TRAITEMENTS DE LA NON-RÉPONSE ET CALAGES POUR L'ENQUÊTE SANTÉ ET ITINÉRAIRE PROFESSIONNEL DE 2010

Nicolas DE RICCARDIS (*)
(*)Drees - Direction de la recherche, des études, de l'évaluation et des statistiques

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 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 marge 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, cet article 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.

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.

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, 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 III.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 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, 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 marge.

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 les 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.

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-réponse 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 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 marge 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 $W_{v1,i}^{TNR,C}$ 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 cet article, 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 la 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 déclaré (voir tableau 2).

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

[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.

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	Tranche d'unité urbaine	Classe d'âge	Niveau de diplôme à la fin des études initiales	Sexe	Santé altérée
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 supérieur à BAC +2		

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 4)

Il existe plusieurs méthodes pour mettre en place des GRH. On utilisera ici l'algorithme CHAID[2]. 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 ; il le fait de proche en proche par paire ou regroupement de modalités. 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.

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 $\mathrm{I}_{Panel,i}^{h}$ (= 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.

Tableau 3 : 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] ; 90% des poids W_{Panel}^{TNR} sont compris entre 1 386 et 7 763 (contre 1 210 et 6 072 initialement).

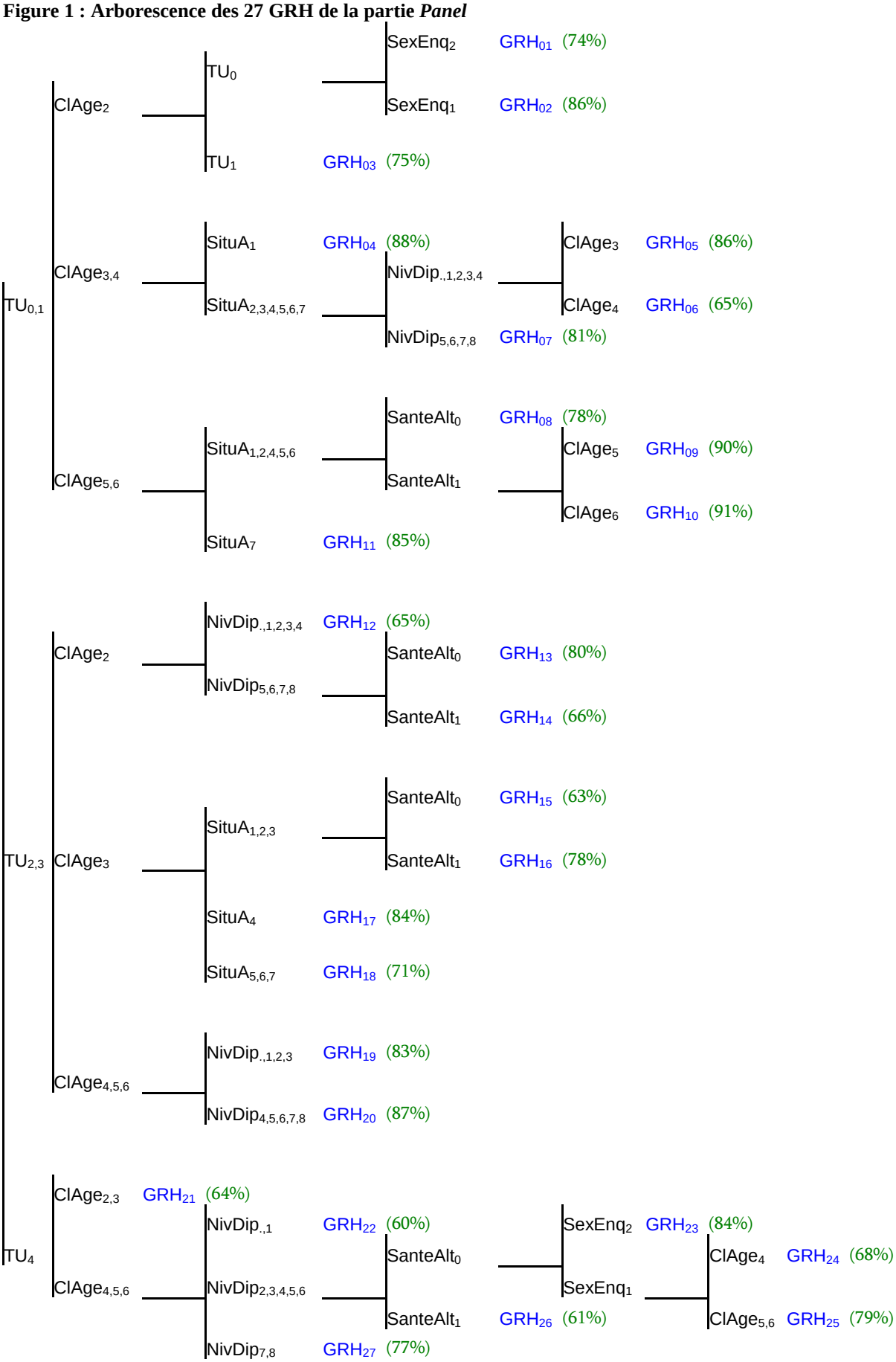
[2] Chi-Square Automated Interaction Detection.

