

2015
–2020

00

Bases
statistiques
et généralités

Neuchâtel 2021

Enquête suisse sur la population active 2015–2020

Révision de la pondération
et de l'estimation de la variance

Rapports de méthodes de la section Méthodes statistiques

Les rapports de méthodes décrivent les méthodes mathématiques et statistiques à la base des résultats et des analyses de la statistique publique. Ils présentent également l'évaluation et le développement de nouvelles méthodes en vue d'une application future. Ces publications visent d'une part à documenter les méthodes utilisées ou envisagées dans un souci de transparence et de rigueur scientifique, et d'autre part à favoriser la collaboration avec le monde scientifique et universitaire.

Les résultats numériques présentés dans les rapports de méthodes illustrent les concepts mathématiques décrits, mais ne sont pas des résultats officiels des enquêtes concernées. De même, les méthodes réellement appliquées peuvent différer légèrement de celles décrites dans ces rapports.

Les rapports de méthodes sont disponibles sous forme électronique sur le site Internet de l'OFS.

Enquête suisse sur la population active 2015–2020

Révision de la pondération
et de l'estimation de la variance

Rédaction	Anne Massiani, OFS
Contenu	Section Méthodes statistiques, OFS
Éditeur	Office fédéral de la statistique (OFS)

Neuchâtel 2021

Éditeur: Office fédéral de la statistique (OFS)

Renseignements: Anne Massiani, tél. 058 463 62 65
anne.massiani@bfs.admin.ch

Rédaction: Anne Massiani, OFS

Contenu: Section méthodes statistiques, OFS

Série: Statistique de la Suisse

Domaine: 00 Bases statistiques et généralités

Langue du texte original: français

Mise en page: section DIAM, Prepress/Print
Ce document a été produit automatiquement à partir de banques de données. Il ne répond donc pas aux normes typographiques des publications de l'OFS.

En ligne: www.statistique.ch

Imprimés: www.statistique.ch
Office fédéral de la statistique, CH-2010 Neuchâtel,
order@bfs.admin.ch, tél. 058 463 60 60
Impression réalisée en Suisse

Copyright: OFS, Neuchâtel 2021
La reproduction est autorisée, sauf à des fins commerciales, si la source est mentionnée.

Numéro OFS: 338-0083

ISBN: 978-3-303-00688-7

Table des matières

Préambule et Résumé	5
1 Introduction	7
2 Plan de sondage	7
3 Principe général de la pondération trimestrielle	8
3.1 Population cible et gestion des hors-cible	8
3.2 Les trois principales étapes de la pondération trimestrielle	10
4 Modélisation du mécanisme de réponse	11
5 Mise en oeuvre pratique de l'estimation des probabilités de réponse initiales	13
5.1 Méthode d'estimation	13
5.2 Performances de la modélisation	14
6 Mise en oeuvre pratique de l'estimation des probabilités d'érosion	15
6.1 Variables auxiliaires disponibles	15
6.2 Méthode d'estimation	16
6.3 Performances de la modélisation	16
7 Facteur de combinaison des panels	17
8 Pondération annuelle	19
9 Estimation de la précision pour la pondération trimestrielle et annuelle	21
10 Evaluation de la précision de l'estimation d'une moyenne sur plusieurs trimestres	26
11 Conclusions et perspectives	29
Bibliographie	30
Annexes	31
A Schéma de rotation des panels	31
A.1 Schéma de rotation standard	31
A.2 Schéma de rotation pour les 75 ans et plus	32
B Variables auxiliaires disponibles	33
C Taux de réponse	38
D Variables auxiliaires entrant dans la segmentation utilisée pour la correction de la non-réponse en vague 1	39
E Variables auxiliaires utilisées pour le calage	40
F Arbre de segmentation de la vague 1 du troisième trimestre 2015	41
Rapports de méthodes de la section méthodes statistiques de l'OFS	43

Préambule

L'enquête suisse sur la population active (ESPA) est une enquête trimestrielle auprès des personnes âgées de 15 ans et plus qui fournit des informations sur la structure de la population active et sur les comportements en matière d'activité professionnelle. L'échantillonnage suit un schéma rotatif de quatre panels, qui présente l'avantage d'être performant pour produire à la fois des estimations transversales et des estimations d'évolutions fiables (pour des niveaux de ventilation adaptés à la taille de l'enquête), mais nécessite une pondération et des méthodes d'estimation appropriées décrites dans ce rapport.

L'auteure souhaite remercier la section METH pour son accompagnement durant ce travail. Ses plus sincères remerciements s'adressent en particulier à Daniel Kilchmann pour son soutien et ses précieux conseils, ainsi qu'à Rémy Mariétan pour sa disponibilité pour de nombreuses et fructueuses discussions. Egalement un immense merci à Jean-Pierre Renfer, Daniel Assoulin, Jann Potterat, Alain Vuille et Sylvie Regli pour leur patiente relecture. Leurs remarques pertinentes ont permis d'améliorer la qualité de ce rapport. Enfin, l'auteure tient à remercier très chaleureusement la section AES pour avoir su créer un climat de travail constructif et surtout très agréable, ainsi que pour tous les échanges enrichissants que nous avons eus.

Résumé

L'enquête suisse sur la population active (ESPA) est une enquête complexe de par son plan d'échantillonnage basé sur un panel rotatif qui nécessite un traitement méthodologique approprié, ainsi que l'exploitation d'informations auxiliaires utiles pour modéliser la non-réponse initiale et les différentes phases d'érosion, mais également pour le calage.

La possibilité d'accéder à des sources de données administratives de plus en plus variées offre de nouvelles perspectives pour la pondération des enquêtes. Dans le cadre du travail présenté dans ce rapport, les gains de précision potentiels permis par les informations sociodémographiques contenues dans le cadre de sondage, mais également les données AVS, ou encore les données provenant des organismes de placement sur le marché du travail et des caisses de chômage, ont été testés et exploités pour la pondération de l'ESPA. De plus, une procédure d'estimation de la variance a également été développée, implémentée et testée, afin d'accompagner d'une mesure de précision les différents types d'exploitations diffusées.

Les analyses effectuées ont montré que les autres sources d'informations auxiliaires que le cadre de sondage ont un impact limité sur les performances de la modélisation du mécanisme de réponse. En revanche, l'impact de ces autres sources est beaucoup plus important si l'on s'intéresse au gain que l'on peut obtenir grâce au calage, qui est par ailleurs, toutes sources d'informations auxiliaires confondues, très appréciable pour certaines estimations. Au niveau national, il permet par exemple une réduction du coefficient de variation d'environ 50% pour l'estimation du nombre d'actifs et de 15% pour l'estimation du nombre de chômeurs.

1 Introduction

L'enquête suisse sur la population active (ESPA) est une enquête trimestrielle auprès des personnes qui fournit des informations sur la structure de la population active et sur les comportements en matière d'activité professionnelle. La population cible est constituée des résidents permanents en ménage privé âgés de 15 ans et plus. L'échantillonnage suit un schéma rotatif de quatre panels selon lequel, chaque trimestre, le panel le plus ancien est renouvelé.

Différents types d'exploitations sont réalisés :

- Exploitations trimestrielles standards
Il s'agit d'estimations, par exemple du nombre ou du taux d'actifs et de chômeurs, basées sur les données récoltées sur les quatre panels interrogés lors du trimestre considéré.
- Exploitations basées sur la moyenne de plusieurs trimestres
La moyenne des estimations des quatre trimestres d'une même année fait également partie des analyses produites, par exemple le taux de chômage moyen sur les quatre trimestres d'une même année. Des moyennes calculées sur un plus grand nombre de trimestres peuvent également être diffusées.
- Exploitations annuelles
Certaines caractéristiques plus stables dans le temps que le statut sur le marché de l'emploi, comme par exemple le niveau de formation, sont demandées seulement à deux des quatre vagues auxquelles un individu participe, et sont exploitées sur une base annuelle.
- Autres types d'exploitations
Certains thèmes plus spécifiques sont enquêtés lors de modules pour lesquels seulement une partie de l'échantillon est conservé. D'autres exploitations moins centrales et impliquant les autres membres du ménage sont basées sur une « pondération ménage ».

La possibilité d'accéder à des sources de données administratives de plus en plus variées, comme par exemple les données AVS, ou encore les données provenant des organismes de placement sur marché du travail et des caisses de chômage, offre de nouvelles perspectives pour la pondération des enquêtes. Les sections « Méthodes statistiques » (METH) et « Travail et vie active » (AES) ont cherché à exploiter ces nouvelles possibilités dans le cadre de l'ESPA. L'objectif de ce rapport est de décrire les méthodes appliquées. A noter que les exploitations plus spécifiques telles que celles basées sur les modules ou sur la pondération ménage ne sont pas couvertes par ce rapport.

2 Plan de sondage

L'échantillonnage suit un schéma rotatif de quatre panels, dont les tirages sont coordonnés négativement dans le SRPH (cadre de sondage pour les enquêtes auprès des personnes et des ménages, Stichprobenregister für Personen- und Haushaltserhebungen en allemand) au niveau du ménage. Chaque trimestre, le panel le plus ancien est renouvelé, et les panels sont réinterrogés selon le schéma 3-9-3 suivant, qui peut être visualisé grâce à la figure présentée en annexe [A.1](#):

- une personne participant à l'enquête une première fois sera à nouveau sollicitée trois mois plus tard pour une deuxième participation.
- Si elle accepte cette deuxième participation, elle sera à nouveau sollicitée neuf mois plus tard pour une troisième participation.
- Si elle accepte cette troisième participation, elle sera à nouveau sollicitée trois mois plus tard pour une quatrième et dernière participation.

Un échantillon trimestriel est donc composé de quatre panels tirés dans quatre cadres différents, et totalise environ 30'000 personnes. Les répondants du panel le plus ancien ont participé quatre fois à l'enquête. Ce schéma rotatif a été choisi car il présente l'avantage d'être performant pour produire à la fois des estimations transversales et des estimations d'évolutions fiables (pour des niveaux de ventilation adaptés à la taille de l'enquête). L'échantillon annuel est quant à lui composé des vagues 1 et 3 des quatre trimestres de l'année considérée. De part le schéma de rotation, cette composition de l'échantillon annuel permet de faire en sorte que les huit panels qui le composent soient issus de huit tirages différents. L'échantillon annuel totalise environ 60'000 personnes.

Par ailleurs, les 75 ans et plus (âge lors de l'enquête) font l'objet d'un traitement spécifique introduit progressivement par AES depuis le premier trimestre 2018. Il a pour but de diminuer la charge liée à l'ESPA pour les 75 ans et plus, qui sont généralement peu concernés par les questions de l'enquête liées à la vie active. Le traitement actuellement appliqué est le suivant :

- Les 90 ans et plus ne seront pas interrogés. Cependant, une exploitation se limitant à la population des moins de 90 ans n'est pas envisageable, notamment car le taux d'actifs doit être calculé sur l'ensemble de la population. La solution pratique retenue est la suivante. Les 90 ans et plus sont sélectionnés selon le plan de sondage prévu, puis :
 - ils sont conservés dans l'échantillon net de la vague 1 ;
 - en revanche, ils ne sont pas interrogés et les variables qui devraient être relevées sont soit imputées (par exemple les 90 ans et plus sont tous considérés comme inactifs), soit exploitées avec une modalité à part pour les valeurs manquantes.
- A partir du moment où une personne atteint l'âge de 75 ans, elle est interrogée durant la vague courante, mais n'est plus sollicitée dans les vagues suivantes. Par exemple si une personne atteint l'âge de 75 ans durant la vague 2 (mais avait encore 74 ans en vague 1), elle est interrogée en vague 2, mais ne le sera plus lors des vagues suivantes. De même, si une personne atteint l'âge de 75 ans en vague 3, elle est interrogée en vague 3, mais ne le sera plus lors de la dernière vague. A noter qu'en rythme de croisière, chaque vague contient des personnes de 75 ans.

De façon résumée, la vague 1 couvre donc la population âgée de 15 ans et plus, tandis que les vagues 2, 3, et 4 couvrent la population âgée de 15 à 75 ans. Le traitement décrit ci-dessus a été introduit depuis le premier trimestre 2019. L'introduction est progressive, dans le sens où elle s'applique aux panels tirés pour une première interrogation lors du premier trimestre 2019 ou des trimestres postérieurs, les panels issus de tirages précédents faisant l'objet d'un autre traitement. Durant l'année 2018, la règle était en effet que les personnes de 75 ans et plus ne pouvaient être interrogées qu'en vague 1. Cette règle et celle appliquée actuellement diffèrent par exemple pour les personnes atteignant l'âge de 75 ans en vague 2. Selon la règle actuelle, ces personnes sont interrogées en vagues 1 et 2, tandis qu'elles ne le seraient qu'en vague 1 avec la règle de 2018. Cette règle avait elle aussi été introduite progressivement, entre le premier et le dernier trimestre 2018. Avant le premier trimestre 2018, les 75 ans et plus ne faisaient l'objet d'aucun traitement spécifique et étaient interrogés comme le reste de la population. L'introduction progressive de ces différents changements peut être visualisée en annexe [A.2](#). Nous décrivons dans ce rapport la pondération des 75 ans et plus développée pour le traitement actuellement appliqué.

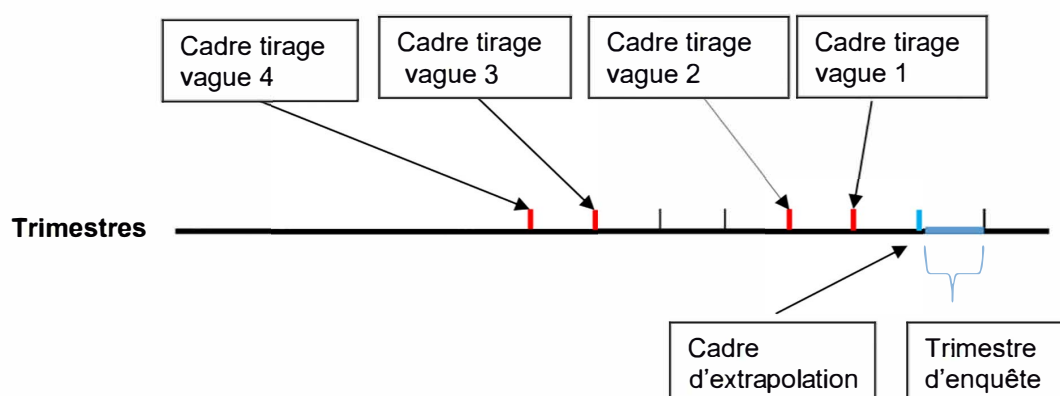
3 Principe général de la pondération trimestrielle

3.1 Population cible et gestion des hors-cible

Pour la pondération trimestrielle, la population cible est constituée des résidents permanents en ménage privé âgés de 15 ans et plus au début du trimestre d'enquête. Pour couvrir la population

cible, un « cadre d'extrapolation » utilisé pour la pondération est construit sur la base du cadre de tirage SRPH, que l'on enrichit en l'appariant à différentes sources de données administratives depuis le troisième trimestre 2015. Le cadre de tirage SRPH utilisé pour construire le cadre d'extrapolation est le cadre de tirage dont la date de référence coïncide avec le début du trimestre d'enquête. Par exemple, pour le troisième trimestre 2015, le cadre d'extrapolation est construit sur la base du cadre de tirage SRPH du 30.06.2015. Même pour la vague 1, le cadre utilisé pour construire le cadre d'extrapolation ne coïncide donc pas avec les cadres de tirage des quatre vagues. Le cadre de tirage de la vague 1 est en effet celui qui est disponible trois mois avant le début de la période d'enquête, tandis la date de référence du cadre d'extrapolation coïncide avec le début de la période d'enquête, afin d'être le plus proche possible de la réalité de l'enquête. Une représentation graphique est proposée dans la figure 1.

