

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

DREES

document
de *travail*

SÉRIE SOURCES ET MÉTHODES

N° 36 • septembre 2012

Traitements de la non-réponse et calages
pour l'enquête santé et itinéraire professionnel de 2010

Nicolas DE RICCARDIS



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

SÉRIE
SOURCES ET MÉTHODES

**DOCUMENT
DE
TRAVAIL**

Traitements de la non-réponse
et calages pour l'enquête Santé
et itinéraire professionnel de 2010

Nicolas DE RICCARDIS*

n° 36 – septembre 2012

* Les travaux présentés dans ce document de travail ont bénéficié du soutien technique
de Pascal Ardilly

« Une première version de ce document a été présentée
aux Journées de méthodologie statistique de 2012 »

Sommaire

Introduction	9
1. L'enquête Santé et Itinéraire Professionnel	11
1.1. Les principaux objectifs de l'enquête	11
1.2. Le champ de l'enquête.....	12
1.3. Mise en œuvre	12
2. Partie longitudinale - Le panel	15
2.1. Les caractéristiques de l'échantillon.....	15
2.2. Traitement de la non-réponse	15
2.2.1. Principe.....	15
2.2.2. Mise en œuvre dans SIP 2006 et 2010	16
2.2.3. Construction de groupes de réponse homogène pour le traitement de la non-réponse en 2010	19
2.3. Calage	23
2.3.1. Principes généraux	23
2.3.2. La mise en œuvre du calage en 2010	24
2.4. Exemple d'utilisation des pondérations du panel pour le calcul d'une moyenne	27
3. Partie transversale - Exposition des personnes ayant un emploi aux risques psychosociaux en 2010 (RPS).....	29
3.1. Pondérations de l'échantillon complémentaire.....	29
3.1.1. Principe.....	29
3.1.2. Mise en œuvre	30
3.2. Partage des poids	36
3.2.1. Rappels théoriques	36
3.2.2. Application dans SIP	37
3.3. Calage	39
3.4. Exemple d'utilisation des pondérations de la partie RPS pour le calcul d'une moyenne	41
Bibliographie.....	43
Annexe I - Les plans de sondages de l'enquête SIP	45
Le plan de sondage de la partie longitudinale	45
Tirage de l'échantillon complémentaire.....	45
Annexe II - Pondération de la première vague de l'enquête SIP	47
Le traitement initial effectué par la Dares en 2006	47
Les corrections apportées en 2010	47
Annexe III - L'algorithme CHAID utilisé par la macro TreeDisc	49
Annexe IV - La correction de la non-réponse du Panel par une régression logistique	51
Annexe V - Les sorties SAS de CalMar pour la partie Panel	55
Annexe VI - Les sorties SAS de CalMar pour la partie RPS	65

Dictionnaire des notations

- ech^{v1} = échantillon enquêté lors de la première vague de l'enquête, en 2006. Il est composé des individus âgés de 20 à 74 ans au 31 décembre 2006, quelle que soit leur situation vis-à-vis du marché du travail (actifs occupés, chômeurs ou inactifs, retraités, etc.).
- *Panel* = personnes ayant répondu aux deux vagues de l'enquête, ou qui ont répondu à la première vague mais qui sont décédées entre la première et la seconde vague.
- ech^C = échantillon complémentaire enquêté en 2010. Cet échantillon a été tiré dans la base de sondage formée par le fichier de la taxe d'habitation de 2009 et concerne les personnes actives en emploi vivant en ménage ordinaire en France métropolitaine.
- *RPS* = échantillon visant à décrire l'exposition aux facteurs de risques psychosociaux au travail. Il est composé des actifs occupés du *Panel* et des répondants de l' ech^C .
- $W_{v1,i}^{TNR,C}$ = poids traité de la non-réponse et calé de l'individu i de l' ech^{v1} . Il s'agit du poids final dont on dispose à l'issue de la première vague.
- $W_{Panel,i}^{TNR}$ = poids longitudinal après traitement de la non-réponse pour l'individu i du *Panel*.
- $W_{Panel,i}^{TNR,C}$ = poids longitudinal traité de la non-réponse et calé pour l'individu i du *Panel*.
- $W_{RPS,i}^{TNR}$ = poids après traitement de la non-réponse pour l'individu i de la partie *RPS*.
- $W_{RPS,i}^{TNR,C}$ = poids traité de la non-réponse et calé pour l'individu i de la partie *RPS*.

Introduction

L'enquête Santé et itinéraire professionnel (SIP) aborde de façon combinée les questions de conditions de travail et d'emploi et les questions de santé. C'est une enquête en deux vagues.

Elle comporte une approche rétrospective biographique dans la première vague de l'enquête, combinée à une approche longitudinale prospective reposant sur une double interrogation à quatre ans d'intervalle (fin 2006 puis fin 2010). De plus, chacune de ces deux vagues décrit de façon détaillée la situation, au moment de l'enquête, des personnes interrogées au regard de la santé et du travail/emploi.

Un deuxième objectif a été assigné à la seconde vague : recueillir des éléments sur les expositions aux risques psychosociaux (pour les personnes ayant un emploi au moment de l'enquête).

Le traitement de la non-réponse ainsi que le calage sur marges sont des étapes essentielles dans une enquête. Elles permettent respectivement de corriger, ou du moins de limiter, le biais introduit par la non-réponse, et de réduire la variance d'échantillonnage en modifiant la pondération (tout se passe comme si l'échantillon avait une structure identique à celle de la population étudiée).

Après avoir présenté l'enquête SIP, ce document de travail abordera la mise en place de deux jeux de pondérations : un premier permettant une exploitation en *Panel* (rassemblant les personnes répondantes aux deux vagues, ainsi que les répondants de la première vague décédés entre les deux vagues) et un second destiné à des analyses transversales sur les facteurs de risques psychosociaux au travail pour la seconde vague de l'enquête (2010). Ces jeux de pondérations sont construits à la suite des processus de traitement de la non-réponse et de calage.

1. L'enquête Santé et Itinéraire Professionnel

L'enquête Santé et itinéraire professionnel (SIP) est une enquête innovante qui cherche à établir les liens entre santé, actuelle et passée, et itinéraire professionnel complet. Cette enquête comporte une approche longitudinale, reposant sur une double interrogation à quatre ans d'intervalle (fin 2006 puis fin 2010), et une approche transversale sur les risques psychosociaux au travail.

1.1. Les principaux objectifs de l'enquête

L'enquête SIP vise un double objectif. D'une part, mieux connaître les déterminants de la santé liés aux grandes caractéristiques des parcours professionnels et des conditions de travail, en situant les états de santé au regard des situations de travail décrites. D'autre part, repérer en retour l'incidence de l'état de santé, au sens le plus large, sur le parcours professionnel des personnes, les aléas de carrière ou les discriminations éventuelles qu'elles ont pu rencontrer.

Afin de remplir ces objectifs, la première vague de l'enquête a abordé les thèmes suivants :

- l'enfance et les études : les grandes caractéristiques et les événements marquants de la vie privée au cours de l'enfance et dans la vie adulte, ainsi que les déménagements liés à la vie professionnelle ;
- l'emploi et le travail : les épisodes successifs de l'itinéraire professionnel (emploi, chômage ou inactivité), les principaux changements de travail et de conditions de travail, le travail actuel (contraintes temporelles, nuisances, dimension collective du travail, autonomie, moyens de travail, type ou absence de reconnaissance) et l'appréciation globale des enquêtés sur leur itinéraire professionnel ;
- les activités sociales, actuelles et antérieures, ainsi que le sentiment d'isolement ;
- les revenus du ménage ;
- la santé : les événements de santé, passés ou actuels (maladies, accidents, handicaps), leur interaction avec l'itinéraire professionnel et leur reconnaissance administrative (affections de longue durée, mise en inaptitude / en invalidité), la santé physique actuelle perçue, ainsi qu'un relevé d'éléments sur la santé mentale, les gênes fonctionnelles, douleurs et restrictions d'activités au moment de l'enquête.

Dans le questionnaire de la vague 2010, les enquêtés de 2006 étaient réinterrogés sur les caractéristiques du travail et de la santé au moment de l'enquête ; les risques psychosociaux faisaient l'objet d'une interrogation approfondie (échantillon complémentaire). Les enquêtés étaient également interrogés succinctement sur les principaux événements vécus entre 2006 et 2010 concernant leur situation professionnelle, leur santé et leur vie privée.

Pour répondre à la demande du Collège d'expertise sur le suivi statistique des risques psychosociaux au travail, un échantillon complémentaire a été tiré dans la base de la taxe d'habitation de 2009 (cf. annexe I). Il vise toutes les personnes ayant un emploi au moment de l'enquête en 2010, sans critère d'âge. Il permet aussi de compléter l'échantillon des répondants aux deux vagues de l'enquête, qui par définition sont nés avant le 31 décembre 1986 et ont donc au moins 24 ans en 2010. Ainsi, l'échantillon complémentaire vise à combler l'absence des enquêtés nés après cette date et à pallier une partie de l'attrition de

l'échantillon. Pour cela, il a été conçu pour surreprésenter les ménages contenant au moins un jeune actif occupé (moins de 24 ans) et, a contrario, sous-représenter les ménages comportant uniquement des personnes de 60 ans et plus (cf. partie 3.1.2.3 concernant les probabilités de tirage d'un individu « Kish »).

1.2. Le champ de l'enquête

La première vague de l'enquête a été effectuée auprès d'individus âgés de 20 à 74 ans au 31 décembre 2006, vivant en ménage ordinaire en France métropolitaine (cf. annexe I) ; il s'agit donc à la fois des actifs et des inactifs au moment de l'enquête. L'âge de 20 ans permet d'avoir des jeunes ayant déjà fini leurs études initiales. Une limite supérieure de 74 ans permet d'observer des itinéraires complets et d'avoir davantage de recul sur les effets différés du travail sur la santé ; concernant les questions sur la santé le choix de limiter le champ à cet âge est aussi dicté par le souci de pallier un risque plus élevé de mémoire défaillante dans l'évocation des données rétrospectives.

Les autres personnes exclues de l'enquête sont les personnes vivant en collectivité (notamment foyers, établissements pour handicapés, pour personnes âgées, prisons) et sans domicile fixe, qui constituent des sous-populations présentant des situations particulièrement critiques tant en termes de parcours professionnel que de santé.

1.3. Mise en œuvre

La vague 2006 de l'enquête SIP s'appuyait sur un questionnaire informatisé (CAPI), une grille biographique et un auto-questionnaire. La grille biographique permettait de reconstituer l'itinéraire professionnel de l'individu et de situer dans le temps certains événements marquants de sa vie familiale et privée (déménagements, naissance d'enfants, etc.). L'entretien, réalisé en face à face par un enquêteur de l'Insee, se poursuivait par la description et la saisie dans le questionnaire informatisé de chaque phase : emplois longs, courts, chômage ou inactivité. À chaque étape décrite, un lien était fait avec l'état de santé. Si l'enquêté déclarait une maladie, un handicap ou un accident, un ensemble de questions relatives à ces problèmes de santé étaient posées à l'enquêté à la suite de l'itinéraire professionnel. Au terme de l'entretien mené par l'enquêteur, l'enquêté se voyait proposer un auto-questionnaire portant notamment sur ses pratiques en matière de consommation d'alcool et de tabac, qu'il devait remplir et renvoyer ultérieurement. Près de 14 000 entretiens ont été réalisés en 2006.

La vague 2010 de l'enquête SIP s'est déroulée intégralement sous CAPI. Après un entretien en face à face avec un enquêteur, un auto-questionnaire sous CAPI avec casque audio a été soumis à l'enquêté ; il comportait les mêmes questions de santé qu'en 2006 ainsi que quelques questions sur des risques psychosociaux (expositions à des violences notamment). La vague 2010 visait à réinterroger l'ensemble des personnes ayant répondu à la première vague de l'enquête en 2006. Dans les faits, 11 016 personnes ont répondu à la deuxième vague.

Pour l'échantillon complémentaire, 4 659 ménages ont été enquêtés. Lors de la collecte, 1 412 ménages ne contenaient aucun individu en emploi, et ont donc été exclus. Compte tenu des informations inhérentes à la Taxe d'habitation, du plan de sondage (*cf.* annexe I) et de la population ciblée, un tel nombre de « hors-champ » n'est pas surprenant. Parmi les 3 247 ménages restants, 2 454 contenaient au moins un actif occupé qui a répondu à l'enquête.

2. Partie longitudinale - Le panel

La partie longitudinale est composée des répondants aux deux vagues de l'enquête et des personnes répondantes à la vague une mais décédées entre les deux vagues. La mise en place des pondérations se fait par un traitement de la non-réponse, suivi d'un calage sur marges.

2.1. Les caractéristiques de l'échantillon

La partie longitudinale de l'enquête SIP est donc composée des répondants, avec un questionnaire exploitable (nombre suffisant de réponses, itinéraire professionnel correctement décrit, etc.) aux vagues une et deux de l'enquête, ainsi que des répondants de la première vague décédés entre les deux vagues. En effet, l'enquête SIP est axée sur les relations entre santé et travail ; attribuer un poids aux personnes décédées permettra des les inclure dans des analyses. On qualifiera cet ensemble de *Panel*. Les individus partis à l'étranger ou en institution au moment de la collecte de la deuxième vague sont considérés comme « hors-champ ».

Sur les 13 648 individus de la première vague de SIP en 2006 ayant un questionnaire exploitable, 11 016 (81 %) restent dans le champ, ont répondu à la deuxième vague en 2010 et présentent un questionnaire exploitable ; 204 individus sont décédés entre les deux vagues. La faible attrition entre les deux vagues de l'enquête est en bonne partie dû à la mise en œuvre d'un suivi d'adresse.

Tableau 1 - Répartition des individus de la partie Panel

Répondants exploitables de la première vague	13 648 ¹
« Hors-champ » de la deuxième vague (parti en institution, à l'étranger...)	128
Non-répondants de la deuxième vague	2 300
Panel	11 220
- dont répondants à la deuxième vague	11 016
- dont décédés entre les deux vagues	204

2.2. Traitement de la non-réponse

2.2.1. Principe

On distingue deux types de non-réponse : la non-réponse partielle (lorsqu'un individu échantillonné ne répond qu'à une partie du questionnaire) et la non-réponse totale (le questionnaire de l'individu échantillonné reste vierge ou n'est que très partiellement rempli). Ces deux non-réponses font l'objet de traitements différents. Bien souvent, la non-réponse partielle est traitée par imputation (les valeurs manquantes sont alors remplacées par des valeurs plausibles), alors que l'on préfère corriger la non-réponse totale par une repondération : le poids initial des non-répondants devient nul, et celui des répondants est

¹ Aux 13 991 répondants de la première vague ont été enlevés quatre individus dont l'âge était hors des tranches imposées, neuf individus dont les années de naissances entre les deux vagues étaient incompatibles, 328 individus dont l'itinéraire professionnel était déclaré comme incomplet après les apurements effectués alors de la première vague, et deux individus dont l'itinéraires professionnel a été jugé incomplet à l'issue des apurements effectués lors de la deuxième vague.

dilaté. Le poids d'échantillonnage de chaque unité répondante est ainsi multiplié par l'inverse de sa probabilité de réponse.

Le traitement de la non-réponse partielle ne sera pas abordé ici ; par la suite, le terme de non-réponse fera donc toujours référence à la non-réponse totale.

2.2.2. Mise en œuvre dans SIP 2006 et 2010

Pour la première vague de l'enquête, il avait été supposé que tous les individus avaient la même probabilité de répondre ; un traitement homogène de la non-réponse a donc été mis en œuvre. Un calage sur marges a ensuite été effectué à l'aide de la fonction linéaire tronquée de la macro CalMar. Le calage portait sur le sexe croisé avec l'âge en tranches, l'activité croisée avec l'âge en tranches, la tranche d'unité urbaine, le nombre d'habitants du ménage, la catégorie socioprofessionnelle, et le secteur d'activité. Afin de traiter l'existence d'une non-réponse supplémentaire (itinéraires dits « incomplets »), une correction a posteriori des pondérations de 2006 a dû être effectuée. (cf. annexe II).

On définit ainsi $W_{v,i}^{TNR,C}$ comme le poids traité de la non-réponse et calé de l'individu i pour la première vague (ech^{v1}). Pour la seconde vague de l'enquête, les phases de correction de la non-réponse et de calage, qui font l'objet de ce document de travail, seront traitées l'une après l'autre. Sauf à supposer que les non-répondants ont le même comportement que les répondants, la non-réponse introduit un biais. Pour le corriger, on met en place des groupes de réponse homogène (GRH), afin d'estimer de façon plus pertinente la probabilité de réponse (et donc le poids) des unités répondantes. La première vague de l'enquête a permis de récolter des informations (géographiques, sociodémographiques...) sur les enquêtés. On définit alors les profils des non-répondants de 2010 (par ailleurs répondants en 2006) à partir des variables de 2006 sur leur situation vis-à-vis de l'emploi, la tranche d'unité urbaine, l'âge regroupé en classes, le niveau de diplôme atteint à la fin des études initiales, le sexe et l'état de santé déclaré² (voir tableau 2).

² L'étude du rôle de l'état de santé déclaré en 2006 dans le comportement de réponse en 2010 fait suite à une demande émanant du Comité du Label du CNIS. Par rapport à un individu ayant déclaré son état de santé comme « bon » ou « très bon », un répondant de la première vague qui a déclaré son état de santé comme altéré en 2006 (c'est-à-dire « moyen », « mauvais » ou « très mauvais ») a, toute chose égale par ailleurs, 1,2 fois plus de chance d'être répondant lors de la seconde vague en 2010.

