Calif Manual

v4.0

CONTENTS

1 CALIBRATION APPROACH

In most cases, parameters derived from statistical surveys are just estimates of real values. Sampling weights that comply with the sampling design play a crucial role, enabling outcomes of the whole population without having knowledge about it. However, some auxiliary variables, at least their total values, are often known and available for the whole population and these are a part of the survey design. An inferential step is then beneficial. The idea is to modify the sampling weights so that the population totals of auxiliary variables match exactly to those inferred using new weights and this modification is minimal. This technique proposed by Devill and Särndal [1] is called calibration and can enhance precision as well as consistence of estimate procedure. As [2] states, "Calibration is a procedure that can be used to incorporate auxiliary data. This procedure adjusts the sampling weights by multipliers known as calibration factors, that make the estimates agree with known totals. The resulting weights are called calibration weights. These calibration weights will generally result in estimates that are design consistent, and that have a smaller variance than the Horvitz-Thompson estimator." The main advantage of calibration is then to enhance estimates precision, especially when auxiliary variables are correlated with the study variable. The calibration brings consistency to the weight system, so that the population totals throughout the several surveys agree with each other and an additional improved accuracy could be achieved (via lower variance and reduced nonresponse bias).

1.1 CALIBRATION ESTIMATOR

Let us consider a population U with N units. The probability sampling S of size n is undertaken. Every unit in S has design sampling weight and it is equal to $d_k = \frac{1}{\pi_k}$ where π_k is the inclusion probability of unit $k \in S$, possibly adjusted for nonresponse. The objective is to estimate the population total of a study variable y, denoted as $Y = \sum_{k=1}^{N} y_k$. The common estimator is the Horvitz-Thompson unbiased estimator $\hat{Y}_{HT} = \sum_{k \in S} d_k y_k$. However, when auxiliary information is available, another estimator could be used to gain efficiency.

Assume J auxiliary variables and their population totals $X_j = \sum_{k \in U} x_{kj}$. These are usual in statistical production when totals are known from administrative sources and censuses. In some cases, also other, broader surveys could be used as a source for the known population totals.

The main objective of the calibration approach is to reproduce the new weights for each $k \in S$ that confirm auxiliary totals and differ minimally from design weights d_k. These weights are independent of y, therefore totals of many study variables could be estimated. Calibration approach doesn't rely on a specific model; it only operates with information to calibrate on.

For almost each case the H-T estimator of auxiliary total is different from the real known value, that means

$$\sum_{k \in S} d_k x_{kj} \neq X_j$$

Let w_k denote the calibration weight of element $k \in S$. The calibration estimator of a study total is

$$\hat{Y}_{CAL} = \sum_{k \in S} w_k y_k$$

while calibration constraints are fulfilled

$$\sum_{k \in S} w_k x_{kj} = X_j$$

for all $j = 1, \dots, J$.

The distance between design and calibration weights is expressed via *distance function*. Let $r_k = \frac{w_k}{d_k}$ denote the **quotients** of these weights (known as calibration factors or g-weights). Then the distance function $G(r_k)$ is a nonnegative convex function of r_k with minimum in 1 (so when calibration and initial weights are equal). As stated in [4], to find calibration weights we have to find a minimum of the equation

$$L = d^T G(r) - \lambda^T (x^T dr - X)$$

where $d^T = (d_1, \dots, d_n)$, $w = (w_1, \dots, w_n)^T$, $X = (X_1, \dots, X_J)^T$, $\lambda^T = (\lambda_1, \dots, \lambda_J)$ is a vector of Lagrange multipliers and x is a $n \times J$ matrix of auxiliary variables.

By taking partial derivatives of L we get

$$w_k = d_k F(\lambda^T x_k)$$

where $F(\cdot)$ is the inverse function to derivative of $G(r_k)$. This gives

$$\sum_{k \in S} d_k F(\lambda^T x_k) x_{kj} = X_j$$

This system can be solved by several optimization methods taking $\left(w_1^0, \dots, w_n^0, \lambda_1^0, \dots, \lambda_J^0\right) = (d_1, \dots, d_n, 0, \dots, 0)$ as starting values.

According to [7], the variance of the Horvitz-Thompson estimate $\hat{Y}_{HT}$ can be estimated by

$$\widehat{Var}\left(\hat{Y}_{HT}\right) = \sum_{i \in S} \sum_{j \in S} \frac{\left(\pi_{ij} - \pi_i \pi_j\right)}{\pi_{ij}} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j}$$

and as stated in [1] variance estimation of calibration estimator is

$$\widehat{Var}\left(\hat{Y}_{CAL}\right) = \sum_{i \in S} \sum_{j \in S} \frac{\left(\pi_{ij} - \pi_i \pi_j\right)}{\pi_{ij}} \left(w_i e_i\right)\left(w_j e_j\right)$$

where e_k are the residuals of k. Second order inclusion probabilities π_{ij} are difficult to compute, but can be approximated by f.i. Hajek approximation, as provided by `Pkl.Hajek.s` function of **samplingVarEst** package.

1.2 DISTANCE FUNCTIONS

Several functions are commonly used for measuring the distance between initial and calibration weights. We consider 4 of them in Calif.

– **linear** – this function is often used due to its ability to find exact solution (if the solution exists). If no solution is found it is worthless to try other functions. On the other hand, resulting weights could be negative, which seems to be inconvenient for statistical production purposes. However, linear distance function is a proper „tester" before applying other functions, just to see e.g. what are the possibilities for the lower and upper bounds or what minimal deviation is achievable. The function itself is defined as

$$G(r) = \frac{1}{2}(r - 1)^2 \qquad \Rightarrow \qquad F(u) = 1 + u$$

– **raking ratio** – nonlinear distance function that circumvents the „negative weights" problem. Not to be so optimistic, also raking ratio brings some difficulties, because weights less than 1 could appear. Also can be used as a good „tester", to see if some acceptable solution with adequate bounds attained is possible

$$G(r) = r \ln r - r + 1 \qquad \Rightarrow \qquad F(u) = e^u$$

- **logit** – bounded version of raking ratio. User is able to enter lower and upper bounds for quotient $r_k = \frac{w_k}{d_k}$, differences between initial and calibration weights as well as the condition that weights are not less than 1 can be controlled. It gives

$$Ld_k \leq w_k \leq Ud_k$$

User must be aware of range allowed for calibration weights, tense bounds often lead to unsolvable system and increase average difference applied to each initial weight d_k. The goal is to seek an appropriate balance between maximum distance applied, its distribution and precision of $\sum_{k \in S} w_k x_{kj} = X_j$. The function is defined as

$$G(r) = \frac{1}{A}\left[(r - L)\ln\frac{r - L}{1 - L} + (U - r)\ln\frac{U - r}{U - 1}\right]$$

$$F(u) = \frac{L(U - 1) + U(1 - L)e^{Au}}{(U - 1) + (1 - L)e^{Au}} \quad \text{where } A = \frac{U - L}{(1 - L)(U - 1)}$$

- **linear bounded** – is the bounded version of the linear method. User has to specify the lower and upper bounds for $r_k = \frac{w_k}{d_k}$

$$G(r) = \begin{cases} \frac{1}{2}(r - 1)^2 & L \leq r \leq U \\ +\infty & \text{otherwise} \end{cases}$$

2   WHAT IS CALIF?

Several software tools deal with calibration. Many of them run under commercial software and are not so user-friendly. The problem of no exact solution is also often encountered. The Statistical Office of the Slovak Republic prepared some open-source versions of Calif in the past that were able to circumvent all these inconveniences and offered user-friendly GUI environment. Moreover, they were very powerful in seeking appropriate and even approximate solutions (some tools can find just the accurate solution but it rarely exists, especially for many auxiliary constraints). However, they reached their limits in graphical user interface appearence and operability, which could have discouraged some users to work with it. Calif 4.0 is a new Shiny web application with modern and attractive design, is very easy to use and very fast, offers many features that can help users to find the best solution whilst maintaining time-proven techniques. The whole application is built under the **shiny** package, while incorporating `calib` function from package

sampling together with other equation solver (function `nleqslv` from package **nleqslv**). The diversity of ways how to find a good solution makes Calif a very interesting and comfortable tool. The various options of Calif require from the user some level of expertise. However, the easy-to-use graphical user interface makes it intuitive and comfortable to work with it. Calif runs in several web browsers locally, without any concerns of leaving sensitive data outside currently used PC.

Calif is the Shiny web application that can be either downloaded from the SO SR's webpage and sourced to the R or run directly from GitHub Repository https://github.com/SO-SR/Calif. The installation process consists of:

– installing R. It can be downloaded from
   https://cran.r-project.org//

– installing required packages – **shiny, sampling, nleqslv** and **haven**. Packages are installed (together with all dependencies) by entering
   `install.packages('`*`package name`*`')` in the console. If you run Calif for the first time, the packages should be installed automatically (if your proxy settings allow it).
   If you have some troubles with proxy settings try first
   `setInternet2(TRUE)`
   or contact your IT department. Installation of packages is needed only once.

