

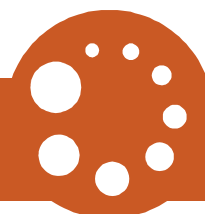


Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Département fédéral de l'intérieur DFI
Office fédéral de la statistique OFS

LABB : pondération pour la banque de données jointes

EXPERIMENTAL STATISTICS



Neuchâtel, 2021

Éditeur:	Office fédéral de la statistique (OFS)	Concept de mise en page:	Section DIAM
Renseignements:	eduperspectives@bfs.admin.ch, Tel. 058 483 93 66	Téléchargement:	www.experimental.bfs.admin.ch
Rédaction:	Section METH	Copyright:	OFS, Neuchâtel 2021 Reproduction autorisée seulement avec l'accord préalable des auteurs
Contenu:	Desislava Nedyalkova, Anne Massiani, Daniel Kilchmann, METH		
Domaine:	00 Bases statistiques des généralités		
Langue du texte original:	Français		
Traduction:	Services linguistiques de l'OFS		

LABB : pondération pour la banque de données jointes

Desislava Nedyalkova, Anne Massiani, Daniel Kilchmann (METH)

17.07.2020

1 Principe général

La population cible pour la banque de données jointes est un sous-groupe de la population résidante permanente en 2016. Elle comprend les personnes âgées de 19 à 64 ans, qui sont salariées dans une entreprise d'au moins trois employés, en dehors du secteur primaire, avec revenu strictement positif en décembre. On exclut de la population les apprentis, les multiactifs, les personnes avec des mouvements professionnels (y compris avec des variations de revenus entre les dates de référence - le mois d'octobre et le 31.12.2016) et les personnes payées à l'heure. Nous disposons de deux échantillons - un obtenu par appariement des données de la Centrale de compensation (CdC) avec le relevé structurel (RS) à l'aide du numéro AVS et l'autre en appariant encore avec l'échantillon de l'enquête suisse sur la structure des salaires (ESS) 2016 à l'aide du numéro AVS. Nous nous intéressons ici au deuxième échantillon, qui est appelé "banque de données jointes" dans ce qui suit.

Pour la pondération de la banque de données jointes, nous proposons de procéder selon l'approche "classique" consistant à appliquer un traitement de la non-réponse suivi d'un calage.

2 Informations disponibles

La section Système de formation (BILD-S) nous a livré trois fichiers qui comprennent la population et les deux échantillons cités plus haut. Dans cette section nous décrivons les informations dont on dispose pour chacun des trois fichiers.

1. fichier SPCDCSSTATENT2016 : population CdC pour l'année 2016 (N= 1 947 192), y compris les apprentis et les personnes payées à l'heure. Les apprentis ne peuvent pas être identifiés dans le fichier, ni les personnes payées à l'heure.

Filtrages :

- 19-64 ans (selon STATPOP),
 - salariés avec revenu strictement positif en décembre selon CdC,
 - sans les multi-actifs CdC,
 - sans les personnes avec des mouvements professionnels selon la CdC (y compris variations de revenus entre les deux dates de référence des enquêtes).
2. (a) fichier CDCRS_sample_f2a_c : échantillon CdC-RS 2016 (n= 67 152), y compris les apprentis et les personnes payées à l'heure. Une variable permet d'identifier les apprentis. Les personnes payées à l'heure ne peuvent pas être identifiées.

Filtrages supplémentaires :

- salariés RS (avec apprentis),
- temps de travail RS valide,
- élimination des revenus aberrants.

- (b) fichier CDCRS_sample_f2b_c : échantillon CdC-RS/ESS 2016 (n= 29 745), sans les apprentis mais avec les personnes pour lesquelles la variable iwaz prend la valeur 0 ce qui correspond approximativement aux personnes payées à l'heure.

Filtrages supplémentaires :

- sans les apprentis,
- sans les multi-actifs ESS.

