



Plus haute formation achevée dans la statistique STATPOP+

Processus et méthode de production

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

Table des matières

1	Introduction	3
2	Objectifs du projet	4
3	Sources de données	5
4	Production de la variable	9
4.1	Préparation des données	10
4.1.1	Seuil pour les retours en formation	11
4.1.2	Reprise et plausibilisation de l'information des différentes sources de données	12
4.2	Priorisation des sources	13
4.2.1	Règles de priorisation pour la reprise des données dans les sources plausibilisées	14
4.2.2	Résultats de la reprise des données dans les sources plausibilisées	15
4.3	Imputations	18
4.3.1	Ensemble d'apprentissage	18
4.3.2	Algorithme d'imputations	19
4.3.2.1	Distinction des sous-populations à imputer	19
4.3.2.2	Application de régressions multinomiales	20
4.3.2.3	Imputations individuelles	21
4.3.3	Variables auxiliaires	24
4.4	Résultats	25
4.4.1	Résultats de la variable plus haute formation achevée pour l'année 2020	26
4.4.2	Evolution des niveaux de formation de 2019 à 2021	27
4.4.3	Performance de la méthode	28
4.4.4	Comparaison avec le relevé structurel	30
4.4.5	Diffusion	32
5	Perspectives	33
	Annexes	34
Annexe 1	LABB : détermination du niveau de la plus haute formation achevée	34
Annexe 2	Analyses de la qualité du relevé structurel	36
Annexe 3	Relevé structurel : la formulation de la question	40
Annexe 4	Comparaison de différentes imputations	42
Annexe 5	Variables auxiliaires pour les arbres de décision	46
Annexe 6	Analyse du relevé structurel pour les personnes en cours d'étude	49

Liste des abréviations

<i>Abréviation</i>	<i>Signification</i>
CdC	Centrale de compensation
ESS	Enquête suisse sur la structure des salaires
HEP	Haute école pédagogique
HES	Haute école spécialisée
HEU	Haute école universitaire
HHPERS	Fichier Ménages du relevé structurel
HSPERS	Personnel des hautes écoles
LABB	Analyses longitudinales dans le domaine de la formation
OFS	Office fédéral de la statistique
PLASTA	Système d'information en matière de placement et de statistique du marché du travail
RS	Relevé structurel de l'OFS (enquête par échantillonnage qui fait partie du nouveau système de recensement fédéral de la population)
SAS	Statistique des bénéficiaires de l'aide sociale
SECO	Secrétariat d'Etat à l'économie
SHAPE	Système de statistiques sur les personnes et les ménages
SNC	Swiss National Cohort (permet de lier les recensements 1990 et 2000 à STATPOP)
SSP	Personnel des écoles
StatBL	Statistique des bâtiments et logements
STATENT	Statistique des entreprises
STATPOP	Statistique de la population et des ménages
STATPOP+	STATPOP enrichie de nouvelles variables dont la plus haute formation achevée
ZPERS	Fichier Personnes cibles du relevé structurel

1 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 changement démographique comme le vieillissement de la population, du maintien de la compétitivité économique et d'un besoin toujours plus accru de personnel qualifié. Le niveau de formation joue un rôle important en termes de santé publique et détermine pour une bonne part les perspectives d'emploi ou de revenu des individus. Il est donc important d'être en mesure de fournir un niveau de formation fiable pour l'ensemble de la population suisse et d'en actualiser les valeurs chaque année.

Jusqu'en 2000, la variable « Plus haute formation achevée » était relevée dans les recensements exhaustifs de la population qui avaient lieu tous les dix ans. Avec la mise en place du nouveau système de recensement à partir de 2010, cette variable est disponible dans le relevé structurel (RS), une enquête annuelle réalisée sur un échantillon d'au moins 200 000 personnes. Le niveau de la plus haute formation achevée est demandé dans le RS non seulement aux personnes tirées dans l'échantillon, mais aussi à tous les autres membres de leur ménage.

L'introduction généralisée en 2011 du numéro AVS dans les différents registres de la Confédération permet, en exploitant l'ensemble des relevés de la formation en Suisse, de suivre le parcours des élèves et étudiants à travers l'ensemble du système de formation en Suisse et donc d'évaluer le plus haut niveau de formation achevée. Le couplage de l'ensemble des relevés de la formation disponibles à l'OFS (relevés des élèves de l'école obligatoire, des élèves et titres du degré secondaire II et des étu-

diants et titres du degré tertiaire) est effectué dans le cadre du programme LABB (analyses longitudinales dans le domaine de la formation). Le niveau de formation obtenu à partir des données LABB est quasi exhaustif pour les personnes jusqu'à 30 ans. Au-delà de 30 ans, une grande part des personnes échappent à LABB étant donnée la disponibilité des enregistrements de la formation limitée à 2011.

Afin de fournir un niveau de formation pour chaque personne de la population résidente permanente, il est alors nécessaire de mettre en place une méthode 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.

2 Objectifs du projet

L'objectif est de fournir un niveau de formation pour l'ensemble de la population suisse en se basant sur différentes sources de données disponibles à l'OFS et en complétant les données manquantes à l'aide d'imputations (voir chapitre 4.3). Grâce à ce projet, la variable du niveau de la plus haute formation achevée est maintenant disponible de manière exhaustive dans STATPOP. Elle permet de faire des analyses à un niveau plus fin que le RS et sur toutes les tranches d'âges, contrairement aux données actuelles de LABB pour lesquelles l'évaluation du niveau de formation est limitée aux personnes jusqu'à 30 ans environ. Il est par exemple possible de comparer le niveau de formation des personnes dans les différents quartiers des grandes villes ou encore de comparer les revenus de travail par niveau de formation à un niveau géographique fin. Elle permet également d'être utilisée au niveau individuel comme variable explicative pour des modèles statistiques et ainsi d'aider à mieux comprendre d'autres problématiques concernant différentes thématiques, par exemple des analyses sur la pauvreté. Les résultats des analyses obtenues avec cette variable doivent être interprétés en considérant l'incertitude liée aux imputations.

L'information sur le niveau de formation est relevée de manière différente entre les sources, via un questionnaire pour les enquêtes, à partir d'analyses longitudinales pour LABB ou sur la base de la présentation d'un certificat attestant de la réussite de la formation pour certains relevés. Le contenu de l'information (c'est-à-dire le nombre de modalités du niveau de formation et leurs désignations) diffère entre les sources, mais il est cependant possible de construire une nomenclature commune sur laquelle les différentes sources peuvent s'aligner. Pour ce faire, il est alors parfois nécessaire d'agréger certaines modalités. La nomenclature commune est celle du système SHAPE qui définit un niveau agrégé à 3 modalités :

1. école obligatoire,
2. niveau secondaire II,
3. niveau tertiaire.

Le système SHAPE définit également un niveau de formation plus détaillé à 5 niveaux. Cette nomenclature, également compatible avec les différentes sources utilisées, distingue les formations professionnelles des formations générales pour le niveau secondaire II et la formation professionnelle supérieure des hautes écoles pour le niveau tertiaire. Le *Tableau 1* ci-dessous présente le détail des différentes formations contenues dans chacun des niveaux de la nomenclature. Dans la suite de ce document, nous considérons uniquement le niveau de formation à 5 niveaux.

Tableau 1 : Niveau de formation à 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

3 Sources de données

De nombreuses sources de données sont utilisées afin de déterminer le niveau de formation de la population. Ces sources, présentées ci-dessous, sont de nature très différente et l'information sur le niveau de formation est soit directement disponible dans la source (pour la grande majorité des cas), soit obtenue en exploitant de manière longitudinale l'ensemble des enregistrements disponibles (données LABB). En plus des variables propres à chaque source de données, on souhaite disposer de variables sociodémographiques. Certaines sources contiennent déjà de telles variables, mais ces valeurs peuvent différer entre elles pour une même personne. Afin de garantir une base commune entre les différentes sources, on ne retient que les informations sociodémographiques issues de STATPOP (p. ex. l'année de naissance, le pays de naissance, la nationalité, le genre, l'âge d'arrivée en Suisse, la commune de domicile, le statut migratoire, etc.). Ces informations sont appariées aux différentes sources à l'aide du numéro AVS (NAVS13) qui est disponible de manière généralisée dans les relevés de l'administration fédérale depuis environ 10 ans. Avant de procéder au couplage des données, les NAVS13 sont mis à jour selon les données les plus récentes, puis pseudonymisés pour garantir la protection des données.

1. Statistique de la population et des ménages (STATPOP)

La statistique de la population et des ménages fait partie du système de recensement annuel de la population depuis 2010. Elle livre des informations concernant l'effectif et la structure de la population résidente à la fin d'une année ainsi que des mouvements de la population pendant l'année en cours.

Le projet STATPOP+ correspond à l'ensemble des variables actuellement disponibles dans STATPOP auxquelles de nouvelles variables sont ajoutées. Le niveau de la plus haute formation achevée pour l'ensemble de la population fait partie des nouvelles variables proposées.

La population résidente permanente de STATPOP est la population cible de ce projet.

2. Relevé structurel (RS)

Depuis 2010, le RS est réalisé dans le cadre du nouveau système de recensement annuel de la population. Il vise avant tout l'observation des structures socioéconomiques et socioculturelles de la population en Suisse. L'univers de base se compose des personnes âgées de 15 ans ou plus faisant partie de

la population résidante permanente vivant dans des ménages privés. Le relevé s'effectue par échantillonnage auprès de 200 000 personnes à l'aide d'un questionnaire en ligne (eCensus) ou d'un questionnaire papier.

Les données du RS sont composées d'une part d'informations recueillies sur la personne interrogée (ZPERS) et d'autre part d'informations sur les autres membres du ménage (HHPERS). Comme le niveau de formation figure parmi les questions posées dans le cadre du RS, nous disposons alors d'un niveau de formation pour chaque personne ayant répondu à un des questionnaires ainsi qu'à l'ensemble des membres de son ménage. Le nombre d'enregistrements avec un niveau de formation disponible en 2020 est de 301 482 pour ZPERS et 584 585 pour HHPERS.

Les variables retenues du RS sont les suivantes :

Vn :	identifiant de la personne
hh_over_niv5 :	niveau de formation à 5 modalités (HHPERS)
hh_year :	dernière année dans le RS (HHPERS)
zp_over_niv5 :	niveau de formation à 5 modalités (ZPERS)
zp_year :	dernière année dans le RS (ZPERS)

3. Analyses longitudinales dans le domaine de la formation (LABB)

Les données LABB (analyses longitudinales dans le domaine de la formation) résultent du couplage, de l'harmonisation et d'un traitement longitudinal de sources de données relatives au système de formation en Suisse (statistique des élèves SdL, statistiques des titres SBA, statistique de la formation professionnelle initiale SBG, statistique des étudiants et examens des hautes écoles SIUS), ainsi que du relevé structurel (RS), de la statistique de la population et des ménages (STATPOP), des données des comptes individuels de la Centrale de compensation (CdC) et du Système d'information en matière de placement et de statistique du marché du travail (PLASTA). Les données LABB contiennent ainsi l'ensemble des enregistrements de la formation en Suisse depuis 2011 (de l'école obligatoire jusqu'aux formations du tertiaire qui mènent à un titre reconnu dans le système de formation suisse). Le plus haut niveau de formation est obtenu en exploitant le parcours des élèves/étudiants et des titres obtenus dans l'ensemble du système de formation en Suisse depuis 2011. Les données LABB couvrent donc très bien les personnes jusqu'à 30 ans environ, mais ne sont pas suffisantes pour évaluer le niveau de formation des personnes plus âgées.

Les enregistrements de LABB sont de 2 types :

- Les enregistrements de type « élève » qui indiquent notamment l'année de formation, le type de formation suivie (p. ex. école obligatoire, CFC ou haute école universitaire) et le domaine de formation (p. ex. CFC en informatique ou formation HEU en médecine humaine).
- Les enregistrements de type « titre » qui indiquent notamment l'année d'obtention du titre, le type de formation achevée et le domaine de formation achevée.

Pour déterminer le plus haut niveau de formation achevée d'une personne dans LABB, on retient le niveau de formation le plus élevé parmi l'ensemble des niveaux associés aux différents enregistrements « élève » et « titre ». Les enregistrements de type « élève » ne donnent pas, contrairement aux enregistrements de type « titre », une indication directe du niveau de la plus haute formation achevée au moment de l'enregistrement. On impute alors dans ce cas le niveau de formation sur la base du titre précédent le plus probable donnant accès aux études en cours. On considère ici que ce titre a été obtenu avant 2011 et qu'il n'a donc pas pu être relevé dans LABB. Par exemple, pour une personne en cours d'études dans une HES en 2011, on suppose qu'elle a obtenu au préalable une maturité professionnelle (on impute alors le niveau de formation associé à cet enregistrement à 3).

Une description détaillée de la procédure mise en place pour déterminer le plus haut niveau de formation sur la base des données LABB est décrite dans l'Annexe 1.

Ajout des variables sociodémographiques de STATPOP et plausibilisation des NAVS13

L'ajout des variables sociodémographiques dans LABB se fait par un couplage avec STATPOP sur le NAVS13 pseudonymisé (identificateur unique). Un algorithme de plausibilisation est mis en place dans cette phase d'appariement pour valider le NAVS13 de LABB, on écarte par exemple les personnes pour lesquelles la différence d'âge entre la valeur de STATPOP et celle dans les relevés de la formation excède 5 ans. On limite également les âges en fonction du degré suivi par l'élève (dans le relevé des hautes écoles, les étudiants ont un âge minimal de 12 ans, dans les relevés du degré secondaire II l'âge minimal est de 10 ans). En 2020, ces restrictions excluent 113 enregistrements du fichier LABB (sur 18 241 461 enregistrements), dont 3 au degré secondaire II et 98 au degré tertiaire. Il n'est pas exclu au final qu'une personne de LABB ne corresponde pas à celle de STATPOP (p. ex. elle pourrait correspondre à un autre membre du ménage), mais ces cas sont très rares et très difficilement quantifiables. Le nombre d'enregistrements LABB avec un niveau de formation disponible en 2020 est de 1 743 915.

Les variables retenues dans LABB sont les suivantes :

Vn :	identifiant de la personne
CountMaxYearLabb :	nb d'années consécutives en formation jusqu'à maxYearofLABBMeasure
FormMaxYearLABB :	niveau correspondant à la dernière formation suivie dans LABB
ID_eo :	indique si la personne a eu un enregistrement à l'école obligatoire
lastyearEO :	année du dernier enregistrement à l'école obligatoire
educyearLastEO :	indique la dernière année d'enseignement à l'école obligatoire suivie
eductypeLastEO :	distingue le degré primaire 1-2, 3-8 et le secondaire I
MaxYearofLABBMeasure :	année du dernier enregistrement dans LABB
Over_niv5 :	plus haut niveau de formation à 5 modalités
YearHighestTitle :	année du plus haut titre obtenu ou imputé
Imputed :	indique si le niveau de formation est imputé ou non

4. La statistique du personnel des écoles (SSP) et des hautes écoles (HSPERS)

La statistique du personnel des écoles (SSP) comprend le personnel des écoles publiques et privées de la scolarité obligatoire à la formation professionnelle supérieure. La statistique du personnel des hautes écoles (HSPERS) recense toutes les personnes qui ont travaillé dans une haute école suisse au cours d'une année donnée. Une partie du personnel de ces statistiques appartient à des catégories qui demandent d'avoir obtenu un titre d'une haute école. Ces catégories concernent par exemple les enseignants, les collaborateurs scientifiques ou les doctorants. On impute alors à ces personnes un niveau de formation maximal (niveau tertiaire).

Ces relevés, du moins une partie d'entre eux, permettent d'élargir l'attribution d'un niveau de formation pour les personnes actives. Le nombre d'enregistrements avec un niveau de formation disponible en 2020 est de 9 629 pour SSP et 10 673 pour HSPERS.

Les variables retenues dans SSP et HSPERS sont les suivantes :

Vn :	identifiant de la personne
e_year :	année du relevé
Over_niv5 :	plus haut niveau de formation à 5 modalités (toujours égal à 5)

5. Statistique des bénéficiaires de l'aide sociale (SAS)

La statistique des bénéficiaires de l'aide sociale, élaborée et mise sur pied par l'Office fédéral de la statistique, recueille des informations permettant entre autres de déterminer le nombre de bénéficiaires de l'aide sociale, le type de prestations sous condition de ressources octroyées, la durée de l'octroi des prestations et la structure familiale des bénéficiaires. Cette statistique comble une lacune dans l'information qui existait auparavant en proposant des résultats détaillés au niveau national, cantonal et régional.

La variable « niveau d'études le plus élevé » est relevée dans SAS pour les personnes de plus de 15 ans qui touchent des prestations dans les catégories suivantes :

- Soutien régulier avec ou sans convention d'objectifs
- Aide sociale aux réfugiés
- Aide sociale aux demandeurs d'asile

Le nombre d'enregistrements avec un niveau de formation disponible en 2020 est de 200 814.

Les variables retenues dans SAS sont les suivantes :

Vn :	identifiant de la personne
year :	année du relevé
Over_niv5 :	plus haut niveau de formation à 5 modalités

6. Statistique du chômage (données du SECO)

Une statistique du chômage existe en Suisse depuis 1936. Elle est menée aujourd'hui par le Secrétariat d'État à l'économie (SECO). Cette statistique enregistre les effectifs, ainsi que les entrées et sorties, des personnes inscrites au chômage à la fin de chaque mois. Depuis 1997, la statistique du chômage est complétée par la publication du nombre de demandeurs d'emploi inscrits.

Les critères suivants sont pris en compte dans la statistique du chômage : sexe, âge, domicile, nationalité/type de permis de séjour, statut sur le plan du chômage (chômeur partiel/chômeur complet), durée du chômage, situation professionnelle antérieure, branche économique, groupe de professions, dernière fonction exercée.

Le nombre d'enregistrements avec un niveau de formation disponible en 2020 est de 458 608.

Les variables retenues dans cette statistique sont les suivantes :

Vn :	identifiant de la personne
year :	année du relevé
Over_niv5 :	plus haut niveau de formation à 5 modalités

7. Enquête suisse sur la structure des salaires (ESS)

L'enquête suisse sur la structure des salaires (ESS) est un sondage réalisé tous les deux ans auprès des entreprises privées et publiques et des administrations en Suisse. Elle permet de décrire régulièrement, à partir de données représentatives, la structure des salaires dans l'ensemble des branches économiques des secteurs secondaire et tertiaire. Elle s'intéresse non seulement à la branche économique et à la taille de l'entreprise concernée, mais aussi aux caractéristiques individuelles des salariés et des postes de travail. Le niveau de la plus haute formation achevée des salariés est indiqué dans l'ESS par

l'employeur et le nombre d'enregistrements avec un niveau de formation disponible en 2020 est de 1 835 297.

Les variables retenues dans ESS sont les suivantes :

Vn :	identifiant de la personne
year :	année du relevé
Over_niv5 :	plus haut niveau de formation à 5 modalités

8. Swiss National Cohort (SNC)

La Swiss National Cohort (SNC) est une cohorte établie sur le long terme et basée sur la population. Elle constitue une vaste plateforme de recherche polyvalente. Sa version actuelle se fonde sur les données des recensements de 1990 et 2000 qui ont été appariées aux registres des décès, des naissances vivantes et des migrations jusqu'en 2019, et appariées à STATPOP et au RS à partir de 2010. La SNC permet toutes sortes de recherches dans le domaine de la santé publique, en particulier en combinaison avec d'autres données longitudinales ou environnementales.