– <u>either</u> sourcing downloadable Calif code and entering **`calif()`** in the console. If you use just the R, choose 'File **->** Source R code' and select the Calif v4.0.R file. If you use R Studio, choose 'Code -> Source File' (or Ctrl + Shift + O) and select the Calif v4.0.R file. Sourcing is needed each time the R is opened (and you are going to use Calif).

– <u>or</u> running Calif directly from GitHub Repository via the `shiny::runGitHub` command from the SO SR's webpage. In order to work always with the latest version, this option is preferable.

– to close the application, just close the browser window and click the **STOP** button in R.

3   DATA PREPARATION

Each calibration process has to be properly prepared, to gain adequate results. You first need to discover available population information that can be used as auxiliary totals in calibration process. Harmonization of several statistical surveys

is often demanded. Parameters chosen as auxiliary information have to be correlated with the study variables as much as possible. E.g. in social surveys, usual auxiliary information is sex, age, region, education, economic status; in business surveys it could be turnover, number of employees, number of enterprises etc. In any case you need to know the actual population totals of selected variables (or at least very precise estimates), possibly on the level of stratification taken into account.

There are no special requirements for the data structure. You have to load the data into Calif in **.txt**, **.csv** or **.sas7bdat** format with the heading in the first row. There is no need to delete unused columns prior to calibration; Calif takes into account just the essential columns, specified in the Data tab.

Required columns in the data are:
- categorical and/or numerical auxiliary calibration variables
- initial weights

Optional columns in the data are:
- stratification
- variables used for computation of various indicators

3.1 TABLE OF TOTALS

The table of auxiliary totals has to be in line with the predefined structure. Separate columns of the table refer to separate auxiliary variables in the data, however, there could be also another columns that are not used for calibration – they will be simply skipped in the process. In the first row there must be a heading with the column names that match exactly to the names of auxiliary variables in the data. In case of categorical auxiliary variable, you have to specify population totals for each category (e.g. number of men and number of women) in separate columns. The names are constructed **by pasting the variable name and the category name with the underscore**, e.g. *sex_male*, *sex_female*. The order of the columns in the table of totals is irrelevant; the only requirement is that if you run stratified calibration, stratum names have to be specified in the first column (as you will see later).

Example 1. Imagine the data with two numerical and two categorical auxiliary variables. The numerical auxiliary variables are Turnover and Salaries, we know the population totals of both of them. The categorical auxiliary variables are NACE and Size with several categories. Columns Type and Prob are just

additional and not interesting for calibration. Then the data and the table of totals can look like

Table 1. Example of some data, just first 6 rows shown

ID	NACE	Size	Turnover	Salaries	Type	Prob	Stratum	Weight
1	C	1	895000	87000	4	0.065	East	15.4
2	D	3	12878000	7254000	8	0.0405	West	24.7
3	C	2	1658000	1200000	2	0.089	East	11.2
4	C	3	11451000	5412000	2	0.04	South	25
5	G	1	960000	241000	2	0.0752	Central	13.3
6	G	3	19630000	13974000	1	0.135	South	7.4

Table 2. Example of table of totals

NACE_C	NACE_D	NACE_G	Turnover	Salaries	Size_1	Size_3	Size_2
412	130	378	560812000	278200000	627	203	90

As you can see, the order of the columns is not equivalent to the order of auxiliary variables in the data; the only criterion is the names matching. If we used stratified calibration, table of totals could look like

Table 3. Example of table of totals with stratification

	NACE_C	NACE_D	NACE_G	Turnover	Salaries	Size_1	Size_3	Size_2
North	87	30	90	196633000	58817000	208	67	30
East	74	42	41	57999000	27143000	99	29	16
South	115	19	91	112541000	63542000	205	71	27
West	81	27	54	93624000	44578000	72	0	12
Central	55	12	102	100015000	84120000	43	14	5

As we can see in West stratum, if some category of a categorical variable is not represented in some stratum, there should be a zero in a corresponding cell. Population total for a numerical variable cannot be equal to zero.

Summary of all data preparation requirements:

– data

- heading with variable names in the first row
- at least one auxiliary calibration variable
- column with initial weights required

- columns irrelevant for calibration can be present; they are omitted in the process

- table of totals
  - order of columns not important
  - heading in the first row
  - column names must match exactly to the names of auxiliary variables in the data
  - there could be columns present in the table of totals that are not used for calibration, in that case, their names must be different from those selected as auxiliary variables in the main window
  - for categorical variables, population totals specified for each category; the column name in the form **variable_category**
  - in case of stratified calibration, separate rows pertain to population totals for each stratum; the names of strata in the first column
  - if for some stratum there is no population representation of a certain category, just insert a zero in the corresponding cell of the table of totals

3.2 Two-stage calibration

Multistage calibration usually pertains to social surveys. If we intend to get so-called *integrated weights*, i.e. weights that are constructed such that each member of the first stage unit (FSU, usually household) has the same weight as the unit itself (these members are second stage units - SSUs; usually members of households). To calibrate both stages simultaneously, an inferential step has to come.

All you need to do is turn the file at the individual level into a file at the household level, and, after that, calibrate just the household file. The process is as follows:

1. at the beginning, you have the sample file at the first stage (household level) with initial design weights (possibly adjusted for nonresponse)
2. join these weights to SSUs (individuals). Each SSU will get the weight of the FSU to which it belongs
3. put together second stage auxiliary variables (and possibly some indicator variables that will be monitored during calibration or other information) in the second stage file (individual level). Now you have the second stage file with some numerical and/or some categorical variables

11

4. for each of the k categories of certain categorical auxiliary variable in the second stage file create k dummy variables (i.e. 1 when $y_i = k$ and 0 otherwise)

5. summarise all auxiliary numerical variables and also dummy variables within each FSU (and possibly some indicator variables), while keeping the essential information on the first stage (like Ids, other variables, ...)

6. now you turned the second stage file into a first stage file (by summarising SSUs within each FSU); each former auxiliary variable is now numerical (usually, summed dummy variables indicate the number of individuals with certain characteristics within each household). Join the first stage auxiliary variables and some other essential information (stratification variables, Ids, ...)

7. the file is prepared for calibration

You can easily do the 5. step by using the `summarise` function along with `group_by` function of the **dplyr** package.

Example 2. Let us focus on the EU-SILC datafile. In Statistical Office of the Slovak republic, the calibration criteria are:

– sex by 6 age groups (12 categories of 1 categorical variable) – second stage

– 5 categorical variables related to economic activity – second stage

– households by members (1 categorical variable with 5 categories) – first stage

– NUTS3 stratification (8 strata)

The individual (second stage) file looks like:

Table 4. Insight into the artificial EU-SILC individual level file, step 3

HH_ID	Sex	Age group	Ec. activity	Region	Income	Weight
1	1	3	1	1	1140	675.42
1	2	3	1	1	977	675.42
2	1	6	5	7	415	624.11
3	1	4	2	3	179	691.74
3	2	4	1	3	1052	691.74
3	2	2	4	3	0	691.74
4	1	3	1	5	841	712.49
4	2	3	3	5	2115	712.49
4	1	1	4	5	0	712.49
4	1	1	4	5	0	712.49

As you can see, there are categorical auxiliary variables (sex, age, economic activity), numerical variable (but not deemed for calibration, it is just used for indicator monitoring), household ID, residence and initial weights column. It is clear that each member of a particular household has the same weight and also the same region. By combining sex with age and executing step 4 of the process, we get

Table 5. Individual level file, dummy variables created for each category

HH_ID	s1a1	s1a2	s1a3	s1a4	s1a5	s1a6	s2a1	s2a2	s2a3	s2a4	s2a5	s2a6
1	0	0	1	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	1	0	0	0
2	0	0	0	0	0	1	0	0	0	0	0	0
3	0	0	0	1	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	1	0	0
3	0	0	0	0	0	0	0	1	0	0	0	0
4	0	0	1	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	1	0	0	0
4	1	0	0	0	0	0	0	0	0	0	0	0
4	1	0	0	0	0	0	0	0	0	0	0	0

econ1	econ2	econ3	econ4	econ5	Region	Income	Weight
1	0	0	0	0	1	1140	675.42
1	0	0	0	0	1	977	675.42
0	0	0	0	1	7	415	624.11
0	1	0	0	0	3	179	691.74
1	0	0	0	0	3	1052	691.74
0	0	0	1	0	3	0	691.74
1	0	0	0	0	5	841	712.49
0	0	1	0	0	5	2115	712.49
0	0	0	1	0	5	0	712.49
0	0	0	1	0	5	0	712.49

