



Plus haute formation achevée STATPOP

Processus de production – description courte

EXPERIMENTAL STATISTICS



Neuchâtel, 2025

Éditeur:	Office fédéral de la statistique (OFS)	Langue du texte original:	Français
Renseignements:	eduperspectives@bfs.admin.ch	Concept de mise en page:	Section PUB
Rédaction:	Sabrina Bravo, OFS ; Pascal Strubi, OFS	Téléchargement:	www.statistique.ch
Contenu:	Sabrina Bravo, OFS ; Pascal Strubi, OFS	Copyright:	OFS, Neuchâtel 2025
Domaine:	00 Bases statistiques et généralités		La reproduction est autorisée, sauf à des fins commerciales, si la source est mentionnée

Introduction

Le plus haut niveau de formation achevée de la population est un indicateur courant pour rendre compte du capital humain d'un pays en termes de qualifications. Son évaluation ainsi que son évolution sont particulièrement importantes dans un contexte de changements démographiques comme le vieillissement de la population, du maintien de la compétitivité économique et d'un besoin toujours plus accru de personnel qualifié. Afin de permettre des analyses multidimensionnelles il est nécessaire de disposer d'un niveau de formation pour chaque personne de la population résidente permanente. Il est alors nécessaire de mettre en place une procédure qui, en plus d'utiliser les données de statistiques et d'enquêtes de l'administration fédérale, complète les valeurs manquantes à l'aide d'un modèle d'imputations. Cette variable peut notamment être utilisée au niveau individuel lors d'appariements pour des analyses statistiques, mais ne peut pas être interprétée au niveau de la personne. La nomenclature du niveau de formation se décline en 3 ou 5 niveaux (Tableau 1).

Tableau 1 : Nomenclature du niveau de formation en 3 et 5 niveaux

Niveaux de formation		
3 niveaux	5 niveaux	Formations contenues dans la nomenclature à 5 niveaux
École obligatoire	1	École obligatoire (en cours ou achevée) / aucune / formation d'un an / offre de transition
Niveau secondaire II	2	Apprentissage en entreprise (CFC, AFP) / formation professionnelle élémentaire / école professionnelle / école de commerce
	3	Ecole de culture générale / maturité gymnasiale / brevet d'enseignement / maturité professionnelle ou spécialisée
Niveau tertiaire	4	Examen professionnel avec brevet fédéral / examen professionnel supérieur avec diplôme fédéral / maîtrise / école professionnelle supérieure
	5	Université / EPF / haute école spécialisée / haute école pédagogique

Sources de données

De nombreuses sources de données sont utilisées dans ce projet afin de déterminer annuellement le niveau de formation de la population résidente permanente : la statistique de la population et des ménages (STATPOP), les relevés structurels (RS) depuis 2010, le fichier des analyses longitudinales dans le domaine de la formation (LABB) augmenté par la statistique du personnel des écoles (SSP) et des hautes écoles (SHIS-PERS), la statistique des bénéficiaires de l'aide sociale (SAS), la statistique du chômage du Secrétariat d'État à l'économie (SECO), l'enquête suisse sur la structure des salaires (ESS) et la Swiss National Cohort (SNC) basé sur les recensements 1990 et 2000.

Préparation des données et imputation déterministe

Dans un premier temps, les données de chaque source sont plausibilisées (LABB sans plausibilisation sera noté LABB_0). D'une part les données potentiellement obsolètes sont supprimées : par exemple, si la personne a repris des études ou si la personne a quitté la Suisse à un âge où elle aurait pu reprendre des études ou encore si la personne annonce un niveau tertiaire alors qu'elle a moins de 18 ans. D'autre part, les données des sources sont corrigées à l'aide des données LABB si le niveau de formation dans LABB_0 est plus élevé ou si la personne est actuellement en cours d'étude et que son cursus depuis sa dernière formation achevée est connu dans LABB.

Une fois les sources plausibilisées, une priorisation de sources est appliquée. Tout d'abord les personnes de **moins de 15 ans** sont mises au niveau 1, sauf si LABB_0 indique un niveau plus élevé. Ensuite l'information de LABB est retenue pour les personnes qui sont en cours d'étude dans LABB et dont le cursus depuis la dernière formation achevée est connu, ainsi que les personnes ayant atteint un

niveau maximal dans LABB (ce sous-ensemble de LABB sera noté **LABB***). Ensuite le niveau de formation est défini selon des **concordances** de sources (si le niveau de formation le plus fréquent indiqué dans les sources a au moins deux sources de plus que le 2^{ème} niveau le plus fréquent). Puis les sources LABB et SECO qui sont issues de données administratives sont retenues : d'abord la formation selon le **SECO** et ensuite celle pas encore utilisée de LABB, donc sans LABB* (**LABB\LABB***), car le SECO contient également des formations obtenues à l'étranger. En dernier lieu, les sources SAS, RS et SNC dont le niveau de formation est renseigné par la personne sont utilisées : **SAS** en priorité, car la personne donne l'information à un interlocuteur qui peut l'aider à déterminer son niveau de formation, suivi du **RS**, et enfin de la **SNC** dont l'information est la plus ancienne.

Tableau 2 : Fréquence des règles d'attribution agrégées du niveau de formation pour 2020

Règles d'attribution	Fréquence	%	Cumul %
< 15 ans	1 306 271	15,1	15,1
LABB*	1 265 351	14,6	29,7
Concordances	2 852 961	32,9	62,6
SECO	556 972	6,4	69,0
LABB\LABB*	269 333	3,1	72,1
SAS	60 549	0,7	72,8
RS	851 471	9,8	82,6
SNC	700 775	8,1	90,7
Manquants	806 617	9,3	100,0
Total	8 670 300	100,0	100,0

Comme le montre le Tableau 2 pour l'année 2020, 90,7% de la population se voient attribuer leur niveau de formation à partir des différentes sources. Il reste alors 9,3% de la population dont le niveau de formation est imputé par un algorithme décrit ci-dessous.