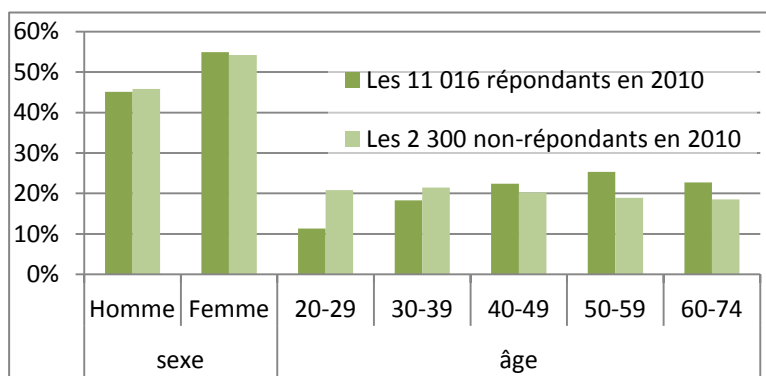
Tableau 2 - Variables utilisées pour définir le comportement de réponse à la seconde vague de l'enquête

Variable	Situation vis-à-vis de l'emploi (SituA)	Tranche d'unité urbaine (TU)	Classe d'âge (ClAge)	Niveau de diplôme à la fin des études initiales (NivDip)	Sexe (SexEnq)	Santé altérée (SanteAlt)
Modalités	1. vous occupez un emploi	0. communes rurales	2. 20-29 ans	. Non renseigné	1. homme	0. si l'enquêté se déclare en bonne ou très bonne santé
	2. vous êtes apprenti(e) sous contrat ou en stage rémunéré	1. communes des unités urbaines de moins de 20 000 habitants	3. 30-39 ans	1. Aucun diplôme	2. femme	1. si l'enquêté se déclare en moyenne, mauvaise ou très mauvaise santé
	3. vous êtes étudiant(e), élève, en formation ou en stage non rémunéré	2. communes des unités urbaines entre 20 000 et 100 000 habitants	4. 40-49 ans	2. CEP (certificat d'études primaires) ou diplôme étranger de même niveau		
	4. vous êtes chômeur (inscrit(e) ou non à l'ANPE)	3. communes des unités urbaines de plus de 100 000 habitants	5. 50-59 ans	3. Brevet des collèges, BEPC, brevet élémentaire ou diplôme étranger de même niveau		
	5. vous êtes retraité(e) ou retiré(e) des affaires ou en préretraite	4. Communes de l'unité urbaine de Paris	6. 60 ans et plus	4. CAP, BEP ou diplôme étranger de même niveau		
	6. vous êtes femme ou homme au foyer			5. Baccalauréat technologique ou professionnel ou diplôme étranger de même niveau		
	7. vous êtes dans une autre situation (personne handicapée...)			6. Baccalauréat général, brevet supérieur, capacité en droit, DAEU, ou diplôme étranger de même niveau		
				7. Diplôme de niveau BAC +2 8. Diplôme de niveau > BAC+2		

Les personnes décédées sont exclues de la mise en place des GRH ; elles conservent leur poids initial (de fait, tout se passe comme si elles étaient systématiquement déclarées répondantes à la seconde vague).

Encadré 1 - Indicateurs de santé et comportement de réponse

Cet encadré présente une analyse du lien entre état de santé en 2006 et comportement de réponse en 2010 qui mobilise d'autres indicateurs de santé que la santé perçue retenue dans la modélisation détaillée *infra* ; il répond notamment à une demande émanant du Comité du Label du CNIS. Ainsi, pour les 13 316³ individus de la première vague, on considère leur comportement de réponse à la seconde vague en fonction de ces indicateurs.



Les répondants de la première vague n'ayant pas répondu lors de la seconde vague sont plus jeunes : 44,2 ans en moyenne au 31/12/2006, contre 48,1 ans pour les répondants aux deux vagues (cf. Figure 1). Les non-répondants présentent plus souvent une santé altérée et un épisode dépressif majeur que les répondants (cf. Tableau 1) ; leur situation est en revanche plus favorable que celle des répondants pour ce qui concerne les troubles musculo-squelettiques (TMS) et les maladies chroniques.

Tableau 1 - Indicateurs de santé en 2006 et comportement de réponse en 2010

Répondants de la première vague ayant déclaré en 2006....:	Les 11 016 répondants en 2010	Les 2 300 non-répondants ⁴ en 2010	Odds-Ratio
Une santé altérée	27,7%	(*)30,4%	0,74
Des limitations depuis au moins six mois dans les activités quotidiennes	16,5%	16,8%	0,86
Avoir au moins une maladie chronique au moment de l'enquête	34,5%	(**)30,7%	NS
Un épisode dépressif majeur (EDM)	45,1%	45,8%	0,78
Un trouble d'anxiété généralisé (TAG)	7,1%	(*)8,6%	NS
Des troubles musculo-squelettiques	5,9%	6,4%	1,14
Des gênes motrices	44,9%	(**)40,0%	0,87
Des troubles du sommeil	46,7%	47,6%	0,91

Note de lecture : Lors de la seconde vague en 2010, 28% des répondants et 30% des non répondants avaient une santé altérée en 2006. Toute chose égale par ailleurs, une personne ayant une santé altérée en 2006 a 0,74 fois moins de chance d'être répondante en 2010.

L'état de santé se dégradant avec l'âge, le comportement de réponse des individus est modélisé, pour chacun des indicateurs de santé, par une régression logistique en tenant compte de l'âge en classes et du sexe. Ainsi, un répondant de la première vague qui a déclaré son état de santé comme altéré en 2006 (c'est-à-dire « moyen », « mauvais » ou « très mauvais ») a, toute chose égale par ailleurs, 0,74 fois moins de chance d'être répondant lors de la seconde vague en 2010 qu'un individu ayant déclaré son état de santé comme « bon » ou « très bon » ; autrement dit, déclarer un état de santé altéré lors de la première vague diminue de 26% la probabilité de répondre à la deuxième vague, ce qui est en ligne avec la statistique descriptive (28% des répondants ont une santé altérée contre 30% des non-répondants). A titre de comparaison, lorsque l'on ajoute des variables de contrôle sur le niveau de diplôme à la fin des études initiales, la tranche d'unité urbaine et la situation vis-à-vis de l'emploi, un individu avec une santé altérée a $1/1,225=0,82$ fois moins de chance de répondre à la seconde vague qu'un individu ayant déclaré un état de santé non altéré (cf. Annexe IV). Le lien entre réponse et présence de TMS est confirmé ; en revanche, le fait d'avoir une maladie chronique en 2006 ne joue plus sur le comportement de réponse en 2010. Pour les autres indicateurs de santé (sauf le trouble d'anxiété généralisé pour lequel le résultat n'est pas significatif), le fait de ne pas répondre en 2010 est associé à un plus mauvais état de santé en 2006.

³ On enlève aux 13 648 répondants exploitables de la première vague les 128 hors-champ et les 204 décédés.

⁴ ** (respectivement *) signale une différence significative entre répondants et non-répondants au seuil de 5% (respectivement 10%) ; dans les autres cas, les écarts ne sont pas significatifs

2.2.3. Construction de groupes de réponse homogène pour le traitement de la non-réponse en 2010

On définit des groupes de réponses homogènes (GRH) à partir de l'analyse du comportement de réponse en fonction de variables connues pour les répondants et les non-répondants. Ces variables sont donc issues des données récoltées lors de la première vague de l'enquête. On s'intéresse donc ici à la probabilité de ne pas répondre en 2010 conditionnellement au fait d'avoir répondu en 2006.

La population est divisée en sous-populations supposées homogènes au sens de la non-réponse. La probabilité de réponse est supposée constante au sein d'un GRH, et le comportement de réponse / non-réponse indépendant d'un GRH à l'autre. Au sein de chaque GRH, le mécanisme de réponse suit donc une loi de Bernoulli. Pour chaque GRH h , cette probabilité (P_{Panel}^h) est estimée par le rapport entre le nombre non pondéré de répondants en 2010 et le nombre non pondéré d'individus de cette sous-population répondante en 2006. On remarquera que les taux de réponse non pondérés et pondérés sont très proche (cf. tableau 3)

Il existe plusieurs méthodes pour mettre en place des GRH. On utilisera ici l'algorithme CHAID⁵. Cet algorithme est basé sur le test du χ^2 et permet d'identifier les caractéristiques qui divisent le mieux l'échantillon en groupes selon leurs propensions à répondre. L'algorithme regroupe pour chaque variable explicative ses modalités les moins liées à la variable à expliquer (réponse en 2010); il le fait de proche en proche par paire ou regroupement de modalités. Cette méthode offre une grande lisibilité des GRH constitués. L'algorithme CHAID est mis en œuvre avec la macro SAS TreeDisc (cf. Annexe III). La taille minimale d'un GRH a été fixée à 50 individus, afin de ne pas avoir d'effectifs trop faibles.

Il est également possible de traiter la non-réponse en estimant, pour les répondants de la première vague, la probabilité individuelle de répondre à la seconde vague de l'enquête. Cette probabilité est estimée par une régression logistique. Cette méthode aboutit à des résultats similaires (cf. annexe IV).

L'algorithme CHAID permet ainsi de déterminer 27 GRH pour la partie longitudinale de l'enquête SIP.

Soit $W_{Panel,i}^{TNR}$ le poids longitudinal traité de la non-réponse de l'individu i pour la partie *Panel*.

En définissant l'indicatrice $I_{Panel,i}^h$ (= 1 si l'individu i du *Panel* appartient au GRH h ; 0 sinon), on a :

$$W_{Panel,i}^{TNR} = \left(\sum_{h=1}^{27} \frac{1}{P_{Panel}^h} I_{Panel,i}^h \right) W_{v1,i}^{TNR,C}$$

Le processus de non-réponse est ainsi assimilé à la réalisation d'une enquête en deux phases, avec post-stratification. En effet, tout se passe comme si l'on tirait, au sein de l'échantillon initial post-stratifié par la mise en place des GRH, un second échantillon avec un taux de sondage égal, pour chaque GRH, à la probabilité qu'a l'individu de répondre.

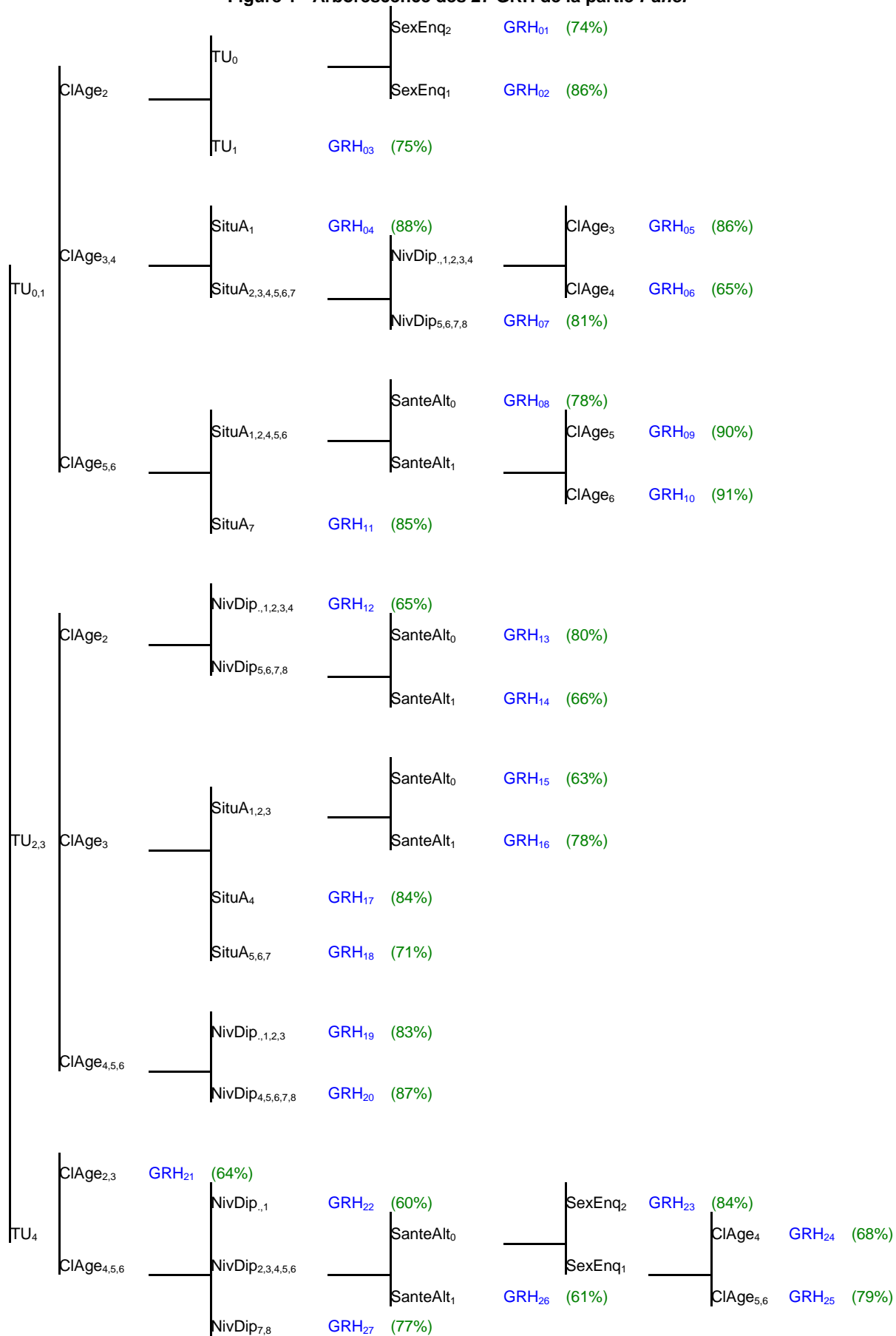
⁵ Chi-Square Automated Interaction Detection.

Tableau 3 - Détails des 27 GRH de la partie *Panel*

GRH	Variable	Obs.	Tx de rép.	Tx de rép. (pondéré)	Moy.	Q1	Q2	Q3	Min	Max	Somme des poids
1	$W_{vl}^{TNR,C}$	240	74%	77%	3 995	2 592	3 285	4 717	1 057	15 410	958 694
	W_{Panel}^{TNR}	177			5 627	3 891	4 454	6 776	1 433	20 895	996 051
2	$W_{vl}^{TNR,C}$	196	86%	86%	4 395	3 037	3 303	5 859	1 199	21 862	861 400
	W_{Panel}^{TNR}	169			5 094	3 523	3 831	6 795	1 441	25 355	860 959
3	$W_{vl}^{TNR,C}$	127	75%	75%	4 492	2 380	3 892	5 258	638	17 594	570 483
	W_{Panel}^{TNR}	95			6 050	3 181	5 203	7 306	950	23 520	574 715
4	$W_{vl}^{TNR,C}$	2 091	88%	88%	3 024	1 913	2 539	3 615	638	21 543	6 323 889
	W_{Panel}^{TNR}	1 837			3 452	2 177	2 891	4 147	726	24 522	6 340 698
5	$W_{vl}^{TNR,C}$	118	86%	89%	3 557	2 287	3 066	4 979	894	12 335	419 745
	W_{Panel}^{TNR}	101			4 344	2 909	3 582	5 823	1 045	14 412	438 728
6	$W_{vl}^{TNR,C}$	108	65%	67%	3 510	2 472	3 084	4 464	824	12 335	379 071
	W_{Panel}^{TNR}	70			5 563	3 813	4 758	7 137	1 577	19 032	389 423
7	$W_{vl}^{TNR,C}$	131	81%	86%	2 903	1 789	2 665	3 148	888	16 101	380 288
	W_{Panel}^{TNR}	106			3 828	2 364	3 316	3 891	1 098	19 898	405 780
8	$W_{vl}^{TNR,C}$	110	78%	80%	2 271	1 497	2 246	2 292	758	9 168	249 760
	W_{Panel}^{TNR}	86			2 966	1 915	2 873	2 932	970	11 727	255 040
9	$W_{vl}^{TNR,C}$	1 741	90%	90%	2 419	1 792	2 354	2 482	638	10 931	4 211 634
	W_{Panel}^{TNR}	1 564			2 694	2 084	2 620	2 763	710	10 973	4 213 104
10	$W_{vl}^{TNR,C}$	432	91%	89%	2 412	1 818	2 246	2 563	638	15 303	1 041 986
	W_{Panel}^{TNR}	391			2 624	2 024	2 481	2 818	704	12 250	1 026 080
11	$W_{vl}^{TNR,C}$	644	85%	86%	2 466	2 354	2 372	2 482	821	14 894	1 588 237
	W_{Panel}^{TNR}	547			2 934	2 771	2 792	2 922	967	17 535	1 604 742
12	$W_{vl}^{TNR,C}$	459	65%	67%	3 984	2 955	3 365	4 666	956	20 189	1 828 702
	W_{Panel}^{TNR}	297			6 348	4 808	5 500	7 379	1 478	29 318	1 885 431
13	$W_{vl}^{TNR,C}$	387	80%	80%	4 294	3 027	3 954	5 365	956	15 477	1 661 585
	W_{Panel}^{TNR}	311			5 290	3 680	4 894	6 455	1 190	19 260	1 645 115

14	$W_{vl}^{TNR,C}$	50	66%	68%	4 024	2 664	3 316	4 939	956	13 645	201 217
	W_{Panel}^{TNR}	33			6 266	3 525	5 061	7 675	1 449	20 674	206 767
15	$W_{vl}^{TNR,C}$	90	63%	60%	3 255	2 049	3 146	3 492	767	6 984	292 911
	W_{Panel}^{TNR}	57			4 867	2 757	4 852	5 514	1 212	11 028	277 437
16	$W_{vl}^{TNR,C}$	86	78%	78%	2 857	2 097	2 515	3 146	767	9 437	245 745
	W_{Panel}^{TNR}	67			3 684	2 692	3 182	4 038	1 061	12 113	246 836
17	$W_{vl}^{TNR,C}$	686	84%	84%	3 169	2 002	2 718	3 826	638	10 941	2 173 654
	W_{Panel}^{TNR}	576			3 781	2 452	3 251	4 556	805	13 030	2 177 957
18	$W_{vl}^{TNR,C}$	126	71%	70%	2 896	1 913	2 540	3 361	761	11 990	364 897
	W_{Panel}^{TNR}	90			3 963	2 678	3 542	4 130	1 066	14 880	356 671
19	$W_{vl}^{TNR,C}$	1 291	83%	83%	2 545	1 782	2 433	2 577	638	15 779	3 286 109
	W_{Panel}^{TNR}	1 077			3 048	2 254	2 917	3 059	764	18 915	3 282 258
20	$W_{vl}^{TNR,C}$	2 361	87%	86%	2 579	1 913	2 416	2 763	638	14 166	6 089 746
	W_{Panel}^{TNR}	2 052			2 946	2 201	2 759	3 158	734	16 299	6 044 706
21	$W_{vl}^{TNR,C}$	672	64%	65%	3 992	2 577	3 694	4 655	914	13 661	2 682 446
	W_{Panel}^{TNR}	432			6 258	4 008	5 707	7 251	1 488	21 250	2 703 553
22	$W_{vl}^{TNR,C}$	141	60%	63%	3 139	2 058	2 982	3 737	638	8 450	442 635
	W_{Panel}^{TNR}	84			5 597	3 941	5 046	6 439	2 502	14 184	470 133
23	$W_{vl}^{TNR,C}$	385	84%	85%	3 164	2 218	3 032	3 595	638	14 480	1 218 191
	W_{Panel}^{TNR}	325			3 766	2 761	3 592	4 259	755	17 153	1 224 100
24	$W_{vl}^{TNR,C}$	220	68%	67%	3 027	2 074	2 982	3 234	952	9 277	666 036
	W_{Panel}^{TNR}	150			4 389	3 023	4 373	4 653	1 396	12 768	658 367
25	$W_{vl}^{TNR,C}$	257	79%	78%	2 985	2 111	3 055	3 173	785	12 597	767 142
	W_{Panel}^{TNR}	203			3 725	2 657	3 895	4 016	994	15 948	756 110
26	$W_{vl}^{TNR,C}$	56	61%	57%	3 618	2 828	3 524	3 887	1 690	8 994	202 612
	W_{Panel}^{TNR}	34			5 622	4 590	5 804	6 403	3 095	12 817	191 136
27	$W_{vl}^{TNR,C}$	112	77%	74%	3 295	2 571	2 982	3 485	956	8 945	369 058
	W_{Panel}^{TNR}	86			4 128	3 499	3 883	4 055	1 246	11 649	354 983

Figure 1 - Arborescence des 27 GRH de la partie Panel



Note de lecture : Le GRH₀₁ est composé des femmes (SexEnq₂) âgées de 20 à 29 ans (CIAge₂) qui vivent en commune rurale(TU₀). Dans ce GRH, 74 % des personnes interrogées ont répondu.

