



Highest level of education completed (STATPOP)

Production process – short description

EXPERIMENTAL STATISTICS



Neuchâtel, 2025

Published by:	Federal Statistical Office (FSO)	Original text:	French
Information:	eduperspectives@bfs.admin.ch	Layout concept:	Section PUB
Editors:	Sabrina Bravo, FSO ; Pascal Strubi, FSO	Downloads:	www.statistics.ch
Contents:	Sabrina Bravo, FSO ; Pascal Strubi, FSO	Copyright:	FSO, Neuchâtel 2025
Topic:	Statistical basis and overviews		Reproduction with mention of source authorised (except for commercial purposes)

Introduction

The population's highest level of education completed serves as crucial indicator for measuring a nation's human capital in terms of qualifications. In the context of demographic shifts such as an ageing population, accurate analysis and monitoring of trends in educational attainment are essential if countries are to remain competitive and meet the increasing demand for skilled labour. To conduct multi-dimensional analyses, it is essential to have data on the educational level of each member of the permanent resident population. To compile such data, a process needs to be put in place that combines existing data from federal administrative statistics and surveys and uses an imputation model to complete the missing datapoints. This variable can be used for statistical analysis at the individual level when combined with other data, but cannot be interpreted at the individual level. The level of education nomenclature can be broken down into 3 or 5 levels (Table 1).

Table 1 : Level of education nomenclature in 3 and 5 categories

Levels of education		
3 levels	5 levels	Educational programmes included in the 5-level nomenclature
Compulsory education	1	Compulsory education (in progress or completed) / none / one year / bridging course
Upper-secondary education	2	In-company apprenticeship (federal VET diploma, federal VET certificate) / basic vocational education / vocational school / upper-secondary level commercial school
	3	upper secondary specialised school / academic baccalaureate school / teaching diploma / vocational or specialised baccalaureate
Tertiary level	4	Advanced federal professional examination with diploma / advanced federal professional examination with diploma / federal professional master diploma
	5	University / federal institute of technology / university of applied sciences / university of teacher education

Data sources

This project uses a wide range of data sources each year to determine the level of education of the Swiss permanent resident population: the Population and Household Statistics (STATPOP), the Structural Surveys (RS) since 2010, the Longitudinal Analyses in the Field of Education (LABB) in addition to the School Staff Statistics (SSP) and the Statistics on the Staff of Higher Education Institutions (SHIS-PERS), the Social Assistance Statistics (SAS), the State Secretariat for Economic Affairs (SECO) unemployment statistics, the Swiss Earnings Structure Survey (ESS) and the Swiss National Cohort (SNC) based on the 1990 and 2000 censuses.

Data preparation and deterministic imputation

First, the data from each source undergoes a validation process (unvalidated LABB data will be labelled LABB_0). This involves deleting data that is potentially obsolete, for example if the person has returned to education or if the person left Switzerland at an age when they could have returned to education, or if the person claims to have a tertiary level of education when they are under the age of 18. In addition, errors in the source data are corrected using LABB data if the person's level of education in LABB_0 is higher or if the person is currently in education and their current course of study since their last completed qualification is listed in the LABB data.

Once the data sources have been validated, they undergo a prioritisation process. People **under the age of 15** are assigned level 1 unless LABB_0 indicates a higher level. The LABB data is then retained if the person listed in the LABB data is currently in education and their current course of study since their last completed qualification is known, in addition to people who have reached a maximum level in LABB (this subset of LABB will be labelled **LABB***). The level of education is then determined based on

concordances in the source data (if the most frequent level of education indicated in the sources has at least two more sources than the 2nd most frequent level). At this stage, the LABB and SECO sources derived from administrative data are retained, firstly the person's education data according to the **SECO** definition and secondly the LABB data not yet used, i.e. without LABB* (**LABB\LABB***), because the SECO definition includes education obtained abroad. Finally, the Social Assistance Statistics (SAS), Structural Survey (RS) and Swiss National Cohort (SNC) data, in which the level of education data was provided by the person, is used. The **SAS** data is prioritised here because the person provided the information to an interlocutor who was able to help them determine their level of education. Next comes the **RS** data, followed by the **SNC** data, which is the oldest.

Table 2: Frequency of aggregated attribution rules for 2020 level of education

Attribution rules	Frequency	%	Cumulative frequency %
<15	1 306 271	15.1	15.1
LABB*	1 265 351	14.6	29.7
Concordances	2 852 961	32.9	62.6
SECO	556 972	6.4	69.0
LABB	269 333	3.1	72.1
SAS	60 549	0.7	72.8
RS	851 471	9.8	82.6
SNC	700 775	8.1	90.7
Missing	806 617	9.3	100.0
Total	8 670 300	100.0	100.0

As Table 2 shows for 2020, a level of education has been assigned to 90.7% of the population from various sources. This leaves 9.3% of the population whose level of education must be imputed using the algorithm described below.

