The Calibration of Weights by Calif Tool in the Practice of the Statistical Office of the Slovak Republic

Róbert VLAČUHA
Boris FRANKOVIČ
Statistical Office of the Slovak Republic

ABSTRACT

The main scope of statistical surveys is to determine sample estimates. If some auxiliary population totals are available, an inferential step could enlarge precision. Calibration estimators are mostly used by statistical agencies. Since 2005, calibration at Statistical Office of the Slovak Republic has gradually moved from intuitive methods to the use of sophisticated tools, mainly SAS macro CALMAR2. The commerce licence, lack of user-friendliness and need for more precise estimates were the important motivations to create own tool Calif, written in the R software. It offers easy-to-use graphical user interface, enhances estimate quality and at last, but not at least, is freely available. It is well suited for each statistical survey and has replaced CALMAR2 in the process of calibration of weights in the Statistical Office of the Slovak Republic. In this paper we present Calif and its characteristics, compared with previous procedure on the example of the Household Budget Survey microdata.

Key words: *calibration, weights, survey, HBS, R*

INTRODUCTION

The main scope of statistical offices is to undertake statistical surveys and estimate some population characteristics. In most cases, parameters derived from the surveys are just the estimates of real values. Sampling weights that comply with the sampling design play a crucial role, enabling outcomes of the whole population without a 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 the additional improved accuracy could be achieved (via lower variance and reduced nonresponse bias). At last but not at least, it could be viewed as easily understandable for users, who appreciate no major differences between several surveys.

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 $k \in S$. 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. Information on values of vectors x_j known for every $k \in U$ is a very effective advantage but unfortunately is less common. Such a knowledge makes it possible to construct a derived auxiliary vector that correlates more with the study variable y, which certainly enhances the accuracy of estimates.

Two approaches are usually considered. The first one is generalized regression estimator (GREG) and is less common in practice. Its principle is a prediction of y values, $\hat{y}_k$, for all population elements via an assisting model and auxiliary information (either population totals or x_j known for every $k = 1,...,N$). A nearly design unbiased GREG estimator is then

$$\hat{Y}_{GREG} = \sum_{k \in U} \hat{y}_k + \sum_{k \in S} d_k (y_k - \hat{y}_k) \tag{2.1}$$

as stated in [2]. Often used in NSIs is *calibration approach*. Its main objective is, as was mentioned above, to reproduce the new weights for each k ∈ S that confirm auxiliary totals and differ minimally from design weights d_k. These weights

are independent of *y*, therefore in contrast to GREG estimator, 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. As [3] point out, "The calibration approach has gained popularity in real applications because the resulting estimates are easy to interpret and to motivate, relying, as they do, on design weights and natural calibration constraints." As [2] notes, "calibration approach seems to be transparent and natural, since the design weights are just slightly modified and unbiasedness only negligibly disturbed. Under this approach a unique weighting system is given, applicable to all study variables".

For almost each case

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

Let w_k denote the calibration weight of element $k \hat{\imath} S$. The calibration estimator of a study total is

$$\hat{Y}_{CAL} = \sum_{k \in S} w_k y_k \tag{2.2}$$

while calibration constraints are fulfilled

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

for all $j=1,...J$. According to [2]

$$\hat{Y}_{CAL} = \hat{Y}_{HT} + \sum_{k \in S} (w_k - d_k) y_k \quad \Rightarrow \quad E\left(\hat{Y}_{CAL}\right) = Y + \sum_{k \in S} E[(w_k - d_k) y_k]$$

so the objective of near design unbiasedness requires $E[(w_k - d_k)y_k] \approx 0$. Following calibration aims, $\frac{w_k}{d_k}$ should be near 1 for each $k \hat{\imath} S$ so the above mentioned mean value is near zero. The distance between design and calibration weights is expressed via distance function. Let $r_k = \frac{w_k}{d_k}$ denote the quotient of these weights. Then the distance function $G(r_k)$ is a nonnegative convex function of r_k with minimum in 1, that is G(1)=0, $\frac{\partial G(1)}{\partial w_k} = 0$, $\frac{\partial^2 G(1)}{\partial w_k^2} > 0$. 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 $d^T=(d_1,...d_n)$, $w = (w_1,..., W_n)^T$, $X=(X_1,...,X_J)^T$, $\mathbf{l}^T=(\mathbf{l}_1,...,\mathbf{l}_J)$ is a vector of Lagrange multipliers and x is a nxJ matrix of auxiliary variables. More precisely,

$$L = \sum_{k\epsilon S} d_k G(r_k) - \lambda^T \left(\sum_{k\epsilon S} r_k d_k x_k - X \right)$$

