



**HAL**  
open science

## Pondérations du panel ELIPSS

Thomas Pilorin, Stéphane Legleye

► **To cite this version:**

Thomas Pilorin, Stéphane Legleye. Pondérations du panel ELIPSS : Version 1, janvier 2018. CDSP (Sciences Po). 2018. hal-03999528

**HAL Id: hal-03999528**

**<https://sciencespo.hal.science/hal-03999528>**

Submitted on 21 Feb 2023

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



ÉTUDE LONGITUDINALE  
PAR INTERNET  
POUR LES SCIENCES SOCIALES

## Pondérations du panel ELIPSS

*CDSP – Centre de données Socio-politiques*

**Version 1, janvier 2018**

T. Pilorin, avec la collaboration de S. Legleye et du Service des Enquêtes et Sondages de l'Ined

**SciencesPo**  
CENTRE DE DONNÉES SOCIO-POLITQUES



*Sciences-Po - CNRS, ANR-10-EQPX-19-01*

# Table des matières

<b>1</b>	<b>Contexte.....</b>	<b>3</b>
1.1	Champ d’ELIPSS et population cible	
1.2	Tirage de l’échantillon pilote	
1.3	Méthodologie de collecte de l’échantillon pilote	
1.4	Tirage de l’échantillon complémentaire	
1.5	Méthodologie de collecte de l’échantillon complémentaire	
1.6	Effectifs recrutés	
<b>2</b>	<b>Méthodologie de redressement.....</b>	<b>6</b>
2.1	Pondération initiale du pilote	
	<i>Correction de la non-réponse totale</i>	
	<i>Calage sur marges de l'EAR 2014</i>	
2.2	Pondération initiale du panel élargi	
	<i>Correction de la non-réponse totale</i>	
	<i>Calage sur marges de l'EAR 2014</i>	
<b>3</b>	<b>Utilisation des pondérations.....</b>	<b>13</b>
3.1	Choix d’une pondération	
3.2	Non-réponse résiduelle	
3.3	Utilisation d’un estimateur pondéré	
3.4	Intervalle de confiance sur ELIPSS	
<b>4</b>	<b>Bibliographie.....</b>	<b>16</b>
<b>5</b>	<b>Annexe 1 : Estimations sur le pilote.....</b>	<b>17</b>
<b>6</b>	<b>Annexe 2 : Estimations sur le panel élargi.....</b>	<b>18</b>
<b>7</b>	<b>Annexe 3 : Modèle de non-réponse.....</b>	<b>19</b>
<b>8</b>	<b>Annexe 4 : Calage sur marge avec Icarus.....</b>	<b>20</b>

## 1 Contexte

Le panel ELIPSS débute en 2012 par la sélection aléatoire d'un échantillon de 4 500 logements constituant une résidence principale en France métropolitaine (hors Corse), dans l'Enquête Annuelle du Recensement (EAR) via l'application Octopusse<sup>1</sup> de l'INSEE. La sélection d'une seule personne par ménage abordé (francophone, de 18 à 75 ans et n'ayant pas l'intention de déménager dans les 3 mois) est effectuée selon une méthode aléatoire ; la personne sélectionnée est celle dont la date d'anniversaire est la plus proche de la date du jour de l'interview. Différents protocoles de recrutement (face-à-face, courrier, téléphone) ont été mis en œuvre pour cette enquête pilote et ont permis la formation d'un premier panel de 1039 participants.

En 2016, un nouvel échantillon aléatoire de 9 000 logements tiré grâce à l'application Octopusse est livré. Le recrutement d'un seul individu par foyer est réalisé selon un unique protocole en face-à-face sur la population 18 à 79 ans à la date du contact. 2 514 personnes acceptent alors de participer à l'étude ELIPSS et viennent compléter l'étude pilote.

### 1.1 Champ d'ELIPSS et population cible

Le champ de l'enquête est l'ensemble des logements ordinaires<sup>2</sup> qui constituent une résidence principale en France métropolitaine, hors Corse, dont au moins un membre est âgé de moins de 80 ans au 1er juillet 2016, c'est-à-dire né à partir du 2 juillet 1936.

La population cible est définie comme les résidents francophones en France métropolitaine (hors Corse), ayant entre 18 et 79 ans le premier juillet 2016, résidant en ménage ordinaire (hors institutions et collectivités, sans domicile et habitats de fortune). La taille de cette population est estimée dans l'Enquête Annuelle du Recensement de 2014 à 45,3 millions d'individus.

### 1.2 Tirage de l'échantillon pilote

La méthode d'échantillonnage utilisée est un sondage à plusieurs degrés effectué par la division Sondages de l'INSEE. Le premier degré est le tirage de la vague du recensement (2011). Il ne s'agit pas à proprement parler d'un tirage aléatoire mais d'une contrainte destinée à fournir des fiches-adresses les plus à jour possible. Pour des raisons pratiques, ce tirage a été complété marginalement par des logements recensés en 2009 et 2010 (moins de 3% de la base fournie).

A partir de cette base, un échantillon de 4 500 logements a été tiré par l'INSEE par sondage stratifié et en grappes sur les critères suivants :

- ➔ Strates : région (sans la Corse) et type de commune (rural/urbain, grande/petite) ;
- ➔ Grappes : les zones d'action enquêteur (ZAE).

Cet échantillon est découpé en trois sous-échantillons dans la phase de recrutement : une partie principale (3 500 logements) et deux réserves, à activer en cas de besoin si l'exploitation de la partie principale ne permet pas d'atteindre l'objectif de taille

---

1 Organisation coordonnée de tirages optimisés pour une utilisation statistique des échantillons. Il s'agit de l'outil de tirage d'échantillon de ménages utilisé depuis 2004 à l'INSEE.

2 Le logement ordinaire est un logement défini par opposition à un logement en résidence offrant des services spécifiques : résidences pour personnes âgées, pour étudiants, de tourisme, à vocation sociale, pour personnes handicapées, ...

d'échantillon fixé, qui était de 1 500 panélistes. Ces réserves ont été activées, ce qui fait au total 4 500 logements exploités.

Après acceptation du ménage (1 349 ménages), le dernier degré du sondage consiste en un tirage aléatoire à probabilités égales d'un individu par ménage parmi ceux qui sont éligibles (c'est-à-dire âgé de 18 à 75 ans, francophone et qui habite dans un logement ordinaire en France métropolitaine et ne devant pas déménager dans les 3 mois suivant le recrutement). Le sondage est donc stratifié à plusieurs degrés, en grappes.

Les poids de sondage des 4 500 logements présentent une dispersion importante (un coefficient de variation de 94 et un rapport Max/Min de 123). Celle-ci ne peut être réduite, et se trouve même encore augmentée lors de la correction de la non-réponse. Ainsi, les poids corrigés qui étaient distribués aux utilisateurs des données présentaient jusqu'à présent une dispersion très importante rendant les estimateurs peu robustes. Dans la suite de cette note, est présentée la solution retenue pour remédier à ce problème.

