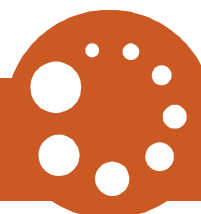




LABB: Weighting of the linked database

EXPERIMENTAL STATISTICS



Neuchâtel, 2021

Published by: Federal Statistical Office (FSO)
Information: eduperspectives@bfs.admin.ch, Tel. 058 483 93 66
Editors: Section METH
Contents: Desislava Nedyalkova, Anne Massiani, Daniel Kilchmann, METH
Topic: 00 Statistical Basis and Overviews
Original text: French
Translation: FSO language services

Layout concept: Section DIAM
Downloads: www.experimental.bfs.admin.ch
Copyright: FSO, Neuchâtel 2021
May only be reproduced with the authors' consent

LABB: Weighting of the linked database

Desislava Nedyalkova, Anne Massiani, Daniel Kilchmann (METH)

17.07.2020

1 General principle

The target population for the linked database is a sub-group of the permanent resident population in Switzerland in 2016. It comprises people aged between 19 and 64 who work in an enterprise with at least three employees outside the primary sector, with an income that must be positive in December. We exclude from the population apprentices, people in multiple employment, people who changed jobs (including variations in income between the reference dates – October and 31.12.2016) and people who are paid by the hour. We have two samples – one is obtained through the linkage of data from the Central Compensation Office (CCO) with the Structural Survey (SS) using the OASI number, and the other by additionally linking with the sample from the 2016 Swiss Earnings Structure Survey (ESS) using the OASI number. Here, we are interested in the second sample, hereinafter referred to as the 'linked database'.

For the weighting of the linked database, we propose adopting the 'traditional' approach, which involves applying a treatment for non-response followed by calibration.

2 Available information

The Education System section (BILD-S) has provided three files which comprise the population and the two samples mentioned above. In this section we describe the information we have for each of the three files.

1. file SPCDCSSTATENT2016: CCO population for 2016 (N= 1 947 192), including apprentices and people paid by the hour. Apprentices cannot be identified in the file, neither can people paid by the hour.

Filters:

- aged 19–64 (according to STATPOP),
 - workers with a strictly positive income in December according to the CCO,
 - excluding people in multiple employment according to the CCO,
 - excluding people who changed jobs according to the CCO (including variations in income between the two survey reference dates).
2. (a) file CDCRS_sample_f2a_c: sample CCO-SS 2016 (n= 67 152), including apprentices and people paid by the hour. One variable allows the apprentices to be identified. People paid by the hour cannot be identified.

Additional filters:

- SS employees (including apprentices),
 - valid working time in the SS,
 - elimination of outlying incomes.
- (b) file CDCRS_sample_f2b_c: sample CCO-SS/ESS 2016 (n= 29 745), excluding apprentices but including people whose iwaz variable equals 0, which corresponds approximately to people paid by the hour.

Additional filters:

- excluding apprentices,
- excluding people in multiple employment according to the ESS.

The issue of apprentices presents a difficulty in terms of the weighting of the linked database. They are not part of the linked database target population, but it is not possible to locate and eliminate them from the calibration frame, which is based on the file SPCDCSSTATENT2016. The same type of issue arises for people paid by the hour. The following section proposes a solution to get around this problem. The solution chosen for apprentices differs from the one chosen for people paid by the hour due to the different situation of these two characteristics in the files CDCRS_sample_f2a_c and CDCRS_sample_f2b_c, as shown in the table below.

Table 1 – Apprentices and people paid by the hour

	Apprentices	People paid by the hour
Target population	not included	not included
File SPCDCSSTATENT2016	included, not identifiable	included, not identifiable
File CDCRS_sample_f2a_c	included, identifiable	included, not identifiable
File CDCRS_sample_f2b_c	not included	included, identifiable

In order to obtain the NOGA sector variable and other STATENT variables that may be useful for the weighting, the file SPCDCSSTATENT2016 was linked with the STATENT 2016 file (u_sta16_ins_20180719_appr) by means of the enterprise identification number present in both files.

3 The weighting steps

This section sets out the steps in the weighting.

3.1 Elimination of the primary sector and enterprises with fewer than three employees

The primary sector is included in the CCO and the SS, but not in the ESS, and thus not in the linked database. This means the primary sector has to be deleted in the three files according to the STATENT 'sector' variable included in the SPCDCSSTATENT2016 file.

Enterprises with fewer than three employees are included in the CCO and the SS, but not in the ESS, and thus not in the linked database. We deleted these enterprises from all three files according to the STATENT emptot variable (total number of employees in an enterprise) which is included in the SPCDCSSTATENT2016 file.

The three files we obtain after eliminating the primary sector and enterprises with fewer than three employees, and which will be used as input files for the weighting, will be named f1, f2a and f2b, and correspond to the files in section 2 after the additional deletions mentioned above.