By taking partial derivatives of L we get

$$\frac{\partial L}{\partial r_k} = d_k \frac{\partial G}{\partial r_k} - \lambda' d_k x_k = \mathbf{0}$$

$$\frac{\partial G}{\partial r_k} = \lambda^T x_k \quad \text{whereas} \tag{2.3}$$

$$\sum_{k\epsilon S} w_k x_k = X$$

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

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

$$\sum_{k\epsilon S} d_k F(\lambda^T x_k) x_{kj} = X_j \tag{2.4}$$

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

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

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

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

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 [15].

DISTANCE FUNCTIONS

Several functions are commonly used for measuring the distance between design and calibration weights. We consider 4 of them here that are most frequently used.

– linear – analogical to linear GREG estimator. This function is often used due to its ability to find exact solution of (2.3) or (2.4) (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 f.i. 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.

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

– 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 design and calibration weights as well as the condition that weights are not less than 1 can be controlled. It gives

$$Ld_k \le w_k \le Ud_k$$

User must be aware of range allowed for calibration weights, tense bounds often lead to unsolvable system and increase average deviation applied to each design 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

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

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

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

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

These functions proposed by [5] are discussed in more detail in [4].

CALIF

Several software tools deal with calibration. They differ in options and environment where they run. Macro CALMAR2 has been used by Statistical Office of the Slovak Republic (SO SR) for calibration for past years but it has met some difficulties. SAS/IML is necessary to run CALMAR2 in SAS, at SO SR there are just two IML licences. Some cheaper and easy-to-use solution has been sought after. R [10] seems to be most eligible environment for statistical computing considering its licence, abilities and online help. SO SR has prepared a free R code called Calif that combines various calibration aspects and offers a user-friendly graphical user interface.

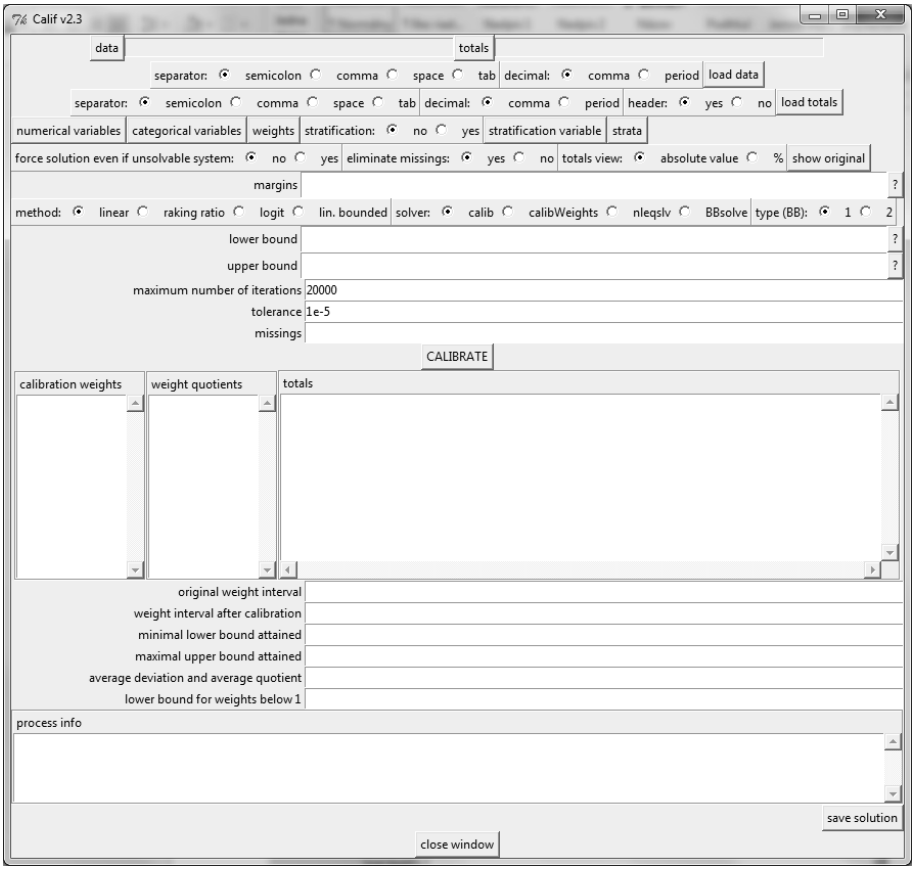
The main differences between CALMAR2 and Calif are price, user-friendliness and in particular ability to find approximate feasible solution. Up to now, by using macro CALMAR2, there has been no possibility to find other than exact solution. However, especially in smaller samples with many auxiliary variables, this solution was never found. The strategy was to calibrate the sample partially, which implied just the weights that did not preserve auxiliary totals. Calif is able to find also approximate solution by many ways, it is up to the user to choose the most sufficient one. It is able to enhance the whole process and to calibrate in very high detail. The various options of Calif requires from user some level of expertise, however, the easy-to-use graphical user interface makes it easier and more comfortable to work with it. On the other hand, in contrast to CALMAR2, Calif considers just the one-stage calibration, therefore data for multi-stage calibration have to be prepared outside before loading into Calif.

