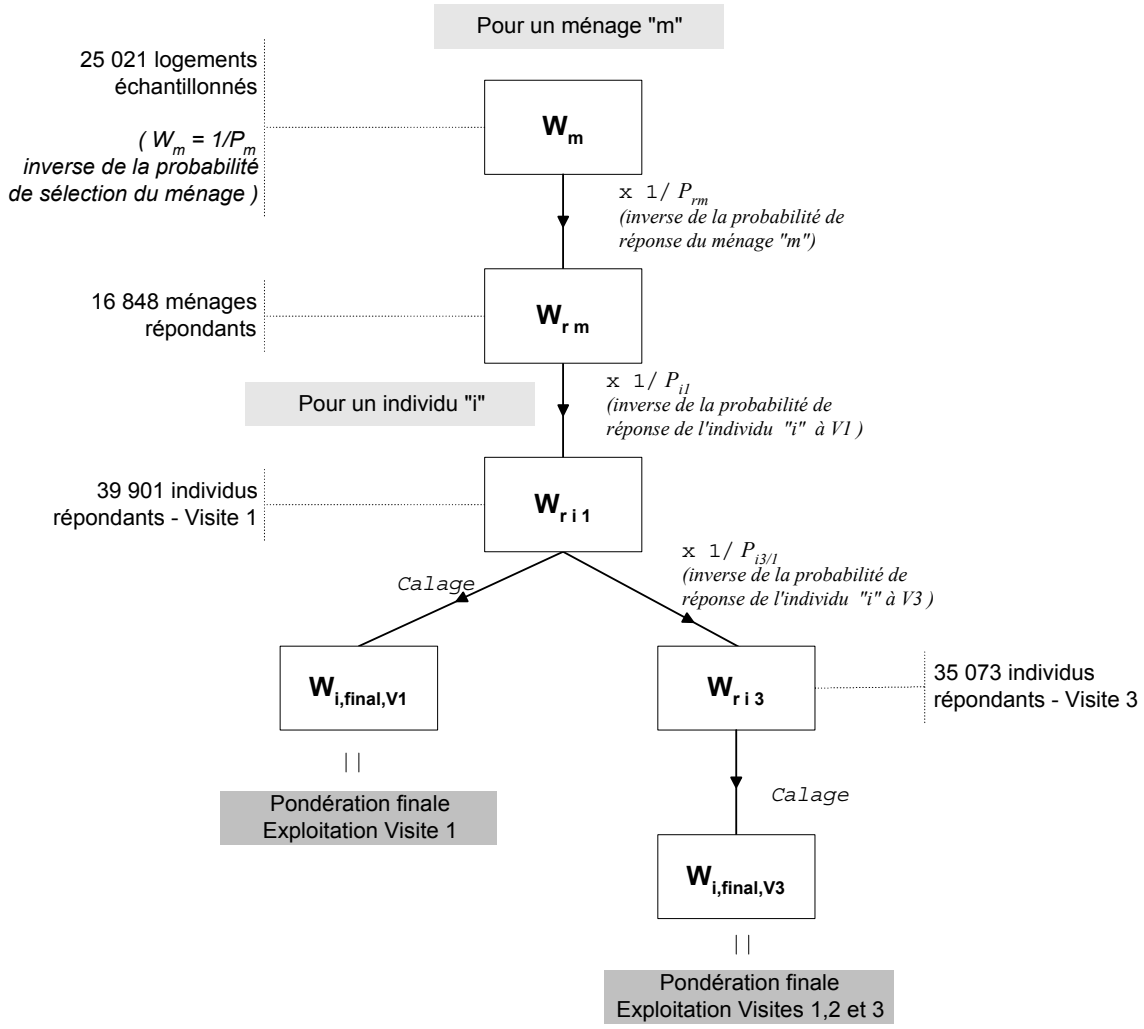
[18] Il s'agit de la vraie taille du ménage et non pas du nombre d'individus répondants. De cette façon, on s'assure que la taille de la population est estimée de manière identique soit à partir des pondérations « individus » soit à partir des pondérations « ménages ». L'annexe n° 1 précise ce point.

Les pondérations issues du calage sont les deux systèmes de pondération proposés :

- **le 1er s'applique aux individus ayant répondu à la 1ère visite et permet d'exploiter les données du 1er questionnaire ;**
- **le 2nd permet d'analyser les informations recueillies au cours des 3 visites (hors auto-questionnaire) et s'applique aux individus ayant répondu jusqu'à la 3ème visite.**

Le schéma 4 résume les différentes étapes du redressement qui ont conduit à ces deux jeux de poids.

Schéma 4 : Les différentes pondérations associées aux données



III. Applications

Préambule : les distributions des différentes pondérations et des rapports de poids à l'issue de chaque traitement sont disponibles en annexe 4.

III.1. Traitement de la 1ère visite

III.1.1. Traitement de la non réponse totale des ménages

La première étape consiste à isoler parmi les logements pour lesquels on n'a aucun questionnaire en retour ceux qui sont considérés comme principaux au moment de la réalisation de l'enquête des autres, ce qui correspond au champ de l'enquête. Cette première étape revient finalement à préciser **la définition d'un ménage répondant**. Pour cette enquête, un ménage non-répondant est un ménage du champ qui n'a pas pu être joint[19] (impossible à joindre, absent de longue durée) ou qui a refusé l'enquête ou pour qui l'enquête était impossible. On a ajouté à cette catégorie de non répondants quelques cas pour lesquels l'exploitation des différents questionnaires était impossible. Au final, 21 655 ménages se sont avérés dans le champ de l'enquête. Parmi ceux-ci, 16 848 sont des ménages répondants.

Définissons **le taux de non réponse** comme le rapport du nombre de ménages non-répondants sur le nombre total de ménages dans le champ de l'enquête. Le taux de non-réponse moyen est de l'ordre 22% ; ce taux, très inégal d'une région à une autre, varie entre 13% pour la région Nord-Pas de Calais et près de 30% pour la région Ile de France.

Dans le but d'affiner la qualité des résultats, nous avons modélisé la non-réponse totale des individus en respectant les contraintes suivantes :

- **appréhender les aspects régionaux** pour intégrer la présence des extensions régionales ;
- **tenir compte de la saisonnalité de la collecte ;**
- **former des sous-populations de taille suffisante**.

Selon les informations présentes dans les bases de sondage, trois populations se dégagent :

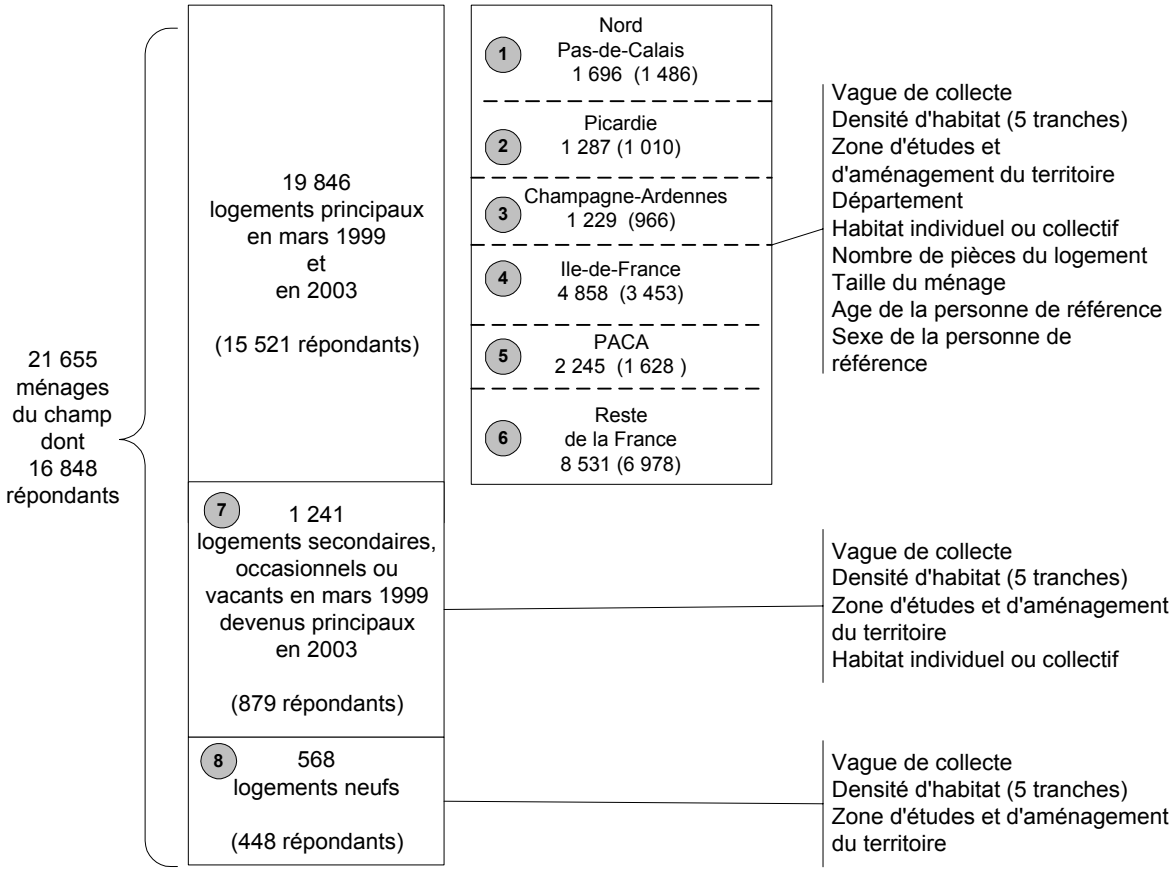
- Les logements recensés comme principaux, pour lesquels de nombreuses informations auxiliaires sont disponibles (localisation géographique, caractéristiques du logement et de ses occupants en mars 1999). Ce groupe consacre plus de 9 cas sur 10 ;
- Les logements recensés non principaux (décrits par des informations géographiques et celles propres au logement) ;
- Les logements construits après le recensement (principalement décrits par leur localisation géographique).