Tableau 4: Détails des 27 GRH de la partie *Panel*

GRH	Variable	Obs.	Tx de rép.	Tx de rép. (pond.)	Moy.	σ	Q1	Q2	Q3	Min	Max	Somme des poids
1	$W_{v1}^{TNR,C}$	240	74%	77%	3 995	2 180	2 592	3 285	4 717	1 057	15 410	958 694
	W_{Panel}^{TNR}	177			5 627	3 109	3 891	4 454	6 776	1 433	20 895	996 051
2	$W_{v1}^{TNR,C}$	196	86%	86%	4 395	2 992	3 037	3 303	5 859	1 199	21 862	861 400
	W_{Panel}^{TNR}	169			5 094	3 379	3 523	3 831	6 795	1 441	25 355	860 959
3	$W_{v1}^{TNR,C}$	127	75%	75%	4 492	3 155	2 380	3 892	5 258	638	17 594	570 483
	W_{Panel}^{TNR}	95			6 050	4 214	3 181	5 203	7 306	950	23 520	574 715
4	$W_{v1}^{TNR,C}$	2 091	88%	88%	3 024	1 753	1 913	2 539	3 615	638	21 543	6 323 889
	W_{Panel}^{TNR}	1 837			3 452	1 992	2 177	2 891	4 147	726	24 522	6 340 698
5	$W_{v1}^{TNR,C}$	118	86%	89%	3 557	1 971	2 287	3 066	4 979	894	12 335	419 745
	W_{Panel}^{TNR}	101			4 344	2 383	2 909	3 582	5 823	1 045	14 412	438 728
6	$W_{v1}^{TNR,C}$	108	65%	67%	3 510	1 715	2 472	3 084	4 464	824	12 335	379 071
	W_{Panel}^{TNR}	70			5 563	2 745	3 813	4 758	7 137	1 577	19 032	389 423
7	$W_{v1}^{TNR,C}$	131	81%	86%	2 903	2 002	1 789	2 665	3 148	888	16 101	380 288
	W_{Panel}^{TNR}	106			3 828	2 654	2 364	3 316	3 891	1 098	19 898	405 780
8	$W_{v1}^{TNR,C}$	110	78%	80%	2 271	1 243	1 497	2 246	2 292	758	9 168	249 760
	W_{Panel}^{TNR}	86			2 966	1 727	1 915	2 873	2 932	970	11 727	255 040
9	$W_{v1}^{TNR,C}$	1 741	90%	90%	2 419	1 181	1 792	2 354	2 482	638	10 931	4 211 634
	W_{Panel}^{TNR}	1 564			2 694	1 308	2 084	2 620	2 763	710	10 973	4 213 104
10	$W_{v1}^{TNR,C}$	432	91%	89%	2 412	1 299	1 818	2 246	2 563	638	15 303	1 041 986
	W_{Panel}^{TNR}	391			2 624	1 265	2 024	2 481	2 818	704	12 250	1 026 080
11	$W_{v1}^{TNR,C}$	644	85%	86%	2 466	1 225	2 354	2 372	2 482	821	14 894	1 588 237
	W_{Panel}^{TNR}	547			2 934	1 520	2 771	2 792	2 922	967	17 535	1 604 742
12	$W_{v1}^{TNR,C}$	459	65%	67%	3 984	2 224	2 955	3 365	4 666	956	20 189	1 828 702
	W_{Panel}^{TNR}	297			6 348	3 515	4 808	5 500	7 379	1 478	29 318	1 885 431
13	$W_{v1}^{TNR,C}$	387	80%	80%	4 294	2 015	3 027	3 954	5 365	956	15 477	1 661 585
	W_{Panel}^{TNR}	311			5 290	2 577	3 680	4 894	6 455	1 190	19 260	1 645 115
14	$W_{v1}^{TNR,C}$	50	66%	68%	4 024	2 267	2 664	3 316	4 939	956	13 645	201 217
	W_{Panel}^{TNR}	33			6 266	3 805	3 525	5 061	7 675	1 449	20 674	206 767