The package `fgui` [11] was used for creating the GUI. Calif covers all four distance functions mentioned in section 3. To solve calibration equations (2.4) or (2.3), various optimization functions are implemented. Function `BBsolve` from the package `BB` [12] and the function `nleqslv` from the package `nleqslv` [13] are able to solve nonlinear systems of equations. The former package uses Barzilai-Borwein spectral methods, while the latter relies on Broyden or Newton method. To make Calif more utilizable, function `calib` from the package `sampling` [14] and its faster version `calibWeights` from the package `laeken` [16] are incorporated, which calculates the Moore-Penrose generalized matrix inverses. The last two are the most powerful and fastest solvers, however, in some very tense scenarios they fail. `BBsolve` and `nleqslv` come with some solution in every situation, however, often a longer time is needed, especially for `BBsolve`.

In figure 4.1 the main Calif window is shown. Let's make a short tour through the particular items. Data for calibration as well as auxiliary totals can be loaded into Calif in .csv and .txt format. Columns of table of totals refer to separate auxiliary variables in data. In case of a categorical variable, a total for each category in alphabetical (or numeric) order has to be defined in the pertaining column. Important is just the order of auxiliary variables, it has to be the same for data and for totals. The names in headings are irrelevant.

The main Calif window

Figure 4.1



Rows of the table of totals refer to separate strata. The first column consists of their identification number. If no strata is present in the data, auxiliary population totals can be entered either by a table of totals (with values in either one row or one column) or directly in the „margins“ entry in the main window. Unlike the table of totals, data structure is completely free, the only necessity is to describe it via the graphical user interface. By pressing the „numerical variables“ button, the list of all variables indicated in the data appears. User is then free to mark auxiliary variables that have to be deemed as numerical. Analogical procedure is applied for categorical variables, design weights column and, if stratification takes part, also for marking variable indicating classification of strata. There is no possibility to construct strata from several variables, just one column can be considered (to be in line with the table of totals structure). Calibration can be processed in selected strata first or in the whole dataset at once. Several ways how to find a feasible solution are implemented in Calif.