G 1 – Ligne du temps du cadre d'extrapolation et des cadres de tirage des différentes vagues d'un trimestre donné



Notons enfin que même si idéalement, le cadre d'extrapolation et le cadre de tirage utilisé pour le construire devraient contenir exactement les mêmes personnes, des exceptions sont possibles en pratique. Les personnes détenant des permis Ci¹ sont en effet exclues du cadre de tirage, car elles sont protégées, mais elles font partie de la population cible et du cadre d'extrapolation. Le cadre d'extrapolation est donc utilisé pour corriger les défauts de couverture des cadres de tirage, en l'occurrence dans ce cas la sous-couverture due à la protection des personnes détenant des permis Ci. Cela revient à les assimiler à des non-répondants.

Les sources de données administratives utilisées pour enrichir le jeu de variables auxiliaires disponibles dans le cadre de tirage sont les données AVS et AI, les données concernant les prestations complémentaires aux rentes AVS-AI, le registre des comptes individuels (CI), le système d'information en matière de placement et de statistique du marché du travail (PLASTA), les systèmes informatisés de paiement des caisses de chômage (SIPAC), les données de l'aide sociale. L'ensemble des variables auxiliaires disponible est détaillé dans l'annexe B.

Les personnes qui ne sont pas retrouvées dans le cadre d'extrapolation sont considérées comme hors-cible. Cette règle est appliquée lors des appariements entre l'échantillon brut du trimestre considéré et le cadre d'extrapolation correspondant. Les personnes qui ne sont pas retrouvées dans le cadre d'extrapolation sont donc supprimées de l'échantillon brut. A noter que la détec-

1. Il s'agit des membres de la famille de toute personne employée de mission permanente ou d'une organisation internationale (ambassadeurs, employés d'ONG, etc.). Sur l'exemple du troisième trimestre 2015, cela concerne 940 personnes. Il n'est pas inhabituel dans les enquêtes de l'OFS que ces personnes soient protégées et ne puissent pas être interrogées, bien qu'elles fassent partie de la population cible.

tion et la suppression des hors-cible est identique pour les répondants et les non-répondants, et concerne chaque trimestre une dizaine de personnes. Les principales étapes de la pondération trimestrielle sont alors les suivantes.

3.2 Les trois principales étapes de la pondération trimestrielle

1. Correction de la non-réponse

Pour un trimestre t donné, l'échantillon net est constitué de quatre panels répondants, en première, deuxième, troisième ou quatrième interrogation. On calcule tout d'abord des poids corrigés pour la non-réponse séparément pour chacun des quatre panels. Le poids corrigé pour la non réponse de l'individu i du panel répondant en $k^{\text{ième}}$ vague est noté $w_{i,k}^{t,NR}$. Pour chaque panel, la somme de ces poids fournit une estimation de la taille de la population qu'il couvre (15 ans et plus pour la vague 1, 15 à 75 ans pour les autres vagues). La façon dont sont obtenus ces poids est décrite dans les sections 4, 5 et 6.

2. Combinaison des panels

Les quatre panels sont ensuite réunis et les poids de l'étape 1 sont ajustés afin que leur somme sur la réunion des quatre panels fournisse une estimation de la taille de la population cible (15 ans et plus). Plus précisément, un facteur de combinaison $f_{i,k}^t$ est appliqué aux individus de chaque panel, de sorte que le poids combiné pour la réunion des panels pour un individu i répondant en vague k est donné par :

$$w_{i,k}^{t,Comb} = w_{i,k}^{t,NR} \cdot f_{i,k}^t.$$

Les facteurs $f_{i,k}^t$ dépendent de l'âge de la personne au moment de l'enquête. Pour les moins de 75 ans, le facteur $f_{i,k}^t$ appliqué au panel répondant en vague k est défini par :

$$f_{i,k}^t = f_{k,75-}^t = \frac{n_{k,75-}^t}{\sum_{k=1}^4 n_{k,75-}^t},$$

où $n_{k,75-}^t$ désigne le nombre de répondants de moins de 75 ans au sein du panel répondant en vague k lors du trimestre t . Pour les plus de 75 ans, le facteur de combinaison tient compte du fait que l'on ne peut être interrogé qu'une seule fois passé cet âge. Cette étape de la pondération est décrite plus en détails dans la section 7.

3. Calage

Les poids de l'étape 2 sont ensuite calés sur des totaux connus pour la population cible. Les poids de départ sont les poids combinés désignés par $w_i^{t,Comb}$ pour l'échantillon de répondants du trimestre t , toutes vagues confondues, et les poids calés finaux sont notés w_i^t . Les motivations pour réaliser ce calage sont les suivantes.

- Cohérence avec d'autres statistiques.
- Correction des défauts de couverture dus au fait que les personnes faisant partie depuis peu de temps de la population cible (par exemple immigration récente, personnes venant d'atteindre l'âge de 15 ans, etc.) ne peuvent être sélectionnées que dans le panel répondant en première vague. A noter que même la vague 1 comporte ce type de défauts de couverture, certes façon moins importante, puisque les personnes entrant dans la population cible entre le trimestre de tirage de la vague 1 et le début du trimestre d'enquête ne peuvent pas être sélectionnées.
- Gain de précision.

Le calage a été développé et implémenté par la section AES grâce à la macro CALMAR de SAS, avec la fonction de distance correspondant au raking ratio. La liste des variables de calage se trouve dans l'annexe E.

4 Modélisation du mécanisme de réponse

Comme mentionné dans la section 2, l'échantillonnage suit un schéma rotatif de quatre panels selon lequel, chaque trimestre, le panel le plus ancien est renouvelé, et les panels sont réinterrogés selon le schéma 3-9-3. On désigne par s_b^t l'échantillon brut de personnes sélectionnées et activées pour une première participation lors du trimestre t , et par π_i^t la probabilité de sélection de l'individu i . L'échantillon de personnes qui acceptent de répondre en $k^{\text{ième}}$ interrogation (ou vague) lors du trimestre t est noté $s_r^{t,k}$, de sorte que l'échantillon s_r^t de répondants lors du trimestre t est égal à l'union des $s_r^{t,k}$, pour $k = 1, \dots, 4$. Dans la notation des échantillons bruts s_b^t , l'exposant t se réfère donc au trimestre pour lequel l'échantillon est sélectionné pour entrer dans le panel, tandis qu'il se réfère au trimestre au cours duquel l'interview a lieu pour les échantillons de répondants $s_r^{t,k}$. Les quatre échantillons composant l'échantillon à pondérer s_r^t sont donc issus d'échantillons bruts entrant lors du trimestre $t - d$, où $d = 0, 1, 4$, ou 5 (trimestres), selon qu'ils répondent en vague $k = 1, 2, 3$ ou 4 . On désigne par $p_{i,k}^t$ la probabilité de réponse de l'individu i en $k^{\text{ième}}$ interrogation lors du trimestre t , sachant qu'il a été sélectionné pour une première participation lors du trimestre $t - d$, où $d = 0, 1, 4$, ou 5 (trimestres), selon que $k = 1, 2, 3$ ou 4 :

$$p_{i,k}^t = P \left(i \in s_r^{t,k} \mid i \in s_b^{t-d} \right). \quad (1)$$

Le poids corrigé pour la non-réponse de l'individu i de s_r^t répondant en vague k (cf. étape 1 de la pondération, section 3) est alors défini par :

$$w_{i,k}^{t,NR} = \frac{1}{\pi_i^{t-d}} \frac{1}{\hat{p}_{i,k}^t}, \quad (2)$$

où $\hat{p}_{i,k}^t$ est une estimation de la probabilité de réponse $p_{i,k}^t$.

Le mécanisme de non-réponse peut être décomposé en une non-réponse initiale lors de la première interrogation, suivie d'une érosion au fil des vagues successives. Par définition, la probabilité de réponse initiale pour les individus répondants en vague 1 lors du trimestre t correspond à $p_{i,1}^t$. La probabilité d'érosion entre la $k^{\text{ième}}$ interrogation lors du trimestre t et l'interrogation précédente est définie par $(1 - p_{i,k-1,k}^t)$, avec :

$$p_{i,k-1,k}^t = P \left(i \in s_r^{t,k} \mid i \in s_r^{t-d,k-1} \right), \quad (3)$$

où $(k, d) \in \{(2, 1), (3, 3), (4, 1)\}$, car l'interrogation précédant celle du trimestre t a lieu un trimestre auparavant pour les personnes répondant en vague $k = 2$, trois trimestres auparavant pour les personnes répondant en vague $k = 3$, un trimestre auparavant pour les personnes répondant en vague $k = 4$. Pour les panels répondants en deuxième, troisième ou quatrième interrogation ($k > 1$), la probabilité de réponse $p_{i,k}^t$ peut alors s'exprimer en fonction de la probabilité de réponse initiale et des probabilités d'érosion successives de la façon suivante :

$$p_{i,k}^t = \begin{cases} p_{i,1}^{t-1} \cdot p_{i,1,2}^t & \text{si } k = 2, \\ p_{i,1}^{t-4} \cdot p_{i,1,2}^{t-1} \cdot p_{i,2,3}^t & \text{si } k = 3, \\ p_{i,1}^{t-5} \cdot p_{i,1,2}^{t-4} \cdot p_{i,2,3}^{t-1} \cdot p_{i,3,4}^t & \text{si } k = 4. \end{cases} \quad (4)$$

Il découle de (2) et de (4) que le poids corrigé pour la non-réponse de l'individu i de s_r^t répondant

en vague k s'obtient de la façon suivante :

$$w_{i,k}^{t,NR} = \begin{cases} \frac{1}{\pi_i^t} \cdot \frac{1}{\hat{p}_{i,1}^t} & \text{si } k = 1, \\ \frac{1}{\pi_i^{t-1}} \cdot \frac{1}{\hat{p}_{i,1}^{t-1}} \cdot \frac{1}{\hat{p}_{i,1,2}^t} & \text{si } k = 2, \\ \frac{1}{\pi_i^{t-4}} \cdot \frac{1}{\hat{p}_{i,1}^{t-4}} \cdot \frac{1}{\hat{p}_{i,1,2}^{t-1}} \cdot \frac{1}{\hat{p}_{i,2,3}^t} & \text{si } k = 3, \\ \frac{1}{\pi_i^{t-5}} \cdot \frac{1}{\hat{p}_{i,1}^{t-5}} \cdot \frac{1}{\hat{p}_{i,1,2}^{t-4}} \cdot \frac{1}{\hat{p}_{i,2,3}^{t-1}} \cdot \frac{1}{\hat{p}_{i,3,4}^t} & \text{si } k = 4. \end{cases} \quad (5)$$

où $\hat{p}_{i,1}^{t-d}$ désigne une estimation des probabilités de réponse initiale, et $(1 - \hat{p}_{i,k-1,k}^{t-d})$ désigne une estimation des probabilités d'érosion. Les probabilités de sélection π_i^{t-1} intervenant dans la formule précédente sont connues et déterminées lors du tirage des échantillons².

Les probabilités de réponse initiale et d'érosion doivent quant à elles être estimées. Les mécanismes de non-réponse initiale et d'érosion sont modélisés par un plan de Poisson sur les individus, de sorte que le mécanisme de réponse « global » (combinaison de la non-réponse initiale et de l'érosion) suit également un plan de Poisson sur les individus. On suppose par ailleurs que les probabilités de non-réponse initiale et d'érosion sont uniformes au sein de groupes de réponse homogènes. Plus précisément, pour toute vague $k = 1, \dots, 4$ d'un trimestre t , on suppose que l'échantillon interrogé ou réinterrogé peut être partitionné en C_k^t groupes de réponse homogènes $g_{c,k}^t$ au sein desquels les probabilités de non-réponse initiale ou d'érosion sont toutes égales à une même valeur commune $(1 - \hat{p}_{c,k}^t)$:

$$\begin{cases} p_{i,1}^t = \hat{p}_{c,1}^t & \text{pour tout } i \in g_{c,1}^t, & \text{si } k = 1 \\ p_{i,k-1,k}^t = \hat{p}_{c,k}^t & \text{pour tout } i \in g_{c,k}^t, & \text{si } k = 2, 3, 4. \end{cases}$$

La probabilité de réponse $\hat{p}_{c,k}^t$ au sein d'un groupe de réponse homogène $g_{c,k}^t$ est alors estimée par la proportion de répondants au sein du groupe. Les méthodes utilisées pour la création des groupes de réponse homogènes sont décrites dans les sections 5 et 6.

D'un point de vue pratique, on observe sur la base de l'équation (5) la relation de récurrence suivante, pour $k \geq 2$:

$$w_{i,k}^{t,NR} = w_{i,k-1}^{t-d,NR} \frac{1}{\hat{p}_{i,k-1,k}^t}, \quad (6)$$

où $d = 1, 4$ ou 5 selon que $k = 2, 3$ ou 4 . La relation (6) provient de la décomposition du mécanisme de réponse en une non-réponse initiale suivie d'érosions successives, et montre que le poids corrigé pour la non-réponse du panel répondant en vague k lors du trimestre t , avec $k \geq 2$, s'obtient en multipliant le poids corrigé pour la non-réponse de ce même panel en vague $k - 1$ par un facteur corrigeant l'érosion entre la vague $k - 1$ et la vague k . Par conséquent, d'un point de vue pratique, pour la pondération d'un trimestre t , quatre probabilités doivent être estimées :

- la probabilité de réponse initiale $p_{i,1}^t$ de la vague 1. Cette estimation est alors multipliée par la probabilité de sélection.
- Les probabilités d'érosion $(1 - p_{i,k-1,k}^t)$ par rapport à la vague précédente pour les trois autres panels. Pour chaque panel, $\hat{p}_{i,k-1,k}^t$ est alors multipliée par le poids corrigé pour la non-réponse de ce même panel lors de l'interrogation précédente, qui a déjà été calculé lors de la pondération du trimestre correspondant.

2. En pratique les probabilités de sélection utilisées pour le tirage sont des probabilités de sélection, réserve comprise. Or seule une partie de la réserve est activée. Les probabilités de sélection utilisées pour la pondération sont donc obtenues en corrigeant les probabilités de sélection utilisées pour le tirage par un taux d'activation uniforme.

Une méthode alternative consisterait à estimer directement les probabilités de réponse $p_{i,k}^t$, sans décomposer le mécanisme de non-réponse en une non-réponse initiale et des érosions successives. De notre point de vue, cette décomposition du mécanisme de non-réponse présente les avantages suivants.

- Cela revient à décrire ce qui se passe effectivement au cours du temps. On dispose ainsi d'une estimation des probabilités de réponse pour chaque phase de l'enquête, ce qui constitue des informations intéressantes en lien avec la façon dont l'enquête est perçue par la population, la qualité de la gestion de l'institut en charge de l'enquête, etc.
- Des variables d'enquête relevées lors de la vague $k - 1$, telles que le statut sur le marché de l'emploi ou le niveau de formation, peuvent être utilisées comme variables auxiliaires lors de l'estimation des probabilités d'érosion entre la vague $k - 1$ et la vague k (cf. section 6 pour plus de détails). De cette façon, on intègre plus de variables auxiliaires dans l'estimation des probabilités de réponse $p_{i,k}^t$ que si elles avaient été estimées directement sans décomposition du mécanisme de non-réponse.

5 Mise en oeuvre pratique de l'estimation des probabilités de réponse initiales

Nous nous intéressons dans cette section à l'estimation de la probabilité de réponse initiale $\hat{p}_{i,1}^t$ des individus de l'échantillon brut s_b^t sélectionnés pour une première participation lors du trimestre t . Pour le troisième trimestre 2015, le taux de réponse en vague 1 était de 44.3% au niveau national. Une analyse plus détaillée des taux de réponse en fonction de la strate de tirage (définie sur la base du canton) est disponible dans l'annexe C. Les informations auxiliaires utilisées pour estimer les probabilités de réponse sont celles contenues dans le cadre d'extrapolation basé sur le cadre de tirage SRPH dont la date de référence coïncide avec le début du trimestre t (cf. section 3). A cet effet, on réalise un appariement entre l'échantillon brut s_b^t et le cadre d'extrapolation.