L'importance numérique du **premier groupe** conjugué à la présence des **extensions régionales** nous a naturellement invitées à **fractionner** ce groupe en 6 sous-populations, propres à chaque région à extension et au reste de la France métropolitaine.

C'est ainsi que nous avons abouti aux **huit modélisations présentées sur le schéma 5**. La figure précise également les motifs de non réponse, le nombre de ménages interrogés dont, entre parenthèses, le nombre de répondants.

[19] Les ménages impossibles à joindre étant relativement peu nombreux, tous ont été considérés comme étant dans le champ de l'enquête (bien qu'une partie d'entre eux peuvent très vraisemblablement s'avérer hors champ. Mais aucune information précise n'est connue sur leur statut véritable).

Schéma 5 ***: les 8 modélisations de la non-réponse totale des ménages***



A l'issue de cette étape, le poids d'échantillonnage (W_m) des ménages répondants a été dilaté en moyenne par 1,3 (cf. annexe 4). Leur nouveau poids (W_{rm}) infère à la population des ménages de France métropolitaine et initialise le traitement de la non-réponse individu.

III.1.2. Traitement de la non réponse totale des individus

Une fois le traitement de la non-réponse ménage effectué, nous avons **modélisé le comportement individuel de non-réponse de manière indépendante au niveau ménage** comme nous l'avons expliqué au paragraphe II.3.1.

Le taux de non-réponse individuel à la 1ère visite, voisin de 2%, **s'avère très faible** : seuls 964 individus n'ont pas répondu parmi les 40 865 personnes qui composent la population des ménages répondants (il s'agit des individus dits « éligibles », c'est-à-dire des individus de plus de 18 ans décrits dans le tronc commun des 16 848 ménages répondants).

En conséquence, la simple modélisation consistant à multiplier les pondérations des répondants par le rapport « nombre d'individus éligibles / nombre de répondants » aurait pu suffire.

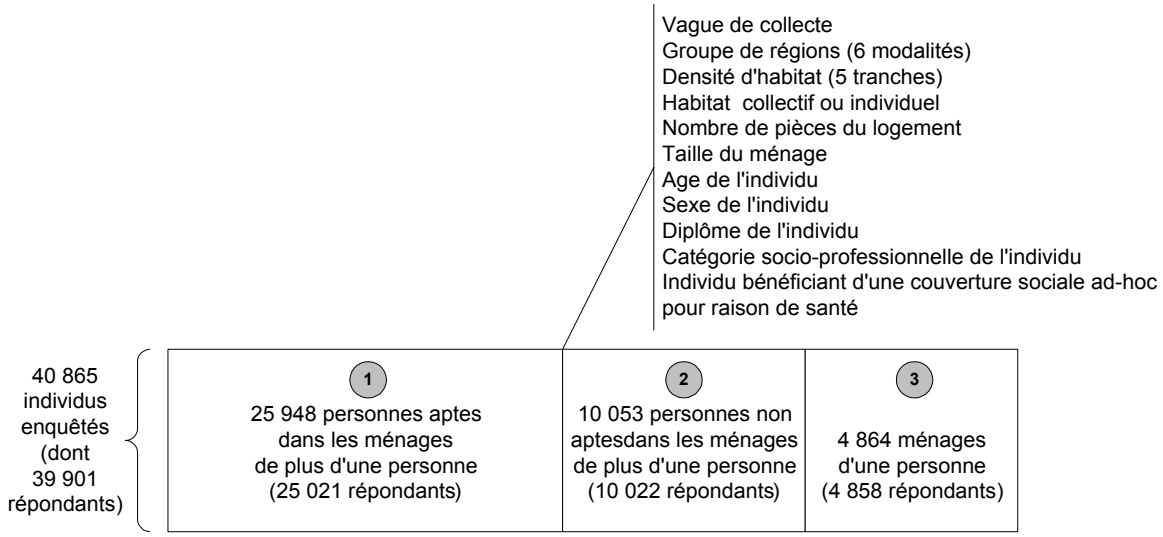
Cependant, le traitement a été affiné selon **l'« aptitude »[20] de l'individu à répondre** à l'enquête et la **taille des ménages** : en effet, si les personnes déclarées aptes répondent elles-mêmes au questionnaire, c'est un proche qui répond pour les personnes non aptes, à moins que ces personnes ne vivent seules. Les raisons d'abandon ont donc été explorées séparément pour les individus aptes, les non aptes et les ménages d'une personne afin de distinguer les motivations propres aux individus de celles de leurs proches, le cas échéant.

[20] Un individu est dit « apte » s'il a plus de 18 ans, ou s'il s'agit d'un mineur qui occupe un emploi et/ou est personne de référence ou conjoint de la personne de référence.

Notons que les motifs de non-réponse individuels explorent ici les caractéristiques des individus telles qu'elles sont décrites dans l'enquête (dans le tronc commun du questionnaire ménage). La modélisation bénéficie donc d'information fraîche et propre aux individus enquêtés. En particulier, elle tient compte de la présence d'une couverture sociale particulière[21] accordée pour raison de santé. Cette dimension est d'autant plus précieuse à capter à ce stade qu'elle ne peut l'être ultérieurement par le calage.

Les **trois modèles** envisagés sont représentés sur le schéma 6.

Schéma 6 ***: les 3 modèles de non-réponse totale des individus pour la visite 1***



La correction introduite à cette étape bouleverse en moyenne relativement peu les poids individuels (cf. annexe 4). Toutefois, à l'issue de cette étape, les membres d'un même ménage ne possèdent plus forcément la même pondération. Le nouveau poids individuel (W_{ri1}) permet une inférence sur la population de France métropolitaine dont la « qualité » est améliorée par le calage réalisé à l'étape suivante.

III.1.3. Calage sur l'enquête emploi

A l'issue de l'étape précédente, nous avons redressé par calage les données relatives à l'ensemble des individus répondants à la 1ère visite. **Le calage s'effectue** donc **sur 39 901 individus munis des pondérations** W_{ri1} ci-dessus.

Comme nous l'avons expliqué précédemment, un traitement **préalable** au calage a été effectué sur les données afin d'**harmoniser le concept d'âge** avec celui de l'enquête Emploi, source des marges de référence, et pour éviter de « perdre » trop de jeunes enfants. C'est ainsi que l'âge des répondants a été porté au 31.12.02 pour les personnes interrogées aux vagues 1 et 2, et au 31.12.03 pour les autres.

[21] Les différentes dispositions sociales prises en compte repèrent les personnes qui :
- relèvent de la couverture maladie universelle ;
- et/ou bénéficient d'une exonération du ticket modérateur de l'assurance maladie pour raison de santé ou de handicap ;
- et/ou exerçent une activité professionnelle à temps partielle, voire n'en exerçent aucune, pour raison de santé.
- et/ou sont titulaires d'une pension accordée pour raison de santé.