Option „type(BB)“ reflects the equations (2.4) and (2.3) and concerns just the solvers BBsolve and nleqslv, it is irrelevant for calib and calibWeights. If logit or linear bounded method is performed, lower and upper bounds for $r_k = \frac{w_k}{d_k}$ must be entered. Specified missing values among auxiliary variables can be omitted from calibration. If a nonlinear system cannot be solved, user is able to force at least an approximate solution. Once the calibration is finished, original and new weight intervals, quotients r_k, totals attained and some other outputs show up. User needs to bear in mind that average quotient, $\frac{1}{n}\sum_{k\in S} r_k$, should be as close to 1 as possible. It is one's favour to choose between absolute value and percentage display of totals obtained. Average deviation of calibration weights is computed simply by $AD = \frac{1}{n}\sum_{k\in S}|w_k - d_k|$ which acts just as a first view on distortion applied to design weights. When using linear distance function, AD attained is minimal, any other bounded solution will have higher value of AD. If inconvenient weights result from the process, "lower bound for weights below 1“ will keep them over 1. When pressing „save solution“ button, input data with added column (calibration weights) is saved.

Prior to calibration, user can look at the totals computed by design weights and compare them with the auxiliary totals by pressing the „show original“ button. It is usefull especially if calibration cannot be well performed due to some non-sampling error in the sample.

SHORT SIMULATION STUDY

In this chapter we look closer at the simple calibration, without considering stratification aspect and compare the Horvitz-Thompson estimate $\bar{Y}_{HT}$ with the calibration estimate $\bar{Y}_{CAL}$. We used the exhaustive survey on labour costs as a population of size 7 068 enterprises. We selected 1 000 from it by simple random sampling without replacement and constructed basic design weights. The study variable is overall labour costs and we estimate its mean.

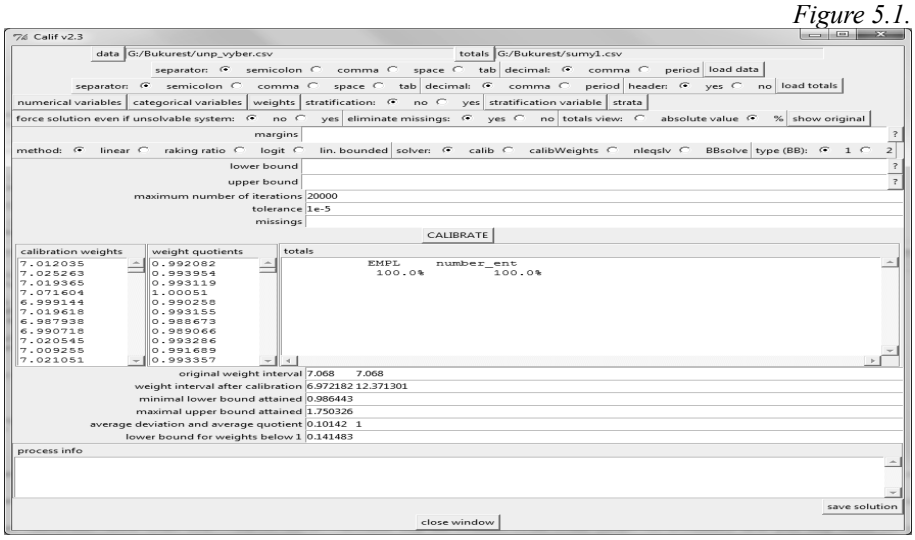
At first, calibration is carried out with auxiliary variables NUTS3 region by size of an enterprise (8 * 3 groups). Just the number of companies is taken into account for each group.

Next, number of employees is considered to be an auxiliary variable, due to its strong correlation with the overall labour costs and general availability from the administrative sources. In order to keep the number of enterprises unchanged, also this variable is added.

Linear distance function and calib function are used for both calibrations. Function calibev from the package sampling estimates the variance of calibration estimate. The real value of mean overall labour costs in our examples' population is 2.016.590€.

The result of the second calibration

Figure 5.1.



In the table 5.1 results are shown. It is clear that in this simulation study, calibration estimator was better than H-T estimator of the mean. Especially, significant difference was reached when strongly correlated variable was chosen as auxiliary for calibration.

The results of the simulation study

Table 5.1.