SNC 1990 et SNC 2000 contiennent, entre autres, l'information sur le niveau de formation. Étant donné l'ancienneté de ces relevés, le niveau de formation pour une partie de ces personnes (les plus jeunes) a encore pu progresser jusqu'en 2020. Une analyse a alors été effectuée pour déterminer un seuil d'âge à partir duquel le niveau de formation ne devrait plus (ou presque plus) évoluer les années suivantes (chapitre 4.1.1). Seules les personnes au-delà des seuils déterminés seront retenues (les niveaux de formations de ces personnes ne devraient donc pas avoir changés). Cette source reste cependant importante, car elle contient toutes les personnes des recensements 1990 et 2000 et le nombre d'enregistrements avec un niveau de formation disponible en 2000 est de 3 792 902.

Les variables retenues dans SNC sont les suivantes :

Vn :	identifiant de la personne
year :	année du relevé (1990 ou 2000)
Over_niv5 :	plus haut niveau de formation à 5 modalités

9. Autres sources

Dans ce projet d'autres sources sont utilisées pour les variables auxiliaires dans le modèle d'imputation (chapitre 4.3.3) :

- STATENT (Statistique structurelle des entreprises),
- CdC (Centrale de Compensation),
- StatBL (Statistique des bâtiments et logements),
- STATPOP (effectif, ménage et typologie des ménages),
- Niveaux géographiques de la Suisse.

4 Production de la variable

Dans un premier temps, le niveau de formation le plus élevé est extrait des données collectées, en fonction d'une hiérarchisation des sources. Ensuite, pour les personnes dont le niveau de formation reste inconnu, des imputations basées sur le RS sont effectuées. La production du niveau de formation a été faite selon la même procédure pour les années 2019, 2020 et 2021. Dans la suite, nous utiliserons le terme « année de référence » pour traiter de manière générale la production de cette variable sans devoir se restreindre à une année particulière.

Le niveau de la plus haute formation achevée utilise dans un premier temps l'information disponible dans les différentes sources collectées (description au chapitre 3). Ces sources sont d'abord plausibilisées : on décide quelles informations sur le niveau de formation peuvent être gardées ou si elles nécessitent un ajustement. Elles sont également hiérarchisées : on classe les sources selon le niveau de confiance qu'on leur attribue, par exemple certaines sources relèvent le niveau de formation sur la base d'un diplôme ou d'un certificat de fin de formation, d'autres sont relevées sur la base de réponses non vérifiées. Si une personne apparaît dans plusieurs sources, on attribue alors un niveau de formation selon la procédure décrite au chapitre 4.2 (priorisation des sources). Un modèle d'imputations est finalement mis en place pour compléter les valeurs du niveau de formation pour les personnes de STATPOP dont aucun niveau de formation n'a pu être évalué pour l'année de référence sur la base des sources disponibles (chapitre 4.3). La suite de cette section décrit en détail les différentes étapes de production.

4.1 Préparation des données

Dans un premier temps, chaque source de données est plausibilisée. En effet, certaines sources peuvent contenir des informations obsolètes, remontant parfois à plusieurs années. Par exemple, une partie des personnes en cours de formation lors du recensement de 2000 ont depuis accru leur niveau de formation. Il est donc nécessaire de ne pas considérer toutes ces données comme représentant leur plus haut niveau de formation achevée.

Depuis le 31 décembre 2010, la population suisse est recensée annuellement au niveau individuel, ce qui permet de suivre, à travers STATPOP, les personnes arrivées ou parties de Suisse à partir de 2011. De même, les informations sur les personnes en formation ou ayant obtenu un diplôme en Suisse sont systématiquement enregistrées dans LABB depuis 2011. C'est pourquoi, dans les règles de plausibilisation du chapitre 4.1.2 ci-dessous, une distinction est faite entre les données à partir de 2011 et les données précédant 2011. À partir de 2011, il est possible de suivre les reprises d'études et les départs de Suisse, ce qui n'est pas réalisable pour les données plus anciennes. On considère qu'avant 2011, une personne aurait pu reprendre des études (en Suisse ou à l'étranger) qui n'ont pas pu être identifiées.

La section suivante définit des seuils d'âge à partir desquels l'information sur le niveau de formation peut potentiellement être conservée. Ces seuils sont obtenus à l'aide d'analyses longitudinales effectuées sur les données de LABB et du RS.

Dans cette section, nous séparons LABB en deux, avec d'un côté LABB* et de l'autre LABB sans LABB* (LABB\LABB*). LABB* (les valeurs les plus sûres de LABB) contient d'une part les personnes qui sont de niveau maximal (niveau 5 dans LABB) et d'autre part les personnes qui sont en cours d'étude durant l'année de référence et dont les « antécédents de formation » sont connus dans LABB, c'est-à-dire si l'une des trois conditions suivantes est remplie :

1. La personne a atteint un niveau 2 ou plus dans LABB et son dernier titre est connu (non imputé dans LABB).
2. La personne est élève à l'école obligatoire (niveau 1 dans LABB).
3. La personne était à l'école obligatoire et a poursuivi sans interruption des études au post-obligatoire, mais n'a pas encore obtenu de titre du degré secondaire II (niveau 1 dans LABB).

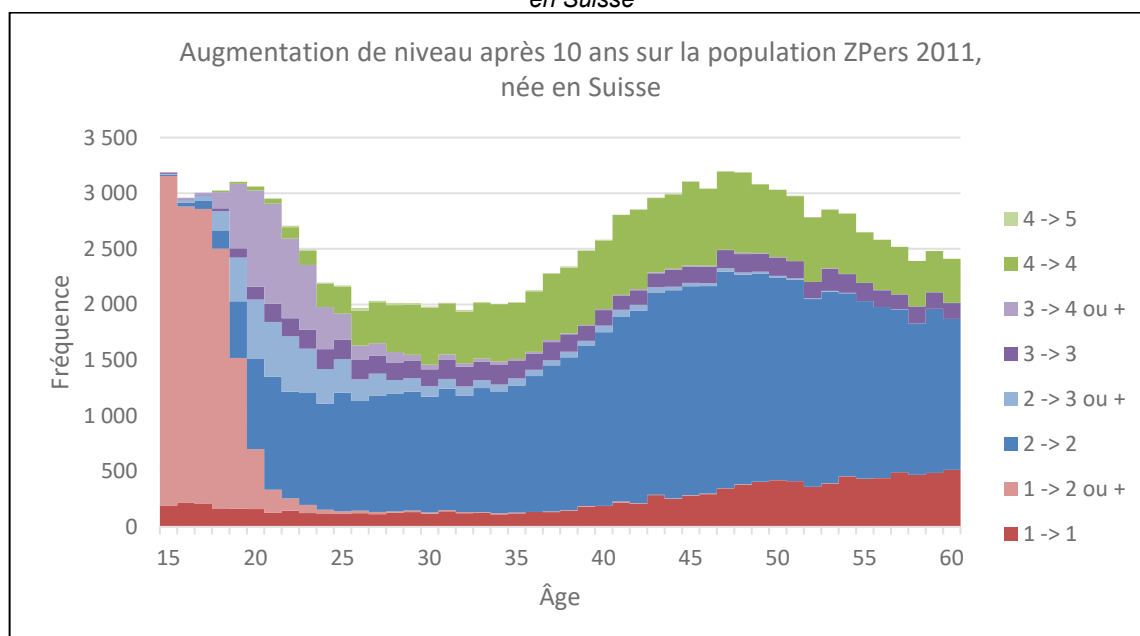
LABB\LABB* contient les personnes qui ne sont plus en formation dans LABB ou dont LABB ne connaît pas les « antécédents de formation ».

4.1.1 Seuil pour les retours en formation

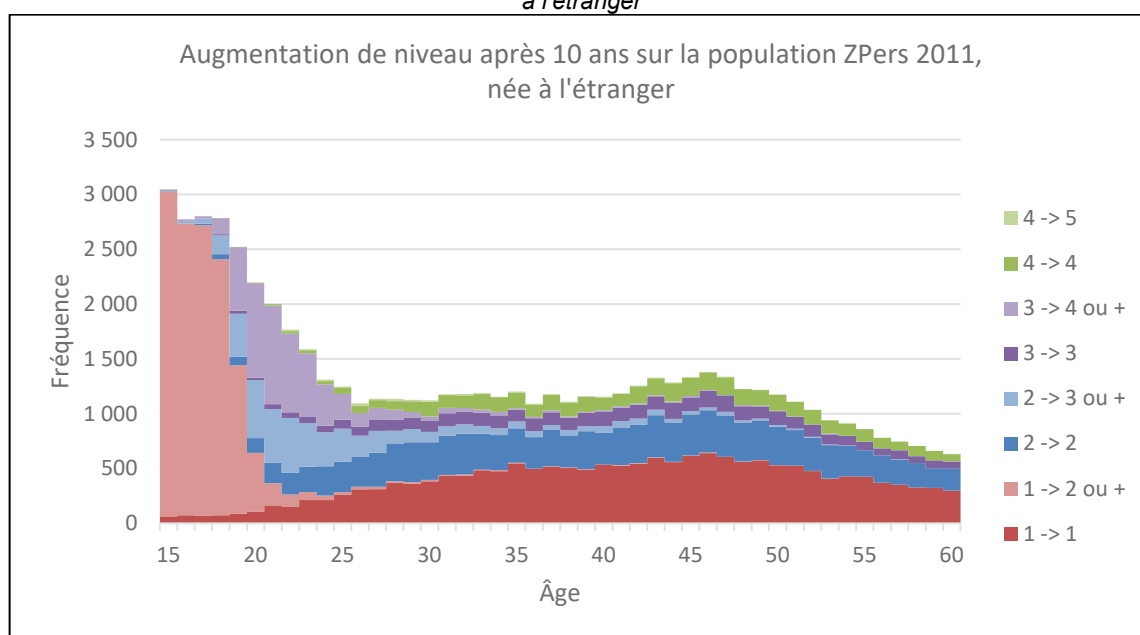
Dans l'analyse ci-dessous, nous définissons des seuils à partir desquels il est supposé qu'une personne ne reprendra plus de formation. Ces seuils prennent en compte l'âge et le lieu de naissance (en Suisse ou à l'étranger). Ils sont utilisés ensuite pour déterminer si l'information est reprise des sources ou non.

On considère ici l'ensemble des personnes du RS en 2011 qui n'ont pas quitté la Suisse jusqu'en 2021 et on analyse l'évolution du niveau de formation sur la période 2011-2021. Le niveau de formation de départ est celui du RS en 2011 et on regarde si le niveau de formation évolue sur la base des enregistrements successifs dans LABB. Si une personne n'a aucun enregistrement dans LABB, elle ne s'est pas formée en Suisse après 2011 et son niveau de formation reste celui du RS en 2011. Si une personne a un niveau de formation « maturité gymnasiale » dans le RS en 2011 et obtient un titre d'une haute école entre 2011 et 2021, son niveau de formation aura donc évolué, passant de 3 à 5. Le *Graphique 1* et le *Graphique 2* illustrent les augmentations de niveau par âge et par niveau.

Graphique 1 : évolution des niveaux de formation sur 10 ans pour la population du relevé structurel de 2011, née en Suisse



Graphique 2 : évolution des niveaux de formation sur 10 ans pour la population du relevé structurel de 2011, née à l'étranger



Les seuils d'âges sont définis pour chaque niveau de formation afin que moins de 2 % des personnes voient leur niveau de formation augmenter après ce seuil. À partir de ces seuils d'âge, il est supposé que les personnes aient peu de chance d'augmenter leur niveau de formation achevée. Même s'il n'est pas possible de quantifier ce pourcentage à l'ensemble de la population, grâce à cette analyse, on suppose que les informations disponibles sont suffisamment fiables.

Tableau 2 : Seuils des âges pour les retours en formation

	Niveau de formation				
	1	2	3	4	5
Né en Suisse	26 ans	32 ans	47 ans	-	-
Né à l'étranger	26 ans	30 ans	34 ans	-	-

Moins de 2% des personnes qui ont atteint le niveau 4 passent au niveau 5. Le niveau 5 étant le plus haut niveau, il n'y a donc pas de seuil d'âge pour ces deux niveaux.

4.1.2 Reprise et plausibilisation de l'information des différentes sources de données

Les règles ci-dessous déterminent les conditions pour lesquelles l'information des sources est retenue ou écartée, afin d'éviter des valeurs qui ne sont plus d'actualité ou peu plausibles.

On **écarte** l'information de la source de données si :

- La personne a moins de 18 ans à la date de référence de la donnée et la donnée indique un niveau tertiaire (niveau 4 ou 5). On suppose que l'information est non plausible, car la personne est trop jeune pour obtenir un titre du degré tertiaire.

Si l'information date **de 2011 ou est plus récente**, elle est **gardée** si :

- Il n'existe pas d'enregistrement plus récent dans LABB indiquant une reprise d'études (c'est-à-dire que le dernier diplôme obtenu indiqué dans LABB est connu et cohérent), et

- La personne n'a pas quitté la Suisse¹ depuis la date de référence de la donnée ou si la personne a quitté la Suisse depuis la date de référence de la donnée à un âge au-delà des seuils indiqués dans le *Tableau 2* ci-dessus (qui tiennent compte de l'âge **au moment de quitter la Suisse**, du niveau de formation annoncé dans la source et du lieu de naissance).

Si l'information date d'**avant 2011**, elle est **gardée** si :

- Il n'existe pas d'enregistrement plus récent dans LABB indiquant une reprise d'études (c'est-à-dire que le dernier diplôme obtenu indiqué dans LABB est connu et cohérent), et
- L'âge de la personne respecte les seuils mentionnés dans le *Tableau 2* ci-dessus (qui tiennent compte de l'âge **à la date de référence**, du niveau de formation annoncé dans la source et du lieu de naissance).

Dans le *Tableau 3* ci-dessous se trouve le nombre d'enregistrements conservés par source de données après l'application des règles de reprise.

Tableau 3 : Nombre d'enregistrements conservés par source de données (année 2020, >= 15 ans)²

	LABB*	SECO	LABB\LABB*	SAS	RS	SNC	ESS
Données brutes	1 265 351	2 014 530	852 138	466 955	4 187 668	4 247 637	3 179 337
Données plaus.	1 265 351	1 710 505	833 759	415 308	3 580 687	2 774 491	3 054 716
% données brutes	100%	85%	98%	89%	86%	65%	96%
n Statpop 2020	17%	23%	11%	6%	49%	38%	41%

La source SNC (RFP1990 et RFP200) perd 35% de ses enregistrements, car il s'agit d'informations anciennes, contrairement à la source LABB pour laquelle 100% des enregistrements sont conservés.*

Les données conservées des différentes sources sont ensuite confrontées à la source LABB. Elles sont corrigées dans deux cas de figure :

- Si la personne est en cours de formation dans LABB* à la date de référence de la source, car ces personnes se trompent plus fréquemment que les autres sur leur niveau de formation (voir Annexe 60). Par exemple, une partie de ces personnes anticipe le fait d'obtenir un certificat ou un titre. Ces personnes peuvent alors être identifiées et corrigées avec LABB.
- Si la personne annonce un niveau plus bas que celui de LABB non plausibilisé (qui est utilisé comme borne inférieure).

Dans les deux cas, on suppose que la personne s'est trompée en indiquant son niveau de formation (anticipation de titre ou erreur) et l'information de LABB* ou LABB est reprise.

4.2 Priorisation des sources

Une fois les données des sources sélectionnées et plausibilisées, elles sont utilisées pour attribuer un niveau de formation à l'ensemble de la population. Dans un premier temps, les sources sont classées par ordre de fiabilité en fonction de la manière dont l'information sur le plus haut niveau de formation a été relevée. Dans les données du SECO, le niveau de formation est relevé et vérifié sur la base d'une présentation d'un diplôme ou d'un certificat attestant de la réussite d'une formation. Le niveau de formation issu de LABB est obtenu à partir de données vérifiées et validées par les écoles. Dans les données du SAS, RS et SNC, les personnes annoncent leur niveau de formation sans avoir besoin de

¹ Quitter la Suisse = annonce d'un départ de Suisse suivi d'un retour avec un intervalle de 365 jours ou plus.

² Les données LABB sont des données qui continuent d'évoluer après la date de référence. Les graphiques présentés dans ce document correspondent aux données 2018, 2019 et 2020 à l'état 2021, ainsi que les données 2021 à l'état 2023.

le justifier par un certificat ou un diplôme attestant la réussite d'une formation. Pour l'ESS c'est l'employeur qui indique le niveau de formation de son employé sans fournir de justificatif.

1. LABB*
2. SECO
3. LABB\LABB*
4. SAS
5. RS
6. SNC
7. ESS

Remarques :

- LABB* est la source la plus sûre (validée par des justificatifs) et la plus récente possible (à l'année de référence).
- Le SECO précède « LABB\LABB* », car il peut contenir des informations hors système suisse (diplôme ou titre obtenu à l'étranger) ce qui n'est pas le cas de « LABB\LABB* ».
- SAS précède le RS, car le relevé des informations du SAS se fait de personne à personne, diminuant ainsi le risque d'une mauvaise compréhension de l'information demandée. L'exactitude de l'information fournie est importante pour le bénéficiaire.
- SNC est utilisée après SAS et le RS, car cette source contient des informations anciennes (recensements de 1990 et 2000).
- ESS est la source la moins fiable, car le niveau de formation est fourni par l'employeur et non par la personne directement.

4.2.1 Règles de priorisation pour la reprise des données dans les sources plausibilisées

Une fois les sources priorisées, la reprise de leurs informations se fait selon des règles définies qui combinent l'information de ces différentes sources. L'information de l'ESS ne va pas être reprise directement, car ce n'est pas la personne qui répond, mais son employeur. Par contre, elle va être utilisée en combinaison avec les autres sources pour les appuyer. De plus, les niveaux de formation 3 et 4 donnés dans le RS ne seront considérés que dans certains cas, car leur fiabilité est inférieure aux autres niveaux (voir Annexe 2).