Model-based imputation

Structural Survey (RS) data is used to impute the level of education of individuals who have not yet been attributed a level. To define the training set for the imputation model, we applied to the reference year's Structural Survey, the same attribution process as applied to the entire population, but excluding data from the reference year's Structural Survey. Our training set is made up of all the people in the Structural Survey who were not attributed a level of education from other data sources. Variables from various registers are used for imputation, including variables from the Central Compensation Office (e.g. salaries), Structural Business Statistics (e.g. NOGA), STATPOP population stock (e.g. gender, nationality, age), STATPOP household data (e.g. household composition), the Buildings and Dwellings Statistics (e.g. apartment blocks/single-family homes), etc.

The analyses have shown that it is difficult to distinguish between levels 2, 3 and 4, so the imputation procedure is carried out in two stages: an initial multinomial regression provides an estimate of the probability of level 1, 234 or 5 (labelled P_A), followed by a second multinomial regression to distinguish between levels 2, 3 and 4 (labelled P_B). Thus for any person k in the population, we calculate the probability of each level:

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

These probabilities are then checked for plausibility: the level must be equal to or higher than LABB_0, and the person must be over 18 to have a tertiary level of education. Once the probabilities have been estimated, an optimisation algorithm is applied to impute a level of education for each person by minimising the expected error (maximising the likelihood), while respecting the totals estimated by summing the probabilities for each level of education based on sub-populations grouped by age and gender combined.

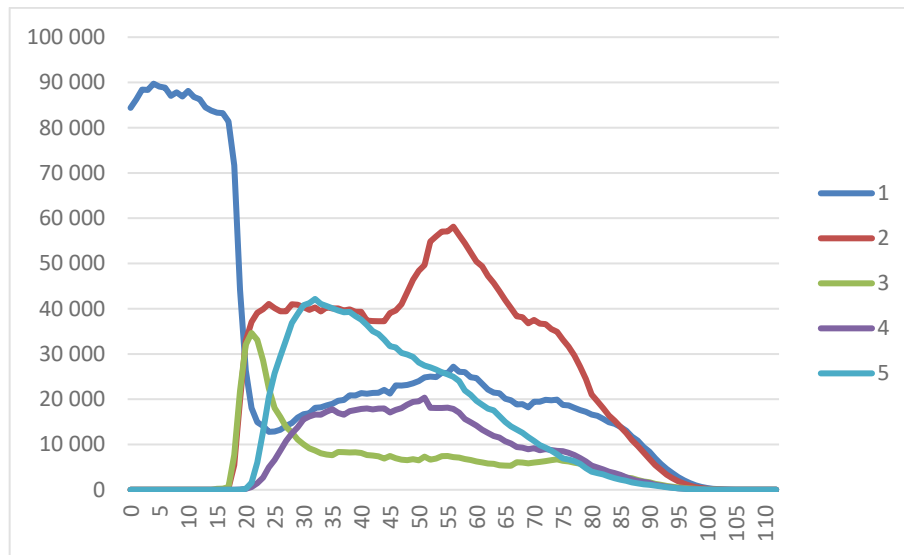
Results

The educational levels of the permanent resident population for 2020 can be found in Table 3 and Graph 1.

Table 3 : Frequency of levels of education in 2020

Levels	0 ≤ age < 115		Age > 14 years		24 < age < 65	
	Frequency	%	Frequency	%	Frequency	%
1	3 065 093	35.4	1 758 825	23.9	848 699	17.6
2	2 690 992	31.0	2 690 992	36.5	1 778 942	36.8
3	638 282	7.4	638 279	8.7	327 617	6.8
4	823 104	9.5	823 104	11.2	642 447	13.3
5	1 452 829	16.8	1 452 829	19.7	1 238 443	25.6
Total	8 670 300	100.0	7 364 029	100.0	4 836 148	100.0

Graph 1: Different levels of education in 2020 distributed by age



Conclusions

The highest level of education variable thus obtained is 90% derived from the various data sources and around 10% from imputations based on a model adjusted using the data from the Structural Survey. This variable will be calculated annually, and the methodology for its construction may still evolve. It will allow users to include the highest level of completed education in multidimensional analyses, provided that any results obtained on the basis of this variable are clearly declared as experimental. Detailed documentation of the methodology is available on the FSO's experimental statistics microsite.