



COMPARAISON DE DIFFÉRENTS SCÉNARIOS DE CALAGE POUR L'ENQUÊTE SRCV

Robin Chatelot(), Maxime Bergeat (**)*

() Insee, Département des ressources et des conditions de vie des ménages*

*(**) Insee, Département des ressources et des conditions de vie des ménages*

robin.chatelot@insee.fr

Mots-clés : Calage, marges de calage, SRCV, bootstrap

Domaine concerné : Théorie des sondages, après collecte – Calage

Résumé

L'ajout d'une marge de calage sur laquelle caler ses poids peut conduire à un dilemme. D'un côté, ajouter une marge permet de s'assurer que les poids permettent de retrouver la valeur de cadrage. Mais de l'autre, l'ajout de cette marge supplémentaire nécessite parfois de déformer fortement les poids. Ce dilemme est particulièrement prégnant dans le cas où le nombre de marges utilisé est trop élevé et/ou la taille de l'échantillon trop faible : faire converger le calage sans revoir les critères de qualité du calage à la baisse peut alors s'avérer impossible. Des contraintes trop fortes sur un calage peuvent donc amener à des ruptures de série (par exemple dans des cas où il est nécessaire de faire varier les bornes de calage d'une année sur l'autre) alors que l'un des intérêts majeurs du calage est de les éviter.

L'enquête « Statistiques sur les ressources et les conditions de vie » (SRCV) est réalisée chaque année par l'Insee. Elle porte sur environ 30 000 individus de 16 ans ou plus, interrogés pendant 4 années consécutives. Cette enquête européenne permet notamment de calculer plusieurs indicateurs de pauvreté ou d'exclusion sociale, tels que la pauvreté monétaire et la privation matérielle et sociale. Les revenus sont collectés par appariement avec des sources administratives (sources fiscales pour les revenus déclarés à l'administration fiscale, sources sociales de la Cnaf, la CCMSA et la Cnav pour les prestations sociales perçues par les ménages). La plupart des variables sur les ressources sont construites à partir des données administratives et les données du questionnaire sont utilisées pour les compléter (le taux d'appariement avec les données fiscales est de 97 % pour les individus de 15 ans et plus).

Aujourd'hui, l'enquête utilise des marges de calage :

- sociodémographiques (estimées à partir de l'enquête emploi en continu, ces marges sont communes à plusieurs enquêtes ménages de l'Insee) ;
- fiscales (masses de revenus estimées à partir des fichiers fiscaux avec lesquels l'enquête SRCV est appariée).

Par ailleurs, depuis le millésime 2020 de l'enquête, le calage est également réalisé sur des distributions de revenu déclaré par unité de consommation, estimées à partir du dispositif Filosofi ou de l'enquête Revenus fiscaux et sociaux.

Les marges utilisées dans l'enquête SRCV sont nombreuses et peuvent conduire à des distorsions de poids importantes. Le but de cet article est d'étudier des scénarios alternatifs de calage de l'enquête SRCV et d'analyser différents critères permettant de choisir les marges à utiliser pour le calage de SRCV. Les tests sont réalisés à partir de l'enquête SRCV 2024 en se concentrant sur les poids transversaux, sur le champ de la France métropolitaine. Les critères retenus pour comparer les scénarios alternatifs de calage sont de plusieurs natures pour déterminer quel est le scénario le plus prometteur :

- étude de critères sur la structure des poids issus du calage : répartition des rapports de poids après/avant calage, accumulation des rapports de poids au bord des bornes utilisées dans le calage
- étude de critères plus conceptuels comme la source des marges utilisées pour le calage, leur champ, ainsi que leurs potentielles variabilités dans le temps. Ces critères sont difficiles à quantifier mais sont essentiels à la construction d'un jeu de marges pertinent ;
- estimation de la variabilité des variables d'intérêt selon les différents poids calés testés à l'aide d'une méthode par bootstrap. Cela permettra d'avoir une idée de la « robustesse » des différents scénarios de calage testés. Il sera aussi important de comparer les résultats obtenus avec ceux des jeux de poids utilisés actuellement dans l'enquête, pour estimer les risques de ruptures de série.

La comparaison entre les différents scénarios de calage testés est effectuée en étudiant l'indicateur de privation matérielle et sociale, indicateur central de l'enquête SRCV. Des différences sur le niveau de l'indicateur sont observées selon le scénario de calage retenu : en particulier, le recours à des marges de statut d'occupation ou sur la distribution de revenus va de pair avec un indicateur plus élevé de privation. La précision de ces estimations est ensuite estimée à partir d'une approche par répliques bootstrap : il est difficile de mettre en évidence des différences significatives selon les scénarios de calage testés.

Abstract

Choosing appropriate margins for data calibration is difficult because the criteria for selecting a given set of margins are diverse. The aim of this paper is to study several criteria in order to compare different sets of calibration margins for the European Union Statistics on Income and Living Conditions (EU-SILC) survey.

The criteria can be grouped into three different categories. The first concerns the structure of the weights: the weights ratios should not be concentrated at the extremes and the weights should not be too variable. The second category relates to precision and bias of the resulting estimates. Precision is assessed here using a bootstrap approach, while evaluating the bias is more complex. The final category covers more conceptual criteria such as the consistency of the margins with the survey scope, their availability, and their stability over time.

1. Contexte et cadrage de l'étude

L'enquête « Statistiques sur les ressources et les conditions de vie » (SRCV) est réalisée chaque année par l'Insee. Elle porte sur environ 30 000 individus de 16 ans ou plus, interrogés pendant 4 années consécutives. Cette enquête européenne permet notamment de calculer plusieurs indicateurs de pauvreté ou d'exclusion sociale, tels que la pauvreté monétaire et la privation matérielle et sociale.

Les revenus sont collectés par appariement avec des sources administratives (sources fiscales pour les revenus déclarés à l'administration fiscale sources sociales de la Cnaf, la CCMSA et la Cnav pour les prestations sociales perçues par les ménages). La plupart des variables sur les ressources sont construites à partir des données administratives et les données du questionnaire sont utilisées pour les compléter (le taux d'appariement avec les données fiscales est de 97 % pour les individus de 15 ans et plus).

Aujourd'hui, l'enquête utilise des marges de calage :

- sociodémographiques (estimées à partir de l'enquête emploi en continu, ces marges sont communes à plusieurs enquêtes ménages de l'Insee) ;
- fiscales (masses de revenus estimées à partir des fichiers fiscaux avec lesquels l'enquête SRCV est appariée).
- de répartition des revenus (vingtiles fournis par Filosofi dans un premier temps puis par l'ERFS suite à des contraintes de calendrier)

L'enquête SRCV est la partie française du dispositif EU-SILC (*European Union – Statistics on Income and Living Conditions*), qui s'inscrit dans le programme d'action communautaire de lutte contre l'exclusion sociale. Ce dispositif permet des comparaisons des distributions de revenus et des conditions de vie entre États membres de l'Union européenne. Depuis 2021, l'enquête SRCV est ainsi régie par le règlement-cadre IESS (*Integrated European Social Statistics*), qui définit les variables à fournir chaque année et donne des consignes sur la méthodologie à suivre pour l'enquête (critères sur la précision requise pour les indicateurs principaux calculés avec l'enquête, notamment).

1.1. Variables d'intérêt de l'enquête

SRCV comporte des variables d'intérêt au niveau des ménages et au niveau des individus. Le questionnaire est donc divisé en un questionnaire ménage suivi d'un questionnaire individu pour chaque membre du ménage à partir de 16 ans.

Les principales variables d'intérêt de l'enquête estimées chaque année sont la pauvreté monétaire, la privation matérielle et sociale, le taux de risque de pauvreté persistante et le risque de pauvreté ou d'exclusion sociale (ARPE ou at risk of poverty or social exclusion).