5.1 Méthode d'estimation

- Des groupes de réponse homogènes sont constitués sur la base d'un algorithme de segmentation appliqué à s_b^t .
- Les poids de sondage (inverse des probabilités de sélection) ne sont pas utilisés lors de l'application de cet algorithme³, pour lequel chaque unité reçoit un poids égal à 1.
- La variable d'intérêt de l'algorithme de segmentation est la variable binaire indicatrice de la participation ou non à l'enquête lors du trimestre t . Les variables auxiliaires entrant dans la segmentation sont issues du cadre d'extrapolation et sont listées dans l'annexe D. De plus, une variable indiquant si la personne dispose ou non d'un raccordement téléphonique fixe (oui=tel / non=altel) est incluse dans le jeu de variables auxiliaires candidates pour la segmentation, puisque le mécanisme de réponse est susceptible d'être assez différent pour ces deux catégories de population. En pratique, cette variable est toujours parmi les plus significatives pour la constitution des groupes de réponse homogènes, depuis le deuxième trimestre 2014 où le panel entrant a été tiré pour la première fois dans le cadre de sondage SRPH, jusqu'au premier trimestre 2021.
- La taille minimale des groupes de réponse homogènes est fixée à 100, afin de garantir une certaine stabilité au fil du temps, et le taux de réponse minimal à 10%.

3. Les poids de sondage sont en revanche bien sûr utilisés pour la pondération, cf. équation (2).

- Au sein de chaque groupe de réponse homogène $g_{c,1}^t$, la probabilité de réponse initiale $\tilde{p}_{c,1}^t$ est estimée par $R_{c,1}^t/N_{c,1}^t$, où $R_{c,1}^t$ désigne le nombre de répondants au sein du groupe de réponse homogène $g_{c,1}^t$, et $N_{c,1}^t$ sa taille.

L'arbre de segmentation obtenu pour la vague 1 du troisième trimestre 2015 est représenté dans l'annexe F.

5.2 Performances de la modélisation

En appliquant une régression logistique avec comme variable d'intérêt la variable indicatrice de la participation ou non à l'enquête, et comme variable auxiliaire l'appartenance aux différents groupes de réponse homogènes, on obtient des estimations des probabilités de réponse initiales égales à celles calculées comme décrit précédemment. L'utilisation de la procédure « logistic » de SAS permet donc d'obtenir différentes mesures calculées automatiquement par SAS qui permettent d'évaluer le pouvoir prédictif de la modélisation, telles que le taux de couples concordants, discordants, et « tied », ainsi que la C-statistic. Une définition de ces différentes mesures se trouve dans le guide d'utilisation de SAS (cf. [SAS Institute Inc., 2018](#), page 5845). Les résultats obtenus sur la base de la vague 1 du troisième trimestre 2015 sont présentés dans le tableau 1. Afin d'évaluer l'apport pour la modélisation des données administratives autres que celles contenues dans le SRPH, nous présentons également les résultats obtenus en utilisant uniquement les variables auxiliaires sociodémographiques du SRPH (les estimations retenues sont celles basées sur la totalité des variables auxiliaires disponibles). Quelques statistiques descriptives sur la distribution des probabilités de réponse estimées pour les répondants sont également présentées dans le tableau 2. Sur la bases de ces résultats, on peut faire les constatations suivantes.

- Si l'on suit la règle pratique⁴ proposée par [Hosmer & Lemeshow \(2000\)](#), le pouvoir prédictif de la modélisation du mécanisme de réponse initiale peut être considéré comme acceptable, puisque la C-statistic se situe dans la classe 0.7-0.8.
- L'exploitation des données administratives autres que celles contenues dans le SRPH semble permettre une amélioration limitée de la qualité du modèle.
- L'utilisation de l'ensemble des variables auxiliaires provenant des différentes sources de données administratives conduit à des poids légèrement plus dispersés que si l'on se base uniquement sur les variables sociodémographiques contenues dans le SRPH. Ceci s'explique par le fait que l'utilisation d'un plus grand nombre de variables auxiliaires conduit à un arbre de segmentation plus détaillé.

Enfin, les résultats présentés dans le tableau 3 permettent d'évaluer l'effet de la correction de la non-réponse en vague 1 sur la distribution des poids, pour le modèle de réponse retenu, à savoir celui basé sur l'exploitation de la totalité des variables auxiliaires.

4. La règle proposée par [Hosmer & Lemeshow \(2000\)](#) page 162 est la suivante.

- C-statistic égale à 0.5 : pas de discrimination,
- C-statistic entre 0.7 et 0.8 : discrimination acceptable,
- C-statistic entre 0.8 et 0.9 : discrimination excellente,
- C-statistic supérieure à 0.9 : discrimination exceptionnelle.

A noter que [Hosmer & Lemeshow \(2000\)](#) proposent cette règle pour analyser le pouvoir prédictif du modèle sur la base du critère de la surface sous la courbe ROC. Les explications page 163 montrent que la surface sous la courbe ROC est égale à la C-statistic.

T1 – Pouvoir prédictif de la modélisation du mécanisme de réponse initiale, vague 1 du troisième trimestre 2015

	Toutes les variables auxiliaires	Seulement les variables auxiliaires du SRPH
Pourcentage concordants	72.56	71.74
Pourcentage discordants	25.18	25.04
Pourcentage tied	2.26	3.22
C-statistic	0.74	0.73

T 2 – Distribution des probabilités de réponse initiales estimées, vague 1 du troisième trimestre 2015

	Min	P10	P25	P50	P75	P90	Max
Toutes les variables auxiliaires	0.10	0.28	0.49	0.59	0.66	0.72	0.85
Seulement les variables auxiliaires du SRPH	0.11	0.23	0.44	0.58	0.69	0.70	0.77

T 3 – Effet de la correction de la non-réponse en vague 1 sur la distribution des poids, vague 1 du troisième trimestre 2015

	Min	P10	P25	P50	P75	P90	Max	Max/Min
Poids de sondage	162	221	240	487	517	523	551	3.4
Poids corrigés pour la non-réponse en vague 1	229	380	505	746	886	1291	4728	21.7

6 Mise en oeuvre pratique de l'estimation des probabilités d'érosion

Il s'agit ici d'estimer la probabilité d'érosion $(1 - p_{i,k-1,k}^t)$ entre la $k^{\text{ième}}$ interrogation lors du trimestre t et l'interrogation précédente. L'échantillon net est ici l'échantillon $s_r^{t,k}$ des personnes répondant en $k^{\text{ième}}$ interrogation lors du trimestre t , tandis que $s_r^{t-d,k-1}$ joue le rôle de l'échantillon brut, avec $d = 1, 4$ ou 5 , selon que $k = 2, 3$ ou 4 . En pratique, la probabilité de réponse entre deux vagues est plus élevée que la probabilité initiale dans la plupart des enquêtes, et dépasse généralement les 90% dans le cadre de l'ESPA. Une analyse plus détaillée des taux de réponse en fonction de la strate de tirage (définie sur la base du canton) est disponible dans l'annexe C.

6.1 Variables auxiliaires disponibles

Les variables auxiliaires utilisées sont de deux types.

- Variables auxiliaires provenant des différentes sources de données administratives et contenues dans le cadre d'extrapolation construit pour le trimestre t . A cet effet, on réalise un appariement entre l'échantillon $s_r^{t-d,k-1}$, qui joue le rôle d'échantillon brut, et le cadre d'extrapolation.
- Certaines informations relevées lors de la précédente participation, qui sont donc disponibles pour l'échantillon brut ressollicité lors du trimestre t , telles que le statut sur le marché de l'emploi, la formation, ou le taux d'occupation (plein temps/temps partiel).

6.2 Méthode d'estimation

La méthode d'estimation est identique à celle utilisée pour la non-réponse initiale.

- Des groupes de réponse homogènes sont constitués sur la base d'un algorithme de segmentation appliqué à l'échantillon brut (personnes ayant répondu lors de la vague précédente).
- Les poids de sondage ne sont pas utilisés lors de l'application de cet algorithme. Chaque unité reçoit un poids égal à 1.
- La variable d'intérêt de l'algorithme de segmentation est la variable binaire indicatrice d'une nouvelle participation ou non à l'enquête lors du trimestre t . Les variables auxiliaires sont décrites dans le paragraphe ci-dessus. La variable indiquant si la personne dispose ou non d'un raccordement téléphonique fixe (tel/altel) est également incluse dans les variables auxiliaires « candidates » pour la segmentation. Cependant, de façon prévisible, elle ne joue pas de rôle important dans le mécanisme d'érosion et n'est généralement pas retenue par l'algorithme de segmentation.
- La taille minimale des groupes de réponse homogènes est fixée à 100, et le taux de réponse minimal à 10%.
- Au sein de chaque groupe de réponse homogène $g_{c,k}^t$, la probabilité de réponse $\tilde{p}_{c,k}^t$ entre la $k^{\text{ième}}$ interrogation et l'interrogation précédente est estimée par $R_{c,k}^t/N_{c,k}^t$, où $R_{c,k}^t$ désigne le nombre de répondants au sein du groupe de réponse homogène $g_{c,k}^t$ et $N_{c,k}^t$ sa taille.

6.3 Performances de la modélisation

Les tableaux 4 et 5 sont l'analogue des tableaux 1 et 2 dans le cas de l'érosion entre la vague 1 et la vague 2, sur l'exemple de la vague 2 du troisième trimestre 2015 (là aussi, les estimations retenues sont celles basées sur la totalité des variables auxiliaires disponibles).

T 4 – Pouvoir prédictif de la modélisation du mécanisme d'érosion entre la vague 1 et la vague 2, sur l'exemple de la vague 2 du troisième trimestre 2015

	Toutes les variables auxiliaires	Seulement les variables auxiliaires du SRPH
Pourcentage concordants	67.99	66.91
Pourcentage discordants	27.13	27.15
Pourcentage tied	4.88	5.93
C-statistic	0.70	0.70

T 5 – Distribution des probabilités de réponse estimées entre la vague 1 et la vague 2, sur l'exemple de la vague 2 du troisième trimestre 2015

	Min	P10	P25	P50	P75	P90	Max
Toutes les variables auxiliaires	0.76	0.87	0.90	0.95	0.96	0.99	0.99
Seulement les variables auxiliaires du SRPH	0.76	0.86	0.91	0.93	0.96	0.99	1.0

Sur la base de ces résultats, on peut faire les constatations suivantes.

- Si l'on suit la règle pratique de [Hosmer & Lemeshow \(2000\)](#), le pouvoir prédictif de la modélisation du mécanisme d'érosion entre la vague 1 et la vague 2 peut être considéré comme acceptable, puisque la C-statistic atteint le seuil de 0.7.

- L'exploitation des données administratives autres que celles contenues dans le SRPH a très peu d'impact sur les résultats.
- Comme cela est fréquemment observé dans les enquêtes par panel, une grande partie de la non-réponse est attribuable à la non-réponse initiale plutôt qu'à l'érosion. En effet, sur l'exemple du troisième trimestre 2015, la médiane de la probabilité de réponse initiale estimée est de l'ordre de 60% (cf. tableau 2), tandis que la médiane de la probabilité de réponse estimée entre la vague 2 et la vague 1 est de l'ordre de 95% (cf. tableau 5).

Les résultats présentés dans le tableau 6 permettent d'évaluer l'effet de la correction de l'érosion entre les vagues 1 et 2, pour le modèle de réponse retenu sur la distribution des poids, à savoir celui basé sur l'exploitation de la totalité des variables auxiliaires.

T 6 – Effet de la correction de l'érosion entre les vague 1 et 2 sur la distribution des poids, vague 2 du troisième trimestre 2015

	Min	P10	P25	P50	P75	P90	Max	Max/Min
Poids corrigées pour la non-réponse en vague 1	194	357	485	716	867	1154	4590	23.7
Poids corrigés pour la non-réponse en vague 2	197	387	532	753	925	1266	5511	28.0

Sur la base des personnes réinterrogées lors du troisième trimestre 2015, les constatations que l'on peut faire pour l'érosion entre la vague 2 et la vague 3, ainsi qu'entre la vague 3 et la vague 4, sont similaires.

7 Facteur de combinaison des panels

Les poids corrigés pour la non-réponse $w_{i,k}^{t,NR}$ obtenus à l'issue de la première étape de la pondération d'un trimestre t (cf. sections 3-6) vérifient les propriétés suivantes.

- Pour le panel répondant en vague 1, la somme des poids $w_{i,1}^{t,NR}$ fournit une estimation de la taille N de la population cible des personnes âgées de 15 ans et plus :

$$\sum_{i \in s_r^{t,1}} w_{i,1}^{t,NR} \simeq N.$$

- Pour les panels répondant en vague $k = 2, 3, 4$, la somme des poids $w_{i,k}^{t,NR}$ fournit une estimation du nombre de personnes âgées de 15 à 75 ans, noté N_{75-} :

$$\sum_{i \in s_r^{t,k}} w_{i,k}^{t,NR} \simeq N_{75-}.$$

Lorsque l'on réunit les quatre panels du trimestre t , on obtient :

$$\sum_{k=1}^4 \sum_{i \in s_r^{t,k}} w_{i,k}^{t,NR} \simeq N_{75+} + 4 \cdot N_{75-},$$

où N_{75+} désigne le nombre de personne âgées de 75 ans et plus. Cette dernière relation met en évidence le fait que les poids $w_{i,k}^{t,NR}$ doivent être ajustés afin que leur somme sur la réunion des quatre panels fournisse une estimation de la taille N de la population cible des 15 ans et plus. Il faut pour cela tenir compte du fait qu'un même individu peut être intégré à l'enquête par l'intermédiaire de différentes vagues et donc de différents tirages. La théorie du partage des

poids (cf. [Lavallée, 2002](#)) propose une solution à cette question. Cette solution se base sur un comptage du nombre de panels par lesquels une personne donnée aurait pu être interrogée. Ce comptage dépend de l'âge de la personne lors de l'enquête.

- **Pour les moins de 75 ans**

Ces personnes auraient pu être atteintes par l'intermédiaire des quatre panels, ce qui correspond à quatre liens dans la théorie du partage des poids. Ces liens sont pondérés par la taille des panels, pour un gain d'efficacité en termes de variance. Les poids du panel k du trimestre t donné sont donc multipliés par un facteur proportionnel à taille de ce panel :

$$f_{i,k}^t = f_{k,75-}^t = \frac{n_{k,75-}^t}{\sum_{k=1}^4 n_{k,75-}^t},$$

où $n_{k,75-}^t$ désigne le nombre de répondants de moins de 75 ans au sein du panel interrogé en vague k lors du trimestre t .

- **Pour les 76 ans et plus**

Comme ils ne sont représentés qu'en vague 1, le facteur $f_{i,k}^t$ est égal à 1, de sorte que les poids sont conservés tels quels.

- **Pour les personnes de 75 ans**

Il faut compter le nombre de panels par lequel une personne de 75 ans pourrait être interrogée dans le trimestre t considéré, en s'appuyant sur le schéma de rotation de l'annexe [A.2](#). On suppose que les interviews ont lieu en fin de trimestre, ce qui n'est pas complètement exact en pratique mais permet de simplifier le raisonnement. En pratique, il devrait donc y avoir des « effets de bords » qui se compensent entre eux.