Dans un premier temps, les personnes de moins de 15 ans se voient attribuer un niveau 1, à l'exception des personnes ayant déjà atteint un niveau plus élevé dans LABB. Pour les personnes de 15 ans et plus, l'attribution du niveau de formation se fait selon la priorisation suivante :

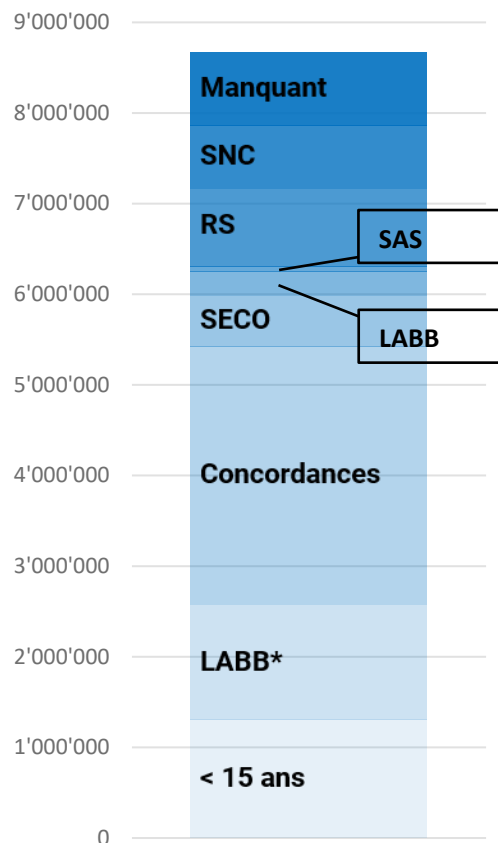
1. LABB*.
2. Contribution de toutes les sources : le niveau de formation achevé est repris si le nombre de sources qui appuient le niveau le plus fréquent a au moins 2 sources de plus que le deuxième niveau le plus fréquent.
(soit c_k : le nombre de sources qui disent k , on garde le niveau i si $c_i - c_j \geq 2, \forall j \in \{1, \dots, 5\}, j \neq i$)
3. Contribution de toutes les sources sans l'ESS : on utilise la même règle qu'au point 2, mais on ne considère pas l'ESS.
4. Contribution de toutes les sources sans RS = (3,4) : on utilise la même règle qu'au point 2, mais on ne considère pas le RS quand sa réponse est 3 ou 4.
5. Contribution de toutes les sources sans l'ESS ni le RS = (3,4) : on utilise la même règle qu'au point 2, mais on ne considère pas l'ESS ni le RS quand sa réponse est 3 ou 4.
6. Si le niveau de formation n'a pas pu être repris selon les règles ci-dessus, mais qu'il y a une information dans LABB ou SECO :
 - 6.1. LABB = SECO : les deux sources sont en accord.
 - 6.2. LABB ou SECO si l'autre source (LABB ou SECO) ne contient pas d'information.

- 6.3. LABB ≠ SECO : si les deux sources sont en désaccord, on regarde parmi les autres sources si une majorité des autres sources tranche en faveur de LABB ou du SECO.
- 6.4. SECO : si les deux sources sont en désaccord, mais que les autres sources n'ont pas pu trancher, c'est le SECO qui l'emporte.
7. Si le niveau de formation n'a pas pu être repris selon les règles ci-dessus mais qu'il y a une information dans RS ou SAS :
- 7.1. RS = SAS : les deux sources sont en accord.
 - 7.2. RS ou SAS si l'autre source (RS ou SAS) ne contient pas d'information.
 - 7.3. RS ≠ SAS : si les deux sources sont en désaccord, on regarde si l'ESS tranche en faveur du RS ou de SAS.
 - 7.4. SAS : si les deux sources sont en désaccord, mais que l'ESS n'a pas pu trancher, c'est SAS qui l'emporte.
8. SNC.

4.2.2 Résultats de la reprise des données dans les sources plausibilisées

Dans les tableaux (*Graphique 3 et Tableau 4 : Fréquence des règles d'attribution agrégées du niveau de formation pour 2020* et *Tableau 5*) et graphiques suivants (*Graphique 3 et Graphique 4*), les règles d'attribution sont agrégées afin de connaître la contribution des différentes sources dans l'attribution des niveaux de formation.

Graphique 3 et Tableau 4 : Fréquence des règles d'attribution agrégées du niveau de formation pour 2020



Règles d'attribution	Fréq.	%	Cumul %
< 15 ans	1 306 271	15.07	15.07
LABB*	1 265 351	14.59	29.66
Concordances	2 852 961	32.90	62.57
SECO	556 972	6.42	68.99
LABB	269 333	3.11	72.10
SAS	60 549	0.70	72.79
RS	851 471	9.82	82.61
SNC	700 775	8.08	90.70
Manquants	806 617	9.30	100.00
Total	8 670 300	100	100

Graphique 4 : Evolution de la fréquence des règles d'attribution agrégées du niveau de formation entre 2018 et 2021

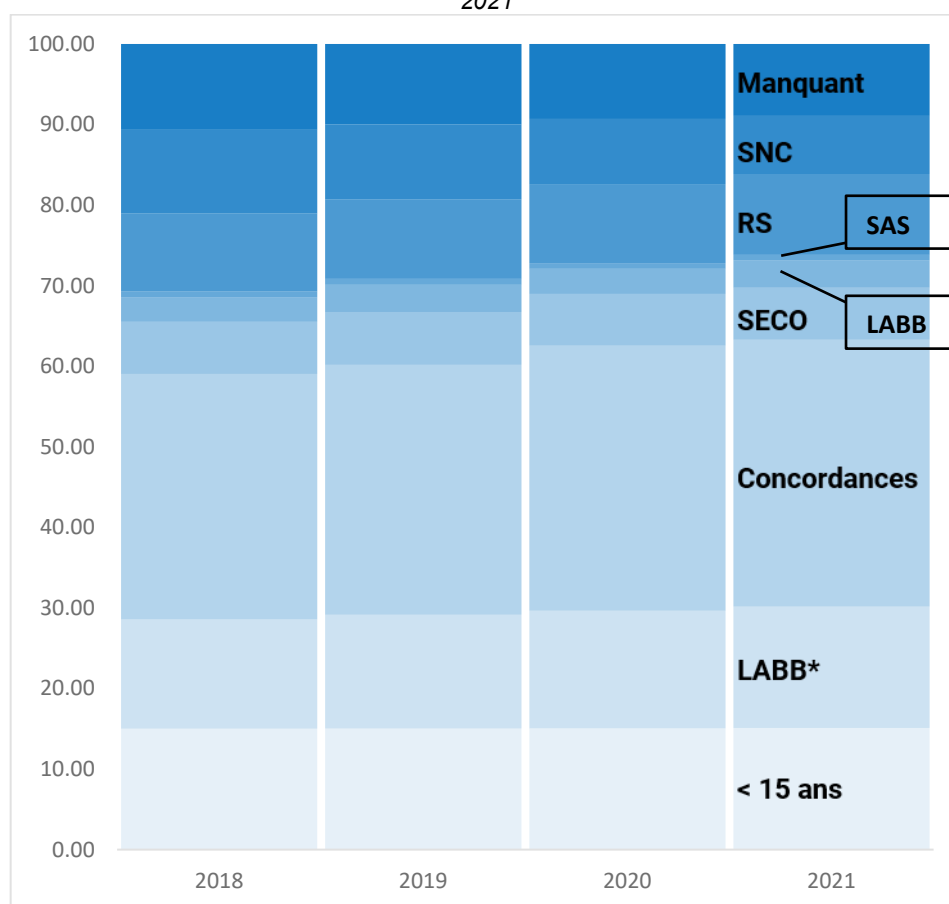


Tableau 5 : Evolution de la fréquence des règles d'attribution agrégées du niveau de formation entre 2018 et 2021

Règles d'attribution	2018	2019	2020	2021	2021-2018
< 15 ans	15.02	15.05	15.07	15.10	0.08
LABB*	13.57	14.11	14.59	15.08	1.51
Concordances	30.45	31.03	32.90	33.11	2.66
SECO	6.49	6.56	6.42	6.47	0.01
LABB	3.03	3.37	3.11	3.39	0.36
SAS	0.74	0.73	0.70	0.69	0.05
RS	9.65	9.88	9.82	10.03	0.38
SNC	10.39	9.30	8.08	7.20	-3.19
Manquants	10.66	9.97	9.30	8.93	-1.73
Total	8 544 527	8 606 033	8 670 300	8 738 791	

Le *Tableau 5* montre que plus le temps passe, plus les sources contiennent de l'information cohérente (les concordances augmentent). LABB* prend de plus en plus de place avec le temps. A l'inverse, comme SNC est la seule source qui ne s'enrichit pas d'années en année et qu'elle est la dernière source utilisée, sa contribution diminue relativement aux autres sources. Les valeurs manquantes sont également de moins en moins présentes, cependant, il en restera toujours jusqu'à un certain seuil dû aux migrants n'ayant pas réalisé leur formation en Suisse.

La couverture du niveau de formation de la population selon la ou les sources de référence est détaillée dans le *Tableau 6* ci-dessous.

Tableau 6 : Fréquence des règles détaillées d'attribution des données des sources, 2020

N°	Description	Fréq.	%	Fréq. Cum.	% Cum.
-	< 15 ans	1 306 271	15.07	1 306 271	15.07
1	LABB*	1 265 351	14.59	2 571 622	29.66
2	Nb de sources avec la valeur la plus fréquente : 2, 2 ^{ème} plus fréq. : 0	1 713 514	19.76	4 285 136	49.42
2	Nb de sources avec la valeur la plus fréquente : 3, 2 ^{ème} plus fréq. : 0	672 318	7.75	4 957 454	57.18
2	Nb de sources avec la valeur la plus fréquente : 3, 2 ^{ème} plus fréq. : 1	88 941	1.03	5 046 395	58.20
2	Nb de sources avec la valeur la plus fréquente : 4, 2 ^{ème} plus fréq. : 0	122 846	1.42	5 169 241	59.62
2	Nb de sources avec la valeur la plus fréquente : 4, 2 ^{ème} plus fréq. : 1	4 765	0.05	5 174 006	59.68
2	Nb de sources avec la valeur la plus fréquente : 4, 2 ^{ème} plus fréq. : 2	7	0.00	5 174 013	59.68
2	Nb de sources avec la valeur la plus fréquente : 5, 2 ^{ème} plus fréq. : 0	5 950	0.07	5 179 963	59.74
2	Nb de sources avec la valeur la plus fréquente : 5, 2 ^{ème} plus fréq. : 1	26	0.00	5 179 989	59.74
3	Nb de sources avec la valeur la plus fréquente : 2, 2 ^{ème} plus fréq. : 0	174 003	2.01	5 353 992	61.75
3	Nb de sources avec la valeur la plus fréquente : 3, 2 ^{ème} plus fréq. : 1	1 184	0.01	5 355 176	61.76
4	Nb de sources avec la valeur la plus fréquente : 2, 2 ^{ème} plus fréq. : 0	64 409	0.74	5 419 585	62.51
4	Nb de sources avec la valeur la plus fréquente : 3, 2 ^{ème} plus fréq. : 1	199	0.00	5 419 784	62.51
5	Nb de sources avec la valeur la plus fréquente : 2, 2 ^{ème} plus fréq. : 0	4 799	0.06	5 424 583	62.57
6.1	SECO = LABB	2 632	0.03	5 427 215	62.60
6.2	LABB - SECO est inconnue	257 831	2.97	5 685 046	65.57
6.2	SECO - LABB est inconnue	522 219	6.02	6 207 265	71.59
6.3	LABB - SECO ≠ LABB, mais plus d'autres sources = LABB	11 502	0.13	6 218 767	71.72
6.3	SECO - SECO ≠ LABB, mais plus d'autres sources = SECO	13 743	0.16	6 232 510	71.88
6.4	SECO	18 378	0.21	6 250 888	72.10
7.1	RS = SAS	1 236	0.01	6 252 124	72.11
7.2	RS - SAS est inconnue	849 025	9.79	7 101 149	81.90
7.2	SAS - RS est inconnue	50 373	0.58	7 151 522	82.48
7.3	RS - RS ≠ SAS, mais ESS = RS	2 446	0.03	7 153 968	82.51
7.3	SAS - SAS ≠ RS, mais ESS = SAS	1 316	0.02	7 155 284	82.53
7.4	SAS	7 624	0.09	7 162 908	82.61
8	SNC	700 775	8.08	7 863 683	90.70
-	Manquants	806 617	9.30	8 670 300	100

Parmi les 8 670 300 personnes de STATPOP en 2020, 15% ont moins de 15 ans et leur niveau de formation est mis par défaut au niveau minimal, même si une ou plusieurs sources indiquent un niveau plus élevé. Cette décision se justifie par le fait qu'il est hautement improbable qu'une personne de moins de 15 ans ait déjà terminé une formation du post-obligatoire. On accepte cependant d'attribuer un niveau plus élevé si la personne de moins de 15 ans est de niveau supérieur dans LABB car cette information est vérifiée (les titres décernés sont validés par les écoles ou les autorités compétentes). On constate

que les sources plausibilisées couvrent 90,7% de la population, les 9,3% restants n'ont donc pas d'information disponible sur le niveau de formation. Un modèle d'imputations a été développé afin de compléter les niveaux manquants et couvrir ainsi l'ensemble de la population. Ce modèle est décrit au chapitre suivant (4.3).

4.3 Imputations

Les personnes pour lesquelles aucun niveau de formation n'a pu être attribué automatiquement selon les règles de priorisation des sources (9.3% de la population résidente permanente de 2020), se verront imputer un niveau de formation selon l'algorithme décrit ci-dessous. Dans le chapitre 4.3.1, nous décrivons l'ensemble d'apprentissage qui est utilisé pour entraîner le modèle d'imputation. Dans le chapitre 4.3.2 sont décrits d'une part l'ensemble à imputer et d'autre part la méthode d'imputation utilisée qui est séparée en deux parties :

- modélisation des probabilités d'être de niveau 1, 2, 3, 4 ou 5 pour chaque personne,
- attribution d'un niveau de formation à chaque personne.

Pour finir, les variables utilisées dans le modèle d'imputation (les variables auxiliaires) sont décrites au chapitre 4.3.3.

4.3.1 Ensemble d'apprentissage

L'ensemble d'apprentissage est défini à partir du RS de l'année de référence. La première étape consiste à attribuer, par priorisation des sources, un niveau de formation aux personnes du RS de l'année de référence. Cette priorisation est la même que celle appliquée au reste de la population et décrite ci-dessus (chapitre 4.2), mais sans utiliser l'information du RS de l'année de référence³. L'ensemble d'apprentissage est constitué des personnes pour lesquelles aucun niveau de formation, hormis le niveau issu du RS de l'année de référence, n'a pu être attribué automatiquement. Cet ensemble d'apprentissage contient le niveau de formation du RS. Il est utilisé pour entraîner le modèle d'imputation qui va servir à déterminer le niveau de formation pour l'ensemble des personnes de STATPOP dont le niveau de formation n'a pas été attribué de manière automatique (via la priorisation des sources en incluant l'information du RS de l'année de référence).

Dans le RS, deux échantillons sont disponibles : ZPERS, qui contient les informations sur les personnes cibles, et HHPERS, qui inclut les personnes cibles ainsi que les membres de leur ménage. Cependant, les ménages dont les personnes diffèrent de celles de STATPOP ne sont pas inclus dans HHPERS, bien que la personne cible figure toujours dans ZPERS. Une pondération ajustée à l'union des deux échantillons (ZPERS + HHPERS) a été créée. Ceci nous permet d'exploiter un unique échantillon contenant l'ensemble des personnes du RS (ZPERS + HHPERS) dont le niveau de formation est renseigné. Les personnes n'ayant pas répondu dans le RS sont exclues de l'échantillon (env. 2%, soit « manquants » dans HHPERS, soit « imputés » dans ZPERS)⁴. L'ensemble d'apprentissage est un sous-ensemble de cet échantillon, il contient les personnes qui ont une information dans l'échantillon du RS de l'année de référence, mais pas d'information dans les autres sources (hormis l'ESS qui n'est pas directement repris dans la priorisation de sources).

³ Ainsi nous apprenons sur une population similaire à la population à imputer. C'est-à-dire les personnes qu'on ne retrouve pas dans les autres sources (les personnes qui n'ont pas été au chômage, qui n'ont pas été à l'aide sociale, qui étaient trop jeunes ou hors de Suisse en 2000, etc.).

⁴ Dans l'ensemble d'apprentissage, la proportion de non-répondants est la même que dans le RS complet.

Les poids recalculés pour l'ensemble du RS (ZPERS et HHPERS) seront utilisés lors de l'imputation. Contrairement à une extrapolation standard par échantillonnage qui exploite directement les poids, dans notre cas, on retranche 1 au poids lors de la modélisation car la population à imputer ne contient pas l'ensemble d'apprentissage.

4.3.2 Algorithme d'imputations

L'algorithme d'imputations s'appuie sur des variables auxiliaires issues de différentes sources (voir chapitre 4.3.3). Étant donné que certaines de ces variables ne sont pas disponibles pour l'ensemble des personnes à imputer, l'algorithme d'imputations s'effectue en distinguant des sous-populations en fonction du nombre de variables auxiliaires disponibles (le détail des sous-populations est décrit au chapitre 4.3.2.1). On commence par réaliser les imputations pour les personnes possédant toutes les variables auxiliaires, puis on applique la même procédure en réduisant le nombre de variables utilisées dans le modèle d'imputation pour les personnes qui en ont moins à disposition.

4.3.2.1 Distinction des sous-populations à imputer

Le modèle d'imputation se base sur le RS dont la population cible ne correspond pas entièrement à la population résidente permanente de STATPOP (par exemple, certaines familles de diplomates internationaux ne sont pas dans la population cible du RS, mais sont dans la population résidente permanente de STATPOP). Par conséquent, il est nécessaire de distinguer différentes sous-populations en prévision des imputations.

On distingue trois sous-populations disjointes dans STATPOP :

1. Population cible du RS (nommée : Pop1). Dans cette population toutes les variables auxiliaires de STATPOP sont disponibles.
2. Population hors cible du RS pour laquelle toutes les variables auxiliaires de STATPOP sont disponibles (nommée : Pop2), par exemple les familles de diplomates internationaux.
3. Population hors cible du RS pour laquelle certaines variables auxiliaires de STATPOP ne sont pas disponibles (nommée : Pop3), par exemple les personnes non formées en ménage ou les personnes en ménages collectifs.

Pour plus d'informations sur les variables auxiliaires, voir le chapitre 4.3.3.

Les personnes de la population Pop1 et Pop2 sont imputées de la même manière. Les personnes de la population Pop3 n'ont pas certaines variables relatives à leur ménage, c'est pourquoi la même procédure d'imputation est appliquée pour ces personnes-là, mais sans les variables de ménage. Le *Tableau 7* ci-dessous montre la répartition des personnes dans les différentes populations.

Tableau 7 : Fréquence des différentes populations dans STATPOP 2020 et dans la population à imputer

Population	STATPOP		A imputer	
	Fréq.	%	Fréq.	%
Pop1	8 482 268	97.83	760 243	94.25
Pop2	22 983	0.27	14 136	1.75
Pop3	165 049	1.90	32 238	4.00
Total	8 670 300	100	806 617	100

De plus, les personnes qui ont une information dans la source ESS sont séparées de celles qui n'en ont pas, car l'information de l'ESS sera utilisée dans le modèle d'imputation. Le *Tableau 8* montre le nombre de personnes dans les 6 sous-ensembles.

Tableau 8 : Fréquence des différentes populations dans l'ensemble à imputer et dans l'ensemble d'apprentissage

	A imputer		Apprentissage	
	Avec ESS	Sans ESS	Avec ESS	Sans ESS
Pop1 %	355 228 46.73	405 015 53.27	30 960 50.34 (pondéré)	31 186 49.66 (pondéré)
Pop2	281	13 855	0	0
Pop3	2 695	29 543	0	0
Total	358 204	448 413	30 960	31 186

4.3.2.2 Application de régressions multinomiales

Pour toutes les personnes pour lesquelles le niveau de formation n'a pas encore pu être attribué, une probabilité d'être de niveau 1, 2, 3, 4, et 5 est calculée. Ce calcul se fait en deux étapes, car selon les analyses réalisées dans le cadre de ce projet (voir Annexe 2), les répondants au RS ont tendance à confondre les niveaux 3 et 4 avec le niveau 2. Par conséquent, les probabilités du niveau de formation vont être modélisées en deux étapes :

1. Régression multinomiale à trois niveaux : 1, 234, 5 (Proba : P_A), qui donne à chaque personne à imputer la probabilité d'être de niveau 1, d'un des niveaux : 2, 3 ou 4, et de niveau 5. L'ensemble d'apprentissage est décrit au chapitre 4.3.1.
2. Régression multinomiale à trois niveaux : 2, 3, 4 (Proba : P_B), qui donne pour chaque personne à imputer, les probabilités conditionnelles d'être de niveaux 2, 3, et 4. L'ensemble d'apprentissage est décrit au chapitre 4.3.1 et est filtré sur les niveaux 2, 3 et 4.