15	$W_{v1}^{TNR,C}$	90	63%	60%	3 255	1 454	2 049	3 146	3 492	767	6 984	292 911
	W_{Panel}^{TNR}	57			4 867	2 317	2 757	4 852	5 514	1 212	11 028	277 437
16	$W_{v1}^{TNR,C}$	86	78%	78%	2 857	1 543	2 097	2 515	3 146	767	9 437	245 745
	W_{Panel}^{TNR}	67			3 684	2 000	2 692	3 182	4 038	1 061	12 113	246 836
17	$W_{v1}^{TNR,C}$	686	84%	84%	3 169	1 615	2 002	2 718	3 826	638	10 941	2 173 654
	W_{Panel}^{TNR}	576			3 781	1 879	2 452	3 251	4 556	805	13 030	2 177 957
18	$W_{v1}^{TNR,C}$	126	71%	70%	2 896	1 850	1 913	2 540	3 361	761	11 990	364 897
	W_{Panel}^{TNR}	90			3 963	2 356	2 678	3 542	4 130	1 066	14 880	356 671
19	$W_{v1}^{TNR,C}$	1 291	83%	83%	2 545	1 321	1 782	2 433	2 577	638	15 779	3 286 109
	W_{Panel}^{TNR}	1 077			3 048	1 590	2 254	2 917	3 059	764	18 915	3 282 258
20	$W_{v1}^{TNR,C}$	2 361	87%	86%	2 579	1 379	1 913	2 416	2 763	638	14 166	6 089 746
	W_{Panel}^{TNR}	2 052			2 946	1 569	2 201	2 759	3 158	734	16 299	6 044 706
21	$W_{v1}^{TNR,C}$	672	64%	65%	3 992	2 071	2 577	3 694	4 655	914	13 661	2 682 446
	W_{Panel}^{TNR}	432			6 258	3 291	4 008	5 707	7 251	1 488	21 250	2 703 553
22	$W_{v1}^{TNR,C}$	141	60%	63%	3 139	1 397	2 058	2 982	3 737	638	8 450	442 635
	W_{Panel}^{TNR}	84			5 597	2 299	3 941	5 046	6 439	2 502	14 184	470 133
23	$W_{v1}^{TNR,C}$	385	84%	85%	3 164	1 532	2 218	3 032	3 595	638	14 480	1 218 191
	W_{Panel}^{TNR}	325			3 766	1 858	2 761	3 592	4 259	755	17 153	1 224 100
24	$W_{v1}^{TNR,C}$	220	68%	67%	3 027	1 303	2 074	2 982	3 234	952	9 277	666 036
	W_{Panel}^{TNR}	150			4 389	1 824	3 023	4 373	4 653	1 396	12 768	658 367
25	$W_{v1}^{TNR,C}$	257	79%	78%	2 985	1 476	2 111	3 055	3 173	785	12 597	767 142
	W_{Panel}^{TNR}	203			3 725	1 867	2 657	3 895	4 016	994	15 948	756 110
26	$W_{v1}^{TNR,C}$	56	61%	57%	3 618	1 490	2 828	3 524	3 887	1 690	8 994	202 612
	W_{Panel}^{TNR}	34			5 622	1 853	4 590	5 804	6 403	3 095	12 817	191 136
27	$W_{v1}^{TNR,C}$	112	77%	74%	3 295	1 485	2 571	2 982	3 485	956	8 945	369 058
	W_{Panel}^{TNR}	86			4 128	1 723	3 499	3 883	4 055	1 246	11 649	354 983

Figure 1 : Arborescence des 27 GRH de la partie *Panel*



<u>Note de lecture</u> : le GRH_{01} est composé des femmes ayant entre 20 et 29 ans qui vivent en commune rurale. Dans ce GRH, 74% des personnes interrogées ont répondu.

