Insee
Mesurer pour comprendre
Direction générale

Sources et méthodes

Enquête Statistiques sur les Ressources et les Conditions de Vie (SRCV)

Présentation générale de l'enquête

L'enquête Statistiques sur les ressources et les conditions de vie (SRCV) permet de produire des statistiques sur les revenus et les conditions de vie et de construire des indicateurs structurels sur la pauvreté et l'exclusion sociale pour la commission européenne en vue d'une comparabilité entre les pays membres de l'Union européenne.

Le dispositif SRCV (Statistiques sur les Ressources et les Conditions de Vie) est la partie française du système communautaire EU-SILC (European Union - *Community Statistics on Income and Living Conditions),* piloté par Eurostat.

Son dispositif s'articule en deux composantes : une composante « transversale » qui s'apparente à une enquête annuelle « traditionnelle », et une composante « longitudinale » qui provient du suivi d'une année sur l'autre des individus enquêtés dans l'enquête transversale. Le règlement européen prévoit le suivi des ménages du panel pendant au moins quatre ans. Pour apporter un réel progrès dans le domaine du suivi de situations de pauvreté et répondre aux besoins de la recherche française, la France a mis en place un panel long sur neuf ans.

Champ de l'enquête

Le champ couvert par l'enquête est celui des ménages (unités de vie) dits « ordinaires » résidant en France métropolitaine. Sont donc exclus les ménages vivant en collectivité (foyers, prisons, hôpitaux…), ainsi que les personnes vivant dans des habitations mobiles (mariniers…) et les sans-domicile. Les individus interrogés sont tous les adultes du ménage âgés de 16 ans ou plus.
Dans la dimension longitudinale, l'unité est l'individu panel, défini comme un individu appartenant à un ménage répondant lors d'une première interrogation. Ce sont ces individus qui sont suivis pendant neuf années. Pour calculer le niveau de vie de l'individu, il est

TIMBRE DG75-F350 - Institut National de la Statistique et des Études Économiques - 18 bd Adolphe Pinard - 75675 PARIS CEDEX 14 - FRANCE - www.insee.fr
Tél. standard : 01.41.17.50.50 - N° SIRET : 120 027 016 00019 - Code APE : 84.11Z - Service Insee Contact : 09 72 72 4000 - (tarification "appel local")

cependant nécessaire d'interroger tous les membres du ménage auquel appartient l'individu panel.

L'échantillon de départ

A son démarrage, en 2004,l'échantillonnage de l'enquête SRCV reposait sur un échantillon maître issu du recensement de 1999 et complété par la base de sondage des logements neufs (BSLN). Depuis 2010, en raison de la nouvelle méthodologie du recensement « en continu », qui assure une couverture partielle du territoire chaque année (et totale sur un cycle de 5 années), une refonte globale du système d'échantillonnage a eu lieu. Le nouveau système d'échantillonnage Octopusse, (Organisation coordonnée de tirages optimisés pour une utilisation statistique des échantillons) permet de tirer les échantillons des enquêtes ménages d'une année donnée dans la liste des logements recensés l'année précédente, assurant ainsi une actualisation permanente de la base de sondage et rendant alors inutile le suivi spécifique des logements neufs.

Le champ de l'enquête est constitué par l'ensemble des ménages ordinaires résidant en France métropolitaine[1]. L'échantillon sélectionné la 1ère année en 2004 pour cette enquête comportait 16 000 logements, soit 16 000 fiches adresses (FA) tirées.

Une partie de cet échantillon a constitué la 1ère vague du Panel sur les Ressources et les Conditions de Vie : les personnes appartenant à ces ménages sont interrogées pendant 9 ans même si elles déménagent, pourvu qu'elles continuent de résider en France métropolitaine et en ménage ordinaire.

A partir de la 2ème année, l'échantillon est constitué des répondants de la vague précédente du Panel sur les Ressources et les Conditions de Vie moins les sortants, et d'un échantillon « entrant », composé de 3 000 logements, destiné à renouveler la population représentée par le panel : c'est le principe de l'échantillon rotatif (Voir ci-après, le schéma des sous-échantillons depuis 2004). Depuis 2012, les échantillons entrants sont composés de 3 200 fiches adresses.

L'enquête est obligatoire pendant les quatre premières interrogations annuelles d'un ménage.

Collecte terrain

La collecte s'effectue sous Capi (collecte assistée par informatique), par visite d'un enquêteur auprès des ménages. Celui-ci a environ huit semaines pour réaliser ses entretiens de début mai à fin juin. L'enquête est composée d'un questionnaire « Ménage » (destiné à l'ensemble du ménage) et d'un questionnaire « Individu » posé à toutes les personnes du ménage âgées de 16 ans ou plus (au 1er janvier de l'année d'enquête).

[1] Le règlement européen *EU-SILC* (n°215/2007 de la Commission du 28/02/2007) prévoit des dérogations pour les régions « ultra périphériques » de l'Union européenne. L'INSEE dispose d'une dérogation pour les départements et territoires d'Outre-mer.