La question des apprentis soulève une difficulté dans le cadre de la pondération de la banque de données jointes. Ils ne font en effet pas partie de la population cible de la banque de données jointes, mais il n'est pas possible de les repérer et de les éliminer du cadre de calage qui se base sur le fichier SPCDCSSTATENT2016. Le même type de problème se pose pour les personnes travaillant à l'heure. Une solution pour contourner cette difficulté est proposée dans la section suivante. La solution retenue pour les apprentis diffère de celle retenue pour les personnes payées à l'heure, du fait de la situation différente de ces deux caractéristiques dans les fichiers CDCRS_sample_f2a_c et CDCRS_sample_f2b_c, comme le montre le tableau ci-dessous.

Table 1 – Apprentis et personnes travaillant à l'heure

	Apprentis	Personnes payées à l'heure
Population cible	non inclus	non incluses
Fichier SPCDCSSTATENT2016	inclus, non identifiables	incluses, non identifiables
Fichier CDCRS_sample_f2a_c	inclus, identifiables	incluses, non identifiable
Fichier CDCRS_sample_f2b_c	non inclus	incluses, identifiable

Afin d'obtenir la variable NOGA secteur ainsi que d'autres variables de la STATENT qui pourraient être utiles pour la pondération, le fichier SPCDCSSTATENT2016 a été apparié avec le fichier de la STATENT 2016 (u_sta16_ins_20180719_appr) à l'aide du numéro identifiant de l'entreprise présent dans les deux fichiers.

3 Les étapes de la pondération

Nous décrivons dans cette section les différentes étapes de la pondération.

3.1 Suppression du secteur primaire et des entreprises de moins de trois employés

Le secteur primaire est inclus dans la CdC et le RS, mais pas dans l'ESS, et donc pas dans la banque de données jointes. Ceci implique la suppression du secteur primaire dans les trois fichiers, selon la variable "secteur" de la STATENT incluse dans le fichier SPCDCSSTATENT2016.

Les entreprises de moins de trois employés sont incluses dans la CdC et le RS, mais pas dans l'ESS, et donc pas dans la banque de données jointes. Nous avons supprimé ces entreprises dans les trois fichiers, selon la variable *emptot* (nombre total d'employés d'une entreprise) de la STATENT incluse dans le fichier SPCDCSSTATENT2016.

Les trois fichiers que nous obtenons après la suppression du secteur primaire et des entreprises de moins de trois employés, et qui vont nous servir comme des fichiers de départ pour la pondération, seront nommés *f1*, *f2a*, *f2b* correspondant aux fichiers de la section 2 après les suppressions supplémentaires mentionnées ci-dessus.

3.2 Calcul des poids corrigés pour la non-réponse

La deuxième étape consiste dans une correction de la non-réponse des poids de l'ESS. Du fait de l'indépendance entre les deux échantillons (RS et ESS), les poids de l'ESS (variable *gewicht*) sont multipliés par les poids du RS (variable *zpers_weight*) dans le fichier *f2b*. Ceci permet d'obtenir un poids "corrigé" pour la non-réponse inhérente aux deux enquêtes, qui est noté w_{NR} . Ce sont ici les poids finaux de l'ESS et du RS qui ont été utilisés, bien qu'idéalement les poids avant calage auraient été préférables.

3.3 Préparation des totaux de calage

Nous avons opté pour un calage en deux phases. Nous disposons comme variables auxiliaires, par exemple, du sexe, des classes d'âge, de la noga, de la taille de l'entreprise dans laquelle travaille la personne (par exemple regroupées en classes en fonction de la variable *emptot*). La première étape du calage consiste à caler les poids du fichier *f2a* (y compris les apprentis) sur les totaux des variables de calage calculés sur le fichier *f1*. A la fin de cette étape, nous obtenons un nouveau jeu de poids, nommés w_{CDCRS} , pour chaque unité du fichier *f2a* (y compris les apprentis). Ensuite, pour chaque modalité des variables de calage, nous calculons la somme des poids w_{CDCRS} sur le fichier *f2a* (sans les apprentis). Cela fournit une approximation des totaux de calage que l'on pourrait calculer sur le fichier *f1* si l'on était en mesure d'en exclure les apprentis (mais y compris les personnes travaillant à l'heure).