Tableau 4 - Quantiles des pondérations avant et après traitement de la non-réponse

Quantile	$W_{v1}^{TNR,C}$	W_{Panel}^{TNR}	$W_{Panel}^{TNR}/W_{v1}^{TNR,C}$
Max	21 863	29 319	1,68
99%	9 483	12 381	1,65
95%	6 072	7 763	1,56
90%	4 984	6 222	1,47
75%	3 285	4 157	1,20
50%	2 479	2 901	1,15
25%	1 923	2 321	1,14
10%	1 336	1 579	1,11
5%	1 210	1 386	1,10
1%	891	1 057	1,00
Min	638	705	1,00

Après traitement de la non-réponse, le rapport des poids entre $W_{Panel}^{TNR}/W_{v1}^{TNR,C}$ se situe dans l'intervalle [1 ; 1,68] ; 98% des poids W_{Panel}^{TNR} sont compris entre 1 057 et 12 381 (contre 891 et 9 483 initialement).

2.3. Calage

Une fois le traitement de la non-réponse effectué, un calage sur marges est mis en œuvre. Après quelques rappels théoriques, la mise en œuvre dans SIP est alors détaillée.

2.3.1. Principes généraux

Le calage permet d'égaliser des totaux pondérés issus de l'échantillon avec les totaux connus de la population, tout en assurant une déformation minimale des pondérations initiales. Le calage permet, par la prise en compte d'informations auxiliaires, d'accroître la précision des estimations pour une variable d'intérêt « expliquée » par ces variables auxiliaires. En effet, le traitement de la non-réponse permet d'éliminer le biais introduit par la non-réponse (en théorie, car on pratique on se contentera de le limiter). Toutefois, à ce stade, il est peu probable que les totaux pondérés issus de l'échantillon répondant coïncident avec les totaux observés pour la population concernée.

Pour l'ensemble des individus du *Panel*, CalMar minimise la distance entre les poids initiaux ($W_{Panel,i}^{TNR}$) et les pondérations recherchées ($W_{Panel,i}^{TNR,C}$), à l'aide d'une fonction G .

Pour l'enquête SIP, les marges seront constituées à partir des quatre enquêtes Emploi trimestrielles de l'année 2006, afin de refléter une situation « moyenne » de l'année.

Encadré 2 - Principe de la méthode de calage sur marges

Soit une population U , dans laquelle on tire un échantillon probabiliste s . Soit y une variable d'intérêt, dont on désire estimer le total sur la population : $Y = \sum_{k \in U} y_k$.

Habituellement, on estime Y par l'estimateur de Horvitz-Thompson. En appelant p_k la probabilité d'inclusion de l'élément k dans l'échantillon s , on a : $\hat{Y} = \sum_{k \in s} \frac{1}{p_k} y_k = \sum_{k \in s} w_k y_k$

On suppose que l'on connaît les totaux sur la population de J variables auxiliaires⁶ $X_1 \dots X_j \dots X_J$, disponibles sur l'échantillon : $X_j = \sum_{k \in U} x_{jk}$

On va chercher de nouvelles pondérations, que l'on qualifie de poids « calés » (w_k^C), qui soient aussi proches que possible, au sens d'une certaine fonction de distance, des pondérations initiales w_k , et qui assurent le calage sur les totaux des variables X_j , i.e. qui vérifient les équations de calage : $\forall j = 1 \dots J \sum_{k \in s} w_k^C x_{jk} = X_j$ (1)

La solution de ce problème est donnée par $w_k^C = w_k F(x'_k \lambda)$, où $x'_k = (x_{1k} \dots x_{jk})$, λ est un vecteur de J multiplicateurs de Lagrange associés aux contraintes (1), et F une fonction dont l'expression dépend du choix de la fonction de distance : elle est appelée fonction de calage.

Le vecteur λ est déterminé par la résolution du système non linéaire de J équations à J inconnues résultant des équations de calage : $\sum_{k \in s} w_k F(x'_k \lambda) x_k = X$

L'estimateur du total d'une variable d'intérêt sera alors l'estimateur « calé » $\hat{Y}^C = \sum_{k \in s} w_k^C y_k$.

2.3.2. La mise en œuvre du calage en 2010

La partie *Panel* doit posséder des totaux cohérents avec ceux de la population française des 20-74 ans en 2006 (cf. partie 1.2). Ainsi, le calage se fera sur la moyenne des quatre enquêtes Emploi⁷ trimestrielles relatives à l'année 2006. Le calage s'effectuera sur six variables (qui conduisent à 29 modalités au total) relatives à la première vague (cf. tableau 5). Le calage n'a pas été fait directement sur la catégorie socioprofessionnelle ; qui a donc pu servir *ex post* de variable de contrôle de la qualité du calage.

La macro CalMar dispose de quatre fonctions G différentes pour effectuer un calage sur marges. Il n'existe pas de critère de choix strict de la fonction de calage à utiliser, ces quatre fonctions étant asymptotiquement équivalentes. La méthode retenue est celle la plus couramment utilisée : la fonction logit tronquée. Les sorties SAS fournies par la macro CalMar sont disponibles en Annexe V.

⁶ Il s'agit de variables quantitatives ou d'indicatrices associées aux modalités de variables catégorielles.

⁷ Même si les résultats des enquêtes annuelles de recensement sont disponibles pour l'année 2006, il reste encore usuel d'effectuer un calage à partir de l'enquête Emploi.

Tableau 5 - Liste des variables utilisées pour le calage de la partie Panel

Variable	Zone d'études et d'aménagement du territoire	Tranche d'unité urbaine	Classe d'âge au 31 décembre 2006 par sexe	Niveau de diplôme à la fin des études initiales	Nationalité de l'enquête	Nombre d'habitants du logement	
Modalités	1. Île-de-France, Bassin parisien	0. communes rurales	1-20. Homme de 20 à 29 ans	0. Aucun diplôme ou CEP ou diplôme équivalent ou diplôme non déclaré	1. Française	1. Une personne	
	2. Nord, Est, Ouest	1. communes des unités urbaines de moins de 20 000 habitants	1-30. Homme de 30 à 39 ans	1. CAP, BEP ou autre diplôme de ce niveau ou BEPC seul	2. Autres	2. Deux personnes	
	3. Sud-ouest, Centre-est, Méditerranée	2. communes des unités urbaines entre 20 000 et 100 000 habitants	1-40. Homme de 40 à 49 ans	2. Bac ou brevet professionnel ou diplôme de ce niveau ou Bac+2 ans,		3. Trois personnes	
		3. communes des unités urbaines de plus de 100 000 habitants	1-50. Homme de 50 à 59 ans	3. Diplôme supérieur		4. Quatre ou cinq personnes	
			4. Communes de l'unité urbaine de Paris	1-60. Homme de 60 à 74 ans			5. Au moins six personnes
				2-20. Femme de 20 à 29 ans			
				2-30. Femme de 30 à 39 ans			
				2-40. Femme de 40 à 49 ans			
	2-50. Femme de 50 à 59 ans						
	2-60. Femme de 60 à 74 ans						

Tableau 6 : répartition des poids et des rapports de poids selon la fonction de calage utilisée

Quantile	Logit tronquée		Raking ratio		Linéaire tronquée		Linéaire	
	Lo=0,63	Up=1,80			Lo=0,57	Up=1,69		
	Poids	Rapport des poids	Poids	Rapport des poids	Poids	Rapport des poids	Poids	Rapport des poids
Max	25 860	1,75	25 597	2,20	25 754	1,69	25 726	1,89
99%	13 403	1,67	13 301	1,80	13 266	1,69	13 253	1,69
95%	8 219	1,55	8 293	1,53	8 274	1,53	8 275	1,53
90%	6 442	1,38	6 433	1,33	6 427	1,36	6 428	1,35
75%	4 348	1,16	4 334	1,14	4 337	1,14	4 329	1,14
50%	2 914	0,96	2 914	0,97	2 934	0,98	2 935	0,98
25%	2 114	0,82	2 125	0,84	2 111	0,84	2 113	0,84
10%	1 589	0,74	1 581	0,74	1 573	0,73	1 570	0,74
5%	1 280	0,71	1 278	0,68	1 282	0,65	1 279	0,65
1%	943	0,68	940	0,63	913	0,57	916	0,57
Min	571	0,66	553	0,56	522	0,57	524	0,46

La catégorie socioprofessionnelle n'entrant pas dans le calage, elle donne un élément d'appréciation de l'impact numérique de l'opération. À l'exception des catégories « non renseigné » et « autres personnes sans activité professionnelle », et dans une moindre mesure

de la catégorie « professions Intermédiaires », les répartitions des catégories socioprofessionnelles, avant et après calage, sont proches de celles de la population. (cf. tableau 7). Il convient de noter que 748 individus (soit environ 2,9 millions en pondéré) n'ont pas de catégorie socioprofessionnelle attribuée.

Tableau 7 - Répartition des catégories socioprofessionnelles dans l'échantillon et dans la population.

Catégories socioprofessionnelles	Pourcentage dans l'échantillon		Pourcentage dans la population
	Avant calage	Après calage	
Non renseigné	2,10%	2,06%	0,01%
Agriculteurs exploitants	1,75%	1,76%	1,56%
Artisans, commerçants et chefs d'entreprise	3,66%	3,67%	4,09%
Cadres et professions intellectuelles supérieures	10,23%	9,43%	10,16%
Professions Intermédiaires	16,42%	16,25%	15,05%
Employés	19,88%	20,10%	19,53%
Ouvriers	15,17%	15,16%	15,67%
Retraités	20,30%	20,69%	20,76%
Autres personnes sans activité professionnelle	10,48%	10,88%	13,17%

Après traitement de la non-réponse et calage, le rapport des poids entre $W_{Panel}^{TNR,C} / W_{Panel}^{TNR}$ est compris dans l'intervalle $[0,66 ; 1,75]$; 98 % des poids finaux de la partie *Panel* $W_{Panel}^{TNR,C}$ sont compris entre 943 et 13 403.

Tableau 8 - Répartition des poids et des rapports de poids finaux de la partie *Panel*

Quantile	$W_{v1}^{TNR,C}$	W_{Panel}^{TNR}	$W_{Panel}^{TNR,C}$	$W_{Panel}^{TNR,C} / W_{Panel}^{TNR}$	$W_{Panel}^{TNR,C} / W_{v1}^{TNR,C}$
Max	21 863	29 319	25 860	1,75	2,68
99%	9 483	12 381	13 403	1,67	2,16
95%	6 072	7 763	8 219	1,55	1,91
90%	4 984	6 222	6 442	1,38	1,75
75%	3 285	4 157	4 348	1,16	1,41
50%	2 479	2 901	2 914	0,96	1,14
25%	1 923	2 321	2 114	0,82	0,96
10%	1 336	1 579	1 589	0,74	0,87
5%	1 210	1 386	1 280	0,71	0,83
1%	891	1 057	943	0,68	0,79
Min	638	705	571	0,66	0,67

2.4. Exemple d'utilisation des pondérations du panel pour le calcul d'une moyenne

Pour une variable d'intérêt Y , on estimera sa moyenne (\bar{Y}) par une moyenne pondérée sur un ensemble s_n , composé de n éléments, appartenant à la partie *Panel* soit :

$$\bar{Y}_{Panel} = \frac{\sum_{i \in s_n} W_{Panel,i}^{TNR,C} Y_i}{\sum_{i \in s_n} W_{Panel,i}^{TNR,C}}$$

3. Partie transversale - Exposition des personnes ayant un emploi aux risques psychosociaux en 2010 (RPS)

L'observation transversale des facteurs de risques psychosociaux au travail en 2010 concerne uniquement les personnes en emploi au moment de l'enquête.

Il s'agit des individus de l'échantillon complémentaire (*ech^C*) et des actifs occupés de la partie *Panel*. On qualifiera cette partie *RPS*. Sur les 4 659 ménages échantillonnés de l'*ech^C*, 1 412 ne contenaient aucun actif occupé au moment de l'enquête, et 787 n'ont pas répondu. On dénombre 6 367 actifs occupés dans le *Panel*. Avec les 2 454 personnes en emploi de l'*ech^C* ayant répondu, on dispose au total de 8 821 actifs occupés pour la partie *RPS*.

Tableau 9 - Répartition de l'échantillon complémentaire et de la partie RPS

Ménage	Ménages enquêtés de l'échantillon complémentaire	4 659
	Ménages de l'échantillon complémentaire ne contenant aucun individu en emploi	1 412
	Ménages non-répondants de l'échantillon complémentaire	787
Individu	individus référencés dans le tableau des habitants du logement	3 971
	- dont individu de moins de 24 ans	318
	- dont individu ayant entre 24 et 59 ans	3 558
	- dont individu de 60 ans et plus	95
	Individus sélectionnés par le tirage Kish	2 460
	- dont individu de moins de 24 ans	208
	- dont individu ayant entre 24 et 59 ans	2 179
- dont individu de 60 ans et plus	73	
	Individus non-répondants	6
	RPS	8 821
	- dont individus répondants de l'échantillon complémentaire	2 454
	- dont actifs occupés de la partie <i>Panel</i>	6 367

La mise en place des pondérations de la partie *RPS* passe également par un traitement de la non-réponse suivi d'un calage sur marges. Toutefois, afin de tenir compte des deux sous-populations qui la composent, une étape supplémentaire de partage des poids sera nécessaire.

3.1. Pondérations de l'échantillon complémentaire

Le *Panel* ayant déjà été traité de la non-réponse (cf. partie 2), on se contentera de corriger la non-réponse de l'échantillon complémentaire.

3.1.1. Principe

Les individus sont enquêtés à partir des ménages tirés dans l'échantillon. Il convient donc de distinguer la cause de non-réponse d'un ménage de celle propre aux caractéristiques de l'individu. On suppose qu'au sein d'un même logement, les comportements individuels sont indépendants (on peut considérer que cette hypothèse est assez forte).

Ainsi, la probabilité qu'un individu de la population soit tiré dans l'échantillon complémentaire et réponde à l'enquête (P_i) dépend :

- de la probabilité initiale de tirage du ménage m auquel il appartient (π_m),
- de la probabilité estimée de réponse du ménage (P_m),
- de la probabilité de tirage de l'individu⁸ i pour le ménage m ($Kish_{m,i}$),
- de la probabilité estimée de réponse de l'individu sélectionné sachant qu'il est tiré ($P_{i/m}$).

On a donc ainsi $P_i = \pi_m \times P_m \times Kish_{m,i} \times P_{i/m}$

Pour les individus de l'échantillon complémentaire, le poids après traitement de la non-réponse sera donc : $W_{ech^c,i}^{TNR} = \frac{1}{P_i}$

3.1.2. Mise en œuvre

3.1.2.1. Les probabilités initiales de tirage du ménage

Les probabilités initiales de tirage d'un ménage π_m sont définies par le plan de sondage, à partir des types de ménages définis dans la taxe d'habitation.

Tableau 10 - Répartition des probabilités d'inclusion π_m

Strates de ménages		Taille dans l'échantillon	$1/\pi_m$
Résidence principale	Ménages dont l'un des membres est né entre le 01/01/1987 et le 31/12/1992 inclus	1 722	1 782
	Ménages ne comportant que des plus de 60 ans (tous les membres sont nés avant le 01/01/1951).	430	20 872
	Autres types de ménages	2 150	6 466
Résidences secondaires		70	6 275
Logements vacants		287	2 468

3.1.2.2. Calcul des probabilités de réponse d'un ménage

Sur les 3 247 ménages interrogés appartenant au champ de l'enquête, 787 sont non-répondants. À partir des données récupérables par la taxe d'habitation de 2009, on cherche à définir un comportement de réponse des ménages sélectionnés. Pour cela, on met de nouveau en place des GRH à partir de la macro TreeDisc (cf. partie 2.2.3), sur les variables portant sur l'unité urbaine en tranches, la taille du ménage en tranches et les revenus du ménage en tranches. Pour chaque GRH h , la probabilité de réponse d'un ménage π_m^h est estimée par le rapport entre le nombre non pondéré de ménages répondants et le nombre non pondéré de ménages de l'échantillon appartenant à cette sous-population.

L'algorithme CHAID permet ainsi de déterminer 16 GRH.

⁸ Une surreprésentation (trois fois plus de chance) des jeunes de moins de 24 ans a été effectuée grâce à une méthode de tirage Kish à probabilité inégale (Cf. partie 3.1.2.3).

Tableau 11 - Variables utilisées pour définir le comportement de réponse des ménages pour l'ech^C.