Liste des sous-échantillons présents par année de collecte

2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
11												
12	12											
13	13	13										
14	14	14	14									
15	15	15	15	15								
16	16	16	16	16	16							
17	17	17	17	17	17	17						
18	18	18	18	18	18	18	18					
19	19	19	19	19	19	19	19	19				
	20	20	20	20	20	20	20	20	20			
		21	21	21	21	21	21	21	21	21		
			22	22	22	22	22	22	22	22	22	
				23	23	23	23	23	23	23	23	23
					24	24	24	24	24	24	24	24
						25	25	25	25	25	25	25
							26	26	26	26	26	26
								27	27	27	27	27
									28	28	28	28
										29	29	29
											30	30
												31

Questionnaire réinterrogation

Questionnaire 1ère interrogation

11 Numéro du sous-échantillon

Entrants

Interrogation obligatoire

La refonte de 2008

En 2008, le dispositif SRCV a fait l'objet d'une refonte. Le projet de refonte SRCV visait, d'une part, à réorganiser l'ensemble du système d'information sur les conditions de vie autour du seul dispositif SRCV (avec l'abandon des anciennes enquêtes permanentes de conditions de vie (EPCV)), en y incluant des thèmes supplémentaires relatifs aux conditions de vie (bloc indicateurs sociaux) ; d'autre part, à remplacer, pour une partie des ressources des ménages, la collecte par voie d'enquête par un recours aux données administratives, déclarations fiscales et fichiers de prestations des organismes gestionnaires (Cnaf, CCMSA et Cnav).

Trois blocs rotatifs d'indicateurs sociaux ont été conçus selon des thèmes généralement présents dans les anciennes enquêtes EPCV. Les thèmes des blocs retenus sont :

- IS1 : Santé, participation sociale, relations professionnelles, vie associative (2008, 2010, 2013, 2016 - hors santé, 2017 - santé)
- IS2 : Participation culturelle et sportive (2009, 2012, 2015…)
- IS3 : Mobilité sociale et contact avec les proches (2011, 2014…)

L'appariement des données avec les fichiers fiscaux à partir de la collecte des données 2008

Le recours aux données administratives, données fiscales et sociales permet d'améliorer la qualité du dispositif en profitant des atouts des sources administratives. Celles-ci offrent une qualité homogène et très satisfaisante des données sur les revenus pour la grande majorité de la population. Ce n'est pas le cas pour les sources déclaratives. En effet, les erreurs entre euros et francs, voire même entre euros et anciens francs, et entre montants annuels et montants mensuels restent non négligeables. Par ailleurs, certains ménages utilisent des documents, d'autres non.
Dans le cas où les documents ne sont pas utilisés, les données sont plus souvent approximatives ou simplement fournies en tranches. Enfin, une fraction des ménages ne souhaite répondre à aucune question relative aux revenus. Tous ces éléments conduisent à une hétérogénéité dans la collecte qui nuit aux objectifs de qualité globale.

Depuis 2008, à de très rares exceptions les ménages ne sont plus questionnés sur le montant des revenus et de leurs impôts, sauf en ce qui concerne les revenus non imposables, la taxe foncière et l'ISF qui restent collectés par voie d'enquête. Les revenus imposables sont obtenus à partir des fichiers fiscaux de la Direction Générale des Finances Publiques ou des organismes gestionnaires de prestations pour les revenus sociaux (CNAF, CCMSA, CNAV depuis 2009). Le rapprochement du fichier d'enquête et du fichier fiscal est opéré à l'aide de l'adresse du ménage. Dans le cas où les adresses ne correspondent pas, les revenus ne seront pas retrouvés. C'est le cas des jeunes en particulier qui peuvent déclarer leurs revenus sur la déclaration d'impôts de leurs parents s'ils sont âgés de moins de 21 ans, ou bien s'ils sont étudiants de moins de 25 ans ; ces derniers restent interrogés suivant l'ancienne méthode.

Enfin, tout appariement n'étant pas exhaustif, un travail de traitement des non-réponses et de redressement des réponses partielles est à réaliser (même s'il est moindre que par le passé).

Pondération

1. *La pondération associée à l'échantillonnage des logements*

L'unité statistique pondérée est le logement. Chaque année, l'échantillon transversal est composé de ménages formés par des individus dont certains sont suivis dans le temps (individus-panels) et d'autres qui sont enquêtés parce qu'ils résident dans un ménage comprenant au moins un individu-panel (on parle alors de cohabitants).

Les logements sont tirés dans le fichier du recensement général de la population le plus récent disponible, selon la méthodologie standard de l'Insee qui s'apparente à une technique d'échantillon-maitre (système d'échantillonnage Octopusse). Le système d'échantillonnage produit des poids variant dans un rapport de 1 à 4 parce qu'on sous-représente au tirage les résidences non-principales au recensement (résidences vacantes et secondaires).

Si on considère l'ensemble des poids d'échantillonnage des sous-échantillons mobilisés une année donnée, on constate qu'ils varient dans un rapport de 1 à 7, les tailles des sous-échantillons entrant chaque année variant selon l'année de tirage.

2. *La pondération transversale utilisant la méthode de partage des poids*

La méthodologie de pondération, décrite en l'absence de non-réponse, figure dans

« Pondération dans les échantillons rotatifs : le cas de l'enquête SILC en France », P. Ardilly et P. Lavallée, Techniques d'enquête, décembre 2007, Vol 33., N°2, pp 149-156.

2.a. *Expression formelle des poids en régime stationnaire*

On a affaire à un échantillonnage indirect de ménages / individus qui associe la réunion de bases de sondage annuelles (au nombre de 9 en régime stationnaire) à la base des ménages / individus considérés à la date courante. Cela conditionne le système des liens qui structure l'échantillonnage indirect : tout individu échantillonné (au travers d'un logement) dans une des neuf bases de sondage est suivi dans le temps et pointe donc sur lui-même (et seulement sur lui-même).