3.4 Calage final

La dernière étape de la pondération consiste en un calage des poids w_{NR} du fichier *f2b* (avec les *iwaz=0*) sur les totaux calculés à l'étape précédente. Cela fournit un nouveau jeu de poids nommé w_{cal} . Le fichier résultant de ce deuxième calage, après suppression des cas pour lesquels *iwaz=0* et sans modification des poids, constitue la banque de données jointes pondérée et a été

livré à la section BILD-S. L'exclusion des personnes payées à l'heure après le calage revient à considérer les analyses réalisées sur la banque de données jointes comme des estimations dans un domaine du fichier f2b.

4 Mise en œuvre pratique

Nous présentons dans cette section, pour chaque étape de la pondération, la mise en œuvre et les principaux résultats.

4.1 Suppression du secteur primaire et des entreprises de moins de trois employés

Le tableau 2 donne les pourcentages de données éliminées parmi les données livrées suite à la suppression du secteur primaire, des entreprises de moins de trois employés puis au total. Nous pouvons constater que dans chacun des cas un faible pourcentage des données est éliminé. Le tableau 3 donne les tailles des fichiers f1, f2a et f2b (après élimination).

Table 2 – Pourcentages d'élimination des données

Elimination des données (en pourcents)	fichier 1	fichier 2.a.	fichier 2.b.
seulement du secteur primaire	0.71%	0.42%	0.02%
seulement des entreprises de moins de trois employés	4.76%	3.38%	0.32%
des deux (total)	5.37%	3.75%	0.35%

Table 3 – Tailles des données de départ pour la pondération

	f1	f2a	f2b
taille	1 842 703	64 633	29 645

4.2 Calcul des poids corrigés pour la non-réponse

Après avoir calculé les poids w_{NR} , nous avons réalisé des analyses sur ces poids, notamment sur leur dispersion. Le tableau 4 donne une vue d'ensemble sur la dispersion des trois jeux de poids (les poids du RS, les poids de l'ESS et les poids w_{NR} qui sont le résultat de la multiplication des deux jeux de poids précédents). On peut constater que la distribution de ces nouveaux poids est en concordance avec les distributions des poids du RS et des poids de l'ESS suite à la multiplication des poids.

Table 4 – Distribution des poids *zpers_weight*, *gewicht* et *w_{NR}*

poids	min	q10	q50	q90	max
<i>zpers_weight</i>	12.81	16.26	21.18	36.40	65.39
<i>gewicht</i>	0.59	1.01	1.16	3.95	138.55
<i>w_{NR}</i>	8.84	17.94	34.55	100.45	4687.63

Nous avons également effectué un contrôle sur la somme des poids. En principe, la somme des poids *w_{NR}* devrait fournir une estimation de la taille de la population cible que l'on n'espère pas trop éloignée de la réalité. Le tableau 5 donne la taille de la population cible y compris les apprentis que l'on n'arrive pas à filtrer dans le fichier *f1b*, la somme arrondie des poids *w_{NR}* sur le fichier *f2b* et la somme arrondie des poids *zpers_weight* (avant calage) sur le fichier *f2a* sans les apprentis.

Table 5 – Somme des poids et taille de la population

Taille de la population cible	1 842 703
Somme des poids <i>w_{NR}</i>	1 618 701
Somme des poids <i>zpers_weight</i>	1 630 812

On constate que l'écart avec la population cible est assez grand, mais il faut tenir compte du fait que la population cible comprend les apprentis, tandis que ce n'est pas le cas du fichier *f2b*.

4.3 Préparation des totaux de calage et premier calage

Afin de déterminer quelles variables (ou croisements de variables) seront utilisées dans les deux calages successifs, nous avons étudié leurs effectifs sur l'échantillon *f2b*. Un seuil minimum de 50 observations a été fixé par cellule et le cas échéant, quand l'effectif d'une catégorie n'était pas suffisant, certains regroupements ont été nécessaires. Dans la suite nous décrivons les choix opérés.

Création de la variable classe de taille

Nous avons étudié la distribution de la variable *emptot* dans la population afin de déterminer une ou plusieurs catégorisations pour la variable classe de taille d'une entreprise. Une première catégorisation basée sur les quartiles nous permettant d'avoir quatre catégories à taille à peu près égale dans la population a été examinée. Le tableau 6 donne les effectifs dans le cadre et dans l'échantillon *f2b*.

