Weighting in the Swiss Household Panel
Technical report

Erika Antal *[1] and Martina Rothenbühler[2]

[1]*Swiss Centre of Expertise in the Social Sciences C/O Université de Lausanne - Bâtiment Géopolis - CH-1015 Lausanne - Switzerland*
[2]*Clinical Trials Unit CTU - University of Bern - Finkenhubelweg 11 - CH-3012 Bern - Switzerland*

September 24, 2015

*email: erika.antal@fors.unil.ch

Contents

1 Introduction

The Swiss Household Panel (SHP) is a survey of households and persons permanently residing in Switzerland. It is a panel survey, more precisely a simple panel (at least at the beginning), which means that no new units are involved nor others abandoned, so the same persons and households are surveyed every year and answer the same questions. This is because the main objective of the SHP is to observe social change, especially the dynamics of changing life conditions of the population in Switzerland.

Compared to cross-sectional surveys, longitudinal household panels face some additional methodological challenges. One of them is the complex weighting scheme. On one side, the main objective of longitudinal surveys is to analyze the evolution of a population over time, which is done using longitudinal weights. On the other side, longitudinal surveys can also be used for cross-sectional analysis, referring to the permanent residents in private households in Switzerland in any given year. For this purpose, there is a need for cross-sectional weights. Furthermore, in a household panel survey there are not only individuals to weight for every wave, but also households. In this report, the current weighting scheme and the construction of each of the weights are described. This discussion is designed to give an idea of how the weights are produced and what techniques are used.[1]

Within the SHP there are three different types of weights:

- cross sectional household weights: WH$$T1P
- cross-sectional individuals weights: WP$$T1P
- longitudinal individuals weights: WP$$L1P (combined panel of SHP_I and SHP_II) and WP$$LP1P (SHP_I)

Additionally, there are transitional factors that enable to create custom made individual longitudinal weights over several waves. [2]

There are three samples in parallel, SHP_I, _II and _III, carried out for the first time in 1999, in 2004 and 2013 respectively. The survey is conducted annually from September to February by the institute M.I.S - Trend in Lausanne and in Berne using the technique of Computer Assisted Telephone Interviewing (CATI), that of Computer Assisted Personal Interviewing (CAPI) or that of Computer Assisted Web Interview (CAWI). The SHP is one of the main longitudinal databases of the population in Switzerland.

[1] This paper is based essentially on the technical report written by Eric Graf (2009). (see http://www.bfs.admin.ch/bfs/portal/de/index/themen/20/22/publ.html?publicationID=3328)

[2] Whereas the current longitudinal weights always refer to the first wave, the transitional factors are thus useful for the development of longitudinal samples that start after the first wave. It also allows for the longitudinal weighting of new household members (cohabitants). You can find more on the construction of these transitional factors, the limits of their application as also on the whole weighting system in the documentation available on the homepage of the Swiss Household Panel.

1.1 Sampling Design

Each of the SHP-s samples is a stratified random sample of private households whose members represent the non-institutional resident population in Switzerland. The first two samples were selected on the basis of the Swiss telephone directory (Stichprobenregister für Haushalterhebungen SRH), whereas the third sample was selected on the basis of the official registers of persons (Stichprobenrahmen für Personen und Haushaltserhebungen SRPH). The stratification is realized according to the seven major statistical regions of Switzerland (NUTS_II), proportionally to the number of households per strata (Graf, 2009). The sampling procedure thus takes into account only the number of household but not the the number of household members, so the size of the household.

Errors concerning the desired survey frame may however occur due to the slight but possible difference between the household population having one and only one telephone number and the true population of household. If not each household has a telephone number, under coverage is present. Contrariwise, it is possible that the telephone number's owner is not a household. Moreover, if a household has more than one telephone number the inclusion probabilities would not be well-defined. Anyhow, in the case of the SHP, this difference between the target and the attainable population is minimal.

1.2 Structure of the SHP

The SHP has three questionnaires: the grid, the household questionnaire and the individual interviews. The grid consists mainly of socio-demographic questions about the household members and helps to determine the eligibility status of the household members. Filling in the grid is imperative before moving to the two other questionnaires. The typical process is to complete the household questionnaire after the grid and to answer to the individual interviews later. However, sometimes the individual interviews are carried out before, or even without the household questionnaire. Generally, it is the - so-called - reference person who answers to the grid and to the household questionnaire while the individual questionnaires are completed by each eligible household member.

1.3 Non-response and Attrition

An individual is considered as respondent if he/she has answered to the individual questionnaire. At the household level, respondents are those having responded to the household questionnaire. Due to the structure of the SHP, the grid of a respondent (both individual and household) is filled out as a rule. The different levels result in non-response rates at different levels. Besides the two evident main levels - individual and household - sub-levels can also be distinguished according to some indicators. At the individual level the most important indicator is the proportion of the number of realized individual interviews and the total number of eligible individuals. At the household level the number of realized grids, the number of realized household questionnaires (with filled grid) and the number of complete questionnaires (realized household questionnaire with filled grid and at least one realized individual interview) are usually calculated. As the total number of activated addresses is also known, several comparisons between these numbers can be computed resulting in response rates at different levels. These rates are essential for the computation of the weights and will be discussed later.

From a longitudinal point of view attrition is a type of non-response. It means the loss of panel members from one wave to the next. Due to its nature as simple panel, SHP is also affected by attrition. The potential non-random patterns of attrition influence mainly the estimation results but it has also an effect on the weighting system. This is the reason why attrition is discussed briefly in this report.