On se situe l'année t, où on dispose d'un échantillon transversal obtenu à partir du suivi dans le temps de 9 sous-échantillons panélisés notés respectivement $a_{t,1}$ à $a_{t,9}$ où $a_{t,k}$ désigne le sous-échantillon entré l'année $t-k+1$ (k varie de 1 à 9). La base de sondage associée au tirage de $a_{t,k}$ est notée $\Omega_{t,k}$. Tout individu de $a_{t,k}$ est (par définition) un individu-panel et tout individu d'un ménage *in fine* enquêté à t mais qui n'est pas un individu-panel est un cohabitant. On trouve des individus-panels de tous âges puisqu'un nouveau-né d'une mère elle-même individu-panel acquiert automatiquement le statut d'individu-panel. Il est alors suivi jusqu'à ce que sa mère sorte du panel. L'année t, le sous-échantillon répondant issu de $a_{t,k}$ est noté $r_{t,k}$.

L'année t, le poids transversal (avant calage) de tout ménage répondant l, et donc de tout individu (de fait répondant) appartenant à ce ménage, sera

$$W_l^{*t} = \frac{\sum_{i \in l} \sum_{k=1}^{9} \sum_{j \in r_{t,k}} \frac{W_j(t,k)}{\Phi_j(t,k) \cdot \theta_j(t,k)} \cdot 1_{j=i}}{\sum_{i \in l} \sum_{k=1}^{9} \sum_{j \in \Omega_{t,k}} 1_{j=i}}$$

avec

$1_{j=i}$, variable indicatrice valant 1 si i et j sont en réalité le même individu, et 0 sinon

$W_j(t,k)$, poids de sondage associé à l'individu-panel j lors du tirage de $a_{t,k}$

$\Phi_j(t,k)$, probabilité de réponse de l'individu-panel j l'année où il a été échantillonné

$\theta_j(t,k)$, probabilité que l'individu-panel j soit répondant à la date t, sachant qu'il est répondant l'année de son échantillonnage (donc l'année du tirage de $a_{t,k}$).

Pour tout t, on a $\theta_j(t,1) = 1$. Par ailleurs, $\Phi_j(t,k)$ est le même pour tous les individus j présents dans le même ménage au moment de son tirage. Un individu cohabitant ne participe pas à la formation du numérateur, en revanche il « compte » dans le dénominateur.

L'expression précédente est une expression générale, valable dans tous les cas. Cela étant, l'Insee a procédé à des simplifications permettant une programmation sensiblement plus facile, adoptant une expression qui lui est (presque) équivalente. En effet, si on néglige les cas - extrêmement rares - où un ménage à la date courante rassemblerait des individus-panels j issus de différents ménages, voire comprendrait un individu-panel tiré à deux occasions[2], alors le poids individuel $W_j(t,k)$ est assimilé à un poids de ménage $W_l(t,k)$, en identifiant par l le ménage contenant les individus j. En outre, il est bien clair que par construction la probabilité de réponse $\Phi_j(t,k)$ est en réalité la probabilité de réponse du ménage dans lequel se trouvaient à l'origine tous les individus-panels, ménage que l'on peut continuer à noter l par convention[3]. Dans ce cas, on peut simplifier un peu l'expression générale selon :

$$W_l^{*t} = \frac{W_{l(}t,k)}{\Phi_l(t,k)} \cdot \frac{\sum_{i \in l} \sum_{k=1}^{9} \sum_{j \in r_{t,k}} \frac{1}{\theta_j(t,k)} \cdot 1_{j=i}}{\sum_{i \in l} \sum_{k=1}^{9} \sum_{j \in \Omega_{t,k}} 1_{j=i}}$$

que l'on peut ré écrire

[2] En pratique, il arrive assez souvent que la composition d'un ménage évolue dans le temps, mais ce serait un très grand hasard qu'il y ait une recomposition qui rapproche des individus-panels provenant initialement de ménages distincts. En revanche, ce qui est courant, c'est de constater au cours du temps une segmentation progressive d'un ménage initial. Mais dans ce cas, les individus-panels qui restent ensemble conservent bien le même poids, qui est celui du ménage initial. Certes, le tirage d'un individu-panel donné à deux occasions pourrait s'imaginer dans le cas d'un déménagement, ou si le hasard de l'échantillonnage s'avère défavorable dans une commune où un même logement est tiré deux fois à plus de 5 années d'intervalle (Octopusse ne gère pas la disjonction au-delà de 5 années consécutives et le panel SILC dure 9 années …).

[3] Par convention effectivement, parce qu'entre la date de tirage et la date courante, le périmètre du ménage a pu changer. Cette notation apparait donc un peu délicate.

$$W_l^{*t} = \frac{W_l(t,k)}{\Phi_l(t,k)} \cdot \frac{\sum_{k=1}^{9} \sum_{\substack{j \in l \\ j \in r_{t,k}}} \frac{1}{\theta_j(t,k)}}{\sum_{i \in l} L_i(t)}$$

en posant $L_i(t) = \sum_{k=1}^{9} \sum_{j \in \Omega_{t,k}} 1_{j=i}$. Cette dernière expression est appelée « nombre de liens » de l'individu i. Elle s'interprète comme le nombre total d'années passées (plafonné à 9) durant laquelle l'individu i a fait partie du champ de l'enquête. C'est aussi le nombre total de sous-échantillons dans lequel il aurait pu se trouver[4].

On peut encore pousser un peu la simplification si le ménage l considéré est un ménage entrant et répondant ($l \in r_{t,1}$). On est toujours sous l'hypothèse - extrêmement faible - où ce ménage ne comprend aucun individu-panel se trouvant également dans un sous-échantillon tiré par le passé. Alors pour tout j de l on a $\theta_j(t,1) = 1$ et

$$W_l^{*t} = \frac{W_l(t,k)}{\Phi_l(t,k)} \cdot \frac{\sum_{\substack{j \in l \\ j \in r_{t,1}}} 1}{\sum_{i \in l} L_i(t)} = \frac{W_l(t,k)}{\Phi_l(t,k)} \cdot \frac{Npan_l(t)}{\sum_{i \in l} L_i(t)}$$

où $Npan_l(t)$ désigne le nombre total d'individus-panels dans le logement l en t, c'est-à-dire de fait la taille du ménage.