En plus des variables d'intérêt que l'on retrouve chaque année, un certain nombre de variables sont présentes dans des modules tournants. Les modules tournants sont des modules qui sont ajoutés périodiquement, tous les 3 ans ou tous les 6 ans, à l'enquête afin d'étudier différentes thématiques. Il arrive aussi que des modules *ad hoc* soient ajoutés à l'enquête, ces modules ne sont posés qu'une fois.

Les divers modules tournants et leur calendrier sont les suivants :

Année de passage	Modules tous les 3 ans	Modules tous les 6 ans	Module <i>ad hoc</i>
2020	Travail et conditions de logement	Surendettement, patrimoine et consommation	

2021	Santé et privations des enfants		Familles recomposées ou enfants de parents séparés
2022	Santé	Qualité de vie, participation sociale et culturelle	
2023	Travail et conditions de logement	Transmission des inégalités – Difficultés pour se loger	Efficacité énergétique du logement
2024	Santé et privations des enfants	Accès aux services	
2025	Santé		Énergie et environnement
2026	Travail et conditions de logement	Surendettement, patrimoine et consommation	
2027	Santé et privations des enfants		Santé mentale et bien-être
2028	Santé	Qualité de vie, participation sociale et culturelle	
2029	Travail et conditions de logement	Transmission des inégalités – difficultés pour se loger	Disponibilité des logements et accessibilité aux services

Le taux de pauvreté monétaire correspond à la proportion d'individus dont le niveau de vie est inférieur à un seuil, dénommé seuil de pauvreté. Le seuil de pauvreté est déterminé de façon relative par rapport à la distribution des niveaux de vie de l'ensemble de la population (60 % du niveau de vie médian). Le taux de risque de pauvreté persistante est quant à lui calculé en rapportant les personnes en situation de pauvreté monétaire une année *N* et également au moins 2 des 3 années précédant l'année *N* à l'ensemble de la population. Le taux de risque de pauvreté ou d'exclusion sociale calcule la part d'individus qui sont soit :

- en situation de pauvreté monétaire ;
- en situation de privation matérielle et sociale sévère (c'est-à-dire déclarant au moins 7 privations sur les 13 mesurées dans l'enquête, voir ci-dessous) ;
- vivant dans un ménage à très faible intensité de travail pour l'ensemble des individus (c'est-à-dire vivant dans un ménage où les membres en âge de travailler ont travaillé en moyenne moins de 20 % de leur potentiel total au cours des douze mois précédents). Une personne en âge de travailler est une personne âgée de 18 à 59 ans, à l'exception des étudiants âgés de 18 à 24 ans.

Les tests réalisés dans cette étude portent uniquement sur la variable de privation matérielle et sociale (PMS). Cet indicateur est privilégié car il est indépendant des variables sur les revenus. En effet, le calcul des poids de la chaîne de production de SRCV se déroule en trois temps majeurs : le traitement de la non-réponse totale et le calage des poids sur marges socio-démographiques, imputation des variables liées aux marges fiscales et aux vingtiles de revenus, calage des poids sur la totalité des marges. Les imputations des revenus qui ne concernent pas les marges utilisées (composantes du revenu déclaré) sont faites par la suite. Ces imputations de revenus dépendent des poids des ménages et individus interrogés. Se fonder sur l'indicateur de PMS permet de ne pas relancer les étapes d'imputations des revenus pour ce travail.

Une personne est en situation de privation matérielle et sociale si elle est concernée par au moins 5 critères parmi cette liste de 13 :

- Ne pas avoir de voiture personnelle pour des raisons financières ;
- Ne pas pouvoir maintenir son logement à bonne température pour des raisons financières ;
- Avoir des impayés de mensualités d'emprunts, de loyer ou de factures d'électricité, d'eau ou de gaz ;
- Ne pas pouvoir dépenser une petite somme d'argent pour soi sans avoir à consulter quiconque ;
- Ne pas pouvoir faire face à des dépenses inattendues ;
- Ne pas avoir accès à internet pour des raisons financières ;
- Ne pas pouvoir retrouver des amis ou de la famille au moins une fois par mois pour boire un verre ou pour un repas pour des raisons financières ;
- Ne pas avoir une activité de loisir régulière pour des raisons financières ;
- Ne pas pouvoir s'offrir une semaine de vacances hors de son domicile pour des raisons financières ;
- Ne pas avoir deux paires de bonnes chaussures pour des raisons financières ;
- Ne pas pouvoir avoir un repas contenant des protéines au moins tous les deux jours pour des raisons financières ;
- Ne pas pouvoir acheter des vêtements neufs.

Certaines questions sont posées au niveau du ménage, et d'autres au niveau des individus. Ainsi, il se peut que certains individus d'un ménage soient en situation de privation matérielle et sociale, et d'autres non [1].

L'objectif principal de l'étude est de comparer la volatilité des estimations de la PMS selon différents scénarios de calage (c'est-à-dire selon plusieurs jeux de marges à utiliser afin de réaliser le calage). Une attention particulière sera aussi portée aux risques de ruptures de séries qui pourraient découler des changements dans les marges de calage utilisées. Les poids utilisés en entrée du calage sont identiques pour tous les scénarios testés ici (les modèles de correction de la non-réponse totale ne sont pas relancés).

1.2. Calcul des pondérations de l'enquête

Plusieurs jeux de poids sont calculés dans l'enquête SRCV. En particulier, pour une année N de collecte, quatre jeux de pondérations sont estimés :

- Un poids pour le calcul en coupe des indicateurs transversaux de l'enquête, comme le taux de pauvreté monétaire ou le taux de privation matérielle et sociale
- Des poids pour les indicateurs longitudinaux permettant de capter des évolutions individuelles des individus panel
 - entre l'année $N-1$ et N
 - entre l'année $N-2$ et N
 - entre l'année $N-3$ et N

Par ailleurs, des poids sont calculés pour être utilisés sur des sous-populations. Cela concerne l'exploitation des questions relatives à la garde des enfants, ainsi que les questions sur le bien-être, qui ne sont pas posées en cas de réponse par proxy (en cas d'absence d'un individu du ménage pour répondre à un questionnaire individuel, une autre personne du ménage peut répondre à sa place).

On se concentre dans ce papier sur le calage réalisé pour les poids transversaux, qui sont les poids mobilisés pour les estimations de la privation matérielle et sociale.

Les différents scénarios de calage testés sont sur le champ de la France métropolitaine, les DROM ne sont pas inclus. À l'heure actuelle, le calage de l'enquête est réalisé séparément pour les DROM : il

comporte moins de contraintes que celui réalisé pour la France métropolitaine. Des tests seront réalisés à l'avenir pour tenter de faire un calage sur l'ensemble de l'échantillon avec les mêmes jeux de marges entre les DROM et l'hexagone, mais cela sort du cadre de cette étude.

L'étude porte donc sur les estimations de la PMS sur la France métropolitaine avec les poids transversaux.

1.3. Scénarios de calage

Les marges alternatives choisies pour ce test correspondent à :

- Des marges qui sont recommandées par Eurostat
- Et/ou des regroupements de marges utilisés pour le calage de l'Enquête sur les Revenus Fiscaux et Sociaux (ERFS), qui permet d'estimer la pauvreté monétaire.

La méthode utilisée pour le calage est la méthode logit, avec des bornes pour fixer les rapports de poids post/pré calage qui sont les mêmes pour chaque scénario (bornes fixées à 0,25 et 4). Il serait possible de chercher à étudier l'impact des bornes et des différentes méthodes de calage avec les tests qui vont suivre mais fixer ces paramètres de calage permet de se concentrer sur la comparaison des jeux de marges utilisés uniquement.

Tous les scénarios comportent un jeu de marges sociodémographiques qui sont habituellement utilisées dans le calage des poids de l'enquête SRCV :

- type de ménage : 5 modalités ;
- sexe x âge : 12 modalités ;
- niveau de diplôme : 4 modalités ;
- PCS : 7 modalités ;
- zone géographique : 3 modalités ;
- nationalité : 2 modalités ;
- quartier prioritaire de la ville (QPV) : 2 modalités.