Les variables de calage sont présentées dans le tableau suivant :

Tableau 2 : les variables de calage

Variables		Nombre de modalités
Niveau Individu	Répartition par croisement sexe et âge	20
	Répartition par groupes de région	6
	Répartition par strate de gestion	5
	Répartition par niveau de diplôme	6
Niveau Ménage	Répartition par groupes de région	6
	Répartition par vague (équi-répartition au 1/5ème)	5
	Répartition d'individus par taille de ménage	3
	Répartition d'individus par type de ménage	5

La méthode retenue utilise la fonction de distance «logit » (M=3) qui présente l'avantage de borner les rapports des poids avant et après calage. Dans le cas présent, les rapports de poids (avant et après calage) varient entre 0,79 et 1,3 (cf.annexe 4).

Nous avons **validé** les pondérations ($W_{i,\ final,\ V1}$) qui résultent de cette étape après avoir vérifié les effectifs extrapolés obtenus.

Cette pondération est celle à employer pour exploiter les réponses recueillies lors de la 1ère visite auprès des 39 901 individus répondants à cette partie de l'enquête.

III.2. Traitement de la 3ème visite

III.2.1. Traitement de la non réponse totale des individus

L'opération entreprise ici traduit le fait que pour qu'un individu réponde à la 3ème visite, il faut que son ménage ait répondu à la 1ère visite et que lui-même ait répondu aux deux premières visites.

Le traitement s'effectue donc sur les 39 901 individus ayant répondu à la 1ère visite et **consiste,** comme pour la première visite, **à successivement** :

- Rechercher les facteurs explicatifs de la non-réponse,
- Former les groupes homogènes de réponse,
- Affecter aux 35 073 individus répondants à la 3ème visite le poids W_{ri3} qui correspond au produit de leur probabilité de réponse estimée par leur poids avant calage $W_{r\,i1}$.

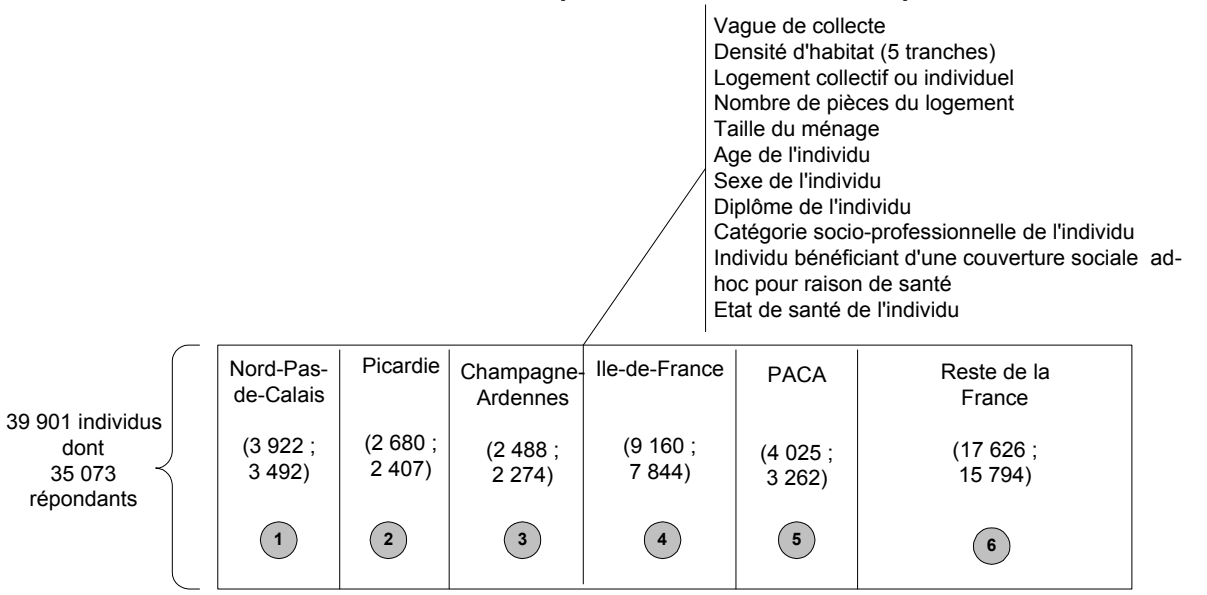
Nous avons retenu **les mêmes impératifs** que ceux déjà présents pour modéliser le comportement de non-réponse lors de la 1ère visite :

- Appréhender les aspects régionaux pour intégrer la présence des extensions régionales
- Tenir compte de la saisonnalité de la collecte
- Former des sous-populations de taille suffisante.

Si la méthodologie employée obéit aux mêmes principes que celle développée pour la 1ère visite, soulignons ici **la plus grande richesse de l'information auxiliaire disponible** : les motifs de non-réponse explorent aussi parmi les réponses obtenues au cours de la 1ère visite en particulier **l'état de santé**[22] des enquêtés est pris en compte. Cette dimension est précieuse à capter à ce stade car elle est susceptible d'influer sur le comportement de réponse et ne peut être appréhendée ultérieurement par le calage.

Les **six modélisations** envisagées sont représentées sur le schéma 7. Les chiffres entre parenthèses précisent respectivement le nombre d'individus interrogés et répondants.

Schéma 7*: les 6 modèles de non-réponse totale des individus pour la visite 3*



Cette étape attribue une pondération W_{ri3} à chacun des 35 073 individus répondants. Elle permet une inférence sur la population de France métropolitaine dont la qualité est améliorée par le calage final réalisé à l'étape suivante.

[22] L'état de santé est mesuré dans le questionnaire individuel de la 1ère visite. Il s'agit de la réponse à la question « Etes-vous en bonne santé ? ».

III.2.2. Calage sur l'enquête emploi

Le calage s'effectue sur l'ensemble des **35 073 individus répondants à la 3ème visite** munis des pondérations W_{ri3} calculées à l'étape précédente .

Les **mêmes principes généraux** que ceux exposés au paragraphe III-1-B prévalent ici aussi (mesure de l'âge, choix de la méthode,etc.). En outre, **les variables de calage sont identiques** à celles utilisées pour le calage relatif à la première visite.

Les rapports de poids varient entre 0,78 et 1,32 et s'avèrent très proches de ceux obtenus pour la 1ère visite.

Nous avons **validé** les pondérations ($W_{i,\ final,\ V3}$) qui résultent de cette étape après avoir vérifié les effectifs extrapolés obtenus.

Cette pondération est celle à employer pour exploiter l'ensemble des réponses -recueillies lors des 3 visites- auprès des 35 073 individus répondants à l'ensemble de l'enquête (hors carnets de soin).

IV. Bibliographie

N. CARON (1996), *« Les principales techniques de correction de la non-réponse et les modèles associés* », Document de travail n° 9604, INSEE.

J.-C. DEVILLE (1999) : « *Calibration with control totals coming from different sources »,* Workshop du colloque de la Société Statistique du Canada, Sherbrooke, Canada.

S. ROUSSEAU, F. TARDIEU (2004), *« La macro SAS CUBE d'échantillonnage équilibré* », Document de travail n°0402, INSEE.

O. SAUTORY (1993), « *La macro CALMAR - Redressement d'un échantillon par calage sur marges* », Document de travail n° 9310, INSEE.

L. WILMS (2000*) « L'échantillon-Maître 1999 et application au tirage des unités primaires par la macro CUBE »,* INSEE Méthodes n°100, INSEE.

Sur le site ***www.insee.fr***

N. CARON (1996), *« Les principales techniques de correction de la non-réponse et les modèles associés* », Document de travail n° 9604, INSEE.

S. ROUSSEAU, F. TARDIEU (2004), *« La macro SAS CUBE d'échantillonnage équilibré* », Document de travail n°0402, INSEE.

O. SAUTORY (1993), « *La macro CALMAR - Redressement d'un échantillon par calage sur marges* », Document de travail n° 9310, INSEE.

Sur l'intranet INSEE,

Questionnaires et protocole de l'enquête Santé 2002-2003.