- Les personnes atteignant l'âge de 75 ans durant le trimestre t peuvent être échantillonnées dans les 4 vagues de t .
- Les personnes atteignant l'âge de 75 ans durant $(t - 1)$ peuvent être échantillonnées dans les vagues 1 et 3 de t , mais pas dans les vagues 2 ou 4 de t , car elles avaient déjà atteint l'âge de 75 ans dans les vagues 1 et 3 de $(t - 1)$. Leur échantillonnage est donc possible dans deux vagues. Pour illustrer ce raisonnement, considérons l'exemple de l'enquête du quatrième trimestre 2020, et des personnes atteignant l'âge de 75 ans durant le troisième trimestre 2020. Ces personnes peuvent être interrogées en vague 1. Elles ne peuvent en revanche pas être interrogées en vague 2 du quatrième trimestre 2020, car cette vague provient de la vague 1 du troisième trimestre 2020, et les personnes concernées avaient déjà atteint l'âge de 75 ans à cette date. Lorsqu'elles arrivent en vague 2 du quatrième trimestre 2020, elles ont déjà participé une fois à l'enquête à l'âge de 75 ans, et ne sont plus réinterrogeables. De façon similaire, une interrogation en vague 4 du quatrième trimestre 2020 n'est pas possible, car cette vague provient de la vague 3 du troisième trimestre 2020, durant laquelle ces personnes ont déjà été interrogées à l'âge de 75 ans. Une interrogation en vague 3 du quatrième trimestre 2020 est en revanche possible, puisque cette vague provient de la vague 2 du premier trimestre 2020, et que ces personnes n'avaient pas encore atteint l'âge de 75 ans à cette date. Lorsqu'elles arrivent en vague 3 du quatrième trimestre 2020, elles n'ont donc pas encore été interrogées à l'âge de 75 ans, de sorte que leur participation est sollicitée.
- Les personnes atteignant l'âge de 75 ans durant $(t - 2)$ peuvent être échantillonnées dans les vagues 1 et 3 de t , mais pas dans les vagues 2 ou 4 de t , car elles avaient déjà atteint l'âge de 75 ans dans les vagues 1 et 3 de $(t - 1)$. Leur échantillonnage est donc possible dans deux vagues.
- Les personnes atteignant l'âge de 75 ans durant $(t - 3)$ ne peuvent être échantillonnées qu'en vague 1 de t . Leur échantillonnage n'est en effet pas possible dans les vagues 2 ou 4

de t , car elles avaient déjà atteint l'âge de 75 ans dans les vagues 1 et 3 de $(t-1)$. De même, leur échantillonnage n'est pas possible dans la vague 3 de t , car elles avaient déjà atteint l'âge de 75 ans dans la vague 2 de $(t-3)$. Leur échantillonnage n'est donc possible que dans une seule vague.

Cela explique pourquoi il y a en pratique très peu de personnes de 75 ans dans les vagues 2 et 4. Par ailleurs, sous l'hypothèse d'une distribution uniforme des naissances, ce décompte justifie pour les 75 ans un facteur de combinaison unique de :

$$f_{i,k}^t = 0.25 \cdot 0.25 + 0.5 \cdot 0.5 + 0.25 \cdot 1 = 0.5625,$$

qui s'obtient en observant que :

- un quart des personnes de 75 ans auraient pu être échantillonnées dans les 4 vagues de t (celles qui atteignent l'âge de 75 ans durant t),
- la moitié des personnes de 75 ans auraient pu être échantillonnées dans deux vagues (celles qui atteignent l'âge de 75 ans durant $t-1$ ou $t-2$),
- un quart des personnes de 75 ans n'auraient pu être échantillonnées que dans une seule vague (celles qui atteignent l'âge de 75 ans durant $t-3$).

Ici, les liens ne sont pas pondérés par les tailles d'échantillon, du fait des faibles effectifs et de l'instabilité que cela pourrait donc causer.

8 Pondération annuelle

Comme mentionné dans les sections 1 et 2, l'échantillon annuel est composé des vagues 1 et 3 des quatre trimestres de l'année considérée. Certaines caractéristiques plus stables dans le temps que le statut sur le marché de l'emploi, comme par exemple le niveau de formation, sont en effet relevées seulement lors des vagues 1 et 3, et sont exploitées sur la base de l'échantillon annuel. De part le schéma de rotation de l'enquête, les huit panels qui composent l'échantillon annuel sont issus de huit tirages différents, et la coordination négative des tirages permet d'éviter les doublons, c'est-à-dire qu'un même individu n'apparaît qu'une seule fois dans cet échantillon. La pondération annuelle, qui développée et implémentée par la section AES, reprend les grands principes de la pondération trimestrielle et suit les mêmes étapes.

1. Correction de la non-réponse

Pour chacun des huit panels composant l'échantillon annuel, on reprend le poids corrigé pour la non-réponse $w_{i,k}^{t,NR}$ calculé lors de la pondération trimestrielle, où t désigne l'un des quatre trimestres de l'année considérée, et $k = 1$ ou 3 désigne la vague.

2. Combinaison des panels

Les huit panels sont ensuite réunis et les poids de l'étape 1 sont ajustés afin que leur somme sur l'échantillon annuel fournisse une estimation de la taille de la population cible (15 ans et plus). Les moins de 75 ans sont traités séparément des 75 ans et plus, sans traitement spécifique des 75 ans.

– Cas des moins de 75 ans

Ils sont présents dans chacun des huit panels de l'échantillon annuel, et pour chacun de ces panels, la somme de leurs poids corrigés pour la non-réponse fournit une approximation du nombre de personnes de moins de 75 ans dans la population cible. Lors de la réunion des huit panels, les poids corrigés pour la non-réponse doivent donc encore être divisés par un facteur approximativement égal à huit. Comme pour la pondération trimestrielle, nous appliquons un facteur proportionnel à la taille des panels. Les poids du panel $k = 1$ ou 3 de l'un des quatre trimestres t de l'année considérée sont donc

multipliés par :

$$f_k^t = \frac{n_{k,75-}^t}{\sum_{t=1}^4 (n_{1,75-}^t + n_{3,75-}^t)},$$

où $n_{k,75-}^t$ désigne le nombre de répondants de moins de 75 ans au sein du panel interrogé en vague k lors du trimestre t .

– Cas des 75 ans et plus

Ils ne sont présents que dans les quatre vagues 1 de l'échantillon annuel. Lors de la réunion des panels, les poids corrigés pour la non-réponse doivent donc être divisés par un facteur approximativement égal à quatre. Ici également, nous appliquons un facteur proportionnel à la taille des panels. Les poids du panel répondant en vague $k = 1$ de l'un des quatre trimestres t de l'année considérée sont donc multipliés par :

$$f_{1,75+}^t = \frac{n_{1,75+}^t}{\sum_{t=1}^4 n_{1,75+}^t},$$

où $n_{1,75+}^t$ désigne le nombre de répondants de plus de 75 ans au sein du panel interrogé en vague 1 lors du trimestre t .

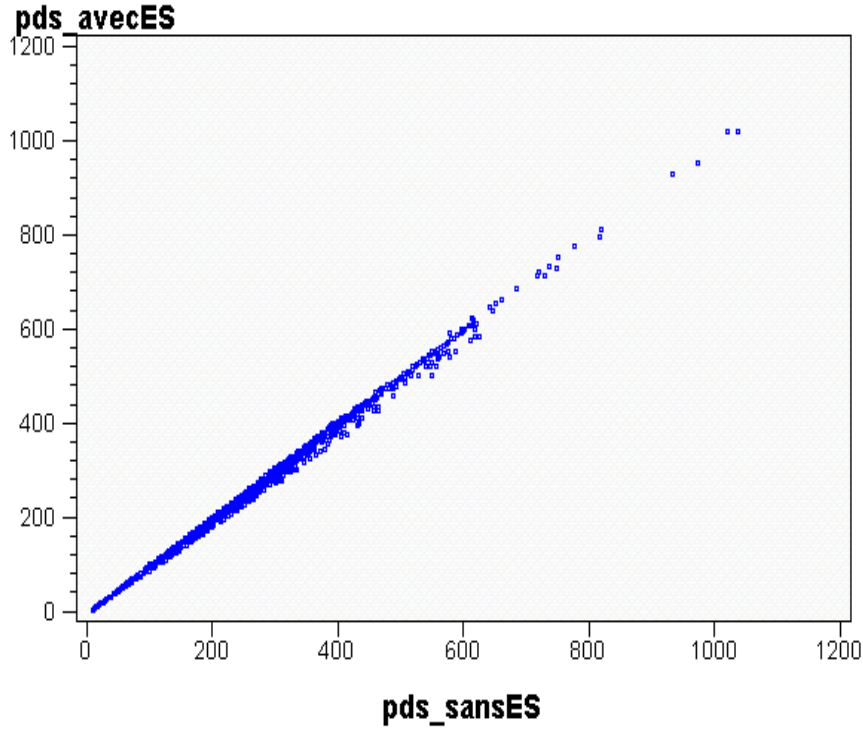
3. Calage

Le calage est effectué grâce à la macro CALMAR de SAS, avec la fonction de distance correspondant au raking ratio, et les variables de calage utilisées pour la pondération trimestrielle sont reprises pour le calage de l'échantillon annuel. La pondération annuelle a cependant la particularité suivante : par souci de cohérence entre les différentes exploitations, les poids sont également calés de façon à ce que l'estimation du nombre d'actifs basée sur la pondération annuelle coïncide avec celle obtenue en calculant la moyenne des quatre estimations trimestrielles de l'année considérée, et ce pour différents niveaux de ventilation. La prise en compte du nombre d'actifs dans le calage est ce qui est appelé le « calage Eurostat ».

L'effet de la prise en compte du nombre d'actifs dans le calage a été analysé. Pour cela deux jeux de poids ont été calculés pour l'année 2011 : un premier jeu de poids (pds_avecES) qui tient compte de toutes les variables de calage, y compris le nombre d'actifs, et un deuxième jeu de poids (pds_sansES) qui ne tient compte que des variables de calage « habituelles ». La figure 2 permet de comparer ces deux jeux de poids. Pour chaque individu, le rapport entre les deux poids a également été calculé, et le tableau 7 contient quelques statistiques descriptives, telle que la moyenne et quelques quantiles, sur ce rapport. On constate que, globalement, la prise en compte du nombre d'actifs a peu d'effet sur les poids.

T 7 – Quelques statistiques descriptives sur le rapport entre les deux jeux de poids

Moyenne	P0	P5	P10	P25	P50	P75	P90	P95	P100
1.00003	0.93928	0.99169	0.9949	0.99847	1.00028	1.00187	1.00414	1.00579	1.07546



9 Estimation de la précision pour la pondération trimestrielle et annuelle

Estimation d'un total

Nous nous intéressons tout d'abord à la variance de l'estimation d'un total, par exemple le nombre d'actifs, obtenue sur la base des poids calés w_i^t d'un échantillon trimestriel s_r^t de taille n^t :

$$\hat{Y} = \sum_{i \in s_r^t} w_i^t y_i. \quad (7)$$

Grâce aux résultats de [Deville & Särndal \(1992\)](#), une approximation de la variance de $\hat{Y}$ est donnée par :

$$\text{var}(\hat{Y}) \simeq \widehat{\text{var}} \left(\sum_{i \in s_r^t} w_i^{t,comb} E_i \right), \quad (8)$$

où $w_i^{t,comb}$ désigne les poids de départ avant calage (cf. étape 3 de la pondération, section 3), et E_i désigne les résidus de la régression de la variable d'intérêt y_i sur les variables de calage, la régression étant ajustée sur l'ensemble de la population cible U . Les résidus E_i étant inconnus en pratique, une estimation de la variance de $\hat{Y}$ basée sur l'approximation (8) est obtenue en remplaçant les E_i par les résidus e_i obtenus en ajustant la régression sur l'échantillon de

répondants s_r^t :

$$\widehat{\text{var}}(\widehat{Y}) \simeq \widehat{\text{var}}\left(\sum_{i \in s_r^t} w_i^{t,comb} e_i\right). \quad (9)$$

Comme le recommandent [Deville & Särndal \(1992\)](#), les poids calés w_i^t sont utilisés pour l'ajustement de cette régression, car ces poids sont considérés comme plus performants que les poids $w_i^{t,comb}$ pour estimer des paramètres sur la base de l'échantillon. L'équation (9) montre que le problème est ramené à l'estimation de la variance sous plan de sondage sans calage de $\sum_{i \in s_r^t} w_i^{t,comb} e_i$. Nous nous basons pour traiter cette question sur l'approximation de Deville rappelée par [Matei & Tillé \(2005\)](#), qui ont également étudié ses propriétés :

$$\widehat{\text{var}}\left(\sum_{i \in s_r^t} w_i^{t,comb} e_i\right) = \sum_{i \in s_r^t} \frac{c_i}{(\pi_i^{comb})^2} (e_i - \hat{e}_i^*)^2, \quad (10)$$

où

$$\hat{e}_i^* = \pi_i^{comb} \frac{\sum_{j \in s_r^t} c_j e_j / \pi_j^{comb}}{\sum_{j \in s_r^t} c_j} \quad (11)$$

avec

$$\pi_i^{comb} = \frac{1}{w_i^{t,comb}} \quad \text{et} \quad c_i = (1 - \pi_i^{comb}) \frac{n^t}{n^t - 1}. \quad (12)$$

Estimation d'un ratio

Des estimations de ratios, tels que le taux d'actifs ou de chômeurs, sont également produits sur la base de cette enquête. Nous désignons par $R = Y/X$ le ratio des totaux de deux variables prenant les valeurs y_i et x_i pour l'individu i . Dans le cas du taux de chômage par exemple, y_i vaut 1 si l'individu est au chômage et 0 sinon, et x_i est constante et égale à 1, de sorte que $X = \sum x_i = N$. Nous désignons par $\widehat{R} = \widehat{Y}/\widehat{X}$ l'estimation de R , avec :

$$\widehat{Y} = \sum_{i \in s_r^t} w_i^t y_i \quad \text{et} \quad \widehat{X} = \sum_{i \in s_r^t} w_i^t x_i.$$

Afin d'estimer la variance de $\widehat{R}$, nous utilisons des techniques de linéarisation dues à [Deville \(1999\)](#), qui montre que :

$$\widehat{\text{var}}(\widehat{R}) = \widehat{\text{var}}\left(\sum_{i \in s_r^t} w_i^t \hat{\ell}_i\right), \quad (13)$$

où la linéarisée $\hat{\ell}_i$ est définie par :

$$\hat{\ell}_i = y_i - \frac{\widehat{Y}}{\widehat{X}} x_i. \quad (14)$$

La formule (13) montre que l'estimation de la variance de $\widehat{R}$ peut être ramenée à l'estimation de la variance d'un total, qui est traitée dans le paragraphe précédent. Il suffit alors d'utiliser les formules (7), (9) et (10), en remplaçant la variable d'intérêt y_i par la linéarisée $\hat{\ell}_i$.

Estimation dans un domaine

Les résultats (totaux ou ratios) sont par ailleurs ventilés au sein de différentes sous-populations, par exemple par classes d'âges ou par sexe (estimation dans un domaine). Le calcul de la variance de l'estimation d'un total dans un domaine est un problème bien connu en théorie de l'échantillonnage, qui se traite en observant qu'un total Y_d calculé sur un domaine d peut s'exprimer sous la forme :

$$Y_d = \sum_{i \in U_d} y_i = \sum_{i \in U} y_i \mathbf{1}_{i \in U_d},$$

où U désigne la population cible de l'enquête, U_d le sous-ensemble de U dans le domaine considéré d , et où $\mathbf{1}_{i \in U_d}$ est une variable indicatrice de l'appartenance au domaine d . De façon similaire, l'estimation $\hat{Y}_d$ de Y_d peut s'exprimer sous la forme :

$$\hat{Y}_d = \sum_{i \in s_r^t \cap U_d} w_i^t y_i = \sum_{i \in s_r^t} w_i^t y_i \mathbf{1}_{i \in U_d} = \sum_{i \in s_r^t} w_i^t y_i^*, \quad (15)$$

où

$$y_i^* = \begin{cases} y_i & \text{si } i \in U_d, \\ 0 & \text{si } i \notin U_d. \end{cases}$$

La formule (15) montre que l'estimation de la variance de $\hat{Y}_d$ s'obtient en appliquant les formules (7), (9) et (10) au niveau national, en remplaçant la variable d'intérêt y_i par la variable y_i^* . De la même façon, l'estimation d'un ratio dans un domaine d peut s'exprimer sous la forme :

$$\hat{R}_d = \frac{\sum_{i \in s_r^t \cap U_d} w_i^t y_i}{\sum_{i \in s_r^t \cap U_d} w_i^t x_i} = \frac{\sum_{i \in s_r^t \cap U_d} w_i^t y_i \mathbf{1}_{i \in U_d}}{\sum_{i \in s_r^t \cap U_d} w_i^t x_i \mathbf{1}_{i \in U_d}} = \frac{\sum_{i \in s_r^t} w_i^t y_i^*}{\sum_{i \in s_r^t} w_i^t x_i^*} = \frac{\hat{Y}^*}{\hat{X}^*}, \quad (16)$$

où

$$y_i^* = \begin{cases} y_i & \text{si } i \in U_d, \\ 0 & \text{si } i \notin U_d, \end{cases} \quad \text{et} \quad x_i^* = \begin{cases} x_i & \text{si } i \in U_d, \\ 0 & \text{si } i \notin U_d. \end{cases}$$

La formule (16) montre que l'estimation de la variance de $\hat{R}_d$ se ramène à l'estimation de la variance d'un ratio défini sur l'ensemble de la population sur la base des variables y_i^* et x_i^* . La procédure suivante peut donc être utilisée pour estimer la variance de $\hat{R}_d$.

