



# Höchste abgeschlossene Ausbildung STATPOP

## Produktionsprozess – Kurzbeschreibung

### EXPERIMENTAL STATISTICS



Neuchâtel, 2025

|                |                                            |                |                                                                               |
|----------------|--------------------------------------------|----------------|-------------------------------------------------------------------------------|
| Herausgeber:   | Bundesamt für Statistik (BFS)              | Originaltext:  | Französisch                                                                   |
| Auskunft:      | seduperspectives@bfs.admin.ch              | Layoutkonzept: | Sektion PUB                                                                   |
| Redaktion:     | Sabrina Bravo, BFS ; Pascal Strubi, BFS    | Download:      | <a href="http://www.statistik.ch">www.statistik.ch</a>                        |
| Inhalt:        | Sabrina Bravo, BFS ; Pascal Strubi, BFS    | Copyright:     | BFS, Neuchâtel 2025                                                           |
| Themenbereich: | 00 Statistische Grundlagen und Übersichten |                | Wiedergabe unter Angabe der Quelle<br>für nichtkommerzielle Nutzung gestattet |

## Einleitung

Die höchste abgeschlossene Ausbildung der Personen der ständigen Wohnbevölkerung wird häufig als Indikator für das bildungsspezifische Humankapital eines Landes herangezogen. Vor dem Hintergrund demografischer Veränderungen wie der Bevölkerungsalterung, der zu wachsenden wirtschaftlichen Wettbewerbsfähigkeit und des damit einhergehenden zunehmenden Fachkräftebedarfs sind die Beurteilung und Entwicklung dieses Humankapitals von grosser Bedeutung. Für multidimensionale Analysen wird das Bildungsniveau aller Personen der ständigen Wohnbevölkerung benötigt. Dazu braucht es zusätzlich zu den statistischen Daten und Erhebungsinformationen der Bundesverwaltung ein Imputationsmodell, das die fehlenden Angaben ergänzt. Die Variable der höchsten abgeschlossenen Ausbildung kann unter anderem auf individueller Ebene bei Verknüpfungen für statistische Analysen verwendet werden, lässt aber keine Rückschlüsse auf Einzelpersonen zu. Die Nomenklatur des Bildungsniveaus besteht aus drei oder fünf Stufen (Tabelle 1).

**Tabelle 1: Nomenklatur des drei- und fünfstufigen Bildungsniveaus**

| Bildungsniveau        |            |                                                                                                               |
|-----------------------|------------|---------------------------------------------------------------------------------------------------------------|
| Dreistufig            | Fünfstufig | Ausbildungen in der fünfstufigen Nomenklatur                                                                  |
| Obligatorische Schule | 1          | Obligatorische Schule (in Ausbildung oder abgeschlossen) / keine / einjährige Ausbildung / Brückenangebot     |
| Sekundarstufe II      | 2          | Berufslehre im Betrieb (EFZ, EBA) / Anlehre / Berufsschule / Handelsschule                                    |
|                       | 3          | Fachmittelschule / gymnasiale Maturität / Lehrkräfteseminar / Berufs- oder Fachmaturität                      |
| Tertiärstufe          | 4          | Berufsprüfung mit eidg. Fachausweis / höhere Fachprüfung mit eidg. Diplom / Meisterdiplom / höhere Fachschule |
|                       | 5          | Universität / ETH / Fachhochschule / pädagogische Hochschule                                                  |

## Datenquellen

Das Bildungsniveau der ständigen Wohnbevölkerung wird jedes Jahr anhand einer Vielzahl von Datenquellen ermittelt: der Statistik der Bevölkerung und Haushalte (STATPOP), der Strukturerhebung (SE) seit 2010, der Datei der Längsschnittanalysen im Bildungsbereich (LABB) ergänzt mit der Statistik des Schulpersonals (SSP) und jener des Hochschulpersonals (SHIS-PERS), der Sozialhilfeempfängerstatistik (SHS), der Arbeitslosenstatistik des Staatssekretariats für Wirtschaft (SECO), der Lohnstrukturerhebung (LSE) und der auf den Volkszählungen 1990 und 2000 basierenden Swiss National Cohort (SNC).