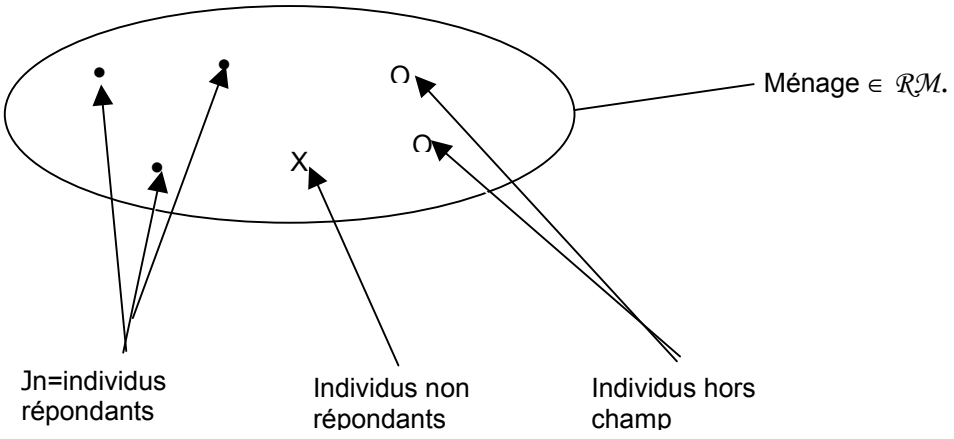
Principales publications à partir de l'enquête Santé 1990-1991.

V. Annexes

Annexe 1 *: Comment reconstituer des poids « ménages » à partir de pondérations « individus » ?*

(note rédigée par Marc Christine en 1999)

On dispose d'un échantillon de ménages répondants, noté $\mathcal{RM}$.. Au sein de chaque ménage de cet échantillon, les individus éligibles sont ceux de 15 ans ou plus. Mais tous ne répondent pas. Soit R_m l'ensemble des individus répondants du ménage m.



Par une procédure ad hoc[23], on a affecté un poids initial α_i à tout individu répondant i.

On cherche maintenant à faire du calage sur marge pour achever de redresser l'échantillon des individus répondants. Outre le calage sur des totaux exogènes fournissant des ventilations de la population totale des individus de 15 ans ou plus selon différents critères, peut-on caler aussi un des répartitions de ménages ?

Il faut, pour cela, pouvoir reconstituer un poids ménage et l'affecter à chaque ménage répondant. Une manière possible est la suivante.

Soit ω_m le poids cherché du ménage m. On impose la contrainte que la population des individus éligibles, estimée soit à partir des pondérations « individus », soit à partir des pondérations « ménages », soit identique (et égale) à la vraie population de référence.

Ceci conduit à l'égalité : $$\sum_{m\in\Re M}\left(\sum_{i\in\Re m}\alpha_i\right)=\sum_{m\in\Re M}\omega_m \mathrm{T}_m$$

où T_m est la vraie taille du ménage m (supposée connue), en termes de nombre de personnes éligibles (i.e., de 15 ans ou plus)

Cette égalité est assurée dès que : $\forall m\in\Re M : \omega_m \mathrm{T}_m=\sum_{i\in\Re m}\alpha_i$

soit : $$\boxed{\omega_m=\frac{1}{\mathrm{T}_m}\sum_{i\in\Re m}\alpha_i}$$

[23] A partir des poids de sondage initiaux, on a fait une correction de la non-réponse, en recherchant les facteurs explicatifs, en deux temps : correction de la non-réponse d'un ménage dans son intégralité, puis de celle d'un individu appartenant à un ménage répondant.

Utilisation de ces poids pour les équations de calage relatives aux ménages

Supposons que l'on cherche à faire un calage sur la répartition des ménages selon une certaine caractéristique. Les totaux relatifs à la ventilation de la population des ménages selon les différentes modalités de la variable retenue sont notés : $z_{1,} z_2,\ldots,z_K$.

Exemple : Variable « type de ménages »

Modalités : 1= personne seule

2= couple seul, sans enfant

3= couple avec enfants

etc …

On notera, pour chaque ménage m, $1_k(m)$ (pour k = 1,….,K) la variable qui désigne l'appartenance du ménage à la modalité k de la variable, valant 1 si le ménage correspond à cette modalité k et 0 sinon.

[Attention : il faut à priori supprimer une modalité si l'on veut éviter les colinéarités entre les différentes équations de calage]

→Les équations de calage pour cette variable s'écriront alors :

$$\sum_{m\in\Re M} \omega_m 1_k(m) = Z_k \ldots \forall k \in \{1,\ldots;K-1\}$$

soit :

$$\sum_{m\in\Re M} \frac{1}{Tm}\left(\sum_{i\in\Re m}\alpha_i\right)1_k(m) = Z_k$$

soit encore : $$\boxed{\sum_i \alpha i \frac{1_k(m_i)}{T_{m_i}} = Z_k \cdots\cdots\cdots \forall k \in \{1,\ldots K-1\}}$$

→ Dans cette formule :

-la somme est étendue à tous les individus répondants appartenant à tous les ménages répondants.

-la notation m_i représente le ménage auquel appartient l'individu i.

Ainsi cette équation est équivalente à une équation de calage au niveau « individu » dans laquelle le total exogène est Z_k (nombre réel de ménages correspondant à la modalité k) et la variable observée au niveau individuel est :

- $\ldots\frac{1_k(m_i)}{T_{m_i}} = 0$ si le ménage auquel appartient l'individu i dans l'échantillon ne correspond pas à la modalité k
- $\ldots\frac{1_k(m_i)}{T_{m_i}} = \frac{1}{T_{m_i}}$ sinon, où T_{m_i} est la taille du ménage auquel appartient l'individu i.

Dans la pratique, on peut donc traiter ces conditions de calage sur des caractéristiques « ménages » comme des conditions de calage sur des caractéristiques « individus », en utilisant les variables individuelles $\ldots\frac{1_k(m_i)}{T_{m_i}}\ldots$ et l'on n'a pas besoin de recalculer explicitement les poids « ménages ».

Attention

Il est même recommandé de ne pas utiliser ces poids « ménages » pour estimer des nombres d'individus ayant une certaine caractéristique car on risque alors une incohérence avec les estimations directes à partir des poids individuels.

Une fois calculés les poids α_i* satisfaisant à l'ensemble des équations de calage, on peut recalculer des poids ménages par les formules précédentes, soit $\omega_m = \frac{1}{T_m} \sum_{i \in \Re m} \alpha_i$

Mais ceux-ci ne doivent être utilisés que pour des estimations relatives aux ménages.

L'intérêt de la procédure est qu'elle assure que la répartition des ménages (et pas seulement des individus) est correcte au vu des caractéristiques retenues

Nota.

Un autre problème, quand on travaille sur des pondérations individuelles est que tous les individus (répondants) d'un même ménage n'ont pas le même poids, et en particulier, que les deux membres d'un couple n'ont pas le même poids.

Annexe 2 *: Fiche récapitulative du plan de sondage*

INSEE
DIRECTION GÉNÉRALE
INSTITUT NATIONAL
DE LA STATISTIQUE
ET DES ÉTUDES
ÉCONOMIQUES

Direction des Statistiques Démographiques et Sociales
Unité "Méthodes Statistiques"
Division "Echantillonnage et Traitement Statistique des Données"
Timbre F410

Paris, le 22 juillet 2002
N° 048/F410

FICHE RECAPITULATIVE DU PLAN DE SONDAGE ENQUETE SANTE (2002)

. Responsable d'enquête : *Madame DUMONTIER*
Monsieur LANOE

. Date du tirage :juillet 2002

. Intitulé de l'enquête : Enquête Santé (2002)

. Nom symbolique de l'enquête : ***SAN02R***

I - EFFECTIFS TIRES - BILAN :

. Nombre de logements tirés par catégorie de logement, France entière.