### **1.3 Méthodologie de collecte de l'échantillon pilote**

Le recrutement du pilote s'est déroulé en deux temps. L'exploitation de l'échantillon principal (3 500 adresses) a eu lieu de juin 2012 à février 2013. Trois modes de contact ont été mis en œuvre. Une invitation à participer a été envoyée par courrier postal. Ce premier contact a fait l'objet d'une expérimentation : la moitié des courriers a été envoyée avec un bon cadeau de 10 euros et l'autre moitié sans bon cadeau<sup>1</sup>. Des relances postales et téléphoniques ont ensuite été réalisées. À l'issue d'un appel d'offres, l'institut de sondage TNS Sofres a été retenu pour exploiter en face-à-face les adresses pour lesquelles il n'y avait pas eu de contact et pour tenter de convaincre certains des ménages qui avaient refusé de participer. De janvier à avril 2013, les 1 000 adresses de réserve ont été exploitées par téléphone et en face-à-face par TNS Sofres. Au terme de cette procédure, 1 039 personnes ont accepté de participer au panel.

### **1.4 Tirage de l'échantillon complémentaire**

En 2016, un nouvel échantillon aléatoire de 10 000 fiches-adresses est tiré via l'application Octopusse. Cet échantillon est découpé en huit sous-échantillons dans la phase de recrutement : une partie principale (6 000 logements) et sept réserves pour un total de 4 000 logements. 9 000 fiches-adresses sont finalement utilisées lors de la phase de recrutement. Le tirage de l'échantillon est similaire à celui du pilote (stratifié à plusieurs degrés, en grappes) mais le calcul des poids de sondage a fait l'objet d'une attention particulière concernant leur dispersion. Les 9 000 poids de sondage présentent un coefficient de variation de 11 (contre 94 pour le pilote) et un rapport max/min de 1.6 (contre 123 pour le pilote).

Dans l'objectif de réunir les deux échantillons (pilote et complémentaire) dans un même ensemble, le champ de l'enquête est modifié par rapport à l'échantillon du pilote pour prendre en compte le vieillissement de 4 ans des individus du pilote. Le

---

<sup>1</sup> La réception des bons cadeaux s'est accompagnée d'une augmentation du taux de recrutement, sans différentiel lié aux caractéristiques des logements et des personnes de références des logements, et n'a pas eu d'incidence sur la participation à long terme aux enquêtes mensuelles. Le détail figure [ici](#) : Legleye, Razakamanana, Cornilleau, Cousteaux, *Intéressement financier, motivation initiale et caractéristiques des enquêtes : effets sur le recrutement et la participation à long terme dans le panel ELIPSS*. Présentation au Colloque francophone sur les sondages 2016.

champ de l'enquête est alors l'ensemble des logements ordinaires qui constituent une résidence principale en France métropolitaine, hors Corse, dont au moins un membre est âgé de moins de 80 ans au 1er juillet 2016, c'est-à-dire né à partir du 2 juillet 1936.

### **1.5 Méthodologie de collecte de l'échantillon complémentaire**

Le recrutement de l'échantillon complémentaire est réalisé par l'entreprise Ipsos du 8 janvier au 18 juillet 2016 selon un unique protocole de face-à-face suivi d'une phase de relance téléphonique. A l'issue d'un entretien d'une soixantaine de minutes en moyenne présentant l'étude et la tablette, 2 514 personnes acceptent de participer et viennent compléter l'étude pilote. Les taux de réponse, tels que définis par l'AAPOR dans l'Outcome Rate Calculator (RR3 In-Person House Hold Surveys), pour le pilote et l'échantillon complémentaire sont respectivement de 26% et 30%. La différence s'expliquant essentiellement par les bénéfices de l'expérience du pilote sur le recrutement complémentaire.

### **1.6 Effectifs recrutés**

L'échantillon du pilote recruté en 2012 a subi une attrition relativement faible : des 1 039 individus inclus initialement, 817 continuent de participer et sont intégrés en mai 2016 dans le panel élargi. La différence (222 individus soit 22 % des inclus du pilote après 4 ans) sont sortis du dispositif, soit par demande explicite, soit par leur exclusion du fait d'une trop faible participation. Ces deux raisons apparaissant en proportion équivalente.

Les variables disponibles dans les bases de sondage permettent d'apprécier la non-réponse lors de l'inclusion. Celle-ci apparaît différentielle, c'est à dire que tous les individus n'ont pas la même probabilité de répondre en fonction de leur caractéristiques. Globalement, le recrutement s'est montré sélectif sur la structure en âge du foyer, sur la situation d'emploi et de diplôme, sur la situation de logement et sur le revenu. En annexe 3 p.19, on peut observer la variabilité des taux de non-réponse sur les variables utilisées pour la modélisation de la non-réponse.

Pour réaliser des estimations, le panel pilote (1 039 inclus) continu d'être mobilisable sur toutes les enquêtes depuis le début du projet.

Le panel élargi agrège les 2 514 individus inclus de l'échantillon complémentaire et les 817 individus restants de l'étude pilote dans un ensemble de 3 331 individus inclus sur les enquêtes administrées depuis 2016.

Ces deux ensembles se sont vu affecter deux pondérations initiales desquelles dérivent toutes les pondérations diffusées aux utilisateurs.

## 2 Méthodologie de redressement

La non-réponse introduit un biais sur les estimateurs lorsque les non répondants n'ont pas le même comportement que les répondants sur le paramètre d'intérêt que l'on cherche à estimer<sup>1</sup>. Elle est ignorable lorsque que le paramètre est indépendant de la non-réponse. Ainsi une non-réponse importante n'implique pas nécessairement une moindre qualité de l'estimation. À l'inverse, la non-réponse n'est pas ignorable lorsque le paramètre estimé apparaît dépendant de la réponse.

La non-réponse est vue comme un mécanisme de sélection des individus qui apparaît à différentes étapes de l'enquête. Le mécanisme considéré ici est en deux étapes. La première phase apparaît lors des recrutements (du pilote et du complémentaire) : elle correspond aux refus et aux non-contacts éligibles qui ne sont donc pas recrutés alors qu'ils auraient dû l'être.

La seconde phase concerne les panélistes recrutés. Elle se décompose en attrition (qui est la sortie définitive du panel, quel qu'en soit le motif, à partir d'une date donnée) et d'une non-réponse ponctuelle propre à une enquête particulière. L'attrition ne peut qu'augmenter au cours du temps : en 2016, elle s'élève à 22% pour les panélistes recrutés durant le pilote, 4 ans après le début de celui-ci. En revanche, la participation à une enquête donnée, parmi les panélistes toujours présents et sollicités, est relativement constante au cours du temps : elle s'élève en moyenne de 85% (écart-type de 3.6 points). Il y a en fait une troisième phase de non-réponse : il s'agit des questionnaires commencés qui ne sont pas terminés. Dans les enquêtes ELIPSS, la proportion de questionnaires concernés est très faible (1,5% des invités à participer en moyenne). Cette troisième phase a été assimilée à de la non-réponse totale de vague et donc confondue avec la seconde phase de non-réponse et traitée en même temps que cette dernière.

