

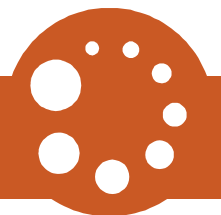


Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Eidgenössisches Departement des Innern EDI
Bundesamt für Statistik BFS

LABB: Gewichtung der verknüpften Datenbank

EXPERIMENTAL STATISTICS



Neuchâtel, 2021

Herausgeber:	Bundesamt für Statistik (BFS)	Layoutkonzept:	Sektion DIAM
Auskunft:	eduperspectives@bfs.admin.ch, Tel. 058 483 93 66	Download:	www.experimental.bfs.admin.ch
Redaktion:	Sektion METH	Copyright:	BFS, Neuchâtel 2021
Inhalt:	Desislava Nedyalkova, Anne Massiani, Daniel Kilchmann, METH		Wiedergabe nur mit Einverständnis der Autoren gestattet
Themenbereich:	00 Statistische Grundlagen		
Originaltext:	Französisch		
Übersetzung:	Sprachdienste BFS		

LABB: Gewichtung der verknüpften Datenbank

Desislava Nedyalkova, Anne Massiani, Daniel Kilchmann (METH)

17.07.2020

1 Grundprinzip

Die Zielpopulation für die verknüpfte Datenbank ist eine Teilgruppe der ständigen Wohnbevölkerung 2016. Sie umfasst Personen im Alter von 19 bis 64 Jahren, die in einem nicht im Primärsektor tätigen Unternehmen mit mindestens drei Beschäftigten angestellt sind und deren Einkommen im Dezember strikt positiv war. Aus der Population ausgeschlossen sind Lernende, Mehrfacherwerbstätige, Personen mit beruflichen Veränderungen (inkl. Einkommensunterschieden zwischen den Referenzzeitpunkten Oktober und 31.12.2016) und Personen im Stundenlohn. Es liegen zwei Stichproben vor. Für die erste wurden die Daten der Zentralen Ausgleichsstelle (ZAS) anhand der AHV-Nummer mit denjenigen der Strukturerhebung (SE) verknüpft, für die zweite wurde, ebenfalls anhand der AHV-Nummer, eine weitere Verknüpfung mit der Stichprobe der Schweizerischen Lohnstrukturerhebung (LSE) 2016 vorgenommen. Von Interesse ist hier die zweite Stichprobe, die sogenannte «verknüpfte Datenbank».

Die verknüpfte Datenbank wird nach der klassischen Methode gewichtet, d.h. zunächst werden die Antwortausfälle bearbeitet, danach erfolgt die Kalibrierung.

2 Verfügbare Informationen

Die drei von der Sektion Bildungssystem (BILD-S) gelieferten Daten umfassen die Grundgesamtheit und die beiden genannten Stichproben. In diesem Kapitel werden die Informationen beschrieben, die für alle drei Dateien vorliegen.

1. Datei SPCDCSSTATENT2016: ZAS-Population im Jahr 2016 (N = 1 947 192), inkl. Lernende und Personen im Stundenlohn. Lernende und Personen im Stundenlohn können in dieser Datei nicht identifiziert werden.

Filter:

- 19 bis 64 Jahre alt (gemäss STATPOP)
 - Arbeitnehmende mit strikt positivem Einkommen im Dezember gemäss ZAS
 - ohne Mehrfacherwerbstätige gemäss ZAS
 - ohne Personen mit beruflichen Veränderungen gemäss ZAS (inkl. Einkommensentwicklungen zwischen den Referenzzeitpunkten der Erhebungen)
2. (a) Datei CDCRS_sample_f2a_c: Stichprobe ZAS-SE 2016 (n= 67 152), inkl. Lernende und Personen im Stundenlohn. Die Lernenden werden mithilfe einer Variablen identifiziert. Personen im Stundenlohn können nicht identifiziert werden.