Les marges sociodémographiques sont calculées à partir de l'Enquête emploi en continu, et sont communes à plusieurs enquêtes ménages de l'Insee portant sur les individus vivant dans des logements ordinaires.

À ce scénario de calage minimal, sont ajoutées différentes combinaisons de jeux de marges que l'on souhaite tester.

Les marges qui sont testées sont les suivantes :

- marges fiscales, ce jeu de marges sera appelé « marges fiscales détaillées » ou simplement « marges fiscales » par la suite ;
 - masse de salaires ;
 - masse d'allocations chômage ;
 - masse de revenus d'activité indépendante ;
 - masse de pensions de retraite ;
 - masse de revenus financiers ;
 - masse de revenus fonciers ;
- marges fiscales agrégées de manière similaire à l'ERFS, ce jeu de marges sera appelé « marges fiscales agrégées » par la suite ;
 - masse de salaire et chômage (agrégats regroupés en une seule marge) ;
 - masse de revenus d'activité indépendante ;
 - masse de pensions de retraite ;
 - masse de revenus financiers et fonciers (agrégats regroupés en une seule marge) ;
- masse d'impôts, une marge qui n'est pas utilisée pour le calage de l'ERFS mais pour SRCV actuellement ;

- vingtiles de revenus déclarés à l'administration fiscale (impact fort soupçonné sur les indicateurs et sur la variance des estimations) ;
- statut d'occupation (locataire ou propriétaire, marge faisant partie des recommandations Eurostat).

Les marges fiscales sont estimées à partir des fichiers exhaustifs fiscaux, après application de coefficients de correction pour se ramener au champ de l'enquête, qui est celui des ménages ordinaires. Ces coefficients, compris entre 0 et 1, sont estimés à partir du dispositif Filosofi pour chaque agrégat. Par exemple, le coefficient « ménages ordinaires » est plus faible que les autres pour l'agrégat « retraites », car un nombre plus important de retraités que de salariés vit hors logement ordinaire, en institutions (EHPAD notamment).

La marge utilisée pour l'agrégat « impôt », actuellement mobilisée dans le calage de l'enquête, est le fruit d'une estimation, qui mobilise les données fiscales N et $N-1$ pour reconstituer l'impôt effectivement payé une année N donnée par les individus de l'enquête, disponible pour les individus interrogés (et non l'impôt dû au titre des revenus de l'année N). L'agrégat n'est pas recalculé sur l'ensemble des données fiscales au vu de la volumétrie des données. La marge utilisée aujourd'hui est estimée à partir d'hypothèses permettant d'estimer l'impôt payé une année N pour l'ensemble des foyers fiscaux.

Tous les scénarios incluent les marges sociodémographiques ainsi que les marges fiscales ou bien les marges fiscales agrégées, avec des combinaisons différentes des trois autres marges (impôts, vingtiles de revenu déclaré, statut d'occupation du logement). Un calage sur marges sociodémographiques uniquement est ajouté.

Les différents scénarios testés sont donc les suivants :

Nom du scénario	Sociodémographiques	Fiscales détaillées	Fiscales agrégées	Impôts	Vingtiles	Statut d'occupation
minimal	X					
Scénarios avec marges fiscales détaillées						
D1	X	X				
D2	X	X		X		
D3	X	X			X	
D4	X	X				X
base	X	X		X	X	
D6	X	X		X		X
D7	X	X			X	X
D8	X	X		X	X	X
Scénarios avec marges fiscales agrégées						
A1	X		X			
A2	X		X	X		
A3	X		X		X	
A4	X		X			X
A5	X		X	X	X	
A6	X		X	X		X

A7	X		X		X	X
A8	X		X	X	X	X

Le scénario actuel de calage est identifié en vert et sera identifié comme « base » dans les graphiques. Le scénario avec seulement les marges socio-démographiques sera référencé comme le scénario « minimal ».

2. Étude des poids et bootstrap

Cette partie analyse les poids obtenus selon les 17 scénarios de calage comparés dans l'étude. Dans un premier temps, la structure des poids obtenus pour l'échantillon de l'enquête SRCV 2024 est comparée dans les différents scénarios, ainsi que leur impact sur l'estimation de l'indicateur de privation matérielle et sociale. Ensuite, *via* une approche par répliques bootstrap, on étudie des critères complémentaires, permettant notamment de donner des estimateurs d'intervalles de confiance pour la privation matérielle et sociale. L'approche par bootstrap est notamment utile pour les indicateurs complexes, cela pourra servir plus tard pour d'autres indicateurs que la PMS présents dans l'enquête SRCV.

Un exemple d'indicateur complexe de l'enquête SRCV est l'indicateur de pauvreté monétaire. En effet, pour le calculer, il faut calculer le seuil de pauvreté dans un premier temps puis observer qui est en dessous de 60 % de ce seuil. Toutefois, les analyses effectuées pour la pauvreté monétaire n'ont pas encore été autant poussées que celles sur la PMS car plus difficiles à interpréter : en effet, les imputations des ressources monétaires, effectuées préalablement au calcul des seuils et taux de pauvreté, dépendent en partie des poids obtenus après calage.

2.1. Premiers éléments quantitatifs

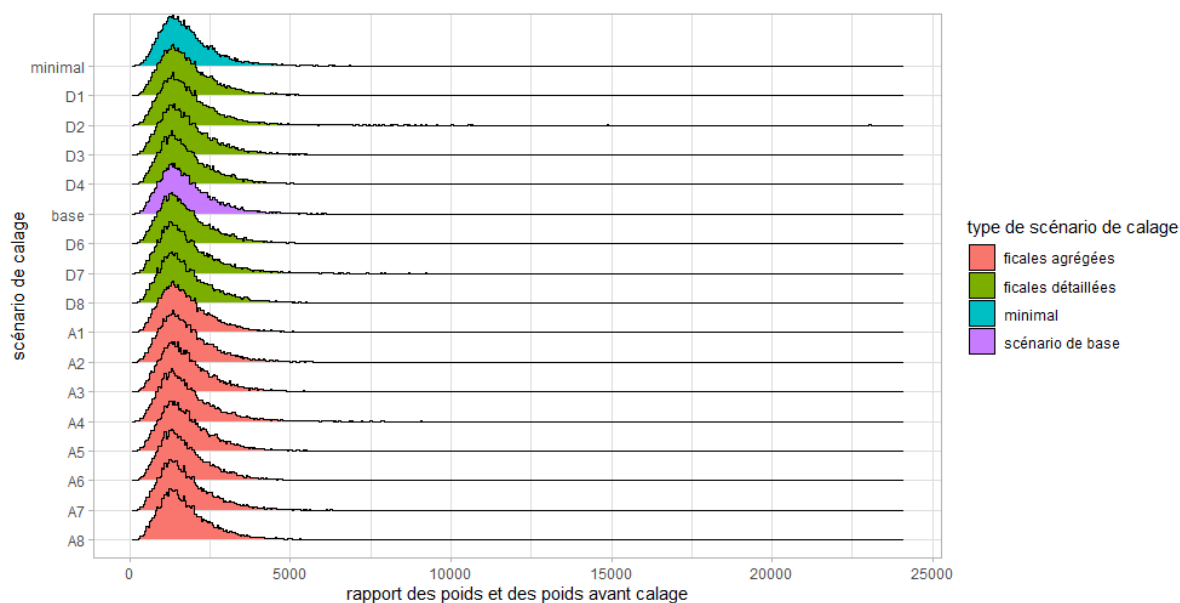


Figure 1 : Répartition des poids selon les scénarios de calage

La répartition des poids obtenus ne diffère pas fondamentalement d'un scénario de calage à un autre. La moyenne des poids est de 1796 pour l'ensemble des scénarios. Le premier quartile des poids se situe entre 1153 et 1187 et le troisième entre 2186 et 2208. Il est aussi possible de s'intéresser aux rapports entre le troisième et le premier quartile pour représenter la dispersion des poids.