Imputation basée sur un modèle

L'imputation du niveau de formation pour les personnes sans niveau de formation déjà attribué, se base sur le RS. Pour définir l'ensemble d'apprentissage du modèle d'imputation, nous appliquons la même procédure d'attribution au RS de l'année de référence que celle utilisée sur l'ensemble de la population (décrite ci-dessus), mais sans le RS de l'année de référence. L'ensemble des personnes du RS sans niveau de formation attribué à l'aide d'autres sources constitue notre ensemble d'apprentissage. Des variables issues de différents registres sont utilisées pour l'imputation, notamment des variables de la Centrale de compensation (p. ex. salaires), Statistique structurelle des entreprises (p. ex. NOGA), STATPOP Effectif (p. ex. sexe, nationalité, âge), STATPOP Ménage (p. ex. composition du ménage), Statistique des bâtiments et logements (p. ex. immeuble/maison individuelle), etc.

Des analyses ont montré que les niveaux de formation 2, 3 et 4 sont difficilement distinguables, c'est pourquoi la procédure d'imputation se fait en deux étapes : une première régression multinomiale fournit une estimation de la probabilité d'être de niveau 1, 234 ou 5 (noté P_A), suivie d'une deuxième régression multinomiale pour distinguer les niveaux 2, 3 et 4 (noté P_B). Ainsi pour une personne k dans la population, nous calculons sa probabilité d'être dans chaque niveau :

- $P_k(i) = P_{A,k}(i)$, pour les niveaux $i \in \{1,5\}$,
- $P_k(i) = P_{A,k}(234) \times P_{B,k}(i|234)$, pour les niveaux $i \in \{2,3,4\}$.

Ces probabilités sont ensuite plausibilisées : le niveau doit être supérieur ou égal au niveau de LABB_0, et il faut avoir plus de 18 ans pour être de niveau tertiaire. Une fois les probabilités estimées, un algorithme d'optimisation est appliqué pour imputer un niveau de formation à chaque personne en minimisant l'espérance de l'erreur (maximiser la vraisemblance), tout en respectant les totaux estimés par la somme des probabilités pour chaque niveau de formation sur des sous-populations selon l'âge croisé avec le sexe.

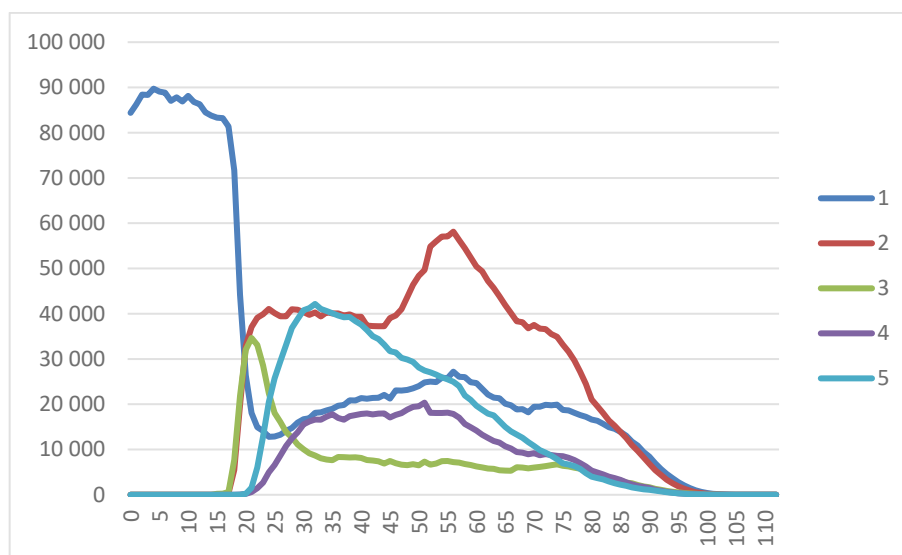
Résultats

Dans le Tableau 3 et le Graphique 1 se trouvent les niveaux de formation de la population résidente permanente pour l'année 2020.

Tableau 3 : Fréquence des niveaux de formation en 2020

Niveaux	0 ≤ Age < 115		Age > 14 ans		24 < Age < 65	
	Fréquence	%	Fréquence	%	Fréquence	%
1	3 065 093	35,4	1 758 825	23,9	848 699	17,6
2	2 690 992	31,0	2 690 992	36,5	1 778 942	36,8
3	638 282	7,4	638 279	8,7	327 617	6,8
4	823 104	9,5	823 104	11,2	642 447	13,3
5	1 452 829	16,8	1 452 829	19,7	1 238 443	25,6
Total	8 670 300	100,0	7 364 029	100,0	4 836 148	100,0

Graphique 1 : Distribution par âge des niveaux de formation en 2020



Conclusions

La variable du plus haut niveau de formation achevée ainsi obtenue est constituée d'environ 90% d'informations provenant des différentes sources et d'environ 10% d'imputations basées sur un modèle ajusté sur les données du RS. Cette variable sera produite annuellement et la méthodologie de sa construction peut encore évoluer. Ainsi, ce plus haut niveau de formation achevée est mis à disposition des utilisateurs pour des analyses multidimensionnelles, pour autant que les résultats impliquant cette variable soient clairement déclarés comme expérimentaux. Une documentation détaillée de la méthodologie est disponible sur le microsite des statistiques expérimentales de l'OFS.