2.b. Expression formelle des poids durant la phase d'initialisation

Le régime stationnaire démarre en 2012. Auparavant, il a fallu adapter chaque année le calcul du nombre de liens, parce que l'échantillon transversal de 2004 a été tiré en une seule fois, dans une unique base de sondage, puis ventilé en 9 sous-échantillons de même structure.

En 2005, il s'avère que 8 sous-échantillons ont été tirés dans la même base (celle qui a servi au tirage de l'échantillon de 2004), ce qui fait que chaque individu d'un ménage enquêté en 2005 possède ou bien 9 liens s'il était tirable en 2004 (donc présent dans la base de sondage de l'époque), ou bien un seul lien s'il était tirable seulement[5] en 2005. Ce dernier cas correspond aux individus du ménage qui n'étaient pas dans le champ de l'enquête en 2004. On a

$$\sum_{i \in l} L_i(2005) = \sum_{i \in l} \sum_{k=1}^{9} \sum_{j \in \Omega_{2005,k}} 1_{j=i} = 8 \cdot Nchamp_l^{2004}(2005) + Npers_l(2005).$$

[4] Le système d'échantillonnage Octopusse assure une disjonction à horizon de 5 années en glissement, mais c'est une propriété « opérationnelle » qu'il faut ignorer pour mener les raisonnements théoriques : évidemment, l'inférence se conçoit toujours sur la population complète et jamais sur une population abstraite amputée des échantillons déjà tirés. C'est pourquoi, pour la grande majorité des individus, le nombre total de liens vaut 9.

[5] Ce sera par exemple le cas d'une personne immigrante, qui arrive en France en 2005.

où $Npers_l(2005)$ est le nombre total d'individus dans le logement *l* en 2005 et $Nchamp_l^{2004}(2005)$ est le nombre total d'individus dans le logement l en 2005 qui étaient déjà dans le champ en 2004.

En 2006, avec le même raisonnement

$$\sum_{i \in l} L_i(2006) = 7 \cdot Nchamp_l^{2004}(2006) + Nchamp_l^{2005}(2006) + Npers_l(2006)$$

etc…
En 2011, dernière année du régime transitoire :

$$\sum_{i \in l} L_i(2011)$$
$$= 2 \cdot Nchamp_l^{2004}(2011) + Nchamp_l^{2005}(2011) + ... + Nchamp_l^{2010}(2011) + Npers_l(2011)$$

Finalement, la formule de pondération du 2.a s'applique en l'état, à l'exception notable du nombre total de liens figurant au dénominateur, qui s'adapte comme nous venons de l'expliquer.

En l'absence d'informations permettant d'apprécier l'année courante la présence ou non de i dans le champ de l'enquête durant les années passées, $L_i(t)$ est automatiquement fixé à la valeur 9. Pour obtenir une valeur pertinente du nombre de liens, il est nécessaire de comptabiliser le nombre d'années de présence de l'individu dans le champ. Pour cela, on prendra en compte l'année de naissance, la date d'arrivée sur le territoire français et le nombre d'années passées à l'étranger ou dans les DOM-TOM.

2.c. L'estimation des probabilités de réponse

On distingue la non-réponse au moment où le logement est échantillonné, qui concerne $\Phi_j(t,k)$ - donc le comportement du ménage à la date du premier contact - et la non-réponse au cours des années ultérieures, en phase de ré-interrogation, qui renvoie aux $\theta_j(t,k)$. Chaque type de non-réponse donne lieu à un modèle spécifique, qui mobilise une information qui lui est propre. Qu'il s'agisse de non-réponse initiale comme de non-réponse en ré-interrogation, le principe général de l'estimation consiste à sélectionner des variables explicatives au moyen d'une régression logistique (non pondérée) puis à fabriquer des groupes de réponse homogènes (GRH) en croisant toutes les modalités des variables retenues (il s'agit exclusivement de variables qualitatives - souvent d'ailleurs des indicatrices). Cette phase donne lieu à des regroupements de classes lorsque l'effectif répondant est jugé trop faible. Dans chaque GRH, on estime *in fine* la probabilité de réponse par le taux de réponse empirique (non pondéré).

Le cas des enfants de moins de 9 ans est spécifique : comme ils ne sont pas à proprement parler interrogés par l'enquêteur, leur « réponse » dépend entièrement de l'attitude des parents. Ces enfants ne participent pas aux modèles et on convient donc, naturellement, de leur affecter la probabilité de réponse de la mère, ou à défaut celle du père.

Le phénomène de non-réponse est un phénomène qui se conçoit au niveau du ménage, du fait du protocole qui autorise le passage par des personnes « proxy ». Néanmoins, ce sont des individus que l'on suit dans le temps (les individus-panels) et l'approche formalisée correcte passe par l'utilisation de probabilités de réponse individuelles[6]. Par convention, on appelle individu répondant tout individu appartenant à un ménage répondant (DB135='1'). On

[6] Du moins pour toutes les dates antérieures à la date de collecte courante t, car à la date t il serait préférable de raisonner exclusivement au niveau ménage en introduisant une probabilité de réponse de l'entité ménage. Mais ce n'est pas l'approche qui a été retenue.

distingue essentiellement 4 raisons de perdre l'individu panel au cours du temps - et donc de diminuer $\theta_j(t,k)$:

- on abandonne définitivement le suivi de tout ménage refusant de répondre (à toute date) ;
- on abandonne définitivement le suivi de tout ménage impossible à joindre une année donnée dès lors qu'on a déjà été confronté à la même situation l'année précédente ;
- on perd la trace de l'individu-panel une année donnée ;
- certains individus-panels sortent du champ de l'enquête.