Zusätzliche Filter:

- Arbeitnehmende SE (mit Lernenden)
- gültige Arbeitszeit SE
- Entfernung von Ausreissern bei den Einkommen

(b) Datei CDCRS_sample_f2b_c: Stichproben ZAS-SE/LSE 2016 (n= 29 745), ohne Lernende, aber mit Personen, deren «iwaz»-Variable gleich 0 ist, was ungefähr den Personen im Stundenlohn entspricht

Zusätzliche Filter:

- ohne Lernende
- ohne Mehrfacherwerbstätige gemäss ZAS

Die Lernenden stellen bei der Gewichtung der verknüpften Datenbank ein Problem dar. Sie gehören nicht zur Zielpopulation der verknüpften Datenbank, können aber nicht identifiziert und aus dem Gewichtungsrahmen, der auf der Datei SPCDCSSTATENT2016 beruht, entfernt werden. Ein ähnliches Problem stellt sich bei Personen im Stundenlohn. Im folgenden Kapitel wird beschrieben, wie dieses Problem umgangen werden könnte. Die Lösung für die Lernenden weicht von derjenigen für die Personen im Stundenlohn ab, da die beiden Personenkategorien in den Dateien CDCRS_sample_f2a_c und CDCRS_sample_f2b_c unterschiedlich erfasst sind (siehe folgende Tabelle).

Tabelle 1 – Lernende und Personen im Stundenlohn

	Lernende	Personen im Stundenlohn
Zielpopulation	nicht enthalten	nicht enthalten
Datei SPCDCSSTATENT2016	enthalten, nicht identifizierbar	enthalten, nicht identifizierbar
Datei CDCRS_sample_f2a_c	enthalten, identifizierbar	enthalten, nicht identifizierbar
Datei CDCRS_sample_f2b_c	nicht enthalten	enthalten, identifizierbar

Zur Bildung der Variable «NOGA-Sektor» sowie weiterer Variablen der STATENT, die für die Gewichtung nützlich sein könnten, wurde mithilfe des in beiden Dateien enthaltenen Unternehmensidentifikators die Datei SPCDCSSTATENT2016 mit der Datei der STATENT 2016 (u_sta16_ins_20180719_appr) verknüpft.

3 Gewichtungsschritte

Nachfolgend wird die Gewichtung Schritt für Schritt beschrieben.

3.1 Entfernung des Primärsektors und der Unternehmen mit weniger als drei Beschäftigten

Der Primärsektor ist in der ZAS und in der SE enthalten, nicht aber in der LSE und somit auch nicht in der verknüpften Datenbank. Das heisst: Der Primärsektor muss anhand der in der Datei SPCDCSSTATENT2016 enthaltenen Variable «Sektor» der STATENT aus den drei Dateien entfernt werden.

Die Unternehmen mit weniger als drei Beschäftigten sind in der ZAS und in der SE enthalten, nicht aber in der LSE und somit auch nicht in der verknüpften Datenbank. Folglich wurden diese Unternehmen anhand der in der Datei SPCDCSSTATENT2016 enthaltenen Variable «emptot» (Gesamtzahl der Beschäftigten in einem Unternehmen) der STATENT aus den drei Dateien entfernt.

Die drei Dateien, die nach Entfernung des Primärsektors und der Unternehmen mit weniger als drei Beschäftigten entstehen und als Ausgangsdateien für die Gewichtung dienen, werden als f1, f2a und f2b bezeichnet. Sie entsprechen den Dateien von Kapitel 2 nach den zusätzlich angewendeten Filtern.

3.2 Berechnung der korrigierten Gewichte für die Antwortausfälle

Der zweite Schritt besteht darin, die Gewichte der LSE für die Antwortausfälle zu korrigieren. Da die SE- und die LSE-Stichprobe voneinander unabhängig sind, werden die Gewichte der LSE (Variable «gewicht») in der Datei f2b mit dem Gewicht der SE (Variable «zpers_weight») multipliziert. Daraus ergibt sich ein Gewicht w_{NR} , das für die Antwortausfälle der beiden Erhebungen korrigiert ist. Obwohl die Gewichte vor der Kalibrierung von Vorteil gewesen wären, wurden hier die Endgewichte der LSE und der SE verwendet.