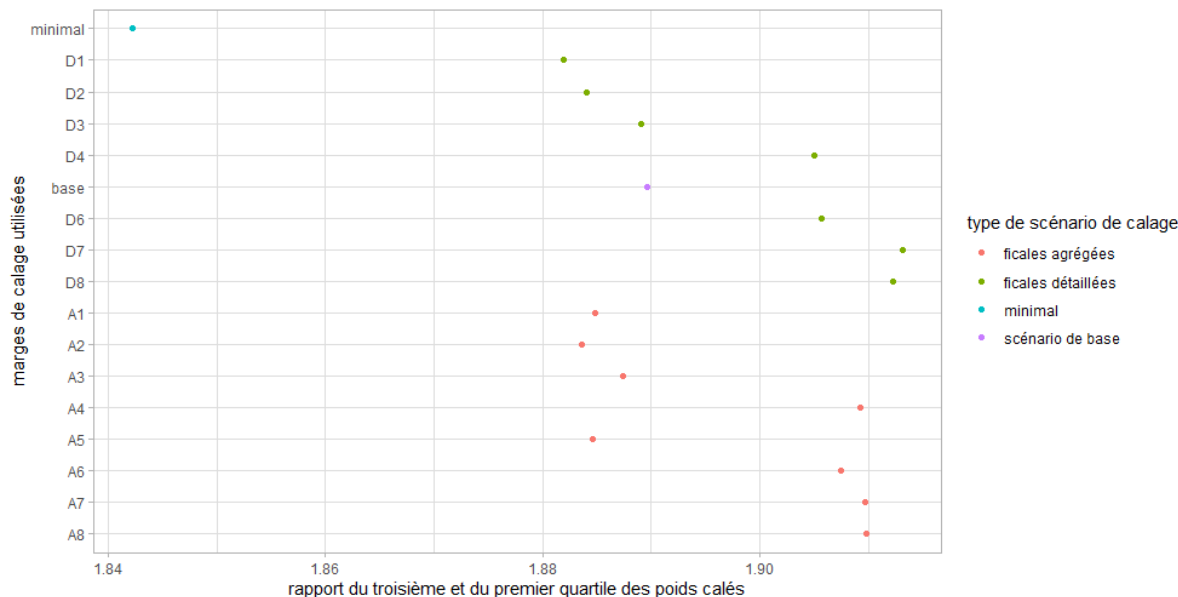


Figure 2 : rapport entre le troisième et le premier quartile des poids selon les scénarios de calage utilisés

Trois groupes se démarquent dans ce graphique :

- le scénario avec seulement les marges socio-démographiques avec une dispersion des poids plus faible probablement liée au nombre de marge plus faible ;
- les scénarios contenant la marge de statut d'occupation, avec une dispersion des poids plus importante ;
- les autres scénarios.

Observer une dispersion des poids plus faible pour un scénario de calage donné n'est pas un gage de qualité des poids. Cependant, la dispersion des poids selon les marges de calage utilisées pourra entrer en compte pour interpréter ou comparer les jeux de poids entre eux quand les autres critères de comparaison auront été étudiés.

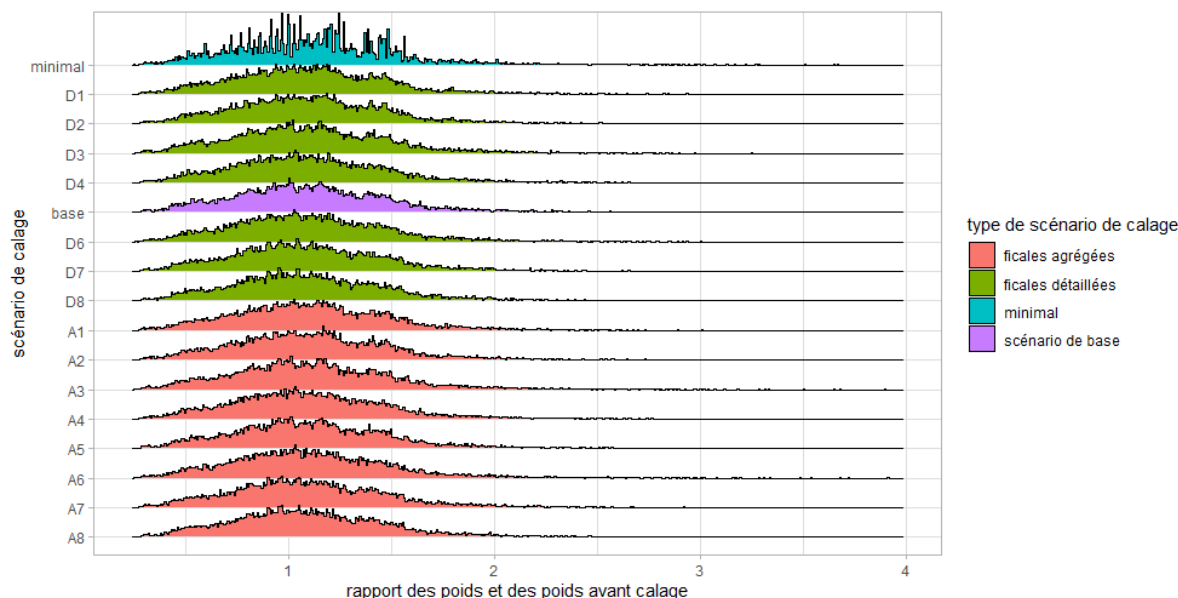


Figure 3 : rapport des poids après et avant calage selon les marges de calage utilisées

De même, les rapports des poids après et avant calage diffèrent peu entre les scénarios de calage. Les rapports de poids du calage avec le scénario mobilisant les marges sociodémographiques uniquement présentent toutefois des modes que l'on n'observe pas pour les autres scénarios. Ceci s'explique par le fait que ce scénario utilise uniquement des marges qualitatives, alors que les autres

scénarios mobilisent aussi des variables continues. Les moyennes des rapports des poids sont proches selon le scénario de calage, entre 1,117 et 1,123, et les quantiles également, entre 0,841 et 0,861 pour le premier quantile et 1,345 et 1,359 pour le troisième.

En plus de la structure des poids, il est intéressant de regarder pour chaque jeu de marges l'estimation obtenue pour certaines variables d'intérêt. Une estimation éloignée de celle obtenue avec le scénario de base, correspondant à ce qui est actuellement implémenté, n'est pas nécessairement mauvaise. Mais faire le choix d'un scénario de calage qui donne des estimations éloignées de celles du scénario de base amène à devoir estimer l'impact de la rupture de série ou d'assumer ladite rupture de série auprès des utilisateurs.