Le cas de sortie du champ n'est pas une non-réponse et se traite naturellement en ignorant purement et simplement l'individu-panel sorti du champ (estimation sur un domaine, celui des individus restant dans le champ).
Si on reprend les trois motifs de non-réponse, on peut considérer que certains déterminants de la non-réponse sont plutôt de nature individuelle, même si la majorité des variables explicatives sont relatives au ménage dans lequel se trouve l'individu. En particulier, une situation individuelle peut justifier (en partie) la perte de contact à un moment donné avec un individu-panel.

Le calcul de $\theta_j(t,k)$, probabilité que l'individu-panel j soit répondant l'année t sachant qu'il est répondant l'année de son échantillonnage, c'est-à-dire l'année $t-k+1$, dépend *a priori* essentiellement de k, et très peu de t dès lors qu'on considère des dates t (assez) voisines. Autrement dit, cette probabilité diminue naturellement lorsque le temps / délai k augmente, de manière *a priori* à peu près régulière et d'une façon qui n'a pas lieu de dépendre de l'année d'échantillonnage. C'est pourquoi les modèles sont estimés en cumulant les données de plusieurs années consécutives en maintenant constante la valeur de k, laquelle désigne la vague d'enquête. On estime donc autant de modèles qu'il y a de valeurs de k possibles, soit 8 en régime stationnaire. Ainsi en 2013, on ajuste 8 modèles :

- un modèle pour les individus interrogés en 2ème vague, utilisé pour estimer les probabilités $\theta_j(t,2)$:
  - sur les entrants 2004 encore dans le champ 2005
  - sur les entrants 2005 encore dans le champ 2006

  etc…
  - sur les entrants 2011 encore dans le champ 2012
  - sur les entrants 2012 encore dans le champ 2013

- un modèle pour les individus interrogés en 3ème vague, utilisé pour estimer les probabilités $\theta_j(t,3)$:
  - sur les entrants 2004 encore dans le champ 2006
  - sur les entrants 2005 encore dans le champ 2007

  etc…
  - sur les entrants 2010 encore dans le champ 2012
  - sur les entrants 2011 encore dans le champ 2013

De la même façon, on ajuste un modèle pour les individus interrogés en 4ème vague, etc… pour terminer par

- un modèle pour les individus interrogés en 8ème vague, utilisé pour estimer les probabilités $\theta_j(t,8)$:
  - sur les entrants 2004 réinterrogés en 2011
  - sur les entrants 2005 réinterrogés en 2012
  - sur les entrants 2006 réinterrogés en 2013

- un modèle pour les individus interrogés en 9ème vague, utilisé pour estimer les probabilités $\theta_j(t,9)$:
    - sur les entrants 2004 réinterrogés en 2012
    - sur les entrants 2005 réinterrogés en 2013

Le cumul d'années ne doit pas remonter trop loin dans le temps afin de ne pas risquer d'occulter d'éventuelles modifications structurelles du comportement de réponse dans le temps. Jusqu'à présent, pour la seconde vague en particulier (là où le risque est le plus fort), on a agrégé toutes les années disponibles, soit 9 années consécutives.

En première interrogation, on a également cumulé tous les sous-échantillons entrants depuis 2004 pour ajuster le modèle. Les variables explicatives retenues sont :

- la typologie dite « Tabard » des quartiers, communes et cantons de France issue du recensement 1999 (coordonnée sur l'axe 1 de l'ACP) ;
- la taille du ménage (ménage d'une personne seule, autres configurations) ;
- la localisation géographique : Ile-de-France, région Méditerranée (Languedoc-Roussillon, Paca, Corse), région Centre Est (Rhône-Alpes, Auvergne), autres régions.

En ré-interrogation, pour chacun des 8 modèles ci-dessus, on a sélectionné :

- le fait d'avoir déménagé avec l'ensemble de son ménage depuis la dernière vague ;
- le fait d'habiter dans une maison ou non ;
- un type de ménage (ménage d'une seule personne, couple avec un enfant, autres configurations) ;
- la localisation géographique (agglomération parisienne, autre) ;
- le quartile de niveau de vie du ménage (1er, 2nd ou au-dessus de la médiane) ;
- le fait qu'il y ait eu un ou plusieurs départs dans le ménage depuis la dernière vague ;
- la catégorie socioprofessionnelle de la personne de référence (plusieurs regroupements de modalités, selon le niveau de réinterrogation) ;
- avoir un contrat à durée indéterminée ou non ;
- la nationalité (plusieurs regroupements de modalités, selon le niveau de réinterrogation) ;
- le diplôme (plusieurs regroupements de modalités, selon le niveau de réinterrogation) ;
- l'âge en tranche ;
- la situation au regard de l'activité à la date de l'enquête (salarié, indépendant, autres situations).

Les 6 dernières variables sont des variables individuelles.

Il est à noter que ces variables sont actuellement celles de la première année d'enquête, quelle que soit la vague k considérée.

3. La pondération transversale finale

L'enquête annuelle est calée sur des marges obtenues à partir de l'enquête Emploi de l'année précédente. Ces marges sont les moyennes des estimations trimestrielles Emploi obtenues à partir des 4 trimestres de l'année antérieure.