This table shows the classification of each individual at the second stage. Number 1 in the table indicates the individual's affiliation to a certain category (obviously, numerical variables remain unchanged, just categorical auxiliary variables are coded into dummy variables). In the next step you have to summarise auxiliary variables within each household, so that categorical variables will become numerical variables, indicating the number of individuals with certain

characteristic (e.g. sex=1, age=4, econ=2) within a household. It can be done by the following command in R, but the number of possible ways is huge. First you need to install the **dplyr** package.

```
file_name %>% group_by(HH_ID) %>%
    summarise(s1a1 = sum(s1a1), s1a2 = sum(s1a2), s1a3 = sum(s1a3),
    s1a4 = sum(s1a4), s1a5 = sum(s1a5), s1a6 = sum(s1a6), s2a1 =
    sum(s2a1), s2a2 = sum(s2a2), s2a3 = sum(s2a3), s2a4 = sum(s2a4),
    s2a5 = sum(s2a5), s2a6 = sum(s2a6), econ1 = sum(econ1), econ2 =
    sum(econ2), econ3 = sum(econ3), econ4 = sum(econ4), econ5 =
    sum(econ5), Income = sum(Income), Region = first(Region), Weight
    = first(Weight))
```

The result is

Table 6. Individual level file summed into household level file, stage 5

HH_ID	s1a1	s1a2	s1a3	s1a4	s1a5	s1a6	s2a1	s2a2	s2a3	s2a4	s2a5	s2a6
1	0	0	1	0	0	0	0	0	1	0	0	0
2	0	0	0	0	0	1	0	0	0	0	0	0
3	0	0	0	1	0	0	0	1	0	1	0	0
4	2	0	1	0	0	0	0	0	1	0	0	0

econ1	econ2	econ3	econ4	econ5	Region	Income	Weight
2	0	0	0	0	1	2117	675.42
0	0	0	0	1	7	415	624.11
1	1	0	1	0	3	1231	691.74
1	0	1	2	0	5	2956	712.49

In the last step you need to add auxiliary variables at the household level. In this example, it is just the variable "household size" (top coded by number 5). You can join it in the last step by standard way, or by inserting `Members = n()` as a last argument of the above-mentioned `summarise` function. The result is

Table 7. File at the household level prepared for two-stage calibration

HH_ID	s1a1	s1a2	s1a3	s1a4	s1a5	s1a6	s2a1	s2a2	s2a3	s2a4	s2a5	s2a6
1	0	0	1	0	0	0	0	0	1	0	0	0
2	0	0	0	0	0	1	0	0	0	0	0	0
3	0	0	0	1	0	0	0	1	0	1	0	0
4	2	0	1	0	0	0	0	0	1	0	0	0

econ1	econ2	econ3	econ4	econ5	Region	Income	Weight	Members
2	0	0	0	0	1	2117	675.42	2
0	0	0	0	1	7	415	624.11	1
1	1	0	1	0	3	1231	691.74	3
1	0	1	2	0	5	2956	712.49	4

Each of the auxiliary variables is now at the household level. The only one categorical variable - Members - indicates the size of the household (in relation to the number of its members). The table of totals could look like:

Table 8. Artificial table of totals for EU-SILC, just 4 NUTS3 levels shown

	Members_1	Members_2	Members_3	Members_4	Members_5	s1a1	s1a2
1	58791	64841	48099	48019	15489	45907	34665
2	48547	50982	40169	47186	26081	43774	42256
3	59378	70510	48030	57141	26692	52052	47017
4	53313	55249	42573	49341	37621	60506	52184

s1a3	s1a4	s1a5	s1a6	s2a1	s2a2	s2a3	s2a4	s2a5	s2a6
99262	39020	38788	29654	43365	33689	101594	45010	47246	48337
94469	43697	37300	29888	41784	40310	89073	43526	40787	47322
109497	50025	43107	32739	48754	44584	104615	50507	49272	57395
109673	48127	39322	28641	57549	50221	103446	47445	43653	48834

econ1	econ2	econ3	econ4	econ5
299642	257426	38793	16336	112073
286562	246088	37265	31657	105236
303936	254999	49319	48456	133943
314721	257019	62527	38998	108660

After the calibration, household's weight will be assigned to each of its individuals. To prove correctness, as [8] presents, if S_M is a sample of households, S_I a sample of individuals, $d_{mi} = d_m$ are design weights, $X = \sum x_m$ are auxiliary population totals at the household level and $Z = \sum z_i$ auxiliary totals at the individual level, combination of values of the individual dummy variables for each household m, i.e. $z_m = \sum z_{mi}$ makes these variables numerical. After this step, on household level there are auxiliary variables x_m (categorical) and z_m (numerical). Resulting weights are $w_m = w_{mi}$ and calibration is correct.

$$\sum_{m \in S_M} w_m x_m = X$$

$$\sum_{m \in S_M} \sum_{i \in S_I} w_{mi} z_{mi} = \sum_{m \in S_M} w_m \sum_{i \in S_I} z_{mi} = \sum_{m \in S_M} w_m z_m = Z$$

4 CALIF TOUR

This chapter will lead you through separate aspects of Calif 4.0 as well as through the calibration process. <u>**Prior to running Calif, make sure you have the latest version of your web browser.**</u>

4.1 OVERVIEW TAB

The first thing you can see after running Calif is the Overview tab. It displays main information on Calif, optimal calibration strategy, set up of the working directory, where the output files will be saved and some comments on bookmarking and output format. This tab can provide you with the general know-how you need to refresh from time to time, without reading this manual.

When working with Calif, if some unexpected error occurs, the application will close without warning. Although Calif can handle almost every possible mistake caused by lack of knowledge or by coincidence, it can't be ruled out that some previously untested error appears. Therefore, it is effective, when calibrating throughout strata, to regularly either bookmark or save your interim solution.
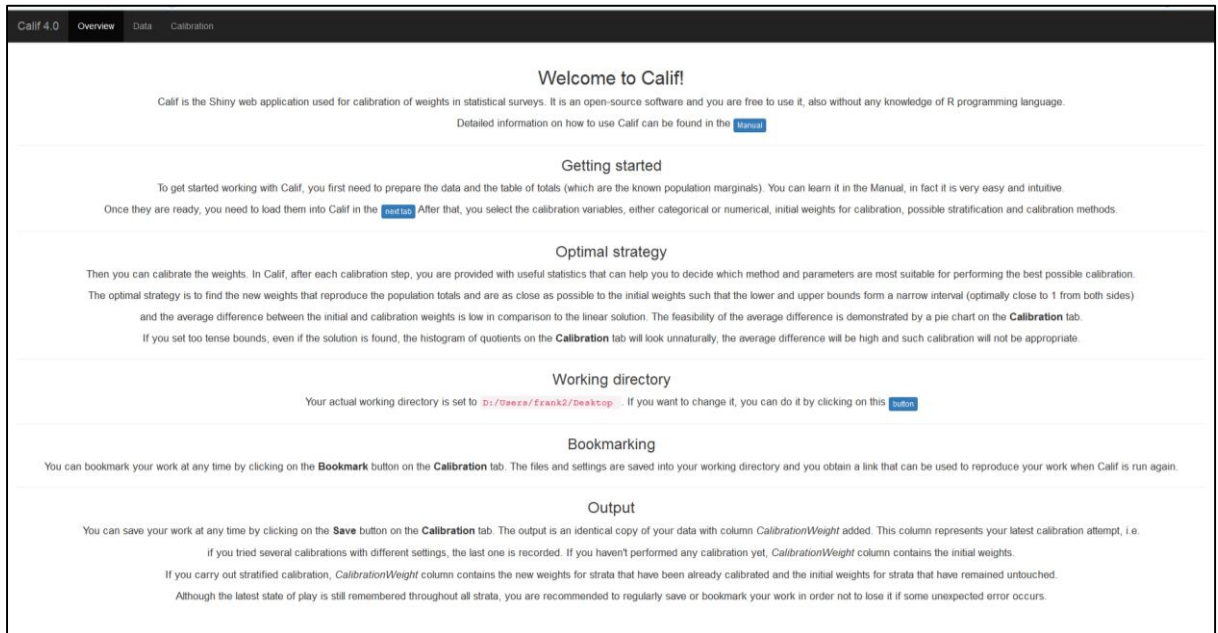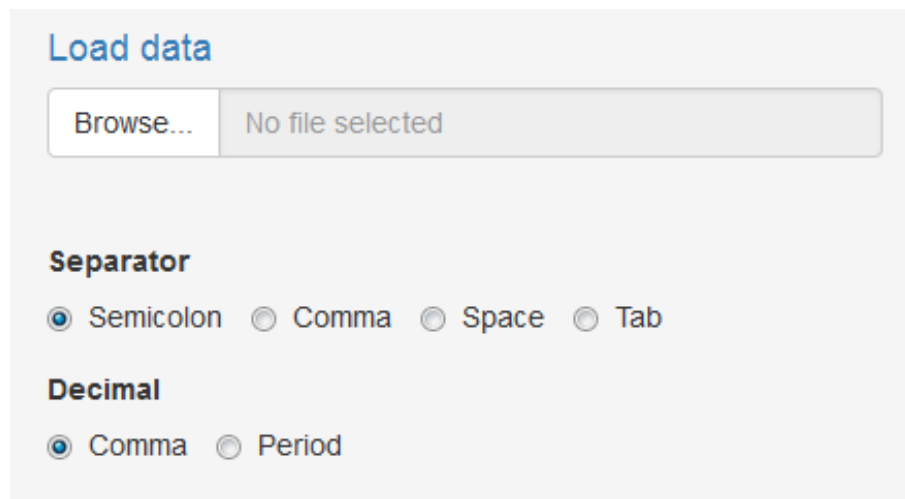
Once read all the information, you can click on the Data tab.