Nom de la variable	TU	TrTail	TrRev
libellé	Tranches d'unité urbaine	Tranches de taille du ménage	Tranches de revenu ⁹ du ménage (en euro)
modalités	0. communes rurales	1. une personne	0. [0 – 10 000[
	1. communes des unités urbaines de moins de 20 000 habitants	2. deux personnes	1. [10 000 – 20 000[
	2. communes des unités urbaines entre 20 000 et 100 000 habitants	3. trois ou quatre personnes	2. [20 000 – 30 000[
	3. communes des unités urbaines de plus de 100 000 habitants	4. au moins cinq personnes	3. [30 000 – 40 000[
	4. Communes de l'unité urbaine de Paris		4. [40 000 – 50 000[
			5. [50 000 – 60 000[
			6. [60 000 – 70 000[
		7. au moins 70 000	

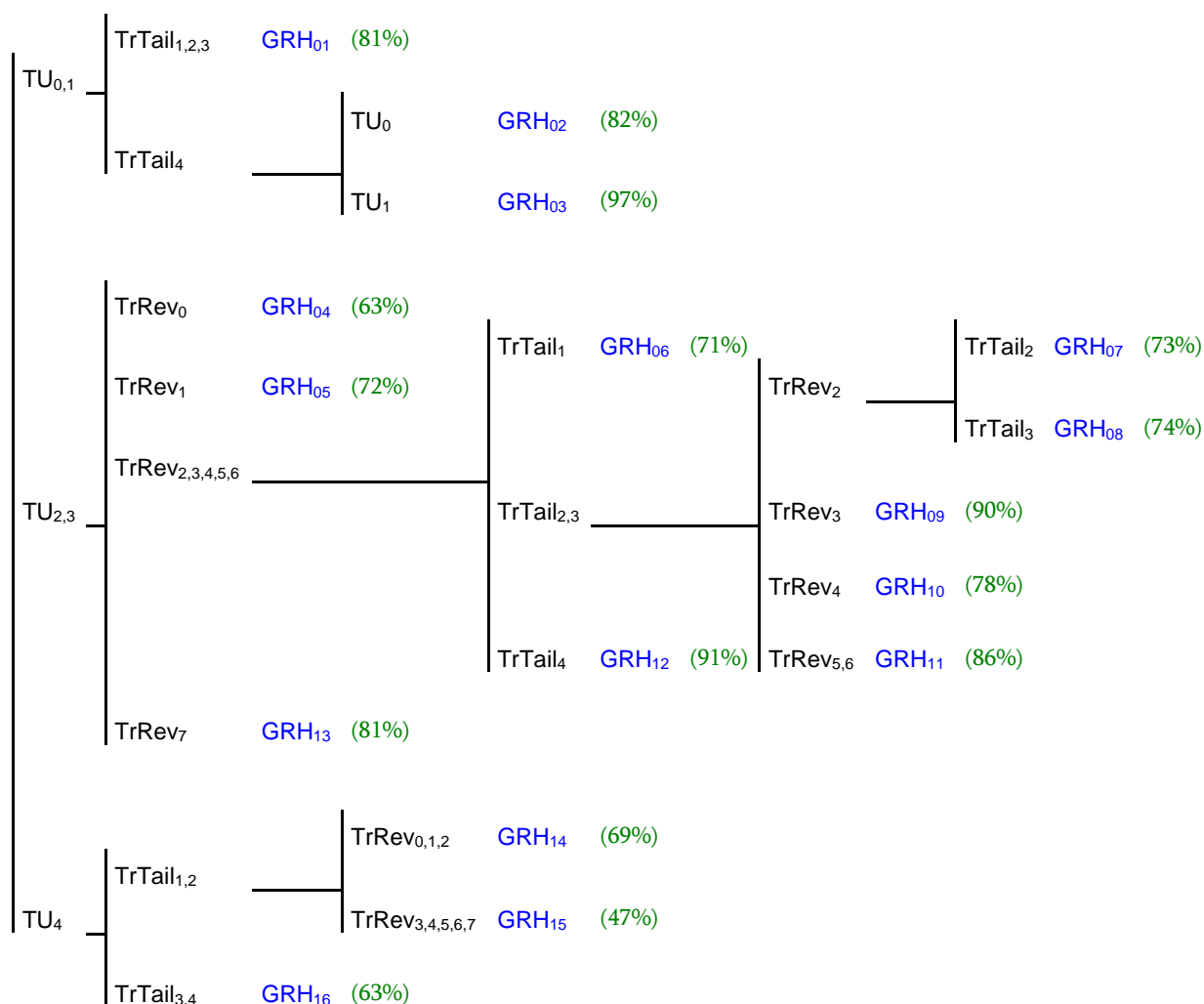
Tableau 12 - Détails des 16 GRH de la partie RPS issus de l'algorithme CHAID

GRH	Variable	Obs.	Tx de rép.	Tx de rép. (pond.)	Moy.	Q1	Q2	Q3	Min	Max	Somme des poids
1	$1/\pi_m^{01}$	1 164	81%	80%	4 614	1 782	6 466	6 466	1 782	20 872	5 370 643
	$1/\left(\frac{\pi_m^{01}}{P_m^{01}}\right)$	942			5 644	2 202	7 989	7 989	2 202	25 791	5 316 971
2	$1/\pi_m^{02}$	114	82%	82%	3 031	1 782	1 782	6 466	1 782	6 466	345 550
	$1/\left(\frac{\pi_m^{02}}{P_m^{02}}\right)$	93			3 753	2 185	2 185	7 926	2 185	7 926	348 993
3	$1/\pi_m^{03}$	72	97%	97%	3 408	1 782	1 782	6 466	1 782	6 466	245 409
	$1/\left(\frac{\pi_m^{03}}{P_m^{03}}\right)$	70			3 485	1 833	1 833	6 650	1 833	6 650	243 937
4	$1/\pi_m^{04}$	169	63%	62%	5 165	1 782	6 466	6 466	1 782	20 872	872 967
	$1/\left(\frac{\pi_m^{04}}{P_m^{04}}\right)$	107			8 046	2 815	10 212	10 212	2 815	32 966	860 969
5	$1/\pi_m^{05}$	337	72%	69%	5 333	1 782	6 466	6 466	1 782	20 872	1 797 144
	$1/\left(\frac{\pi_m^{05}}{P_m^{05}}\right)$	242			7 145	2 482	9 004	9 004	2 482	29 066	1 729 153
6	$1/\pi_m^{06}$	96	71%	75%	3 540	1 782	1 782	6 466	1 782	20 872	339 795
	$1/\left(\frac{\pi_m^{06}}{P_m^{06}}\right)$	68			5 313	2 516	2 516	9 128	2 516	29 466	361 307
7	$1/\pi_m^{07}$	102	73%	70%	6 486	6 466	6 466	6 466	1 782	20 872	661 535
	$1/\left(\frac{\pi_m^{07}}{P_m^{07}}\right)$	74			8 566	8 912	8 912	8 912	2 457	28 769	633 875
8	$1/\pi_m^{08}$	90	74%	69%	2 987	1 782	1 782	6 466	1 782	6 466	268 810
	$1/\left(\frac{\pi_m^{08}}{P_m^{08}}\right)$	67			3 722	2 394	2 394	2 394	2 394	8 685	249 404

⁹ Le revenu du ménage correspond au revenu fiscal de référence, sommé sur les foyers fiscaux rattachés.

9	$1/\pi_m^{09}$	155	90%	91%	4 066	1 782	1 782	6 466	1 782	20 872	630 278
	$1/\left(\pi_m^{09} P_m^{09}\right)$	140			4 514	1 973	1 973	7 158	1 973	23 108	631 913
10	$1/\pi_m^{10}$	98	78%	74%	4 285	1 782	4 467	6 466	1 782	20 872	419 928
	$1/\left(\pi_m^{10} P_m^{10}\right)$	76			5 250	2 298	2 741	8 337	2 298	8 337	398 995
11	$1/\pi_m^{11}$	95	86%	86%	3 915	1 782	1 782	6 466	1 782	6 466	371 883
	$1/\left(\pi_m^{11} P_m^{11}\right)$	82			4 532	2 065	2 065	7 491	2 065	7 491	371 664
12	$1/\pi_m^{12}$	55	91%	91%	5 384	1 782	6 466	6 466	1 782	20 872	296 125
	$1/\left(\pi_m^{12} P_m^{12}\right)$	50			5 907	1 961	7 112	7 112	1 961	22 959	295 329
13	$1/\pi_m^{13}$	110	81%	81%	3 749	1 782	1 782	6 466	1 782	20 872	412 338
	$1/\left(\pi_m^{13} P_m^{13}\right)$	89			4 624	2 203	2 203	7 991	2 203	25 797	411 512
14	$1/\pi_m^{14}$	302	69%	69%	3 340	1 782	1 782	6 466	1 782	6 466	1 008 696
	$1/\left(\pi_m^{14} P_m^{14}\right)$	209			4 839	2 575	2 575	9 343	2 575	9 343	1 011 322
15	$1/\pi_m^{15}$	187	47%	46%	6 112	6 466	6 466	6 466	1 782	20 872	1 142 975
	$1/\left(\pi_m^{15} P_m^{15}\right)$	87			12 860	5 306	13 897	13 897	3 831	44 863	1 118 847
16	$1/\pi_m^{16}$	101	63%	60%	6 550	6 466	6 466	6 466	1 782	20 872	661 508
	$1/\left(\pi_m^{16} P_m^{16}\right)$	64			9 766	10 204	10 204	10 204	2 813	32 939	625 040

Figure 2 - Arborescence des 16 GRH du comportement de réponse des ménages de l'ech^C



Note de lecture : le GRH₀₁ est composé des ménages de moins de cinq personnes (TrTail_{1,2,3}) vivants en commune rurale ou en commune urbaine de moins de 20 000 habitants (TU_{0,1}). Dans ce GRH, 81% des personnes interrogées ont répondu.

3.1.2.3. Calcul des probabilités de tirage d'un individu au sein d'un ménage

Compte tenu de la faible proportion de jeunes de moins de 24 ans ayant un emploi dans la population¹⁰, un tirage Kish à probabilités égales à l'intérieur du ménage aurait conduit à un nombre trop faible de jeunes actifs dans l'échantillon complémentaire. Une surreprésentation des jeunes actifs occupés de moins de 24 ans au 31 décembre 2010 a donc été effectuée grâce à une méthode de tirage Kish à probabilités inégales. Ainsi, ils ont trois fois plus de chance d'être sélectionné qu'un actif occupé de 24 ans ou plus. En effet, se limiter au champ de l'enquête de 2006 aurait conduit à écarter les jeunes (les personnes interrogées en 2006 étant nées entre le 01/01/1932 et le 31/12/1986).

¹⁰ Dans l'enquête Emploi 2007, 30 % des 15-24 ans ont un emploi, avec de fortes variations de ce taux d'emploi selon l'âge détaillé.

$$\text{On veut donc : } \begin{cases} 24_m^- P_{24^-} + 24_m^+ P_{24^+} = 1 \\ P_{24^-} = 3P_{24^+} \end{cases} \Leftrightarrow Kish_{m,i} = \frac{I_{m,i}}{3 \times 24_m^- + 24_m^+} \text{ avec :}$$

$I_{m,i} = 3$ si l'individu i du ménage répondant m est un jeune en emploi de [16-24 ans[; 1 si l'individu i du ménage répondant m est une personne en emploi d'au moins 24 ans.

24_m^- (24_m^+) = nombre actifs occupés de moins de 24 ans (d'au moins 24 ans) en emploi dans le ménage m .

P_{24^-} (P_{24^+}) = probabilité de sélection d'un actif occupé de moins de 24 ans (d'au moins 24 ans) en emploi dans le ménage m .

Ainsi, une fois le tableau des habitants du logement rempli, on attribue une probabilité de tirage à chaque individu du ménage interrogé.

3.1.2.4. Calcul des probabilités de réponse d'un individu sélectionné sachant qu'il appartient à un ménage répondant

Les 2 460 individus sélectionnés par le tirage Kish ont répondu, au moins partiellement, à l'enquête. Toutefois, six d'entre eux ont fourni trop peu de réponses et sont ainsi considérés comme non-répondant. La probabilité de réponse d'un individu sélectionné sachant qu'il

appartient à un ménage répondant est donc $P_{i/m} = \frac{2454}{2460} \approx 1$.

Tableau 13 - Répartition des pondérations de l'échantillon complémentaire

Quantile	$1/\pi_m$	$1/P_m$	$1/Kish_{m,i}$	w_{ech^C}
Max	20 872	2,15	8,00	65 932
99%	20 872	2,15	5,00	31 958
95%	6 466	1,58	2,00	18 685
90%	6 466	1,58	2,00	15 982
75%	6 466	1,39	2,00	14 224
50%	6 466	1,24	1,67	7 989
25%	1 782	1,24	1,00	3 990
10%	1 782	1,11	1,00	2 298
5%	1 782	1,11	1,00	2 202
1%	1 782	1,03	1,00	1 961
Min	1 782	1,03	1,00	1 833

Après traitement, 98 % des pondérations de l'échantillon complémentaire se situent entre 1 961 et 31 958. Il y a 103 observations dans les cinq derniers percentiles de la population.

Tableau 14 - Répartition des cinq derniers percentiles des pondérations de l'échantillon complémentaire

Percentile	$W_{ech^c}^{TNR}$
100,0%	65 932
99,5%	44 863
99,0%	31 958
98,5%	28 028
98,0%	27 795
97,5%	25 012
97,0%	22 472
96,5%	20 407
96,0%	20 407
95,5%	18 685
95,0%	18 685

Dans l'échantillon complémentaire, 23 individus âgés de moins de 60 ans appartiennent par erreur à la strate des « sénior » (ménage composé exclusivement des plus de 60 ans). Ils possèdent des poids traités de la non-réponse élevés compris entre 22 959 et 65 932. Pour les corriger, on les remplace par le poids moyen du GRH auquel ils appartiennent. Les $W_{ech^c,i}^{TNR}$ de ces 23 individus sont alors comprises entre 7 581 et 15 958.

Les autres pondérations extrêmes ne sont pas modifiées. Il s'agit d'individus dont les $1/\pi_m$, $1/P_{m,sel}$ et $1/Kish_{m,i}$ sont plus élevés que pour le reste des individus.

Tableau 15 - Traitement des valeurs extrêmes sur les pondérations de l'échantillon complémentaire

Quantile	$W_{ech^c}^{TNR}$	
	Avant traitement des poids extrêmes	Après traitement des poids extrêmes
Max	65 932	65 932
99%	31 958	27 795
95%	18 685	18 256
90%	15 982	15 979
75%	14 224	13 897
50%	7 989	7 989
25%	3 990	3 990
10%	2 298	2 298
5%	2 202	2 202
1%	1 961	1 961
Min	1 833	1 833

3.2. Partage des poids

Le partage des poids intervient dès lors que les individus peuvent être présents dans différents échantillons alimentant la même enquête. On tient ainsi compte du nombre de fois où un individu aurait pu être sélectionné.

3.2.1. Rappels théoriques

La méthode du partage des poids permet de tenir compte des liens existants entre deux populations. Ce point s'appuie sur une présentation faite par Pascal Ardilly et David Le Blanc (cf. bibliographie).

On dispose d'une population U de n unités, et d'une population V de m unités. On suppose qu'il existe des liens entre les unités des deux populations. On note $r_{ji} = 1$ si l'unité i de V est reliée à l'unité j de U ; $r_{ji} = 0$ sinon.

U représente la population dans laquelle on échantillonne initialement, et V la population dans laquelle on considère l'échantillon déduit de l'échantillon initial par le système de liens.

Toutes les unités de V ont au moins un lien avec une unité de U . Soit Y le total d'une variable d'intérêt y sur V . On a : $Y = \sum_{i \in V} y_i$.

On note $r_i = \sum_{j \in U} r_{ji}$.

On a alors $Y = \sum_{i \in V} y_i = \sum_{j \in U} \sum_{i \in V} \frac{r_{ji}}{r_i} y_i$. Si l'on définit pour tout j de U la variable $z_j = \sum_{i \in V} \frac{r_{ji}}{r_i} y_i$

on a $Z = \sum_{j \in U} z_j = \sum_{j \in U} \sum_{i \in V} \frac{r_{ji}}{r_i} y_i = \sum_{i \in V} y_i = Y$.

Soit s_U un échantillon issu de la population U , auquel est associé un jeu de poids w_j $j \in s_U$.

Par le système de liens, cet échantillon définit un échantillon s_V dans V tel que : $s_V = \{i \in V; \exists j \in s_U, r_{ji} = 1\}$. On suppose que l'on a collecté les r_{ji} pour tous les $i \in s_V$, c'est-à-dire que tous les liens des individus de s_V avec l'univers U sont connus.

Le total $Z = Y$ est estimé par $\hat{Z} = \sum_{j \in s_U} w_j z_j$. Si les poids sont sans biais, \hat{Z} estime sans biais Y .

On a donc $\hat{Z} = \sum_{j \in s_U} w_j \sum_{i \in V} \frac{r_{ji}}{r_i} y_i = \hat{Y}$.

En posant $\tilde{w}_i = \frac{1}{r_i} \sum_{j \in s_U} w_j r_{ji}$, on a $\hat{Y} = \sum_{i \in V} \tilde{w}_i y_i = \sum_{i \in s_V} \tilde{w}_i y_i$ (car $\tilde{w}_i = 0$ si $i \in V - s_V$)

3.2.2. Application dans SIP

La partie *RPS* est composé des actifs occupés du *Panel* et des répondants de l'*ech^C* ; les répondants de l'*ech^C* étant tous, par définition, des actifs occupés. Il convient alors de prendre en compte les liens existants entre ces deux sous-populations.

La partie *RPS* contient des jeunes actifs occupés de moins de 24 ans (24^-) et des actifs occupés de 24 ans et plus (24^+). Les 24^- ne peuvent appartenir qu'à l'*ech^C*, car les individus du *Panel* ont au moins 24 ans en 2010 (car ils avaient entre 20 et 74 ans en 2006). Cependant, les 24^+ peuvent provenir du *Panel* comme de l'*ech^C*. Suite à une procédure d'élimination des doublons (cf. tirage de l'échantillon complémentaire dans l'Annexe I), aucun individu n'a été tiré à la fois dans le *Panel* et dans l'*ech^C*, si bien qu'on a :

$$\tilde{W}_{ech^C,i}^{TNR} = W_{ech^C,i}^{TNR} \times I_{24}$$

avec $I_{24} = 1$ si l'enquêté est un jeune actif occupé de moins de 24 ans ; $I_{24} = \frac{1}{2}$ si l'enquêté est un actif occupé de 24 ans et plus.

Tableau 16 - Répartition des pondérations de l'échantillon complémentaire après partage des poids

Quantile	$W_{ech^C}^{TNR}$	$\tilde{W}_{ech^C}^{TNR}$
Max	65 932	32 966
99%	27 795	14 533
95%	18 256	9 343
90%	15 979	8 337
75%	13 897	7 158
50%	7 989	3 995
25%	3 990	2 202
10%	2 298	1 197
5%	2 202	1 101
1%	1 961	987
Min	1 833	917

Après partage des poids, 98 % des pondérations de l'échantillon complémentaire se situent entre 987 et 14 533.

La population d'inférence de la partie *RPS* étant les actifs occupés, on peut également observer la distribution des pondérations des actifs occupés issus de la partie *Panel*, avant ($W_{Panel,i}^{TNR,C}$) et après partage des poids ($\tilde{W}_{Panel,i}^{TNR,C}$). Les individus du *Panel* ont tous entre 24 et 78 ans en 2010. Or, les individus âgés de 24 ans et plus peuvent provenir en théorie du *Panel* comme de l'*ech^C* (il y a donc toujours deux liens, soit $r_i = 2$). En réalité, du fait de la procédure d'élimination des doublons mentionnée supra, les échantillons *Panel* et *ech^C* sont par construction totalement disjoints (pour faire la liaison avec la partie 3.2.1, cela signifie

que la sommation définissant le poids \tilde{w}_i ne comprend qu'un seul terme) si bien que l'application du partage des poids conduit toujours à :

$$\tilde{W}_{Panel,i}^{TNR,C} = \frac{1}{2} \times W_{Panel,i}^{TNR,C}$$

On a ainsi $W_{RPS,i}^{TNR} = \tilde{W}_{ech^C,i}^{TNR} \times I_{ech^C,i} + \tilde{W}_{Panel,i}^{TNR,C} \times I_{Panel,i}$ avec $I_{ech^C} \cup I_{Panel} = 1$ si l'individu appartient à l'échantillon complémentaire (*Panel*); 0 sinon.