Calcul des probabilités (pour la personne k) :

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

En séparant la procédure en deux étapes, les niveaux 3 et 4, qui sont moins représentés dans la population, ont la possibilité de s'exprimer. En effet, en plus d'avoir une population plus petite dans la deuxième régression, les variables utilisées sont également ciblées pour la distinction 2, 3 et 4.

Suite à l'attribution des probabilités d'être dans chacun des 5 niveaux, une plausibilisation est réalisée. D'une part, LABB non plausibilisé est utilisé comme borne inférieure, c'est-à-dire que si le niveau de formation est trouvé dans LABB, les probabilités des niveaux inférieurs à la valeur dans LABB sont mises à 0 (4 463 personnes modifiées en 2020 (0.55% des imputations)) :

$$P_k(i) = 0, \forall i < niv_{LABB}.$$

D'autre part, les personnes de moins de 18 ans reçoivent une probabilité de 0 pour les niveaux 4 et 5, car elles sont trop jeunes pour être de niveau tertiaire (1 756 personnes modifiées en 2020 (0.22% des imputations)) :

$$P_k(4) = 0, P_k(5) = 0, \forall k | age_k < 18.$$

Les probabilités sont ensuite « rescalées » pour que la somme des probabilités d'être de niveau 1, 2, 3, 4, et 5 soit égale à 1.

4.3.2.3 Imputations individuelles

Une fois les probabilités connues pour chaque personne à imputer, il est encore nécessaire d'attribuer un niveau de formation parmi les 5 afin d'avoir une variable « plus haute formation achevée » directement utilisable. Plusieurs approches sont comparées en Annexe 4 et nous détaillons ci-dessous l'approche utilisée.

L'imputation par le maximum de vraisemblance impute la catégorie ayant la plus grande probabilité estimée. Malheureusement, cette approche tend à créer des biais sur les totaux. En effet, on observe que les totaux de la variable catégorielle peuvent parfois s'éloigner fortement des totaux calculés par sommes des probabilités qui sont sans biais si le modèle est correct.

Pour rester proche des sommes des probabilités, une imputation a été développée. L'idée de cette imputation est de chercher une qualité proche de celle obtenue en utilisant le maximum de vraisemblance, tout en rendant les totaux sur certains domaines cohérents avec la somme des probabilités sur ces domaines. Nous avons choisi de croiser la variable classe d'âge et la variable sexe pour définir les domaines, voir *Tableau 9*.

Tableau 9 : Description des variables classe d'âge et sexe

Sexe	Description
1	Homme
2	Femme

Cat_age	Description
0	< 15 ans
1	15 – 25 ans
2	26 – 35 ans
3	36 – 45 ans
4	46 – 55 ans
5	56 – 65 ans
6	> 65 ans

Les moins de 15 ans ne sont pas imputés, autrement dit, nous allons travailler sur $6 \times 2 = 12$ domaines.

Le raisonnement qui a permis de définir la procédure d'imputation est expliqué ci-dessous en quatre étapes. La première donne un aperçu du cas théorique sur un seul domaine dans un contexte idéal, la deuxième est une adaptation de ce cas à nos données. La troisième étape est l'extension aux 12 domaines définis dans notre projet et la dernière décrit l'algorithme d'optimisation.

1. Cas idéal avec son approche théorique

D'un côté, on aimerait choisir le maximum de vraisemblance qui minimise les erreurs d'imputations individuelles et, d'un autre côté, on souhaiterait que les totaux soient proches des totaux estimés par les probabilités. Idéalement, un compromis entre ces deux objectifs est alors trouvé par optimisation sous contrainte.

Dans le cas d'une régression multinomiale unique, l'attribution du niveau de formation se ferait selon la procédure ci-dessous (où U est la population à imputer et β les coefficients d'une régression multinomiale classifiant les niveaux de formation $i \in \{1,2,3,4,5\}$) :

1. Le maximum de vraisemblance (MLE) permet d'estimer pour une personne $k \in U$ des probabilités $\hat{P}_k(i)$ de niveau de formation en trouvant le paramètre $\hat{\beta}_1$ qui maximise la vraisemblance $L(\beta)$:

$$MLE \rightarrow \hat{P}_k(i) \forall k \in U, i \in \{1,2,3,4,5\} \mid \hat{\beta}_1 = \operatorname{argmax}_{\beta}(L(\beta)).$$

2. Une optimisation sous contrainte permet de trouver le paramètre $\hat{\beta}_2$ qui maximise la vraisemblance

$$\hat{\beta}_2 = \operatorname{argmax}_{\beta}(L(\beta)), \forall i \in \{1, \dots, 5\},$$

sous contrainte que

$$\sum_{k \in U} 1_{\tilde{P}_k(i)=\max_{j=1,2,3,4,5} \{\tilde{P}_k(j)\}} = \sum_{k \in U} \hat{P}_k(i), \forall i \in \{1, \dots, 5\},$$

avec $\tilde{P}$ estimé par la régression multinomiale sous contrainte de paramètre $\hat{\beta}_2$ (on rappelle que $\hat{P}$ est estimé par la régression multinomiale de paramètre $\hat{\beta}_1$).

Autrement dit, un premier modèle modéliserait des $\hat{P}_k(i)$ pour chaque unité k , avec lesquels des totaux seraient estimés (sans biais si le modèle est correct) comme suit :

$$\hat{N}_i = \sum_{k \in A} \hat{P}_k(i), i \in \{1, \dots, 5\}.$$

Ensuite une optimisation sous contrainte assurerait des totaux imputés sans biais valant exactement $\hat{N}_1, \hat{N}_2, \dots, \hat{N}_5$, tout en tentant de maximiser la vraisemblance.

Le cas idéal décrit ci-dessus est difficilement applicable à notre situation :

Dans ce projet, les probabilités ne sont pas directement modélisées par une seule régression. En effet, nous les modélisons avec deux régressions multinomiales suivies d'une plausibilisation. De plus, l'optimisation sous contrainte est numériquement difficile, car les totaux des contraintes sont en escalier. En effet, si les β bougent, il se peut que $\sum_{k \in U} 1_{\tilde{P}_k(i)=\max_{j=1,2,3,4,5} \{\tilde{P}_k(j)\}}$ fasse un saut ou, au contraire, ne change pas.

2. Solution choisie pour l'imputation du niveau de formation

La solution choisie pour pallier ces difficultés est de produire tout d'abord les $\hat{P}_k(i)$ définis au chapitre 0.456.1. On rappelle que ces probabilités sont construites à partir de deux MLE consécutifs produits par deux régressions multinomiales suivies d'une plausibilisation. Ensuite, ces $\hat{P}_k(i)$ sont exploités pour trouver un vecteur $\vec{\alpha} = (\alpha_1, \alpha_2, \alpha_3, \alpha_4, 1)$ qui vérifie

$$\sum_{k \in U} 1_{\alpha_i \hat{P}_k(i)=\max_{j=1,2,3,4,5} \{\alpha_j \hat{P}_k(j)\}} = \sum_{k \in U} \hat{P}_k(i), \forall i \in \{1, 2, 3, 4, 5\}.$$

Il y a donc 4 conditions imposées et 4 paramètres à trouver.

Cette méthode peut être vue comme une optimisation post-MLE. Elle devrait dans la pratique permettre une performance proche d'une imputation par MLE tout en respectant les conditions imposées. On s'attend donc à une perte de performance d'imputation faible pour que les totaux soient respectés (Annexe 4).

Remarque : Une imputation ajustée par un vecteur $\vec{\alpha}$ peut être vue comme une approche basée sur des coûts de mauvaises classifications, où les coûts ont été fixés pour que les totaux soient respectés⁵.

3. Extension de l'attribution du niveau de formation « catégoriel » à plusieurs domaines

Cette méthode d'attribution du niveau de formation est étendue aux 12 domaines, soit

⁵ Trevor Hastie, Robert Tibshirani, Jerome Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Second Edition, Springer Series in Statistics, 2009.

$$D_{r,s} = \{k \in U \mid \text{sexe} = r, \text{cat_age} = s\}, \text{ avec } r \in \{1, 2\} \text{ et } s \in \{1, 2, \dots, 6\}.$$

Pour chaque domaine $D_{r,s}$, on cherche $\vec{\alpha}_{D_{r,s}} = (\alpha_{D_{r,s},1}, \alpha_{D_{r,s},2}, \alpha_{D_{r,s},3}, \alpha_{D_{r,s},4}, 1)$ tel que

$$\sum_{k \in D_{r,s}} 1_{\{\alpha_{D_{r,s},i} \hat{P}_k(i) = \max_{j=1,2,3,4,5} \{\alpha_{D_{r,s},j} \hat{P}_k(j)\}\}} = \sum_{k \in D_{r,s}} \hat{P}_k(i), \forall i \in \{1, 2, 3, 4, 5\}.$$

Ainsi, pour chaque domaine, pour faire coïncider les totaux des 5 niveaux sur les 12 domaines, on vérifie une perte de performance faible par rapport au maximum de vraisemblance en Annexe 4.

4. Algorithme d'optimisation numérique

En pratique, pour calculer les paramètres $\vec{\alpha}_{D_{r,s}}$, nous parcourons un maillage dans le but de trouver un point qui minimise la distance euclidienne entre les totaux de l'imputation et les totaux estimés. Autrement dit, trouver $\vec{\alpha}_{D_{r,s}}$ qui minimise

$$\sum_{i=1}^5 \left(\sum_{k \in D_{r,s}} 1_{\{\alpha_{D_{r,s},i} \hat{P}_k(i) = \max_{j=1,2,3,4,5} \{\alpha_{D_{r,s},j} \hat{P}_k(j)\}\}} - \sum_{k \in D_{r,s}} \hat{P}_k(i) \right)^2.$$

Nous commençons par estimer les totaux souhaités ($\hat{N}_1, \hat{N}_2, \dots, \hat{N}_5$) (estimation non biaisée si le modèle est correct) sur chaque domaine $D_{r,s}$

$$\hat{N}_{D_{r,s},i} = \sum_{k \in D_{r,s}} \hat{P}_k(i), i \in \{1, \dots, 5\}.$$

Ensuite l'algorithme d'optimisation suivant est appliqué :

1. Initialisation de la boucle : un maillage à 5 nœuds, de précision 0.4, centré en 1 (les paramètres de $\vec{\alpha}_{D_{r,s}}$ valent 0.2, 0.6, 1, 1.4, 1.8).
2. Faire parcourir le maillage aux 4 paramètres ($\alpha_{D_{r,s},1}, \alpha_{D_{r,s},2}, \alpha_{D_{r,s},3}, \alpha_{D_{r,s},4}$) et pour chaque combinaison de paramètres (5^4 combinaisons, car chaque paramètre parcourt 5 points du maillage) :
 - a. Compter le nombre d'imputations de chaque niveau de formation qu'on ferait sur le domaine :

$$\tilde{N}_i = \sum_{k \in D_{r,s}} 1_{\{\alpha_{D_{r,s},i} \hat{P}_k(i) = \max_{j=1,2,3,4,5} \{\alpha_{D_{r,s},j} \hat{P}_k(j)\}\}}, i \in \{1, \dots, 5\}.$$

- b. Calculer la distance :

$$d = \sum_{i=1}^5 (\tilde{N}_i - \hat{N}_i)^2.$$

3. Choisir la combinaison de paramètres $\vec{\alpha}_{min,D_{r,s}}$ qui minimise d .
4. Refaire les étape 2 et 3, avec un maillage à 5 nœuds centrés en $\vec{\alpha}_{min,D_{r,s}}$ (les paramètres qui optimisent le maillage précédent), avec les précisions suivantes : 0.3, 0.2, 0.15 et 0.1.

Pour finir, l'imputation se fait selon les paramètres finaux $\vec{\alpha}_{min,D_{r,s}}$ de l'algorithme d'optimisation

$$\text{Niveau imputé} = \operatorname{argmax}_{j=1,2,3,4,5} (\alpha_{min,D_{r,s},j} P_k(j)).$$

Cette méthode constitue un compromis pour avoir des totaux peu biaisés et une performance individuelle proche du maximum de vraisemblance. En effet, cette méthode d'imputation permet d'obtenir une qualité individuelle proche de celle du maximum de vraisemblance (env. 1.5% d'erreurs en plus), tout en assurant des totaux cohérents dans les domaines choisis.

La vision intuitive derrière cette méthode est la suivante : lorsque les probabilités ne permettent pas de trancher clairement en faveur d'un niveau (comme c'est le cas en ayant par exemple deux probabilités élevées proches), l'algorithme pourrait favoriser le deuxième niveau le plus vraisemblable afin de garantir les totaux.

Par exemple, si pour une unité, nous avons les probabilités suivantes :

$$\hat{P}(1) = 0.1, \hat{P}(2) = 0.1, \hat{P}(3) = 0.1, \hat{P}(4) = 0.34, \hat{P}(5) = 0.36,$$

une imputation par maximum de vraisemblance attribuerait le niveau 5. Cependant, si l'on constate que dans ce domaine il y a trop de niveaux 5 et pas assez de niveaux 4 par rapport aux estimations non-biaisées, l'algorithme pourrait ici imputer le niveau 4 au lieu du 5. Ainsi, l'estimation de l'erreur change peu (dans l'exemple ci-dessus, il augmente de 0.02) et les totaux réduisent leur biais.

4.3.3 Variables auxiliaires

Le modèle décrit ci-dessus utilise des variables auxiliaires reprises de différentes sources : CdC, STATENT, STATPOP (effectif, ménage et typologie des ménages), niveaux géographiques de la Suisse et StatBL. Ces variables sont utilisées de deux manières : d'une part, les plus importantes sont directement intégrées dans le modèle d'imputation ; d'autre part, l'ensemble des variables pertinentes sont utilisées dans des arbres de décision afin d'identifier les croisements optimaux pour créer de nouvelles variables auxiliaires. Ces arbres de décision sont appliqués à l'ensemble d'apprentissage de l'année précédant l'année de référence, afin d'éviter un surapprentissage⁶. En effet, c'est un ensemble indépendant de l'ensemble d'apprentissage de l'année de référence. Les variables utilisées dans ces arbres sont décrites à l'Annexe 5.

Si la fréquence d'une modalité d'une variable dans l'ensemble d'apprentissage n'est pas assez élevée, les modalités sont fusionnées entre elles (soit par une modalité qui chapeaute deux modalités, soit par deux modalités similaires, voir *Tableau 10* ci-dessous). Les variables auxiliaires créées par arbre de décision n'ont pas de modalités trop petites (contrainte lors de la création de l'arbre).

Tableau 10 : Liste des variables auxiliaires du modèle d'imputation

Variable	Cat ⁷ .	Description
cat_sal_age	1	Croisement entre classe de salaire et âge. Si une modalité est trop petite, deux modalités avec un salaire proche sont fusionnées
Age_sal_noga_&niv. ⁸	1	Croisement entre classe de salaire, âge et la branche de l'entreprise (NOGA), défini automatiquement par un arbre de décision
STALAN	1	Type de commune : urbaine, intermédiaire, rurale
nb_logement	1	Type de maison : individuelle ou à plusieurs logements
compBasics*	1	Composition du ménage selon l'âge et le sexe
FOREIGNHOUSEHOLDPERM*	1	Composition du ménage selon la nationalité, si une

⁶ Surapprentissage (« overfitting » en anglais) : quand une analyse correspond trop précisément à un ensemble de données spécifique et, par conséquent, ne serait plus pertinent pour un autre ensemble de données. Cela se produit par exemple quand une modélisation exploite trop de variables relativement à son nombre de données.

⁷ Variable catégorielle : 1 = oui, 0 = non.

⁸ &niv. : correspond à l'agrégation de la première ou de la deuxième régression multinomiale.

		modalité est trop petite, elle est regroupée avec une modalité proche
HOUSEHOLDTYP25AGGR*	1	Type de ménage
COUNTOFPERSONPERM*	0	Nombre de personnes dans le ménage
flag_salaire	1	Est salarié (oui / non)
flag_ind	1	Est indépendant (oui / non)
r_tot_log	0	Log du revenu annuel total (CdC)
Emptot	0	Nombre d'employés dans l'entreprise (0 si pas employé)
noga_grp	1	Branche de l'entreprise (NOGA regroupé, automatiquement et manuellement)
SEX	1	Sexe : homme, femme
nb_enfant	1	Nombre d'enfants (0, 1, 2, 3, 4 ou +), si la dernière modalité est trop petite, elle est regroupée avec la précédente
RESIDPERMIT	1	Permis de résidence, si une modalité est trop petite, elle est regroupée avec la modalité voisine
Cat_AgePremierEnfant	1	Âge à la naissance du premier enfant
Celibataire	1	Etat civil : célibataire (oui / non)
Nat	1	Nationalité (regroupé manuellement)
Birth	1	Pays de naissance (regroupé manuellement)
justArrived	1	Durée de séjour en Suisse (regroupé manuellement)
age1	0	Âge continu par morceaux (> 20 ans)
age2	0	Âge continu par morceaux (> 25 ans)
age3	0	Âge continu par morceaux (> 30 ans)
age4	0	Âge continu par morceaux (> 45 ans)
age5	0	Âge continu par morceaux (> 60 ans)
age6	0	Âge continu par morceaux (> 75 ans)
cat_age	1	Classe d'âge
Job_avant20	1	A eu un revenu soumis à cotisation avant 20 ans
Plaus_niv&niv._lse	1	Niveau de formation dans l'ESS, si la valeur ESS est connue.
11 variables	1	Définies par arbre de décision sans utiliser l'ESS, la description se trouve à l'Annexe 5.
11 variables	1	Définies par arbre de décision si la valeur ESS est connue, la description se trouve à l'Annexe 5.

* Variable indisponible pour la population Pop3.

4.4 Résultats

Le niveau de formation est maintenant déterminé pour l'ensemble de la population résidente permanente. La construction de cette variable, reproductible chaque année, peut se résumer ainsi :

- Enfants de moins de 15 ans : niveau de formation forcé à 1,
- Plausibilisation des sources selon la méthode décrite au chapitre 4.1,
- Personnes avec au moins un enregistrement dans les sources LABB, RS, SECO, SAS ou SNC : niveau de formation selon la priorisation décrite au chapitre 4.2,
- Personnes sans niveau de formation issu des sources : niveau de formation imputé selon la méthode décrite au chapitre 4.3.

Ce chapitre présente quelques résultats en distinguant certaines sous-populations (p. ex. en fonction des classes d'âge ou de la nationalité) ainsi que l'évolution du niveau de formation de la population entre 2019 et 2021.