4.2 DATA TAB

4.2.1 IMPORT

You can import the data and the table of totals into Calif in **.txt** format (text files), **.csv** (comma separated files) or **.sas7bdat** (SAS datasets). Just click on the *Browse* button and then select corresponding separator and decimal (not needed for SAS data). The data displayed on the main panel will change responding to the change of the separator and decimal. Consequently, you can easily see if they are loaded properly or not. If the import or data structure is not correct, you will be informed by a message. As the table of totals needs to follow a pre-defined structure, if you choose unsuitable separator, a warning message could appear. After loading your data, feel free to play with them, filter or sort the columns.

Load data
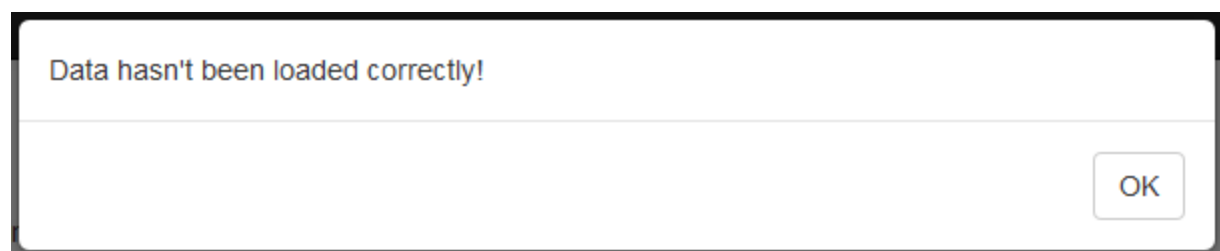
Browse...   No file selected

Separator

◉ Semicolon   ○ Comma   ○ Space   ○ Tab

Decimal

◉ Comma   ○ Period

Data hasn't been loaded correctly!

OK

Incorrect form of the Totals!

OK

4.2.2 EXPLORE VARIABLES

The new feature of Calif is the option to explore data variables. You can find the option on the main panel below the table of totals. In spite of being just a cosmetic service, it can give you a nice view on the variables' structure. They can be explored either as numerical or categorical. Displayed are the histogram (or barplot) and summary statistics (or frequency tables) with the number of missing values (NA's).
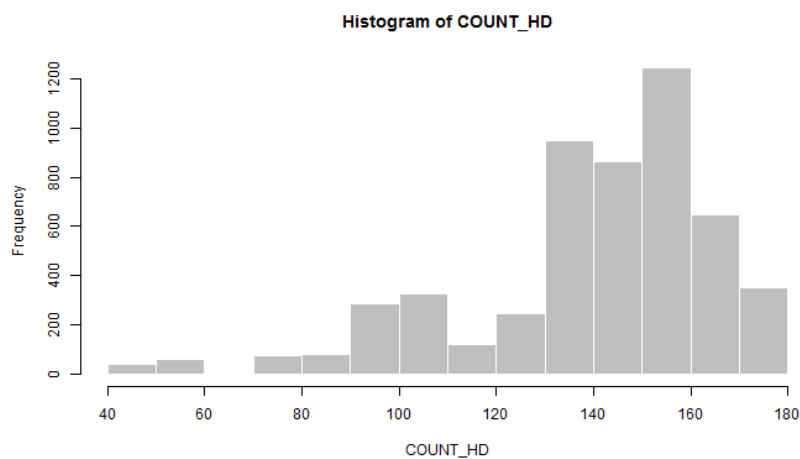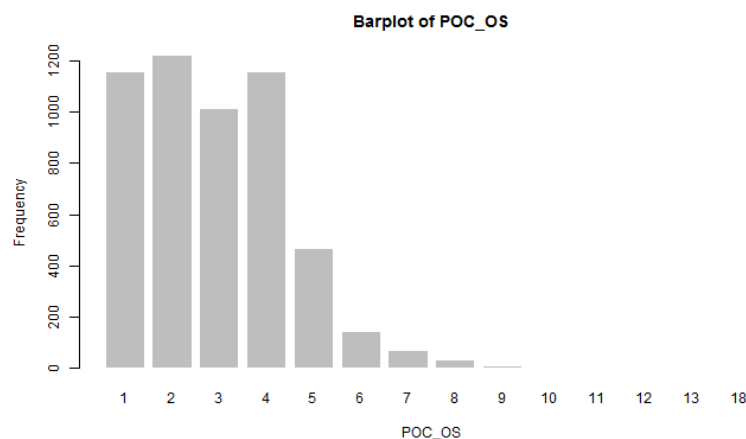
4.2.3 SPECIFICATION OF CALIBRATION VARIABLES

It is necessary to tell Calif which variables are deemed auxiliary numerical or categorical, which designate strata allocation, initial weights or are used for indicator calculation. Please bear in mind that other variables, which serve only as an additional information and are irrelevant for calibration, cannot be selected in the Calif window; they will be left out of the process. **Choose only those variables that are relevant for calibration** and have a counterpart in the table of totals (except indicators monitoring).

– *numerical variables* - select variables from the list that are deemed as numerical auxiliary calibration variables. For multiple selection use Shift/Ctrl keys or just move the mouse over several items. If the list seems to be incorrect, you have probably marked the wrong separator. Each selected variable has to have a matching column in the table of totals. If you operate with some other parts of Calif, selected items may gray out but it doesn't mean they are deselected – they are still chosen.

– *categorical variables* - select variables from the list that are deemed as categorical auxiliary calibration variables. Each category of selected variables has to have a matching column in the table of totals (see Chapter 3.1). Dummy variables in your data are not categorical but numerical.

– *weights* - choose the column that represents the initial weights.

– *stratification* - is stratification aspect taken into account? If so, stratification variable list will appear.

– *stratification variable* - if stratification, choose the column that represents the stratum allocation. Just one column can be selected.
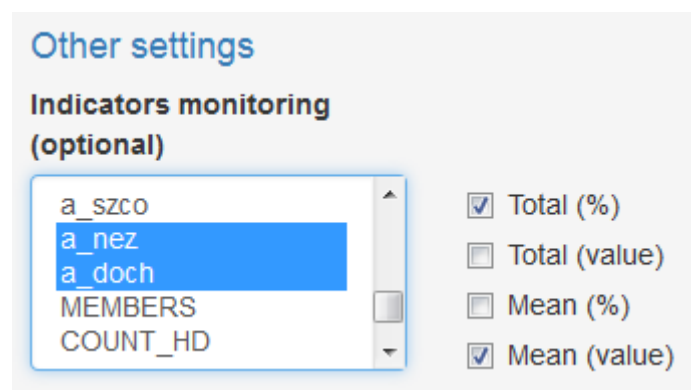
4.2.4 OTHER SETTINGS

– *indicators monitoring* - if you would like monitor some key indicators you can choose them from the list along with statistics that will be calculated. Weighted means and totals of selected columns can be calculated anytime, taking into account stratification aspect. Percentages can be calculated only if there is corresponding column in the table of totals that is used for computation (it will be the percentage of your estimate against the population value). In case you want to calculate a percentage mean, you need a column with population means in the table of totals. Bear in mind that only numerical variables can be selected for indicators monitoring.

– *missings* – if you consider some values in categorical variables as missings (f.i. 0), you can specify them here, separated by commas, so that Calif knows which values don't denote categories. These values will not be taken into account in calibration process. NA's are considered as missings in any case and sould not be specified at this place.

– *eliminate rows with missings* - should rows with missings, specified in the *missings* entry, be completely deleted? Be very careful with this option as it can cause severe problems in calibration process. Not recommended option.

– *tolerance* – desired accuracy for the iteration procedure

If you have specified all the necessary inputs, you can move to the next tab either by clicking the *Proceed* button or the *Calibration* tab. Your settings are checked

and if some mistake or inconsistence is detected, a warning is shown and you are requested to correct the settings.