Tableau 17: répartition des pondérations des actifs occupés de la partie *Panel* après partage des poids

Quantile	W_{Panel}^{TNR}	\tilde{W}_{Panel}^{TNR}
Max	29 319	14 660
99%	13 177	6 588
95%	8 447	4 224
90%	6 801	3 401
75%	4 808	2 404
50%	3 175	1 588
25%	2 317	1 159
10%	1 752	876
5%	1 444	722
1%	1 065	532
Min	705	352

Au final, les pondérations de la partie *RPS* sont distribuées de la façon suivante :

Tableau 18 - Répartition des pondérations de la partie *RPS* après partage des poids

Quantile	W_{RPS}^{TNR}
Max	32 966
99%	10 212
95%	7 989
90%	5 506
75%	3 276
50%	1 890
25%	1 269
10%	987
5%	759
1%	550
Min	352

Après traitement de la non-réponse et partage des poids, 98 % des pondérations de la partie *RPS* se situent entre 550 et 10 212.

3.3. Calage

Il est souhaitable d'associer à la partie *RPS* une pondération qui permette d'estimer parfaitement bien certaines structures connues sur la population complète des actifs occupés ayant entre 16 ans et 78 ans en 2010 (notion de "représentativité" d'un plan de sondage). Ainsi, le calage se fera sur la moyenne des quatre enquêtes Emploi trimestrielles relatives à l'année 2010. L'échantillon complémentaire (*ech^C*), qui est l'un des éléments de la partie *RPS*, ne contient pas de données sur le diplôme et la nationalité de l'enquêté. De plus, la tranche d'unité urbaine des actifs occupés du *Panel* n'a pas été actualisée lors de la seconde vague de l'enquête. Ainsi, elle ne tient pas compte des mouvements des enquêtés entre 2006 et 2010 ; la tranche d'unité urbaine ne sera donc pas utilisée pour le calage de la partie *RPS*.

Ainsi, les variables de calage retenues seront différentes de celles utilisées pour la partie *Panel*. Le calage s'effectuera finalement sur des variables collectées à la fois sur le *Panel* et sur l'*ech^C* lors de la seconde vague (cf. tableau 19). Les sorties SAS fournies par la macro CalMar sont disponibles en Annexe VI.

Le calage ainsi effectué porte sur quatre variables qui conduisent à 29 modalités au total. Ici encore, on utilisera la fonction de calage logit tronquée de la macro CalMar (cf. point 2.3.2).

Tableau 19 - Liste des variables utilisées pour le calage de la partie *RPS*

Variable	Zone d'études et d'aménagement du territoire	Classe d'âge au 31 décembre 2010 par sexe	Catégorie socioprofessionnelle	Nombre d'habitants du logement
Modalités	1. Île-de-France, Bassin parisien	1-16. Homme de 16 à 19 ans	0. Non renseigné	1. Une personne
	2. Nord, Est, Ouest	1-20. Homme de 20 à 29 ans	1. Agriculteurs exploitants	2. Deux personnes
	3. Sud-ouest, Centre-est, Méditerranée	1-30. Homme de 30 à 39 ans	2. Artisans, commerçants, et chefs d'entreprise	3. Trois personnes
		1-40. Homme de 40 à 49 ans	3. Cadres et professions intellectuelles supérieures	4. Quatre ou cinq personnes
		1-50. Homme de 50 à 59 ans	4. Professions intermédiaires	5. Au moins six personnes
		1-60. Homme de 60 à 65 ans	5. Employés	
		1-66. Homme de 66 à 78 ans	6. Ouvriers	
		2-16. Femme de 16 à 19 ans		
		2-20. Femme de 20 à 29 ans		
		2-30. Femme de 30 à 39 ans		
		2-40. Femme de 40 à 49 ans		
		2-50. Femme de 50 à 59 ans		
		2-60. Femme de 60 à 65 ans		
		2-66. Femme de 66 à 78 ans		

Tableau 20 - Répartition des poids et des rapports de poids selon la fonction de calage utilisée

Quantile	Logit tronquée		Raking ratio		Linéaire tronquée		Linéaire	
	Lo=0,1	Up=1,5	Poids	Rapport des poids	Lo=0,1	Up=1,5	Poids	Rapport des poids
Max	40 536	1,47	39 847	1,77	40 128	1,50	40 122	1,66
99%	12 479	1,41	12 507	1,52	12 454	1,49	12 432	1,48
95%	8 508	1,34	8 418	1,36	8 436	1,35	8 439	1,35
90%	6 193	1,29	6 306	1,29	6 288	1,28	6 282	1,28
75%	3 632	1,18	3 626	1,16	3 633	1,16	3 632	1,16
50%	1 962	1,06	1 971	1,04	1 966	1,05	1 968	1,05
25%	1 271	0,95	1 274	0,95	1 273	0,95	1 273	0,95
10%	944	0,83	962	0,86	960	0,85	961	0,85
5%	781	0,78	790	0,83	785	0,81	785	0,81
1%	555	0,68	552	0,75	554	0,71	553	0,71
Min	216	0,11	238	0,10	226	0,10	-18	-0,01

Tableau 21 - Répartition des poids et des rapports de poids finaux de la partie RPS

Quantile	W_{RPS}^{TNR}	$W_{RPS}^{TNR.C}$	Rapport des poids
Max	32 966	40 536	1,47
99%	10 212	12 479	1,41
95%	7 989	8 508	1,34
90%	5 506	6 193	1,29
75%	3 276	3 632	1,18
50%	1 890	1 962	1,06
25%	1 269	1 271	0,95
10%	987	944	0,83
5%	759	781	0,78
1%	550	555	0,68
Min	352	216	0,11

Au final, après redressements et calage, 98% des pondérations de la partie *RPS* se situent entre 555 et 12 479.

Encadré 3 - Pondérations du questionnaire auto-administré.

Parmi les répondants de la partie *Panel* (respectivement *RPS*), seuls 94% (respectivement 97%) ont répondu au questionnaire auto-administré. Une correction de la non-réponse pour chaque partie, *Panel* et *RPS*, est effectuée à l'aide de la méthode CHAID, à partir des variables de 2010 concernant le sexe, l'âge en classes et la santé altérée. Ce sont ces jeux de pondérations qu'il faut utiliser lorsque l'on mobilise des variables provenant du questionnaire auto-administré.

Quantile	Pondérations du questionnaire auto-administré	
	Partie <i>Panel</i>	Partie <i>RPS</i>
Max	27 294	45 406
99%	14 171	13 503
95%	8 885	8 771
90%	6 899	6 335
75%	4 737	3 705
50%	3 164	2 013
25%	2 286	1 305
10%	1 715	976
5%	1 385	804
1%	1 013	563
Min	593	225

3.4. Exemple d'utilisation des pondérations de la partie *RPS* pour le calcul d'une moyenne

Pour une variable d'intérêt Y , on estimera sa moyenne (\bar{Y}) par une moyenne pondérée sur un ensemble s_n composé de n éléments appartenant à la partie *RPS* par :

$$\bar{Y}_{RPS} = \frac{\sum_{i \in s_n} W_{RPS,i}^{TNR,C} Y_i}{\sum_{i \in s_n} W_{RPS,i}^{TNR,C}}$$

Bibliographie

Ardilly P., « Les techniques de sondage », *Technip*, 2006.

Ardilly P., Le Blanc D., « Échantillonnage et pondération d'une enquête auprès de personnes sans domicile : un exemple français », *Techniques d'enquête*, vol. 27, n° 1, 2001.

Bahu M., Coutrot T., Mermilliod C., Rouxel C., « Appréhender les interactions entre la santé et la vie professionnelle et leur éventuel décalage temporel, premier bilan d'une enquête innovante : SIP », *Document de travail, Série Sources et méthodes*, n° 35, septembre 2012.

Bahu M., Coutrot T., Herbet J.-B., Mermilliod C., Rouxel C., « Parcours professionnels et état de santé », *Drees, Dossiers solidarité santé*, n°14, 2010.

Bodier M., Gollac M. (sous la direction de), « Mesurer les facteurs psychosociaux de risque au travail pour les maîtriser », *Rapport du Collège d'expertise sur le suivi des risques psychosociaux au travail*, 2011.

Brisebois F., « Enquêtes répétées dans le temps », *Notes de cours FCDA*, 2011.

Buisson B., Neiter B., « Comment redresser une enquête thématique ? », *Insee, Document de travail*, n° E2010/01, 2010.

Caron N., « La correction de la non-réponse par repondération et par imputation. », *Insee, Document de travail*, n°M0502, 2005.

Caron N., Rousseau S., « Correction de la non réponse et calage de l'enquête santé 2002 », *Insee, Document de travail*, n° M0501, 2005.

Christofari M.F, « Bilan des sources quantitatives dans le champ de la santé et de l'itinéraire professionnel », *CEE*, 2003.

Haziza D., « Traitement de la non-réponse », *Notes de cours FCDA*, 2008.

Mermilliod C., « L'impact de la mise en place d'un suivi d'adresse entre les deux vagues de l'enquête Santé et Itinéraire Professionnel (SIP) », *Documents de travail, série Sources et méthodes*, n° 37, septembre 2012.

Sautory O., « Redressement d'échantillon et méthode de calage », *Notes de cours FCDA*, 2011.

Sautory O., « Redressement d'un échantillon par calage sur marges », *Insee, Document de travail*, n° F9313, 1993.

Tuffery S., « Étude de cas en statistique décisionnelle », *Technip*, 2009.

Tuffery S., « Data mining et statistique décisionnelle: L'intelligence des bases de données », *Technip*, 2005.

Annexe I - Les plans de sondages de l'enquête SIP

Le plan de sondage de la partie longitudinale

L'échantillon résulte d'un tirage aléatoire d'adresses de ménages dans l'échantillon maître issu du recensement de la population de 1999 (RP99). On a cherché à représenter la population cible en utilisant des probabilités de tirage par individu qui soient aussi peu variables que possible. Pour cela, les résidences principales ont été tirées dans l'échantillon maître issu du RP99 avec une probabilité de sélection qui augmente avec la taille de la population du champ au moment du recensement, notée X . On a donc calculé, pour chaque résidence principale au RP99, le nombre de personnes par ménage qui ont entre 13 ans et 67 ans en mars 1999 (pour cibler les 20-74 ans en 2006), et affecté les probabilités de tirage suivantes :

Résidence principale au RP99, $X = 0$ ou 1	f
Résidence principale au RP99, $X = 2$ ou 3	$2 \times f$
Résidence principale au RP99, $X \geq 4$	$3 \times f$
Résidence secondaire, ou vacante rurale	$f/2$

Un tirage à taux uniforme f de logements achevés entre mars 1999 et août 2004 a complété l'échantillon. Le changement de statut des logements ainsi que le changement de composition des ménages depuis mars 1999 sont à l'origine d'une inévitable dispersion des poids des individus in fine tirés.

Une fois le ménage contacté, l'individu à enquêter était tiré dans le tableau de composition des ménages en prenant le premier prénom par ordre alphabétique sous réserve de l'application du critère d'âge. Le tirage des ménages visait à assurer une représentation équiprobable de la population du champ.

Tirage de l'échantillon complémentaire

L'échantillon complémentaire SIP a été sélectionné, à partir des bases de la Taxe d'habitation 2009, dans les 349 UP de l'échantillon-maître 1999.

Un premier tirage de 238 408 logements a eu lieu dans les UP de façon à obtenir un poids uniforme de 119,4828.

Ces 238 408 logements ont été stratifiés selon le statut d'occupation (vacant, résidence secondaire, résidence principale) et, au sein des résidences principales, selon la catégorie du ménage (au moins une personne de moins de 24 ans au 01/01/2010, uniquement constitué de personnes de plus de 60 ans au 01/01/2011, autres types de ménages).

Un taux de sondage spécifique à chaque strate a été appliqué en tirant, par sécurité, un échantillon double de celui qui a été effectivement enquêté.

Strates de ménages		Taux sondage arrondi
Résidence principale	Ménages dont l'un des membres est né entre le 01/01/1987 et le 31/12/1992 inclus	1/7,46
	Ménages ne comportant que des plus de 60 ans (tous les membres sont nés avant le 01/01/1951).	1/87,34
	Autres types de ménages	1/27,05
Résidences secondaires		1/26,26
Logements vacants		1/10,32

Une procédure d'élimination des doublons a fait au final baisser la somme des poids de 28 485 656 à 27 084 361.

Au final la répartition par catégorie de logement est la suivante :

Strates de ménages		Taille
Résidence principale	Ménages dont l'un des membres est né entre le 01/01/1987 et le 31/12/1992 inclus	1 720
	Ménages ne comportant que des plus de 60 ans (tous les membres sont nés avant le 01/01/1951).	430
	Autres types de ménages	2 150
Résidences secondaires		70
Logements vacants		285
<i>Ensemble</i>		4 665 ¹¹

¹¹ Lors de la collecte, 4 659 logements ont été enquêtés. Les quatre logements supplémentaires proviennent de logements éclatés : deux logements supplémentaires parmi la strate des ménages de résidences principales comprenant au moins un jeune de moins de 24 ans, et deux autres parmi la strate des logements vacants.

Annexe II - Pondération de la première vague de l'enquête SIP

Le traitement initial effectué par la Dares en 2006

Il a été supposé que tous les individus avaient la même probabilité de répondre ; un traitement homogène de la non-réponse a donc été mis en œuvre.

Le traitement de la non-réponse accompli, le calage a été effectué selon la méthode linéaire tronquée (macro SAS Calmar). On obtient alors les pondérations finales, mises en œuvre.

Les marges retenues sont : le sexe*âge, âge*activité, la tranche d'unité urbaine en cinq postes, le nombre d'habitants en trois tranches, la catégorie socioprofessionnelle en six postes, le secteur d'activité à un niveau agrégé (agriculture, industrie, BTP, commerce, services, administration).

Elles ont été calculées à partir des moyennes des quatre fichiers de l'enquête Emploi 2006.

Les rapports de poids obtenus sont bornés et se situent dans l'intervalle [0,7 ; 2,0]. Ainsi, 98 % des pondérations sont comprises entre 866 et 9 250, autour d'une pondération médiane égale à 2 421, et 90% d'entre elles entre 1 175 et 6 038.

Les tests réalisés après calage confirment que les volumes et les distributions obtenus après calage de l'enquête SIP 2006 sont cohérents avec les totaux issus de l'enquête Emploi 2006, pour :

- les actifs occupés, les chômeurs et les inactifs, par sexe ;
- les actifs occupés, les chômeurs et les inactifs, par tranche d'âge ;
- les actifs occupés par grande catégorie socioprofessionnelle ;
- l'ensemble de la population selon l'âge et le sexe ;
- les actifs occupés par grand secteur d'activité ;
- grandes catégories socioprofessionnelles par grands secteurs d'activité.

Les corrections apportées en 2010

En 2006, 328 individus (2,3 % des effectifs pondérés) parmi 13 991 ont été sortis des bases d'études, car ils présentaient des itinéraires professionnels incohérents ou mal renseignés ; on parle alors d'« itinéraires incomplets ». Ces « itinéraires incomplets » ont été exclus des analyses effectuées, et n'ont pas été ré-interrogés lors de la seconde vague. Toutefois, cet apurement a été effectué une fois les pondérations mise en place : un poids leur avait donc été déjà attribué. Il convient donc de corriger les pondérations initiales de la première vague de l'enquête $\left(V_0^{TNR,C} \right)$, en traitant ces 328 individus comme des non-répondants.

De plus, lors de l'apurement des bases, 15 répondants de 2006 doivent être considérés comme non-répondants (itinéraire incomplet non repéré initialement, âge hors des limites définies, etc.). Au final, 343 individus répondants en 2006 seront, a posteriori, traités comme non-répondants.

On effectue la correction suivante : $W_{v1,i}^{TNR,C} = W_0^{TNR,C} \times \frac{\sum_{i=1}^{13991} W_0^{TNR,C}}{\sum_{i=1}^{13991} W_0^{TNR,C} \times I_{complet}}$ avec $I_{complet} = 1$ si l'individu fait parti des 13 648 « itinéraires complets »; 0 sinon. Les $W_0^{TNR,C}$ seront ainsi multipliés par $\frac{1}{97,5\%}$.

Tableau 22 - Répartition des pondérations de la première vague entre itinéraires complets/incomplets

Quantile	Pondérations des 343 « itinéraires incomplets »	Pondérations des 13 648 « itinéraires complets »	$W_0^{TNR,C}$	$W_{v1}^{TNR,C}$
Max	10 673	21 326	21 326	21 863
99%	8 884	9 254	9 251	9 483
95%	6 090	6 033	6 038	6 072
90%	5 100	4 952	4 955	4 984
75%	3 189	3 286	3 282	3 285
50%	2 429	2 421	2 421	2 479
25%	1 973	1 887	1 887	1 923
10%	1 450	1 300	1 301	1 336
5%	1 206	1 170	1 176	1 210
1%	804	867	867	891
Min	622	622	622	638

Annexe III - L'algorithme CHAID utilisé par la macro TreeDisc

Ce point reprend les explications de l'algorithme CHAID présentées par Stéphane Tufféry dans son ouvrage « Data mining et statistique décisionnelle: L'intelligence des bases de données ». La macro SAS TreeDisc s'appuie sur cet algorithme.

L'algorithme CHAID utilise le test du χ^2 pour définir la variable la plus significative de chaque nœud (un nœud est une configuration associée à un certain regroupement de modalités). Il doit être utilisé avec des variables discrètes ou qualitatives. Le test du χ^2 est utilisé dans les étapes successives de division de chaque nœud, les étapes 1 à 4 étant les étapes de fusion des modalités des variables explicatives, et l'étape 5 étant une étape de scission du nœud. Ces étapes sont appliquées de façon itératives, jusqu'à ce qu'une condition d'arrêt soit rencontrée. L'effectif du nœud à scinder doit être au moins égal à une valeur seuil initialement prescrite (dans notre cas, 50 individus).

1 – Pour chaque variable explicative X ayant au moins trois modalités, le χ^2 est utilisé pour regrouper les modalités de X en les croisant avec les k modalités de la variable à expliquer (dans notre cas, $k = 2$: réponse / non-réponse). Après balayage, on commence par sélectionner la paire admissible¹² de modalités de X dont le sous-tableau $2 \times k$ correspondant à la plus petite valeur associée au χ^2 . Si ce χ^2 n'est pas significatif au seuil choisi (0,0001 par défaut), on fusionne les deux modalités, et on considère le résultat de cette fusion comme une nouvelle modalité composée.

2 – On répète l'étape 1 jusqu'à ce que toutes les paires de modalités (simples ou composées) aient un χ^2 significatif, ou jusqu'à ce qu'il n'y ait plus de modalités distinctes. Si l'une de ces modalités a un effectif inférieur au minimum prescrit lors du paramétrage de l'arbre, cette modalité sera fusionnée avec la modalité la plus proche en terme de χ^2 , même si ce χ^2 était déjà significatif.

3 – Si la variable explicative est nominale et présente des valeurs manquantes, l'ensemble des valeurs manquantes est considéré comme une modalité qui est traitée comme les autres. Si, au contraire, la variable est ordinale ou quantitative, la modalité des valeurs manquantes ne prend pas part aux traitements de fusions évoqués précédemment. Ce n'est qu'a posteriori que CHAID tente de la fusionner avec une autre, celle qui est la plus proche en terme de χ^2 .