3.3 Vorbereitung der Kalibrierungstotale

Die Kalibrierung erfolgt in zwei Schritten. Als Hilfsvariablen liegen zum Beispiel das Geschlecht, die Altersklasse, der NOGA-Code und die Grösse des Unternehmens, in dem die Person arbeitet (z.B. anhand der Variable «emptot» nach Klassen gruppiert), vor. Der erste Kalibrierungsschritt besteht darin, die Gewichte der Datei f2a (inkl. Lernende) auf die Totale der für die Datei f1 berechneten Kalibrierungsvariablen zu kalibrieren. Daraus resultiert für jede Einheit der Datei f2a (inkl. Lernende) ein neuer Gewichtungssatz w_{CDCRS} . Anschliessend wird in der Datei f2a (ohne Lernende) für jede Modalität der Kalibrierungsvariablen die Summe der Gewichte w_{CDCRS} berechnet. Dies ergibt einen ungefähren Wert der Kalibrierungstotale, die in der Datei f1 berechnet werden könnten, wenn es möglich wäre, die Lernenden auszuschliessen (die Personen im Stundenlohn aber beizubehalten).

3.4 Abschliessende Kalibrierung

Beim letzten Gewichtungsschritt werden die Gewichte w_{NR} der Datei f2b (mit $iwaz = 0$) auf die im vorangehenden Schritt berechneten Totale kalibriert. Daraus resultiert der neue

Gewichtungssatz w_{cal} . Die Datei, die sich aus dieser zweiten Kalibrierung und nach Entfernung der Fälle mit $iwaz = 0$ sowie ohne Änderung der Gewichte ergibt, entspricht der gewichteten verknüpften Datenbank, die der Sektion BILD-S geliefert wurde. Da Personen im Stundenlohn nach der Kalibrierung entfernt wurden, handelt es sich bei den Analysen auf Basis der verknüpften Datenbank um Schätzungen in einem Bereich der Datei f2b.

4 Praktische Umsetzung

Im Folgenden werden für jeden Gewichtungsschritt die Umsetzung und die wichtigsten Ergebnisse präsentiert.

4.1 Entfernung des Primärsektors und der Unternehmen mit weniger als drei Beschäftigten

In Tabelle 2 sind die Anteile der entfernten Daten an den gelieferten Daten nach Ausschluss des Primärsektors, der Unternehmen mit weniger als drei Beschäftigten sowie insgesamt aufgeführt. Wie sich zeigt, wurde in allen drei Fällen nur ein kleiner Anteil der Daten entfernt. Tabelle 3 gibt die Grösse der Dateien f1, f2a und f2b (nach Entfernung) an.

Tabelle 2 – Anteile der entfernten Daten

Entfernte Daten (in %)	Datei 1	Datei 2.a	Datei 2.b
Nur Primärsektor	0,71%	0,42%	0,02%
Nur Unternehmen mit weniger als drei Beschäftigten	4,76%	3,38%	0,32%
Beide (Total)	5,37%	3,75%	0,35%

Tabelle 3 – Grösse der für die Gewichtung verwendeten Populationen

	f1	f2a	f2b
Grösse	1 842 703	64 633	29 645

4.2 Berechnung der korrigierten Gewichte für die Antwortausfälle

Nach der Berechnung der Gewichte w_{NR} wurden diese Gewichte, insbesondere ihre Streuung, analysiert. Tabelle 4 gibt einen Überblick über die Streuung der beiden Gewichtungssätze (Gewichte der SE, der LSE und Gewichte w_{NR} , die aus der Multiplikation der beiden vorangehenden Gewichtungssätze resultieren). Wie sich zeigt, stimmen diese neuen Gewichte mit der Streuung der SE- und der LSE-Gewichte nach der Multiplikation der Gewichte überein.