## Datenaufbereitung und deterministische Imputation

In einem ersten Schritt werden die Daten der einzelnen Quellen plausibilisiert (LABB ohne Plausibilisierung wird als LABB\_0 bezeichnet). Dabei werden zum einen die potenziell veralteten Daten gelöscht, zum Beispiel, wenn eine Person das Studium wieder aufgenommen oder die Schweiz in einem Alter verlassen hat, in dem sie das Studium wieder hätte aufnehmen können, oder wenn eine Person unter 18 Jahren die Tertiärstufe angegeben hat. Zum anderen werden die Datenquellen mithilfe der LABB-Daten korrigiert, wenn das Bildungsniveau in LABB\_0 höher ist oder wenn sich die Person in Ausbildung befindet und ihre Bildungslaufbahn seit der letzten abgeschlossenen Ausbildung in LABB bekannt ist.

Nach der Plausibilisierung der Quellen werden diese priorisiert. Zunächst werden die Personen **unter 15 Jahren** auf das Niveau 1 gesetzt, ausser, wenn in LABB\_0 ein höheres Niveau vermerkt ist. Dann wird die Information aus LABB für Personen, die in LABB in Ausbildung sind und deren Bildungslaufbahn seit der letzten abgeschlossenen Ausbildung bekannt ist, sowie für Personen, die in LABB das höchste

Niveau erreicht haben (in LABB als **LABB\*** bezeichnet), übernommen. Anschliessend wird das Bildungsniveau anhand der **Konkordanz** der Quellen festgelegt (wenn das in den Quellen am häufigsten angegebene Bildungsniveau in mindestens zwei Quellen mehr enthalten ist als das zweithäufigste Bildungsniveau). In einem weiteren Schritt werden die auf Administrativdaten basierenden Daten aus den Quellen LABB und SECO verwendet: zunächst das Bildungsniveau gemäss **SECO**, danach das noch nicht verwendete Bildungsniveau gemäss LABB, d. h. ohne LABB\* (**LABB\LABB\***), da SECO auch im Ausland erworbene Ausbildungen enthält. Zuletzt werden die Quellen SHS, SE und SNC, die das von der Person selbst angegebene Bildungsniveau enthalten, verwendet: an erster Stelle die **SHS**, da die Person, die die Befragung durchführt, der befragten Person helfen kann, das Bildungsniveau zu bestimmen, gefolgt von der **SE** und schliesslich von der **SNC**, deren Informationen am ältesten sind.

**Tabelle 2: Häufigkeit der aggregierten Zuweisungsregeln für das Bildungsniveau 2020**

| Zuteilungsregeln | Häufigkeit | %     | Kumuliert % |
|------------------|------------|-------|-------------|
| < 15 Jahre       | 1 306 271  | 15,1  | 15,1        |
| LABB*            | 1 265 351  | 14,6  | 29,7        |
| Konkordanzen     | 2 852 961  | 32,9  | 62,6        |
| SECO             | 556 972    | 6,4   | 69,0        |
| LABB\LABB*       | 269 333    | 3,1   | 72,1        |
| SHS              | 60 549     | 0,7   | 72,8        |
| SE               | 851 471    | 9,8   | 82,6        |
| SNC              | 700 775    | 8,1   | 90,7        |
| Fehlend          | 806 617    | 9,3   | 100,0       |
| Total            | 8 670 300  | 100,0 | 100,0       |

Wie Tabelle 2 für das Jahr 2020 zeigt, wird 90,7% der Bevölkerung das Bildungsniveau anhand verschiedener Quellen zugeordnet. Das Bildungsniveau der verbleibenden 9,3% der Bevölkerung wird mithilfe des im Folgenden beschriebenen Algorithmus imputiert.

## Modellbasierte Imputation

Das Bildungsniveau von Personen ohne zugeordnetes Bildungsniveau wird auf Basis der SE imputiert. Zur Definition der Trainingsdaten des Imputationsmodells wird auf die SE des Referenzjahres das gleiche Zuordnungsverfahren angewendet wie auf die Gesamtbevölkerung (oben beschrieben), aber ohne die SE des Referenzjahres. Die Trainingsdaten bestehen aus allen Personen der SE, denen nicht anhand anderer Quellen ein Bildungsniveau zugeordnet wurde. Für die Imputation werden Variablen aus verschiedenen Registern verwendet, insbesondere aus der Zentralen Ausgleichsstelle (z. B. Löhne), der Statistik der Unternehmensstruktur (z. B. NOGA-Code), der STATPOP-Variablenliste Bestand (z. B. Geschlecht, Nationalität, Alter), der STATPOP-Variablenliste Haushalt (z. B. Haushaltszusammensetzung) sowie der Gebäude- und Wohnungsstatistik (z. B. Ein- und Mehrfamilienhäuser).