2.3. Calage

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 1: principe de la méthode de calage sur marge

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 auxiliaires[3] $X_1 \ldots X_j \ldots X_J$, 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 :
$\forall j = 1 \ldots J \ \sum_{k \in s} w_k^C x_{jk} = X_j$ (1)

La solution de ce problème est donnée par $w_k^C = w_k F(x_k' \lambda)$, où $\mathrm{x}_k' = (\mathrm{x}_{1k} \ldots \mathrm{x}_{Jk})$, 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 » $\hat{Y}^C = \sum_{k \in s} w_k^C y_k$.

[3] Il s'agit de variables quantitatives ou d'indicatrices associées aux modalités de variables catégorielles.

2.3.2. La mise en œuvre du calage

La partie $Panel$ doit posséder des totaux cohérents avec ceux de la population française des 20-74 ans en 2006 (Cf. partie I.2). Ainsi, le calage se fera sur la moyenne des quatre enquêtes Emploi[4] 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 ; elle sert en effet de variable de contrôle.

La macro CalMar dispose de quatre fonctions G différentes pour effectuer un calage sur marge. 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 IV.

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'enquêté	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			

[4] 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 6 : répartition des poids et des rapports de poids selon la fonction de calage utilisée

	Logit tronquée		Raking ratio		Linéaire tronquée		Linéaire	
	Lo=0,63	Up=1,80			Lo=0,57	Up=1,69		
Quantile	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. A l'exception des catégories 0 (non renseigné) et 8 (Autres personnes sans activité professionnelle), et dans une moindre mesure de la catégorie 4, 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 pondérés) n'ont pas de catégorie socioprofessionnelle attribuée.

Tableau 7 : répartition des catégories socioprofessionnelles dans l'échantillon et dans la population.

	Pourcentage dans l'échantillon		Pourcentage dans la population
catégories socioprofessionnelles	avant calage	après calage	
0	2,10%	2,06%	0,01%
1	1,75%	1,76%	1,56%
2	3,66%	3,67%	4,09%
3	10,23%	9,43%	10,16%
4	16,42%	16,25%	15,05%
5	19,88%	20,10%	19,53%
6	15,17%	15,16%	15,67%
7	20,30%	20,69%	20,76%
8	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] ; 90% des poids finaux de la partie $Panel$ $\left(W_{Panel}^{TNR,C}\right)$ sont compris entre 1 280 et 8 219.

Tableau 8 : répartition des poids et des rapports de poids finaux de la partie *Panel*

Quantile	$W_{v1,i}^{TNR,C}$	$W_{Panel,i}^{TNR}$	$W_{Panel,i}^{TNR,C}$	$W_{Panel}^{TNR,C} / W_{Panel}^{TNR}$	$W_{Panel}^{TNR,C} / W_{v1}^{TNR}$
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

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

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

3. Partie transversale : exposition des personnes en 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.

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 Panel	*6 367*

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.

3.1. Pondérations 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'individu[5] 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