Tabelle 4 – Verteilung der Gewichte *zpers_weight*, *gewicht* und *w_{NR}*

Gewicht	Min.	q10	q50	q90	Max.
<i>zpers_weight</i>	12,81	16,26	21,18	36,40	65,39
<i>gewicht</i>	0,59	1,01	1,16	3,95	138,55
<i>w_{NR}</i>	8,84	17,94	34,55	100,45	4687,63

Die Summe der Gewichte wurde ebenfalls überprüft. Die Summe der Gewichte *w_{NR}* müsste grundsätzlich einer realitätsnahen Schätzung der Zielpopulationsgrösse entsprechen. Tabelle 5 enthält die Grösse der Zielpopulation inkl. der in der Datei f1b nicht identifizierbaren Lernenden, die gerundete Summe der Gewichte *w_{NR}* in der Datei f2b sowie die gerundete Summe der Gewichte *zpers_weight* (vor der Kalibrierung) in der Datei f2a ohne Lernende.

Tabelle 5 – Summe der Gewichte und Grösse der Population

Grösse der Zielpopulation	1 842 703
Summe der Gewichte <i>w_{NR}</i>	1 618 701
Summe der Gewichte <i>zpers_weight</i>	1 630 812

Die Summe der Gewichte weicht relativ stark von der Grösse der Zielpopulation ab. Dabei ist aber zu berücksichtigen, dass in der Zielpopulation die Lernenden enthalten sind, in der Datei f2b hingegen nicht.

4.3 Vorbereitung der Kalibrierungstotale und erste Kalibrierung

Um die Variablen (oder die Variablenkombinationen) festzulegen, die in den beiden aufeinanderfolgenden Kalibrierungen verwendet werden sollen, wurden ihre Bestände in der Stichprobe f2b untersucht. Pro Zelle galt eine Mindestanzahl von 50 Beobachtungen. Bei ungenügenden Beständen wurden bestimmte Kategorien gruppiert. Die getroffenen Auswahlen sind im Folgenden beschrieben.

Bildung der Variable Unternehmensgrössenklasse

Um eine oder mehrere Kategorien für die Variable «Unternehmensgrössenklasse» zu bilden, wurde die Verteilung der Variable «emptot» in der Population untersucht. Hierzu wurde anhand der Quartile eine erste Kategorisierung vorgenommen, die vier Kategorien vergleichbarer Grösse ergibt. Die Bestände des Kalibrierungsrahmens und der Stichprobe f2b sind in Tabelle 6 ersichtlich.

Tabelle 6 – Die Variable «Grösse»

Unternehmensgrösse	Wert	Bestand Rahmen	Bestand Stichprobe f2b
[3; 25]	1	463 486	2 576
(25; 170]	2	458 969	4 842
(170; 1700]	3	455 911	8 787
≥ 1700	4	464 337	13 440

Eine detaillierte Kategorisierung ergab die folgenden Kategorien und Bestände:

Tabelle 7 – Die Variable «detaillierte Grösse»

Unternehmensgrösse	Wert	Bestand Rahmen	Bestand Stichprobe f2b
≤ 8	1	185 072	594
[8; 25]	2	278 414	1 982
(25; 100]	3	329 935	3 002
(100; 200]	4	168 982	2 493
(200; 500]	5	203 411	3 504
(500; 1700]	6	212 552	4 630
(1700; 10 000]	7	235 262	5 948
≥ 10 000	8	229 075	7 492

Bildung der Variable Altersklasse

In den vorliegenden Daten ist die Variable ageclass bereits enthalten. Sie wurde anhand der folgenden Kategorien in numerische Werte umgewandelt:

- wenn ageclass gleich «19–30 Jahre», dann ist die neue Variable cl age = 1
- wenn ageclass gleich «31–40 Jahre», dann ist die neue Variable cl age = 2
- wenn ageclass gleich «41–50 Jahre», dann ist die neue Variable cl age = 3
- wenn ageclass gleich «51–64 Jahre», dann ist die neue Variable cl age = 4