	H-T estimate	Calibration estimate 1	Calibration estimate 2
Estimate	1 822 897	1 942 081	2 039 776
Standard error	180 197	170 311	86 147

CASE STUDY – HOUSEHOLD BUDGET SURVEY

Household Budget Survey are part of the most requested data in Europe. It describes the expenditure structure of different types of households. This information is mainly used at EU level in the context of Consumer Protection Policy [8]. All household members are surveyed, but only those aged 16 and more are interviewed.
HBS is calibrated on two stages. The first stage concerns household level and the second the individual level. Condition that members of the household should have equal weight than household itself (integrated weights) imply the need of simultaneous calibration. All the auxiliary variables at the individual and household level are categorical. First we construct dummy variables at the individual level. Summation of all these variables results in the database of households where all previous individual dummy variables are numerical. Each households' sampling (or calibration) weight is assigned to its particular members. To be specific, as [9] presents, if S_M is a sample

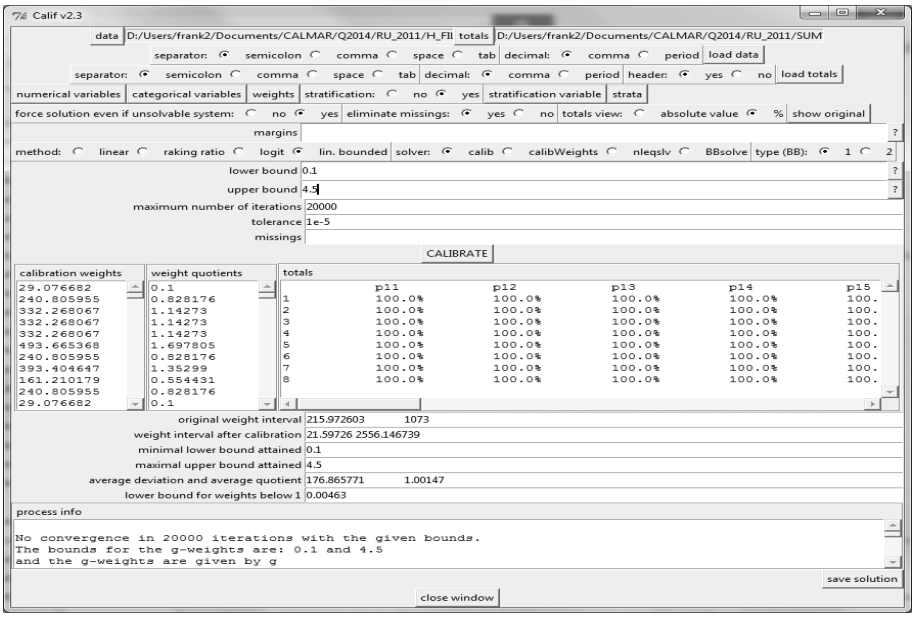
of households, S_I a sample of individuals, $d_{mi}=d_m$ are design weights, $X = \sum x_m$ are auxiliary population totals at the household level and $Z = \sum z_i$ auxiliary totals in individual population, combination of values of the individual dummy variables for each household m , i.e. $z_m = \sum z_{mi}$ 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.

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

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

Calibration of the HBS 2011

Figure 6.1.



Auxiliary totals attained by CALMAR2 and Calif

Table 6.1

	CALMAR2				Calif			
Individuals by sex and age groups	Totals matched	0 / 48	Min	93,0%	Totals matched	96 / 96	Min	100,0%
			Max	117,4%			Max	100,0%
			Average	100,6%			Average	100,0%
Individuals by sex	Totals matched	16 / 16	Min	100,0%	Totals matched	16 / 16	Min	100,0%
			Max	100,0%			Max	100,0%
			Average	100,0%			Average	100,0%
Households by size	Totals matched	0 / 40	Min	78,3%	Totals matched	25 / 40	Min	97,9%
			Max	121,9%			Max	105,7%
			Average	99,6%			Average	100,2%
Households overall	Totals matched	8 / 8	Min	100,0%	Totals matched	5 / 8	Min	99,4%
			Max	100,0%			Max	101,5%
			Average	100,0%			Average	100,2%
6 economic activity classes	Totals matched	0 / 6	Min	97,8%	Totals matched	48 / 48	Min	100,0%
			Max	101,7%			Max	100,0%
			Average	98,6%			Average	100,0%
Overall	Totals matched	24 / 118	Min	78,3%	Totals matched	190 / 208	Min	97,9%
			Max	121,9%			Max	105,7%
			Average	100,1%			Average	100,0%