Pour assurer la correction du biais la plus efficace possible, il a été retenu de procéder à un redressement en deux étapes<sup>2</sup> pour la première phase de non-réponse (celle du recrutement). La première étape de ce redressement consiste en une correction de la non-réponse totale par la méthode des groupes de réponse homogènes sur la base d'une régression logistique. La deuxième étape est un calage des poids corrigés sur des marges de l'Enquête Annuelle du Recensement (EAR) de 2014.

Une troncature des poids (2 % à gauche, 2 % à droite) est réalisée à l'issue de chacune de ces étapes. Deux pondérations initiales, correspondant aux ensembles *Pilote* et *Élargi* sont alors mises à disposition des utilisateurs.

La deuxième phase de non-réponse totale (celle concernant les panélistes recrutés) est effectuée en une seule étape : elle consiste à ajuster les pondérations corrigeant la non-réponse issue de la première phase sur les seuls répondants complets à l'enquête considérée par un nouveau calage sur les marges de l'EAR 2014.

Dans la suite du document sont présentées les variables et méthodes utilisées pour la correction de la non-réponse totale des pondérations initiales (correction de la première phase de non-réponse). En annexe 3, un exemple de calage d'une pondération initiale sur une enquête quelconque est présenté (correction de la deuxième phase de non-réponse). Cette dernière étape de correction aboutie aux pondérations telles qu'elles sont diffusées aux utilisateurs lors d'une livraison de données. Le code mis à disposition en annexe 4 p.20 peut être réutilisé pour produire des pondérations sur des ensembles de répondants à l'intersection de plusieurs enquêtes.

---

1 Groves 2006, Groves, Couper et al. 2006

2 Haziza et Beaumont 2007, Haziza et Lesage 2016

## 2.1 Pondération initiale du pilote

La procédure de tirage de l'échantillon du pilote avait sous-estimé l'impact croisé que pouvait avoir la petite taille de l'échantillon et le processus de sélection des unités primaires sur la dispersion des poids. Pour limiter cette dispersion, il est proposé de ne pas prendre en compte la phase de tirage de premier degré (sélection des unités primaires). Cela revient dans le cas présent à fixer un poids de tirage uniformisé pour l'ensemble de l'échantillon.

Les poids de sondage du pilote fournis par l'application Octopusse en 2012 sont en effet apparus très dispersés, rendant les estimateurs peu robustes. L'uniformisation des poids de tirage des logements résout le problème au prix de l'introduction d'un biais, partiellement compensé par la redressement. Les estimations de paramètres d'intérêt de l'enquête avec les différents poids finaux montre des différences relativement faibles. Deux distributions de variables extraites de la première enquête annuelle du pilote (924 répondants) sont présentées ci-dessous.

	Poids de sondage <b>originaux</b> corrigés et calés	Poids de sondage <b>uniformisés</b> corrigés et calés
<b>Etat de santé</b>		
Très bonne	12,6%	14,2%
Bonne	53,5%	48,6%
Assez bonne	29,5%	31,7%
Mauvaise	4,3%	5,4%
Très mauvaise	0,0%	0,1%
<b>Intérêt pour la politique</b>		
Beaucoup	12,6%	13,8%
Assez	40,9%	39,2%
Peu	33,0%	35,7%
Pas du tout	13,5%	11,2%

Les écarts à la base de sondage constatés entre les estimations pondérées suivant des poids tirage uniformisés ou non (respectivement nouvelles vs. anciennes pondérations) illustrent les distorsions engendrées par l'uniformisation des poids et nous confortent dans ce choix. Ils sont présentés en annexe 1, p.17.

Les poids de tirage des logements sont uniformes. Multipliés par le nombre d'individus éligibles du logement ils forment les poids de sondage qui sont corrigés en deux étapes.



### 2.1.a Correction de la non-réponse totale

La correction de la non-réponse totale à l'inclusion dans le panel pilote (soit la première étape du redressement de la première phase de non-réponse) est réalisée par la méthode des groupes de réponse homogènes à partir de variables de la base de sondage au niveau ménage. 9 groupes sont formés en suivant la méthode des quantiles pour le panel pilotes. Les 10 variables utilisées ont été sélectionnées parmi 28 variables en fonction de leur lien avec la non-réponse globale. Parmi les recrutés, elles sont toutes liées aux variables d'intérêt « état de santé perçu » et « intérêt pour la politique ». Ces variables, décrites dans le tableau en annexe, sont les suivantes :

- ➔ Nombre de pièces du logement (3 classes) ;
- ➔ Taille du foyer (5 classes) ;
- ➔ Taille d'unité urbaine (5 classes) ;
- ➔ Présence de personnes de plus de 60 ans dans le ménage ;
- ➔ Présence de personnes de moins de 25 ans dans le ménage ;
- ➔ Revenus du foyer (3 classes) ;
- ➔ Présence de diplômes supérieurs (aucun, un seul, plusieurs) ;
- ➔ Présence de fonctionnaires ;
- ➔ Présence de personnes en contrat à durée indéterminée ;
- ➔ Présence de voitures dans le ménage.

Les distributions de ces variables sont présentées en annexe 3 p.19 Les groupes formée par la méthode des quantiles sont de même taille et sont homogènes au sens de la non-réponse. Les poids de sondage des répondants sont corrigés d'un coefficient égal au taux de réponse au sein de leur groupe. La distribution de ces neuf coefficients est présentée ci dessous.

<b>Coefficients d'ajustement de la non-réponse du panel pilote</b>					
Min	Max	Max/Min	Moyenne	Ecart-type	CV (%)
0,06	0,45	7,50	0,22	0,09	41

## 2.1.b Calage sur marges de l'EAR 2014

Le calage est réalisé sur 5 marges de l'enquête annuelle du recensement 2014 (EAR) par la méthode du *raking ratio* implémentée dans le package *Icarus*<sup>1</sup>. L'algorithme converge avec une dispersion des coefficients de calage raisonnable. Ces coefficients ajustent les poids de sorte à retrouver les marges choisies avec l'estimateur pondéré, leur distribution est la suivante :

### **Distribution des coefficients de calage**

0%	1%	10%	25%	50%	75%	90%	99%	100%
0,43	0,49	0,66	0,75	0,88	1,11	1,53	2,47	3,88

Au final, le poids initial à l'inclusion dans le panel pilote est le produit du poids de tirage du logement (uniformisé), du nombre d'éligible dans le logement, du coefficient d'ajustement de la non-réponse et du coefficient de calage sur les marges du recensement.