- Résidences principales au RP99 :	*21511*
- Résidences secondaires au RP99	*1123*
- Résidences occasionnelles au RP99	*118*
- Résidences vacantes au RP99:	*1619*
- Logements achevés après le RP99:	*650*
TOTAL	*25021*

*. Nombre de logements tirés par **région de gestion** ×strate, toutes catégories de logements confondues :*

	Rural(0)	- 20 000 (1)	20 000 à 100 000 (2)	+ 100 000 (3)	Agglo paris (4)	Total
11. Ile de France	49	79	17	0	4243	4388
21. Champagne-Ardenne	677	336	384	403	232	2032
22 Picardie	586	377	391	138	0	1492
23. Haute-Normandie	158	76	83	188	427	932
24. Centre	173	217	154	171	459	1174
25 Basse-Normandie	249	75	70	59	0	453
26. Bourgogne	243	70	142	76	0	531
31. Nord-Pas de Calais	244	240	286	1232	0	2002
41. Lorraine	168	129	122	229	0	648
42. Alsace	35	145	67	202	0	449
43 Franche-Comté	169	99	30	76	0	374
52. Pays de la Loire	318	170	105	344	0	937
53. Bretagne	254	302	158	191	0	905
54. Poitou-Charentes	223	155	50	113	0	541
72. Aquitaine	302	228	138	358	0	1026
73. Midi-Pyrénées	254	119	123	247	0	743
74. Limousin	153	42	21	58	0	274
82. Rhône-Alpes	363	230	318	777	0	1688
83. Auvergne	181	40	89	84	0	394
91. Languedoc-Roussillon	179	329	136	203	0	847
93. PACA-Corse	411	464	376	1940	0	3191
Total	5389	3922	3260	7089	5361	25021

RAPPEL :

***Strate 0** : Communes rurales au RP99*

***Strate 1** : Communes appartenant à des Unités Urbaines ayant moins de 20 000 habitants au RP99.*

***Strate 2** : Communes appartenant à des Unités Urbaines ayant entre 20 000 et 100 000 habitants au RP99.*

***Strate 3** : Communes appartenant à des Unités Urbaines de plus de 100 000 habitants au RP99, sauf l'Unité Urbaine de PARIS.*

***Strate 4** : Unité Urbaine de PARIS.*

II - PLAN DE SONDAGE :

Sur les 17 régions sans extension :
Echantillon tiré à partir de la base de Sondage Echantillon-Maître 99 et de la Base de Sondage des Logements Neufs.
Sur les 5 régions avec extension
Echantillon tiré à partir de la base de Sondage EMEX 99 (base EM et base d'extension RP) et de la Base de Sondage des Logements Neufs EMEX(base BSLN et base d'extension des logements neufs).

Les 5 régions ayant bénéficié d'une extension sont :

Ile-de-France, Picardie, Nord Pas-de-Calais, Champagne-Ardenne, Provence-Alpes-Côte d'Azur.

A) Taux de sondage :

Logements RP

Le taux de sondage d'un logement, sur une région sans extension, est donné par la formule :

$$f \times c$$

Le taux de sondage d'un logement de strate 2,3 ou 4, sur une région avec extension, est donné par la formule :

$$f \times c$$

Le taux de sondage d'un logement de strate 0, sur une région avec extension, est donné par la formule :

$$\frac{f \times c}{RED0}$$

Le taux de sondage d'un logement de strate 1, sur une région avec extension, est donné par la formule :

$$\frac{f \times c}{RED1}$$

avec

f *, taux de sondage général brut (sans redressement)*
c *: coefficient de sous représentation de catégorie de logements*
Résidences principales : c=1
Résidences vacantes (hors strate 0) : c=1
Résidences vacantes (strate 0) : c=0.5
Résidences secondaires et occasionnelles : c=0.5

RED0 *et* ***RED1*** *sont des coefficients de redressements relatifs aux tirages régionaux des UP-EM et UP-EMEX dans les strates de gestion 0 et 1.*

Logements neufs

Le taux de sondage d'un logement, sur une région sans extension, est donné par la formule :

$$f_{neuf}$$

Le taux de sondage d'un logement de strate 2,3 ou 4, sur une région avec extension, est donné par la formule :

$$f_{neuf}$$

Le taux de sondage d'un logement de strate 0, sur une région avec extension, est donné par la formule :

$$\frac{f_{neuf}}{RED0}$$

Le taux de sondage d'un logement de strate 1, sur une région avec extension, est donné par la formule :

$$\frac{f_{neuf}}{RED1}$$

avec

f_{neuf} *, taux de sondage général brut (sans redressement)*
RED0 *et* ***RED1,*** *coefficients de redressements relatifs aux tirages régionaux des UP-EM et UP-EMEX dans les strates de gestion 0 et 1.*

tableau de reconstitution du taux de sondage d'un logement

Région	***f***	f_{neuf}	***RED0***	***RED1***
Régions sans extension	1/1550	1/1240	1	1
11. Ile de France	1/855	1/926	1	12/13
21. Champagne-Ardenne	1/414	1/399	15/17	6/5
22. Picardie	1/526	1/536	15/16	1
31. Nord-Pas de Calais	1/820	1/843	1	1
93. PACA	1/792	1/833	12/15	12/13

***B) <u>Autres paramètres</u>** :*

• Une seule phase de tirage.

• 5 vagues (cf annexe pour la répartition par région de gestion des vagues).

• Aucune enquête précédente n'a été réintégrée.

• Les résidences vacantes (hors strate 0) au RP99 ont été assimilées à des résidences principales.

• Les résidences occasionnelles au RP99 ont été assimilées à des résidences secondaires.

• Tris effectués lors du tirage des logements RP dans les groupes de commune selon :

- Résidences principales recensées : CATL, DEP, COM, CILILFIL, NPER AGEM.

- Résidences secondaires, occasionnelles : CATL, DEP, COM, ACHI, NLOG.

• Logements achevés après le RP99 : :

- *Champ de l'enquête : logements déclarés achevés entre le 9 mars 1999 et le 31 Décembre 2001.*
- *Périodes disponibles au moment du tirage : logements achevés entre le 9 mars 1999 et le 31 Décembre 2001.*
- *Estimation (brute) du nombre de logements faisant partie du champ de l'enquête, France entière :*
- *Pas d'introduction de sous-périodes de représentativité lors du tirage*

<u>IMPORTANT</u> :

Vous disposez maintenant d'un fichier plat sur le serveur ftp contenant, pour chaque logement tiré, selon le dessin habituel :

- l'identifiant (plus une clef de contrôle).

- le poids de sondage (coefficient d'extrapolation)

- les variables RP99 ou BSLN99.

Le nom de ce fichier est : ***SAN02R.FICFLCLE***
disponible par le serveur ftp, site ftpappli sous EME/SAN02R

On rappelle que la procédure de saisie ne doit concerner que le cadre "identifiant informatique" de la fiche de repérage, ainsi que les codes de gestion.

Les logements issus de la base EM 99 ont reçu le numéro de sous-échantillon ***10,*** *ceux issus de la base d'extension, le numéro* ***11****.*

Pour améliorer la qualité des estimations, il faut procéder à des redressements adaptés à cette enquête (voir CALMAR).

L.WILMS

DESTINATAIRES :

- Mmes. DUMONTIER, THIESSET
- M. LANOE
- MM. VERGER, CHRISTINE, ARDILLY, WILMS, BOURDALLE
- MM. SEYS, Mme GUILLEMOT
- MM. Les Chefs des Services Statistiques (Toutes DR)
- Mme. M-A MERCIER, M. D. BLAIZEAU

ANNEXE : Répartition des FA par vague et strate de gestion

RGES	VAGUE					
Frequency	01	02	03	04	05	Total
11	880	859	873	876	900	4388
21	403	417	428	388	396	2032
22	303	323	262	316	288	1492
23	180	170	196	190	196	932
24	232	223	239	229	251	1174
25	96	56	91	107	103	453
26	123	110	129	91	78	531
31	411	406	419	391	375	2002
41	136	131	111	120	150	648
42	101	100	85	81	82	449
43	74	80	54	69	97	374
52	213	202	173	171	178	937
53	151	174	191	211	178	905
54	92	103	102	117	127	541
72	192	216	209	205	204	1026
73	135	164	160	160	124	743
74	60	57	52	63	42	274
82	313	334	359	349	333	1688
83	86	67	81	75	85	394
91	170	167	156	183	171	847
93	649	639	624	634	645	3191
Total	5000	4998	4994	5026	5003	25021

Annexe 3 *: Liste des variables et des modalités retenues dans le calage avec les marges de référence associées*

- **Les variables de calage** sont les suivantes :

Variables		Nombre de modalités
Niveau Individu	Répartition par croisement sexe et âge	20
	Répartition par groupes de région	6
	Répartition par strate de gestion	5
	Répartition par niveau de diplôme	6
Niveau Ménage	Répartition par groupes de région	6
	Répartition par vague (équi-répartition au 1/5ème)	5
	Répartition d'individus par taille de ménage	3
	Répartition d'individus par type de ménage	5

- **Le nombre total d'individus** sur lequel le fichier est calé est 58 438 795.