The most important thing that leads to attrition is the longitudinal aspect of the survey. As the main purpose of a panel survey is the longitudinal analysis, only the initially selected households and individuals are followed year by year. As has been already mentioned above, basically there are no new units included nor existing abandoned. However, not all households and persons respond each year. According to some given rules of the current weighting system these units might not be followed anymore and attrition appears. If, in order to contend with attrition, the follow-up rules change, the longitudinal aspect of the survey becomes less evident and the weighting system would also be severely affected.

Several measures can be taken to stabilize attrition rates and try to increase participation. Most of them concern the fieldwork, in particular the communication with the households, but there are some possible proceedings that refer to the weighting system. The most evident is resulting from the fact that there is always a part of attrition which is natural and can not be influenced. To compensate this loss in participants the survey is "refreshed" by a new sample (SHP_II, SHP_III, ...). Every new sample needs to be combined with the preceding ones. As the longitudinal target population of this additional panel is different from the previous samples, the different weights also need to be combined. As this combination may be done in many different ways, its rules have to be determined. This is how new samples - needed mainly because of the natural attrition - affect the weighting system.

1.4 Data Preparation

The data used for the weighting process is derived from five different files: the grid (G), the household (H), the individual (P), the master household (MH), and the master person (MP) files. The grid contains basic information about the household members that are used for the construction of the weights. The household file contains information at the household level and in the individual file there are variables coming from the individual questionnaire. The master person file contains variables at the individual level that do not change from one year to another, like the sex or the date of birth. Finally the household master file contains the basic information about the household that do not change from wave to wave.

The first step is to create a plausibilised person file that contains all the personal and household variables that are used for the weighting. Missing values for an individual are imputed using the previous year's values inductively. If a value is missing from the current and also from the previous year, the value from two years earlier is used, etc. If no information is available concerning this variable, the value is imputed using the information of previous years.

The weighting scheme of the SHP differentiates three types of individuals. The first group represents the original sample members (OSM). These individuals were selected and could already have been interviewed at the first wave of one of the panels (in 1999, in 2004 or in 2013). An original sample member has both weights, cross-sectional and also longitudinal, so its personal identifier (variable IDPERS) has to be listed in the data set containing the initial weights. Vice versa, if the IDPERS of an individual is listed in this file, we know that this person is an OSM.

Children who were not yet born at the time of the sample selection, but whose parents are OSM, represent the second category, the children of OSM. As each child whose parent has participated in one of the waves receives an identification of the mother and or the father, this means that, in order to be treated as a child of an OSM, the identifier of the parent has to be in the file containing the initial weights. Once the child is 14 years old, it becomes an OSM.

The third group contains the non-OSM, that is, cohabitants who joined the household after the first wave and that are not children of any OSM.

The variable "OSM" informs about the different types of individuals within the SHP.

Table 1: OSM values and their meaning

OSM	Meaning	Condition
1	OSM	IDPERS in file with initial weights
2	Child of OSM	IDFATH/IDMOTH in file with initial weights
3	Non-OSM	Otherwise

This differentiation of types of individuals is essential for a correct weighting scheme. The whole procedure is based on this information.

2 Conceptual Aspects

In this section the philosophy and the terminologies related to the construction of weights are summarized. Regarding the conceptual aspects of the weighting procedure of the SHP, two classes of concepts can be distinguished. Those concerning cross-sectional weighting and those with reference to the longitudinal weighting. The two parts are presented separately in the following subsections.

2.1 Concepts Related to the Longitudinal Weighting

Because households represent a dynamic unit that could change constantly year by year, households do not have longitudinal weight nor transitional factors. Therefore, only individuals get a longitudinal weight and a transitional factor within the SHP. All individuals who were aged at least 14 and who lived in a selected household at the time the sample was drawn represent longitudinal persons, independently of the fact if they answer to the individual questionnaire or not (Graf, 2009). These people are panel members as long as the panel exists, even if they do not participate in any wave of the survey. Children of longitudinal persons become longitudinal members once they turn 14. For each panel, the longitudinal target population stays the same as long as the panel exists. The reference population is the resident population in Switzerland, that is all the people living in private households in Switzerland in 1999 for the SHP_I, 2004 for the SHP_II and 2013 for the SHP_III. All individuals of the original sample are eligible for longitudinal weighting of a specific wave if they were in the target population at time of the sample selection and if they answered to the individual questionnaire at the given wave. This rule implies that if a longitudinal individual did not answer to the individual questionnaire at wave 1, he/she gets a longitudinal weight zero at wave 1, but if he/she answered at a subsequent wave, the longitudinal weight at this wave will be non-zero.

2.2 Concepts Related to the Cross-sectional Weighting

The cross-sectional target population is the population of individuals living in private household in Switzerland in the given year. However, the population that can be observed is slightly different. By definition, it is composed of individuals living in private households containing at least one member who lived in a private household having at least one telephone number in Switzerland at the time of the sample selection.