3.2 Calculation of weights adjusted for non-response

The second stage involves adjusting the ESS weights for non-response. Due to the independence of the two samples (SS and ESS), the ESS weights (gewicht variable) are multiplied by the SS weights (zpers_weight variable) in file f2b. This provides a weight adjusted for the non-response inherent in both surveys, which is denoted by w_{NR} . The final ESS and SS weights have been used here, although ideally the weights before calibration would have been preferable.

3.3 Preparation of calibration totals

We opted for a two-stage calibration. We have a number of auxiliary variables, such as sex, age group, NOGA, size of enterprise where the person works (e.g. grouped according to the emptot variable). The first stage of calibration involves calibrating the weights in file f2a (including apprentices) to the totals of the calibration variables calculated on file f1. At the end of this stage, we obtain a new set of weights, denoted by w_{DCRS} , for each unit in the f2a file (including apprentices). Next, for each modality of the calibration variables, we calculate the sum of the weights w_{DCRS} on the f2a file (excluding apprentices). This provides approximate calibration totals that we could calculate on file f1 if we were able to exclude apprentices (but keep people paid by the hour).

3.4 Final calibration

The final stage of weighting involves calibrating the weights w_{NR} in the f2b file (with iwaz=0) to the totals calculated in the previous stage. This provides a new set of weights, denoted by w_{cal} . The file resulting from this second calibration, after deleting cases where iwaz=0 and without modifying the weights, constitutes the weighted linked database and was delivered to the BILD-S section. The exclusion of people paid by the hour after calibration means the analyses conducted on the linked database should be considered as estimations in a domain of file f2b.

4 Practical implementation

In this section, we present the implementation and the main results for each stage of weighting.

4.1 Elimination of the primary sector and enterprises with fewer than three employees

Table 2 gives the percentages of data eliminated from the data at hand after deletion of the primary sector, enterprises with fewer than three employees, and in total. We see that in each case, a small percentage of data is eliminated. Table 3 shows the sizes of files f1, f2a and f2b (after elimination).

Table 2 – Percentages of data eliminated

Elimination from the data (in %)	file 1	file 2.a	file 2.b
of the primary sector only	0.71%	0.42%	0.02%
of enterprises with fewer than three employees only	4.76%	3.38%	0.32%
of both (total)	5.37%	3.75%	0.35%

Table 3 – Sizes of input data for weighting

	f1	f2a	f2b
size	1 842 703	64 633	29 645

4.2 Calculation of weights adjusted for non-response

After calculating the weights w_{NR} , we conducted analyses on these weights, in particular on their dispersion. Table 4 provides an overview of the dispersion of the three sets of weights (SS weights, ESS weights and the weights w_{NR} that result from multiplying the two previous sets of weights). We can see that the distribution of these new weights is in line with the distributions of SS weights and ESS weights following multiplication of the weights.

Table 4 – Distribution of weights *zpers_weight*, *gewicht*, and w_{NR}

weight	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
w_{NR}	8.84	17.94	34.55	100.45	4 687.63

We also carried out a check on the sum of the weights. In principle, the sum of the weights w_{NR} should provide an estimate of the size of the target population, which we hope, is not too far from reality. Table 5 shows the size of the target population including apprentices, whom we are unable to filter out in file f1b, the rounded sum of the weights w_{NR} on file f2b, and the rounded sum of the weights *zpers_weight* (before calibration) on file f2a excluding apprentices.

Table 5 – Sum of weights and population size

Size of target population	1 842 703
Sum of <i>weights</i> w_{NR}	1 618 701
Sum of weights <i>zpers_weight</i>	1 630 812

We see that the deviation from the target population is quite large, but it should be borne in mind that the target population includes apprentices, whereas this is not the case in file f2b.

4.3 Preparation of calibration totals and first calibration

In order to determine which variables (or crossed variables) will be used in the two successive calibrations, we studied their sizes on the f2b sample. A minimum threshold of 50 observations was set per cell and if necessary when the size of a category was not sufficient, some aggregation was necessary. We describe the choices made below.

Creation of the enterprise size category variable

We studied the distribution of the *emptot* variable in the population to determine one or more categorisations for the enterprise size category variable. An initial categorisation, based on the quartiles and thus resulting in four categories of roughly equal size in the population, was examined. Table 6 shows the number of observations in the frame and in the f2b sample.

Table 6 – The size variable			
size of enterprise	value	Number of observations in the frame	Number of observations in the f2b sample
[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

A more refined categorisation was defined with the following categories and respective number of observations from Table 7.