- Calcul de la linéarisée $\hat{\ell}_i^*$ du ratio $\hat{R}_d = \hat{Y}^*/\hat{X}^*$ sur la base de la formule (14), en remplaçant y_i et x_i par y_i^* et x_i^* :

$$\widehat{\text{var}}(\hat{R}_d) = \widehat{\text{var}}\left(\sum_{i \in s_r^t} w_i^t \hat{\ell}_i^*\right)$$

où

$$\hat{\ell}_i^* = y_i^* - \frac{\hat{Y}^*}{\hat{X}^*} x_i^* = y_i \mathbf{1}_{i \in U_d} - \frac{\hat{Y}_d}{\hat{X}_d} x_i \mathbf{1}_{i \in U_d}. \quad (17)$$

- Application des formules (7), (9) et (10), en remplaçant la variable d'intérêt y_i par la linéarisée $\hat{\ell}_i^*$.

Opérationnalisation

De façon résumée et d'un point de vue opérationnel, l'implémentation de la méthode d'estimation de variance présentée dans cette section suit les étapes suivantes.

1. S'il s'agit d'estimation dans un domaine, introduction des variables indicatrices $1_{i \in U_d}$ de l'appartenance au domaine d , comme dans les formules (15) et (16).
2. S'il s'agit de l'estimation d'un ratio, calcul de la linéarisée ($\hat{\ell}_i$ ou $\hat{\ell}_i^*$ selon que le ratio soit estimé au niveau national ou bien dans un domaine).
3. Calcul des résidus de la régression par rapport aux variables de calage, afin de tenir compte de l'effet du calage dans le calcul de la variance. La variable expliquée de cette régression dépend du type d'estimation envisagée, comme résumé dans le tableau 8.

T 8 – Régression sur les variables de calage

Type d'estimation	Variable expliquée par la régression
Estimation d'un total Y au niveau national	y_i
Estimation d'un total Y_d dans un domaine d	$y_i^* = y_i 1_{i \in U_d}$
Estimation d'un ratio $R = Y/X$ au niveau national	$\hat{\ell}_i$, cf. formule (14)
Estimation d'un ratio $R_d = Y_d/X_d$ dans un domaine d	$\hat{\ell}_i^*$, cf. formule (17)

4. Application de la formule (10), en remplaçant e_i par les résidus obtenus à l'étape 3.

Une macro SAS implémentant cette procédure a été développée, testée et livrée à la section AES.

Illustration

Le tableau 9 présente les estimations de la précision obtenues sur la base des formules précédentes pour quelques paramètres d'intérêt de l'enquête. La précision est ici exprimée en termes de coefficient de variation (CV). Un deuxième calcul est également réalisé sans tenir compte du calage, c'est-à-dire en sautant l'étape de calcul des résidus de la régression par rapport aux variables de calage. Le gain de précision dû au calage est alors exprimé comme la différence relative entre le CV sans tenir compte du calage et celui en tenant compte. Enfin, le tableau 9 contient également un troisième calcul de précision, qui consiste à calculer le coefficient de variation qui serait obtenu si le calage ne se basait que sur des variables auxiliaires issues du SRPH, ainsi que le gain correspondant par rapport à l'absence de calage. La comparaison du gain permis par les deux types de calage (variables auxiliaires issues du SRPH uniquement ou exploitation d'autres sources d'informations auxiliaires) permet d'évaluer l'apport des autres sources d'informations auxiliaires utilisées.

Pour l'estimation du nombre d'actifs, on constate que le calage sur les variables auxiliaires issues des différentes sources (pas uniquement le SRPH) permet un gain de précision allant d'environ 50% à 65% dans les différentes sous-populations considérées dans le tableau 9, à l'exception des personnes âgées de 65 ans et plus pour lesquelles le gain est beaucoup plus faible. Des gains de 50% et 65% sont très conséquents et correspondent approximativement à une multiplication par 4 et 8 respectivement de la taille de l'échantillon⁵. Le gain pour l'esti-

5. Sous l'hypothèse que le plan est proche d'un plan simple, et que le taux de sondage est négligeable, le CV de l'estimation $\hat{T}$ d'un total est donné par $CV(\hat{T}) = \sqrt{N^2/n} \cdot CV_y$ où N désigne la taille de la population, n celle de l'échantillon, et CV_y le coefficient de variation de la variable d'intérêt. Par conséquent, la taille d'échantillon $\tilde{n}$ nécessaire pour atteindre un gain de $x\%$ sur la précision vérifie $\sqrt{N^2/\tilde{n}} \cdot CV_y = (1 - x/100) \sqrt{N^2/n} \cdot CV_y$, c'est-à-dire $\tilde{n}/n = 1/(1 - x/100)^2$. Par exemple, un gain de précision de 65% correspond à $\tilde{n}/n = 1/(1 - 0.65)^2 = 8.16$, c'est à dire à une multiplication par environ 8 de la taille de l'échantillon.

T 9 – Effet du calage sur la précision

	Estimation	CV (en %)	CV (en%) sans calage	CV (en%) avec calage uniquement sur SRPH	Gain (en %) dû au calage	Gain (en %) dû au calage uniquement sur SRPH
Nombre d'actifs	4765751.30	0.27	0.56	0.31	51.7	44.1
Hommes	2550151.32	0.33	0.94	0.38	64.9	59.8
Femmes	2215599.98	0.45	1.01	0.53	55.3	47.7
15-24 ans	633680.77	1.03	2.13	1.08	51.5	49.3
25-64 ans	3960399.41	0.24	0.66	0.30	63.4	55.2
65 ans et plus	171671.12	3.39	4.05	3.66	16.2	9.6
Nombre de chômeurs	245790.51	3.02	3.55	3.44	15.0	3.3
Hommes	124522.38	4.14	5.14	4.90	19.5	4.7
Femmes	121268.13	4.44	4.95	4.81	10.4	2.8
15-24 ans	68583.32	5.44	5.82	5.58	6.5	4.1
25-64 ans	175906.74	3.63	4.41	4.25	17.7	3.6
65 ans et plus	1300.45	60.72	62.04	60.82	2.1	2.0

mation du nombre de chômeurs est moins important mais reste appréciable, s'élevant de 7% à 20% (toujours à l'exception des personnes âgées de 65 ans et plus). Ces gains correspondent approximativement à une augmentation respective de 15% et 55% de la taille de l'échantillon. Le faible gain observé chez les 65 ans et plus peut s'expliquer par le fait que faire partie de la population active au-delà de 65 ans est relativement rare et peu corrélé aux variables auxiliaires à disposition.

Si l'on s'intéresse à l'effet du calage sur les variables auxiliaires issues uniquement du SRPH, on constate que le gain de précision varie entre 44% et 60% pour l'estimation du nombre d'actifs, à l'exception des personnes âgées de 65 ans. Le gain pour l'estimation du nombre de chômeurs varie quant à lui de 3% à 5%. Ces résultats mettent en évidence l'apport des autres sources d'informations auxiliaires que le SRPH pour l'estimation du nombre d'actifs. Le gain permis par ces autres sources est encore beaucoup plus net pour l'estimation du nombre de chômeurs.

Généralisation à la pondération annuelle

Nous abordons ici la question de l'estimation de la variance d'un total ou d'un ratio calculés sur la base de la pondération annuelle. Nous excluons de cette étude les variables relatives aux revenus qui comportent beaucoup de valeurs manquantes et ne sont pas exploitées directement par la section AES. Les cas traités permettent de calculer la précision des estimations présentées dans les tableaux standards. Comme le montrent les sections 3 et 8, les pondérations trimestrielles et annuelles sont construites selon les mêmes principes, à l'exception de la particularité de la pondération annuelle appelée « calage Eurostat » et décrite dans la section 8, qui a pour but de garantir la cohérence entre l'estimation du nombre d'actifs basée sur la pondération annuelle et celle basée sur une moyenne des quatre trimestres de l'année considérée, et ce pour différents niveaux de ventilation. Cette particularité est le seul élément qui pourrait éventuellement poser problème pour étendre à la pondération annuelle la méthode d'estimation de variance développée pour la pondération trimestrielle et décrite précédemment. Compte tenu du faible impact de cette particularité sur les poids obtenus (cf. figure 2 et tableau 7), elle est négligée dans les calculs de variances, de sorte que la méthode d'estimation de variance développée pour la pondération trimestrielle et la macro SAS correspondante peuvent être utilisées.

10 Evaluation de la précision de l'estimation d'une moyenne sur plusieurs trimestres

Principe général

Nous décrivons ici la méthode utilisée pour évaluer la précision de l'estimation d'une moyenne sur les quatre trimestres d'une même année, la méthode se généralisant facilement pour un plus grand nombre de trimestres. Nous nous intéressons tout d'abord à la variance de l'estimation de la moyenne de totaux Y_t sur les quatre trimestres t d'une même année, par exemple le nombre moyen de chômeurs sur les quatre trimestres :

$$\hat{Y} = \frac{1}{4} \sum_{t=1}^4 \hat{Y}_t.$$

La variance de $\hat{Y}$ peut être développée sous la forme :

$$\begin{aligned} \text{var}(\hat{Y}) = \frac{1}{16} & \left[\text{var}(\hat{Y}_1) + 2\text{cov}(\hat{Y}_1, \hat{Y}_2) + 2\text{cov}(\hat{Y}_1, \hat{Y}_3) + 2\text{cov}(\hat{Y}_1, \hat{Y}_4) \right. \\ & + \text{var}(\hat{Y}_2) + 2\text{cov}(\hat{Y}_2, \hat{Y}_3) + 2\text{cov}(\hat{Y}_2, \hat{Y}_4) \\ & + \text{var}(\hat{Y}_3) + 2\text{cov}(\hat{Y}_3, \hat{Y}_4) \\ & \left. + \text{var}(\hat{Y}_4) \right]. \end{aligned}$$

Comme le montre le schéma de rotation (cf. annexe A.1), le recouvrement entre les trimestres d'une même année est d'environ 50% entre deux trimestres consécutifs t et $t+1$, de 25% entre les trimestres t et $t+3$, et 0% entre t et $t+2$. Par conséquent, les termes $\text{cov}(\hat{Y}_1, \hat{Y}_3)$ et $\text{cov}(\hat{Y}_2, \hat{Y}_4)$ s'annulent⁶, de sorte que :

$$\begin{aligned} \text{var}(\hat{Y}) = \frac{1}{16} & \left[\text{var}(\hat{Y}_1) + \text{var}(\hat{Y}_2) + \text{var}(\hat{Y}_3) + \text{var}(\hat{Y}_4) \right. \\ & + 2\text{cov}(\hat{Y}_1, \hat{Y}_2) + 2\text{cov}(\hat{Y}_1, \hat{Y}_4) \\ & \left. + 2\text{cov}(\hat{Y}_2, \hat{Y}_3) + 2\text{cov}(\hat{Y}_3, \hat{Y}_4) \right]. \end{aligned} \quad (18)$$

Les quatre termes de variance peuvent être estimés selon la méthode décrite dans la section 9 consacrée à l'estimation de la variance des exploitations trimestrielles. Pour estimer les quatre termes de covariance, nous nous basons sur les travaux de [Qualité & Tillé \(2008\)](#). Nous faisons donc l'approximation que les termes de covariance peuvent s'exprimer sous la forme suivante :

$$2\text{cov}(\hat{Y}_{t_1}, \hat{Y}_{t_2}) \simeq \left[\text{var}(\hat{Y}_{t_1}) + \text{var}(\hat{Y}_{t_2}) \right] \rho^{t_1, t_2} \cdot q^{t_1, t_2}, \quad (19)$$

où q^{t_1, t_2} désigne le taux de recouvrement tel que défini par [Qualité & Tillé \(2008\)](#) entre les échantillons des trimestres t_1 et t_2 , et ρ^{t_1, t_2} désigne le coefficient de corrélation entre la mesure de la « variable d'intérêt pour le calcul de variance » lors du trimestre t_1 et la mesure lors du trimestre t_2 . A noter que l'on entend ici par « variable d'intérêt pour le calcul de variance » les résidus de la régression sur les variables de calage, comme le montre la formule (9). Pour rappel, la variable expliquée de cette régression dépend du type d'estimation envisagée, cf. tableau 8.

6. Du moins approximativement, car du fait des faibles taux de sondage, les tirages des panels, qui sont coordonnés négativement, peuvent être considérés comme indépendants.

Pour finaliser l'implémentation d'une estimation de variance basée sur la formule (18), il ne reste alors plus qu'à estimer les quantités q^{t_1, t_2} et ρ^{t_1, t_2} . L'estimation du taux de recouvrement est donnée par :

$$q^{t_1, t_2} = \frac{2n^{t_1, t_2}}{n^{t_1} + n^{t_2}}, \quad (20)$$

où n^t désigne la taille de l'échantillon net lors du trimestre t et où n^{t_1, t_2} désigne la taille de l'intersection des échantillons des trimestres t_1 et t_2 . La corrélation ρ^{t_1, t_2} est estimée sur la base de la « proc corr » du logiciel SAS appliquée sur l'intersection des échantillons des trimestres t_1 et t_2 . Puisque deux jeux de poids sont disponibles pour cette intersection (ceux du trimestre t_1 et ceux du trimestre t_2), se pose alors la question duquel choisir. Le choix s'est porté sur les poids finaux du trimestre le plus récent, bien qu'il n'y ait pas d'argument marquant permettant de trancher entre les deux. Nous avons cependant constaté qu'en pratique, ce choix a un impact extrêmement limité.

Autres paramètres

Les moyennes sur plusieurs trimestres ne sont pas calculées seulement pour des totaux au niveau national, mais également pour d'autres paramètres tels que des ratios, par exemple le taux de chômage, ou encore des estimations dans des domaines. Nous désignons par $\widehat{R}$ la moyenne de ratios $\widehat{R}_t = \widehat{Y}_t / \widehat{X}_t$ sur les quatre trimestres d'une même année :

$$\widehat{R} = \frac{1}{4} \sum_{t=1}^4 \widehat{R}_t = \frac{1}{4} \sum_{t=1}^4 \frac{\sum_{i \in s_r^t} w_i^t y_{i,t}}{\sum_{i \in s_r^t} w_i^t x_{i,t}}.$$

On désigne par U le pooling avec comptage à double des unités communes des populations cibles U_t des quatre trimestres t , et par s le pooling des échantillons trimestriels s_r^t . On définit sur s les poids $w_i = w_i^t$ si $i \in s_r^t$, et sur U les quatre variables $y_i^{t,*} = y_i^t \mathbf{1}_{i \in U_t}$, ainsi que les quatre variables $x_i^{t,*} = x_{i,t} \mathbf{1}_{i \in U_t}$. La ratio moyen $\widehat{R}$ peut alors s'exprimer sous la forme :

$$\widehat{R} = \frac{1}{4} \sum_{t=1}^4 \frac{\sum_{i \in s} w_i y_i^{t,*}}{\sum_{i \in s} w_i x_i^{t,*}}.$$

La linéarisée d'une somme étant égale à la somme des linéarisées, on obtient l'approximation suivante de la variance de $\widehat{R}$:

$$\widehat{\text{var}}(\widehat{R}) \simeq \widehat{\text{var}}\left(\sum_{i \in s} w_i \widehat{\ell}_i\right) \quad (21)$$

où

$$\widehat{\ell}_i = \frac{1}{4} \sum_{t=1}^4 \left(y_{i,t} - \widehat{R}_t x_{i,t} \right) \mathbf{1}_{i \in U_t}.$$

La formule (21) peut se réécrire sous la forme suivante :

$$\widehat{\text{var}}(\widehat{R}) \simeq \widehat{\text{var}}\left(\frac{1}{4} \sum_{t=1}^4 \sum_{i \in s_r^t} w_i^t \widehat{\ell}_i^t\right), \quad (22)$$

où

$$\hat{\ell}_i^t = y_{i,t} - \hat{R}_t x_{i,t}.$$

La formule (22) montre que, grâce aux techniques de linéarisation, l'évaluation de la précision de la moyenne de ratios sur plusieurs trimestres se ramène à l'évaluation de la précision de la moyenne de totaux, qui est traitée dans le paragraphe précédent.