Diese Variable wurde mit der Variable Geschlecht kombiniert. In Tabelle 8 sind die Bestände des Kalibrierungsrahmens und der Stichprobe f2b dieser kombinierten Variable ersichtlich:

Tabelle 8 – Bestand nach Geschlecht und Altersklasse

Geschlecht	Altersklasse	Bestand Rahmen	Bestand Stichprobe f2b
1	1	240 688	2 730
1	2	265 672	4 518
1	3	255 098	4 552
1	4	258 332	4 580
2	1	206 770	2 678
2	2	195 035	3 401
2	3	208 744	3 587
2	4	212 364	3 599

Die verschiedenen NOGA-Variablen

Ebenfalls analysiert wurden die Bestände im Kalibrierungsrahmen und in der Stichprobe für mehrere NOGA-Variablen sowie ihre Kombination mit der Variable Geschlecht.

- Variable «noga082 STATENT» (zweistelliger NOGA-Code)
In der Stichprobe waren zu viele kleine Bestände enthalten. Diese Variable wurde verworfen.
- Variable «noga081 STATENT» (NOGA-Abschnitt: 18 Kategorien für die Kalibrierung)
Der Bestand in der NOGA Abschnitt B ist mit 19 Beobachtungen zu klein. Die Kombination der Variable «noga081 STATENT» mit dem Geschlecht (noga*sexe) ergab einige Nullbestände. Um auch bei der Kombination mit dem Geschlecht über genügend grosse Bestände zu verfügen, wurden daher die NOGA-Abschnitte B und C sowie die NOGA-Abschnitte D und E gruppiert.
- Variable «noga082 50» (NOGA BFS50: 47 Kategorien für die Kalibrierung)
Diese Variable mit detaillierteren Kategorien als die Variable «noga081 STATENT» hat nur in der NOGA «05.9» einen Bestand unter 50 Beobachtungen. Die NOGA «05.9» wurde daher mit der NOGA «31.3» zu einer neuen Kategorie zusammengelegt.

Berücksichtigte Variablen

Folgende Variablen wurden für die Kalibrierung verwendet:

- gruppierte NOGA BFS50 (nog50_cal_gr)
- Geschlecht kombiniert mit der Altersklasse (sexe*cl_age)
- Geschlecht kombiniert mit der Unternehmensgrösse nach Klasse mit je acht Kategorien (sexe*taille_fine)
- Geschlecht kombiniert mit dem gruppierten NOGA-Abschnitt (sexe*noga_gr)

Erste Kalibrierung

In Kapitel 3.3 wurde der erste Kalibrierungsschritt beschrieben. Daraus resultierte für jede Einheit der Datei f2a ein Gewichtungssatz w_{CDCRS} . Tabelle 9 zeigt die Verteilung der Anfangsgewichte ($zpers_weight$) und der kalibrierten Gewichte (w_{CDCRS}). Die beiden Verteilungen sind sehr ähnlich.

Tabelle 9 – Verteilung der Gewichte $zpers_weight$ und der kalibrierten Gewichte w_{CDCRS}

Gewicht	Min.	q10	q50	q90	Max.
$zpers_weight$	12,81	16,26	21,18	36,40	65,39
w_{CDCRS}	15,18	17,26	25,29	41,79	81,43

4.4 Abschliessende Kalibrierung

Durch die Kalibrierung der Gewichte w_{NR} der Datei f2b (mit $iwaz = 0$) auf die im vorangehenden Schritt berechneten Totale ergab sich ein neuer Gewichtungssatz (w_{cal}) für 29 645 Beobachtungen. Tabelle 10 zeigt die Verteilung der Anfangsgewichte (w_{NR}) und der kalibrierten Gewichte (w_{cal}). Die beiden Verteilungen sind sehr ähnlich und die Gewichte w_{cal} und w_{NR} ziemlich stark gestreut. Grund dafür ist insbesondere die Korrektur der Antwortausfälle, genauer gesagt die Art, wie die Stichprobe f2b gebildet wurde, nämlich durch Verknüpfung zweier Stichproben (SE und LSE) mit voneinander unabhängigen und sehr unterschiedlichen Antwortmechanismen, die durch die Verknüpfung kombiniert werden.