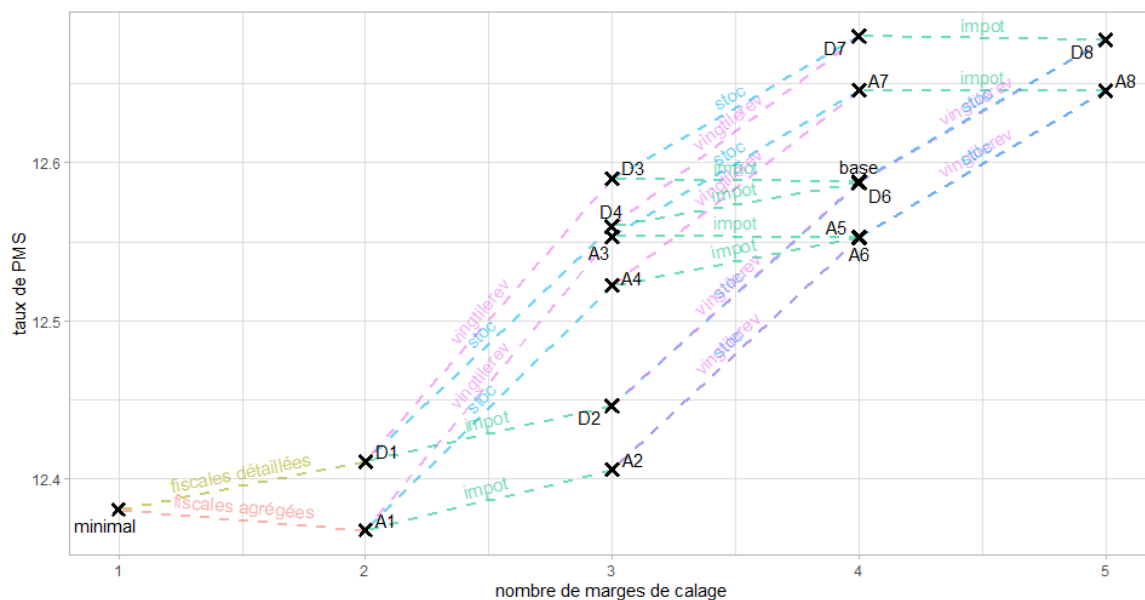


Figure 4 : Estimations de la PMS selon le nombre de marges du scénario de calage

Le scénario de calage actuel est identifié par le libellé « base ». Il comporte 4 marges et on peut faire aboutir à ce scénario avec plusieurs chemins selon l'ordre dans lequel les marges sont ajoutées. Avec ce scénario de calage, le taux de privation matérielle et sociale est de 12,58 %.

La figure 4 présente la valeur de l'indicateur de privation matérielle et sociale selon le scénario de calage utilisé. Le point le plus à gauche correspond au calage avec jeu de marges socio-démographiques uniquement, auxquelles les marges supplémentaires sont ajoutées successivement. Le texte au-dessus des lignes tracées entre les points correspond à la marge qui a été ajoutée par rapport au point plus à gauche. Ainsi, plus le point est à droite du graphique, plus le scénario associé utilise de marges.

L'effet d'ajouter une marge au jeu de calage a la majorité du temps un effet à la hausse sur l'estimation de la privation matérielle et sociale. Une explication possible pourrait être que le taux de réponse des personnes ayant de plus faibles ressources serait plus faible, et que cela ne serait pas complètement capté et corrigé par le traitement de la non-réponse totale de l'enquête. Il est possible de penser que les informations fournies par le calage permettent de corriger cela, c'est notamment l'une des intuitions qui avait été mises en avant lors de l'ajout des marges de vingtiles de revenus déclarés dans SRCV 2020.

L'ajout des marges de vingtiles de revenu déclaré a un effet fort à la hausse sur l'estimation de la PMS, de l'ordre de 0,14 point de pourcentage sur la valeur de l'indicateur, de même que les marges de statut d'occupation du logement (0,12 point). La marge d'impôt a un effet plus faible que les autres. Les deux jeux de marges fiscales ont un effet modéré, mais de sens contraire : Le seul jeu de marges à

avoir un effet à la baisse est celui du jeu de marges fiscales agrégées. Il a ainsi un effet opposé à celui des marges fiscales détaillées, alors que leurs différences sont uniquement liées à l'agrégation ou non de certaines composantes des ressources. Ainsi, agréger les marges de salaire et de chômage et les marges des revenus financiers et fonciers diminue le taux de PMS estimé dans la population.

Les poids calculés avec les marges agrégées pourraient conduire à valoriser plus les salaires que le chômage. Cependant les masses de salaires et d'allocations chômage respectivement obtenues sont proches des cibles, y compris avec la marge agrégée.

En revanche, les poids calculés avec les marges fiscales agrégées aboutissent à une proportion sensiblement plus grande de masse monétaire sur le foncier (par rapport au financier) que les marges fiscales détaillées. On observe en moyenne une hausse de 1,5 milliards d'euros de la masse de revenus fonciers pour les scénarios avec marges fiscales agrégées par rapport à leur équivalent avec marges fiscales détaillées. Les marges sur les revenus fonciers et financiers étant combinée dans ces scénarios, la masse de revenus financiers baisse de 1,5 milliards en moyenne par un effet de vases communicant.

Outre les effets des marges de calage sur les estimations, un autre critère de choix peut être les effets de ces marges sur la précision des estimations. Pour les analyser, une méthode d'estimation de la précision par bootstrap est mise en place. Ici, le but n'est pas de fournir une estimation précise de la variance des estimations, mais d'avoir des estimations de la variance comparables entre les différents scénarios de calage. Dans un cadre où des estimations plus fiables sont requises, une méthode de bootstrap plus avancée ou des méthodes analytiques, comme réalisées par la Division sondages, chaque année pour l'enquête SRCV, pourraient être mobilisées, notamment pour l'envoi des estimateurs de précision à Eurostat.

2.2. Bootstrap

2.2.1. Description du bootstrap

Le bootstrap est une méthode statistique basée sur la création de répliques d'un échantillon de base sur chacune desquelles une variable d'intérêt sera calculée [2,3]. Les estimations pour chaque réplique bootstrap sont ensuite étudiées pour estimer la dispersion de celles-ci par exemple. Il y a plusieurs manières de réaliser des estimations à l'aide de méthode de bootstrap, la méthode qui sera utilisée ici est une méthode dite « naïve » qui reste sur les fondamentaux du bootstrap sans y ajouter de raffinement supplémentaire.

En notant N la taille de l'échantillon d'origine et B un nombre de répliques bootstrap choisi de manière à obtenir la convergence des moyennes des estimations bootstrap, la méthode naïve peut-être résumée de la manière suivante :

1. Pour b allant de 1 à B :
 - a) Tirer $N-1$ individus par tirage équiprobable avec remise dans l'échantillon d'origine ;
 - b) Estimer la variable d'intérêt pour la réplique bootstrap b ;
2. Observer la convergence des moyennes et la répartition des B estimations bootstrap.

Dans le cas étudié ici, la méthode mobilisée, pour chacun des scénarios est la suivante avec $B = 1500$:

1. Pour b allant de 1 à B :
 - a) Tirer $N-1$ individus par tirage équiprobable avec remise dans l'échantillon d'origine ;
 - b) Calculer les poids des individus avec la formule suivante : $poids_CNR * N / (N-1)$ où $poids_CNR$ est le poids corrigé de la non-réponse qui est habituellement utilisé dans le cadre du calcul des poids transversaux (le poids corrigé de la non-réponse totale) ;
 - c) Estimer le taux de privation matérielle et sociale (PMS) pour la réplique bootstrap b ;
 - d) Calculer la moyenne des estimations bootstrap 1 à b .
2. Observer la convergence et la répartition des estimations bootstrap.

La méthode est ici volontairement simple pour assurer une vitesse d'exécution rapide. D'autres méthodes plus fines pourront éventuellement être utilisées à l'avenir.

2.2.2. Poids bootstrap

Avant de chercher à estimer la précision des estimations, on observe dans un premier temps la convergence, la répartition et la cohérence des estimations bootstrap.