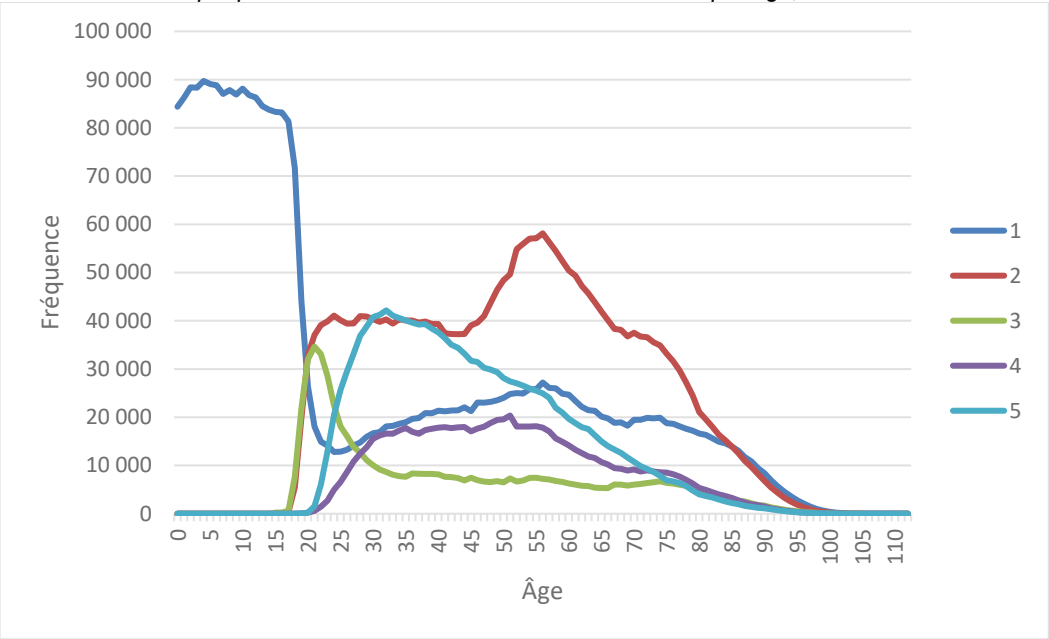
4.4.1 Résultats de la variable plus haute formation achevée pour l'année 2020

Les développements de la méthode ainsi que les analyses ont principalement été réalisés avec les données 2020. Dans le *Tableau 11* et le *Graphique 5* ci-dessous se trouvent des fréquences du niveau de formation par âge.

Tableau 11 : Fréquence des niveaux de formation pour la population résidente permanente, en 2020

Niveaux	0 ≤ Age < 115		Age > 14 ans		24 < Age < 65	
	Fréq.	%	Fréq.	%	Fréq.	%
1	3 065 093	35.35	1 758 825	23.88	848 699	17.55
2	2 690 992	31.04	2 690 992	36.54	1 778 942	36.78
3	638 282	7.36	638 279	8.67	327 617	6.77
4	823 104	9.49	823 104	11.18	642 447	13.28
5	1 452 829	16.76	1 452 829	19.73	1 238 443	25.61
Total	8 670 300	100	7 364 029	100	4 836 148	100

Graphique 5 : Distribution des niveaux de formation par âge, en 2020



4.4.2 Evolution des niveaux de formation de 2019 à 2021

Le *Tableau 12* ci-dessous montre l'évolution des niveaux de formation dans la population résidente permanente entre 2019 et 2021.

Tableau 12 : Evolution des niveaux de formation de 2019 à 2021

Niveaux	2019		2020		2021		2021-2019	
	Fréq.	%	Fréq.	%	Fréq.	%	%	
1	3 066 373	35.63	3 065 093	35.35	3 064 071	35.06		-0.57
2	2 718 585	31.59	2 690 992	31.04	2 681 502	30.69		-0.90
3	634 021	7.37	638 282	7.36	642 366	7.35		-0.02
4	795 365	9.24	823 104	9.49	835 810	9.56		0.32
5	1 391 689	16.17	1 452 829	16.76	1 515 042	17.34		1.17
Total	8 606 033		8 670 300		8 738 791			

Dans le *Tableau 12* ci-dessus, nous remarquons une tendance à l'augmentation du niveau de formation dans la population suisse avec le temps.

Regardons cette évolution en séparant des données attribuées par priorisation de sources de celles imputées par algorithme d'imputation. Dans le *Tableau 13* ci-dessous, se trouve l'évolution des niveaux de formation attribués automatiquement à l'aide des sources plausibilisées.

Tableau 13 : Evolution des niveaux de formation attribués par priorisation des sources, de 2019 à 2021

Niveaux	2019		2020		2021		2021-2019	
	Fréq.	%	Fréq.	%	Fréq.	%	%	
1	2 911 464	37.58	2 913 053	37.04	2 917 620	36.66		-0.92
2	2 451 596	31.64	2 469 737	31.41	2 473 349	31.08		-0.56
3	531 027	6.85	541 279	6.88	547 814	6.88		0.03
4	695 265	8.97	719 061	9.14	742 326	9.33		0.35
5	1 158 255	14.95	1 220 553	15.52	1 277 067	16.05		1.10
Total	7 747 607		7 863 683		7 958 176			
%	90.03		90.70		91.07			

La tendance à l'augmentation du niveau de formation dans la population en Suisse se constate également sur les niveaux de formation attribués par priorisation de sources. Nous constatons aussi que les sources couvrent d'année en année une plus grande part de la population, ce qui implique une diminution des valeurs à imputer. La couverture des sources ne pourra jamais atteindre 100%, car il y a chaque année de nouvelles migrations ainsi que des personnes qui étudient à l'étranger qui ne seront pas entièrement présentes dans les sources et resteront à imputer.

Après avoir regardé l'évolution des niveaux de formation attribués par priorisation des sources, le *Tableau 14* ci-dessous montre l'évolution des niveaux de formation qui ont été imputés à l'aide du RS (env. 10% de la population).

Tableau 14 : Evolution des niveaux de formation attribués par imputations, de 2019 à 2021

	2019		2020		2021		2021-2019
Niveaux	Freq.	%	Freq.	%	Freq.	%	%
1	154 909	18.05	152 040	18.85	146 451	18.76	0.72
2	266 989	31.10	221 255	27.43	208 153	26.67	-4.44
3	102 994	12.00	97 003	12.03	94 552	12.11	0.11
4	100 100	11.66	104 043	12.90	93 484	11.98	0.31
5	233 434	27.19	232 276	28.80	237 975	30.49	3.29
Total	858 426		806 617		780 615		
%	9.97		9.30		8.93		

Le *Tableau 14* montre une forte diminution de l'imputation du niveau 2 entre 2019 et 2020, passant de 31,1% à 27,4%. Cette baisse semble être liée à la modification des libellés dans le questionnaire du RS. En effet, en 2020, les formulations concernant les niveaux de formation ont été ajustées afin d'améliorer la clarté pour les répondants. Selon les analyses effectuées dans le cadre de ce projet, ces ajustements ont entraîné un changement dans les réponses, en particulier entre les catégories 2 et 4. À partir de 2020, les confusions entre les niveaux 4 et 2 ont augmenté (plus de personnes répondent 4), tandis que les confusions entre les niveaux 1 et 2 se sont atténuées (plus de personnes répondent 1), ce qui conduit à une diminution du nombre de personnes se classant au niveau 2 (voir Annexe 3). Ces modifications dans le RS impactent ainsi directement l'imputation.

4.4.3 Performance de la méthode

Comme mentionné précédemment, nous avons identifié des incohérences dans le RS concernant les personnes déclarant un niveau de formation 3 ou 4. En effet, en comparant le niveau de formation indiqué dans le RS avec plusieurs autres sources sur une sous-population représentant environ 40 % de l'ensemble, environ la moitié des personnes concernées n'aurait en réalité pas un niveau 3 ou 4, mais un niveau 2 (voir Annexe 2).

L'analyse de fiabilité du modèle d'imputation repose sur un ensemble d'apprentissage basé sur les données du RS. Autrement dit, nous utilisons un ensemble contenant du bruit à la fois pour imputer les données et pour estimer la qualité de l'imputation. Cela rend l'estimation de la qualité de l'imputation peu crédible, en particulier pour les niveaux 3 et 4. Cependant, si le modèle était correct, ces estimations seraient justes. L'estimation de la qualité est calculée selon la formule suivante :

$$\sum_{k \in U} 1 - \hat{P}_k(i_k),$$

où $i_k \in \{1, 2, 3, 4, 5\}$ est la valeur imputée pour la personne k . Le résultat de l'estimation de la qualité pour l'ensemble de la population résidente permanente est présenté au *Tableau 15*.

Tableau 15 : Estimation d'erreur pour l'ensemble de la population résidente permanente, en 2020

	STATPOP	%	Estimation d'erreur
Priorisation des sources	7 863 683	90.7%	-
Imputation	806 617	9.3%	39%
Total	8 670 300	100%	3.6%

Dans *Tableau 15* se trouve l'estimation d'erreur sur l'ensemble des valeurs de la variable niveau de formation achevée. Si la qualité ne peut pas être déterminée pour 90.7% de la population en raison de l'attribution du niveau de formation par priorisation des sources, les 9.3% qui ont été imputés ont une estimation d'erreur de 39% en se basant sur le RS. Ce qui représente une erreur de 3.6% sur l'ensemble de la population.

Afin de mieux comprendre les limites du modèle d'imputation, nous analysons les erreurs d'imputation du niveau de formation qu'il produit. Pour identifier les confusions entre les niveaux de formation créées par le modèle, les données de l'ensemble d'apprentissage sont imputées. Les réponses du RS (plausibilisées) sont ensuite confrontées à cette imputation.

Le *Tableau 16* et le *Tableau 17* permettent de mettre en évidence les différences entre le résultat des imputations et les réponses du RS.

Tableau 16 : Matrice de confusion du niveau de formation entre les imputations et le RS pondéré, en 2020

Niveau du RS	Niveau des données imputées					Total
	1	2	3	4	5	
1	78 808	20 508	17 690	4 652	10 349	132 007
2	17 460	159 442	19 403	35 796	9 887	241 988
3	13 735	18 389	33 575	13 192	15 489	94 379
4	4 276	36 818	9 122	47 969	16 535	114 719
5	11 542	7 286	17 012	19 009	182 080	236 929
Total	125 821	242 443	96 802	120 617	234 340	820 022

Tableau 17 : Matrice de confusion du niveau de formation entre les imputations et le RS pondéré (en %)

Niveau du RS	Niveau des données imputées				
	1	2	3	4	5
1	9.6	2.5	2.2	0.6	1.3
2	2.1	19.4	2.4	4.4	1.2
3	1.7	2.2	4.1	1.6	1.9
4	0.5	4.5	1.1	5.8	2.0
5	1.4	0.9	2.1	2.3	22.2

En rouge les deux valeurs hors diagonale les plus élevées.

Comme le *Tableau 17* ci-dessus le montre, la méthode d'imputation tend à créer des symétries entre les confusions (autant de confusions 2 vers 4 que de confusions 4 vers 2). Ainsi, les totaux par niveau de formation imputé restent proches de ceux estimés par le RS, c'est-à-dire que la somme par ligne est proche de la somme par colonne pour chaque niveau. Notons aussi que la plus grande confusion observée est celle entre les niveaux de formation 2 et 4, qui est responsable de 8.9% d'erreurs (4.5% + 4.4%), soit environ un quart de l'ensemble des erreurs.

La faible qualité de l'imputation s'explique par plusieurs facteurs. Notamment, la population à imputer est particulière (migrants récents, personnes âgées, etc.), ce qui limite la disponibilité d'informations dans les variables auxiliaires. Cependant, la principale cause de cette qualité réduite réside dans la présence importante d'erreurs dans l'ensemble d'apprentissage (le RS), qui se répercute sur les valeurs imputées.

Bien que le RS soit bruité au niveau des réponses (voir Annexe 2), son utilisation est indispensable, car il s'agit de la seule source couvrant de manière suffisamment large l'ensemble de la population. Cependant, en raison des problèmes de qualité, l'imputation est limitée au strict nécessaire et l'approche fondée sur la priorisation des sources est privilégiée.

4.4.4 Comparaison avec le relevé structurel

L'information du niveau de formation au niveau suisse peut déjà être calculée directement à l'aide du RS. Dans ce chapitre, nous comparons les extrapolations faites à partir du RS et les niveaux de formation obtenus dans STATPOP+.

Après plusieurs analyses, il est apparu que les personnes qui répondaient dans le RS avaient tendance à avoir du mal à se placer dans certaines catégories (notamment les 3 et 4) ce qui a pour conséquence que le RS a beaucoup plus de personnes de niveaux 3 et 4 et moins de personnes de niveau 2 que dans STATPOP+ (dû aux priorisations de source). Dans ces analyses, nous avons aussi constaté que les personnes qui ne répondaient pas à la question de la plus haute formation achevée dans le RS (valeur -9 dans le *Graphique 6* et le *Tableau 18* ci-dessous) avaient une tendance à être principalement des personnes de niveau 1. Par conséquent, si l'information relevée dans le RS est extrapolée pour la Suisse, en excluant les non-répondants ou en les répartissant dans les autres niveaux de formation, le niveau 1 est sous-évalué (voir Annexe 2). Nous avons aussi noté que le libellé de la question avait son importance et que la manière dont la question est posée dans le RS a de l'influence dans les réponses (voir Annexe 3).

Suite à ces analyses, nous nous attendons, à avoir dans nos résultats, significativement moins de personnes dans les niveaux 3 et 4 et un peu plus de personnes dans les niveaux 1 et 2 que dans le RS (voir *Graphique 6* et *Tableau 18*).

Graphique 6 : Répartition du niveau de formation dans STATPOP (filtré sur la population cible du RS) et le RS pondéré, par âge en 2020

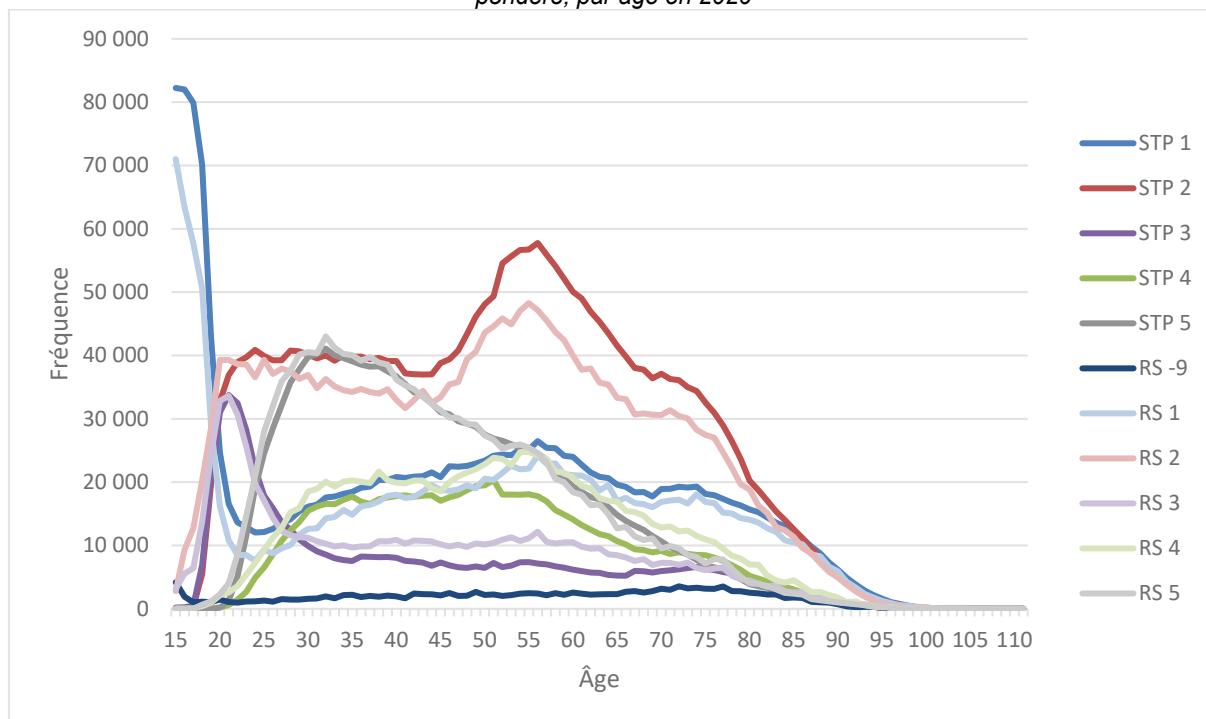


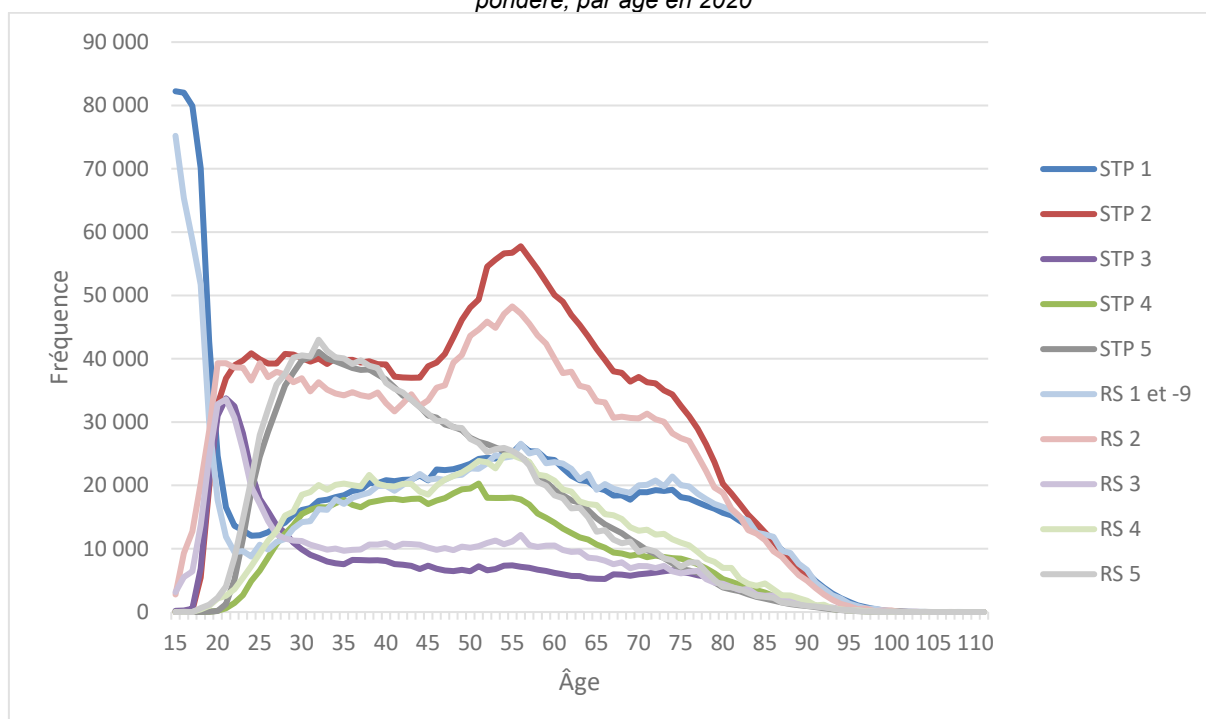
Tableau 18 : Répartition du niveau de formation dans STATPOP (filtré sur la population cible du RS) et le RS pondéré en 2020

Niveaux de formation	RS		STATPOP+	
	Fréq.	%	Fréq.	%
-9	162 357	2.3	0	0
1	1 408 355	19.6	1 682 830	23.4
2	2 358 704	32.8	2 650 826	36.9
3	761 434	10.6	622 594	8.7
4	1 049 460	14.6	816 320	11.4
5	1 447 405	20.1	1 415 145	19.7
Total	7 187 715	100	7 187 715	100

Les résultats obtenus sont cohérents avec les analyses décrites ci-dessus. En effet, dans la variable de STATPOP+, il y a plus de niveaux de formation 1 et 2 et moins de niveaux 3 et 4 que dans le RS. Le nombre de niveaux 5 par âge est similaire pour les deux méthodes.