HBS 2011 comprises 12 633 individuals within 4 705 households. 19 auxiliary variables within 8 regional strata are considered for calibration. 18 of them are numerical at the individual level (summed dummy variables) and the rest one is categorical with 5 categories at the household level. To be more precise, information on sex by age groups (12 variables) and economic status (6 variables) are taken incto account at the individual level and size of the household by members at the household level. Up to now, calibration has been done by CALMAR2 but the results have been insufficient. An exact simultaneous solution has never been found due to high number of calibration constraints (8 * 23 = 184 totals to attain), so the calibration process ran by iterative manner. Some of the variables were calibrated first, resulting $w_m w_m$ weights were taken as design weights and after that the calibration of another variables was executed. This was done several times, separately for each stratum. The result was some kind of an approximate solution, closer to the exact one after each iteration. Therefore, CALMAR2 must have been run for over 100 times with inadequate result on the highest detail. On the lower detail (many constraints ommited), CALMAR2 found a feasible solution.

Calif is able to calibrate the whole table in a few mouse clicks. The bounds for quotients were set to 0,1 and 4,5 for each stratum in both cases but Calif can handle even more tense bounds. Bounded linear function with calib as a solver were used. In table 6.1 results of the calibration processes are compared. Lower level of detail was used for CALMAR2 in order to find more suitable solution (decrease in number of totals at the individual level).

CONCLUSIONS

In this paper we have introduced the free R code Calif designed for calibration of weights of statistical surveys and showed its funcionality on the example of the HBS

dataset. It was made by the Statistical Office of the Slovak Republic in reaction to growing demands for some freeware user-friendly tool that could enhance the ability to match the known population totals. Before using Calif, SAS macro CALMAR2 had been used, but it had met several complications, like time consuming calibration, necessary experience with SAS and impossibility of calibration in high detail. Often no solution had been found. Calif is able to circumvent all these difficulties and offers an easy-to-use powerful tool. Its advantages are reached mainly thanks to the several R packages, in particular fgui, sampling, laeken, BB and nleqslv. As it is shown in the examples, the level of detail of calibration process can be considerably increased by Calif, and, moreover, by using graphical user interface.

Calif can be found on the webpage of the SO SR and is free to use.
http://slovak.statistics.sk/wps/portal/ext/products/software.tools/

References

1. DEVILLE, J.-C., SÄRNDAL, C.-E. (1992). Calibration estimators in survey sampling. Journal of the American Statistical Association, 87, 376-382
2. SÄRNDAL, 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. FRANKOVIČ, 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-OPITZOVÁ, 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. EUROSTAT. (2010). Household Budget Surveys
9. SAUTORY, O. (2003). A new version of the Calmar calibration adjustment program. Statistics Canada International Symposium Series – Proceedings.
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. Thomas J. Hoffmann, Nan M. Laird (2009). fgui: A Method for Automatically Creating Graphical User Interfaces for Command-Line R Packages. Journal of Statistical Software 30(2), 1-14. URL http://www.jstatsoft.org/v30/i02/
12. Ravi Varadhan, Paul Gilbert (2009). BB: An R Package for Solving a Large System of Nonlinear Equations and for Optimizing a High-Dimensional Nonlinear Objective Function. Journal of Statistical Software, 32(4), 1-26. URL http://www.jstatsoft.org/v32/i04/
13. Berend Hasselman (2014). nleqslv: Solve systems of non linear equations. R package version 2.1.1. http://CRAN.R-project.org/package=nleqslv
14. Yves Tillé and Alina Matei (2013). sampling: Survey Sampling. R package version 2.6. http://CRAN.R-project.org/package=sampling
15. Emilio Lopez Escobar and Ernesto Barrios Zamudio (2012). samplingVarEst: SamplingVariance Estimation. R package version 0.9-9
16. Andreas Alfons, Matthias Templ (2013). Estimation of Social Exclusion Indicators fromComplex Surveys: The R Package laeken. Journal of Statistical Software, 54(15), 1-25. URL http://www.jstatsoft.org/v54/i15/