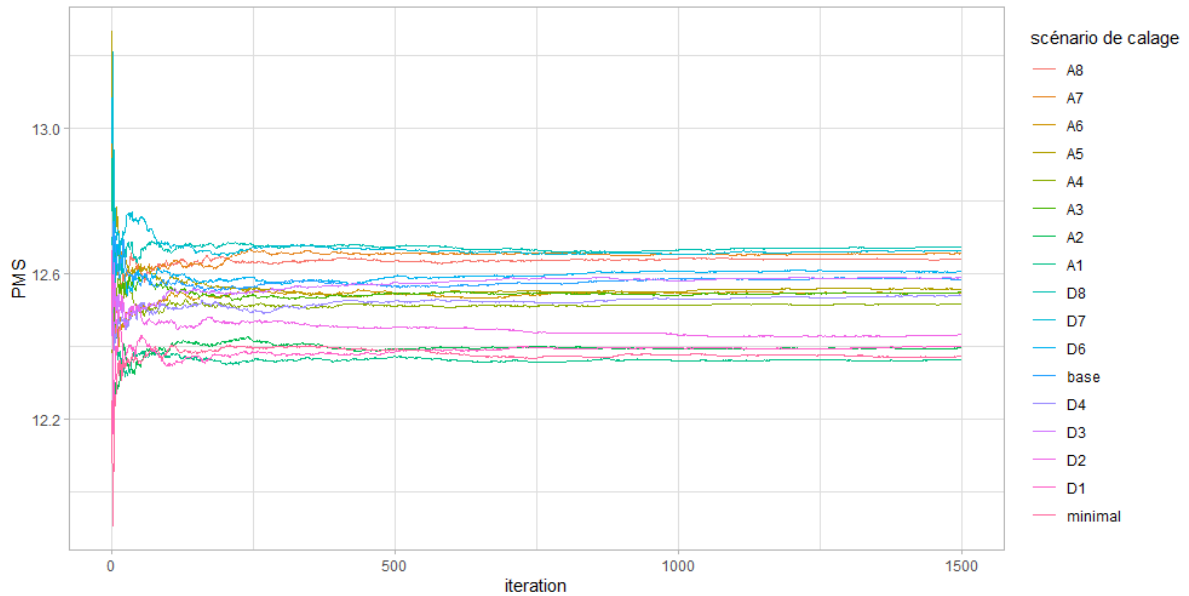


Figure 5 : Graphe de convergence des estimations bootstrap

Graphiquement, la moyenne des estimations bootstrap de la PMS semble bien converger vers une valeur, on peut cependant remarquer que la convergence des moyennes des estimations bootstrap ne semble pas tout à fait terminée. Il en est de même pour la convergence des quantiles.

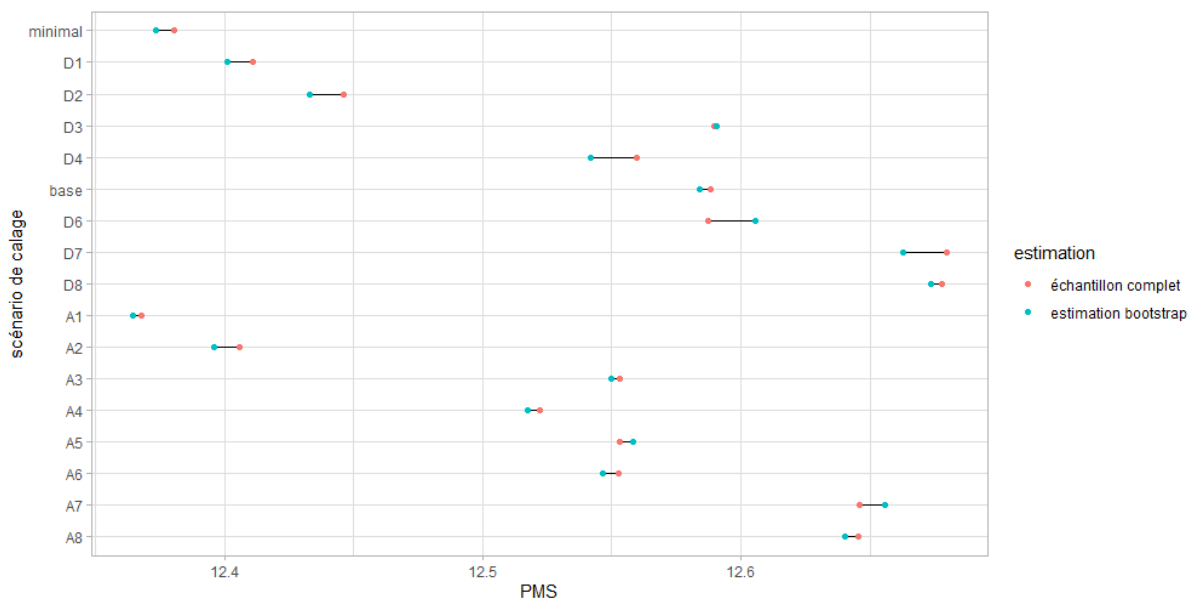


Figure 6 : Estimation avec échantillon complet et moyenne des estimations bootstrap

Malgré les doutes sur la convergence des estimations, les moyennes des estimations bootstrap restent proches des estimations sur l'échantillon. Il est également possible de regarder directement la répartition des estimations pour voir si quelque chose apparaît suspect dans leur répartition.

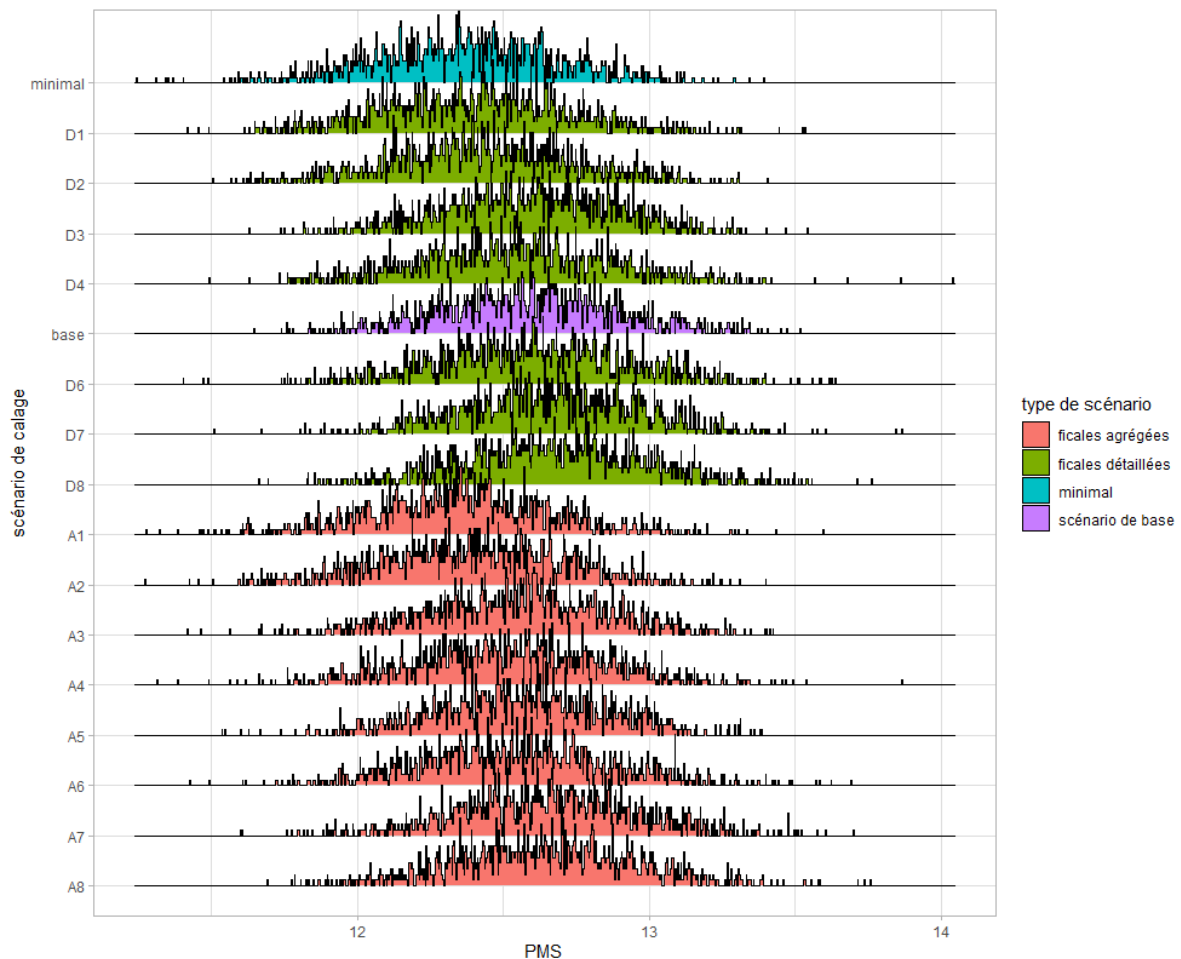


Figure 7 : Répartition des estimations bootstrap selon les scénarios de calage

Il ne semble pas y avoir d'estimations très éloignées des autres ou bien plusieurs modes (ou « pointes ») dans la répartition de celle-ci. En effet, s'il y avait plusieurs modes dans la répartition des estimations, cela pourrait entraîner des problèmes de convergence de la moyenne des estimations bootstrap.