Table 6 – La variable taille

taille de l'entreprise	valeur	effectifs cadre	effectifs échantillon f2b
[3; 25]	1	463 486	2 576
(25; 170]	2	458 969	4 842
(170; 1700]	3	455 911	8 787
≥ 1700	4	464 337	13 440

Une catégorisation plus fine a également été définie avec les catégories et effectifs suivants du tableau 7.

Table 7 – La variable taille fine

taille de l'entreprise	valeur	effectifs cadre	effectifs échantillons f2b
≤ 8	1	185 072	594
[8; 25]	2	278 414	1 982
(25; 100]	3	329 935	3 002
(100; 200]	4	168'982	2 493
(200; 500]	5	203 411	3 504
(500; 1 700]	6	212 552	4 630
(1 700; 10 000]	7	235 262	5 948
≥ 10 000	8	229 075	7 492

Création de la variable classe d'âge

Dans les données, nous disposons déjà de la variable ageclass que nous avons transformée en numérique avec les catégories suivantes :

- si ageclass est égale à "19-30 ans" alors la nouvelle variable cl_age= 1;
- si ageclass est égale à "31-40 ans" alors la nouvelle variable cl_age= 2;
- si ageclass est égale à "41-50 ans" alors la nouvelle variable cl_age= 3;
- si ageclass est égale à "51-64 ans" alors la nouvelle variable cl_age= 4;

Nous avons pris en compte le croisement de cette variable avec la variable sexe. Le tableau 8 donne les effectifs dans le cadre et dans l'échantillon f2b de cette variable croisée :

Table 8 – Les effectifs pour le sexe par classe d'âge

sexe	cl âge	effectifs cadre	effectifs échantillon f2b
1	1	240 688	2 730
1	2	265 672	4 518
1	3	255 098	4 552
1	4	258 332	4 580
2	1	206 770	2 678
2	2	195 035	3 401
2	3	208 744	3 587
2	4	212 364	3 599

Les différentes variables NOGA

Nous avons également analysé les effectifs dans le cadre de calage et dans l'échantillon de plusieurs variables NOGA, ainsi que de leur croisement avec la variable sexe.

- variable noga082 Statent (NOGA à deux positions)
Beaucoup trop de petits effectifs dans l'échantillon. Cette variable a été abandonnée.
- variable noga081 Statent (NOGA section : 18 catégories dans le cadre de calage)
L'effectif dans la NOGA section B est de 19, donc trop petit. Aussi, en croisant la variable noga081 Statent avec le sexe, nous avons constaté des effectifs nuls dans certains croisement noga*sexe. Cela a mené à un regroupement des sections NOGA B avec C et D avec E ce qui a permis de garder assez d'effectifs même en croisant avec le sexe.
- variable noga082 50 (NOGA OFS50 : 47 catégories dans le cadre de calage).
Cette variable, à catégorisations plus fines que la variable noga081 Statent, présente seulement un effectif plus petit que 50 dans la NOGA "05.9". Cette NOGA a été donc regroupée avec la NOGA "31.3" pour créer une catégorie à part.

Les variables retenues

Les variables que nous avons retenues pour le calage sont finalement :

- la NOGA OFS50 regroupée (nog50_cal_gr),
- le sexe croisé avec la classe d'âge (sexe*cl_age),
- le sexe croisé avec la taille de l'entreprise par classe à 8 catégories (sexe*taille_fine) et
- le sexe croisé avec la NOGA section regroupée (sexe*noga_s_gr).

Premier calage

Dans la Section 3.3 nous avons décrit la première étape du calage. A la fin de cette étape, nous avons obtenu un jeu de poids, nommés $w_{CD CRS}$, pour chaque unité du fichier f2a. Le tableau 9 montre la distribution des poids initiaux (zpers_weight) et des poids calés ($w_{CD CRS}$). Nous pouvons constater que les deux distributions sont assez semblables.