		Marges de l'échantillon non-pondérées	Marges de l'échantillon corrigées avant calage	Marges EAR 2014 (estimateur corrigé après calage)
		Proportions %	Proportions %	Proportions %
<b>Age</b>	18 à 24 ans	4,7	5,5	11,0
	25 à 34 ans	15,0	15,3	16,2
	35 à 44 ans	25,3	18,8	17,5
	45 à 54 ans	22,0	18,4	19,6
	55 à 64 ans	18,1	17,6	17,2
	65 à 79 ans	14,9	24,5	18,6
<b>Diplôme</b>	Aucun / CEP / BEPC	15,9	19,2	27,8
	Bac+3 et plus	26,5	19,6	15,2
	Bac / bac+2	36,5	38,3	33,6
	CAP / BEP	21,1	22,9	23,4
<b>Sexe</b>	Femme	48,1	50,4	51,0
	Homme	51,9	49,6	49,0
<b>Nationalité</b>	Étranger	4,2	4,2	6,1
	Français de naissance	90,6	90,0	88,3
	Français par acquisition	5,1	5,9	5,6
<b>Région</b>	Bassin parisien	18,8	17,9	16,3
	Centre-Est	12,3	13,6	12,2
	Est	10,0	9,7	8,6
	Méditerranée	11,1	11,8	12,3
	Nord	5,5	4,7	6,4
	Ouest	14,7	15,0	13,9
	Région parisienne	17,7	16,9	18,9
	Sud-ouest	10,0	10,5	11,4

1 Rebecq, Antoine (2016). *Icarus* : an R package for calibration in survey sampling. R package version 0.2.0.

## **2.2 Pondération initiale du panel élargi**

Le panel élargi est rendu opérationnel par la mise à disposition d'une pondération sur l'ensemble élargi des 3 331 individus participants. L'ensemble des individus du pilote continuant à participer aux enquêtes en 2016, et des nouveaux recrutés de l'échantillon complémentaire sont agrégés dans cet ensemble. Deux méthodes d'agrégation des poids de sondage ont été testées.

La première divise les poids par le nombre de lien avec la population cible. Les individus ayant entre 18 et 22 ans ne sont en effet présents en 2016 que dans le deuxième échantillon (complémentaire), leur poids ne sont pas modifiés. Les autres individus sont présents dans les deux échantillons et voient leur poids de sondage divisés par deux.

La deuxième méthode fait l'hypothèse que les deux échantillons couvrent la même population des 18-79 ans. L'échantillon pilote comptant environ deux fois moins d'individus que l'échantillon complémentaire, les poids du pilote sont multipliés par 1/3 environ, les poids du complémentaire par 2/3.

C'est cette deuxième solution qui a été retenue en raison de la dispersion des poids du pilote, qui une fois agrégés par cette méthode avec les poids du complémentaire, n'apparaissent plus problématique pour la robustesse des estimateurs.

Au final, le poids initial à l'inclusion dans le panel élargi est le produit du poids de tirage du logement (fourni par Octopusse), du nombre d'éligible dans le logement, du coefficient d'ajustement de la non-réponse et du coefficient de calage sur les marges du recensement.

## 2.2.a Correction de la non-réponse totale

La correction de la non-réponse totale à l'inclusion dans le panel élargi (première étape du redressement de la première phase de non-réponse) est réalisée par la méthode des groupes de réponse homogènes. 12 groupes sont formés en suivant la méthode des quantiles pour le panel élargi. Les variables utilisées, et décrites dans le tableau en annexe 3 p.19, sont les suivantes :

- ➔ Nombre de pièce du logement (3 classes) ;
- ➔ Taille du foyer (5 classes) ;
- ➔ Taille d'unité urbaine (5 classes) ;
- ➔ Type d'aire urbaine (4 classes) ;
- ➔ Présence de personnes de plus de 60 ans dans le ménage ;
- ➔ Présence de personnes de moins 25 ans dans le ménage ;
- ➔ Revenus du foyer (3 classes) ;
- ➔ Diplôme supérieur dans le ménage (aucun, un seul, plusieurs) ;
- ➔ Présence de fonctionnaires ;
- ➔ Présence de personnes en contrat à durée indéterminée ;
- ➔ Présence de personnes à temps partiel ;
- ➔ Présence de personnes travaillant hors du département d'habitation ;
- ➔ Présence de voitures dans le ménage.

Comme précédemment, ces variables ont été sélectionnées en raison de leur lien avec la non-réponse totale et avec les deux variables d'intérêt disponibles (parmi les répondants) « état de santé perçu » et « intérêt pour la politique ». Les groupes sont de même taille et les coefficients d'ajustement sont décrits ci-après :

### Coefficients d'ajustement de la non-réponse du panel élargi

Min	Max	Max/Min	Moyenne	Ecart-type	CV (%)
0,08	0,45	5,63	0,25	0,09	36

## 2.2.b Calage sur marges de l'EAR 2014

Le calage est réalisé sur 5 marges de l'enquête annuelle du recensement 2014 (EAR) par la méthode du *raking ratio*. La variable âge voit ses modalités adaptées par rapport au calage du pilote pour prendre en compte l'absence d'individus de moins de 22 ans dans l'échantillon pilote. L'algorithme converge avec une dispersion des coefficients de calage raisonnable. Leur distribution est la suivante :

### ***Distribution des coefficients de calage***

0%	1%	10%	25%	50%	75%	90%	99%	100%
0,24	0,39	0,57	0,73	0,88	1,15	1,49	2,33	3,12

		Marges de l'échantillon non-pondérées	Marges de l'échantillon corrigées avant calage	Marges EAR 2014 (estimateur corrigé après calage)
		Proportions %	Proportions %	Proportions %
<b>Age</b>	18 à 22 ans	5,2	7,4	8,1
	23 à 34 ans	16,3	15,9	19,1
	35 à 44 ans	22,1	18,7	17,5
	45 à 54 ans	23,3	20,9	19,6
	55 à 64 ans	18,3	17,9	17,2
	65 à 75 ans	13,0	16,8	14,6
	75 à 79 ans	1,9	2,5	3,9
<b>Diplôme</b>	Aucun / CEP / BEPC	15,3	18,3	27,8
	Bac+3 et plus	25,6	22,2	15,2
	Bac / bac+2	36,6	36,4	33,6
	CAP / BEP	22,6	23,2	23,4
<b>Sexe</b>	Femme	52,4	51,2	51,0
	Homme	47,6	48,8	49,0
<b>Nationalité</b>	Étranger	7,4	9,1	6,1
	Français de naissance	88,3	86,1	88,3
	Français par acquisition	4,3	4,8	5,6
<b>Région</b>	Bassin parisien	18,3	17,2	16,3
	Centre-Est	12,8	12,1	12,2
	Est	9,5	8,6	8,6
	Méditerranée	11,5	12,5	12,3
	Nord	4,4	4,0	6,4
	Ouest	17,3	16,4	13,9
	Région parisienne	13,9	16,8	18,9
	Sud-ouest	12,2	12,3	11,4

## 3 Utilisation des pondérations

### 3.1 Choix d'une pondération

ELIPSS est une enquête longitudinale structurée en vagues d'enquêtes mensuelles. Pour chaque estimation sur ELIPSS, on souhaiterait pouvoir bénéficier d'une pondération où tous les individus répondants ont un poids strictement positif et tous les individus non-répondants ont un poids nul. Les pondérations décrites ci-avant sont fondées sur deux ensembles de répondants formés lors des phases d'inclusions dans le panel pilote et élargi en 2012 et 2016.