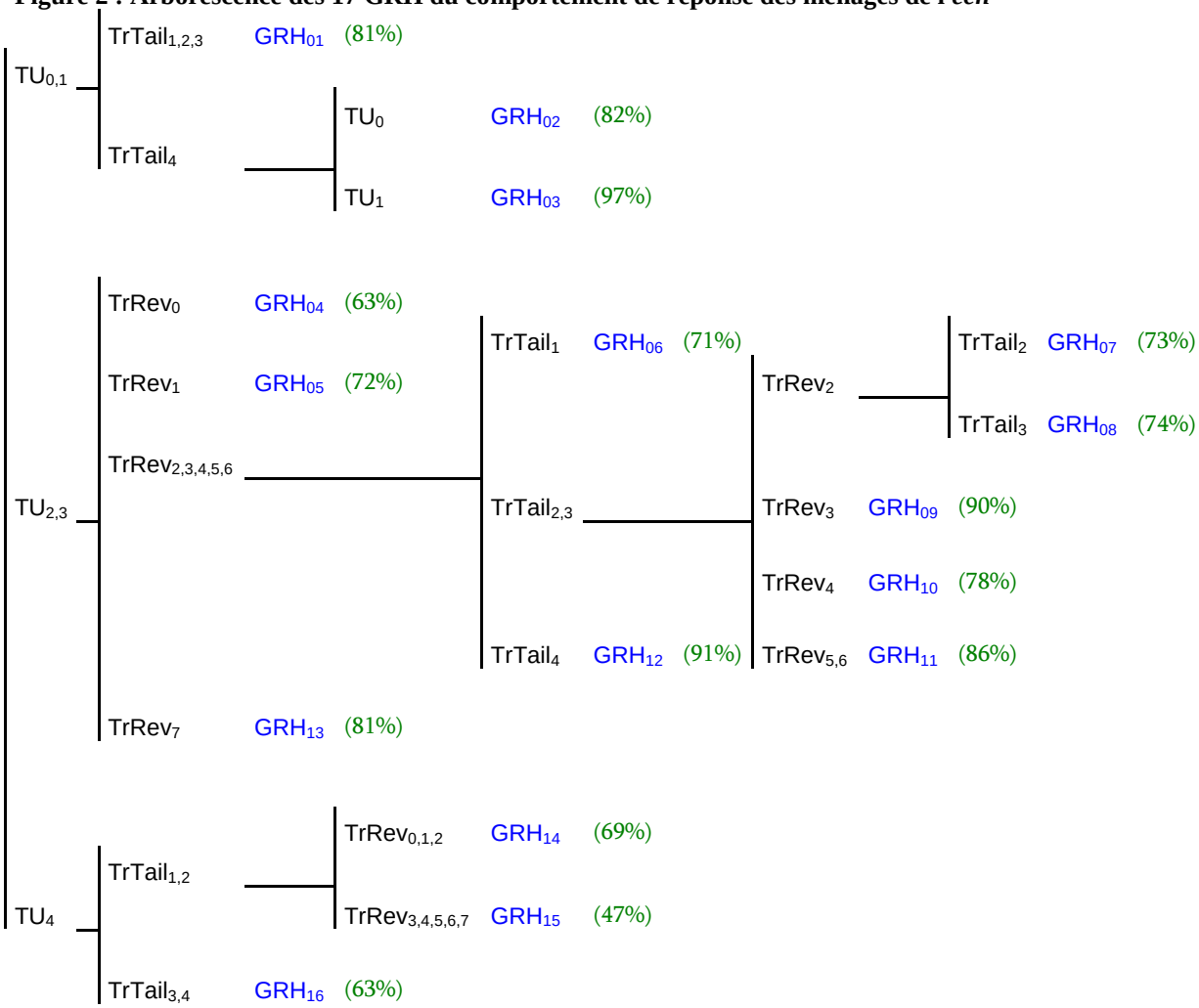
Strates de ménages		Taille dans l'échantillon	$1/\pi_m$
Résidence principale	Ménages dont l'un des membres est né entre le 01/01/1987 et le 31/12/1992 inclus	1 722	1 782
	Ménages ne comportant que des plus de 60 ans (tous les membres sont nés avant le 01/01/1951).	430	20 872
	Autres types de ménages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

[5] 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 III.1.2.3).

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. A partir des données récupérables avec 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 II.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 $\left(P_m^h\right)$ 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.

Figure 2 : Arborescence des 17 GRH du comportement de réponse des ménages de l'*ech*^{C}



Note de lecture : le GRH_{01} est composé des ménages de moins de cinq personnes vivants en commune rurale ou en commune urbaine de moins de 20 000 habitants. Dans ce GRH, 81% des personnes interrogées ont répondu.

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[6] 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

L'algorithme CHAID permet ainsi de déterminer 16 GRH.