- **Le nombre total de ménages** est fixé à 24 737 732.

- **Les modalités de chacune de ces variables ainsi que les effectifs** sont précisés ci-dessous :

  - **Variable croisement sexe et âge (niveau individu)**

Nom	*Libellé*	*Marge*
ageh_05	Homme de moins de 5 ans	2 123 244,50
ageh_10	Homme entre 6 et 10 ans	1 789 990,00
ageh_15	Homme entre 11 et 15 ans	1 967 791,69
ageh_1	Homme entre 16 et 25 ans	3 751 134,17
ageh_2	Homme entre 26 et 35 ans	3 946 392,95
ageh_3	Homme entre 36 et 45 ans	4 201 897,86
ageh_4	Homme entre 46 et 55 ans	4 043 383,23
ageh_5	Homme entre 56 et 65 ans	2 834 723,04
ageh_6	Homme entre 66 et 75 ans	2 257 930,06
ageh_7	Homme de plus de 76 ans	1 478 410,26
agef_05	Femme de moins de 5 ans	2 000 209,50
agef_10	Femme entre 6 et 10 ans	1 714 371,96
agef_15	Femme entre 11 et 15 ans	1 860 972,78
agef_1	Femme entre 16 et 25 ans	3 715 762,55
agef_2	Femme entre 26 et 35 ans	3 993 798,34
agef_3	Femme entre 36 et 45 ans	4 313 357,28
agef_4	Femme entre 46 et 55 ans	4 197 027,17
agef_5	Femme entre 56 et 65 ans	3 006 298,30
agef_6	Femme entre 66 et 75 ans	2 756 464,79
agef_7	Femme de plus de 76 ans	2 485 634,34

- **Variable groupes de région (niveau individu)**

Nom	*Libellé*	*Marge*
Rega11	Région parisienne	10 671 691,42
Rega21	Région Champagne Ardennes	1 370 833,05
Rega22	Région Picardie	1 886 544,49
Rega31	Région Nord/ Pas-de-Calais	4 159 840,60
Rega93	Région PACA	4 211 077,48
Regaautre	Les autres régions	36 138 807,72

- **Variable strate de gestion (niveau individu)**

Nom	*Libellé*	*Marge*
Strate0	Rural	14 800 106,75
Strate1	Unité Urbaine de moins de 20 000 habitants	10 306 341,82
Strate2	Unité urbaine entre 20 000 et 100 000 habitants	7 568 183,58
Strate3	Unité urbaine de plus de 100 000 habitants	16 299 308,72
Strate4	Unité urbaine de Paris	9 464 853,89

- **Variable niveau de diplôme (niveau individu - pour 16 ans et plus - le nombre d'individus de 16 ans et plus est 46 982 214)**

Nom	*Libellé*	*Marge*
Dipl7_1	Diplôme supérieur	4 553 615,29
Dipl7_3	Bac + 2 ans	4 091 307,19
Dipl7_4	Bac ou brevet professionle	6 767 388,20
Dipl7_5	CAP, BEP	10 435 099,29
Dipl7_6	BEPC seul	4 333 487,38
Dipl7_7	Aucun diplôme ou CEP (ou non déclaré)	16 801 316,96

- **Variable taille de ménage (niveau ménage)**

Nom	*Libellé*	*Marge*
Ind_14	Ménage de moins de 5 personnes	22 934 343,62
Ind_5	Ménage de 5 personnes	1 332 068,75
Ind_678	Ménage de plus de 5 personnes	471 319,79

- **Variable type de ménage (niveau ménage)**

Nom	*Libellé*	*Marge*
Typmen51	Ménage d'une personne	7 410 962,35
Typmen52	Famille monoparentale	1 916 283,44
Typmen53	Couple sans enfant	6 823 218,32
Typmen54	Couple avec enfants	7 883 919,67
Typmen55	Ménage complexe de plus d'une personne	703 348,36

- **Variable groupes de région (niveau ménage)**

Nom	*Libellé*	*Marge*
Regam11	Région parisienne	4 550 710,11
Regam21	Région Champagne Ardennes	580 991,62
Regam22	Région Picardie	748 578,27
Regam31	Région Nord/ Pas-de-Calais	1 626 285,76
Regam93	Région PACA	1 855 145,33
Regamautre	Les autres régions	15 376 021,07

- **Variable vague (niveau ménage)**

Nom	*Libellé*	*Marge*
vag1m	Vague n°1	4 947 546,43
vag2m	Vague n°2	4 947 546,43
vag3m	Vague n°3	4 947 546,43
vag4m	Vague n°4	4 947 546,43
vag5m	Vague n°5	4 947 546,43

Annexe 4 : *Distribution des poids à l'issue de chaque traitement*

4.1. Distribution des pondérations initiales d'échantillonnage des logements

Moments

N	25021	Sum Weights	25021
Mean	1163.8286	Sum Observations	29120155.3
Std Deviation	541.922899	Variance	293680.429
Skewness	1.32594342	Kurtosis	3.01477029
Uncorrected SS	4.12388E10	Corrected SS	7347884330
Coeff Variation	46.5638068	Std Error Mean	3.42598275

Basic Statistical Measures

Location		Variability	
Mean	1163.829	Std Deviation	541.92290
Median	857.949	Variance	293680
Mode	1550.292	Range	2748
		Interquartile Range	730.37070

Tests for Location: Mu0=0

Test	-Statistic-		-----p Value------	
Student's t	t	339.7065	Pr > |t|	<.0001
Sign	M	12510.5	Pr >= |M|	<.0001
Signed Rank	S	1.5652E8	Pr >= |S|	<.0001

Quantiles (Definition 5)

Quantile	Estimate
100% Max	3100.583
99%	3100.583
95%	1550.292
90%	1550.292
75% Q3	1550.292
50% Median	857.949
25% Q1	819.921
10%	526.551
5%	493.642
1%	365.865
0% Min	352.106

Extreme Observations

------Lowest-----		-----Highest-----	
Value	Obs	Value	Obs
352.106	24490	3100.58	21857
352.106	24489	3100.58	21861
352.106	24488	3100.58	21863
352.106	24487	3100.58	22291
352.106	24486	3100.58	22321

4.2. Distribution des rapports de poids après correction de la non-réponse ménages

Moments

N	16848	Sum Weights	16848
Mean	1.28530417	Sum Observations	21654.8046
Std Deviation	0.15269813	Variance	0.02331672
Skewness	1.11416996	Kurtosis	1.93767602
Uncorrected SS	28225.8273	Corrected SS	392.816744
Coeff Variation	11.8803105	Std Error Mean	0.00117641

Basic Statistical Measures

Location		Variability	
Mean	1.285304	Std Deviation	0.15270
Median	1.257229	Variance	0.02332
Mode	1.108402	Range	0.92143
		Interquartile Range	0.20180

Tests for Location: Mu0=0

Test	-Statistic-		-----p Value------	
Student's t	t	1092.563	Pr > |t|	<.0001
Sign	M	8424	Pr >= |M|	<.0001
Signed Rank	S	70967988	Pr >= |S|	<.0001

Quantiles (Definition 5)

Quantile	Estimate
100% Max	1.96309
99%	1.64393
95%	1.55618
90%	1.50989
75% Q3	1.37703
50% Median	1.25723
25% Q1	1.17523
10%	1.11458
5%	1.08696
1%	1.06406
0% Min	1.04167

Extreme Observations

------Lowest-----		-----Highest-----	
Value	Obs	Value	Obs
1.04167	8236	1.96309	4645
1.04167	8204	1.96309	4681
1.04167	8196	1.96309	4682
1.04167	8147	1.96309	4685
1.04167	8073	1.96309	4725

4.3. Distribution des rapports de poids après / avant La correction de la non-réponse individus V1

Moments

N	39901	Sum Weights	39901
Mean	1.0241598	Sum Observations	40865
Std Deviation	0.03265557	Variance	0.00106639
Skewness	2.99596307	Kurtosis	16.4539428
Uncorrected SS	41894.8389	Corrected SS	42.5488196
Coeff Variation	3.18852324	Std Error Mean	0.00016348