Tabelle 10 – Verteilung der Gewichte w_{NR} und w_{cal}

Gewicht	Min.	q10	q50	q90	Max.
w_{NR}	8,84	17,94	34,55	100,45	4687,63
w_{cal}	13,79	18,28	35,26	107,71	4629,00

Nach der Entfernung der Fälle mit $iwaz = 0$ (ohne Veränderung der Gewichte) ergibt sich ein Gewichtungssatz für 28 210 Beobachtungen. Diese Datei und diese Gewichte gelten als Endgewichte, die der Sektion geliefert wurden.

5 Berechnungsbeispiel für die Varianz

Als Beispiel wurden die Medianwerte des mbles- Einkommens (aus der LSE) und des standardisierten $cdrsref2$ -Einkommens berechnet. Sie wurden auf Basis der Kalibrierungsgewichte w_{cal} gewichtet. Diese Medianwerte sowie ihre Varianzen wurden mit dem Verfahren «proc surveymeans» von SAS berechnet, wobei für beide Einkommen der Clustereffekt der LSE berücksichtigt wurde (durch Verwendung der Variable $entidn$ der LSE). Für die Berechnung der Varianz der berechneten Variable $cdrsref2$ muss auch der Clustereffekt

berücksichtigt werden, da nicht alle Personen der SE in der Stichprobe übernommen werden, sondern nur diejenigen, die auch in der LSE enthalten sind. Auf diese Weise wird der Clustereffekt der LSE auf die ganze verknüpfte Datenbank übertragen. Bei der Berechnung der Varianz wird der Kalibrierungseffekt zudem nicht berücksichtigt. Tabelle 11 zeigt die geschätzten (gerundeten) Werte der gewichteten Mediane, den Standardfehler (SE) und die Grenzen der geschätzten Vertrauensintervalle (ICL und ICU).

Mit den für die Berechnung der Varianz verwendeten Verfahren kann die Varianz geschätzt werden, die auf den Stichprobenplan zurückzuführen ist, so wie das im BFS üblicherweise bei interessierenden Variablen gehandhabt wird, die nicht auf einer Zufallsbasis gebildet werden. Die berechnete Variable `cdcrsref2` (standardisiertes Einkommen) ist aufgrund ihrer Berechnung hingegen eine Zufallsvariable. Der durch die Berechnung dieses Einkommens bedingte Effekt fließt nicht in die Berechnung der Varianz ein. Diese zusätzliche Varianz konnte in der vorgegebenen Frist nicht integriert werden. Es wäre jedoch interessant, in einem weiteren Projektschritt zu untersuchen, ob diese Quelle zusätzlicher Varianz in die Präzisionsberechnungen integriert werden könnte.

Tabelle 11 – Gewichtete Medianwerte (Schätzung)

Variable	Geschätzter Medianwert	SE	ICL	ICU
MBLS	6726,07	56,56	6615,20	6836,94
cdcrsref2	6739,10	59,01	6623,42	6854,77

6 Grenzen der Gewichtung

Diese Arbeit ist im Rahmen der LABB-Analysen im Hinblick auf Analysen der verknüpften Datenbank erfolgt. Die Gewichtung hat aber ihre Grenzen, u.a.:

- Mit der Gewichtung können die Daten nur für die Zielpopulation hochgerechnet werden. Diese entspricht allerdings nur 52% der Bevölkerung der unverknüpften Datenbanken.
- Die Endgewichte sind stark gestreut. Dies könnte sich für die Präzision einiger Teilpopulationen als problematisch erweisen.
- Für die Korrektur der Antwortausfälle wurden die Endgewichte der LSE und der SE verwendet, obwohl die Gewichte vor der Kalibrierung von Vorteil gewesen wären.