As has been mentioned above, individuals join the household as a consequence of family dynamics. For cross-sectional purposes, data is usually collected at each survey wave for all household members if at least one of them is an OSM (Merkouris, 2001). Within the SHP, all members of the household (who are more than 14 years old) are interviewed, regardless of the fact if they are longitudinal persons (OSM) or whether they joined the household at a later stage. Cohabitants are people who joined a household containing at least one longitudinal person after the first wave of the panel. Cohabitants initially present and those who were initially absent must be distinguished. The first group refers to the individuals who were in the longitudinal target population at time of the sample selection. They have therefore an unknown but positive probability of inclusion. The cohabitants who are initially absent are those who were not in the target population at the time of the sample selection, that is in 1999, in 2004 and in 2013 respectively. This includes babies, immigrants or people who lived in an institutional household (prison, hospital, etc) instead of a private household at the time of sample selection. The reference cross-sectional population consists of all the private households in Switzerland in a given wave. The observable cross-sectional population is composed of all the households of the panel, which contains at least one longitudinal individual (OSM). This implies that households consisting only of immigrants who arrived after the first wave - they are not part of the longitudinal target population - are not taken into account in the cross-sectional weighting scheme.

Each longitudinal individual belonging to the reference population at a given wave is also part of the cross-sectional sample of this year and is therefore eligible for the cross-sectional individual weighting if he/she answers to the individual questionnaire.

In contrast to the longitudinal weighting, the cross-sectional weighting scheme not only applies to individuals, but also to households. A household is part of the cross-sectional universe of a given wave if he lives in one of the seven major statistical regions[3] at the time of the respective wave. The household is eligible for the cross-sectional household weighting if a member of the household answers to the household questionnaire.

3 Theoretical Approaches

This section is devoted to the main theoretical approaches of the weighting procedure of the SHP.

First the sampling weights are calculated. The basis of these weights are the inclusion probabilities, the probabilities of being selected for the sample. These probabilities are entirely determined by the sampling design. In order to compute the sampling weights from these probabilities, only their inverses need to be taken.

Households are selected for the sample, but not all answer, consequently not all units are finally part of the sample. The method of segmentation developed by Kass (1980) is used to compute a factor to adjust the sampling weights for this total non-response. These adjusted weights are called initial weights.

The Generalized Weight Share Method of Lavallée (2002) is widely used when constructing cross-sectional weights within longitudinal surveys. For the

[3]The seven major statistical regions are the following:

1. Lake Geneva (GE, VD and VS)
2. Middleland (BE, FR, JU, NE and SO)
3. North-West Switzerland (AG, BL and BS)
4. Zurich
5. East Switzerland (AI, AR, GL, GR, SH, SG and TG)
6. Central Switzerland (LU, NW, OW, SZ, UR and ZG)
7. Ticino

SHP it is used for both the cross-sectional individual and household weights in order to allocate a weight to cohabitants, of whom the inclusion probability is not known.

The third approach concerns the combination of the panels. It is done according to a factor allocating a relative importance to each of the samples. The method applied is the method proposed by Merkouris (2001).

Finally, the calibration method of Deville and Särndal (1992) is shortly presented. Calibration is applied to all the weights of the SHP, to adjust them so that certain population sums are correct (equal to the sums of the non-institutionalized Swiss population). The adjustments due to calibration are chosen to be as small as possible so that the introduction of bias for non-correlated variables is minimized.

3.1 Adjustment for Non-response

In order to compensate the effect of non-response, the weights are adjusted. In the SHP this adjustment is needed at several level: at the grid, at the household questionnaire and also at the individual questionnaire level. The method used for modeling non-response is the analysis by segmentation proposed by Kass (1980). Kass suggests to use the Chi-squared Automatic Interaction Detector (CHAID) procedure. CHAID represents a classification tree algorithm for selecting the set of auxiliary variables to model non-response (Kalton and Brick, 2000). The dependent variable is the response status, thus binary, and the independent variables are also treated as binary variables. The choice of the variables used to adjust non-response is rather limited (principally at the grid level and at the first wave) as information is needed not only for respondents but also for non-respondents. Information, especially about non-respondents, can come from official registers from where the sample was drawn, as well as from filled questionnaires of previous waves. Consequently in panel surveys, from wave 2 onwards, the number of available variables is larger than in cross-sectional surveys, as one can also rely on information collected at the previous waves. CHAID proceeds in consecutive steps and thus represents an iterative process. The algorithm first chooses the variable for the partition of the data that is most highly associated with the response status according to the highest χ^2. The data is then divided into two groups according to this chosen predictor. Each of these subgroups is then analyzed separately and independently of the other, to produce further subdivisions Kass (1980). It means that the separator variable should not be the same for each of the two subgroups and the predictors can be used several times to partition the data (Kalton and Brick, 2000).

The CHAID algorithm assumes that the dependent variable has $a \geq 2$ categories, and do so the predictors, predictor j has $b_j \geq 2$ categories.

The first step is to simultaneously determine the number of categories and the way to combine the categories of each predictor in order to find the best χ_j^2 statistics. It means that the objective of this step is to reduce the given $a \times b_j$ contingency table to a $a \times c_j$ table ($c_j \leq b_j \ \forall j$) which is the most significant (highest χ^2) independently of c_j.

If $T_{c_j}^{(i)}$ represents the χ^2 statistic for the $a \times c_j$ contingency table formed in the way i, $T_{c_j}^{(*)} = max_i \ T_{c_j}^{(i)}$ is the χ^2 of the best way to combine categories in term of independence, resulting in c_j categories. After having found the best combination with regard to the number of categories, it is sufficient to maximize these $T_{c_j}^{(*)}$ in c_j to find the best possible way to split the data set into two subgroups. Kass proposes a stepwise algorithm to implement the procedure:

Step 1. The categories of a predictor are cross-tabulated with the categories of the dependent variable.