Panel	Nom	N	Somme	Moyenne	CV	Min.	Max.	Max / Min
Pilote	PDSPLT_INIT	994	994	1	71 %	0,29	3,5	11,9
Élargi	POIDS_INIT	3 331	3 331	1	77 %	0,04	3,5	96,1

Trois remarques peuvent être indiquées :

- ➔ Les poids de tirage des logements de la pondération du panel pilote sont uniformisés avant redressement et calage ;
- ➔ Les différents indicateurs de dispersion des deux pondérations (coefficients de variation et rapport des extrema) indiquent une dispersion nettement plus faible que celles des poids précédemment diffusés ;
- ➔ Leur somme est égale à la taille de l'échantillon et non à la taille de la population cible estimée à 45,3 millions par l'EAR 2014.

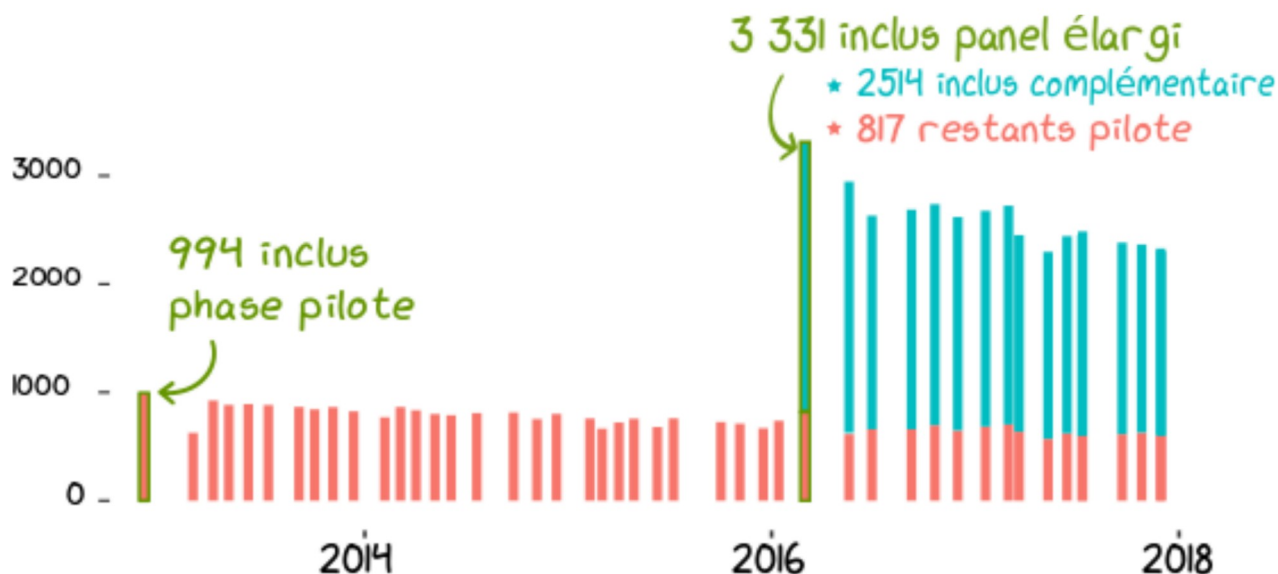
Pour bénéficier d'une correction de la non-réponse de vague, les poids initiaux des seuls répondants<sup>1</sup> de chaque vague sont recalés sur les marges de l'EAR 2014. Les utilisateurs bénéficient ainsi pour chaque enquête d'une pondération (nommée *POIDS [enquête]*) fondée sur l'ensemble des répondants pour une enquête choisie.

Pour que les enquêtes postérieures à 2016 puissent être appariées aux enquêtes de la phase pilote, une pondération *PDSPLT [enquête]* issue de la pondération du pilote est également mise à disposition et permet de mobiliser les seuls répondants du panel inclus dès la phase du pilote.

Quand les fichiers transmis résultent de l'appariement de différentes vagues d'enquête, des pondérations sont fournies sur l'intersection des répondants à toutes les vagues concernées suivant le même principe. Elles sont nommées *PDSLNG* suivi d'un identifiant propre aux enquêtes mobilisées.

Les pondérations nécessaires aux utilisateurs sont choisies par l'équipe ELIPSS et ajoutées à chaque fichier de données.

<sup>1</sup> Ayant complètement terminé le questionnaire. Le taux de questionnaires incomplets à l'issue d'une enquête ELIPSS étant généralement très faible (1,5 % des inclus en moyenne).

*Effectifs répondants aux enquêtes au fil du temps*

En moyenne, 85% des personnes invitées à répondre, terminent complètement l'enquête (écart-type de 3,6 points). L'attrition était de 22% en quatre ans pour le pilote. Elle suit la même évolution et atteint 16 % au bout de 2 ans sur le panel complémentaire.

### 3.2 Non-réponse résiduelle

Le cas général d'utilisation des données ELIPSS implique la transmission d'une vague d'enquête appariée avec l'enquête annuelle<sup>1</sup> précédente et un ensemble de variables de pondérations et de calage. D'autres types d'appariements peuvent aussi être proposés. La non-réponse aux enquêtes annuelles est imputée sur une quinzaine de variables des réponses aux enquêtes antérieures. L'année de l'enquête source de l'imputation est renseignée dans la variable nommée *VAGUE*.

Les variables de poids mises à disposition avec le fichier sont choisies pour traiter la présence de données manquantes résultant d'un croisement quelconque de variable du fichier. On souhaite que chaque valeur renseignée dispose d'un poids strictement positif et chaque valeur manquante d'un poids nul.

Du fait de la présence de non-réponse partielle aux enquêtes, il peut néanmoins apparaître des valeurs manquantes avec un poids positif. Et dans ce cas, la non-réponse peut introduire un biais sur l'estimateur. Si la non-réponse observée n'apparaît ni négligeable ni ignorable, il est toujours possible de calculer une nouvelle pondération grâce aux variables de calage mises à disposition et au code indiqué en annexe 4 p.20

Les poids initiaux (à l'inclusion dans le panel) *POIDS\_INIT* ou *PDSPLT\_INIT* des seuls répondants effectifs sont alors calés avec les cinq variables de calage *CAL\_ZEAT*, *CAL\_SEXE*, *CAL\_DIPL*, *CAL\_AGE*<sup>2</sup>, *CAL\_NAT* mises à dispositions dans les fichiers transmis. Les marges qui sont proposées (directement dans le code de l'annexe) proviennent de l'enquête annuelle du recensement de 2014.

1 Les enquêtes annuelles agrègent les questions utiles à l'ensemble des autres enquêtes.

2 Deux versions sont disponibles : *CAL\_AGE1* et *CAL\_AGE2* utilisées respectivement pour les enquête des phases du pilote et du panel élargi.

### 3.3 Utilisation d'un estimateur pondéré

L'estimation d'un total, d'une moyenne ou d'une proportion peut être réalisée simplement en utilisant un estimateur pondéré, c'est à dire en indiquant dans une procédure logiciel la variable de pondération fournie.