4 – Afin d'éviter la surévaluation de la significativité des variables à modalités multiples, on multiplie les p-values associées aux différentes modalités par un coefficient correctif dit de Bonferroni. Ce coefficient est le nombre de possibilités de regrouper les m modalités d'une variable explicative en g groupes ($2 \leq g \leq m$).

5 – Après avoir regroupé les modalités de façon optimale pour chaque variable explicative, l'algorithme retient la variable dont la valeur associée au χ^2 est la plus élevée. Si la p-value

¹² Une paire admissible est une paire adjacente si X est ordinale ou quantitative (rendue qualitative, donc ordinale), n'importe quelle paire si X est nominale.

associée est inférieure au seuil choisi, on peut diviser le nœud en autant de nœud-fils que la variable a de modalités après regroupement. Si ce χ^2 n'atteint pas le seuil spécifié, le nœud n'est pas divisé.

Annexe IV - La correction de la non-réponse du Panel par une régression logistique

À partir des mêmes variables utilisées pour la construction des GRH, on établit, pour chaque répondant de la première vague, sa probabilité estimée de réponse pour la seconde vague. Cette estimation se fait à l'aide d'une régression logistique. La variable relative au sexe de l'enquête n'est pas significative dans cette régression.

Les pondérations traitées de la non-réponse par cette méthode sont proches de celles fournies par la méthode CHAID.

Tableau 23 - Comparaison des $W_{Panel,i}^{TNR}$ selon la méthode utilisée

Quantile	$W_{Panel,i}^{TNR}$	
	CHAID	Régression logistique
Max	29 319	28 721
99%	12 381	12 407
95%	7 763	7 917
90%	6 222	6 187
75%	4 157	4 185
50%	2 901	2 877
25%	2 321	2 328
10%	1 579	1 611
5%	1 386	1 382
1%	1 057	1 045
Min	705	687

The LOGISTIC Procedure

Model Information	
Data Set	WORK.TEMP
Response Variable	Repondant_Panel
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	13316
Number of Observations Used	13316

Response Profile		
Ordered Value	Repondant_Panel	Total Frequency
1	1	11016
2	0	2300

Probability modeled is Repondant_Panel='1'.

Model Convergence Status	
Convergence criterion (GCONV=1E-8) satisfied.	

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	12257.517	11783.922
SC	12265.014	11963.843
-2 Log L	12255.517	11735.922

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	519.5955	23	<.0001
Score	551.0944	23	<.0001
Wald	514.9459	23	<.0001

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
2.3797	1	0.1229

Summary of Stepwise Selection							
Step	Effect		DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
	Entered	Removed					
1	tu		4	1	237.5732		<.0001
2	CIAge		4	2	167.3266		<.0001
3	fnivdip		8	3	90.6137		<.0001
4	fsitua		6	4	43.2196		<.0001
5	SanteAlt		1	5	13.5302		0.0002

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
SanteAlt	1	13.5050	0.0002
fnivdip	8	59.1897	<.0001
CIAge	4	148.0862	<.0001
tu	4	220.2485	<.0001
fsitua	6	34.4032	<.0001

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	1.1948	0.0879	184.7115	<.0001
SanteAlt	0 1	0.2028	0.0552	13.5050	0.0002
fnivdip	. 1	-0.1135	0.1493	0.5784	0.4470
fnivdip	1 1	-0.4054	0.0852	22.6198	<.0001
fnivdip	2 1	-0.0882	0.0925	0.9096	0.3402
fnivdip	3 1	-0.1215	0.0986	1.5175	0.2180
fnivdip	5 1	0.0957	0.0969	0.9759	0.3232
fnivdip	6 1	0.00479	0.0954	0.0025	0.9599
fnivdip	7 1	0.2279	0.0903	6.3688	0.0116
fnivdip	8 1	0.3044	0.0814	13.9680	0.0002
CIAge	2 1	-0.4102	0.0807	25.8192	<.0001
CIAge	4 1	0.3072	0.0732	17.6155	<.0001
CIAge	5 1	0.5764	0.0767	56.5162	<.0001
CIAge	6 1	0.4094	0.1220	11.2617	0.0008
tu	0 1	0.3749	0.0667	31.6358	<.0001
tu	1 1	0.2774	0.0740	14.0553	0.0002
tu	2 1	0.1653	0.0790	4.3759	0.0365
tu	4 1	-0.6401	0.0681	88.4394	<.0001
fsitua	2 1	0.2705	0.3917	0.4769	0.4898
fsitua	3 1	-0.0142	0.1831	0.0060	0.9381
fsitua	4 1	-0.3606	0.0857	17.6889	<.0001
fsitua	5 1	0.0818	0.1158	0.4998	0.4796
fsitua	6 1	-0.2186	0.0958	5.2039	0.0225
fsitua	7 1	-0.4885	0.1383	12.4839	0.0004

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
SanteAlt 0 vs 1	1.225	1.099	1.365
fnivdip . vs 4	0.893	0.666	1.196
fnivdip 1 vs 4	0.667	0.564	0.788
fnivdip 2 vs 4	0.916	0.764	1.098
fnivdip 3 vs 4	0.886	0.730	1.074
fnivdip 5 vs 4	1.100	0.910	1.331
fnivdip 6 vs 4	1.005	0.833	1.211
fnivdip 7 vs 4	1.256	1.052	1.499
fnivdip 8 vs 4	1.356	1.156	1.590
CIAge 2 vs 3	0.663	0.566	0.777
CIAge 4 vs 3	1.360	1.178	1.569
CIAge 5 vs 3	1.780	1.531	2.068
CIAge 6 vs 3	1.506	1.186	1.913
tu 0 vs 3	1.455	1.277	1.658
tu 1 vs 3	1.320	1.142	1.526
tu 2 vs 3	1.180	1.010	1.377
tu 4 vs 3	0.527	0.461	0.602
fsitua 2 vs 1	1.311	0.608	2.824
fsitua 3 vs 1	0.986	0.689	1.412
fsitua 4 vs 1	0.697	0.589	0.825
fsitua 5 vs 1	1.085	0.865	1.362
fsitua 6 vs 1	0.804	0.666	0.970
fsitua 7 vs 1	0.614	0.468	0.805

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	63.6	Somers' D	0.284
Percent Discordant	35.2	Gamma	0.288
Percent Tied	1.3	Tau-a	0.081
Pairs	25336800	c	0.642

Annexe V - Les sorties SAS de CalMar pour la partie Panel

```

*****
*** Paramètres de la macro ***
*****

```

Table en entrée	DATA	=	MATABLE
Pondération initiale	POIDS	=	W01_10_GRH_CHAID
Pondération Qk	PONDQK	=	UN
Identifiant	IDENT	=	IDENT_IND
Table des marges	DATAMAR	=	MARGE_PANEL
Marges en pourcentages	PCT	=	NON
Effectif de la population	EFFPOP	=	
Méthode utilisée	M	=	3
Borne inférieure	LO	=	0.63
Borne supérieure	UP	=	1.80
Seuil d'arrêt	SEUIL	=	0.0001
Nombre maximum d'itérations	MAXITER	=	15
Table contenant la pond. finale	DATAPOI	=	SORTIE_PANEL
Mise à jour de la table DATAPOI	MISAJOUR	=	OUI
Pondération finale	POIDSFIN	=	PONDANEL
Label de la pondération finale	LABELPOI	=	Pondération de la partie Panel
Contenu de la table DATAPOI	CONTPOI	=	OUI
Edition des poids	EDITPOI	=	NON
Statistiques sur les poids	STAT	=	OUI
Contrôles	CONT	=	OUI
Table contenant les obs. éliminées	OBSELI	=	OUI
Notes SAS	NOTES	=	OUI

**Comparaison entre les marges tirées de l'échantillon (avec la pondération initiale)
et les marges dans la population (marges du calage)**

Variable	Modalité ou variable	Marge échantillon	Marge population	Pourcentage échantillon	Pourcentage population
CLASEXE	01	3697804.51	3711123.44	9.21	9.14
	02	4029144.50	4077002.48	10.04	10.05
	03	4070877.61	4200074.89	10.14	10.35
	04	4073276.91	4029634.37	10.15	9.93
	05	3670772.33	3717119.27	9.15	9.16
	06	3694175.44	3809763.75	9.20	9.39
	07	4308547.92	4178529.61	10.73	10.30
	08	4366106.09	4402272.57	10.88	10.85
	09	4130064.64	4213894.52	10.29	10.38
	10	4095466.55	4244167.37	10.20	10.46
DIP	1	9500270.52	11049448.36	23.67	27.23
	2	14030731.29	12904821.72	34.96	31.80
	3	10637320.24	11453925.76	26.50	28.22
	4	5967914.45	5175386.43	14.87	12.75
HABMEN	1	5662421.03	6237244.70	14.11	15.37
	2	13826142.89	14053279.29	34.45	34.63
	3	8101014.27	8110132.66	20.18	19.98
	4	11371205.74	11016586.66	28.33	27.15
	5	1175452.56	1166338.96	2.93	2.87
NAT	1	37830715.81	38212086.61	94.26	94.16
	2	2305520.68	2371495.66	5.74	5.84
TU	1	10567467.23	10535789.30	26.33	25.96
	2	6792171.95	6799285.14	16.92	16.75
	3	4982027.26	7341184.03	12.41	18.09
	4	11364616.61	9197197.63	28.32	22.66
	5	6429953.43	6710126.17	16.02	16.53
ZEAT3	1	14299748.46	14706667.47	35.63	36.24
	2	12462462.63	11563444.27	31.05	28.49
	3	13374025.41	14313470.53	33.32	35.27

Méthode : logit, inf=0.63, sup=1.80

Premier tableau récapitulatif de l'algorithme :

la valeur du critère d'arrêt et le nombre de poids négatifs après chaque itération

Itération	Critère d'arrêt	Poids négatifs
1	0.73043	0
2	0.09183	0
3	0.00703	0
4	0.00010	0
5	0.00000	0

Méthode : logit, inf=0.63, sup=1.80

Deuxième tableau récapitulatif de l'algorithme :

les coefficients du vecteur lambda de multiplicateurs de Lagrange après chaque itération

Variable	Modalité	lambda1	lambda2	lambda3	lambda4	lambda5
CLASEXEF	01	-0.06029	-0.09809	-0.10206	-0.10211	-0.10211
CLASEXEF	02	-0.01284	-0.04798	-0.05205	-0.05211	-0.05211
CLASEXEF	03	0.01588	-0.01247	-0.01589	-0.01593	-0.01593
CLASEXEF	04	-0.08773	-0.13823	-0.14389	-0.14395	-0.14395
CLASEXEF	05	-0.13193	-0.19309	-0.19993	-0.20001	-0.20001
CLASEXEF	06	-0.05102	-0.09238	-0.09681	-0.09686	-0.09686
CLASEXEF	07	-0.04706	-0.08913	-0.09386	-0.09392	-0.09392
CLASEXEF	08	-0.02518	-0.06010	-0.06408	-0.06413	-0.06413
CLASEXEF	09	-0.07894	-0.12678	-0.13214	-0.13220	-0.13220
CLASEXEF	10	-0.12010	-0.17624	-0.18251	-0.18259	-0.18259
DIP	1	0.31886	0.38192	0.38901	0.38909	0.38909
DIP	2	0.04147	0.05903	0.06160	0.06164	0.06164
DIP	3	0.20329	0.25405	0.25958	0.25964	0.25964
DIP	4
HABMEN	1	0.14392	0.17785	0.18155	0.18160	0.18160
HABMEN	2	0.07636	0.09802	0.10040	0.10042	0.10042
HABMEN	3	0.05257	0.07110	0.07317	0.07320	0.07320
HABMEN	4	0.01269	0.02154	0.02264	0.02265	0.02265
HABMEN	5
NAT	1	0.03891	0.04820	0.04913	0.04914	0.04914
NAT	2
TU	1	-0.09112	-0.10658	-0.10809	-0.10811	-0.10811
TU	2	-0.08809	-0.10267	-0.10409	-0.10410	-0.10410
TU	3	0.37607	0.36202	0.36308	0.36309	0.36309
TU	4	-0.30492	-0.38676	-0.39725	-0.39739	-0.39739
TU	5
ZEAT3	1	-0.13029	-0.15920	-0.16202	-0.16205	-0.16205
ZEAT3	2	-0.14863	-0.18413	-0.18782	-0.18786	-0.18786
ZEAT3	3

Méthode : logit, inf=0.63, sup=1.80
Comparaison entre les marges finales dans l'échantillon (avec la pondération finale)
et les marges dans la population (marges du calage)

Variable	Modalité ou variable	Marge échantillon	Marge population	Pourcentage échantillon	Pourcentage population
CLASEXE	01	3711123.44	3711123.44	9.14	9.14
	02	4077002.48	4077002.48	10.05	10.05
	03	4200074.89	4200074.89	10.35	10.35
	04	4029634.37	4029634.37	9.93	9.93
	05	3717119.27	3717119.27	9.16	9.16
	06	3809763.75	3809763.75	9.39	9.39
	07	4178529.61	4178529.61	10.30	10.30
	08	4402272.57	4402272.57	10.85	10.85
	09	4213894.52	4213894.52	10.38	10.38
	10	4244167.37	4244167.37	10.46	10.46
DIP	1	11049448.36	11049448.36	27.23	27.23
	2	12904821.72	12904821.72	31.80	31.80
	3	11453925.76	11453925.76	28.22	28.22
	4	5175386.43	5175386.43	12.75	12.75
HABMEN	1	6237244.70	6237244.70	15.37	15.37
	2	14053279.29	14053279.29	34.63	34.63
	3	8110132.66	8110132.66	19.98	19.98
	4	11016586.66	11016586.66	27.15	27.15
	5	1166338.96	1166338.96	2.87	2.87
NAT	1	38212086.61	38212086.61	94.16	94.16
	2	2371495.66	2371495.66	5.84	5.84
TU	1	10535789.30	10535789.30	25.96	25.96
	2	6799285.14	6799285.14	16.75	16.75
	3	7341184.03	7341184.03	18.09	18.09
	4	9197197.63	9197197.63	22.66	22.66
	5	6710126.17	6710126.17	16.53	16.53
ZEAT3	1	14706667.47	14706667.47	36.24	36.24
	2	11563444.27	11563444.27	28.49	28.49
	3	14313470.53	14313470.53	35.27	35.27

Méthode : logit, inf=0.63, sup=1.80
 Statistiques sur les rapports de poids (= pondérations finales / pondérations initiales)
 et sur les pondérations finales

The UNIVARIATE Procedure
 Variable: _f_ (Rapport de poids)

Moments			
N	11220	Sum Weights	11220
Mean	1.01469477	Sum Observations	11384.8753
Std Deviation	0.24897486	Variance	0.06198848
Skewness	0.88923494	Kurtosis	0.06709672
Uncorrected SS	12247.6221	Corrected SS	695.448788
Coeff Variation	24.5369221	Std Error Mean	0.00235049

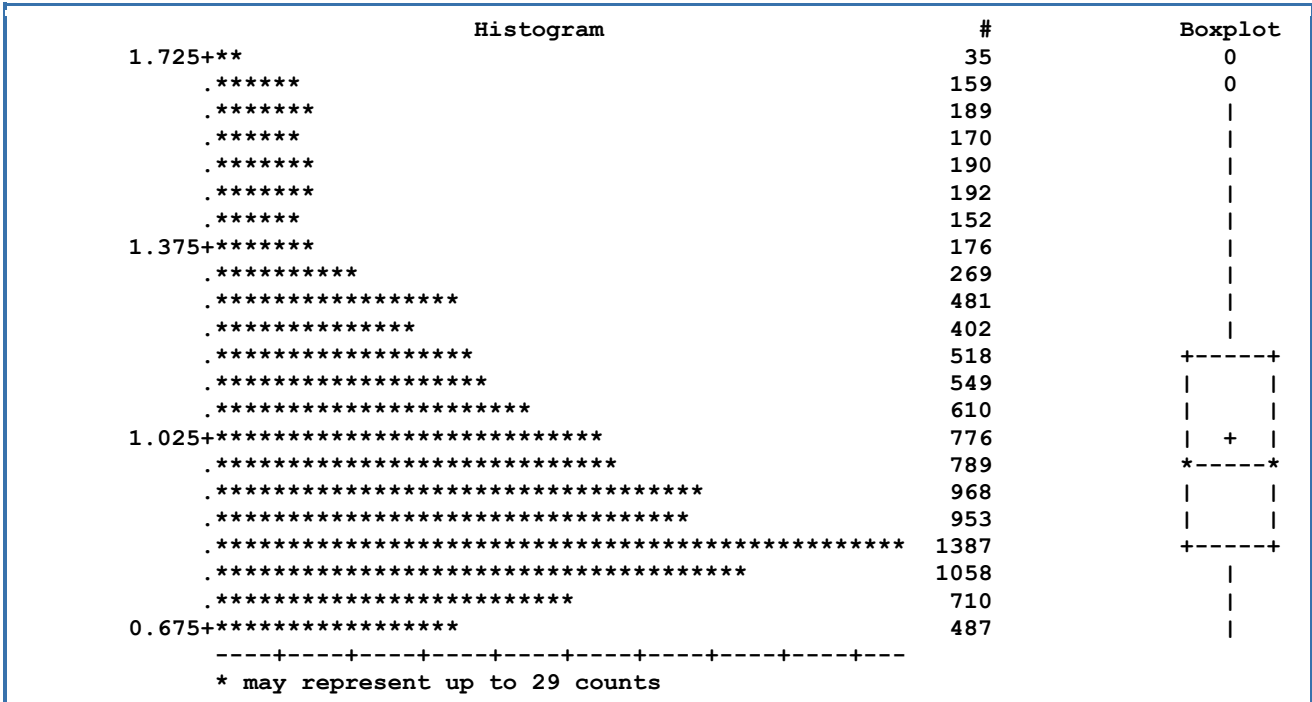
Basic Statistical Measures			
Location		Variability	
Mean	1.014695	Std Deviation	0.24897
Median	0.956950	Variance	0.06199
Mode	1.275910	Range	1.08697
		Interquartile Range	0.34354

Tests for Location: Mu0=0				
Test		Statistic	p Value	
Student's t	t	431.6943	Pr > t	<.0001
Sign	M	5610	Pr >= M	<.0001
Signed Rank	S	31474905	Pr >= S	<.0001