Table 9 – Distribution des poids *zpers_weight* et des poids calés *w_{CDCRS}*

poids	min	q10	q50	q90	max
<i>zpers_weight</i>	12.81	16.26	21.18	36.40	65.39
<i>w_{CDCRS}</i>	15.18	17.26	25.29	41.79	81.43

4.4 Calage final

Le calage des poids w_{NR} du fichier *f2b* (avec *iwaz*=0) sur les totaux calculés à l'étape précédente nous a fourni un nouveau jeu de poids (w_{cal}) pour 29 645 observations. Le tableau 10 montre la distribution des poids initiaux (w_{NR}) et des poids calés (w_{cal}). Nous pouvons constater que les deux distributions sont très proches et que les poids w_{cal} , tout comme les poids w_{NR} , sont assez dispersés. Cela provient essentiellement de la correction de la non-réponse et plus exactement de la façon dont l'échantillon *f2b* est obtenu - comme un appariement de deux échantillons (RS et ESS), dont les mécanismes de réponse sont indépendants, très différents, et se combinent.

poids	min	q10	q50	q90	max
w_{NR}	8.84	17.94	34.55	100.45	4687.63
w_{cal}	13.79	18.28	35.26	107.71	4629.00

Table 10 – Distribution des poids w_{NR} et w_{cal}

Après suppression des cas pour lesquels *iwaz*=0 (sans modification des poids) nous obtenons un jeu de poids pour 28 210 observations. C'est ce fichier et ces poids qui sont considérés comme les poids finaux et sont livrés à la section.

5 Exemple de calcul de variance

Nous avons calculé à titre d'illustration les médianes du revenu *mbls* (provenant de l'ESS) et du revenu standardisé *CdCrsref2*. Il s'agit de médianes pondérées sur base des poids de calage w_{cal} . Elles sont calculées, ainsi que leurs variances respectives, avec la procédure *proc surveymeans* de SAS en tenant compte, pour les deux revenus, de l'effet cluster de l'ESS (en utilisant la variable *entid_n* de l'ESS). Pour le calcul de la variance de la variable construite *CdCrsref2*, on doit également tenir compte de l'effet cluster du fait que nous ne gardons pas toutes les personnes du RS dans l'échantillon, mais seulement les quelques-unes que l'on retrouve aussi dans l'ESS. Ainsi l'effet cluster de l'ESS est transmis à toute la banque de données jointes. De plus, l'effet du calage n'est pas pris en compte dans le calcul de la variance. Le tableau 11 donne les valeurs estimées (arrondies) des médianes pondérées, l'erreur standard (SE) et les limites des intervalles de confiance estimés, notés, respectivement, ICL et ICU.

Il est à noter que les procédures que nous utilisons pour le calcul de la variance permettent d'estimer la variance due au plan d'échantillonnage, comme on le fait dans le cadre "classique" de l'OFS où les variables d'intérêt ne sont en général pas construites sur une base aléatoire. Toutefois

la variable construite CdCrsref2 (revenu standardisé) est par construction une variable aléatoire. Dans le calcul de la variance nous n'intégrons pas l'effet dû à la construction de ce revenu. L'intégration de cette part de variance supplémentaire n'a en effet pas pu se faire dans le délai imparti, mais il pourrait être intéressant dans une étape ultérieure du projet d'examiner la faisabilité d'intégrer dans les calculs de précision cette source de variabilité supplémentaire.

Table 11 – Médianes pondérées estimées

variable	médiane estimée	SE	ICL	ICU
MBLS	6726.07	56.56	6615.20	6836.94
CdCrsref2	6739.10	59.01	6623.42	6854.77

6 Limitations de la pondération

Ce travail a été fait dans le cadre des analyses LABB afin de permettre d'effectuer des analyses sur la banque de donnée jointes. Cette pondération a toutefois ses limites. Parmi celles-ci nous pouvons citer :

- La pondération ne permet qu'une extrapolation à sa population cible qui ne représente que 52 % de celle des banques de données disjointes.
- Les poids finaux sont très dispersés. Ceci pourrait s'avérer problématique pour la précision dans certaines sous-populations.
- Les poids utilisés dans la correction de la non-réponse sont les poids finaux de l'ESS et du RS, bien qu'idéalement les poids avant calage auraient été préférables.