Comme on le voit dans le *Graphique 7* ci-dessous, en attribuant la non-réponse du RS au niveau 1, la courbe du RS ainsi obtenue est très proche de celle de STATPOP+ à partir de 35 ans.

Graphique 7 : Répartition du niveau de formation dans STATPOP (filtré sur la population cible du RS) et le RS pondéré, par âge en 2020



4.4.5 Diffusion

Il est prévu pour juin 2025 de mettre à disposition une carte interactive du niveau de formation pour l'ensemble de la Suisse. Cette carte, détaillée au niveau de l'hectare, permettra d'évaluer le niveau de formation en Suisse à un niveau fin de granularité. Il sera par exemple possible de comparer le niveau de formation des personnes dans les différents quartiers des grandes villes ou encore de comparer les revenus de travail par niveau de formation à un niveau géographique fin.

Une autre utilisation consiste à mettre cette variable à disposition des chercheurs qui pourront, via un couplage de leurs données avec cette variable, enrichir leurs analyses d'une information inédite et fondamentale au niveau sociétal. Elle permettra d'être utilisée au niveau individuel comme variable explicative pour des modèles statistiques et ainsi aider à mieux comprendre d'autres problématiques concernant différentes thématiques, comme par exemple des analyses sur la pauvreté. Il faudra tenir compte du fait que cette variable est imputée pour environ 9% de la population et que les analyses doivent être interprétées en considérant l'incertitude liée à ces imputations.

La construction de cette variable est reproductible chaque année et permet d'actualiser les résultats à un rythme annuel. La carte interactive ainsi que la variable mise à disposition pour d'éventuels appariements pourront être actualisées chaque année.

Les micro-données concernant le niveau de formation seront livrées avec les variables suivantes :

- numéro AVS (pseudonymisé),
- niveau de la plus haute formation achevée (*Tableau 1*),
- P1, P2, P3, P4 et P5 : les probabilités issues des régressions multinomiales (chapitre 4.3.2),
- flags : indiquent comment le niveau de formation a été attribué (chapitre 4.2).

5 Perspectives

Publier la variable « Niveau de la plus haute formation achevée » en tant que statistique expérimentale nous permettra d'avoir des retours d'utilisateurs sur les données et d'améliorer cette statistique en fonction. De notre côté, nous pourrions également continuer de travailler sur des améliorations. Nous voudrions, de plus, faire des tests et analyses pour savoir depuis quelle année, nous pourrions produire cette statistique (disponibilité et qualité des sources).

À moyen terme, la variable « Niveau de la plus haute formation achevée » devrait être intégrée dans la production courante de STATPOP+.

Nous prévoyons également de faire des tests afin de trouver une partition plus adaptée de la population sur laquelle appliquer l'algorithme d'optimisation qui fait coïncider les totaux de l'imputation (nombre de personnes pour chaque niveau) avec les totaux calculés par la somme des probabilités (chapitre 4.3.2). Actuellement, cet algorithme est appliqué sur le produit cartésien entre les variables « catégorie d'âge » et « sexe » qui définit cette partition. Il s'agit donc d'une partition discrète (la plus fine) de la population avec ces deux variables. L'idée serait alors de trouver une partition de la population, en fonction de plus de variables choisies (sans en faire le produit cartésien, qui contiendrait beaucoup trop d'éléments). On espère ainsi détecter des sous-populations plus utiles afin de réduire les biais sur les totaux.

De futurs travaux pourraient être prochainement investis pour le domaine de la plus haute formation achevée en fonction de sa faisabilité et des besoins. En disposant de l'information du domaine de formation correspondant au niveau de la plus haute formation achevée, il est alors possible de répondre à certaines questions liées au manque de personnel qualifié. Plusieurs sources disposent d'informations concernant le domaine de formation (LABB, RS, SECO et SNC). Au même titre que le niveau de formation, les domaines ne sont pas relevés de la même manière entre les différentes sources et les nomenclatures ne correspondent pas toujours. Dans LABB, on dispose des domaines selon la nomenclature CITE, alors que pour les autres sources, il s'agit des codes de la profession apprise (ISCO). Il est cependant possible, via une table de correspondance (Berufsdatenbank) disponible à l'OFS, de faire correspondre les codes de professions aux domaines de formation. La procédure d'attribution des domaines de formation ainsi que la méthode pour gérer les valeurs manquantes restent à réfléchir et à développer.

Annexes

Annexe 1 LABB : détermination du niveau de la plus haute formation achevée

Pour chaque enregistrement de LABB (élève ou titre) on attribue un niveau de formation. Pour les enregistrements de type « titre », c'est trivial : on définit le niveau de formation en fonction du titre obtenu. Pour tous les enregistrements de type « élève », on impute le niveau de formation correspondant au titre précédent le plus probable. Par exemple, si une personne est en cours d'études dans une HEU, on fait l'hypothèse très probable qu'elle a préalablement obtenu une maturité gymnasiale. Pour cet enregistrement de type « élève » dans une haute école suisse, on impute alors le niveau de formation en lui attribuant un niveau de formation = 3 (correspondant au niveau « formations générales »). Ces imputations donnent une valeur minimale du vrai niveau de formation de la personne au moment de l'enregistrement de type « élève », c'est-à-dire qu'elle peut avoir obtenu un titre de niveau plus élevé mais pas inférieur.

Le niveau de formation d'une personne dans LABB est finalement déterminé par le plus haut niveau (observé ou imputé) correspondant à ses différents enregistrements depuis 2011. Si plusieurs enregistrements fournissent une même valeur du niveau de formation maximal, on privilégie d'abord les enregistrements dont le niveau est non imputé (donc réellement observé). Ce choix ne joue pas de rôle sur le niveau de formation, car il est identique dans les deux cas, mais il a un impact sur la valeur des autres variables retenues dans LABB (par exemple l'année à partir de laquelle la personne obtient ce niveau de formation). Si cette condition ne suffit pas, c'est-à-dire qu'il existe plusieurs enregistrements de même type (observés ou imputés) pour déterminer le niveau de formation, on retiendra alors le niveau correspondant à l'enregistrement le plus ancien.

Les règles d'imputations sont les suivantes (les niveaux indiqués entre parenthèses correspondent à la nomenclature détaillée à 5 niveaux décrite plus haut) :

- Si la personne est étudiante dans une formation menant à une maturité professionnelle, alors on impute un niveau de formation achevée égal à celui de l'obtention d'un CFC (niveau = 2).
- Si la personne est étudiante dans une formation menant à une maturité spécialisée, on impute un niveau de formation achevée égal à celui de l'obtention d'un titre d'une école de culture générale (niveau = 3).
- Si la personne est étudiante dans une formation menant à une passerelle ou à une formation transitoire, on impute un niveau de formation achevée égal à celui de l'obtention d'une maturité (niveau = 3).
- Si la personne est étudiante dans une formation menant à un CFC, une maturité gymnasiale ou à un certificat d'une école de culture générale, on impute un niveau de formation achevée égal à l'école obligatoire (niveau = 1).
- Si la personne est étudiante en formation professionnelle supérieure, on impute un niveau de formation achevée égal à celui de l'obtention d'un CFC (niveau = 2).
- Si la personne est étudiante dans une HES au niveau bachelor, on impute un niveau de formation achevée égal à celui de l'obtention d'une maturité professionnelle (niveau = 3).
- Si la personne est étudiante dans une HEU ou une HEP au niveau bachelor, on impute un niveau de formation achevée égal à celui de l'obtention d'une maturité gymnasiale (niveau = 3).
- Si la personne est étudiante dans une haute école (HEU, HES ou HEP) au niveau formation continue (on parle ici des Master of Advanced Studies qui demandent un titre des hautes écoles), master ou doctorat, on impute un niveau de formation achevée égal à celui de l'obtention d'un titre d'une haute école (niveau = 5).

Le *Tableau 19* ci-dessous indique la proportion de niveaux de formation de LABB obtenus sur la base d'un titre réellement observé (valeur non imputée) et de ceux obtenus sur la base d'un enregistrement imputé (pas de titre observé dans LABB).

Tableau 19 : Valeurs imputées et non imputées selon le plus haut niveau de formation obtenu dans LABB

	École obligatoire	Degré secondaire II : formation professionnelle	Degré secondaire II : formation générale	Formation professionnelle supérieure	Hautes écoles
Valeur Non imputée	0	440 490	215 808	197 376	395 449
Valeur Imputée	1 415 177	67 075	59 183	0	54 934
Total	1 415 177	507 565	274 991	197 376	450 383
% imputé	100%	13%	22%	0%	12%

Remarques :

- Les personnes avec un niveau de formation « École obligatoire » ont toutes un niveau imputé, car il n'y a pas de titre décerné à l'école obligatoire.
- Pour les personnes de niveau « degré secondaire II : formation professionnelle », 13% sont imputées et la très grande majorité d'entre eux (98%) sont en formation menant à un titre de la formation professionnelle supérieure (2% suivent une formation menant à la maturité professionnelle).
- Pour les personnes de niveau « degré secondaire II : formation générale », 22% sont imputées et la très grande majorité d'entre eux (96%) sont étudiants dans une haute école (4% en formation menant à un titre de passerelle ou de maturité spécialisée).
- Les personnes avec un niveau « Formation professionnelle supérieure » ne sont jamais imputées, car aucune règle d'imputation n'engendre un titre imputé pour ce niveau.
- Pour le niveau de formation « Hautes écoles », 12% des personnes ont un niveau imputé :
 - 53% de ces valeurs imputées concernant des personnes en formation continue (les formations continues relevées sont des MAS et exigent d'avoir obtenu au préalable un titre d'une haute école). Il est important de noter qu'au niveau de la formation continue des hautes écoles, seuls les enregistrements de type « élève » sont relevés.
 - 26% de ces valeurs imputées concernent des étudiants au niveau master pour qui on ne retrouve pas de titre de bachelor dans LABB. Ce sont en grande majorité des étudiants qui entrent au master avec un bachelor obtenu à l'étranger.
 - 20% de ces valeurs imputées concernent des personnes en formation doctorat pour qui on ne retrouve pas de titre de master dans LABB. Ce sont en grande majorité des étudiants qui entrent au doctorat avec un master obtenu à l'étranger.

Annexe 2 Analyses de la qualité du relevé structurel

Si on compare les réponses des personnes dans le RS avec les informations dans les différentes autres sources de données, on remarque qu'il y a des différences non négligeables, principalement concernant les catégories 3 et 4.

Dans le *Tableau 20* ci-dessous, se trouvent en colonne les sources en excluant le RS 2020 de ces sources et en ligne le RS 2020.

Tableau 20 : Comparaison du niveau de formation entre 4 sources ou plus concordantes (sans contradiction) et le relevé structurel 2020

RS	Niveau de formation des sources sans RS 2020					Total
	1	2	3	4	5	
-9	141	35	3	6	7	192
1	2 242	91	8	3	0	2 344
2	97	4 748	60	29	1	4 935
3	46	122	257	49	4	478
4	9	462	11	462	17	961
5	3	12	4	58	924	1 001
Total	2 538	5 470	343	607	953	9 911

Ce tableau met en évidence plusieurs points intéressants. La plupart des valeurs manquantes dans le RS (-9) correspondent à un niveau de formation 1 selon les autres sources, ce qui contribue à l'augmentation du nombre de personnes classées au niveau 1 dans le total des « autres sources » par rapport au RS. En examinant les personnes ayant déclaré un niveau 4 dans le RS, on observe que ce niveau est confirmé par les 4 autres sources pour 462 (case non-colorée) personnes sur 961. Cependant, pour 462 (case colorée) d'entre elles, les sources indiquent plutôt un niveau 2. De manière similaire, parmi les personnes ayant déclaré un niveau 3 dans le RS, seulement 257 sur 478 voient ce niveau confirmé par les 4 autres sources.

Dans le *Tableau 21* ci-dessous se trouve cette même analyse de comparaison des valeurs du RS les valeurs des 4 autres sources concordantes entre elles (sans source contradictoire) mais en pourcentage (par ligne).

Tableau 21 : Comparaison du niveau de formation entre 4 sources ou plus concordantes (sans contradiction) et le relevé structurel 2020, pourcentage par ligne

RS	Niveau de formation des sources sans RS 2020				
	1	2	3	4	5
-9	73.4	18.2	1.6	3.1	3.6
1	95.6	3.9	0.3	0.1	0.0
2	2.0	96.2	1.2	0.6	0.0
3	9.6	25.5	53.8	10.3	0.8
4	0.9	48.1	1.1	48.1	1.8
5	0.3	1.2	0.4	5.8	92.3

Le *Tableau 21* illustre clairement que, lorsqu'une personne déclare un niveau 3 ou 4 dans le RS, environ 50 % des cas montrent une contradiction avec les quatre sources cohérentes. Ces divergences entre le RS et les autres sources affectent principalement les totaux des niveaux de formation 3 et 4, mais également celui du niveau 2.

Il est important de souligner que nous avons appliqué des critères très stricts concernant les « sources », exigeant que 4 sources soient cohérentes et sans contradiction. En conséquence, seule une quantité limitée de données du RS est disponible pour la comparaison (9 911 personnes pour l'année 2020). Cette sous-population est peut-être trop restreinte pour permettre de tirer des conclusions.

Dans le *Tableau 22* et le *Tableau 23* ci-dessous, les critères sur les sources sont assouplis. Tous les cas, avec au moins 3 sources concordantes sans contradictions, sont pris en considération pour être comparés aux données du RS. La sous-population ainsi considérée est de 65 064 personnes.

Tableau 22 : Comparaison du niveau de formation entre 3 sources ou plus concordantes (sans contradiction) et le relevé structurel 2020

RS	Niveau de formation des sources sans RS 2020					Total
	1	2	3	4	5	
-9	831	295	10	38	34	1 208
1	11 698	777	28	11	6	12 520
2	753	28 999	262	347	14	30 375
3	314	841	1 351	330	46	2 882
4	49	3 446	80	5 174	147	8 896
5	35	60	41	663	8 384	9 183
Total	13 680	34 418	1 772	6 563	8 631	65 064

Tableau 23 : Comparaison du niveau de formation entre 3 sources ou plus concordantes (sans contradiction) et le relevé structurel 2020, pourcentage par ligne

RS	Niveau de formation des sources sans RS 2020				
	1	2	3	4	5
-9	68.8	24.4	0.8	3.1	2.8
1	93.4	6.2	0.2	0.1	0.0
2	2.5	95.5	0.9	1.1	0.0
3	10.9	29.2	46.9	11.5	1.6
4	0.6	38.7	0.9	58.2	1.7
5	0.4	0.7	0.4	7.2	91.3

En assouplissant les critères de sélection pour les données issues des sources, les mêmes conclusions persistent. Les valeurs manquantes dans le RS correspondent majoritairement à des personnes de niveau 1 d'après les autres sources. De plus, comme déjà observé avec les comparaisons basées sur 4 sources cohérentes, pour les personnes déclarant un niveau 3 ou 4 dans le RS les sources contredisent ce niveau dans 50 % des cas.

Si on élargit encore un peu l'échantillon, en reprenant cette fois les règles de reprises des sources décrites au chapitre 4.2, on constate que les mêmes conclusions peuvent être tirées alors que près de 230 000 personnes sont entrées dans l'analyse (Tableau 24 et Tableau 25).

Tableau 24 : Comparaison du niveau de formation reprise des sources selon les règles de concordance décrites au chapitre 4.2 (sans le RS 2020) et comparé au relevé structurel 2020

RS	Niveau de formation des sources sans RS 2020					Total
	1	2	3	4	5	
-9	3 034	1 421	98	215	160	4 928
1	38 498	3 801	259	71	54	42 683
2	4 125	90 095	1 154	1 944	71	97 389
3	1 814	4 201	6 476	1 723	366	14 580
4	355	13 467	579	21 824	1 002	37 227
5	268	331	507	4 020	27 472	32 598
Total	48 094	113 316	9 073	29 797	29 125	229 405

Tableau 25 : Comparaison du niveau de formation reprise des sources selon les règles de concordance décrites au chapitre 4.2 (sans le RS 2020) et comparé au relevé structurel 2020, pourcentage par ligne

RS	Niveau de formation des sources sans RS 2020				
	1	2	3	4	5
-9	61.6	28.8	2.0	4.4	3.2
1	90.2	8.9	0.6	0.2	0.1
2	4.2	92.5	1.2	2.0	0.1
3	12.4	28.8	44.4	11.8	2.5
4	1.0	36.2	1.6	58.6	2.7
5	0.8	1.0	1.6	12.3	84.3

En conclusion, que l'on applique la règle de concordance des sources présentée au chapitre 4.2 pour attribuer le niveau de formation, ou que l'on adopte des critères très stricts concernant ce qui est retenu (4 sources concordantes sans contradiction) pour les comparer au RS, les résultats obtenus sont finalement similaires. Cela confirme la pertinence de la règle de concordance des sources, mais soulève également des interrogations sur la qualité du RS pour la question sur la plus haute formation achevée.

Les résultats de l'analyse montrent que la qualité des résultats du RS est plus faible lorsque les répondants se situent aux niveaux 3 ou 4, que dans les autres niveaux. De plus, ces observations renforcent notre confiance dans les informations issues des sources concordantes plutôt que dans celles du RS, ce qui justifie notre approche consistant à privilégier les sources concordantes plutôt que le RS.

Annexe 3 Relevé structurel : la formulation de la question

Les modifications apportées à la formulation des différents niveaux de formation dans le questionnaire du RS en 2020 (voir l'*Image 1* et l'*Image 2*) ont influencé la façon dont les répondants ont émis leurs réponses.