Step 1a. Find the pair of categories (only amongst allowable pairs) of the predictor j whose $2 \times a$ sub-table is the least significantly different, i.e. has the smallest χ^2. If this significance does not reach a critical value, merge the two categories and consider them as a single compound one. Repeat this step as many times as possible/needed.

Step 1b. For each compound category consisting of three or more of the original categories, find the most significant and allowable binary split. If the significance is beyond a critical value, implement the split and return to *step 1a*.

Step 2. Calculate the significance of each merged predictor and isolate the most significant one. If this significance is greater than a critical value, subdivide the data according to the categories of the chosen predictor.

Step 3. For each partition of the data that has not been yet analyzed, return to the *step 1*.

The last step can be modified by putting restrictions. Within the SHP there is two additional restraints to the adjustment model for non-response, that is a minimum group size of 30 units and a minimum response rate within each subgroup of 30%. The partitioning process goes so on until no significant χ^2 can be found anymore. The result is a classification tree showing at each step the variable that is most predictive for the response status. CHAID thus partitions the data into mutually exclusive, exhaustive subsets that best describe the dependent variable (Kass, 1980). These subsets represent homogenous response groups (HRG). The HRG contain units that have similar characteristics that explain the non-response to a specific wave. Adjustment for non-response is based on these HRG. If the non-response among one HRG is completely at random, then the bias that occurs because of non-response can be ignored (Graf, 2009). Within the SHP we use the segmentation method presented above for all the adjustments for non-response. The segmentation is performed using an SAS macro developed for the SHP by Jean-François Naud of Statistics Canada, as well as Eric Graf and Bryce Weaver, both former collaborators of the SHP. The SAS macro enables to perform various types of adjustments, that is:

- Adjustment for non-response to the grid questionnaire
- Adjustment for non-response to the household questionnaire for the cross-

sectional weighting procedure

- Adjustment for non-response to the individual questionnaire for the longitudinal weighting procedure
- Adjustment for non-response to the individual questionnaire for the cross-sectional weighting procedure
- Adjustment for non-response to the individual questionnaire for the individual transitional factors.

3.2 Generalized Weight Share Method

The problem with panel surveys is that the inclusion probabilities of new household members are not known, since the cohabitants have not been selected at the time the sample was drawn. The inverse selection probability weighting scheme can therefore not be applied (Kalton and Brick, 1995). An alternative strategy consists of using only the inclusion probabilities of OSM and allocating parts of these weights within a household to cohabitants (Naud, 2004). This method thus requires information only on the initial inclusion probabilities of the OSM, but not on the inclusion probabilities of new cohabitants. It allows to incorporate cohabitants into cross-sectional analysis and to account therefore for an important part of the population dynamics. More generally, the method is used to enable cross-sectional analysis of the longitudinal samples. In order to construct cross-sectional weights both for individuals and households, the SHP actually uses the generalized weight share method (GWSM) of Lavallée (2002).

The method provides estimation weight for each unit surveyed in the cross-sectional target population U^B, by assigning the average of the sampling weights of units of the population U^A. U^A is the population from which the sample was selected (Lavallée, 2002).

A sample s^A of size m^A is selected from a population U^A that contains M^A units. π_i^A is the inclusion probability of unit $i \in U^A$ and assumed to be positive. The target population U^B contains M^B units and is divided into N clusters (for example households), each of them containing M_k^B units. Within the GWSM, links are assumed between units i of U^A and units j of cluster k of U^B (Deville and Lavallée (2006), Lavallée (2002)). This relationship is identified by an indicator variable $I_{i,jk}$, where $I_{i,jk} = 1$ if there exists a link and 0 otherwise. Ω^B represents the set of units $j \in U^B$ that are linked with at least one unit $i \in U^A$. For any unit j there can be zero, one or several links. The sum of these links is defined by $L_i^B = \sum_{k=1}^{N} \sum_{j=1}^{M_k^B} I_{i,jk}$ which can take the values 0, 1 or > 1. In order to apply the GWSM, each cluster k of U^B must have at least one link with the U^A, that is:

$$L_k^B = \sum_{j=1}^{M_k^B} I_{i,jk} > 0 \;\; \forall \;\; k \ldots N \tag{1}$$

The purpose of the GWSM is to attribute an estimation weight w_{jk} to each unit j of a surveyed cluster k. As the sample was not drawn from U^B but from U^A, only the sampling weight of unit i is known, that of unit j in the cross-sectional target population is assumed to be unknown. In order to compute it, the sampling weight of unit i of s^A, that are linked with unit j, are used. More precisely, it is not directly the sampling weights, but the sampling weights corrected for the total non response of unit i, so the initial weights are used.

The method contains four steps to assign a final weight to all units j within the cluster k (Lavallée, 2002):

Step 1. For each unit j of cluster k of U^B, the weight W'_{jk} is:

$$W'_{jk} = \sum_{j=1}^{M^A} I_{i,jk} \tag{2}$$

and in the same way, the initial weight of unit j of cluster k in Ω^B can be calculated as follows:

$$w'_{jk} = \sum_{j=1}^{M^A} I_{i,jk} \times \frac{t_i}{\pi_i^A} = \sum_{j=1}^{m^A} \frac{I_{i,jk}}{\pi_i^A} \tag{3}$$

where $t_i = 1$ if $i \in s^A$ and 0 otherwise. If unit j of cluster k has no link to any unit i of s^A, its initial weight w'_{jk} is equal to zero.