4.3 CALIBRATION TAB

4.3.1 CHOOSE STRATA

If you set stratified calibration on the Data tab, the Choose strata list becomes available. If you wish to calibrate one or more strata separately and independently of each other, you can select them by using this option. If you left it unchanged, all strata are calibrated at once. Remember that fine tuning of your calibration can be accomplished by processing each stratum separately. **Calif always remembers the last calibration along with its settings that has been performed in each stratum.** If you omit some stratum, its weights remain unchanged. **If you calibrate some stratum several times, just its last setting is remembered. Further calibration of another stratum will not affect the previous result obtained for different stratum. If you want to confirm a calibration setting for some strata, run it and move to another strata.**

Hint: If you are not satisfied with the calibration results after several trials and want to keep the initial weights unchanged, try to run another calibration by using logit method with calib solver with some very strict bounds (e.g. 0.99 and 1.01). It often comes with the solution where all quotients are equal to 1, i.e. calibration weights are equal to initial weights. Therefore, in the output file the resulting weights will remain unchanged, however, the calif_settings file will contain the abovementioned method.

4.3.2 SHOW WITH INITIAL WEIGHTS

This is a very useful feature. It is recommended to use it each time before calibration of the whole file or specific strata, as it gives you a good view on the situation in your data with initial weights only. By clicking this button totals calculated prior to calibration (Horvitz-Thompson estimates) are shown, primary as a percentages (H-T estimate/known population total) but also displaying absolute values is a possibility. It is a very helpful utility to check for eventual incorrectness of the survey data or the population totals. Ideally, the proportions should be around 100% (in case of no non-sampling error).

Totals obtained

☐ Show obtained totals as values

Stratum	a_nez.Total (%)	a_nez.Mean (value)	MEMBERS_2	p11	p12	p13	p14	p15	p16	p21	p22	p23	p24	p25	p26	MEMBERS_1	MEMBERS_3
1.00	74.09%	0.05	100.00%	72.54%	129.90%	79.59%	104.73%	96.21%	119.40%	67.19%	153.85%	91.56%	123.42%	111.18%	100.42%	100.00%	100.00%
3.00	74.09%	0.11	100.00%	65.85%	134.87%	81.03%	103.47%	102.37%	86.91%	81.51%	148.21%	93.07%	120.66%	126.40%	104.72%	100.00%	100.00%
4.00	74.09%	0.14	100.00%	95.31%	131.96%	85.78%	105.14%	74.15%	114.25%	83.25%	134.44%	101.09%	117.76%	116.54%	106.53%	100.00%	100.00%
5.00	74.09%	0.12	100.00%	69.63%	134.72%	78.81%	117.46%	94.91%	88.51%	75.11%	138.32%	99.70%	136.37%	106.36%	104.30%	100.00%	100.00%
6.00	74.09%	0.20	100.00%	65.79%	142.37%	78.37%	106.81%	105.43%	112.01%	71.86%	114.25%	88.08%	130.83%	123.43%	116.45%	100.00%	100.00%
2.00	74.09%	0.11	100.00%	71.63%	113.08%	83.63%	101.06%	101.11%	107.92%	85.79%	130.22%	89.52%	110.99%	129.82%	121.59%	100.00%	100.00%
7.00	74.09%	0.20	100.00%	63.99%	124.81%	87.10%	108.29%	116.53%	117.30%	73.17%	118.58%	86.19%	116.95%	119.66%	130.63%	100.00%	100.00%
8.00	74.09%	0.20	100.00%	47.34%	115.65%	81.76%	110.13%	111.02%	119.06%	62.79%	122.09%	81.95%	119.02%	129.71%	119.24%	100.00%	100.00%

4.3.3 Method & Solver

Select one of the four distance functions mentioned in Chapter 1.2. Linear and raking ratio are unbounded whereas logit and linear bounded need to have specified lower and upper bounds. Once you select some bounded distance function, lower and upper bound entries appear. They are expressed by proportion **calibration weight/initial weight.** It is recommended to run linear first, just to see, if some feasible solution exists. If not (due to for example negative weights), continue with raking ratio. It is very likely that you end up with some bounded function, but remember to keep the average difference as low as possible and bounds close to 1. Bounded version will always have the average difference higher than unbounded version, therefore the linear's difference is a good navigation. Feasibility of the average difference is represented by a pie chart, which will be discussed below. Tense bounds imply high average difference, as there is no space to move and many weights are therefore pressed to them. Furthermore, if you see that by running linear method the bounds obtained are extreme (f.i. -500 and 1400) you can forget about a good bounded solution. In such a case, check the correctness of your population totals and eventually relax some of the calibration constraints. Don't forget that Calif always remembers the last calibration undertaken in each stratum. **If you want to confirm a calibration setting for some strata, run it and move to other strata.**

Two powerful optimizers are available in Calif. Function **calib** (from package sampling) is very fast and powerful, therefore set as default solver. Function **nleqslv** (package nleqslv) is a little bit slower, however in some scenarios can perform better – especially in business surveys with numerical variables, when small strata with just a few units are calibrated and differences between H-T estimates and known population totals are significant. For social surveys (strata with many units and auxiliary variables) calib performs better. Each of the solvers is able to find also an approximate solution (not only exact).

Regarding interaction between methods and solvers, following best practices are learnt:

– linear and raking ratio works equally with both solvers, where calib is faster

– logit performs better with nleqslv, as calib often comes with no solution (then calibration weights are equal to initial weights). Therefore nleqslv is set as default solver for logit, however it can be changed to calib

– linear bounded is most suitable with calib, analogically set as default solver

– lower bound should always be smaller than 1 and upper bound greater than 1

– the only allowed exception is the linear bounded method with calib when also other combinations for lower and upper bounds can be set (e.g. both less than or greater than 1)

– if some unsuitable combination of methods, solvers and/or bounds is submitted, warning appears and you are requested to re-specify your settings

4.3.4 CALIBRATE

If you are fine with strata you selected, totals with initial weights, method and solver, you can click the Calibrate button and wait for the result.
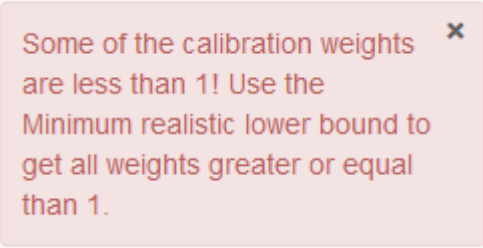
4.3.5 RESULTS – SUMMARY STATISTICS

Once calibration is done, several outputs are displayed. The first of them is the Results table with some important statistics that can help you to find most feasible solution.

Results

Initial weights interval	286.604	337.132
Calibration weights interval	85.981	761.426
Lower bound obtained	0.300	
Upper bound obtained	2.657	
Average weight quotient	1.000	
Average difference	98.330	
Minimum realistic lower bound	0.004	

- *initial weights interval* – the minimum and maximum values of initial weights.

- *calibration weights interval* – the minimum and maximum value of calibration weights.

- *lower and upper bound obtained* – the minimum and maximum value of the weight quotients. These tell you if your bounds were kept.

- *average weight quotient* – the mean of the weight quotients, usually close to 1 (if your data were correctly adjusted for nonresponse).

- *average difference* – calculated as $AD = \frac{1}{n}\sum|\text{calibration weight - initial weight}|$ and should be as low as possible. The lowest AD is usually for linear or sometimes raking ratio method.

- *minimum realistic lower bound* – with this value set as a lower bound, your weights will always be greater or equal than 1. In some cases, if you use it as a lower bound, new (higher) minimum realistic lower bound is calculated. If a calibration process ends up with some calibration weights below 1, a notification appears and your lower bound is automatically set to the minimum realistic lower bound for the next calibration. Nevertheless, you are allowed to change it to another value.

Some of the calibration weights are less than 1! Use the Minimum realistic lower bound to get all weights greater or equal than 1.

4.3.6 TOTALS OBTAINED

Totals obtained after calibration are shown in this window, together with calculated indicators. They should be as close to 100% as possible. If some totals are far away from 100%, you should try to relax the bounds a bit or to test some other solver/method. **This is the most important guidance of the calibration quality** and should be checked prior to other statistics. You can also check the absolute values in lieu of percentages, by clicking the *Show obtained totals as values* option.

4.3.7 AVERAGE DIFFERENCE FEASIBILITY

This is the new feature of Calif and it has to be treated with utmost care. Formerly, you were recommended to run linear calibration first, check and remember the average difference, then run bounded calibration and compare these differences. As the linear's AD is optimal for this set up (if it is too high, you have specified too many and too strict calibration constraints – population totals), the bounded's AD is higher but should not be too much. In such a case, the bounds are very strict and your resulting weights are too different from initial weights, which is not a good scenario. In Calif 4.0 the linear calibration is run automatically in the background, its' AD is remembered and then compared to the current AD, simply by calculating their quotient. The result is represented by a pie chart, it should not be significantly less than 100%. If the ADF is slightly greater than 100%, you have probably used the raking ratio method and it can be considered better than linear calibration in this case (if other statistics satisfy requirements). If the ADF is significantly greater than 100% (let's say 200%) you'd probably deal with insufficient calibration when totals obtained are not matched against the population totals and are close to the totals obtained with initial weights (therefore AD for this setting is low and ADF is high).