Différents types de poids peuvent être impliqués dans les procédures des logiciels. Ils peut s'agir notamment de poids sous forme de probabilités ou d'effectifs. Les poids choisis pour être diffusés aux utilisateurs ELIPSS représentent des effectifs dans la population cible ramenés à la taille de l'échantillon des répondants. Leur somme est égale au nombre de répondants mobilisés. Si besoin, on obtiendra des poids d'effectifs représentant la population cible en ramenant simplement leur somme à 45,3 millions de personnes.

Dans le logiciel R, on trouvera des estimateurs pondérés à travers les packages *Hmisc* ou *questionr* qui proposent des fonctions utiles aux analyses d'enquête par sondage. Sur SAS, on précisera l'argument *weight* d'une procédure. Avec Stata, l'argument *fweights*.

### 3.4 Intervalle de confiance sur ELIPSS

Dans une inférence sous le plan de sondage, les intervalles de confiance qui encadrent les estimations dérivent d'un estimateur de variance. Les estimateurs de variance implémentés dans les logiciels statistiques (notamment *wtd.var* des packages R ou *Proc\_Survey* pour SAS) ne correspondent pas au plan de sondage ELIPSS<sup>1</sup> qui n'a pas été caractérisé. Par conséquent, aucun estimateur de variance n'est actuellement conseillé pour produire un intervalle de confiance sur les données d'enquêtes ELIPSS.























---

1 Les deux échantillons sont issus de l'application Octopusse. Cf. *Les méthodes d'estimation de la précision pour les enquêtes ménages de l'Insee tirées dans Octopusse*. Méthodologie statistique. M 2015/03. Emmanuel Gros - Karim Moussallam. Document de travail. Institut National de la Statistique et des Études Économiques.



A titre illustratif, une comparaison des intentions de vote au premier tour de l'élection présidentielle française de 2017 avec les résultats effectivement obtenus, invite à considérer l'ensemble des sources d'erreur d'une estimation. En plus des erreurs d'échantillonnage et de non-réponse, on trouvera les erreurs de couverture et les biais inhérents aux enquêtes d'opinion fondées sur le déclaratif. On remarquera aussi les différences entre la population cible (métropolitaine) et la population concernée par l'élection.

*Déclarations des intentions de vote dans ELIPSS vs. Résultats de l'élection présidentielle 2017*

Liste des candidats au premier tour des présidentielles 2017	% déclaré dans l'échantillon ELIPSS (Avril 2017)	Écart relatif au résultat	Estimation du % déclaré dans la population cible ELIPSS (Avril 2017)	Écart relatif au résultat	Résultat de l'élection : % exprimés lors du vote (Ministère de l'Intérieur)
	Estimation naïve		Estimation pondérée		Résultats exhaustifs
M. Emmanuel MACRON	27,52	 15%	25,29	 5%	24,01
Mme Marine LE PEN	17,72	 17%	22,20	 4%	21,3
M. François FILLON	16,37	 18%	15,05	 25%	20,01
M. Jean-Luc MÉLENCHON	20,13	 3%	20,30	 4%	19,58
M. Benoît HAMON	10,86	 71%	9,05	 42%	6,36
M. Nicolas DUPONT-AIGNAN	3,40	 28%	3,47	 26%	4,7
M. Jean LASSALLE	1,47	 21%	1,26	 4%	1,21
M. Philippe POUTOU	1,29	 18%	1,37	 26%	1,09
M. François ASSELINEAU	0,47	 49%	0,72	 21%	0,92
Mme Nathalie ARTHAUD	0,70	 10%	1,23	 92%	0,64
M. Jacques CHEMINADE	0,06	 67%	0,05	 73%	0,18

Les utilisateurs des données ELIPSS rencontrant des difficultés pour utiliser les pondérations peuvent contacter l'équipe ELIPSS à l'adresse suivante : elipss.cdsp [at] sciencespo.fr

## 4 Bibliographie

- ➔ Groves, R. M. (2006). "Nonresponse Rates And Nonresponse Bias In Household Surveys." *Public Opinion Quarterly* 70(5): 646-675.
- ➔ Groves, R. M., M. P. Couper, S. Presser, E. Singer, R. Tourangeau, G. P. Acosta and L. Nelson (2006). "Experiments in producing nonresponse bias." *Public Opinion Quarterly* 70(5): 720-736.
- ➔ Frank E Harrell Jr, with contributions from Charles Dupont and many others. (2017). *Hmisc: Harrell Miscellaneous*. R package version 4.0-3.
- ➔ Haziza, D. and J.-F. Beaumont (2007). "On the Construction of Imputation Classes in Surveys." *International statistical review* 75(1): 25-43.
- ➔ Haziza, D. and E. Lesage (2016). "A discussion of weighting procedures for unit nonresponse." *Journal of Official Statistics* 32(1): 129-145.
- ➔ Julien Barnier, François Briatte and Joseph Larmarange (2017). *questionr: Functions to Make Surveys Processing Easier*. R package version 0.6.0.
- ➔ Rebecq, Antoine (2016). *Icarus: an R package for calibration in survey sampling*. R package version 0.2.0.

## 5 Annexe 1 : Estimations sur le pilote

		Base de sondage (n = 4 500)	Inclus Pilote (n = 1039)			
		estimations pondérées par les <b>poids de tirage</b>	estimations pondérées par les <b>nouveaux</b> poids	estimations pondérées par les <b>anciens</b> poids	écarts à la base de sondage	écarts à la base de sondage
		E <sub>0</sub>	E <sub>2</sub>	E <sub>1</sub>	E <sub>2</sub> - E <sub>0</sub>	E <sub>1</sub> - E <sub>0</sub>
Taille d'unité urbaine	Agglomération parisienne	16,4	18,1	18,4	1,7	2
	Plus de 200 mille hab hors Paris.	28,3	22,8	25,9	5,5	2,4
	100 à 200 mille hab.	5,1	6,6	4,3	1,5	0,8
	50 à 100 mille hab.	6,4	7,9	5,1	1,5	1,3
	20 à 50 mille hab.	5,9	7,7	5,4	1,8	0,5
	10 à 20 mille hab.	5,7	7,1	4,7	1,4	1
	5 à 10 mille hab.	8,5	5,7	7,1	2,8	1,4
	2 à 5 mille hab.	2,2	1,3	1,1	0,9	1,1
Commune rurale	21,5	22,9	28	1,4	6,5	
Nombre de pièce du logement	20 % plus petits	40,3	34,1	33,6	6,2	6,7
	60 % medians	43,9	48,8	47,5	4,9	3,6
	20 % plus grands	15,9	17,1	18,9	1,2	3
Taille du foyer	1 personne	35,5	23,2	18	12,3	17,5
	2 personnes	32,7	31,6	32,9	1,1	0,2
	3 personnes	14,4	18,5	22	4,1	7,6
	4 personnes	12,1	16,2	19,7	4,1	7,6
	5 personnes et plus	5,3	10,5	7,3	5,2	2
Type d'unité urbaine	Commune monopolarisée	6,2	4,9	8,1	1,3	1,9
	Commune multipolarisée	14,5	15,3	16,4	0,8	1,9
	Commune rurale	17,8	16	17	1,8	0,8
	Pôle urbain	61,6	63,7	58,4	2,1	3,2
Zone Urbaine Sensible	Logement non situé en ZUS	94,1	92,1	93,4	2	0,7
	Logement situé en ZUS	5,9	7,9	6,6	2	0,7
Statut d'occupation du logement	HLM	15,9	20,1	16,6	4,2	0,7
	Location vide	20,7	20,2	18,5	0,5	2,2
	Location meublée	2,6	2,5	3,1	0,1	0,5
	Logement gratuit	2,5	2	2,4	0,5	0,1
	Propriétaire	58,3	55,3	59,4	3	1,1
Type de logement	Appartement	43,7	40,9	36,1	2,8	7,6
	Chambre d'hôtel	0	0	0	0	0
	Habitation de fortune	0,1	0	0	0,1	0,1
	Logement-foyer	0,3	0,4	0,4	0,1	0,1
	Maison	55,8	58,3	63,2	2,5	7,4
	Pièce indépendante (ayant sa propre entrée)	0,2	0,4	0,3	0,2	0,1
Présence de + 60 ans dans le ménage	Absence plus de 60 ans	60,7	69,6	71,6	8,9	10,9
	Présence plus de 60 ans	39,3	30,4	28,4	8,9	10,9
Présence de - de 25 ans dans le ménage	Absence moins de 25 ans	63,5	50,8	46,1	12,7	17,4
	Présence moins de 25 ans	36,5	49,2	53,9	12,7	17,4
Revenus du ménage	25 % plus petits	24,5	24,4	20,3	0,1	4,2
	50 % médians	51	50,1	50,2	0,9	0,8
	25 % plus grands	24,5	25,5	29,5	1	5