La question de l'estimation dans un domaine se traite comme dans la section 9, en introduisant une variable indicatrice de l'appartenance au domaine.

Illustration

Une macro SAS implémentant la méthode décrite dans cette section a été développée, testée et livrée à la section AES. Cette macro fait appel à la macro d'estimation de variance des exploitations trimestrielles qui apparaissent dans les formules (18) et (19). A titre d'illustration, nous présentons dans le tableau 10 les résultats obtenus pour la moyenne des estimations du nombre d'actifs sur les quatre trimestres de 2016, qui est la première année pour laquelle les sources d'informations auxiliaires autres que le SRPH ont été intégrées aux quatre pondérations trimestrielles. On constate que le CV des estimations trimestrielles se situe aux alentours

T 10 – Evaluation de la précision de l'estimation du nombre d'actifs en 2016

Trimestre 2016	Estimation	CV (en %)
1	4831541.17	0.26
2	4815069.54	0.27
3	4847527.92	0.28
4	4876106.92	0.27
Moyenne	4842561.39	0.17

de 0.27%, tandis que le CV de la moyenne est de 0.17%. On observe donc que le gain permis par le « pooling » des quatre trimestres est un peu moins important que si les quatre échantillons trimestriels étaient sélectionnés indépendamment, situation dans laquelle le CV de la moyenne serait approximativement égal la moitié du CV trimestriel. Ceci a trait aux panels rotatifs, qui sont très performants pour obtenir à la fois des estimations transversales et des estimations d'évolution, mais permettent un gain moins important en cas de pooling qu'un échantillonnage indépendant d'un trimestre à l'autre. Cette perte d'efficacité du pooling dépend du taux de recouvrement entre les échantillons trimestriels, cf. formule (20), ainsi que de la corrélation entre la « variable d'intérêt pour le calcul de variance » mesurée lors du trimestre t_1 et celle mesurée lors du trimestre t_2 . Les valeurs de ces paramètres obtenues pour l'évaluation de la précision de l'estimation du nombre d'actifs en 2016 sont présentées dans le tableau 11. Pour rappel, la « variable d'intérêt pour le calcul de variance » est dans ce cas le résidu de la régression de la variable binaire indiquant si l'individu est actif ou non sur les variables de calage. On constate que ces valeurs ne sont pas extrêmement élevées, particulièrement entre le premier et le quatrième trimestre. Par ailleurs, les échantillonnages des trimestres 1 et 3, ainsi que ceux des trimestres 2 et 4 sont considérés comme indépendants. Cela explique que la perte d'efficacité du pooling est ici modérée.

Limitations

L'évaluation de la précision de la moyenne de paramètres sur plusieurs trimestres nécessite l'estimation de corrélations sur la base de l'intersection de couples de trimestres. Cette intersection

T 11 – Taux de recouvrement et de corrélation utilisés pour l'évaluation de la précision de l'estimation du nombre d'actifs en 2016

Trimestres	Corrélation	Recouvrement
1-2	0.66	0.50
2-3	0.63	0.48
3-4	0.69	0.48
1-4	0.54	0.23

peut être plus ou moins réduite et est en tout cas plus petite que l'échantillon trimestriel. Par conséquent, si le pooling est utilisé pour produire des résultats à un niveau de ventilation plus détaillé que ce qui est fait sur la base d'un seul trimestre, les estimateurs de variance risquent d'être entachés d'une certaine instabilité.

11 Conclusions et perspectives

L'ESPA est une enquête complexe de par son plan d'échantillonnage basé sur un panel rotatif, ainsi que sa gestion spécifique des personnes âgées de 75 ans et plus. Cela nécessite un traitement méthodologique approprié, ainsi que l'exploitation d'informations auxiliaires, qui sont utiles pour modéliser la non-réponse initiale et les différentes phases d'érosion, mais également pour le calage. Par ailleurs, une procédure d'estimation de la variance est implémentée afin d'accompagner d'une mesure de précision les estimations publiées sur la base des exploitations trimestrielles ou annuelles. Une méthode d'estimation de la variance de l'estimation d'une moyenne sur plusieurs trimestres est également développée.

Le cadre de sondage (SRPH) contient des variables sociodémographiques qui peuvent être utilisées comme informations auxiliaires. Ces informations sont par ailleurs enrichies par d'autres sources de données administratives, telles que les données AVS et AI, les données concernant les prestations complémentaires aux rentes AVS-AI, le registre des comptes individuels (CI), le système d'information en matière de placement et de statistique du marché du travail (PLASTA), les systèmes informatisés de paiement des caisses de chômage (SIPAC), les données de l'aide sociale. Les analyses effectuées ont montré que les autres sources d'informations auxiliaires que le cadre de sondage ont un impact limité sur les performances de la modélisation du mécanisme de réponse. En revanche, l'impact de ces autres sources est beaucoup plus important si l'on s'intéresse au gain que l'on peut obtenir grâce au calage, qui est par ailleurs, toutes sources d'informations auxiliaires confondues, très appréciable pour certaines estimations. Au niveau national, il permet par exemple une réduction du coefficient de variation d'environ 50% pour l'estimation du nombre d'actifs et de 15% pour l'estimation du nombre de chômeurs.

Depuis le premier trimestre 2021, un passage au « mixed-mode » est progressivement introduit pour la réalisation des interviews, alors qu'elles étaient jusqu'alors exclusivement réalisées par téléphone. Ce changement est progressif, dans le sens où il n'est introduit que pour les panels sélectionnés pour une première participation à partir du premier trimestre 2021, mais ne concerne pas encore les panels sélectionnés antérieurement. Le passage au « mixed-mode » ne change rien aux principes des méthodes décrites dans ce rapport, qui peuvent continuer à être appliquées de la même manière. Il permet en revanche d'augmenter le taux de réponse. Il conviendra d'analyser en détails l'impact de ce changement sur le mécanisme de réponse et la distribution des poids lorsqu'il sera pleinement introduit, c'est-à-dire lors du deuxième trimestre 2022.

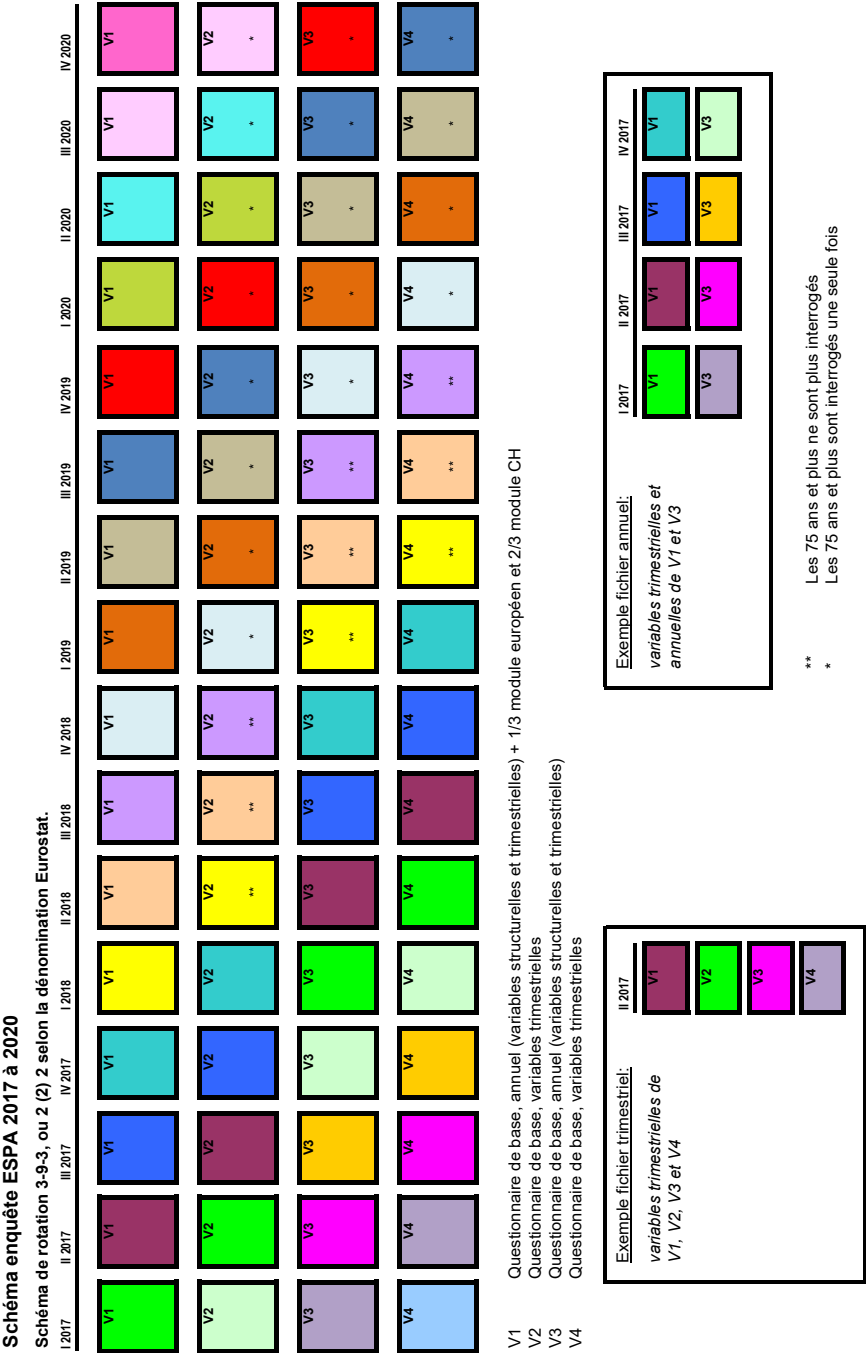
Bibliographie

- DEVILLE, J.-C. (1999). Estimation de variance pour des statistiques et des estimateurs complexes : techniques de résidus et de linéarisation. *Techniques d'enquête* **25**, 219–230.
- DEVILLE, J.-C. & SÄRNDAL, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association* **87**, 376–382.
- HOSMER, P. W. & LEMESHOW, S. (2000). *Applied Logistic Regression*. New York: Wiley.
- LAVALLÉE, P. (2002). *Le sondage indirect ou la méthode généralisée du partage des poids*. Paris: Ellipses.
- MATEI, A. & TILLÉ, Y. (2005). Evaluation of variance approximations and estimators in maximum entropy sampling with unequal probability and fixed sample size. *Journal of Official Statistics* **21**, 543–570.
- QUALITÉ, L. & TILLÉ, Y. (2008). Estimation de la précision d'évolutions dans les enquêtes répétées, application à l'enquête suisse sur la valeur ajoutée. *Techniques d'enquête* **41**, 173–181.
- SAS Institute Inc. (2018). *SAS/STAT 15.2 User's guide*.

A.1 Schéma de rotation standard



A.2 Schéma de rotation pour les 75 ans et plus



B Variables auxiliaires disponibles

Indic_phone : indique si l'OFS a connaissance ou non d'un numéro de téléphone fixe (tel / altel)

Variables issues de STATPOP

AgeFin : classe d'âge définie en fonction de « agerev », plus grand degré de détail

- 1 = 15 ans
- 2 = 16 ans
- 3 = 17 ans
- 4 = 18 ans
- 5 = 19 ans
- 6 = 20 - 24 ans
- 7 = 25 - 29 ans
- 8 = 30 - 34 ans
- 9 = 35 - 39 ans
- 10 = 40 - 44 ans
- 11 = 45 - 49 ans
- 12 = 50 - 54 ans
- 13 = 55 - 59 ans
- 14 = 60 - 64 ans
- 15 = 65 - 69 ans
- 16 = 70 - 74 ans
- 17 = 75 - 79 ans
- 18 = 80 ans et plus

cl_Age : classe d'âge, niveau de détail intermédiaire

- 1 = moins de 19 ans
- 2 = 19-24 ans
- 3 = 25-29 ans
- 4 = 30-39 ans
- 5 = 50-54 ans
- 6 = 55-64 ans
- 7 = 65-79 ans
- 8 = 80 et plus

Age : classe d'âge, faible niveau de détail

- 1 = 15 - 24 ans
- 2 = 25 - 39 ans
- 3 = 40 - 54 ans
- 4 = 55 - 64 ans
- 5 = 65 ans et plus

AgeEuro : classe d'âge, niveau de détail minimal

- 1 = 15 - 24 ans
- 2 = 25 - 39 ans
- 3 = 40 - 54 ans
- 4 = 55 ans et plus

civil : état civil

- 1 = célibataire
- 2 = marié(e)
- 3 = veuf/veuve
- 4 = divorcé(e)

sex : genre

- 1 = hommes
- 2 = femmes

Nation : nationalité

- 1 = Suisse
- 2 = France
- 3 = Allemagne
- 4 = Italie
- 5 = Autriche
- 6 = Espagne
- 7 = Portugal
- 8 = Pays-Bas / Belgique / Luxembourg
- 9 = Irlande / Royaume-Uni
- 10 = Danemark / Finlande / Suède / Norvège / Islande
- 11 = Grèce / Chypre / Malte / Liechtenstein / Monaco / Vatican / Andorre /
St-Marin
- 12 = Estonie / Lettonie / Lituanie / Pologne / République Tchèque/ Slovaquie / Hongrie /
Slovénie
- 13 = Bulgarie / Roumanie
- 14 = Croatie
- 15 = Macédoine
- 16 = Bosnie
- 17 = Kosovo / Serbie / Monténégro / Albanie
- 18 = Turquie
- 19 = USA + Canada
- 20 = Reste Europe
- 21 = Afrique
- 22 = Reste Monde

Natio_cl4_SILC : nationalité, avec les mêmes regroupements que dans SILC

- 1 = Suisse
- 2 = Europe du Nord et de l'Ouest
- 3 = Europe du Sud
- 4 = Autres pays

Permis : type de permis de séjour

- 1 = Suisses
- 2 = Permis C
- 3 = Permis B
- 4 = Autres

duree : durée du séjour en Suisse

- 1 = Suisses
- 2 = Étrangers depuis 0 à 1 an en Suisse
- 3 = Étrangers depuis 2 à 4 ans en Suisse
- 4 = Étrangers depuis plus de 5 ans en Suisse

Origine : origine

- 1 = Étranger
- 2 = Suisses

Gros : grande région

Kanton : 24 cantons avec regroupement AR-AI et OW-NW

Typo_men : typologie du ménage, définie en fonction de la variable

« compBasicAg25ofPHHperm »

- 101 = Homme seul de 25 ans ou plus
- 102 = Femme seule de 25 ans ou plus
- 110 = Deux personnes de 25 ans ou plus de sexe différent
- 111 = Deux personnes de 25 ans ou plus de même sexe
- 120 = Trois personnes ou plus de 25 ans ou plus
- 201 = Une ou plusieurs personnes de moins de 25 ans (regroupement des modalités 201 et 202)
- 210 = Homme seul de 25 ans ou plus avec personne(s) de moins de 25 ans
- 211 = Femme seule de plus de 25 ans avec personne(s) de moins de 25 ans
- 220 = Deux personnes de 25 ans ou plus avec personne(s) de moins de 25 ans (regroupement des modalités 221 et 220)
- 230 = Trois personnes ou plus de 25 ans ou plus avec personne(s) de moins de 25 ans
- 8 = Valeur manquante

HHsize : taille du ménage en fonction de la variable « HouseholdsizeSRPH »