Hovering over the plot, a short information can be noticed on the left side.

Average difference feasibility

96.4%
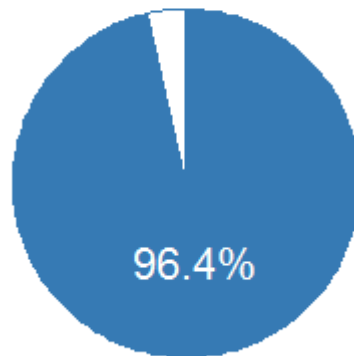
4.3.8 HISTOGRAM OF QUOTIENTS

It is possible to check if weight quotients (g-weights) calculated as *calibration weights/initial weights* are not pushed too much to predetermined bounds. If so, the bounds should be relaxed, since the solution like this is too distorted and the average difference is probably very high. Ideally, especially in social surveys, the histogram should follow the normal distribution.

Histogram of quotients
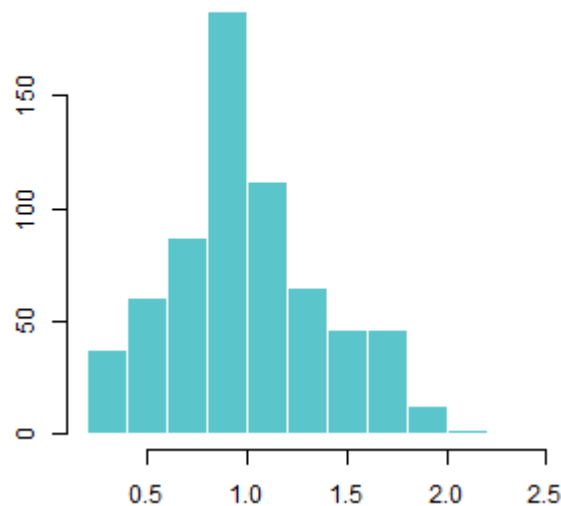
4.3.9 BOXPLOTS OF WEIGHTS

It illustrates the differences between initial and calibration weights. In the optimal calibration, the second box should be as narrow as possible, with some few outliers (in this case average difference is low, but bounds could be broad). Nevertheless,

these two boxplots should not be very different from each other. In the plot the line at point 1 is highlighted, to quickly see if some of the weights are below 1.

Boxplots of weights
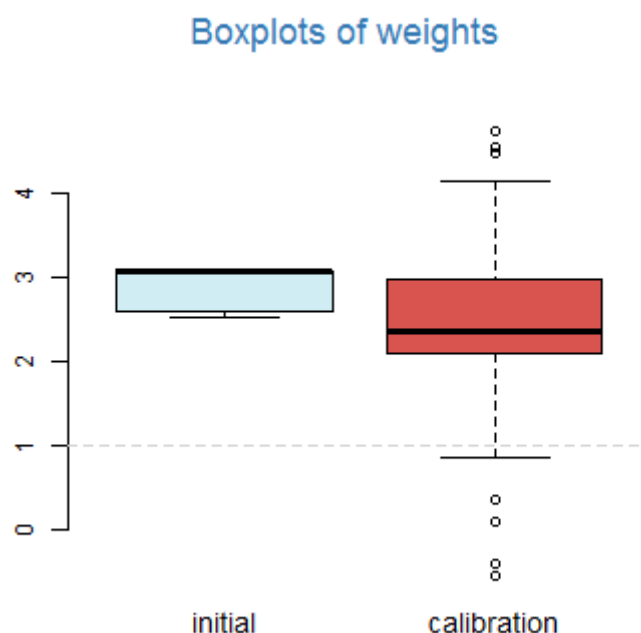
4.3.10 WEIGHTS & QUOTIENTS

At the bottom of the main panel the table with the row number, the initial and calibration weights and the weight quotients is shown. You are able to explore the weight and quotients for each row of the table, in order to see e.g. how many weights are below 1, which rows have the highest resulting weights or anything else. The table is sorted by the Calibration weights column but you are free to sort it anyhow. If this Quotients column is full of 1's, no calibration was performed (calibration weight equal to initial weight).

4.3.11 SAVE

You can save your work at any time, two output files are saved:

- the same file as it was loaded enriched with the last column containig calibration weights
- calibration settings that have been used in each stratum

After clicking the *Save* button, a modal dialog will pop up where you can set your Working directory (outputs will be saved there) and names of the outputs. By default, outputs with already existing names are overwritten without warning, but you are free to disable this option. In stratified calibration, if you save the outputs

Weights & Quotients

Show 20 ▾ entries

Row	Initial	Calibration	Quotients
158	286.6044	80.69619	0.28156
141	286.6044	83.12684	0.29004
1121	325.42069	83.23454	0.25578
13	275.12821	87.63458	0.31852
82	286.6044	89.85295	0.31351
5	275.12821	95.3722	0.34665
2345	340.74359	97.02143	0.28473
165	286.6044	101.77241	0.3551
31	275.12821	105.82202	0.38463
143	286.6044	111.45084	0.38887
967	324.98473	111.60212	0.34341

e.g. after each stratum, the resulting files will contain the calibration performed in each stratum, not just latest work done, i.e. all changes carried out between two saves are recorded and added to the previously saved files. If some strata were not calibrated, CalibrationWeight column in the output file would contain the initial weights for these strata. Although the latest state of play is still remembered throughout all strata, you are recommended to regularly save or bookmark your work in order not to lose it if some unexpected error occurs.

4.3.12 BOOKMARKING

Calif 4.0 allows for bookmarking. In contrast to saving, with the bookmark option you don't get the outputs but rather a URL. After clicking the *Bookmark* button, current state of the application is saved and the URL will restore the application with that state.

This is useful when you need to interrupt your work and continue with it later (mostly in stratified calibration) but do not want to lose your intermediate results. The state of Calif and back-end values are saved into shiny_bookmarks folder in your Working directory (when running from GitHub, it will be the default R's working directory and cannot be changed) and you get a URL, which can be used to restore the latest state. To proceed with your previous work just open a new session of Calif in your browser and paste the URL to the address bar. The latest state will be restored immediately and allows you to continue with your calibration.

In case you are using the **calif v4.0.R** file from the SO SR's webpage and you set a different working directory in previous session, you would need to copy and paste the shiny_bookmarks folder from the default directory to your working directory (set in the previous session), in order to restore the latest state properly.
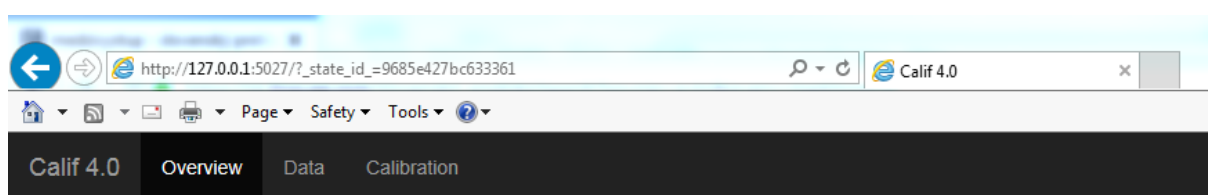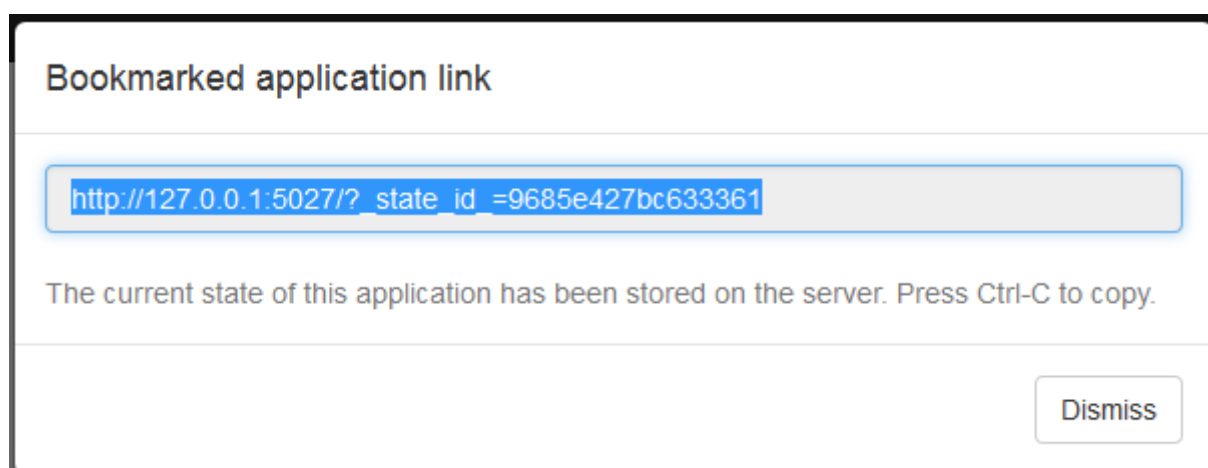
In a restored state, do not be concerned about empty *Load data* and *Load totals* fields, it is a normal behaviour. Data and totals are correctly loaded.