L'enquête Emploi ignorant la notion de budget séparé, l'unité statistique en vigueur pour la phase de calage est le logement. Il est donc nécessaire, à partir des données SILC, de reconstituer un ménage-logement en déterminant au préalable une personne de référence. De fait, on abandonne au moment du calage la notion de ménage SILC et lorsqu'un

logement abrite plusieurs ménages, tous les ménages du logement ont le même poids après calage.

Les marges sont formées à partir des variables suivantes :

- nombre de ménages par tranche d'âge de la personne de référence (5 modalités, des moins de 31 ans aux 76 ans et plus). L'âge est mesuré au 31 décembre.
- nombre de ménages par tranche de densité d'habitat : rural, unité urbaine de moins de 20 000 habitants, unité urbaine de 20 000 à 100 000 habitants, unité urbaine de plus de 100 000 habitants, région parisienne.
- nombre de ménages par type :
  - personne seule,
  - couple sans enfant,
  - couple avec 1 enfant,
  - couple avec 2 enfants ou plus,
  - famille monoparentale,
  - autre configuration ;
- nombre d'hommes par tranche d'âge (6 modalités, des moins de 15 ans aux 76 ans et plus) ;
- nombre de femmes par tranche d'âge (6 modalités, des moins de 15 ans aux 76 ans et plus) ;
- nombre de ménages selon le diplôme de la personne de référence :
  - sans diplôme, non déclaré,
  - diplôme inférieur baccalauréat (CAP, BEPC),
  - baccalauréat, bac+2,
  - diplôme supérieur ;
- nombre de ménages selon la catégorie socioprofessionnelle de la personne de référence (activité actuelle ou ancienne activité) :
  - agriculteurs (retraités ou non),
  - indépendants et professions libérales (retraités ou non),
  - professeurs et instituteurs actifs,
  - professeurs et instituteurs retraités,
  - employés et ouvriers actifs,
  - employés et ouvriers retraités,
  - autres.

C'est la macro %Calmar, avec l'option Logit (rapports de poids bornés), qui est utilisée pour produire les poids calés.

Imputation

Depuis l'enquête réalisée en 2008, à de très rares exceptions près, les ménages ne sont plus questionnés sur le montant des revenus. Ces derniers sont obtenus à partir des fichiers fiscaux de la Direction Générale des Finances Publiques ou des organismes gestionnaires de prestations pour les revenus sociaux. Le rapprochement du fichier d'enquête et du fichier fiscal est opéré à l'aide de l'adresse du ménage et d'informations sur le déclarant (nom, prénom, date de naissance, département de naissance, pays de naissance). Dans le cas où les adresses ne correspondraient pas, les revenus ne seront pas retrouvés. C'est le cas des jeunes en particulier qui peuvent déclarer leur impôt avec leurs parents s'ils sont âgés de moins de 21 ans ou bien s'ils sont étudiants de moins de 25 ans. Si ces jeunes ont décohabité, le risque de ne pas retrouver leurs revenus est élevé, c'est pourquoi ils sont interrogés toujours suivant l'ancienne méthode (collecte complète des revenus et impôts en face à face).
Les prestations familiales et les minima sociaux sont obtenus auprès des organismes les détenant, c'est toujours l'adresse du ménage qui permet la recherche et les caractéristiques démographiques de l'allocataire (sexe, date de naissance) : pour diverses raisons certaines adresses ne correspondent pas, ces revenus sont alors imputés.

L'imputation est d'abord nécessaire parce qu'il existe des données manquantes ou en tranches. Les données manquantes proviennent des individus dont la déclaration fiscale n'a pu être retrouvée. Les données en tranches concernent les enfants rattachés fiscalement au foyer fiscal de leurs parents et ne résidant pas avec ceux-ci. Les revenus dans l'EU-SILC 2016 sont relatifs à plusieurs dates. Les revenus appariés et collectés sont relatifs à l'année 2015 et l'impôt payé a pour assiette les revenus imposables perçus au cours de l'année 2014. Nous décrivons maintenant les opérations concernant les revenus principaux.

1. Les deux méthodes retenues pour les imputations

L'imputation des revenus individuels est menée de deux façons différentes, selon que le ménage est enquêté pour la première fois ou non. Dans le premier cas, l'imputation est transversale : une équation du revenu est estimée sur les répondants et permet d'imputer le revenu des non-répondants (pour les jeunes) ou non appariés. Dans le second cas, nous utilisons le revenu donné par l'individu à une date précédente pour estimer le revenu manquant perçu à l'autre date. Pour ce faire, nous estimons une équation du ratio entre les revenus des deux années sur les répondants aux deux vagues. Ce ratio est ensuite estimé pour les individus n'ayant répondu qu'à une enquête afin d'attribuer le revenu manquant. Cette méthode est appliquée pour les imputations des salaires et des retraites.
Si les montants de revenus ou d'impôts ne sont presque plus collectés dans l'enquête, le type de revenu perçu et d'impôt payé est systématiquement demandé. Ces informations sont particulièrement utiles pour les imputations.