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.	σ	Q1	Q2	Q3	Min	Max	Somme des poids
1	$1/\pi_m^{01}$	1 164	81%	80%	4 614	3 066	1 782	6 466	6 466	1 782	20 872	5 370 643
	$1/\left(\pi_m^{01} P_m^{01}\right)$	942			5 644	3 583	2 202	7 989	7 989	2 202	25 791	5 316 971
2	$1/\pi_m^{02}$	114	82%	82%	3 031	2 061	1 782	1 782	6 466	1 782	6 466	345 550
	$1/\left(\pi_m^{02} P_m^{02}\right)$	93			3 753	2 543	2 185	2 185	7 926	2 185	7 926	348 993
3	$1/\pi_m^{03}$	72	97%	97%	3 408	2 245	1 782	1 782	6 466	1 782	6 466	245 409
	$1/\left(\pi_m^{03} P_m^{03}\right)$	70			3 485	2 303	1 833	1 833	6 650	1 833	6 650	243 937
4	$1/\pi_m^{04}$	169	63%	62%	5 165	4 371	1 782	6 466	6 466	1 782	20 872	872 967
	$1/\left(\pi_m^{04} P_m^{04}\right)$	107			8 046	6 608	2 815	10 212	10 212	2 815	32 966	860 969
5	$1/\pi_m^{05}$	337	72%	69%	5 333	3 368	1 782	6 466	6 466	1 782	20 872	1 797 144
	$1/\left(\pi_m^{05} P_m^{05}\right)$	242			7 145	4 177	2 482	9 004	9 004	2 482	29 066	1 729 153
6	$1/\pi_m^{06}$	96	71%	75%	3 540	2 805	1 782	1 782	6 466	1 782	20 872	339 795
	$1/\left(\pi_m^{06} P_m^{06}\right)$	68			5 313	4 310	2 516	2 516	9 128	2 516	29 466	361 307
7	$1/\pi_m^{07}$	102	73%	70%	6 486	3 251	6 466	6 466	6 466	1 782	20 872	661 535
	$1/\left(\pi_m^{07} P_m^{07}\right)$	74			8 566	4 004	8 912	8 912	8 912	2 457	28 769	633 875
8	$1/\pi_m^{08}$	90	74%	69%	2 987	2 051	1 782	1 782	6 466	1 782	6 466	268 810
	$1/\left(\pi_m^{08} P_m^{08}\right)$	67			3 722	2 572	2 394	2 394	2 394	2 394	8 685	249 404

[6] 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	3 006	1 782	1 782	6 466	1 782	20 872	630 278
	$1/\left(\pi_m^{09} P_m^{09}\right)$	140			4 514	3 398	1 973	1 973	7 158	1 973	23 108	631 913
10	$1/\pi_m^{10}$	98	78%	74%	4 285	2 880	1 782	4 467	6 466	1 782	20 872	419 928
	$1/\left(\pi_m^{10} P_m^{10}\right)$	76			5 250	3 029	2 298	2 741	8 337	2 298	8 337	398 995
11	$1/\pi_m^{11}$	95	86%	86%	3 915	2 330	1 782	1 782	6 466	1 782	6 466	371 883
	$1/\left(\pi_m^{11} P_m^{11}\right)$	82			4 532	2 702	2 065	2 065	7 491	2 065	7 491	371 664
12	$1/\pi_m^{12}$	55	91%	91%	5 384	3 755	1 782	6 466	6 466	1 782	20 872	296 125
	$1/\left(\pi_m^{12} P_m^{12}\right)$	50			5 907	4 286	1 961	7 112	7 112	1 961	22 959	295 329
13	$1/\pi_m^{13}$	110	81%	81%	3 749	2 807	1 782	1 782	6 466	1 782	20 872	412 338
	$1/\left(\pi_m^{13} P_m^{13}\right)$	89			4 624	3 602	2 203	2 203	7 991	2 203	25 797	411 512
14	$1/\pi_m^{14}$	302	69%	69%	3 340	2 174	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	3 144	2 575	2 575	9 343	2 575	9 343	1 011 322
15	$1/\pi_m^{15}$	187	47%	46%	6 112	3 637	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	8 236	5 306	13 897	13 897	3 831	44 863	1 118 847
16	$1/\pi_m^{16}$	101	63%	60%	6 550	4 312	6 466	6 466	6 466	1 782	20 872	661 508
	$1/\left(\pi_m^{16} P_m^{16}\right)$	64			9 766	5 909	10 204	10 204	10 204	2 813	32 939	625 040

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 population[7], 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 probabilité inégale. 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).

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

- $\mathrm{I}_{m,i}$ = 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.
- $24_m^-\left(24_m^+\right)$ = nombre actifs occupés de moins de 24 ans (d'au moins 24 ans) en emploi dans le ménage m .
- $P_{24^-}\left(P_{24^+}\right)$ = probabilité de sélection d'un actif occupé de moins de 24 ans (d'au moins 24 ans) en emploi dans le ménage m .

[7] 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.

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 probabilités 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	$1/\pi_m$	$1/P_{m/sel}$	$1/Kish_{m,i}$	$w_{ech^C,i}$
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, 90% des pondérations de l'échantillon complémentaire se situent entre 2 202 et 18 685. 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	$W_{ech^C,i}^{TNR}$
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

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 $\hat{Z} = \sum_{s_U} w_j z_j$. Si les poids sont sans biais, $\hat{Y}$ estime sans biais Y. On a donc $\hat{Z} = \sum_{s_U} w_j \sum_{i \in V} \frac{r_{ji}}{r_i} y_i = \hat{Y}$.