Step 2. L^B_{jk} represents the number of links between the units of U^A and the unit j of cluster k of U^B:

$$L^B_{jk} = \sum_{i=1}^{M^A} I_{i,jk}. \tag{4}$$

The quantity $L^B_k = \sum_{j=1}^{M^B_k} L^B_{jk}$ corresponds to the number of links present in cluster k.

Step 3. The final weight of each cluster k, W_k is obtained by calculating the ratio of the sum of the initial weights for the cluster over the total number of links for that cluster:

$$W_k = \frac{\sum_{j=1}^{M^B_k} W'_{jk}}{\sum_{j=1}^{M^B_k} L^B_{jk}}. \tag{5}$$

and

$$w_k = \frac{\sum_{j=1}^{M^B_k} w'_{jk}}{\sum_{j=1}^{M^B_k} L^B_{jk}}. \tag{6}$$

Step 4. Finally, $w_{jk} = w_k$ is assigned for all units $j \in U_i^B$, that is within the cluster k.

The GWSM offers a relatively simple solution to attribute a weight to the unit in the set Ω^B by using the probabilities of selected units in sample s^A.

In the context of longitudinal weighting, the sampling frame U^A can be associated to the initial population at wave 1, while the target population U^B represents the population a few years later (Lavallée, 2002). The linkage between population U^A and population U^B is established by the same individuals in these two populations, that is by longitudinal individuals, and forms therefore a one-to-one relation. $I_{i,jk} = 1$ if individual i of U^A corresponds to individual j of household k in U^B (longitudinal individual or OSM), and $I_{i,jk} = 0$ otherwise (cohabitant). As will be shown later in Chapter 4.2, within the SHP an adapted version of equation 6 is used, as a differentiation is needed between the households (of the longitudinal person) with at least one initially present and those with only initially absent cohabitants.

3.3 Combination of Multiple Panels

When using multiple panels, the way the panels are combined has to be considered to enable valuable cross-sectional estimations. The combination of the two first panels from 2004 on and also the inclusion of the third panel from 2013 on, are performed using the method of Merkouris (2001). As each panel has a different sampling frame due to population dynamics, each separate panel will, at time t refer to a slightly different population. At the same time, each panel covers, at the time of its selection, the entire survey population represented by the preceding panels (Merkouris, 2001). As a consequence, a common frame is formed each time a new panel is selected. The common frame of multiple panels is always the one of the most recent panel minus a non-overlap part. In the case of the SHP, the last common sampling frame is the one of the SHP_III.

In his contribution, Merkouris (2001) assumes that, in the case of a multiple-panel household survey, the inclusion probabilities (π_{1i}, π_{2i} … π_{Ki}) of all samples are small and hence that the probability of duplicated units is negligibly small. The combination of the panels requires however the development of weights for the sampling overlap. In order to avoid to define the inclusion probability π_{1i} of the units in the following panels, Merkouris (2001) suggests to apply an alternative strategy that requires information on the probability of inclusion in only one of the samples, thus avoiding the difficulties noted above. The strategy consists of associating with the i^{th} unit from the common frame a value p_{ki} ($0 \leq p_{ki} \leq 1$) when unit i is part of the k^{th} sample. Of course the sum of p_{ki}s equal to 1: $\sum_{k=1}^{K} p_{ik} = 1$. The weights for the combined samples can therefore be defined as follows:

$$w_i * = \sum_{k=1}^{K} p_{ki} \times w_{ki} \tag{7}$$

where w_{ki} denotes the weights associated with the k^{th} sample. The formula

7 defines the weighted combinations of the weights of several panels (Merkouris, 2001). Because of the limits of the factor of allocation p_{ki}, it is guaranteed that the weights will be non-negative. Furthermore, when setting p_{ki}s, some criterion of optimal weighting for the combined sample has also been considered (Latouche et al. (2000), Merkouris (2001), Graf (2009)).

In the SHP, the allocation factor is defined as to minimize the variance of a cross-sectional estimator for a variable of interest $\hat{Y}$. In order to estimate Y, we want to use a combined estimator, that allocates a relative importance to each of the panel according to the allocation factor p:

$$\hat{Y} = \sum_{k=1}^{K} p_k \times \hat{Y}_k \tag{8}$$

where $\hat{Y}_k, k = 1 \dots K$ denotes the Horwitz-Thompson estimators from the K panels (Latouche et al., 2000). The variance of the cross-sectional combined estimator $\hat{Y}$ is minimized if:

$$p_k = \frac{n_k \Big/ \text{deff}_k}{\sum\limits_{l=1}^{K} n_l \Big/ \text{deff}_l} \tag{9}$$

where n_k and n_l denotes the number of units (individuals or households in the case of the SHP) in the panel k and deff_k refers to the design effects of the panel k. Because the sampling procedure of the panels are similar in SHP, it can be assumed that the ratio of the design effects is 1 and the allocation factor is therefore

$$p_k = \frac{n_k}{\sum_{k=1}^{K} n_k}, \tag{10}$$

where p_k represents the allocation factor of the k^{th} panels and n_k is the number of units of whom we have a valid questionnaire in the k^{th} panel. Thus, $p_{ki} = p_k$ for each unit i in the k^{th} panel. As the sample size of the SHP_I is the largest, this method gives the most importance to the first panel.