Basic Statistical Measures

Location		Variability	
Mean	1.024160	Std Deviation	0.03266
Median	1.013889	Variance	0.00107
Mode	1.000536	Range	0.31633
		Interquartile Range	0.03130

Tests for Location: Mu0=0

Test	-Statistic-		-----p Value------	
Student's t	t	6264.729	Pr > |t|	<.0001
Sign	M	19950.5	Pr >= |M|	<.0001
Signed Rank	S	3.9803E8	Pr >= |S|	<.0001

Quantiles (Definition 5)

Quantile	Estimate
100% Max	1.31633
99%	1.12249
95%	1.09263
90%	1.06542
75% Q3	1.03297
50% Median	1.01389
25% Q1	1.00166
10%	1.00054
5%	1.00000
1%	1.00000
0% Min	1.00000

4.4. Distribution des rapports de poids
Poids après CALAGE V1/ poids après non-réponse V1

Moments

N	39901	Sum Weights	39901
Mean	1.02290334	Sum Observations	40814.8661
Std Deviation	0.17774105	Variance	0.03159188
Skewness	0.15372759	Kurtosis	-1.4480244
Uncorrected SS	43010.1789	Corrected SS	1260.51605
Coeff Variation	17.3761336	Std Error Mean	0.00088981

Basic Statistical Measures

Location		Variability	
Mean	1.022903	Std Deviation	0.17774
Median	1.006898	Variance	0.03159
Mode	1.141328	Range	0.51000
		Interquartile Range	0.34765

Tests for Location: Mu0=0

Test	-Statistic-		-----p Value------	
Student's t	t	1149.579	Pr > |t|	<.0001
Sign	M	19950.5	Pr >= |M|	<.0001
Signed Rank	S	3.9803E8	Pr >= |S|	<.0001

Tests for Normality

Test	--Statistic---		-----p Value------	
Kolmogorov-Smirnov	D	0.103919	Pr > D	<0.0100
Cramer-von Mises	W-Sq	166.2412	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	1173.464	Pr > A-Sq	<0.0050

Quantiles (Definition 5)

Quantile	Estimate
100% Max	1.299997
99%	1.299303
95%	1.293112
90%	1.277356
75% Q3	1.194960
50% Median	1.006898
25% Q1	0.847314
10%	0.794462
5%	0.790103
1%	0.790007
0% Min	0.790000

4.5. Distribution des rapports de poids
Poids après CALAGE V1/ poids initial

Variable: rat1b

Moments

N	39901	Sum Weights	39901
Mean	1.32455022	Sum Observations	52850.8781
Std Deviation	0.26251251	Variance	0.06891282
Skewness	0.69843448	Kurtosis	0.51143257
Uncorrected SS	72753.2634	Corrected SS	2749.6214
Coeff Variation	19.8189926	Std Error Mean	0.00131419

Basic Statistical Measures

Location		Variability	
Mean	1.324550	Std Deviation	0.26251
Median	1.290422	Variance	0.06891
Mode	1.600230	Range	1.97663
		Interquartile Range	0.38357

Tests for Location: Mu0=0

Test	-Statistic-		-----p Value------	
Student's t	t	1007.883	Pr > |t|	<.0001
Sign	M	19950.5	Pr >= |M|	<.0001
Signed Rank	S	3.9803E8	Pr >= |S|	<.0001

Quantiles (Definition 5)

Quantile	Estimate
100% Max	2.799547
99%	2.034189
95%	1.790412
90%	1.678398
75% Q3	1.500681
50% Median	1.290422
25% Q1	1.117110
10%	1.010026
5%	0.967616
1%	0.901942
0% Min	0.822920

4.6. Distribution des poids finaux après CALAGE V1

Moments

N	39901	Sum Weights	39901
Mean	1464.59474	Sum Observations	58438794.8
Std Deviation	652.949039	Variance	426342.448
Skewness	0.92658886	Kurtosis	2.40309868
Uncorrected SS	1.026E11	Corrected SS	1.70111E10
Coeff Variation	44.5822329	Std Error Mean	3.26879283

Basic Statistical Measures

Location		Variability	
Mean	1464.595	Std Deviation	652.94904
Median	1374.576	Variance	426342
Mode	1268.983	Range	5778
		Interquartile Range	977.05298

Tests for Location: Mu0=0

Test	-Statistic-		-----p Value------	
Student's t	t	448.0537	Pr > |t|	<.0001
Sign	M	19950.5	Pr >= |M|	<.0001
Signed Rank	S	3.9803E8	Pr >= |S|	<.0001

Tests for Normality

Test	--Statistic---		-----p Value------	
Kolmogorov-Smirnov	D	0.086448	Pr > D	<0.0100
Cramer-von Mises	W-Sq	66.14488	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	397.156	Pr > A-Sq	<0.0050

Quantiles (Definition 5)

Quantile	Estimate
100% Max	6116.112
99%	2926.890
95%	2519.044
90%	2344.214
75% Q3	1929.410
50% Median	1374.576
25% Q1	952.357
10%	728.354
5%	566.531
1%	412.184
0% Min	338.243

4.7. Distribution des rapports de poids correction de la non réponse individus V3 / individus V1

The UNIVARIATE Procedure
Variable: pondv3

Moments

N	35073	Sum Weights	35073
Mean	1.13765575	Sum Observations	39901
Std Deviation	0.07616234	Variance	0.0058007
Skewness	1.98152026	Kurtosis	5.48834822
Uncorrected SS	45597.0441	Corrected SS	203.442197
Coeff Variation	6.69467329	Std Error Mean	0.00040668

Basic Statistical Measures

Location		Variability	
Mean	1.137656	Std Deviation	0.07616
Median	1.116208	Variance	0.00580
Mode	1.076640	Range	0.51545
		Interquartile Range	0.08494

Tests for Location: Mu0=0

Test	-Statistic-		-----p Value------	
Student's t	t	2797.416	Pr > |t|	<.0001
Sign	M	17536.5	Pr >= |M|	<.0001
Signed Rank	S	3.0754E8	Pr >= |S|	<.0001

Quantiles (Definition 5)

Quantile	Estimate
100% Max	1.54015
99%	1.43651
95%	1.29050
90%	1.24074
75% Q3	1.16158
50% Median	1.11621
25% Q1	1.07664
10%	1.07664
5%	1.07316
1%	1.03889
0% Min	1.02469

4.8. Distribution des rapports de poids
Poids après CALAGE V3/ poids après non-réponse V3

Moments

N	35073	Sum Weights	35073
Mean	1.02400507	Sum Observations	35914.9298
Std Deviation	0.18237592	Variance	0.03326098
Skewness	0.18450901	Kurtosis	-1.3892615
Uncorrected SS	37943.5991	Corrected SS	1166.52896
Coeff Variation	17.8100603	Std Error Mean	0.00097383

Basic Statistical Measures

Location		Variability	
Mean	1.024005	Std Deviation	0.18238
Median	1.006288	Variance	0.03326
Mode	1.052402	Range	0.53998
		Interquartile Range	0.34773

NOTE: The mode displayed is the smallest of 2 modes with a count of 32.

Tests for Location: Mu0=0

Test	-Statistic-		-----p Value------	
Student's t	t	1051.529	Pr > |t|	<.0001
Sign	M	17536.5	Pr >= |M|	<.0001
Signed Rank	S	3.0754E8	Pr >= |S|	<.0001

Tests for Normality

Test	--Statistic---		-----p Value------	
Kolmogorov-Smirnov	D	0.099138	Pr > D	<0.0100
Cramer-von Mises	W-Sq	125.9997	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	898.3985	Pr > A-Sq	<0.0050

Quantiles (Definition 5)

Quantile	Estimate
100% Max	1.319983
99%	1.318270
95%	1.308551
90%	1.290164
75% Q3	1.196536
50% Median	1.006288
25% Q1	0.848803
10%	0.789279
5%	0.781064
1%	0.780096
0% Min	0.780002

4.9. Distribution des rapports de poids
Poids après CALAGE V3/ poids initial

The UNIVARIATE Procedure
Variable: rat3b

Moments

N	35073	Sum Weights	35073
Mean	1.50676359	Sum Observations	52846.7194
Std Deviation	0.34465838	Variance	0.1187894
Skewness	0.93300665	Kurtosis	1.19514936
Uncorrected SS	83793.6945	Corrected SS	4166.18172
Coeff Variation	22.8740845	Std Error Mean	0.00184036