La comparaison des estimations obtenues avec les anciens et les nouveaux poids illustre les distorsions induites par l'uniformisation des poids de tirage. Si la proportion d'habitants dans les villes de plus de 200 mille habitants s'écarte plus de la base de sondage, la proportion d'habitants des communes rurales s'en trouve améliorée. Les autres écarts constatés apparaissent comparables.

## 6 Annexe 2 : Estimations sur le panel élargi

Estimateurs	Base de sondage (n = 14 000)	Inclus panel élargi (n = 3 331)		écarts	
	Pondéré (poids de tirage agrégés)	Non pondéré	Poids corrigé et calé		
	EAR		HT	HT - EAR	
Taille d'unité urbaine	Agglomération parisienne	16,6	12,7	17,4	0,8
	Plus de 200 mille hab hors Paris.	26,8	23,4	23,4	-3,4
	100 à 200 mille hab.	4,8	5,6	4,5	-0,3
	50 à 100 mille hab.	6,1	6	5,7	-0,4
	20 à 50 mille hab.	6,8	7,3	8,1	1,3
	10 à 20 mille hab.	5,3	5,2	4,6	-0,7
	5 à 10 mille hab.	7,9	7,1	7,5	-0,4
	2 à 5 mille hab.	3,9	6,1	6,3	2,4
Commune rurale	21,8	26,7	22,5	0,7	
Nombre de pièce du logement	20 % plus petits	39,3	30,1	32,2	-7,1
	60 % médians	44,3	47,1	48,4	4,1
	20 % plus grands	16,4	22,8	19,4	3
Taille du foyer	1 personne	33,1	24,3	21,3	-11,8
	2 personnes	33,3	31,6	34,1	0,8
	3 personnes	15	17,4	18,8	3,8
	4 personnes	12,8	18,1	17,1	4,3
	5 personnes et plus	5,9	8,6	8,6	2,7
Logement HLM	Le logement appartient à un organisme HLM	15,8	13,2	16,6	0,8
	Le logement n'appartient pas à un organisme HLM	84,2	86,8	83,4	-0,8
Type d'unité urbaine	Commune monopolarisée	5,7	6,1	6,1	0,4
	Commune multipolarisée	16,1	19,9	18,8	2,7
	Commune rurale <sup>1</sup>	17,2	18,7	17	-0,2
	Pôle urbain	60,9	55,3	58,1	-2,8
Zone Urbaine Sensible	Logement non situé en ZUS	94,6	96,1	94,8	0,2
	Logement situé en ZUS	5,4	3,9	5,2	-0,2
Statut d'occupation du logement	HLM	15,3	12,9	16,2	0,9
	Location vide	22,2	20,2	19,8	-2,4
	Location meublée	2,7	1,8	1,7	-1
	Logement gratuit	2,2	1,4	1,2	-1
	Propriétaire	57,6	63,7	61,1	3,5
Présence de + 60 ans	Absence plus de 60 ans	64,5	76,4	69,5	5
	Présence plus de 60 ans	35,5	23,6	30,5	-5
Présence de - de 25 ans	Absence moins de 25 ans	60,5	48,5	50,4	-10,1
	Présence moins de 25 ans	39,5	51,5	49,6	10,1
Revenus du foyer	Revenu 25 % plus petits	25	24,3	24,4	-0,6
	Revenu 50 % médians	50,3	52,2	49,8	-0,5
	Revenu 25 % plus grands	24,7	23,5	25,8	1,1
Présence d'actifs	Absence d'actif	54,5	47,3	51,9	-2,6
	Présence d'actif	45,5	52,7	48,1	2,6
Présence de diplôme supérieur	Pas de diplôme supérieur	79,6	72,9	81,7	2,1
	Un seul diplôme supérieur	14,8	18,9	13,1	-1,7
	Plus d'un diplôme supérieur	5,6	8,2	5,2	-0,4

La pondération initiale du panel élargi résulte de la multiplication des poids de tirage agrégés sur les deux bases de sondage, par le nombre d'individus du ménage, par le coefficient correcteur des groupes de réponse homogène et par le coefficient de calage. Les écarts à la base de sondage sont réduits sans être complètement neutralisés, notamment vis-à-vis de la taille et de la structure en âge des foyers.