Analysen haben gezeigt, dass sich die Bildungsniveaus 2, 3 und 4 nur schwer unterscheiden lassen. Aus diesem Grund erfolgt die Imputation in zwei Schritten. Eine erste multinomiale Regression liefert eine Schätzung der Wahrscheinlichkeit, ein Bildungsniveau 1, 234 oder 5 zu haben (als  $P_A$  bezeichnet), eine zweite multinomiale Regression unterscheidet zwischen den Niveaus 2, 3 und 4 (als  $P_B$  bezeichnet). Auf diese Weise wird für eine Person  $k$  in der Bevölkerung die Wahrscheinlichkeit berechnet, den einzelnen Niveaus anzugehören:

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

Diese Wahrscheinlichkeiten werden anschliessend plausibilisiert: Das Niveau muss höher oder gleich hoch sein wie in LABB\_0 und die Person muss älter sein als 18 Jahre, um eine Tertiärstufe zu haben. Nach der Schätzung dieser Wahrscheinlichkeiten wird ein Optimierungsalgorithmus angewendet, um

jeder Person ein Bildungsniveau zuzuordnen. Dabei wird der Erwartungswert des Fehlers unter Beibehaltung der geschätzten Totale, die sich aus der Summe der Wahrscheinlichkeiten für die einzelnen Bildungsniveaus nach Alter kombiniert mit dem Geschlecht ergeben, minimiert (Maximierung der Likelihood-Schätzung).

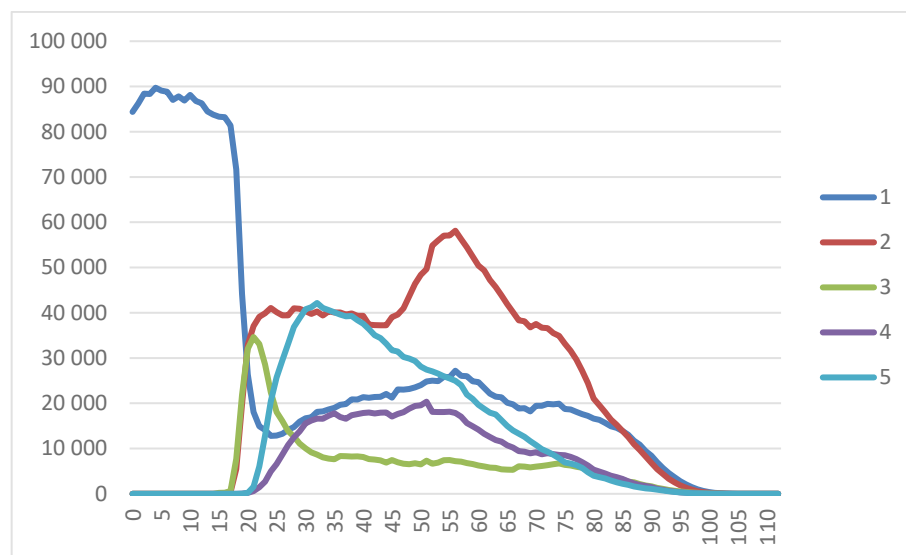
## Ergebnisse

In Tabelle 3 und Grafik 1 sind die Bildungsniveaus der ständigen Wohnbevölkerung 2020 ersichtlich.

**Tabelle 3: Häufigkeit der Bildungsniveaus, 2020**

| Niveau       | 0 ≤ Alter < 115  |              | Alter > 14 Jahre |              | 24 < Alter < 65  |              |
|--------------|------------------|--------------|------------------|--------------|------------------|--------------|
|              | Häufigkeit       | %            | Häufigkeit       | %            | Häufigkeit       | %            |
| 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         |
| <b>Total</b> | <b>8 670 300</b> | <b>100,0</b> | <b>7 364 029</b> | <b>100,0</b> | <b>4 836 148</b> | <b>100,0</b> |

**Grafik 1: Verteilung der Bildungsniveaus nach Alter, 2020**



## Fazit

Die so ermittelte Variable der höchsten abgeschlossenen Ausbildung besteht zu rund 90% aus Informationen aus den verschiedenen Quellen und zu rund 10% aus Imputationen anhand eines auf die SE-Daten abgestimmten Modells. Diese Variable wird jährlich produziert, wobei sich die Methodik noch verändern kann. Die höchste abgeschlossene Ausbildung wird für multidimensionale Analysen zur Verfügung gestellt, sofern die entsprechenden Ergebnisse eindeutig als experimentell deklariert sind. Ausführliche Erläuterungen zur Methodik sind auf der BFS-Microsite «Experimentelle Statistiken» zu finden.