Image 1 : Question sur la formation dans le questionnaire des relevés structurels de 2015 à 2019

8. Indiquez toutes les formations achevées dans la colonne a)... (cochez tous les niveaux terminés)	a) Formations achevées	b) Formation en cours
9. ... et la formation en cours dans la colonne b). (une seule réponse)		
1 aucune	1 ☐	1 ☐
2 jusqu'à maximum 7 ans d'école obligatoire	2 ☐	2 ☐
3 école obligatoire (école secondaire, cycle d'orientation, pré-gymnase, école spéciale)	3 ☐	3 ☐
4 1 an de préapprentissage, 1 an d'école de formation générale, 10 ^e année scolaire, 1 an d'école d'orientation professionnelle, année d'enseignement ménager, séjour linguistique (au minimum 1 an) avec diplôme, année d'engagement social, formation transitoire ou équivalente	4 ☐	4 ☐
5 2 à 3 ans d'école de degré diplôme, d'école d'administration et des transports, école de culture générale ECG ou formation équivalente	5 ☐	5 ☐
6 formation professionnelle élémentaire ou initiale (2 à 4 ans d'apprentissage ou d'école professionnelle à plein temps, école supérieure de commerce, école de métiers ou formation équivalente)	6 ☐	6 ☐
7 maturité gymnasiale, école normale (préparant à l'enseign. aux niveaux jardin d'enfants, école primaire, travaux manuels, économie ménagère)	7 ☐	7 ☐
8 maturité professionnelle ou maturité spécialisée	8 ☐	8 ☐
9 formation professionnelle supérieure avec brevet ou diplôme fédéral ou maîtrise, école technique ou autre école professionnelle, ET ou ES, ESGC (2 ans d'études à plein temps ou 3 ans à temps partiel)	9 ☐	9 ☐
10 école professionnelle supérieure (prédécesseur des hautes écoles spécialisées p. ex. ETS, ESCEA, ESSA, IES) y compris diplômes postgrades (3 ans d'études à plein temps ou 4 ans à temps partiel)	10 ☐	10 ☐
11 bachelor (université, EPF, haute école spécialisée, haute école pédagogique)	11 ☐	11 ☐
12 master, licence, diplôme, examen d'Etat, diplôme postgrade (université, EPF, haute école spécialisée, haute école pédagogique)	12 ☐	12 ☐
13 doctorat, habilitation	13 ☐	13 ☐

Image 2 : Question sur la formation dans le questionnaire des relevés structurels 2020 et suivants

8. Indiquez la formation la plus élevée que vous avez achevée dans la colonne a)...	a) Formation la plus élevée achevée	b) Formation en cours
9. ... et la formation en cours dans la colonne b).		
1 aucune	1 ☐	1 ☐
2 école obligatoire inachevée	2 ☐	2 ☐
3 école obligatoire	3 ☐	3 ☐
4 formation d'un an / offre de transition	4 ☐	4 ☐
5 école de culture générale / école de degré diplôme	5 ☐	5 ☐
6 apprentissage en entreprise (CFC, AFP) / formation professionnelle élémentaire / école professionnelle / école de commerce	6 ☐	6 ☐
7 maturité gymnasiale / brevet d'enseignement	7 ☐	7 ☐
8 maturité professionnelle / spécialisée	8 ☐	8 ☐
9 examen professionnel avec brevet fédéral / examen professionnel supérieur avec diplôme fédéral/maîtrise	9 ☐	9 ☐
10 école professionnelle supérieure (ES)	10 ☐	10 ☐
11 Bachelor: université, EPF, haute école spécialisée, haute école pédagogique (y compris diplôme HES/HEP)	11 ☐	11 ☐
12 Master: université, EPF, haute école spécialisée, haute école pédagogique (y compris licence/diplôme université / EPF)	12 ☐	12 ☐
13 doctorat / habilitation	13 ☐	13 ☐

Dans le *Tableau 26*, le *Tableau 27* et le *Tableau 28* ci-dessous, nous comparons les réponses dans le RS à des sources concordantes de la priorisation des sources (chapitre 4.2), en excluant les réponses qui ont été corrigées par LABB lors de la plausibilisation (chapitre 4.1).

Tableau 26 : Comparaison entre le relevé structurel et les sources concordantes de la priorisation de sources (sans le RS de l'année de référence), en excluant les données du relevé structurel corrigées par LABB, 2020

Niveau du RS	Niveau selon la concordance de sources (Fréq.)					Total	Niveau selon la concordance des sources (% par ligne)				
	1	2	3	4	5		1	2	3	4	5
1	38 498	3 584	209	52	54	42 397	90.8	8.5	0.5	0.1	0.1
2	4 125	90 095	541	1 400	70	96 231	4.3	93.6	0.6	1.5	0.1
3	1 814	4 201	6 476	1 342	365	14 198	12.8	29.6	45.6	9.5	2.6
4	355	13 467	579	21 824	1 002	37 227	1.0	36.2	1.6	58.6	2.7
5	268	331	507	4 020	27 472	32 598	0.8	1.0	1.6	12.3	84.3
Total	45 060	111 678	8 312	28 638	28 963	222 651					

Tableau 27 : Comparaison entre le relevé structurel et les sources concordantes de la priorisation de sources (sans le RS de l'année de référence), en excluant les données du relevé structurel corrigées par LABB, 2019

Niveau du RS	Niveau selon la concordance de sources (Fréq.)					Total	Niveau selon la concordance des sources (% par ligne)				
	1	2	3	4	5		1	2	3	4	5
1	33 091	4 426	190	93	57	37 857	87.4	11.7	0.5	0.2	0.2
2	3 638	76 546	462	1 338	73	82 057	4.4	93.3	0.6	1.6	0.1
3	1 560	5 798	5 195	745	269	13 567	11.5	42.7	38.3	5.5	2.0
4	220	6 127	709	19 510	1 382	27 948	0.8	21.9	2.5	69.8	4.9
5	185	309	447	1 625	23 219	25 785	0.7	1.2	1.7	6.3	90.0
Total	38 694	93 206	7 003	23 311	25 000	187 214					

Tableau 28 : Comparaison entre le relevé structurel et les sources concordantes de la priorisation de sources (sans le RS de l'année de référence), en excluant les données du relevé structurel corrigées par LABB, 2018

Niveau du RS	Niveau selon la concordance de sources (Fréq.)					Total	Niveau selon la concordance des sources (% par ligne)				
	1	2	3	4	5		1	2	3	4	5
1	31 207	4 272	161	81	65	35 786	87.2	11.9	0.4	0.2	0.2
2	3 781	72 477	487	1 295	90	78 130	4.8	92.8	0.6	1.7	0.1
3	1 609	5 619	4 951	767	266	13 212	12.2	42.5	37.5	5.8	2.0
4	407	5 693	665	17 652	1 229	25 646	1.6	22.2	2.6	68.8	4.8
5	476	299	453	1 520	21 462	24 210	2.0	1.2	1.9	6.3	88.6
Total	37 480	88 360	6 717	21 315	23 112	176 984					

Dans les tableaux ci-dessus (Tableau 26, Tableau 27 et Tableau 28), nous constatons que les réponses pour les années 2018 et 2019 se comportent de manière similaire. Alors qu'en 2020, après le changement dans le libellé des niveaux de formation dans le questionnaire du RS, des différences dans les réponses se remarquent. Notamment, en 2020, il y a plus de cohérences entre le RS et les sources pour les niveaux 3 (env. 45% en 2020, contre env. 38% en 2018 et 2019). A l'inverse, les niveaux 4 sont moins cohérents en 2020 (env. 59% en 2020, contre env. 69% en 2018 et 2019). Nous pouvons également constater une légère amélioration de la cohérence pour le niveau 1 en 2020 et une légère détérioration pour le niveau 5.

Annexe 4 Comparaison de différentes imputations

Pour choisir la méthode pour imputer le niveau de formation sur la base des probabilités $\hat{P}(1), \hat{P}(2), \hat{P}(3), \hat{P}(4), \hat{P}(5)$ calculées grâce à deux régressions multinomiales (chapitre 0.456.1), les trois méthodes d'imputations suivantes ont été testées et comparées :

1. Maximum de vraisemblance : imputer le niveau $\operatorname{argmax}_{i \in \{1,2,3,4,5\}} (\hat{P}_k(i))$,
2. Optimisation : méthode décrite au chapitre 0,
3. Aléatoire : imputer aléatoirement en fonction des probabilités. Par exemple, si la personne k a les probabilités estimées par $\hat{P}_k(1) = 0.1, \hat{P}_k(2) = 0.1, \hat{P}_k(3) = 0.1, \hat{P}_k(4) = 0.1, \hat{P}_k(5) = 0.6$, alors la personne aura 10% de chance d'être imputée 1, 10% de chance d'être imputée 2, 10% de chance d'être imputée 3, 10% de chance d'être imputée 4 et 60% d'être imputée de niveau 5.

La première méthode par maximum de vraisemblance a l'avantage d'être meilleure sur les imputations individuelles (si le modèle est correct), mais le désavantage d'être biaisée sur les totaux (*Tableau 31*). La méthode aléatoire a l'avantage de ne pas être biaisée sur les totaux (si le modèle est correct), mais d'être moins bonne pour les imputations individuelles. La méthode d'optimisation que nous utilisons dans ce projet, est une méthode de compromis entre ces deux critères (voir *Tableau 31* ci-dessous). En effet, elle s'éloigne peu du maximum de vraisemblance et c'est pourquoi sa qualité individuelle en est proche. En même temps, elle permet de réduire les biais sur des domaines choisis (dans notre cas : le croisement entre les variables « catégorie d'âge » et « sexe »).

Le calcul de l'estimation de l'erreur individuelle pour les trois méthodes testées est présenté dans le *Tableau 29*.

Tableau 29 : Formule d'estimation de l'erreur individuelle pour les trois méthodes d'imputation

	Méthodes d'imputations	
	Optimisation et Maximum de vraisemblance	Aléatoire
Formule d'estimation de l'erreur individuelle	$\sum_{k \in U} 1 - \hat{P}_k(i_k),$ où $i_k \in \{1, 2, 3, 4, 5\}$ est la valeur imputée pour la personne k.	$\sum_{k \in U} \sum_{i=1}^5 \hat{P}_k(i)(1 - \hat{P}_k(i))$

Comme la méthode « aléatoire » ne produit pas toujours les mêmes résultats, en vue des comparaisons, nous calculons ici son espérance. En effet, le résultat d'une réalisation de la méthode aléatoire peut être meilleur ou moins bon. Dans le *Tableau 30* se trouve la répartition du niveau de formation selon les trois méthodes d'imputation ainsi que selon la somme des probabilités.

Tableau 30 : Répartition du niveau de formation des trois méthodes d'imputations et de la somme des probabilités

		Méthodes d'imputations			
		Somme des probabilités (MLE)	Maximum de Vraisemblance	Aléatoire (une réalisation)	Optimisation
Niveau de formation	1	18.85	21.06	18.84	18.85
	2	27.57	33.22	27.58	27.43
	3	11.95	6.45	11.91	12.03
	4	12.92	5.99	12.95	12.90
	5	28.70	33.29	28.71	28.80

Tableau 31 : Comparaison des trois méthodes d'imputations avec les indicateurs de performance au niveau individuel et sur les totaux

		Méthodes d'imputations		
		Maximum de Vraisemblance	Aléatoire	Optimisation
Distance euclidienne avec la somme des probabilités		94 143	Moy. de 100 réal. : 603	1 531
Estimation d'erreurs individuelles		37.3%	En espérance : 50.3%	39%

Comme attendu, les totaux obtenus avec la méthode du maximum de vraisemblance s'écartent des totaux estimés par la somme des probabilités. La méthode aléatoire présente un taux d'erreur individuelle inférieur de 13% par rapport au maximum de vraisemblance (soit une perte de qualité de 21%). La méthode du maximum de vraisemblance est la meilleure pour les imputations individuelles, tandis que la méthode aléatoire est la plus adaptée pour obtenir des totaux proches de ceux estimés par la somme des probabilités. La méthode d'optimisation, quant à elle, est calée sur des totaux estimés par les probabilités. Dans le *Tableau 31* ci-dessus, pour des raisons de précision numérique du maillage, une distance de 1 531 est observée entre les totaux optimisés et ceux définis par la somme des probabilités, bien que, théoriquement, cette distance devrait être nulle. Cet écart est jugé négligeable comparé à la distance de 94 143 observée avec la méthode du maximum de vraisemblance. Enfin, avec une perte de performance de seulement 1,7% sur l'erreur individuelle et un biais réduit sur les totaux clés, cette méthode d'optimisation semble être un bon compromis.

Dans le *Tableau 32*, le *Tableau 33* et le *Tableau 34* ci-dessous se trouvent les matrices de confusion entre l'ensemble d'apprentissage basé sur les réponses du RS et les trois méthodes d'imputation. Comme au chapitre 4.4.3, l'ensemble d'apprentissage est imputé selon la méthode d'imputation testée et le résultat nous permet de voir où sont les confusions.

Tableau 32 : Matrice de confusion sur l'ensemble d'apprentissage pondéré entre les réponses du relevé structurel et l'imputation par la méthode d'imputation « Optimisation », en 2020

Niveau selon la méthode « Optimisation »										
Niveau du RS	%					% par ligne				
	1	2	3	4	5	1	2	3	4	5
1	9.6	2.5	2.2	0.6	1.3	59.7	15.5	13.4	3.5	7.8
2	2.1	19.4	2.4	4.4	1.2	7.2	65.9	8.0	14.8	4.1
3	1.7	2.2	4.1	1.6	1.9	14.6	19.5	35.6	14.0	16.4
4	0.5	4.5	1.1	5.8	2.0	3.7	32.1	8.0	41.8	14.4
5	1.4	0.9	2.1	2.3	22.2	4.9	3.1	7.2	8.0	76.9

Tableau 33 : Matrice de confusion sur l'ensemble d'apprentissage pondéré entre les réponses du relevé structurel et l'imputation par la méthode d'imputation « Maximum de vraisemblance », en 2020

Niveau selon la méthode « Maximum de vraisemblance »										
Niveau du RS	%					% par ligne				
	1	2	3	4	5	1	2	3	4	5
1	10.4	2.9	1.0	0.1	1.7	64.9	18.0	6.0	0.7	10.4
2	2.4	22.6	1.2	1.4	1.9	8.3	76.5	3.9	4.8	6.4
3	2.3	3.2	2.6	0.7	2.7	19.7	27.9	22.7	5.9	23.8
4	0.7	6.2	0.6	3.7	2.8	4.7	44.2	4.4	26.7	19.9
5	1.6	1.4	0.9	0.9	24.0	5.4	4.9	3.3	3.2	83.2

Tableau 34 : Matrice de confusion sur l'ensemble d'apprentissage pondéré entre les réponses du relevé structurel et l'imputation par la méthode d'imputation « Aléatoire », en 2020

Niveau selon la méthode « Aléatoire »										
Niveau du RS	%					% par ligne				
	1	2	3	4	5	1	2	3	4	5
1	7.9	3.2	2.2	1.0	1.9	49.0	19.8	13.3	6.2	11.6
2	3.1	16.2	2.9	5.1	2.2	10.5	55.0	9.8	17.2	7.5
3	2.1	2.9	2.7	1.5	2.4	18.0	24.9	23.9	12.8	20.5
4	1.0	5.0	1.4	4.0	2.6	7.0	35.7	10.3	28.8	18.2
5	2.0	2.2	2.4	2.5	19.8	6.8	7.7	8.4	8.6	68.5

Dans le Tableau 32, le Tableau 33 et le Tableau 34 ci-dessus, les cases vertes indiquent pour chaque niveau la meilleure correspondance des trois méthodes, entre les valeurs imputées et le RS. La moins bonne correspondance est en rouge et la correspondance intermédiaire n'a pas de couleur. Les cases en violet mettent en avant le fait que les méthodes « Optimisation » et « Aléatoire » tendent à créer des

symétries entre les confusions (autant de confusions 2 vers 4 que de confusions 4 vers 2), ce qui a pour conséquence d'avoir des totaux, suite à l'imputation, proches des totaux estimés par le RS (la somme par ligne est proche de la somme par colonne pour chaque niveau). Ces symétries ne se retrouvent pas pour la méthode « Maximum de vraisemblance ».

Annexe 5 Variables auxiliaires pour les arbres de décision

Dans le *Tableau 35* se trouve la liste des variables auxiliaires utilisées dans les arbres de décision (chapitre 4.3.3).

Tableau 35 : Liste des variables utilisées dans les arbres de décision

Arbre ⁹	Variable	Cat ¹⁰	Description
Tous	Sex	1	Sexe : homme, femme
1	AgePremierEnfant	0	Âge à la naissance du premier enfant
1, 5, 6, 7	CLASSOFMARITALSTATUS	1	4 modalités (célibataire, marié/partenaire, veuf, divorcé/partenariat dissous)
1	nb_enfant	0	Nombre d'enfants
1, 5	HOUSEHOLDTYP25AGGR	1	Type de ménage en 5 modalités (une personne, couple sans enfant, couple avec enfant(s) de moins de 25 ans, parent seul avec enfant(s) de moins de 25 ans, autres)
1, 2, 4, 5, 8	STALAN	1	Type de commune (urbaine, rurale, intermédiaire)
Tous	AGE	0	Âge de la personne
1, 2, 4, 5, 8	DEGURB	0	Type de commune (Zone densément peuplée, Zone urbanisée intermédiaire, Zone faiblement peuplée)
1	POPULATIONGROUP	1	Type de population (Suisse né en Suisse, Suisse né à l'étranger, étranger né en Suisse, étranger né à l'étranger)
1, 5, 8	Job_avant20	1	A eu un revenu entre 18 ans et 20 ans
2, 4	STALAN_birth	1	Type de la commune de naissance (urbaine, rurale, intermédiaire)
2	r_tot_log	0	Log du revenu annuel total (CdC)
2	WAZIM	0	Nombre de pièces dans le logement
2	WAREA	0	Surface du logement
2, 4, 9	nb_logement	1	Nombre de logements dans le bâtiment (1, 2, 3+)
2, 6	Warea_ppers_log	0	Log de la surface par personne dans le logement
2	r_sal_log	0	Log du revenu salarial annuel total (CdC)
2, 4	DEGURB_birth	1	Type de la commune de naissance (Zone densément peuplée, Zone urbanisée intermédiaire, Zone faiblement peuplée)
3, 6, 7, 9	HOUSEHOLDTYP25	1	Type de ménage en 9 modalités (une personne, couple sans enfant (marié de sexe opposé, union libre de sexe opposé, de même sexe), couple avec enfant(s) de moins de 25 ans (marié de sexe opposé, union libre de sexe opposé, de même sexe), parent seul avec enfant(s) de moins de 25 ans, autres)
3, 7, 9	FOREIGNHOUSEHOLDPERM	1	Composition du ménage en 6 modalités (tous suisses, suisses-étrangers de même nationalité, suisses-étrangers de nationalités différentes, étrangers de même nationalité, étrangers de nationalités différentes)

⁹ Numéro des arbres dans lesquelles la variable est utilisée (10 arbres en tout).

¹⁰ Variable catégorielle : 1 = oui, 0 = non.