If you prefer saving over bookmarking, you can save your intermediate outputs in a usual way, then load them as a new data and choose the CalibrationWeight column in the *Weight* field. As the CalibrationWeight column is equal to the initial weight column in strata that have not been calibrated, the calibration will continue

correctly and the CalibrationWeight will be further adjusted. The output file will contain the CalibrationWeight2 column after another save.

5   OPTIMAL STRATEGY

At this place we would like to summarise all the information about the optimal strategy for calibration. This is a very important part because finding some kind of an optimal solution among plenty of possibilities is not straightforward and uniquely determined. Different solutions may yield different results and estimates, where some of them impose less bias than the others.

In Calif, after each calibration step, you are provided with useful statistics that can help you to decide which method and parameters are most suitable for performing the best possible calibration. **The optimal strategy is to find the new weights that reproduce the population totals and are as close as possible to the initial weights** such that the lower and upper bounds form a narrow interval (optimally close to 1 from both sides) and the average difference between the initial and calibration weights is low in comparison to the linear solution. If you set too tense bounds, even if the solution is found, the histogram of quotients on the Calibration tab will look unnaturally, the average difference will be high and such calibration will not be appropriate.

The procedure can be described as follows:

1. The linear method is run in current stratum. Despite negative weights, this calibration is optimal, yielding minimal average difference (in some cases raking ratio returns lower average difference than linear), i.e. even if the AD is high, it is the lowest possible for the current scenario. We get a picture of what to expect further in bounded calibration. If bounds obtained by linear method are e.g. -0.3 and 7, we can expect that e.g. 0.3 and 3 for logit or linear bounded could work. However, if bounds obtained by linear method are e.g. -10 and 20, we can hardly reach the population totals by some bounded method; they are too restrictive. In that case, it is necessary to adjust the calibration scenario, relax the population totals (reduce the number of auxiliary variables) or merge some strata. If some of the totals obtained is not equal to 100%, there is definitely something wrong in your data or table of totals. If bounds obtained by linear method are more than satisfactory and population totals completely reproduced, you can accept it and move to another strata.

2. The raking ratio method is run, in order to find out whether it is possible to find some solution with all weights greater or equal to 1. If so and bounds obtained are satisfactory, it is an acceptable solution.

3. Bounded methods are used. The linear method can give you a hint of where to start. You should try several bounds (and methods) such that:

   – the population totals are completely reproduced (100%)

   – the bounds interval should be as narrow as possible (ideally close to 1 from both sides)

   – **whereas** the average difference as close to linear solution as possible. This is calculated by the average difference feasibility and it should be as close to 100% as possible (explained in more detail in Chapter 4.3.7)

   – weight quotients should not be pushed off to the bounds; the histogram of weight quotients should resemble normal distribution (check the figures below)

4. When a feasible trade-off is found, solution is admitted.

The figures below describe the linear calibration (average difference lowest possible, broad bounds), the calibration with very strict bounds (too distorted, average difference high, narrow bounds) and feasible calibration (average difference not very high and bounds kept quite tight).

Linear method – weight quotients normally distributed, calibration weight similar to initial weights with some outliers, resulting weight below 1
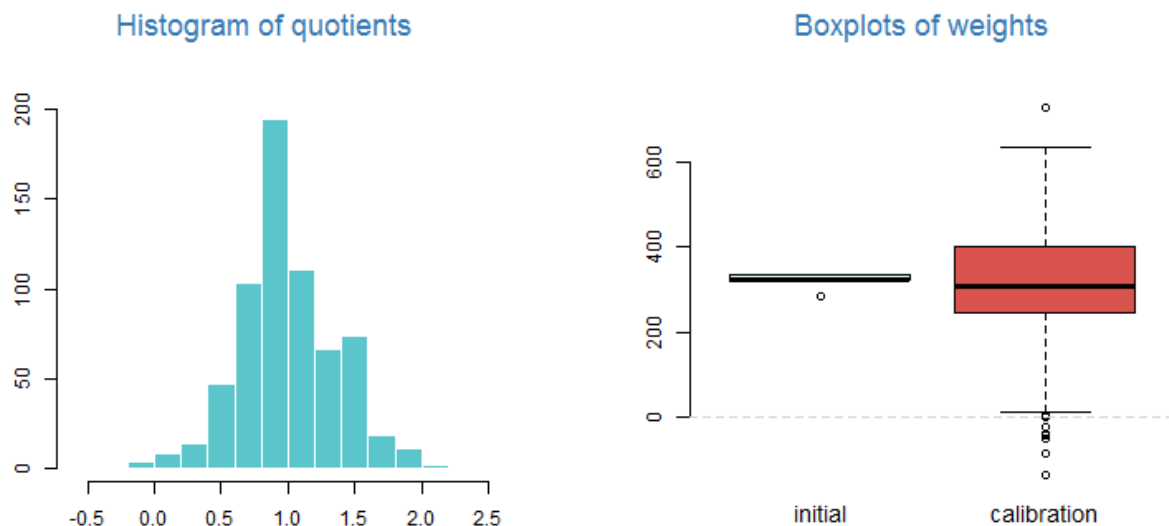
Strictly bounded method – weight quotients pushed off to the bounds, calibration weights very different from initial weights although greater than 1, unacceptable solution
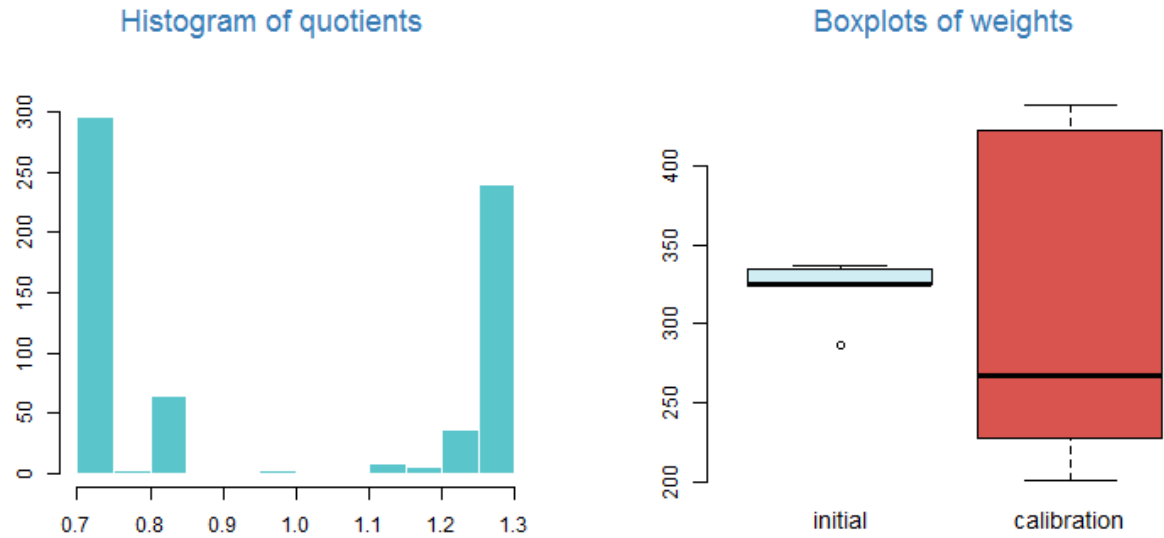
Finely bounded method – weight quotients resemble normal distribution but not as much as linear method, calibration weights not as different from initial weights, still greater than 1, feasible solution
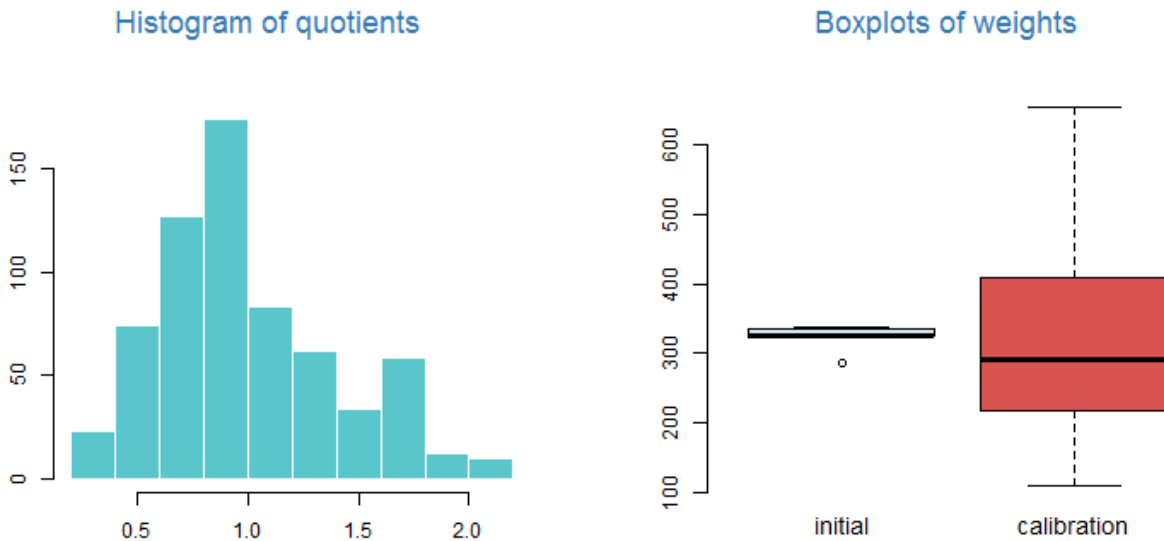
6 EXAMPLE – EU-SILC