Pour étudier la précision des indicateurs de PMS grâce au bootstrap, on analyse la distribution empirique des estimations de PMS pour chaque échantillon bootstrapé. Les intervalles de confiance présentés dans la Figure 8 sont calculés à partir des quantiles de la distribution des indicateurs de PMS présentés dans la Figure 7. La figure 8 classe les scénarios par taille d'intervalle de confiance croissante, avec un rappel de la valeur de la PMS obtenue sur l'échantillon complet pour chaque scénario.

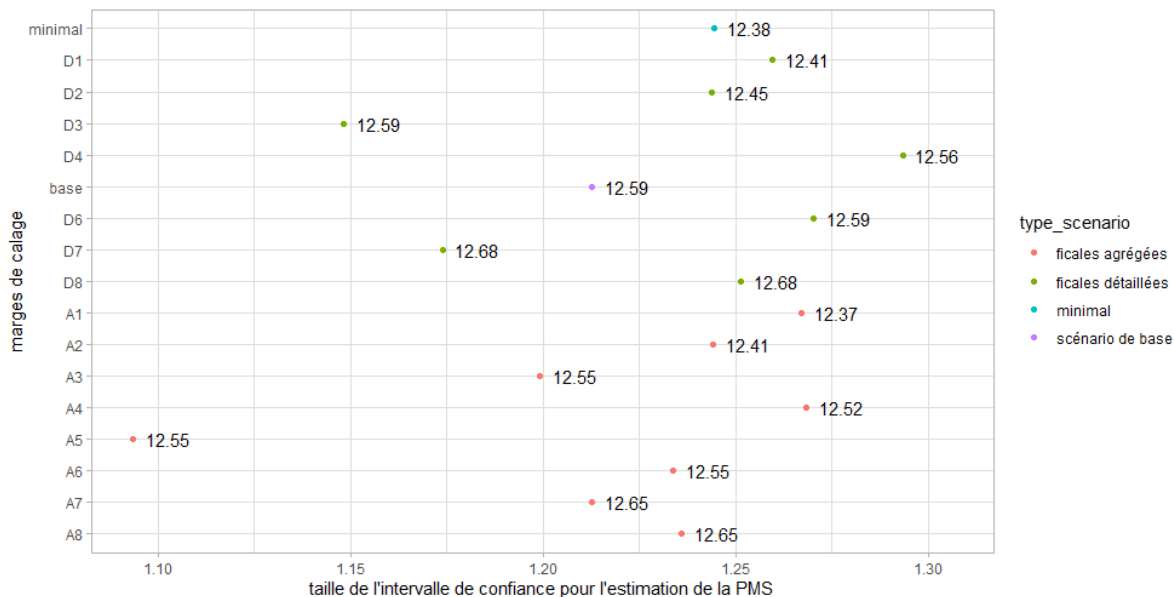


Figure 8 : Taille de l'intervalle de confiance à 95 % de l'estimation de la PMS selon le jeu de marges (estimation de la PMS sur l'échantillon complet précisée à côté des points)

Les tailles des intervalles de confiance bootstrap pour l'estimation de la PMS sont relativement variables selon le jeu de marges choisi. Par exemple, pour le calage réalisé actuellement dans l'enquête, la taille de l'intervalle de confiance estimé par bootstrap est de 1,21 point de pourcentage. D'autres estimations de la précision obtenues par étude analytique ne sont pas disponibles actuellement pour vérifier la vraisemblance de ces valeurs, mais il est possible de comparer les estimations bootstrap de la précision des estimations de la privation matérielle et sociale sévère (au moins sept privations de la vie quotidienne parmi la liste présentée dans la partie 1) avec celles qui sont fournies par la Division sondages.

Un élément rassurant est que les estimations de précision de l'estimation de la PMS sévère fournies par la Division sondages (calcul analytique de précision) et celles calculées par bootstrap pour le scénario de base de SRCV 2024 sont du même ordre de grandeur. Les estimations de précision obtenus avec le bootstrap naïf donnent toutefois un intervalle de confiance légèrement plus large que les estimations fournies par la Division sondages. Ces résultats sont cohérents avec une analyse comparative d'estimation de la variance (approche analytique ou approche par bootstrap) effectuée pour certaines variables issues de l'enquête Histoire de vie et patrimoine [4].

Cependant, les différents scénarios de calage ne sont pas classés dans le même ordre quand on refait un jeu de répliquions bootstrap. Cette instabilité pourrait indiquer que la précision des estimations d'un bootstrap naïf n'est pas suffisante pour départager correctement les jeux de marges de calage entre eux. Ainsi, des analyses plus poussées, par exemple mobilisant des calculs analytiques de variance ou des méthodes d'analyse par bootstrap plus raffinées, pourraient compléter l'analyse présentée ci-dessus.

Il est possible de visualiser les effets sur l'estimation de la précision (taille de l'intervalle de confiance) de l'ajout successif d'une marge :

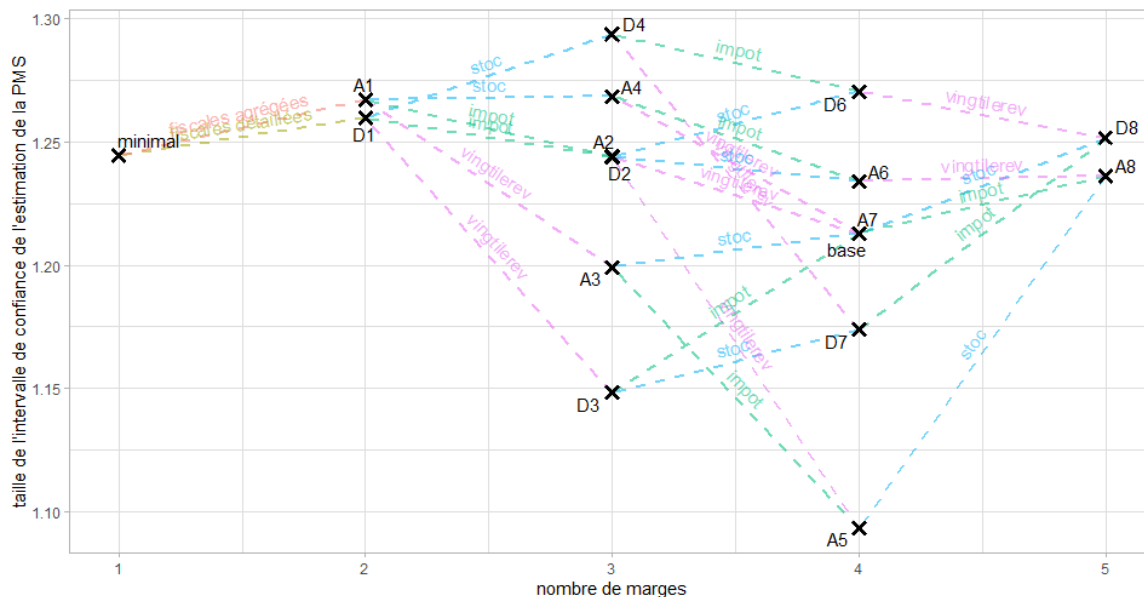


Figure 9 : Taille de l'intervalle de confiance de la PMS estimée par bootstrap, selon le scénario et le nombre de marges de calage

Les effets de l'ajout d'une marge de calage sur la taille de l'intervalle de confiance bootstrap sont beaucoup plus flous que pour les estimations directes de la PMS. Cela est potentiellement dû aux défauts du bootstrap évoqués précédemment. Cela dit, ce graphique met en relief des éléments qui seront à garder en tête lors de l'exploitation d'un bootstrap plus fiable ou même dans le cadre de choix de marges de calage en général. Par exemple, on observe sur le graphique ci-dessus que l'ajout d'une même marge de calage peut avoir des effets différents selon les autres marges utilisées. Cela était déjà observable avec le graphe des estimations directes mais est potentiellement d'autant plus remarquable quand on s'intéresse aux largeurs des intervalles de confiance. Ici, la lecture du graphique amènerait à choisir les scénarios avec une faible largeur de l'intervalle de confiance. Il est aussi possible de prendre en compte les valeurs des estimations pour mettre de côté les scénarios qui ont un impact trop fort sur l'estimation des variables d'intérêt par rapport aux estimations actuelles.