- 1 = une personne
- 2 = deux personnes
- 3 = trois personnes
- 4 = quatre personnes
- 5 = cinq personnes ou plus

Variables issues d'autres sources

gkats : type de domicile

- 1021 = Maison individuelle
- 1025 = Bâtiment exclusivement à usage d'habitation comportant plusieurs logements
- 1030 = Bâtiment d'habitation à usage annexe
- 1040 = Bâtiment partiellement à usage d'habitation
- 99 = pas disponible

taille_silc : surface du domicile en fonction du nombre de personnes, définie comme dans SILC

- 1 = moins de 20m² par personne
- 2 = 20m² et plus par personne
- 99 = pas disponible

ci_typact_1 : statut sur le marché du travail

- 99 = Pas encore présent dans la population de référence ou non concerné
- 2 = Actif occupé

ci_typact_2 : statut sur le marché du travail, plus détaillé

- 1 = Salarié
- 2 = Indépendant ou salarié et indépendant
- 10 = non concerné
- 11 = Pas encore présent dans la population de référence

ci_nact : nombre d'activités lucratives

- 1 = 1 activité lucrative
- 2 = 2 activités lucratives et plus
- 10 = non concerné
- 11 = Pas encore présent dans la population de référence

ci_rmprof_decile : déciles du revenu selon la ZAS

A noter que dans le cas où une personne a plusieurs revenus, ses revenus sont additionnés.

rr_pinv : rente d'invalidité

- 99 = Pas encore présent dans la population de référence ou non concerné
- 1 = Jusqu'à 100% d'invalidité

el_typepc : type de prestation complémentaire

- 99 = Pas encore présent dans la population de référence ou non concerné
- 1 = PC AVS
- 2 = PC AI

p_statorp_1 : inscription dans un organisme de placement

- 2 = Chômeurs inscrits et demandeurs d'emploi non-chômeurs
- 10 = non concerné

p_statorp_2 : inscription dans un organisme de placement

- 1 = Chômeur inscrit
- 2 = Autre demandeur d'emploi non-chômeur
- 10 = non concerné

duree_inscr : durée de l'inscription dans un organisme de placement

- 1 = moins de 12 mois
- 2 = 12 à moins de 24 mois
- 3 = 24 mois à moins de 5 ans
- 10 = non concerné

A noter qu'au-delà de 5 ans, le code -10 est attribué.

duree_afd : durée depuis l'arrivée en fn de droits

- 1 = jusqu'à moins de 5 ans
- 10 = non concerné

A noter qu'au-delà de 5 ans, le code -10 est attribué.

noga08_section : type d'activité économique de l'établissement

-11 = Pas encore présent dans la population de référence

-10 = non concerné

A = agriculture, sylviculture et pêche

B = B/C/D/E. industries extractives/industrie manufacturière/produ. et distr. d'électricité/produ. et distr. d'eau ; assainissement

F = construction

G = commerce ; réparation d'automobiles et de motocycles

H = transports et entreposage

I = hébergement et restauration

J = information et communication

K = activités financières et d'assurance

L = L+N. activités immobilières+activités de services administratifs et de soutien

M = activités spécialisées, scientifiques et techniques

O = O/U. administration publique/activités extra-territoriales

P = enseignement

Q = santé humaine et action sociale

R = R/S/T. arts, spectacles et activités récréatives/autres

rechtsform : forme juridique de l'établissement

-10 = non concerné

-11 = Pas encore présent dans la population de référence

01 = chef indépendant

02 = sociétés (chef salarié)

03 = entreprises publiques

04 = administrations publiques

99 = non attribué

C Taux de réponse

T 12 – Taux de réponse pour le troisième trimestre 2015

Strate de tirage	Vague 1	De vague 1 à vague 2	De vague 2 à vague 3	De vague 3 à vague 4
Global	0.44	0.93	0.91	0.96
AG	0.45	0.91	0.92	0.96
AI-AR	0.52	0.97	0.92	0.97
BE	0.48	0.93	0.93	0.97
BL	0.40	0.92	0.89	0.96
BS	0.38	0.89	0.88	0.97
FR	0.47	0.96	0.92	0.97
GE	0.41	0.90	0.89	0.95
GL	0.30	0.84	0.85	1.00
GR	0.37	0.91	0.90	0.96
JU	0.49	0.95	0.93	0.98
LU	0.46	0.93	0.92	0.95
NE	0.45	0.91	0.92	0.97
NW-OW	0.41	0.92	0.95	0.95
SG	0.44	0.94	0.90	0.94
SH	0.39	0.93	0.80	0.96
SO	0.41	0.90	0.91	0.95
SZ	0.46	0.94	0.90	0.97
TG	0.45	0.93	0.88	0.97
TI	0.48	0.95	0.93	0.94
UR	0.49	0.97	0.96	0.89
VD	0.46	0.92	0.90	0.95
VS	0.43	0.92	0.87	0.96
ZG	0.46	0.91	0.95	0.98
ZH00	0.44	0.93	0.90	0.96
ZH01	0.37	0.94	0.89	0.97

D Variables auxiliaires entrant dans la segmentation utilisée pour la correction de la non-réponse en vague 1

Indic_phone : indique si l'OFS a connaissance ou non d'un numéro de téléphone fixe (tel / altel)

Variables issues de STATPOP

Gros : grande région

Kanton : canton

sex : genre

cl_Age : classe d'âge définie en fonction de « agerev »

civil : état civil

Natio_cl4_SILC : nationalité, avec les mêmes regroupements que dans SILC

Permis : type de permis de séjour

duree : durée du séjour en Suisse

Typo_men : typologie du ménage

HHsize : taille du ménage

Variables issues d'autres sources

gkats : type de domicile

taille_silc : surface du domicile en fonction du nombre de personnes, définie comme dans SILC

ci_typact_1 : statut sur le marché du travail

ci_typact_2 : statut sur le marché du travail, plus détaillé

ci_nact : nombre d'activités lucratives

ci_rmprof_decile : déciles du revenu selon la ZAS

rr_pinv : rente d'invalidité

el_typepc : type de prestation complémentaire

p_statorp_1 : inscription dans un organisme de placement

p_statorp_2 : inscription dans un organisme de placement, plus détaillée

duree_inscr : durée de l'inscription dans un organisme de placement

duree_afd : Dauer seit Aussteuerung

noga08_section : type d'activité économique de l'entreprise (ou de l'établissement ?)

rechtsform : forme juridique de l'entreprise

La définition des modalités des variables énumérées ci-dessus se trouve dans l'annexe [B](#).

E Variables auxiliaires utilisées pour le calage

Variables issues de STATPOP

agefin
nation
duree
permis
civil*sexe
nation*sexe
kanton*sexe
sexe*origine*age
gros*sexe*origine
gros*age*origine
gros*age*sexe

Variables issues d'autres sources

P_statorp_2
rechtsform

Croisements entre les sources

CL_typact_1*kanton
CL_typact_2*sexe
CL_typact_2*origine
CL_typact_2*Age
CL_nact*sexe
CL_nact*origine
ci_rmprof_decile*sexe
ci_rmprof_decile*origine
RR_pinv*sexe
RR_pinv*origine
EL_typepc*sexe
P_statorp_1*sexe
P_statorp_1*origine
P_statorp1_*ageEuro
P_statorp_1*gros
DUREE_inscr*sexe
DUREE_inscr*origine
DUREE_afd*sexe
DUREE_afd*origine
nog08_section*sexe
nog08_section*origine

La définition des modalités des variables énumérées ci-dessus se trouve dans l'annexe [B](#).

F Arbre de segmentation de la vague 1 du troisième trimestre 2015

Indic_phone=18.00%	Natio_cl4_Sil=14.43%	Natio_cl4_Sil=12.75%	Gros7=0 (198) 11.90%	p_statorp_12=11.38%	Gros1=0 (1326) 10.41%	rechtsform2=0 (317) 10.09%	rechtsform2=1 (203) 19.70%
				p_statorp_12=1 (137) 18.98%			
			Gros7=1 (228) 20.18%				
		Natio_cl4_Sil=19.85%	civil2=0 (479) 16.28%	sex1=0 (202) 20.79%			
				sex1=1 (277) 13.00%			
			civil2=1 (206) 28.16%				
	Natio_cl4_Sil=22.70%	rechtsform4=21.08%	taille_silc2=0 (242) 12.40%	ci_rmprof_de=21.48%	ci_rmprof_de=20.77%	Gros4=0 (124) 18.99%	HHsize2=0 (783) 20.82%
			taille_silc2=1 (22.28%)				HHsize2=1 (360) 15.00%
						ci_rmprof_de=26.57%	HHsize5=1 (143)
					Gros4=1 (226) 26.11%		
				ci_rmprof_de=30.33%	ci_rmprof_de=1 (122)		
				ci_rmprof_de=33.93%	ci_rmprof_de=1 (112)		
		rechtsform4=1 (215) 37.67%					
Indic_phone=56.66%	gkats1021=0 53.17%	Natio_cl4_Sil=55.03%	Typo_men22=52.01%	cl_Age6=0 (3) 49.62%	ci_rmprof_de=45.56%	Typo_men21=44.14%	noga08_sect=45.01%
							el_typepc999=0 (130) 30.77%
							rechtsform4=45.74%
							Gros7=0 (22) 44.87%
							Natio_cl4_Sil=44.15%
							civil2=0 (131) 45.49%
							noga08_sect=43.14%
							cl_Age4=0 (9) 44.30%
							cl_Age4=1 (2) 46.29%
							noga08_sectionG=1 (136) 35.73%
							civil2=1 (583) 50.77%
							Natio_cl4_Sil=36.61%
							sex1=0 (136) 43.38%
							sex1=1 (200) 32.00%
							Gros7=1 (107) 59.81%
							rechtsform4=1 (207) 55.56%
							noga08_sectionF=1 (132) 26.52%
							Typo_men21=1 (310) 58.39%
							ci_rmprof_de=63.43%
							cl_Age6=1 (1) 58.25%
							cl_nact910=0 (421) 66.75%
							sex1=1 (376) 56.65%
							cl_nact910=1 (324) 49.07%
							cl_Age7=1 (1) 59.59%
							Natio_cl4_SilC3=0 (1239) 60.61%
							Natio_cl4_SilC3=1 (164) 51.83%
			Typo_men22=62.82%	Natio_cl4_Sil=65.55%	cl_Age2=0 (1) 67.13%	HHsize3=0 (1) 69.61%	cl_Age1=0 (8) 68.11%
							sex1=0 (411) 71.78%
							sex1=1 (423) 64.54%
							cl_Age1=1 (173) 76.88%
							HHsize3=1 (4) 62.12%
							noga08_sect=65.52%
							sex1=0 (190) 70.53%
							sex1=1 (187) 60.43%
							noga08_section910=1 (122) 51.64%
							cl_Age2=1 (2) 57.09%
							HHsize4=0 (169) 62.13%
							HHsize4=1 (113) 49.56%
							Natio_cl4_Sil=52.70%
							rechtsform2=0 (224) 59.82%
							rechtsform2=1 (258) 46.51%
		Natio_cl4_Sil=41.35%	Typo_men22=34.85%	taille_silc2=0 (184) 26.09%			
				taille_silc2=1 (539) 37.85%			
			Typo_men22=49.82%	HHsize3=0 (387) 52.71%			
				HHsize3=1 (167) 43.11%			
	gkats1021=1 65.22%	Typo_men22=61.22%	rechtsform4=59.90%	civil2=0 (811) 53.88%	noga08_section910=0 (429) 48.95%		
					noga08_section910=1 (382) 59.42%		
				civil2=1 (125) 63.78%	HHsize5=0 (1) 65.11%	cl_Age5=0 (9) 67.32%	Gros1=0 (807) 65.92%
							Gros1=1 (160) 74.38%
							cl_Age5=1 (188) 53.72%
							HHsize5=1 (104) 49.04%
							rechtsform4=1 (163) 77.91%
		Typo_men22=70.79%	Natio_cl4_Sil=61.85%	Natio_cl4_SilC2=0 (159) 55.97%			
				Natio_cl4_SilC2=1 (166) 67.47%			
			Natio_cl4_Sil=73.06%	rechtsform4=71.59%	ci_nact1=0 (5) 75.53%	cl_Age2=0 (418) 73.21%	
						cl_Age2=1 (101) 85.15%	
						ci_nact1=1 (5) 68.17%	civil2=0 (133) 57.89%
							civil2=1 (464) 71.12%
							rechtsform4=1 (161) 83.23%

Rapports de méthodes de la section méthodes statistiques de l'OFS
Methodenberichte der Sektion Statistische Methoden des BFS
Methodology reports published by the FSO's Statistical Methods Section

- Massiani, A. (2021). Enquête suisse sur la population active 2015-2020 : Révision de la pondération et de l'estimation de la variance. Numéro de commande : 338-0083
- Chevalier, C., Qualité, L. (2021). Enquête suisse sur la structure des salaires 2018. Révision de la pondération. Numéro de commande : 338-0082
- Qualité, L., Potterat, J. (2020). Enquête suisse sur la structure des salaires 2018. Révision du plan de sondage. Numéro de commande : 338-0081
- Potterat, J., Qualité, L., Assoulin, D. (2019). Strukturhebung der eidgenössischen Volkszählung. Stichprobenplan und -ziehung, Gewichtung, Schätzverfahren und Pooling 2010-2018. Bestellnummer: 338-0080
- Potterat, J., Assoulin, D., Nicoletti, J.-M. (2019). Beschäftigungsstatistik BESTA. Revision 2015: Stichprobenrahmen, Stichprobenplan und Hochrechnung. Bestellnummer: 338-0079
- Massiani, A. (2017). Estimation de la couverture du nouveau recensement en Suisse en 2012. Numéro de commande : 338-0078
- Nedyalkova, D., Assoulin, D. (2017). Construction of full-time equivalents for the Swiss structural business statistics. Order number: 338-0077
- Ferster, M. (2016). Energieverbrauchsstatistik EVS2014 - Stichprobe, Hochrechnung und Vergleichbarkeit mit der EVS2013. Bestellnummer: 338-0076
- Panchard, C. (2014). Enquête sur la participation aux activités culturelles 2008. Plan de sondage et estimations. Bestellnummer: 338-0073
- Assoulin, D. (2014). Wertschöpfungsstatistik. Revision 2009: Statistische Datenaufbereitung und Hochrechnung. Bestellnummer: 338-0072
- Wilhelm, M. (2014). Echantillonnage boule de neige. La méthode de sondage déterminé par les répondants. Numéro de commande: 338-0071
- Assoulin, D. (2013). Wertschöpfungsstatistik. Revision 2009: Stichprobenrahmen und Stichprobenplan. Bestellnummer: 338-0070
- Ferster, M. (2013). EVS I - Energieverbrauchsstatistik 2002 bis 2007: Stichprobenplan und Hochrechnung. Bestellnummer: 338-0069
- Assoulin, D. (2013). Zusatzerhebung für die landwirtschaftliche Betriebszählung 2010: Stichprobenplan und Hochrechnung. Bestellnummer: 338-0068
- Potterat, J. (2012). Use of conversion keys for NOGA2002-2008. Order number: 338-0067-05
- Andrade, B., Salamin P.-A. (2012). Enquête sur la situation sociale et économique des étudiant-e-s des hautes écoles suisses 2009. Cadre de sondage, plan d'échantillonnage et méthodes d'estimation. Numéro de commande: 338-0066
- Potterat, J., Panchard, C., Kilchmann, D. (2012). Umweltschutzausgaben der Unternehmen 2009 (UWSA2009). Stichprobenplan, Einsetzungen, Gewichtung und Schätzverfahren. Bestellnummer: 338-0065
- Potterat, J. (2012). Benutzung der Umsteigeschlüssel NOGA 2002-2008. Bestellnummer: 338-0064
- Potterat, J. (2011). Kosten und Nutzen der Berufsbildung aus Sicht der Betriebe im Jahr 2009 (KNBB09). Stichprobenplan, Gewichtung und Schätzverfahren. Bestellnummer: 338-0063
- Kilchmann, D., Potterat, J., Genoud, S. (2011). Gütertransporterhebung 2008. Stichprobenplan, Datenaufbereitung, Gewichtung und Schätzverfahren. Bestellnummer: 338-0062