3.4 Calibration

Within the construction of the cross-sectional weights, to fully control the population dynamics is almost impossible. Because of birth, immigration, entering and leaving an institution or death, the weights can be, in a final step, adjusted to known population totals. The method used for the marginal calibration within the SHP is the calibration by generalized regression. Technically it is implemented by using a SAS macro developed by Jean-François Naud and Caroline Cauchon of Statistics Canada. This calibration by generalized regression method has been proposed by Deville and Särndal (1992), and the idea behind their method is to derive new, calibrated weights w_i that modify as little as possible the sampling weights d_i (Deville and Särndal, 1992). The objective is therefore to minimize for any sample s the quantity

$$\sum_s \frac{(w_i - d_i)^2}{2d_i q_i} = \sum_s \frac{d_i \times (w_i/d_i - 1)^2}{q_i}. \tag{11}$$

The distance function G_i used by the method is defined as follow:

$$G_i(d_i, w_i) = \frac{(w_i - d_i)^2}{2d_i q_i} \tag{12}$$

where d_i represents the weight before calibration for the unit i, w_i the weight after the calibration and q_i an adjustment factor independent of d_i. In order to overcome the presence of some extreme values, the distribution of the cross-sectional weights are winsorised at the first and the 99^{th} percentile (winsorisation). Until now, this has only needed to be done for the household cross-sectional weights. This adjustments of the extreme weights are done within the calibration process.

4 Weight Construction in SHP

4.1 Weighting Procedure Steps

Currently, cross-sectional and longitudinal weights are assigned to individuals, whereas households are given only cross-sectional weights. This is due to the dynamic nature of households. Because the household composition changes over time, it does not make sense to have longitudinal weights for them. The longitudinal weights are used to adjust for the population the panel was sampled from, while the target population changes every year for cross-sectional estimations. Therefore, the longitudinal weights within the SHP are constructed with regards to the population in 1999, in 2004 and in 2013 (for the SHP_I, SHP_II and SHP_III respectively).

In practice, first the inclusion probabilities are determined for every unit of the reference population and then their inverse is taken as the sampling weights. The second step of the weighting procedure refers to the computation of the initial weights. Initial weights are set by adjusting the sampling weights for the total non-response. Total non-response is, when no survey data are collected for a specific unit, so in the case of the SHP it is adjustment for the non-response to the grid. These two first steps are the same for all the weights within the SHP and calculated only once, at the first year of each panel. Third step is, from second wave on, to adjust the initial weights to the total non-response at the actual year, so to the grid of the actual year. This step is also common for all the weights. For the non-response adjustment information, especially about non-respondent can come from official registers from where the sample was drown, as well as from filled questionnaires. In the SHP, non-response survey is conducted at the first year of the panel, then for the second wave on the information of the previous years is used. In order to compute the cross-sectional individual and household weights, weight sharing is then applied on the adjusted initial weight of the Original Sample Member (panel member) in the same household. Another adjustment for the non-response is carried out after, either on the individual or on the household level, depending on which weight is to develop. The next step is to combined panels. As there is more than one panel, their combination has to be done for the common frame. The final stage consists of the calibration which is applied in order to adjust the weight to known totals of the target population. In rare case, if negative or extreme weights occur, winsorisation is also needed.

4.1.1 Sampling Weights and Initial Weights

These weights attempts to correct the bias introduced by an imperfect sampling frame or an over representation of certain groups. The sampling weights are design weights. That is, they are entirely determined by the sampling design, as they are simply the inverses of the inclusion probabilities. Therefore they are not altered by subsequent changes of the composition of the panel. As it was already mentioned above, sampling weights are then adjusted to the total non-response at the first year resulting in the initial weights. Within the SHP, the initial weights are named POIDINIT. The part of the non-response adjustment of their calculation is slightly different regarding the SHP_I, the SHP_II and the SHP_III. The reason of this difference is that in 2004 and 2013 a non-response survey was performed among the households who did not want to participate. By that mean several information used to calculate the initial weights were collected.

Within the SHP_I, the initial weight corresponds to the household weight calculated at wave 1 in 1999. It is equal to the sampling weight that is adjusted for non-response to the grid and the household questionnaire in 1999. The initial weight is calculated as follows:

$$\text{POIDINIT}_i^{\text{SHP\_I}} = \frac{1}{\pi_i \times p_i^1 \times p_i^2 \times p_i^3 \cdot f}$$

where π_i corresponds to the sampling weight of household i, p_i^1 is the response probability to answer to the grid at wave 1, p_i^2 is the probability to answer to the household questionnaire at wave 1 and p_i^3 is the probability that at least one individual questionnaire within a household is filled out at wave 1. Finally, f is an adjustment factor that was applied at wave 1 so that the household weights multiplied by the size of the household corresponds to the population size in Switzerland in 1999 (Graf, 2009).

For the SHP_II and for the SHP_III, the initial weights were calculated using the sampling weight (π_i) multiplied by a factor f_i^{NRQ} that is defined by information of the non-response questionnaire and that is calculated using the segmentation method by Kass (1980) described above.

$$\text{POIDINIT}_i^{\text{SHP\_II, SHP\_III}} = \frac{1}{\pi_i \cdot f_i^{\text{NRQ}}}.$$

Note that the initial weights remain the same for the whole life of the panel.

4.1.2 Adjustment for Non-response to the Grid

From the second wave on the initial weights should be adjusted for the total non-response at the actual year. That is, to the grid of the actual year. The adjustment factor of each answering unit corresponds to the inverse of the response rate of the HRG the unit belongs to (Graf, 2009).

$$\text{f\_aj\_nrgril} = \frac{1}{\pi_{\text{HRG}}},$$

where π_{HRG} is the response rate for the given HRG.