En posant $\tilde{w}_i = \frac{1}{r_i} \sum_{s_U} w_j r_{ji}$, 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 :

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

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

Tableau 15 : répartition des pondérations de l'échantillon complémentaire après partage des poids

Quantile	$W_{ech^C,i}^{TNR}$	$\tilde{W}_{ech^C,i}^{TNR}$
Max	65 932	38 359
99%	31 958	16 483
95%	18 685	10 204
90%	15 982	8 337
75%	14 224	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, 90% des pondérations de l'échantillon complémentaire se situent entre 1 101 et 10 204.

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 III.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 $W_{RPS,i}^{TNR} = \tilde{W}_{ech^C,i}^{TNR} \times \mathrm{I}_{ech^C,i} + \tilde{W}_{Panel,i}^{TNR,C} \times \mathrm{I}_{Panel,i}$ avec $\mathrm{I}_{ech^C}(\mathrm{I}_{Panel}) = 1$ si l'individu appartient à l'échantillon complémentaire ($Panel$); 0 sinon.

Tableau 16: répartition des pondérations des actifs occupés de la partie *Panel* après partage des poids

Quantile	$W_{Panel,i}^{TNR}$	$\tilde{W}_{Panel,i}^{TNR}$
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 17 : répartition des pondérations de la partie *RPS* après partage des poids

Quantile	$W_{RPS,i}^{TNR}$
Max	38 359
99%	11 554
95%	7 989
90%	5 544
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, 90% des pondérations de la partie RPS se situent entre 759 et 7 989.

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 18).

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 II.3.2).

Tableau 18 : 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 19 : répartition des poids et des rapports de poids selon la fonction de calage utilisée

	Logit tronquée		Raking ratio		Linéaire tronquée		Linéaire	
	Lo=0,1	Up=1,48			Lo=0,1	Up=1,46		
Quantile	Poids	Rapport des poids	Poids	Rapport des poids	Poids	Rapport des poids	Poids	Rapport des poids
Max	44 222	1,46	44 094	1,75	44 037	1,46	43 904	1,65
99%	13 049	1,39	13 060	1,49	13 079	1,46	13 083	1,46
95%	8 486	1,31	8 426	1,33	8 468	1,32	8 466	1,32
90%	6 170	1,26	6 285	1,26	6 254	1,27	6 260	1,26
75%	3 587	1,16	3 566	1,14	3 573	1,15	3 573	1,15
50%	1 942	1,05	1 949	1,02	1 945	1,03	1 945	1,03
25%	1 258	0,94	1 260	0,94	1 259	0,94	1 260	0,94
10%	938	0,83	954	0,86	947	0,85	946	0,85
5%	770	0,78	778	0,83	774	0,81	774	0,81
1%	543	0,67	545	0,76	544	0,71	544	0,72
Min	215	0,11	237	0,10	225	0,10	-76	-0,03

Au final, après redressements et calage, 90% des pondérations de la partie RPS se situent entre 770 et 8 486.

Tableau 20 : répartition des poids et des rapports de poids finaux de la partie *RPS*

Quantile	$W_{RPS,i}^{TNR}$	$W_{RPS,i}^{TNR,C}$	Rapport des poids
Max	38 359	44 222	1,46
99%	11 554	13 049	1,39
95%	7 989	8 486	1,31
90%	5 544	6 170	1,26
75%	3 276	3 587	1,16
50%	1 890	1 942	1,05
25%	1 269	1 258	0,94
10%	987	938	0,83
5%	759	770	0,78
1%	550	543	0,67
Min	352	215	0,11

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

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

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 enter la santé et la vie professionnelle et leur éventuel décalage temporel, premier bilan d'une enquête innovante : SIP », article présenté aux Journées de méthodologie statistique, 2009

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.

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

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
Ensemble		*4 665*[8]

[8] 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 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 $\left(W_0^{TNR,C}\right)$, 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 $\mathrm{I}_{complet} = 1$ 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 21: 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	9 487
95%	6 090	6 033	6 038	6 185
90%	5 100	4 952	4 955	5 077
75%	3 189	3 286	3 282	3 369
50%	2 429	2 421	2 421	2 482
25%	1 973	1 887	1 887	1 935
10%	1 450	1 300	1 301	1 333
5%	1 206	1 170	1 176	1 199
1%	804	867	867	889
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 admissible[9] 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 $(1 \leq g \leq m)$.

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 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.

[9] 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.