Table 7 – The refined size variable			
size of enterprise	value	Number of observations in the frame	Number of observations in the f2b sample
≤ 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; 1'700]	6	212 552	4 630
(1 700; 10 000]	7	235 262	5 948
≥ 10 000	8	229 075	7 492

Creation of the age category variable

In the data, we already have the ageclass variable that we have converted numerically with the following categories:

- if ageclass equals "19-30 years" the new cl_age variable = 1;
- if ageclass equals "31-40 years" the new cl_age variable = 2;
- if ageclass equals "41-50 years" the new cl_age variable = 3;
- if ageclass equals "51-64 years" the new cl_age variable = 4;

We took into account the crossing of this variable with the sex variable. Table 8 shows the number of observations in the frame and in the f2b sample of this crossed variable:

Table 8 – Number of observations for sex by age category

sex	cl_age	Number of observations in frame	Number of observations in f2b sample
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

The various NOGA variables

We also analysed the number of observations in the calibration frame and in the sample for several NOGA variables, as well as for the NOGA by sex variables.

- variable noga082_Statent (NOGA two-digit)
The sample contained too many NOGA categories with a low number of observations. This variable was therefore discarded.
- variable noga081_Statent (NOGA section: 18 categories in the calibration frame)
The size of NOGA section B is 19, so too small. Also, when crossing the variable noga081_Statent with sex, we noted some cases where we had zero observations in certain NOGA*sex intersections. This resulted in the grouping of NOGA sections B with C and D with E, which allowed a sufficient number of observations to be retained, even when crossing with sex.
- variable noga082_50 (NOGA FS050: 47 categories in the calibration frame).
For this variable, with more refined categorisations than the variable noga081_Statent, only for the category NOGA "05.9" was the number of observations lower than 50. This NOGA was therefore grouped with NOGA "31.3" to create a separate category.

The selected variables

The variables we ultimately selected for the calibration are:

- grouped NOGA FS050 (nog50_cal_gr),
- sex by age category (sexe*cl_age),
- sex by size of enterprise by class with 8 categories (sexe*taille_fine) and
- sex by the grouped NOGA section (sexe*noga_s_gr).

First calibration

We described the first stage of calibration in section 3.3. At the end of this stage, we obtained a set of weights, denoted by w_{CDCRS} , for each unit in the f2a file. Table 9 shows the distribution of initial weights (zpers_weight) and calibrated weights (w_{CDCRS}). We see that the two distributions are quite similar.

Table 9 – Distribution of weights zpers_weight and calibrated weights w_{CDCRS}

weight	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 Final calibration

The calibration of the weights w_{NR} from the f2b file (with iwaz=0) to the totals calculated in the previous step provided us with a new set of weights (w_{cal}) for 29 645 observations. Table 10 shows the distribution of initial weights (w_{NR}) and calibrated weights (w_{cal}). We see that the two distributions are very close and that the weight w_{cal} and the weight w_{NR} are quite dispersed. This primarily results from the non-response adjustment and more precisely from the way in which the sample f2b is obtained – as a linkage of two samples (SS and ESS), whose response mechanisms are independent, very different, and are combined.

Table 10 – Distribution of weights w_{NR} and w_{cal}

weight	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

After deleting the cases where iwaz=0 (without modifying the weights) we obtain a set of weights for 28 210 observations. This file and these weights are then considered the final weights and are delivered to the section.

5 Example of variance calculation

By way of illustration, we calculated the median of the *mbls* income (from the ESS) and of the standardised income *cdcrsref2*. These are weighted medians (using the calibration weights w_{cal}). They are calculated, as are their respective variances, using the SAS PROC SURVEYMEANS procedure, taking into account, for both incomes, the cluster effect of the ESS (using the variable *entid_n* from the ESS). To calculate the variance of the constructed *cdcrsref2* variable, we also

need to take account of the cluster effect due to the fact that we do not retain all the persons from the SS in the sample, but only those who we also find in the ESS. The ESS cluster effect is therefore transmitted to the whole linked database. In addition, the calibration effect is not taken into account in the variance calculation. Table 11 shows estimated (rounded) values for the weighted medians, standard error (SE) and the limits of the estimated confidence intervals, which are denoted by ICL and ICU, respectively.

It should be noted that the procedures we use to calculate the variance allow us to estimate the variance due to the sampling design, as we do in the 'conventional' FSO setting, where the variables of interest are not generally constructed on a random basis. However, the constructed variable `cdcrsref2` (standardised income) is by construction a random variable. When calculating the variance, we do not integrate the effect due to the construction of this income. It was not possible to integrate this part of additional variance in the time allotted, but it could be interesting – in a subsequent stage of the project – to examine the feasibility of including this source of additional variability in the precision calculations.

Table 11 – Estimated weighted medians

variable	estimated median	SE	ICL	ICU
MBLS	6726.07	56.56	6615.20	6836.94
cdcrsref2	6739.10	59.01	6623.42	6854.77

6 Limitations of the weighting

This work was done as part of the LABB analyses to enable us to conduct analyses on the linked database. However, this weighting has its limits. These include:

- The weighting only allows an extrapolation to the target population, which only represents 52% of that of the separate databases.
- The final weights are very dispersed. This may prove problematic in terms of accuracy in certain sub-populations.
- The weights used for non-response adjustment are the final ESS and SS weights, although ideally weights prior to calibration would have been preferable.