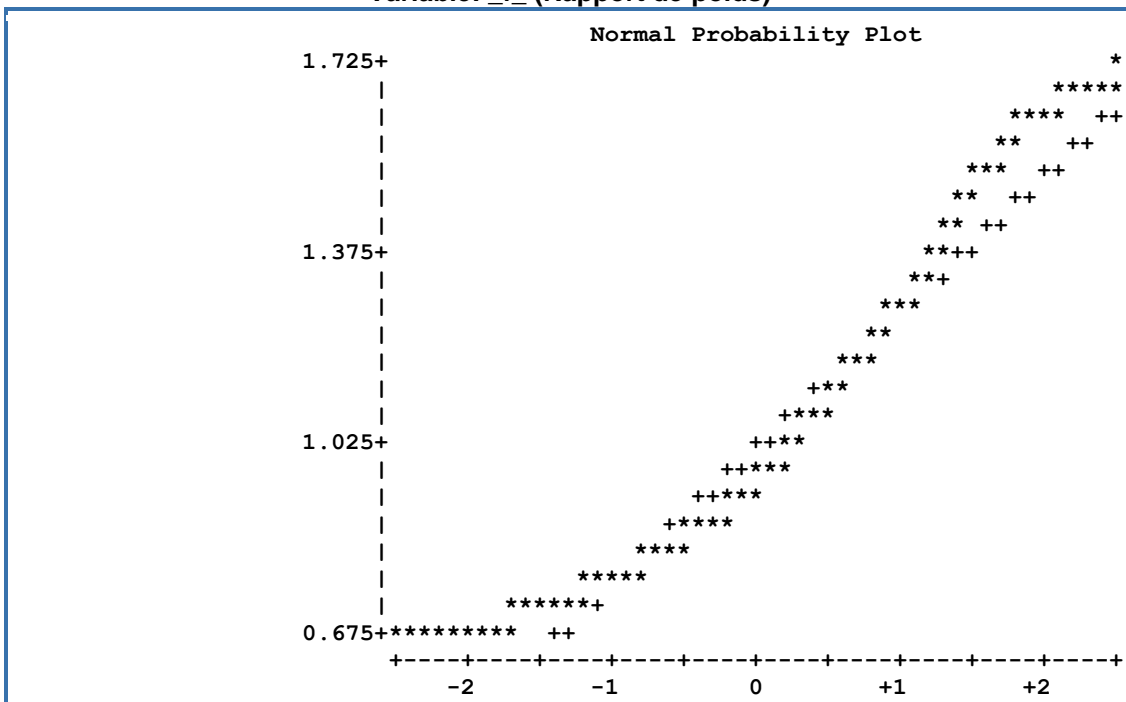
Tests for Normality				
Test		Statistic	p Value	
Kolmogorov-Smirnov	D	0.100847	Pr > D	<0.0100
Cramer-von Mises	W-Sq	39.36061	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	251.5644	Pr > A-Sq	<0.0050

Quantiles (Definition 5)	
Quantile	Estimate
100% Max	1.747151
99%	1.669082
95%	1.547620
90%	1.381557
75% Q3	1.161844
50% Median	0.956950
25% Q1	0.818302
10%	0.743721
5%	0.706216
1%	0.682896
0% Min	0.660182

Extreme Observations					
Lowest			Highest		
Value	IDENT_IND	Obs	Value	IDENT_IND	Obs
0.660182	430065800100001	5311	1.72396	730101990100001	8460
0.661385	520070190100001	5682	1.72845	720089130100001	7346
0.665680	520075290100001	6135	1.73666	830117120100001	9705
0.668911	520071010100001	5754	1.73945	820106020100001	8809
0.668911	520069510100001	5626	1.74715	910122240100001	10121



Variable: f (Rapport de poids)



The UNIVARIATE Procedure
Variable: __wfin (Pondération finale)

Moments			
N	11220	Sum Weights	11220
Mean	3617.07507	Sum Observations	40583582.3
Std Deviation	2460.77585	Variance	6055417.81
Skewness	2.73716471	Kurtosis	12.1367093
Uncorrected SS	2.1473E11	Corrected SS	6.79357E10
Coeff Variation	68.032203	Std Error Mean	23.2314131

Basic Statistical Measures			
Location		Variability	
Mean	3617.075	Std Deviation	2461
Median	2913.976	Variance	6055418
Mode	2753.398	Range	25289
		Interquartile Range	2234

Note: The mode displayed is the smallest of 2 modes with a count of 23.

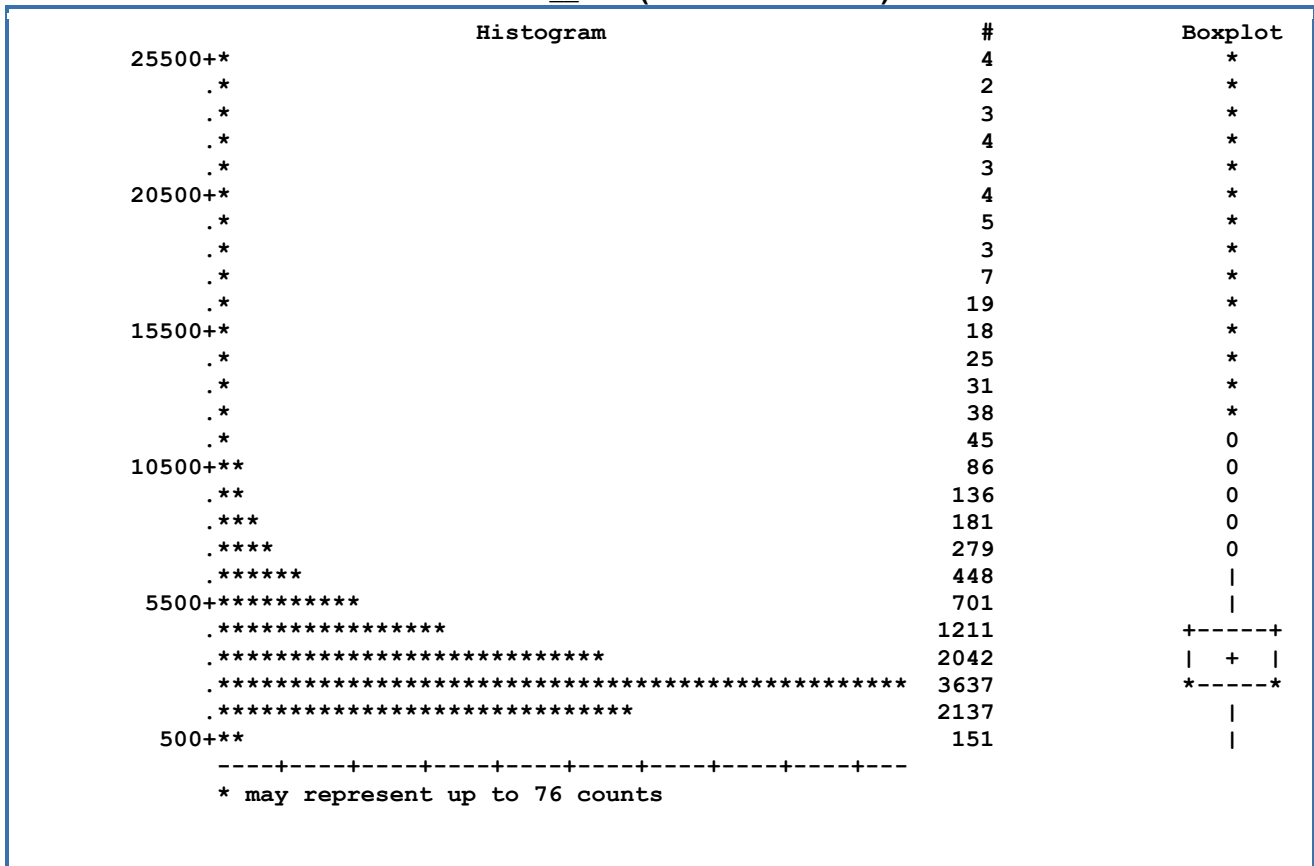
Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	155.6976	Pr > t	<.0001
Sign	M	5610	Pr >= M	<.0001
Signed Rank	S	31474905	Pr >= S	<.0001

Tests for Normality				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.155961	Pr > D	<0.0100
Cramer-von Mises	W-Sq	109.4531	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	615.4301	Pr > A-Sq	<0.0050

Quantiles (Definition 5)	
Quantile	Estimate
100% Max	25860.360
99%	13402.789
95%	8219.272
90%	6442.030
75% Q3	4348.039
50% Median	2913.976
25% Q1	2113.813
10%	1588.740
5%	1280.209
1%	942.516
0% Min	570.928

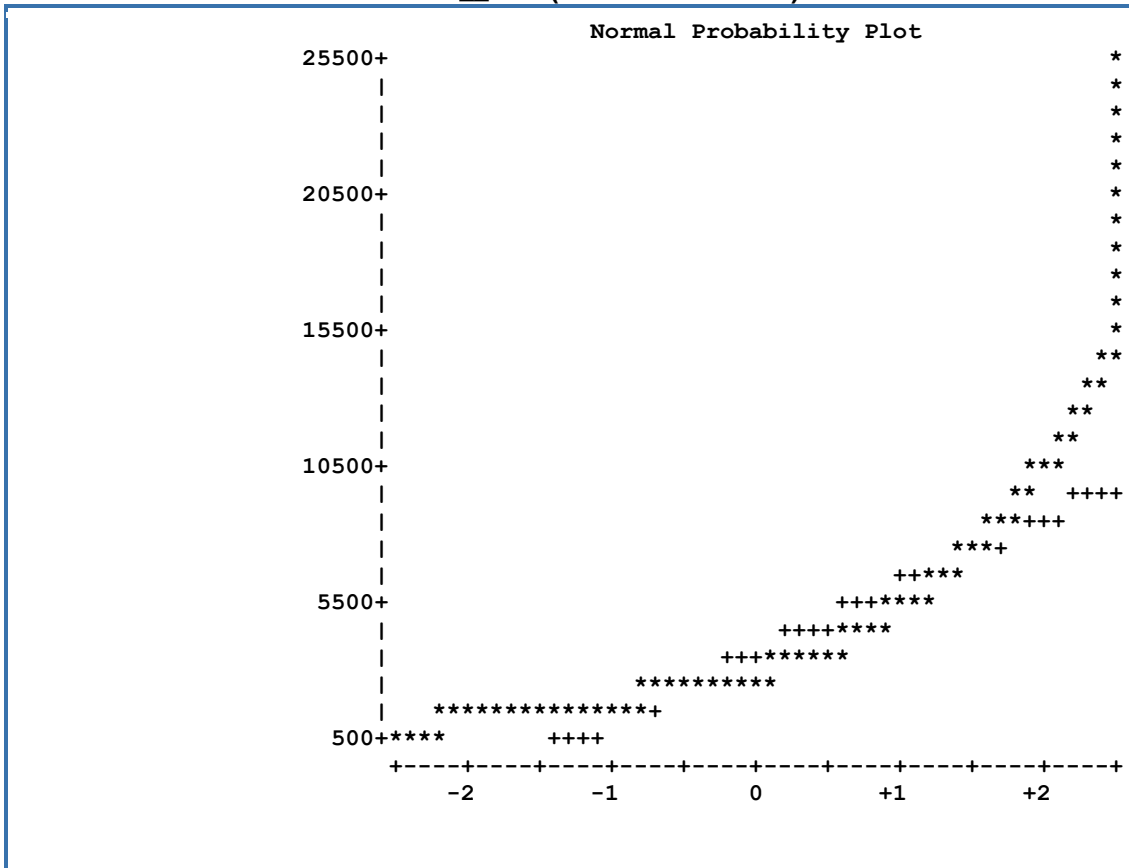
Extreme Observations					
Lowest			Highest		
Value	IDENT_IND	Obs	Value	IDENT_IND	Obs
570.928	240035930100001	2766	24755.0	830117930100001	9776
594.375	310050430100001	4022	25514.7	740103750100001	8621
616.395	530078860100001	6456	25517.1	910121740100001	10079
630.779	540084730100001	6971	25598.6	740103180100001	8567
641.634	520069780100001	5648	25860.4	230028680100001	2176

Variable: __wfin (Pondération finale)



The UNIVARIATE Procedure

Variable: __wfin (Pondération finale)



Data Set Name	WORK.SORTIE_PANEL	Observations	11220
Member Type	DATA	Variables	2
Engine	V9	Indexes	0
Created	lundi 09 janvier 2012 10 h 06	Observation Length	24
Last Modified	lundi 09 janvier 2012 10 h 06	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	WINDOWS_32		
Encoding	wlatin1 Western (Windows)		

Engine/Host Dependent Information	
Data Set Page Size	4096
Number of Data Set Pages	68
First Data Page	1
Max Obs per Page	168
Obs in First Data Page	116
Number of Data Set Repairs	0
File Name	D:\TEMP\SASWork\TD1004\sortie_panel.sas7bdat
Release Created	9.0101M3
Host Created	XP_PRO

Alphabetic List of Variables and Attributes					
#	Variable	Type	Len	Format	Label
1	IDENT_IND	Char	15	\$15.	
2	PondPanel	Num	8		Pondération de la partie Panel

```

*****
***      BILAN      ***
*****

*
*   Date : 09 JANVIER 2012           Heure : 09:55
*
*   Table en entrée : MATABLE
*
*   Nombre d'observations dans la table en entrée : 11220
*   Nombre d'observations éliminées : 0
*   Nombre d'observations conservées : 11220
*
*   Variable de pondération : W01_10_GRH_CHAID
*
*   Nombre de variables catégorielles : 6
*   Liste des variables catégorielles et de leurs nombres de modalités :
*     CLASEXEF (10 ) DIP (4 ) HABMEN (5 ) NAT (2 ) TU (5 ) ZEAT3 (3 )
*   Taille de l'échantillon (pondéré) : 40136236
*   Taille de la population : 40583582
*
*   Méthode utilisée : logit, borne inférieure = 0.63, borne supérieure = 1.80
*   Le calage a été réalisé en 5 itérations
*   Les poids ont été stockés dans la variable PONDANEL de la table SORTIE_PANEL

```

Annexe VI - Les sorties SAS de CalMar pour la partie RPS

```

*****
***   Paramètres de la macro   ***
*****

Table en entrée          DATA      =  MATABLE
Pondération initiale    POIDS      =  W02_10_GRH_CHAID_PARTAGEPOIDS
Pondération Qk          PONDQK     =  UN
Identifiant             IDENT      =  IDENT_IND

Table des marges        DATAMAR     =  MARGE_RPS
Marges en pourcentages  PCT        =  NON
Effectif de la population EFFPOP    =

Méthode utilisée        M          =  3
Borne inférieure        LO          =  0.1
Borne supérieure        UP          =  1.48
Seuil d'arrêt           SEUIL      =  0.0001
Nombre maximum d'itérations MAXITER   =  15

Table contenant la pond. finale DATAPOI  =  SORTIE_RPS
Mise à jour de la table DATAPOI MISAJOUR =  OUI
Pondération finale      POIDSFIN  =  PONDRPS
Label de la pondération finale LABELPOI  =  Pondération de la partie RPS
Contenu de la table DATAPOI CONTPOI  =  OUI

Edition des poids       EDITPOI   =  NON
Statistiques sur les poids STAT      =  OUI

Contrôles               CONT      =  OUI
Table contenant les obs. éliminées OBSELI =  OUI
Notes SAS               NOTES     =  OUI

```

Variable	Modalité ou variable	Marge échantillon	Marge population	Pourcentage échantillon	Pourcentage population
CLASEXE	01	175918.68	207011.55	0.73	0.80
	02	1814638.93	2431281.83	7.50	9.45
	03	3029767.00	3491330.53	12.53	13.57
	04	3582199.88	3766790.97	14.81	14.65
	05	3086561.92	3078094.67	12.76	11.97
	06	478465.52	449567.45	1.98	1.75
	07	55774.86	79459.88	0.23	0.31
	08	77179.15	102935.60	0.32	0.40
	09	1870813.08	2192783.73	7.73	8.53
	10	3103774.61	3068084.88	12.83	11.93
	11	3450219.28	3488283.44	14.27	13.56
	12	3037079.49	2883311.85	12.56	11.21
	13	348441.96	430389.03	1.44	1.67
	14	75730.17	50288.97	0.31	0.20
CSP	1	12103.48	1420.40	0.05	0.01
	2	583112.65	515835.15	2.41	2.01
	3	1430498.99	1710002.71	5.91	6.65
	4	3997386.81	4302761.10	16.53	16.73
	5	6068353.16	6295614.95	25.09	24.48
	6	7038447.17	7428626.33	29.10	28.88
	7	5056662.28	5465353.74	20.91	21.25
HABMEN	1	3277221.64	3766934.22	13.55	14.65
	2	6426063.22	7238493.68	26.57	28.14
	3	5403925.82	5933232.08	22.34	23.07
	4	8359167.19	8139713.97	34.56	31.65
	5	720186.68	641240.43	2.98	2.49
ZEAT3	1	8835483.07	9475820.85	36.53	36.84
	2	7428276.62	7244735.89	30.71	28.17
	3	7922804.85	8999057.64	32.76	34.99

Itération	Critère d'arrêt	Poids négatifs
1	0.81034	0
2	0.14182	0
3	0.05350	0
4	0.01722	0
5	0.00335	0
6	0.00016	0
7	0.00000	0

Variable	Modalité	lambda1	lambda2	lambda3	lambda4	lambda5	lambda6	lambda7
CLASEXE2	01	0.08949	0.12604	0.12728	0.12728	0.12728	0.12728	0.12728
CLASEXE2	02	0.18337	0.30990	0.33502	0.33587	0.33587	0.33587	0.33587
CLASEXE2	03	0.03845	0.06037	0.06052	0.06050	0.06050	0.06050	0.06050
CLASEXE2	04	-0.05167	-0.04563	-0.04612	-0.04614	-0.04614	-0.04614	-0.04614
CLASEXE2	05	-0.16643	-0.17363	-0.17478	-0.17480	-0.17480	-0.17480	-0.17480
CLASEXE2	06	-0.26017	-0.27538	-0.27702	-0.27704	-0.27704	-0.27704	-0.27704
CLASEXE2	07	0.19706	0.41934	0.53503	0.55930	0.56020	0.56020	0.56020
CLASEXE2	08	0.23162	0.36069	0.38392	0.38463	0.38463	0.38463	0.38463
CLASEXE2	09	0.01791	0.03697	0.03684	0.03682	0.03682	0.03682	0.03682
CLASEXE2	10	-0.11902	-0.11888	-0.11963	-0.11965	-0.11965	-0.11965	-0.11965
CLASEXE2	11	-0.10548	-0.10537	-0.10615	-0.10616	-0.10616	-0.10616	-0.10616
CLASEXE2	12	-0.24036	-0.25353	-0.25507	-0.25509	-0.25509	-0.25509	-0.25509
CLASEXE2	13	0.03076	0.06678	0.06855	0.06853	0.06853	0.06853	0.06853
CLASEXE2	14	-0.63856	-0.65871	-0.66205	-0.66208	-0.66208	-0.66208	-0.66208
CSP	1	-0.86513	-1.17912	-1.40869	-1.53325	-1.56403	-1.56559	-1.56560
CSP	2	-0.12658	-0.12455	-0.12441	-0.12441	-0.12441	-0.12441	-0.12441
CSP	3	0.15981	0.19970	0.20253	0.20255	0.20255	0.20255	0.20255
CSP	4	0.01669	0.02223	0.02269	0.02269	0.02269	0.02269	0.02269
CSP	5	-0.02150	-0.02099	-0.02078	-0.02077	-0.02077	-0.02077	-0.02077
CSP	6	0.02393	0.03057	0.03109	0.03109	0.03109	0.03109	0.03109
CSP	7
HABMEN	1	0.27729	0.30835	0.31079	0.31082	0.31082	0.31082	0.31082
HABMEN	2	0.29252	0.32220	0.32436	0.32439	0.32439	0.32439	0.32439
HABMEN	3	0.23676	0.25406	0.25541	0.25543	0.25543	0.25543	0.25543
HABMEN	4	0.10219	0.10649	0.10720	0.10722	0.10722	0.10722	0.10722
HABMEN	5
ZEAT3	1	-0.06572	-0.07746	-0.07809	-0.07809	-0.07809	-0.07809	-0.07809
ZEAT3	2	-0.16970	-0.19185	-0.19313	-0.19314	-0.19314	-0.19314	-0.19314
ZEAT3	3