The weights P_NRGRIL_i are thus defined as:

$$\text{P\_NRGRIL}_i = \frac{\text{POIDINIT}_i}{\pi_{\text{HRG}}} = \text{POIDINIT}_i \cdot \text{f\_aj\_nrgril}$$

where f_aj_nrgril represents the adjustment factor to the non-response for the grid of the actual year. This adjusted weight becomes the base weight of all the weights of a given wave.

After these two first steps, different further steps are needed for different weights. The subsequent procedure of the construction of the weights is presented separately for each of the final weights. First, the development of the cross-sectional individual weights and then the cross-sectional household weights are presented: Then the focus is put on the individual longitudinal weights. In a final section the individual transitional factor, which enables to construct alternative longitudinal weights, is introduced.

4.2 Cross-sectional Individual Weights

4.2.1 Generalized Weight Share

The current weighting scheme of the SHP assigns cross-sectional weights to non-OSM, as long as they are cohabitants of an OSM, but not after having left this household.[4] For the cross-sectional individual weights, a weight sharing is performed in households that have non-OSM. This is because cohabitants do not have any initial weight, as only longitudinal units have a non-zero weight P_NRGRIL. The weight sharing depends on whether or not the non-OSM was eligible when the sample was drawn, e.g., lived in an independent household in Switzerland at the time of the selection and was at least 14 years old. If the

[4] Since 2008, the rules of the follow-up changed and such individuals are now also included into the SHP. They do not however have any weights yet. The weight sharing method requires at least one OSM in the household. Part of this could be overcome by using an alternative definition of a cluster.

household contains one or more OSM and at least one initially present, thus eligible cohabitant, the shared weight is the same for all the individuals of the household and is equal to:

$$\text{PTI\_PAR}_i = \frac{\sum_{l=1}^{L} \text{P\_NRGRIL}_i}{L+P}$$

where L denotes the number of longitudinal household members and P represents the number of non-OSM initially present. If none of the cohabitants was initially present, then the weights of the longitudinal individuals stay the same, while the weights of the cohabitants initially absent equal the mean of the weights of the longitudinal sample members:

$$\text{PTI\_PAR}_i = \begin{cases} \text{P\_NRGRIL}_i & \text{for longitudinal person } i \\ \frac{\sum_{l=1}^{L} \text{P\_NRGRIL}_i}{L} & \text{for initially absent cohabitant } i \end{cases}$$

The weight share is performed using a SAS macro developed by Jean-François Naud of Statistics Canada and adapted by Bryce Weaver. Because OSM keep their weight if there are only initially absent cohabitants, the weights can be different among the individuals within a household.

4.2.2 Adjustment for Non-response

The weights are then adjusted for non-response to the individual questionnaire. The segmentation is done on the response to the individual questionnaire, conditional on having responded to the grid. The weights after the adjustment for the individual questionnaire are noted by PTI_NRQI. Both the adjustment for non-response to the grid and the weight sharing are performed separately for the three panels, as the samples have been selected at different moments (Naud, 2004).

4.2.3 Combination of Panels

There are three representative panels of the first wave population. The two first panels are combined from the 6^{th} wave onwards, the three panels are combined from the 15^{th} wave of the SHP_I onwards. The combinations of the panels are performed using the method of Merkouris (2001) presented above. The allocation factors are calculated separately for each of the seven regions. Using the allocation factors for each region r, the combined cross-sectional individual weights PTI_COMBPAN1_REG and PTI_COMBPAN2_REG are computed, which represent the population in Switzerland at any given wave.

$$\text{PTI\_COMBPAN1\_REG} = \begin{cases} \text{PTI\_NRQI} \cdot p_1 & \text{for the SHP\_I} \\ \text{PTI\_NRQI} \cdot p_2 & \text{for the SHP\_II} \end{cases}$$

and

$$\text{PTI\_COMBPAN2\_REG} = \begin{cases} \text{PTI\_NRQI} \cdot p_1 & \text{for the SHP\_I} \\ \text{PTI\_NRQI} \cdot p_2 & \text{for the SHP\_II} \\ \text{PTI\_NRQI} \cdot p_3 & \text{for the SHP\_III} \end{cases}$$

.

Note that the p_k of the first combination is not the same that for the second one, it always depends on how many panels are combined and in which year.

4.2.4 Calibration

In a final step, in order to fit the population totals, the weights are calibrated according to several socio-demographic variables, like age, civil status, nationality and the region of residence. The applied data sets are the ESPOP data of the year 1999 for the SHP_I, the ESPOP data of the year 2004 for the combined panel SHP_I and SHP_II and the ESPOP data of the year 2013 for the combined panel containing all three panels. The final cross-sectional individual weights are called:

- *WP$$T1P* for the cross-sectional individual weight adjusted to the size of the population in Switzerland in the year *$$*
- *WP$$T1S* for the cross-sectional individual weight keeping the sample size for the year *$$*.

4.3 Cross-sectional Household Weights

4.3.1 Generalized Weight Share

The weight sharing method is also applied within the household cross-sectional weighting procedure. The household cross-sectional weight after the weight sharing is called PTM_PAR. This weight is equal for all individuals within a household (Graf, 2009), independently of their age and cohabitant status. The shared weight is, similar to the individual cross-sectional weight, equal to the sum of all the longitudinal weights within a household divided by the number of longitudinal individuals and cohabitants initially present:

$$\text{PTM\_PAR}_i = \frac{\sum_{l=1}^{L} \text{P\_NRGRIL}_i}{L + P}$$

where L denotes the number of longitudinal household members and P represents the number of non-OSM initially present.