3. Choix des scénarios

3.1. Critères quantitatifs

Le choix d'utiliser ou non une marge pour le calage peut se fonder sur de nombreux critères, de nature quantitative ou conceptuelle. Il est ici question des critères quantitatifs révélés par le bootstrap. Les critères conceptuels seront ensuite pris en compte dans le choix final des marges à utiliser pour le calage.

Les critères quantitatifs étudiés sont :

- la valeur des estimations ;
- la précision des estimations ;
- la variance des marges ;
- la répartition des poids ;
- les rapports des poids calés sur les poids en entrée du calage.

La distribution des poids ne semble pas un critère déterminant dans notre cas, car elle diffère peu d'un scénario à l'autre. Le critère de variance des estimations obtenues est très important (des seuils de précision minimale réglementaires sont fixés par Eurostat), mais il n'est pas possible d'accorder une confiance suffisante au bootstrap naïf pour sélectionner un scénario de calage, car les estimations de précision réalisées avec le bootstrap sont relativement instables.

Les marges qui impactent le plus la précision de l'indicateur de PMS ne sont pas nécessairement celles qui impactent la valeur de son estimation. Les marges qui permettent d'avoir des estimations plus précises tout en impactant faiblement la valeur de l'estimation sont idéales dans notre cas, car elles limitent les risques de rupture de série.

S'il était supposé que les estimations de précision par bootstrap présentés dans la partie 2, sont suffisamment précises, les jeux de marges suivants se dégageraient selon ces critères quantitatifs :

- marges socio-démographiques, marges fiscales détaillées, vignettes → variance des estimations faible, estimations parmi les plus proches de ce qui est fait actuellement
- marges socio-démographiques, marges fiscales agrégées, vignettes, impôts → variances des estimations de la PMS la plus faible, estimation pas trop éloignée par rapport au calage actuel

Les critères quantitatifs aboutissent à effectuer un arbitrage entre éloignement de l'estimation d'origine et réduction de la variance des estimations. Par la suite, ce sont des critères plus qualitatifs qui sont discutés afin de départager différents scénarios de calage candidats.

3.2. Prise en compte de critères plus qualitatifs

Les critères qualitatifs et conceptuels étudiés sont :

- conformité aux recommandations Eurostat ;
- méthodologie de calage proche de celle utilisée pour l'ERFS, source de référence sur les revenus et la pauvreté monétaire ;
- disponibilité des marges ;
- champ des marges et cohérence avec les variables disponibles dans l'échantillon ;
- facilité avec laquelle l'algorithme de calage converge.

Par rapport à ces critères, quelques scénarios se dégagent :

- marges socio-démographiques, marges fiscales détaillées, impôts, vignettes → scénario utilisé actuellement, pas de changements à apporter et évite les ruptures de série.
- marges socio-démographiques, marges fiscales détaillées, vignettes → scénario proche du scénario de base, retrait de la marge impôts qui résulte d'une estimation et pas d'observation.
- marges socio-démographiques, marges fiscales agrégées, vignettes, impôts → scénario plus éloigné du scénario de base que les autres mais qui a le mérite de se rapprocher des marges de calage de l'enquête ERFS et qui devrait améliorer la comparabilité entre SRCV et ERFS, notamment pour certains points clefs comme les masses fiscales.
- marges socio-démographiques, marges fiscales agrégées, statut d'occupation du logement → scénario qui est le plus proche des recommandations européennes et des marges de calage de l'ERFS

Enfin, un critère supplémentaire essentiel ne peut pas être directement étudié avec l'analyse statique, réalisée sur un seul millésime de l'enquête, effectuée dans cette étude : la stabilité dans le temps. L'objectif est en effet d'avoir une méthode de calage qui permet de donner des intervalles de précision suffisamment stables dans le temps, et dans le respect de ce qui est requis par Eurostat. La persistance des résultats observés dans le temps permet de renforcer la validité de la méthode de calage utilisée. Par ailleurs, cela concerne également la disponibilité des marges utilisées sur plusieurs millésimes : par exemple, les vignettes de distribution utilisés pour le calage, initialement calculés à partir du dispositif Filosofi, ont à partir du millésime 2023 de l'enquête SRCV été calculés avec l'ERFS, car le dispositif Filosofi n'a pas été produit pour le millésime 2022, ce qui crée une rupture conceptuelle dans la cible utilisée pour le calage.

Conclusion

Le choix du jeu de marges de calage est le fruit de plusieurs arbitrages. Les méthodes présentées ici permettent de le faciliter, sans toutefois fournir d'outils clef en main pour sélectionner automatiquement le jeu de marges à utiliser. La précision des estimations peut être confirmée par l'utilisation de techniques de bootstrap. Le bootstrap permet de faire l'estimation de la précision des indicateurs pour un grand nombre de scénarios, et pour des indicateurs potentiellement complexes.

À défaut d'utiliser les techniques de bootstrap, il est aussi possible d'utiliser des techniques analytiques pour calculer la précision. La technique de bootstrap permet de tenir compte des spécificités des individus de l'échantillon, de calculer la précision d'indicateurs complexes et permet également de vérifier que le calage est répliquable sans encombre, car celui-ci est effectué à nouveau pour chaque réplique bootstrap. La méthode analytique est plus classique et permet de prendre en compte plus en détails le mécanisme d'échantillonnage.

Pour l'enquête SRCV, les tests réalisés dans le cadre de cette étude ne sont pas encore conclusifs. Il sera notamment nécessaire de poursuivre les travaux sur les calculs de précision selon les scénarios de calage. Si une modification des marges de calage était actée pour l'enquête SRCV, celle-ci pourrait être mise en place à partir du millésime 2026 de l'enquête, dont la diffusion est prévue début 2027. En effet, des modifications du questionnaire de l'enquête (liées notamment à l'intégration dans les outils de la filière d'enquête et à la prise en compte de la modularité des questions du tronc commun des enquêtes ménage) sont prévues pour la collecte de 2026, qui pourraient occasionner une rupture de série d'origines multiples qui ne sera pas quantifiable, dont Eurostat est prévenu. Ce millésime pourrait alors absorber des changements méthodologiques supplémentaires, comme une révision de la méthode de calage.

Bibliographie

[1] François Gleizes et Anne Pla, « En 2021, un enfant sur dix ne part pas en vacances pour des raisons financières », Insee Focus, n°294, mars 2023.

[2] Pascal Bessonneau, Gwennaëlle Brillhaut, Guillaume Chauvet et Cédric Garcia, « Estimation de la variance par le bootstrap avec remise pour les enquêtes auprès des ménages. Principes, exemples et mise en œuvre », Techniques d'enquête, Statistique Canada, n° 12-001, décembre 2021.

[3] Guillaume Chauvet, Gwennaëlle Brillhaut et Cédric Garcia, « Bootstrap avec remise pour l'estimation de variance des enquêtes auprès de ménages et exemples », 14^e Journées de méthodologie statistique de l'Insee (JMS), mars 2022.

[4] Khaled Larbi et Jean Rubin, « Estimation de la variance pour l'enquête longitudinale Histoire de vie et Patrimoine : une comparaison des approches analytique et par bootstrap », Documents de travail de l'Insee, n° 2025/07, octobre 2025.