2. Salaire ou revenu assimilé (PY010N)

Il est nécessaire dans un premier temps de définir sur quelles données l'imputation va porter puis de comparer le salaire imputé avec les maxima observés dans l'ERFS en tenant compte du sexe du salarié et de sa catégorie socioprofessionnelle (sur une position). Un salaire est attribué aux individus déclarant en percevoir et ne figurant pas dans le fichier des impôts, ainsi qu'aux individus ayant répondu par un montant en tranches (cas des jeunes).
L'imputation est menée par strates. Huit strates sont créées à partir du sexe, de l'emploi, qualifié ou non, et du secteur d'emploi, privé ou public. Nous sélectionnons différentes variables pouvant expliquer le salaire dans chaque strate. Un tronc commun de variables explicatives est formé par l'ancienneté dans la profession, l'emploi atypique ou non et le diplôme du salarié. Pour les salariés du privé nous y ajoutons le type de contrat, le fait d'avoir un emploi en Île-de-France ou pas, la proportion de femmes dans le secteur et le fait d'être cadre ou pas. Enfin pour les salariés du public, en plus des variables du tronc commun, sont ajoutés : le fait d'être enseignant ou pas, fonctionnaire d'État ou pas et le grade.
Le salaire mensuel ou le ratio entre les salaires des deux années consécutives est imputé. Le nombre de mois d'activité déclaré à l'enquête est pris en compte pour estimer le salaire annuel. Un travail particulier est nécessaire pour les salariés à temps partiel.

3. Les salaires des non-salariés

Les salaires des non-salariés sont imputés à partir de salaires moyens de personnes ayant les mêmes caractéristiques.

4. Préretraites

Un petit nombre de préretraites sont à imputer : les imputations sont réalisées avec le montant moyen de préretraites d'individus ayant des caractéristiques similaires.

5. Allocations de vieillesse (PY100N) ou pension au survivant (PY110N)

Lorsque le montant de la pension est manquant, le montant de la retraite est imputé. Deux strates sont utilisées, suivant que le conjoint de la personne retraitée est vivant ou pas. Pour les personnes dont le conjoint n'est pas décédé les variables explicatives du montant de la retraite perçue sont le sexe, le secteur d'activité (privé ou public), la qualification, le diplôme,

l'âge, et l'ancienneté dans la profession. Pour les retraités dont le conjoint est décédé, ces variables sont complétées par le secteur d'activité de l'ex-conjoint ainsi que sa qualification.
Selon le rang d'interrogation de l'individu, la retraite ou le ratio des retraites des deux années consécutives est estimé afin d'imputer un montant.
Les retraites étant quasi stables, il est possible par ailleurs de contrôler le montant imputé. Comme nous ne disposons pas d'une retraite courante (relative au mois de l'enquête), nous utilisons le revenu courant comme élément de contrôle et nous le comparons à la somme des revenus de l'année de référence du ménage. Pour les ménages concernés, la retraite est un élément prépondérant du revenu total, ce qui justifie la comparaison. Nous ajoutons un autre contrôle, cette fois entre les années de revenu antérieures, toujours sous la même hypothèse nous comparons les impôts sur le revenu déclarés que nous calculons. Deux contrôles sont donc possibles avant de prendre une décision.

6. *Bénéfices en espèces ou perte de trésorerie en rapport avec une activité indépendante*

Ces revenus sont collectés sous deux formes : d'une part la forme fiscale, comprenant les amortissements et autres abattements, et d'autre part la forme privée, correspondant au revenu net déterminé par le ménage. Le revenu fiscal, parfois négatif, est jugé peu réaliste, c'est pourquoi le revenu privé collecté par voie d'enquête est privilégié (les prélèvements privés que la personne a effectué sur les ressources de son activité d'indépendant pour ses besoins de consommation ou d'épargne). Ainsi si les deux revenus sont renseignés et si le revenu privé est vraisemblable, le revenu privé est le seul pris en compte. À défaut de revenu privé, le revenu fiscal est retenu. Si les deux types de revenus sont manquants ou peu crédibles, l'imputation se fait par hot-deck.

7. *Prestations familiales*

Les prestations familiales sont, depuis la collecte 2008 sur les prestations 2007, obtenues par appariement auprès des organismes de prestations sociales. Dans le cas où l'allocataire n'a pu être apparié, les prestations sont calculées sur barème. La principale difficulté est la période de référence des revenus pour les prestations sous conditions de ressources. Pour les Caisses d'allocations familiales (Caf), jusqu'à juillet d'une année N, les revenus retenus pour le calcul des aides est celui de l'année N-2 ; à partir de juillet N, les revenus retenus sont ceux de N-1, Nous utilisons uniquement les revenus de l'année N pour imputer les prestations de l'année N.
Les résultats sont conformes aux données des CAF. Les valeurs calculées sont imputées si l'individu n'a pas été retrouvé dans les fichiers sociaux.

8. *Aides au logement*

Pour les allocataires non appariés, les aides au logement sont calculées sur barème pour tous les locataires et les accédants à la propriété. Nous utilisons le loyer déclaré à l'enquête, et à défaut, un loyer imputé. La masse des allocations collectées est inférieure de 10 % aux données de la Cnaf corrigées de la différence de champ (ménages en institutions). Des aides sont donc attribuées à certains ménages, de façon aléatoire, afin de disposer du bon nombre de bénéficiaires. Le calcul des aides pour les locataires ne pose pas de problème majeur même si, comme pour les prestations familiales, la période des revenus n'est pas exactement celle retenue par les CAF. Le calcul des aides aux accédants à la propriété diffère néanmoins du calcul des mêmes aides dans l'ERFS. Nous avons en effet choisi d'appliquer le barème locatif aux accédants, en nous aidant du loyer fictif imputé.

9. *Minima sociaux*

En cas de non-appariement, le montant du Revenu de Solidarité Active (RSA) est imputé sur barème si la personne a déclaré être bénéficiaire. Un calage sur les sources Cnaf est réalisé.
Pour les allocataires non appariés, trois minima sociaux sont imputés dans SILC : le RSA, l'ASPA ou ex-minimum vieillesse et l'allocation de parent isolé (API). Selon les données brutes, SILC comprend 90 % des bénéficiaires du RSA et 60 % des bénéficiaires du minimum vieillesse.