Méthode : logit, inf=0.1, sup=1.48
Comparaison entre les marges finales dans l'échantillon (avec la pondération finale)
et les marges dans la population (marges du calage)

Variable	Modalité ou variable	Marge échantillon	Marge population	Pourcentage échantillon	Pourcentage population
CLASEXE	01	207011.55	207011.55	0.80	0.80
	02	2431281.83	2431281.83	9.45	9.45
	03	3491330.53	3491330.53	13.57	13.57
	04	3766790.97	3766790.97	14.65	14.65
	05	3078094.67	3078094.67	11.97	11.97
	06	449567.45	449567.45	1.75	1.75
	07	79459.88	79459.88	0.31	0.31
	08	102935.60	102935.60	0.40	0.40
	09	2192783.73	2192783.73	8.53	8.53
	10	3068084.88	3068084.88	11.93	11.93
	11	3488283.44	3488283.44	13.56	13.56
	12	2883311.85	2883311.85	11.21	11.21
	13	430389.03	430389.03	1.67	1.67
	14	50288.97	50288.97	0.20	0.20
CSP	1	1420.40	1420.40	0.01	0.01
	2	515835.15	515835.15	2.01	2.01
	3	1710002.71	1710002.71	6.65	6.65
	4	4302761.10	4302761.10	16.73	16.73
	5	6295614.95	6295614.95	24.48	24.48
	6	7428626.33	7428626.33	28.88	28.88
	7	5465353.74	5465353.74	21.25	21.25
HABMEN	1	3766934.22	3766934.22	14.65	14.65
	2	7238493.68	7238493.68	28.14	28.14
	3	5933232.08	5933232.08	23.07	23.07
	4	8139713.97	8139713.97	31.65	31.65
	5	641240.43	641240.43	2.49	2.49
ZEAT3	1	9475820.85	9475820.85	36.84	36.84
	2	7244735.89	7244735.89	28.17	28.17
	3	8999057.64	8999057.64	34.99	34.99

Méthode : logit, inf=0.1, sup=1.48
Statistiques sur les rapports de poids (= pondérations finales / pondérations initiales)
et sur les pondérations finales

Variable: _f_ (Rapport de poids)

Moments			
N	8821	Sum Weights	8821
Mean	1.04630728	Sum Observations	9229.47653
Std Deviation	0.16055754	Variance	0.02577872
Skewness	-0.1294147	Kurtosis	0.0498367
Uncorrected SS	9884.23684	Corrected SS	227.368347
Coeff Variation	15.3451614	Std Error Mean	0.00170951

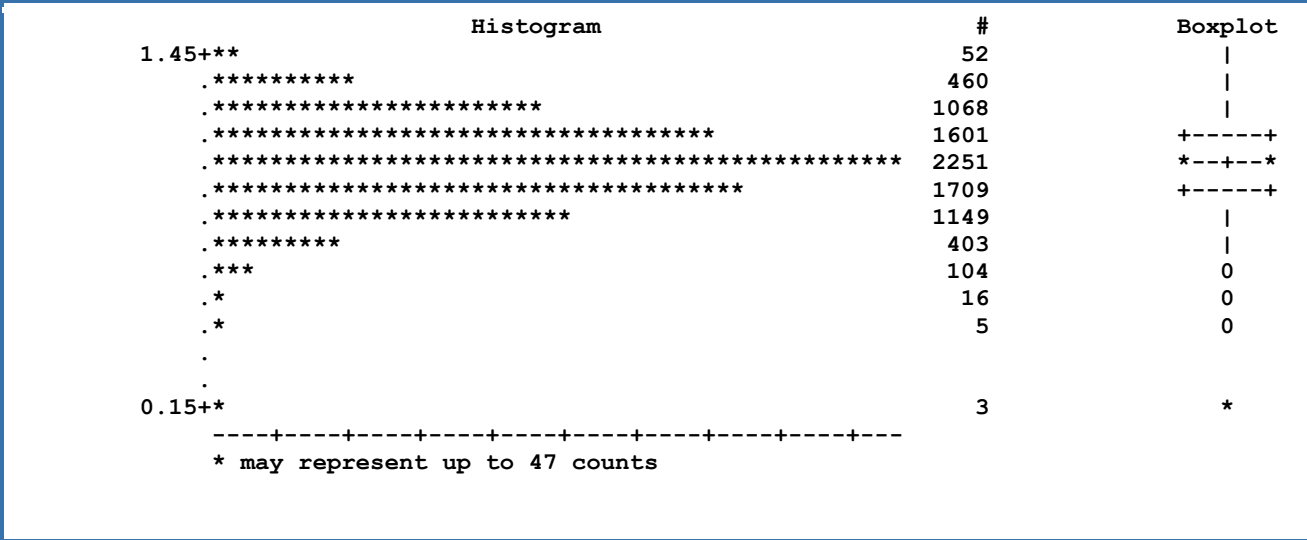
Basic Statistical Measures			
Location		Variability	
Mean	1.046307	Std Deviation	0.16056
Median	1.046098	Variance	0.02578
Mode	0.903605	Range	1.34764
		Interquartile Range	0.21645

Tests for Location: Mu0=0				
Test		Statistic	p Value	
Student's t	t	612.0508	Pr > t	<.0001
Sign	M	4410.5	Pr >= M	<.0001
Signed Rank	S	19454716	Pr >= S	<.0001

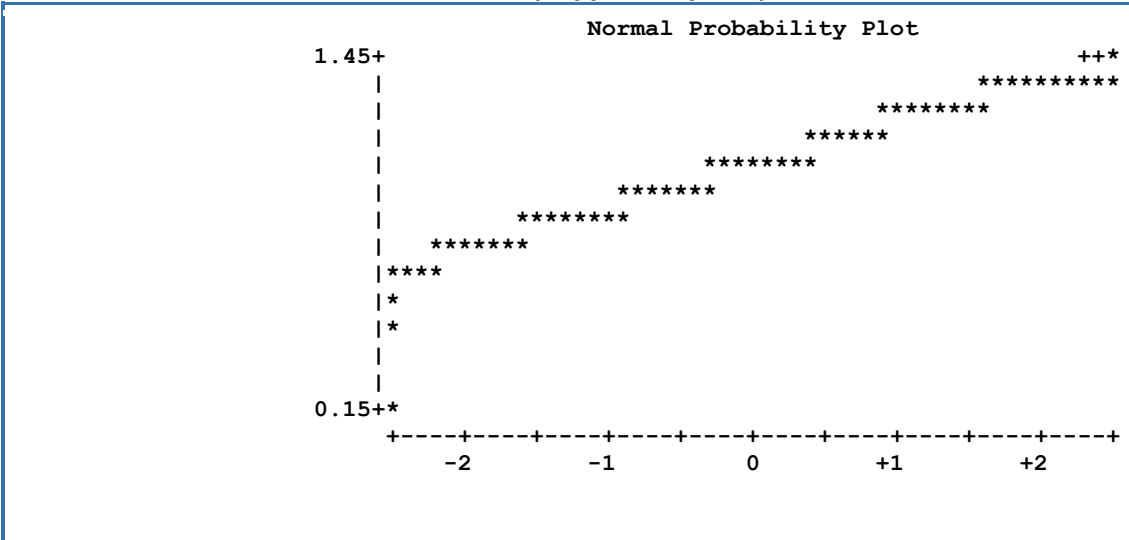
Tests for Normality				
Test		Statistic	p Value	
Kolmogorov-Smirnov	D	0.021368	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.510289	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	3.860143	Pr > A-Sq	<0.0050

Quantiles (Definition 5)	
Quantile	Estimate
100% Max	1.457534
99%	1.392644
95%	1.309234
90%	1.255394
75% Q3	1.160027
50% Median	1.046098
25% Q1	0.943574
10%	0.829392
5%	0.782840
1%	0.666644
0% Min	0.109891

Extreme Observations					
Lowest			Highest		
Value	IDENT_IND	Obs	Value	IDENT_IND	Obs
0.109891	820177380200001	7572	1.44059	720093230100001	6039
0.116297	110007200100001	320	1.44059	930130510100001	8398
0.133165	720097010100001	6235	1.44059	930133450100001	8511
0.425166	520070020100001	4560	1.45753	930126910100001	8248
0.468127	410057860100001	3744	1.45753	930128400100001	8312



The UNIVARIATE Procedure
Variable: _f_ (Rapport de poids)



The UNIVARIATE Procedure
Variable: __wfin (Pondération finale)

Moments			
N	8821	Sum Weights	8821
Mean	2915.72547	Sum Observations	25719614.4
Std Deviation	2834.1929	Variance	8032649.39
Skewness	3.96181984	Kurtosis	31.7638153
Uncorrected SS	1.45839E11	Corrected SS	7.0848E10
Coeff Variation	97.2036952	Std Error Mean	30.1766131

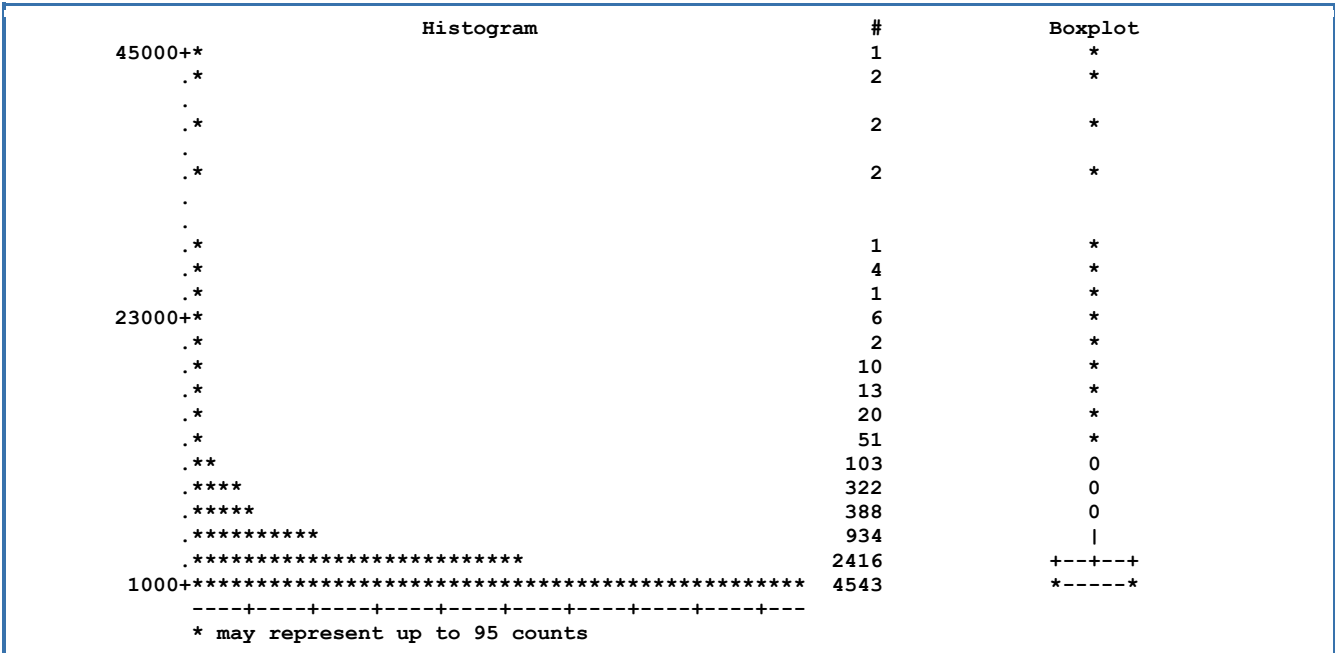
Basic Statistical Measures			
Location		Variability	
Mean	2915.725	Std Deviation	2834
Median	1942.427	Variance	8032649
Mode	1153.152	Range	44007
		Interquartile Range	2329

Tests for Location: Mu0=0				
Test		Statistic	p Value	
Student's t	t	96.62203	Pr > t	<.0001
Sign	M	4410.5	Pr >= M	<.0001
Signed Rank	S	19454716	Pr >= S	<.0001

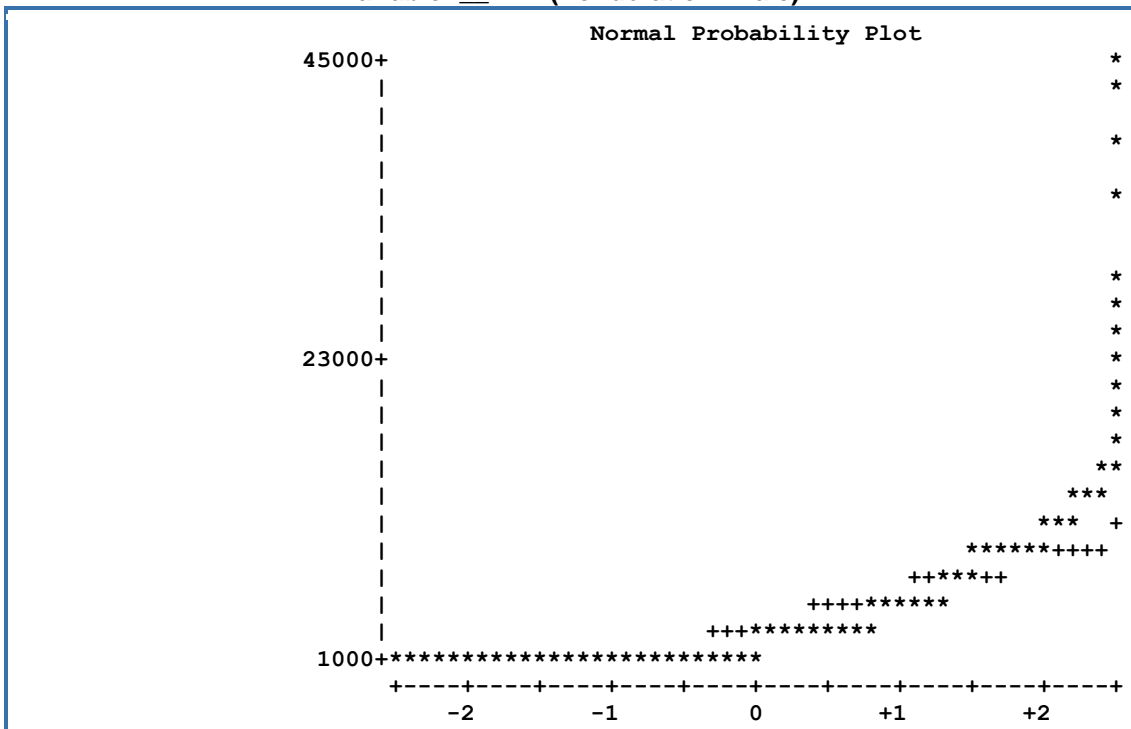
Tests for Normality				
Test		Statistic	p Value	
Kolmogorov-Smirnov	D	0.191839	Pr > D	<0.0100
Cramer-von Mises	W-Sq	132.8765	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	715.797	Pr > A-Sq	<0.0050

Quantiles (Definition 5)	
Quantile	Estimate
100% Max	44221.812
99%	13049.433
95%	8486.121
90%	6170.054
75% Q3	3586.948
50% Median	1942.427
25% Q1	1257.742
10%	937.507
5%	770.257
1%	543.199
0% Min	214.908

Extreme Observations					
Lowest			Highest		
Value	IDENT_IND	Obs	Value	IDENT_IND	Obs
214.908	520070020100001	4560	38774.6	540169230200002	5803
222.120	720097010100001	6235	39706.3	240152590200002	2393
263.096	820177380200001	7572	42546.6	820176910200001	7543
300.463	410057860100001	3744	43543.1	310157860200002	3503
301.275	530079810100001	5226	44221.8	240154380200003	2505



Variable: __wfin (Pondération finale)



Méthode : logit, inf=0.1, sup=1.48

Contenu de la table sortie_RPS contenant la nouvelle pondération PondRPS

The CONTENTS Procedure

Data Set Name	WORK.SORTIE_RPS	Observations	8821
Member Type	DATA	Variables	2
Engine	V9	Indexes	0
Created	lundi 09 janvier 2012 18 h 11	Observation Length	24
Last Modified	lundi 09 janvier 2012 18 h 11	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	WINDOWS_32		
Encoding	wlatin1 Western (Windows)		

Engine/Host Dependent Information	
Data Set Page Size	4096
Number of Data Set Pages	53
First Data Page	1
Max Obs per Page	168
Obs in First Data Page	117
Number of Data Set Repairs	0
File Name	D:\TEMP\SASWork\TD2396\sortie_rps.sas7bdat
Release Created	9.0101M3
Host Created	XP_PRO

Alphabetic List of Variables and Attributes					
#	Variable	Type	Len	Format	Label
1	IDENT_IND	Char	15	\$15.	
2	PondRPS	Num	8		Pondération de la partie RPS

```

*****
***      BILAN      ***
*****

*
*   Date : 09 JANVIER 2012           Heure : 14:48
*
*   Table en entrée : MATABLE
*
*   Nombre d'observations dans la table en entrée   : 8821
*   Nombre d'observations éliminées                 : 0
*   Nombre d'observations conservées                 : 8821
*
*   Variable de pondération : W02_10_GRH_CHAID_PARTAGEPOIDS
*
*   Nombre de variables catégorielles : 4
*   Liste des variables catégorielles et de leurs nombres de modalités :
*     CLASEXE2 (14 ) CSP (7 ) HABMEN (5 ) ZEAT3 (3 )
*   Taille de l'échantillon (pondéré) : 24186565
*   Taille de la population           : 25719614
*
*   Méthode utilisée : logit, borne inférieure = 0.1, borne supérieure = 1.48
*   Le calage a été réalisé en 7 itérations
*   Les poids ont été stockés dans la variable PONDRPS de la table SORTIE_RPS

```


DOCUMENT
De
Travail

Série
Sources et Méthodes

ISSN : 2100-3963 (électronique)

Internet : <http://www.drees.sante.gouv.fr>

Reproduction autorisée sous réserve de la mention des sources

DIRECTION DE LA RECHERCHE, DES ÉTUDES DE L'ÉVALUATION ET DES STATISTIQUES
MINISTÈRE DE L'ÉCONOMIE ET DES FINANCES
MINISTÈRE DES AFFAIRES SOCIALES ET DE LA SANTÉ
MINISTÈRE DU TRAVAIL, DE L'EMPLOI, DE LA FORMATION PROFESSIONNELLE ET DU DIALOGUE SOCIAL