In this chapter, we will run the calibration of the SO SR's synthetic (i.e. fully anonymized) EU-SILC 2012 cross-sectional file that has been adjusted to serve only as an example for two-stage calibration. For the sake of simplicity, we will omit economic variables from the list of auxiliary calibration variables. The data file and the table of totals can be found on GitHub repository.

Consider a file with the same structure as the file in chapter 3.2. We prepared the data for two-stage calibration, with 5023 households after summation of auxiliary variables at the individual level. Total number of individuals is equal to 13627. Auxiliary calibration variables are:

– sex combined with 6 age groups – individual level, numerical after summation

– households by members - household level, categorical, 5 categories

In this example we do not want to calibrate the s2a6 variable but rather see its estimates with the weights inferred from calibration of the other variables. In ideal case, they should resemble the population totals for s2a6 under the condition that the auxiliary variables are correlated. This will demostrate the *indicators monitoring* functionality.

The data and the table of totals are loaded first.

Then, auxiliary numerical and categorical variables are selected from the list. The same is done for weight and stratification column. In some browsers selected items may gray out, it is a common behaviour.

In order to monitor the s2a6 variable (normally it would be among calibration variables) we select it in the next list and tick some statistics to compute.


After proceeding to the next tab, prior to calibration, we try to look at the proportions of the H-T estimates of totals to known population totals by clicking on the *Show with initial weights* button. We can see considerable difference, therefore calibration is necessary. As a first step, linear method with calib solver is chosen. All totals obtained are equal to 100%, however, negative weights appeared with bounds obtained at -1.797 and 4.728, which gives us a clue that we can try to use some bounds around 0.3 and 3. After some trials we found bounds 0.12 and 3 with the nleqslv + logit option to be appropriate, taking into account the totals obtained, the average difference and the distribution of quotients.
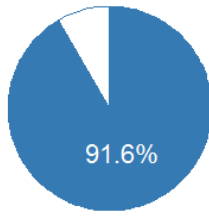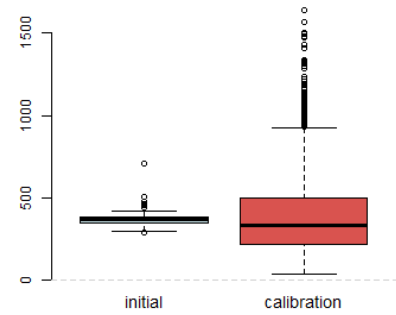
**Average difference feasibility**   **Histogram of quotients**   **Boxplots of weights**
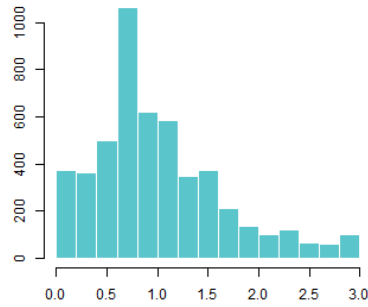
However, this solution is not sufficient for strata 7 and 8, where the totals obtained are not equal to 100%. Further fine tuning is necessary. After selecting strata 7 and 8 in the *Choose strata* list and running again the nleqslv + logit option with the bounds equal to 0.12 and 3, we can see the obstacle − lower bound is too strict and the weight quotients are pushed off to it. We need to relax it a little bit.

Histogram of quotients
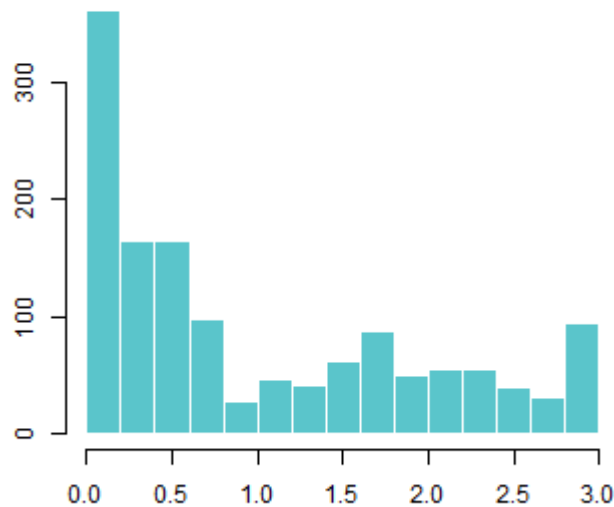
By trying some settings, we decided to use the calib + linear bounded solution with the bounds 0.1 and 3.5. Weight quotients are still leant to the left and totals obtained are not 100%, however, we cannot set lower bound to be even less than 0.1, distortion would be too high. Finally, the totals obtained are not very far away from 100% and could be deemed satisfactory. This approximate solution for strata 7 and 8 is not perfect but feasible. Calif now remembers the former calibration of strata 1 − 6 and the latter calibration of strata 7 − 8. Solution can be saved by clicking on the *Save* button.

For questions, comments and bug fixes visit https://github.com/SO-SR/Calif or contact the SO SR.

Boris Frankovic
Statistical Surveys and Methodology Dep.
tel: +421 2 50236 304
e-mail: boris.frankovic@statistics.sk

Statistical Office of the Slovak republic
info@statistics.sk
www.statistics.sk

References

[1] DEVILLE, J.-C., SARNDAL, C.-E. (1992). *Calibration estimators in survey sampling.* Journal of the American Statistical Association, 87, 376-382

[2] SARNDAL, C.-E. (2007). *The calibration approach in survey theory and practice.* Statistics Canada, Business Survey Methods Division. Catalogue no. 12-001-X, Vol. 33, No. 2, pp. 99-119

[3] HARMS, T., DUCHENSE, P. (2006). *On calibration estimation for quantiles.* Survey Methodology, 32, 37-52

[4] FRANKOVIC, B. (2013). *Calibration of weights of statistical surveys in R language.* Bratislava: Forum Statisticum Slovacum 5/2013, p. 19-37

[5] SAUTORY, O. (1993). *La macro CALMAR.* Paris: INSEE

[6] GLASER-OPITZOVA, H. et al. (2014). *The Calibration of Weights Using Calmar2 and Calif in the Practice of the Statistical Office of the Slovak Republic.* Vienna: European conference on quality in official statistics Q2014, paper.

[7] KIM, J.-K. (2013). *Chapter 2: Horvitz – Thompson estimation.* Iowa State University. Spring, 2013.

[8] SAUTORY, O. (2003). *A new version of the Calmar calibration adjustment program.* Statistics Canada International Symposium Series – Proceedings.

[9] VLACUHA, R., FRANKOVIC, B. 2015. *The Calibration of Weights by Calif tool in the Practice of the Statistical Office of the Slovak republic.* Bucharest: Romanian Statistical Review 2/2015, The International Conference New Challenges for Statistical Software - The Use of R in Official Statistics, paper

[10] R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/

[11] Winston Chang, Joe Cheng, JJ Allaire, Yihui Xie and Jonathan McPherson (2017). shiny: Web Application Framework for R. R package version 1.0.5.

https://CRAN.R-project.org/package=shiny

[12] Berend Hasselman (2014). nleqslv: Solve systems of non linear equations. R package version 2.1.1. http://CRAN.R-project.org/package=nleqslv

[13] Yves Tillé and Alina Matei (2013). sampling: Survey Sampling. R package version 2.6. http://CRAN.R-project.org/package=sampling

[14] Emilio Lopez Escobar and Ernesto Barrios Zamudio (2012). samplingVarEst: SamplingVariance Estimation. R package version 0.9-9

[15] Hadley Wickham and Evan Miller (2015). haven: Import SPSS, Stata and SAS Files. R package version 0.2.0.
http://CRAN.R-project.org/package=haven