Les méthodes utilisées sont proches des méthodes d'imputation utilisées dans l'ERFS. Elles présentent une limite. Le revenu retenu pour le calcul du RSA est un revenu trimestriel que nous ne connaissons pas. Le RSA imputé est égal à la différence entre le plafond du RSA et les revenus de l'année. Ce plafond dépend du type de famille et du nombre de personnes à charge. Ainsi calculé, aucun minimum ne peut être imputé à une famille dont les revenus annuels sont supérieurs au plafond. Cette famille a pourtant pu être éligible, si les revenus d'un trimestre se sont avérés insuffisants.
N'ont pas été traités les dispositifs d'intéressement à la prise ou à la reprise d'un emploi. Néanmoins, une étude sur l'impact de l'intéressement du RSA dans l'enquête ERFS a montré que les taux de pauvreté restaient inchangés.

10. Loyer imputé

Un loyer fictif est imputé aux propriétaires, aux accédants, aux usufruitiers, aux logés gratuitement et aux locataires du parc social payant un loyer inférieur au prix du marché.

La méthode retenue comprend quatre étapes, dont deux régressions :

1) Estimation d'une équation de loyer à partir des données de l'enquête logement 2006. L'estimation porte sur les logements locatifs du parc privé hors loi de 1948. Les variables explicatives sont les caractéristiques du logement (surface, confort, sanitaires, équipement, état du logement …) de localisation (tranche de taille d'agglomération, zone climatique, typologie socio-économique de Nicole Tabard …) et le quintile de niveau de vie. Deux équations distinctes sont estimées, l'une pour les appartements (variable expliquée : le logarithme du loyer au m²) et l'autre pour les maisons (variable expliquée : le logarithme du loyer, la surface figurant parmi les variables explicatives)
2) Les équations précédemment estimées sont utilisées pour imputer un loyer fictif aux propriétaires occupants ainsi qu'aux ménages logés gratuitement et un loyer de marché aux locataires du parc social ou en Loi de 1948 de l'enquête logement. On a rajouté à la valeur issue de l'équation un résidu tiré selon une procédure de hot-deck stratifié.
3) Ce loyer imputé est régressé sur deux types de variables : des variables du tronc commun des enquêtes ménages de l'Insee d'une part, et des variables géographiques d'autre part. À caractéristiques sociodémographiques et de localisation identiques, les logements occupés par les accédants à la propriété sont d'une qualité moyenne supérieure à ceux des propriétaires sans charge de remboursement, qui sont eux-mêmes de meilleure qualité que ceux du parc social. Estimer une seule équation aurait pu biaiser les estimations. Aussi huit régressions distinctes ont-elles été estimées sur des segments relativement homogènes du parc :
   - appartements, propriétaires sans charge de remboursement et ménages logés gratuitement
   - appartements, accédants à la propriété ;
   - appartements, locataires du parc social ou loi de 1948 ;
   - appartements, locataires du parc libre louant vide ;
   - maisons, propriétaires sans charge de remboursement et ménages logés gratuitement ;
   - maisons, accédants à la propriété ;
   - maisons, locataires du parc social ou loi de 1948 ;
   - maisons, locataires du parc libre louant vide.
4) Les huit équations estimées sont exportées vers l'enquête SILC pour y imputer soit un loyer fictif soit un loyer manquant. Lors de l'imputation on rajoute à la valeur prédite un résidu tiré selon une procédure de hot-deck stratifié.

Glossaire

L’unité de vie – ménage

L’enquête SRCV utilise la définition européenne du ménage, qui s‘appuie davantage sur la communauté de budgets que sur l’occupation d’un même logement. Ainsi deux unités de vie faisant budgets séparés au sein d’un même logement renseigneront deux questionnaires ménages.
Est considéré comme un ménage l'ensemble des personnes (apparentées ou non) qui partagent de manière habituelle un même logement et qui font budget commun, c'est-à-dire :

1) qui apportent des ressources servant à des dépenses faites pour la vie du ménage ;
2) et/ou qui bénéficient simplement de ces dépenses.

- Dans la définition du budget commun, on ne tient pas compte des dépenses faites pour le logement ;
- La participation occasionnelle à des dépenses communes ne suffit pas à former un budget commun ;
- Avoir plusieurs comptes en banque différents dans un ménage ne signifie pas faire budget séparé.

La personne de référence

La personne de référence du ménage au sens des enquêtes ménages est la personne qui apporte le plus de ressources.

Le couple

Un couple, au sens des enquêtes ménages, est composé de deux personnes de 15 ans ou plus, habitant le même **logement**, déclarant actuellement être en couple, quel que soit leur état matrimonial légal (qu'elles soient donc mariées ou non).

Le revenu disponible et le niveau de vie

Le **revenu disponible** d'un ménage comprend les revenus d'activité, les revenus du patrimoine, les transferts en provenance d'autres ménages et les prestations sociales (y compris les pensions de retraite et les indemnités de chômage), nets des impôts directs. Quatre impôts directs sont pris en compte : l'impôt sur le revenu, la taxe d'habitation et les contribution sociale généralisée (CSG) et contribution à la réduction de la dette sociale (CRDS).

Le **niveau de vie** est égal au revenu disponible du ménage divisé par le nombre d'unités de consommation (UC). Le niveau de vie est donc le même pour tous les individus d'un même ménage. Les unités de consommation sont généralement calculées selon l'échelle d'équivalence dite de l'OCDE modifiée qui attribue 1 UC au premier adulte du ménage, 0,5 UC aux autres personnes de 14 ans ou plus et 0,3 UC aux enfants de moins de 14 ans.