Basic Statistical Measures

Location		Variability	
Mean	1.506764	Std Deviation	0.34466
Median	1.457312	Variance	0.11879
Mode	1.396704	Range	2.77411
		Interquartile Range	0.46593

Tests for Location: Mu0=0

Test	-Statistic-		-----p Value------	
Student's t	t	818.7338	Pr > |t|	<.0001
Sign	M	17536.5	Pr >= |M|	<.0001
Signed Rank	S	3.0754E8	Pr >= |S|	<.0001

Quantiles (Definition 5)

Quantile	Estimate
100% Max	3.646077
99%	2.538004
95%	2.142782
90%	1.957473
75% Q3	1.708349
50% Median	1.457312
25% Q1	1.242422
10%	1.111502
5%	1.053316
1%	0.973440
0% Min	0.871967

4.10. Distribution des poids finaux après CALAGE V3

Moments

N	35073	Sum Weights	35073
Mean	1666.20462	Sum Observations	58438794.8
Std Deviation	751.06606	Variance	564100.226
Skewness	0.94685551	Kurtosis	2.56335952
Uncorrected SS	1.17155E11	Corrected SS	1.97841E10
Coeff Variation	45.07646	Std Error Mean	4.01043682

Basic Statistical Measures

Location		Variability	
Mean	1666.205	Std Deviation	751.06606
Median	1584.809	Variance	564100
Mode	2165.299	Range	8573
		Interquartile Range	1064

Tests for Location: Mu0=0

Test	-Statistic-		-----p Value------	
Student's t	t	415.4671	Pr > |t|	<.0001
Sign	M	17536.5	Pr >= |M|	<.0001
Signed Rank	S	3.0754E8	Pr >= |S|	<.0001

Tests for Normality

Test	--Statistic---		-----p Value------	
Kolmogorov-Smirnov	D	0.068146	Pr > D	<0.0100
Cramer-von Mises	W-Sq	38.66643	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	243.198	Pr > A-Sq	<0.0050

Quantiles (Definition 5)

Quantile	Estimate
100% Max	8933.341
99%	3611.862
95%	2905.466
90%	2654.443
75% Q3	2154.678
50% Median	1584.809
25% Q1	1090.971
10%	800.210
5%	610.182
1%	436.286
0% Min	360.231

Série des Documents de Travail 'Méthodologie Statistique'

9601 : 'Une méthode synthétique, robuste et efficace pour réaliser des estimations locales de population'
G. DECAUDIN, J.-C. LABAT

9602 : 'Estimation de la précision d'un solde dans les enquêtes de conjoncture auprès des entreprises'
N. CARON, P. RAVALET, O. SAUTORY

9603 : 'La procédure FREQ de SAS® - Tests d'indépendance et mesures d'association dans
un tableau de contingence'
J. CONFAIS, Y. GRELET, M. LE GUEN

9604 : 'Les principales techniques de correction de la non-réponse et les modèles associés'
N. CARON

9605 : 'L'estimation du taux d'évolution des dépenses d'équipement dans l'enquête de conjoncture : analyse et voies d'amélioration'
P. RAVALET

9606 : 'L'économétrie et l'étude des comportements. Présentation et mise en œuvre de modèles de régression qualitatifs. Les modèles univariés à résidus logistiques ou normaux (LOGIT, PROBIT)'
S. LOLLIVIER, M. MARPSAT, D. VERGER

9607 : 'Enquêtes régionales sur les déplacements des ménages : l'expérience de Rhône-Alpes'
N. CARON, D. LE BLANC

9701 : 'Une bonne petite enquête vaut-elle mieux qu'un mauvais recensement ?'
J.C. DEVILLE

9702 : 'Modèles univariés et modèles de durée sur données individuelles'
S. LOLLIVIER

9703 : 'Comparaison de deux estimateurs par le ratio stratifiés et application aux enquêtes auprès des entreprises'
N. CARON, J.C. DEVILLE
9704 : 'La faisabilité d'une enquête auprès des ménages

1. au mois d'août. 2. à un rythme hebdomadaire'
C. LAGARENNE, C. THIESSET

9705 : 'Méthodologie de l'enquête sur les déplacements dans l'agglomération toulousaine'
P. GIRARD

9801 : 'Les logiciels de désaisonnalisation TRAMO & SEATS : philosophie, principes et mise en œuvre sous SAS'
K. ATTAL-TOUBERT, D. LADIRAY

9802 : 'Estimation de variance pour des statistiques complexes : technique des résidus et de linéarisation'
J.C. DEVILLE

9803 : 'Pour essayer d'en finir avec l'individu Kish'
J.C. DEVILLE

9804 : 'Une nouvelle (encore une !) méthode de tirage à probabilités inégales'
J.C. DEVILLE

9805 : 'Variance et estimation de variance en cas d'erreurs de mesure non corrélées ou de l'intrusion d'un individu Kish'
J.C. DEVILLE

9806 : 'Estimation de précision de données issues d'enquêtes : document méthodologique sur le logiciel POULPE'
N. CARON, J.C. DEVILLE, O. SAUTORY

9807 : 'Estimation de données régionales à l'aide de techniques d'analyse multidimensionnelle'
K. ATTAL-TOUBERT, O. SAUTORY

9808 : 'Matrices de mobilité et calcul de la précision associée'
N. CARON, C. CHAMBAZ

9809 : 'Echantillonnage et stratification : une étude empirique des gains de précision'
J. LE GUENNEC

9810 : 'Le Kish : les problèmes de réalisation du tirage et de son extrapolation'
C. BERTHIER, N. CARON, B. NÉROS

9811 : 'Vocabulaire statistique Français - Chinois - Anglais'
LIU Xiaoyue, CUI Bin

9901 : ‘Perte de précision liée au tirage d'un ou plusieurs individus Kish’
N. CARON

9902 : ‘Estimation de variance en présence de données imputées : un exemple à partir de l’enquête Panel Européen’
N. CARON

0001 : ‘L’économétrie et l’étude des comportements. Présentation et mise en oeuvre de modèles de régression qualitatifs. Les modèles univariés à résidus logistiques ou normaux (LOGIT, PROBIT) ’ (version actualisée)
S. LOLLIVIER, M. MARPSAT, D. VERGER

0002 : ‘Modèles structurels et variables explicatives endogènes’
Jean-Marc Robin INRA-LEA et CREST-INSEE.

0003 : ‘L’enquête 1997-1998 sur le devenir des personnes sorties du RMI- Une présentation de son déroulement’
D. ENEAU, D. GUILLEMOT

0004 : ‘Plus d’amis, plus proches? Essai de comparaison de deux enquêtes peu comparables’
O. GODECHOT

0005 : ‘Estimation dans les enquêtes répétées : Application à l’Enquête Emploi en Continu’
N. CARON, P. RAVALET

0006 : ‘Non-parametric approach to the cost-of-living index’
F. MAGNIEN, J. POUGNARD

0101 : ‘Diverses Macros SAS : Analyse exploratoire des données, Analyse des séries temporelles’
D. LADIRAY

0102 : ‘Econométrie linéaire des panels : une introduction’
T. MAGNAC

0201 : 'Application des méthodes de calage à l'enquête EAE-Commerce'
N. CARON

0203 : 'General principles for data editing in business surveys ans how to optimise it'
P. RIVIERE

0301 : 'Les modèles logit polytomiques non ordonnés :théorie et applications'
C. AFSA ESSAFI

0401 : 'Enquête sur le patrimoine des ménages - Synthèse des entretiens monographiques'
V. COHEN, C. DEMMER

0402 : 'La macro SAS CUBE d'échantillonnage équilibré'
S. ROUSSEAU, F. TARDIEU

M0501 : 'Correction de la non-réponse et calage de l'enquête Santé 2002'
N. CARON, S. ROUSSEAU

Série des Documents de Travail 'Méthodologie de Collecte'

C0201 : 'Comportement face au risque et à l'avenir et accumulation patrimoniale - Bilan d'une expérimentation'
L. ARRONDEL, A. MASSON, D. VERGER

0202 : 'Enquête Méthodologique Information et Vie Quotidienne - Tome 1 : bilan du test 1, novembre 2002'
L-A. VALLET, G. BONNET, J-C. EMIN, J. LEVASSEUR, T. ROCHER, P. VRIGNAUD, X. D'HAULTFOEUILLE, F. MURAT, D. VERGER, P. ZAMORA