- Graf, E. (2010). Enquête suisse sur la santé 2007. Plan d'échantillonnage, pondérations et analyses pondérées des données. Numéro de commande: 338-0061
- Eichenberger P., Hulliger B., Potterat J. (2010). Describing the Anticipated Accuracy of the Swiss Population Survey. Order number: 338-0060
- Graf, E. (2010). Étude empirique de l'attrition du Panel Suisse de Ménages : vers une caractérisation du profil du non-répondant. Numéro de commande: 338-0059
- Graf, E. (2009). Weightings of the Swiss Household Panel: SHP_I wave 9, SHP_II wave 4, SHP_I et SHP_II combined. Order number: 338-0058
- Graf, E. (2009). Pondérations du Panel Suisse de Ménages: PSM_I vague 9, PSM_II vague 4, PSM_I et PSM_II combinés. Numéro de commande: 338-0057-05
- Qualité, L., Tillé, Y. (2009). Estimation de la précision d'évolutions dans l'enquête sur la valeur ajoutée. Numéro de commande: 338-0056
- Renaud, A., Panchard, C. et Potterat, J. (2008). Statistique de l'emploi. Révision 2007 : méthodes d'estimation. Numéro de commande: 338-0055
- Graf, E. (2008). Pondérations du PSM. PSM_I vague 8, PSM_II vague 3, PSM_I et PSM_II combinés. Numéro de commande: 338-0054
- Andrade, B., Graf, M. (2008). Enquête suisse sur la structure des salaires 2006. Aspects méthodologiques du modèle des salaires "Salarium". Numéro de commande: 338-0053
- Renaud, A. (2008). Statistique de l'emploi. Révision 2007 : cadre de sondage et échantillonnage. Numéro de commande: 338-0052
- Graf, E. (2008). Pondérations du SILC pilote. SILC_I vague 2, SILC_II vague 1, SILC_I et SILC_II combinés. Numéro de commande: 338-0051
- Kilchmann, D. (2008). Statistik der sozialmedizinischen Institutionen 1999-2004 und Krankenhausstatistik 1999-2002. Einsetzungen für fehlende Daten. Bestellnummer: 338-0050
- Renaud, A. (2008). Technologies de l'information et de la communication. Estimations sur la base de la statistique de la valeur ajoutée. Numéro de commande: 338-0049
- Assoulin, D. (2007). Wertschöpfungsstatistik. Einsetzungsversuche für fehlende Antworten grosser Unternehmen. Bestellnummer: 338-0048
- Kilchmann, D. (2007). Beherbergungsstatistik Campingplätze. Stichprobenrahmen und Schätzverfahren 2005/06. Bestellnummer: 338-0047
- Gabler, S., Häder, S. (2007). Haushalts- und Personenerhebungen. Machbarkeit von Random Digit Dialing in der Schweiz. Bestellnummer: 338-0046
- Ferrez, J., Graf, M. (2007). Enquête suisse sur la structure des salaires. Programmes R pour l'intervalle de confiance de la médiane. Numéro de commande: 338-0045
- Renaud, A. (2007). Harmonisation de la scolarité obligatoire en Suisse (HarmoS). Design général de l'enquête et échantillon des écoles. Numéro de commande: 338-0044
- Potterat, J. (2007). Betriebszählung 2005. Statistische Methoden zur Schätzung der provisorischen Ergebnisse. Bestellnummer: 338-0043
- Hulliger, B. (2006). Umweltschutzausgaben der Unternehmen 2003, Stichprobenplan, Datenaufbereitung und Schätzverfahren. Bestellnummer: 338-0042
- Renfer, J.-P. (2006). Enquête sur les chiffres d'affaires du commerce de détail. Plan d'échantillonnage et méthodes d'estimation. Numéro de commande: 338-0041
- Salamin, P.-A. (2006). Statistique de l'aide sociale dans le domaine de l'asile. Plan de sondage et extrapolations pour l'enquête pilote 2005. Numéro de commande: 338-0040
- Renaud, A. (2006). Statistique suisse des bénéficiaires de l'aide sociale. Pondération des communes 2004. Numéro de commande: 338-0039

- Graf, M. (2006). Swiss Earnings Structure Survey 2002-2004. Compositional data in a stratified two-stage sample: Analysis and precision assessment of wage components. Order number: 338-0038
- Potterat, J. (2006). Pensionskassenstatistik 2004. Statistische Methoden zur Schätzung der provisorischen Ergebnisse. Bestellnummer: 338-0037
- Potterat, J. (2006). Kosten und Nutzen der Berufsbildung aus Sicht der Betriebe im Jahr 2004. Stichprobenplan, Gewichtung und Schätzverfahren. Bestellnummer: 338-0036
- Kilchmann, D. (2006). Vierteljährliche Wohnbaustatistik. Stichprobenplan, statistische Datenaufarbeitung und Schätzverfahren 2005. Bestellnummer: 338-0035
- Kilchmann, D. (2006). Erhebung über Forschung und Entwicklung in der schweizerischen Privatwirtschaft 2004. Bereinigung der Stichprobe, Ersatz fehlender Werte und Schätzverfahren. Bestellnummer: 338-0034
- Kilchmann, D., Eichenberger, P., Potterat, J. (2005). Volkszählung 2000. Statistische Einsetzungsverfahren Band 2. Bestellnummer: 338-0033
- Kilchmann, D., Eichenberger, P., Potterat, J. (2005). Volkszählung 2000. Statistische Einsetzungsverfahren Band 1. Bestellnummer: 338-0032
- Graf, M., Matei, A. (2005). Enquête suisse sur la structure des salaires 2002. La précision du salaire brut standardisé médian. Numéro de commande: 338-0031
- Graf, E., Renfer, J.-P. (2005). Enquête suisse sur la santé 2002. Plan d'échantillonnage, pondération et estimation de la précision. Numéro de commande: 338-0030
- Potterat, J. (2005). Mietpreis-Strukturerhebung 2003. Gewichtung und Schätzverfahren. Bestellnummer: 338-0029
- Potterat, J. (2005). Landwirtschaftliche Betriebszählung 2003. Schätzverfahren für die Zusatzerhebung. Bestellnummer: 338-0028
- Renaud, A. (2004). Coverage estimation for the Swiss population census 2000. Estimation methodology and results. Order number: 338-0027
- Kilchmann, D. (2004). Revision des Schweizerischen Lohnindex. Schätzmethoden der Lohnindices und deren Varianzschätzer. Bestellnummer: 338-0026
- Graf, M. (2004). Enquête suisse sur la structure des salaires 2002. Plan d'échantillonnage et extrapolation pour le secteur privé. Numéro de commande: 338-0025
- Renaud, A. (2004). Analyse de données d'enquêtes. Quelques méthodes et illustration avec des données de l'OFS. Numéro de commande 338-0024
- Renaud, A., Potterat, J. (2004). Estimation de la couverture du recensement de la population de l'an 2000. Echantillon pour l'estimation de la sous-couverture (P-sample) et qualité du cadre de sondage des bâtiments. Numéro de commande: 338-0023
- Graf, M. (2004). Fusion de données. Etude de faisabilité. Numéro de commande: 338-0022
- Potterat, J. (2003). Mietpreis-Strukturerhebung 2003. Entwicklung des Stichprobenplans und Ziehung der Stichprobe. Bestellnummer: 338-0021
- Potterat, J. (2003). Landwirtschaftliche Betriebszählung 2003. Stichprobenplan der Zusatzerhebung. Bestellnummer: 338-0020.
- Renaud, A. (2003). Estimation de la couverture du recensement de la population de l'an 2000. Echantillon pour l'estimation de la sur-couverture (E-sample). Numéro de commande: 338-0019
- Hulliger, B. (2003). Bereinigung der Stichprobe, Ersatz fehlender Werte und Schätzverfahren. Erhebung über F+E in der schweizerischen Privatwirtschaft 2000. Bestellnummer: 338-0018
- Renfer, J.-P. (2003). Enquête 2000 sur la recherche et le développement dans l'économie privée en Suisse. Plan d'échantillonnage. Numéro de commande: 338-0017
- Potterat, J. (2003). Kosten und Nutzen der Berufsbildung aus Sicht der Betriebe. Schätzverfahren. Bestellnummer: 338-0016

- Graf, M., Matei, A. (2003). Stratégie de choix des modèles de désaisonnalisation. Application aux séries de l'emploi total. Numéro de commande: 338-0015
- Potterat, J., Salamin, P.A. (2002). Betriebszählung 2001. Methoden für die Datenbereinigung. Bestellnummer: 338-0014
- Renaud, A. (2002). Programme international pour le suivi des acquis des élèves (PISA). Plans d'échantillonnage pour PISA 2000 en Suisse. Numéro de commande: 338-0013
- Renfer, J.-P. (2002). Enquête 2001 sur les coûts et l'utilité de la formation des apprentis du point de vue des établissements. Plan d'échantillonnage. Numéro de commande: 338-0012
- Potterat, J., Salamin, P.A. (2002). Betriebszählung 2001. Stichprobenplan und Schätzverfahren für die provisorischen Ergebnisse. Bestellnummer: 338-0011
- Graf, M. (2002). Enquête suisse sur la structure des salaires 2000. Plan d'échantillonnage, pondération et méthode d'estimation pour le secteur privé. Numéro de commande: 338-0010
- Renaud, A., Eichenberger P. (2002). Estimation de la couverture du recensement de la population de l'an 2000. Procédure d'enquête et plan d'échantillonnage de l'enquête de couverture. Numéro de commande: 338-0009
- Kilchmann, D., Hulliger, B. (2002). Stichprobenplan für die Obstbaumzählung 2001. Bestellnummer: 338-0008
- Graf, M. (2002). Passage du concept établissement au concept entreprise. Numéro de commande: 338-0007
- Salamin, P.A. (2001). La technique de la double enquête pour la statistique du transport routier de marchandise. Numéro de commande: 338-0006
- Peters, R., Renfer, J.-P. et Hulliger, B. (2001). Statistique de la valeur ajoutée 1997-1998. Procédure d'extrapolation des données. Numéro de commande: 338-0005
- Potterat, J., Hulliger, B. (2001). Schätzung der Sägereiproduktion mit der Sägerei-Erhebung PAUL. Bestellnummer: 338-0004
- Graf, M. (2001). Désaisonnalisation. Aspects méthodologiques et application à la statistique de l'emploi. Numéro de commande: 338-0003
- Hüsler, J., Müller, S. (2001). Schlussbericht Betriebszählung 1995 (BZ 95), Mehrfach imputierte Umsatzzahlen. Bestellnummer: 338-0002
- Renaud, A. (2001). Statistique suisse des bénéficiaires de l'aide sociale. Plan d'échantillonnage des communes. Numéro de commande: 338-0001
- Hulliger, B., Eichenberger, P. (2000). Stichprobenregister für Haushalterhebungen: Umstellung auf Telefonnummern ohne Namen und Adressen, Abläufe für Erstellung und Stichprobenziehung. Bestellnummer: 338-0000

Programme des publications de l'OFS

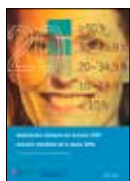
En tant que service statistique central de la Confédération, l'Office fédéral de la statistique (OFS) a pour tâche de rendre les informations statistiques accessibles à un large public. Il utilise plusieurs moyens et canaux pour diffuser ses informations statistiques par thème.

Les domaines statistiques

- 00 Bases statistiques et généralités
- 01 Population
- 02 Espace et environnement
- 03 Travail et rémunération
- 04 Économie nationale
- 05 Prix
- 06 Industrie et services
- 07 Agriculture et sylviculture
- 08 Énergie
- 09 Construction et logement
- 10 Tourisme
- 11 Mobilité et transports
- 12 Monnaie, banques, assurances
- 13 Sécurité sociale
- 14 Santé
- 15 Éducation et science
- 16 Culture, médias, société de l'information, sport
- 17 Politique
- 18 Administration et finances publiques
- 19 Criminalité et droit pénal
- 20 Situation économique et sociale de la population
- 21 Développement durable, disparités régionales et internationales

Les principales publications générales

L'Annuaire statistique de la Suisse



L'Annuaire statistique de la Suisse de l'OFS constitue depuis 1891 l'ouvrage de référence de la statistique suisse. Il englobe les principaux résultats statistiques concernant la population, la société, l'État, l'économie et l'environnement de la Suisse.

Le Mémento statistique de la Suisse



Le mémento statistique résume de manière concise et attrayante les principaux chiffres de l'année. Cette publication gratuite de 52 pages au format A6/5 est disponible en cinq langues (français, allemand, italien, romanche et anglais).

Le site Internet de l'OFS: www.statistique.ch

Le portail «Statistique suisse» est un outil moderne et attrayant vous permettant d'accéder aux informations statistiques actuelles. Nous attirons ci-après votre attention sur les offres les plus prisées.

La banque de données des publications pour des informations détaillées

Presque tous les documents publiés par l'OFS sont disponibles gratuitement sous forme électronique sur le portail Statistique suisse (www.statistique.ch). Pour obtenir des publications imprimées, vous pouvez passer commande par téléphone (058 463 60 60) ou par e-mail (order@bfs.admin.ch).
www.statistique.ch → Trouver des statistiques → Catalogues et banques de données → Publications

Vous souhaitez être parmi les premiers informés?



Abonnez-vous à un Newsmail et vous recevrez par e-mail des informations sur les résultats les plus récents et les activités actuelles concernant le thème de votre choix.
www.news-stat.admin.ch

STAT-TAB: la banque de données statistiques interactive



La banque de données statistiques interactive vous permet d'accéder simplement aux résultats statistiques dont vous avez besoin et de les télécharger dans différents formats.
www.stattab.bfs.admin.ch

Statatlas Suisse: la banque de données régionale avec ses cartes interactives



L'atlas statistique de la Suisse, qui compte plus de 4500 cartes, est un outil moderne donnant une vue d'ensemble des thématiques régionales traitées en Suisse dans les différents domaines de la statistique publique.
www.statatlas-suisse.admin.ch

Pour plus d'informations

Centre d'information statistique

058 463 60 11, info@bfs.admin.ch

L'enquête suisse sur la population active (ESPA) est une enquête complexe de par son plan d'échantillonnage basé sur un panel rotatif qui nécessite un traitement méthodologique approprié, ainsi que l'exploitation d'informations auxiliaires utiles pour modéliser la non-réponse initiale et les différentes phases d'érosion, mais également pour le calage.

La possibilité d'accéder à des sources de données administratives de plus en plus variées offre de nouvelles perspectives pour la pondération des enquêtes. Dans le cadre du travail présenté dans ce rapport, les gains de précision potentiels permis par les informations sociodémographiques contenues dans le cadre de sondage, mais également les données AVS, ou encore les données provenant des organismes de placement sur le marché du travail et des caisses de chômage, ont été testés et exploités pour la pondération de l'ESPA. De plus, une procédure d'estimation de la variance a également été développée, implémentée et testée, afin d'accompagner d'une mesure de précision les différents types d'exploitations diffusées.

Les analyses effectuées ont montré que les autres sources d'informations auxiliaires que le cadre de sondage ont un impact limité sur les performances de la modélisation du mécanisme de réponse. En revanche, l'impact de ces autres sources est beaucoup plus important si l'on s'intéresse au gain que l'on peut obtenir grâce au calage, qui est par ailleurs, toutes sources d'informations auxiliaires confondues, très appréciable pour certaines estimations. Au niveau national, il permet par exemple une réduction du coefficient de variation d'environ 50% pour l'estimation du nombre d'actifs et de 15% pour l'estimation du nombre de chômeurs.

En ligne

www.statistique.ch

Imprimés

www.statistique.ch

Office fédéral de la statistique

CH-2010 Neuchâtel

order@bfs.admin.ch

tél. 058 463 60 60

Numéro OFS

338-0083

ISBN

978-3-303-00688-7

La statistique
compte pour vous.

www.la-statistique-compte.ch