## 7 Annexe 3 : Modèle de non-réponse

Variables incluses dans les modèles	Effectifs mobilisés et non-réponse globale	Effectifs panel élargi			Effectifs panel pilote		
		14 000	%	% de non-réponse	4 500	%	% de non-réponse
				3 331 <b>76%</b>			1 039 <b>77%</b>
Nombre de pièce du logement	Nb. pièce log. 20 % plus petits	5 363	40%	81%	1 793	40%	82%
	Nb. pièce log. 60 % médians	5 955	44%	74%	2 001	44%	77%
	Nb. pièce log. 20 % plus grands	2 181	16%	65%	706	16%	71%
Taille du foyer	1 personne	4 374	32%	81%	1 580	35%	85%
	2 personnes	4 525	34%	77%	1 491	33%	80%
	3 personnes	2 022	15%	71%	632	14%	73%
	4 personnes	1 719	13%	65%	525	12%	62%
	5 personnes et plus	859	6%	67%	272	6%	69%
Présence de + 60 ans	Absence plus de 60 ans	8 983	67%	72%	2 755	61%	71%
	Présence plus de 60 ans	4 516	33%	83%	1 745	39%	88%
Présence de - de 25 ans	Absence moins de 25 ans	7 981	59%	80%	2 860	64%	83%
	Présence moins de 25 ans	5 518	41%	69%	1 640	36%	68%
Revenus du foyer	Revenu 25 % plus petits	3 413	25%	67%	1 136	25%	67%
	Revenu 50 % médians	6 733	50%	77%	2 258	50%	81%
	Revenu 25 % plus grands	3 353	25%	81%	1 106	25%	83%
Présence de diplôme supérieur	Pas de diplôme sup.	10 629	79%	77%	3 621	80%	81%
	Un seul diplôme sup.	2 077	15%	70%	649	14%	69%
	Plus d'un diplôme sup.	793	6%	66%	230	5%	62%
Présence de fonctionnaire	Absence fonctionnaire	9 324	69%	78%	3 206	71%	82%
	Présence fonctionnaire	4 175	31%	69%	1 294	29%	68%
Présence d'un CDI	Absence CDI	6 201	46%	82%	2 219	49%	86%
	Présence CDI	7 298	54%	70%	2 281	51%	70%
Présence travailleur hors département	Pas de travail hors département	7 824	58%	80%	2 654	59%	84%
	Travail hors département	5 675	42%	69%	1 846	41%	69%
Présence de voiture	Absence voiture	2 354	17%	86%	866	19%	87%
	Présence voiture	11 145	83%	73%	3 634	81%	76%
Type d'unité urbaine	Commune monopolarisée	657	5%	69%			
	Commune multipolarisée	2 190	16%	70%			
	Commune rurale	2 122	16%	71%			
	Pôle urbain	8 530	63%	78%			
Présence de temps partiel	Absence temps partiel	11 332	84%	77%			
	Présence temps partiel	2 167	16%	68%			

Les variables sont choisies dans la base de sondage en fonction de leur lien avec la non-réponse et de leur lien supposé avec les variables d'intérêt de l'enquête. Différentes méthodes de modélisation et de formation des groupes de réponse homogène ont été testées. Les modélisations par *boosting* ou par régression logistique, les classifications hiérarchiques, ou par l'algorithme *Mclust*, ou la méthode des quantiles donnent des résultats comparables sur ces données. Le choix de la méthode s'est donc porté sur celle qui nous est apparue la plus simple.

Différents modèles ont été testés pour estimer les probabilités de réponse. La qualité de prédiction de la non-réponse et la qualité d'explication supposée avec les variables d'intérêt de l'enquête a présidé au choix des modèles.

## 8 Annexe 4 : Calage sur marge avec Icarus

```
# Marges estimées de la population cible sur EAR.2014 (n = 10 000)
MARGES_EAR.2014 = rbind(
MAR1=c("CAL_AGE1", "6", "0.1099", "0.162", "0.175", "0.1956", "0.172", "0.1855", "0", "0"),
MAR2=c("CAL_AGE2", "7", "0.0806", "0.1914", "0.175", "0.1956", "0.172", "0.1464", "0.039",
"0"),
MAR3=c("CAL_DIPL", "4", "0.2783", "0.1518", "0.3364", "0.2335", "0", "0", "0", "0"),
MAR4=c("CAL_SEXE", "2", "0.5098", "0.4902", "0", "0", "0", "0", "0", "0"),
MAR5=c("CAL_NAT", "3", "0.0608", "0.8828", "0.0564", "0", "0", "0", "0", "0"),
MAR6=c("CAL_ZEAT", "8", "0.1632", "0.1217", "0.0862", "0.1234", "0.0642", "0.1385",
"0.1886", "0.1142"))

# Choix du poids à caler : pondération initiale POIDS_INIT ou PDSPLT_INIT
# Suppression des individus sans poids initial
POND_INITIALE = subset(POND_INITIALE, !is.na(POND_INITIALE$POIDS_INIT))

# Sous ensemble de répondants (fictif)
POND_INITIALE$repondant = rbinom(dim(POND_INITIALE)[1], 1, 1)

# Table à caler
POND_INITIALE = subset(POND_INITIALE, POND_INITIALE$repondant == 1,
c("POIDS_INIT", "CAL_AGE2", "CAL_NAT", "CAL_DIPL", "CAL_SEXE", "CAL_ZEAT"))
## Pour un calage des enquêtes antérieure à 2016 on utilisera CAL_AGE1

# Total de la population estimé, ou total de l'échantillon suivant l'utilisation de la
pondération
N = 43500000

# Ordonner les modalités des facteurs
tri_levels = fonction(variable){
return(factor(variable, levels(variable)[order(levels(variable))]))
}
for (i in c("CAL_AGE2", "CAL_NAT", "CAL_DIPL", "CAL_SEXE", "CAL_ZEAT")) {
POND_INITIALE[,i] = tri_levels(POND_INITIALE[,i])}

# Calage sur marges
library(icarus)
poids = calibration(
data = POND_INITIALE,
colWeights = "POIDS_INIT",
marginMatrix = MARGES_EAR.2014[-1,],
pct = T,
popTotal = N,
method = "raking")
```

```
##### Summary of before/after weight ratios #####
```

```
Calibration method : raking
```

```
Mean : 0.9854
```

```
0% 1% 10% 25% 50% 75% 90% 99% 100%
```

```
0.8943 0.9120 0.9537 0.9620 0.9761 1.0018 1.0302 1.1007 1.1634
```

```
##### Comparison Margins Before/After calibration #####
```

```
Careful, calibration may not be exact
```

```
$Total
```

```
Before calibration After Calibration Margin
```

```
3331 42270000 42270000
```

```
$CAL_AGE_2
```

```
Before calibration After Calibration Margin
```

```
(17,22] 7.96 8.06 8.06
```

```
(22,34] 18.99 19.14 19.14
```

```
(34,44] 17.67 17.50 17.50
```

```
(44,54] 19.82 19.56 19.56
```

(54,64]	17.18	17.20	17.20
(64,75]	14.66	14.64	14.64
(75,79]	3.73	3.90	3.90

\$CAL\_DIPL

	Before calibration	After Calibration	Margin
Aucun/CEP/BEPC	26.68	27.83	27.83
Bac+3 et plus	15.50	15.18	15.18
Bac_bac+2	34.11	33.64	33.64
CAP/BEP	23.71	23.35	23.35

\$CAL\_SEXE

	Before calibration	After Calibration	Margin
Femme	51.08	50.98	50.98
Homme	48.92	49.02	49.02

\$CAL\_NAT

	Before calibration	After Calibration	Margin
Etranger		6.21	6.08 6.08
Français de naissance		88.22	88.28 88.28
Français par acquisition		5.57	5.64 5.64

\$CAL\_ZEAT

	Before calibration	After Calibration	Margin
Bassin parisien	16.53	16.32	16.32
Centre-Est	12.34	12.17	12.17
Est	8.75	8.62	8.62
Méditerranée	12.25	12.34	12.34
Nord	5.88	6.42	6.42
Ouest	14.10	13.85	13.85
Région parisienne	18.63	18.86	18.86
Sud-ouest	11.53	11.42	11.42