3	COUNTOFPERSON- PERM	0	Nombre de personnes dans le ménage (population résidente permanente au domicile principal)
3	compBasics	1	Composition du ménage (homme seul, femme seule, deux adultes de sexes différents, deux adultes de même sexe, trois adultes ou plus, adulte seul avec mineur(s), deux adultes avec mineur(s), trois adultes ou plus avec mineur(s))
3	HHBYNATIOPLACE- BIRTHPERM	1	Composition du ménage (tous suisses nés en Suisse, tous suisses nés à l'étranger, tous suisses dont au moins une personne née en Suisse et une personne née à l'étranger, tous étrangers nés en Suisse, tous étrangers nés à l'étranger, tous étrangers dont au moins une personne née en Suisse et une personne née à l'étranger, tous nés en Suisse avec au moins une personne de nationalité suisse et étrangère, tous nés à l'étranger avec au moins une personne de nationalité suisse et étrangère, au moins une personne de nationalité suisse et étrangère née en Suisse et à l'étranger)
3	COUNTOFPER- SONUNDER18	0	Nombre de personnes mineures dans le ménage
3, 4, 5, 6, 7	RESIDPERMIT	1	Permis de séjour (Suisse, C, diplomates ou fonctionnaires internationaux avec leurs conjoints ou permis B UE/AELE, diplomates ou fonctionnaires internationaux avec leurs conjoints ou permis B hors UE/AELE ou permis F ou permis L ou permis N)
3, 6, 7, 9	justArrived	1	Durée de séjour (<=1 an, 1-3 ans, 3-7 ans, >12 ans, toujours)
4, 8	cat_rtot	1	Catégorie de revenu annuel CdC (<1, 1-20 000, 20 000-60 000, 60 000-90 000, 90 000-120 000, 120 000-170 000, 170 000-240 000, >240 000)
4, 6, 7, 9	Naturalise	1	Naturalisé depuis 2010 (oui, non)
5	Cat_ageJob	1	Age du premier salaire plus élevé qu'un seuil ¹¹
5	cat_sal_age	1	Croisement entre classe de salaire et âge. Si une modalité est trop petite, deux modalités avec un salaire proche sont fusionnées.
5, 9	noga_grp	1	Branche de l'entreprise (NOGA regroupé, automatiquement et manuellement)
5	ageJob	0	Âge du premier salaire plus élevé qu'un seuil
6, 8, 9	Age_sal_noga_&niv. ¹²	1	Croisement entre classe de salaire, âge et la branche de l'entreprise (NOGA), défini automatiquement par un arbre de décision.
6, 8	f_multi_income	1	Plusieurs emplois (oui, non)
6, 8, 9	emptotClass	1	Classe d'emplois selon le nombre d'emplois de l'entreprise (emploi)
6, 8	pub_orPriv	1	Entreprise publique ou privée (emploi)

¹¹ Seuil défini chaque année en fonction de l'inflation.

¹² &niv. : correspond à l'agrégation de la première ou de la deuxième régression multinomiale.

6, 7	Nat	1	Nationalité (agrégation)
6, 7	Birth	1	Pays de naissance (agrégation)
6, 8	nb_income	0	Nombre de revenu
8, 9	emptot	0	Nombre d'emplois dans l'entreprise (emploi)
8	flag_emptot	1	Emptot > 0 : oui ou non
8, 9	indic_statent	1	La personne est dans la STATENT
9	Voyage	1	Nombre de voyage > 30 jours depuis 2011
9	cat_sal	1	Catégorie de salaire annuel CdC (<1, 1-20 000, 20 000-50 000, 50 000-80 000, 80 000-110 000, 110 000-180 000, 180 000-250 000, >250 000)
10	niv_mere	1	Niveau de la plus haute formation achevée de la mère (selon SECO, LABB, SAS, RS ou SNC)
10	niv_pere	1	Niveau de la plus haute formation achevée du père (selon SECO, LABB, SAS, RS ou SNC)
10	niv_conjoint	1	Niveau de la plus haute formation achevée du conjoint (selon SECO, LABB, SAS, RS ou SNC)

Les arbres de décision permettent d'optimiser automatiquement des croisements entre des groupements de variables pour expliquer le niveau de formation (dans notre cas 10 groupements présentés dans le *Tableau 35*). Ces 10 arbres sont créés pour toutes les sous-populations (chapitre 0.456.1). Un 11^{ème} arbre, exploitant les mêmes variables que les régressions multinomiales, est aussi défini (chapitre 4.3.3). Les 11 mêmes arbres sont également créés en ajoutant la variables source de l'ESS pour les sous-populations pour lesquelles le niveau de formation est disponible. Ces arbres de décision sont paramétrés pour ne pas créer de catégories de moins de 1% de l'ensemble. Un algorithme de pruning (cost-complexity) est utilisé pour éviter le surapprentissage.

Annexe 6 Analyse du relevé structurel pour les personnes en cours d'étude

Cette analyse est faite sur la population des personnes de nationalité suisse, nées en Suisse et n'ayant pas quitté la Suisse depuis le 31.12.2010. Cette population est choisie pour éviter notamment les parcours de formations atypiques comme ceux des migrants ou les formations à l'étranger. L'analyse porte sur les personnes ayant répondu au RS 2020 et ayant une information dans LABB. Les personnes de niveau 1 dans LABB sont exclues de l'analyse.

La population de l'analyse est séparée en deux groupes, ceux qui sont en cours d'étude en 2020 et ceux ne le sont pas (*Tableau 36*).

Tableau 36 : Comparaison du niveau de formation entre LABB et RS en % en fonction des groupes

	RS ≠ LABB	RS = LABB	Total	
En cours d'étude	13.28	86.72	39599	36.54%
Pas en cours d'étude	10.27	89.73	54201	50.01%
Total	10 825	82 975	93 800	100%
	11.54%	88.46%	100%	

Dans les graphiques (*Graphique 8*, *Graphique 9* et *Graphique 10*) et tableaux (*Tableau 37*, *Tableau 39* et *Tableau 40*) ci-dessous, les données présentées concernent les âges pour lesquels, dans chaque groupe comparé, plus de 1% des personnes sont représentées.

Dans le *Graphique 8* et le *Tableau 37* ci-dessous, les données du RS sont comparées à celles de LABB afin de détecter une éventuelle différence de qualité dans le RS entre les personnes « en cours d'étude » et celles « pas en cours d'étude ». Nous émettons l'hypothèse que les personnes en cours d'étude commettent généralement plus d'erreurs dans le RS que les autres, potentiellement en raison d'une anticipation de leur diplôme.

Graphique 8 : Taux de correspondance du niveau de formation entre le RS et LABB selon l'âge

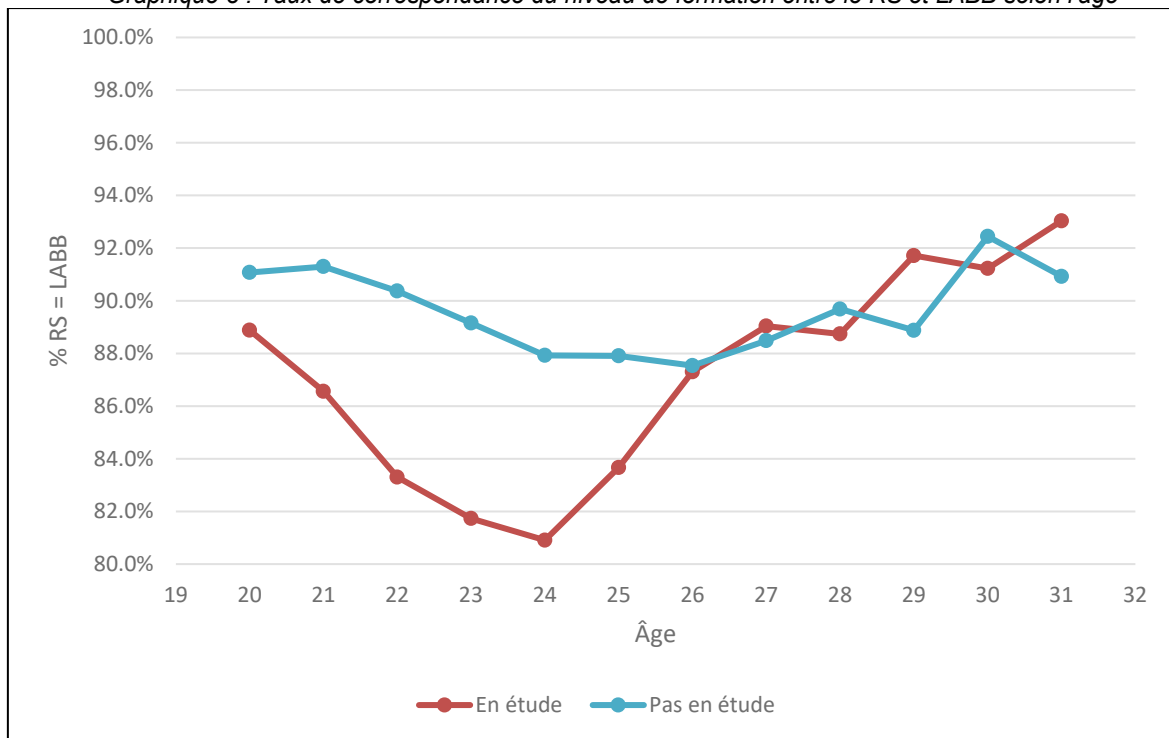


Tableau 37 : Taux de correspondance du niveau de formation entre le RS et LABB selon l'âge

Âge	En étude			Pas en étude		
	N	Mean	Std	N	Mean	Std
20	5288	88.9%	0.4%	1826	91.1%	0.7%
21	4666	86.6%	0.5%	2758	91.3%	0.5%
22	4438	83.3%	0.6%	2835	90.4%	0.6%
23	4056	81.7%	0.6%	2867	89.2%	0.6%
24	3593	80.9%	0.7%	2932	87.9%	0.6%
25	2792	83.7%	0.7%	3126	87.9%	0.6%
26	2205	87.3%	0.7%	3322	87.5%	0.6%
27	1560	89.0%	0.8%	3526	88.5%	0.5%
28	1190	88.7%	0.9%	3481	89.7%	0.5%
29	857	91.7%	0.9%	3012	88.9%	0.6%
30	604	91.2%	1.2%	2620	92.4%	0.5%
31	445	93.0%	1.2%	2314	90.9%	0.6%

La courbe des personnes en cours de formation s'éloigne de celle des personnes pas en cours de formation entre 20 et 26 ans. Durant cet intervalle de temps, les réponses dans le RS et dans LABB ont plus de divergences pour les personnes en cours de formation alors que dans les autres classes d'âge les différences semblent similaires, que les personnes soient en cours de formation ou non.

Nous avons essayé de montrer que les différences étaient liées à une anticipation de diplôme, mais les résultats obtenus ne l'ont pas confirmé (Tableau 38). Pour les personnes de moins de 30 ans, parmi les divergences entre LABB et RS, on trouve :

Tableau 38 : Type de divergences de niveau de formation entre les données du RS et de LABB, en 2020

	Niveau de formation RS < LABB	Niveau de formation RS > LABB
En cours d'étude	36.30%	63.70%
Pas en cours d'étude	33.46%	66.54%

On note cependant que les personnes en formation ont plus de divergences entre le RS et LABB que celles qui ont fini leur formation.

On pourrait aussi se demander si c'est correct de supposer que les personnes en cours d'étude de notre population d'analyse (nationalité suisse, né en Suisse, pas quitté la Suisse depuis le 31.12.2010) étendu aux personnes en cours d'étude dans toute la population se comportera de la même manière. Les résultats se trouvent dans le *Graphique 9*, le *Graphique 10*, le *Tableau 39* et le *Tableau 40*.

Dans *Graphique 9* et le *Tableau 39* ci-dessous, les données du RS sont comparées à celles de LABB afin de détecter pour les personnes en cours d'étude une éventuelle différence de qualité entre la population d'analyse et toute la population.

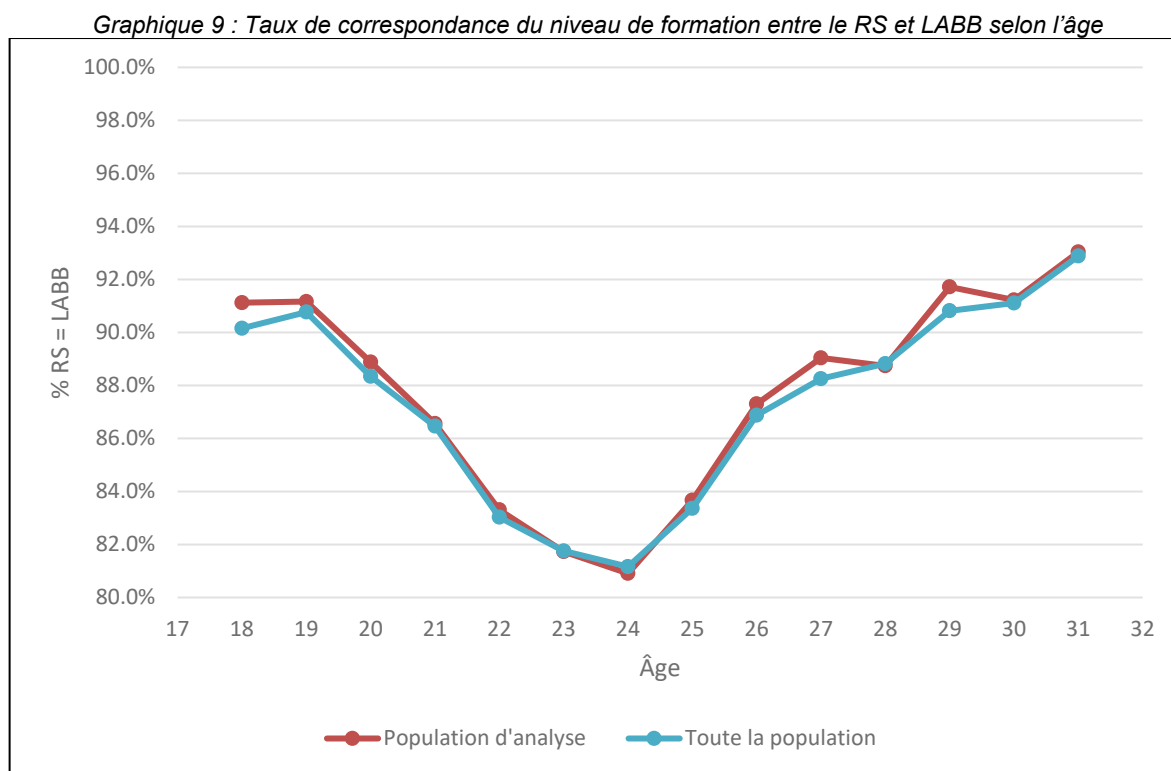


Tableau 39 : Taux de correspondance du niveau de formation entre le RS et LABB selon l'âge

Âge	Population d'analyse			Toute la population		
	N	Mean	Std	N	Mean	Std
18	1329	91.1%	0.8%	1615	90.2%	0.7%
19	4233	91.2%	0.4%	4939	90.8%	0.4%
20	5288	88.9%	0.4%	6218	88.3%	0.4%
21	4666	86.6%	0.5%	5518	86.5%	0.5%
22	4438	83.3%	0.6%	5261	83.0%	0.5%
23	4056	81.7%	0.6%	4807	81.8%	0.6%
24	3593	80.9%	0.7%	4210	81.2%	0.6%
25	2792	83.7%	0.7%	3324	83.4%	0.6%
26	2205	87.3%	0.7%	2621	86.9%	0.7%
27	1560	89.0%	0.8%	1907	88.3%	0.7%
28	1190	88.7%	0.9%	1476	88.8%	0.8%
29	857	91.7%	0.9%	1067	90.8%	0.9%
30	604	91.2%	1.2%	810	91.1%	1.0%
31	445	93.0%	1.2%	604	92.9%	1.0%

On note que ces deux courbes sont effectivement très similaires. C'est peut-être dû au fait que la grande majorité des personnes de toute la population est aussi dans la population d'analyse. Le *Graphique 10* et le *Tableau 40* présentent cette même analyse, mais en comparant la population d'analyse à sa population complémentaire (population totale sans la population d'analyse).

Graphique 10 : Taux de correspondance du niveau de formation entre le RS et LABB selon l'âge

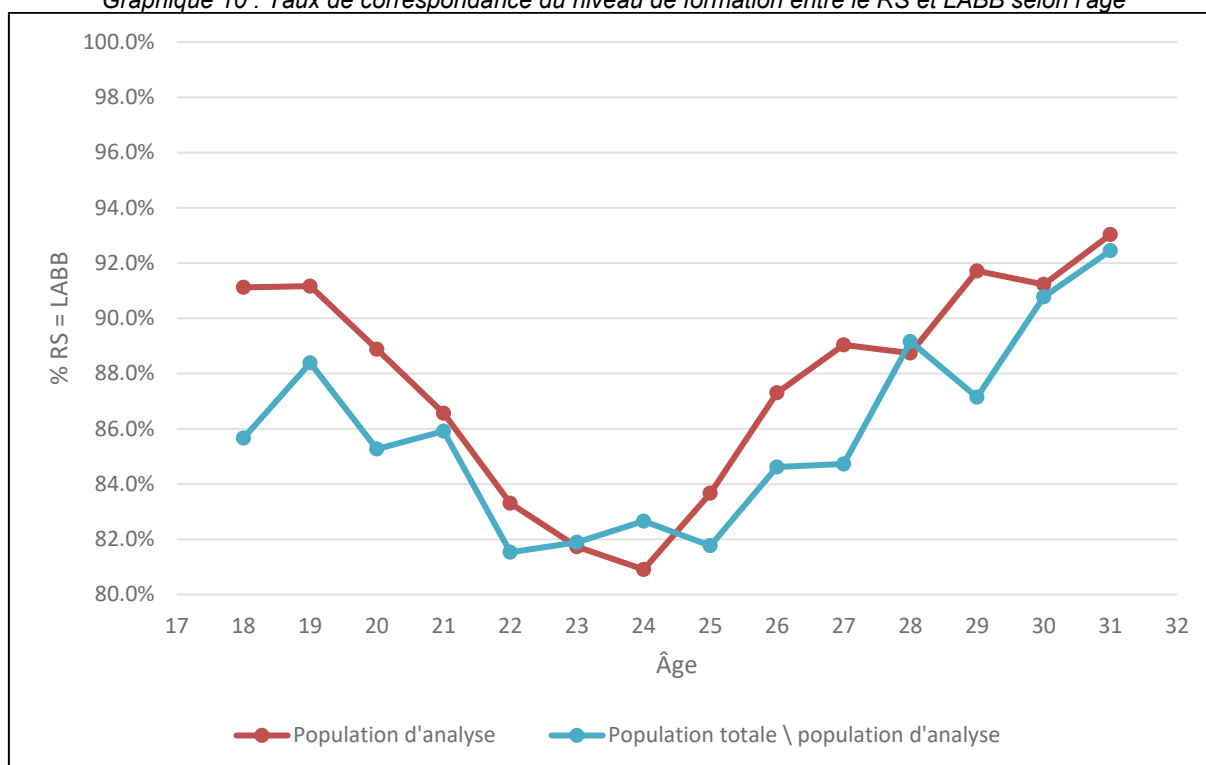


Tableau 40 : Taux de correspondance du niveau de formation entre le RS et LABB selon l'âge

Âge	Population d'analyse			Population totale \ Population d'analyse		
	N	Mean	Std	N	Mean	Std
18	1329	91.1%	0.8%	286	85.7%	2.1%
19	4233	91.2%	0.4%	706	88.4%	1.2%
20	5288	88.9%	0.4%	930	85.3%	1.2%
21	4666	86.6%	0.5%	852	85.9%	1.2%
22	4438	83.3%	0.6%	823	81.5%	1.4%
23	4056	81.7%	0.6%	751	81.9%	1.4%
24	3593	80.9%	0.7%	617	82.7%	1.5%
25	2792	83.7%	0.7%	532	81.8%	1.7%
26	2205	87.3%	0.7%	416	84.6%	1.8%
27	1560	89.0%	0.8%	347	84.7%	1.9%
28	1190	88.7%	0.9%	286	89.2%	1.8%
29	857	91.7%	0.9%	210	87.1%	2.3%
30	604	91.2%	1.2%	206	90.8%	2.0%
31	445	93.0%	1.2%	159	92.5%	2.1%

On note qu'effectivement ces deux populations se comportent de manière similaire. Par conséquent, pour toute la population des personnes en cours d'étude, on constate que cette population semble faire plus d'erreurs en répondant au RS. C'est pour cette raison que les personnes en cours d'étude sont corrigées par les informations de LABB.