4.3.2 Adjustment for Non-response

This weight PTM_PAR_i is then adjusted for non-response to the household questionnaire using the method of segmentation of Kass (1980). The adjusted weight is called PTM_NRQM.

4.3.3 Combination of Panels

The next step of this weighting procedure consists of combining the panels using the same method as for the individual cross-sectional weights. The allocation factors are again calculated separately for each of the seven major statistical regions. We get the combined cross-sectional household weights as follows:

$$\text{PTM\_COMBPAN1\_REG} = \begin{cases} \text{PTM\_NRQM} \cdot p_1 & \text{for the SHP\_I} \\ \text{PTM\_NRQM} \cdot p_2 & \text{for the SHP\_ II} \end{cases}$$

and

$$\text{PTM\_COMBPAN2\_REG} = \begin{cases} \text{PTM\_NRQM} \cdot p_1 & \text{for the SHP\_I} \\ \text{PTM\_NRQM} \cdot p_2 & \text{for the SHP\_II} \\ \text{PTM\_NRQM} \cdot p_3 & \text{for the SHP\_III} \end{cases}$$

4.3.4 Calibration

After having combined the samples, the weights are calibrated using the same known population totals as for the cross-sectional individual weights. However this time under the restriction that all household members have the same weight. Thus the calibration is performed, in order to satisfy the marginal totals at individual level by keeping the final weight of each individual within a household equal to the one of the other household members.

4.3.5 Winsorisation

Because of some extreme weights, the final step consists of winsorising the weights at the first and the 99th percentile. The resulting final cross-sectional household weights are:

- *WH$$T1P* for the cross-sectional household weight adjusted to the size of the population in Switzerland in the year *$$*
- *WH$$T1S* for the cross-sectional household weight keeping the sample size for the year *$$*.

4.4 Longitudinal Individual Weights

The rules for the longitudinal weighting are similar to the cross-sectional ones. As cohabitants, independently of the fact of being initially present or absent, do not receive a longitudinal weight, there is no weight sharing in this weighting procedure. As the three samples within the SHP have been selected at different times, we have three different kinds of longitudinal individual weights. While the longitudinal individual weights of the SHP_I refer to the population in 1999, the year the first sample was selected, the first combined sample makes reference to the population in 2004 and respectively the second combined sample refers to the population in 2013.

4.4.1 Adjustment for Non-response

The changes about the eligibility mentioned above have to be taken into account when doing the non-response adjustment to the individual questionnaire of the SHP_I for the combined panel. From P_NRGRIL a basic longitudinal weight is computed:

$$\text{PL\_NRQI}_i = \text{P\_NRGRIL}_i \cdot \text{f\_aj\_nrlqi}$$

where f_aj_nrlqi represents the non-response adjustment factor to the individual questionnaire. This adjustment factor is calculated using the segmentation method of Kass (1980). In order to build the HRG the same rules are applied as for the cross-sectional weights.

4.4.2 Combination of Panels

The combination of the three samples is again done according to the method of Merkouris (2001), but is somehow more complex than for the cross-sectional weights. As the longitudinal weighting of the combined samples refers to 2004 or 2013, there is a need to define first the eligibility of the members of the first sample for the year 2004 and 2013, and second also the eligibility of the members of the first and the second sample for the year 2013. Concerning the first combination, the cohabitants of the SHP_I are still treated as cohabitants in the combined longitudinal sample, independently of their eligibility in 1999 or 2004. However, children of a household member of the SHP_I who are 14 and over in 2004 become longitudinal (combined) sample members and get therefore a longitudinal weight for the combined weighting of the panels. The allocation factor p_k is applied, but compared to the cross-sectional combination it is necessary to slightly adapt the rules. p_k are applied the following way:

- p_1 for the individuals of the SHP_I who were eligible in 2004
- p_2 for the individuals of the SHP_II who were eligible in 1999
- 1 for the individuals of the SHP_II who were not eligible in 1999.

The last group contains individuals who turned 14 years old, left an institution or immigrated between 1999 and 2004.

The second combination, that is when all three panels are combined, uses the same idea to determine the allocation factors.

4.4.3 Calibration

In order to produce the final longitudinal individual weights, the weights are calibrated to the same population totals as in the case of the cross-sectional weights. The final weights are noted as:

- *WP$$LP1* for the longitudinal individual weight of the year *$$* adjusted to the size of the population in Switzerland in 1999 concerning the SHP_I,

- *WP$$LP1S* fore the longitudinal individual weight of the year *$$* keeping the sample size concerning the SHP_I,
- *WP$$LP2P* for the longitudinal individual weight of the year *$$* adjusted to the size of the population in Switzerland in 2004, concerning the first combined panel (SHP_I and SHP_II),
- *WP$$LP2S* for the longitudinal individual weight of the year *$$* keeping the sample size concerning the first combined panel (SHP_I and SHP_II),
- *WP$$LP3P* for the longitudinal individual weight of the year *$$* adjusted to the size of the population in Switzerland in 2013 concerning the second combined panel (SHP_I, SHP_II and SHP_III), and finally
- *WP$$LP3S* for the longitudinal individual weight of the year *$$* keeping the sample size concerning the second combined panel (SHP_I, SHP_II and SHP_III).

4.5 Transitional Individual Factors

Whereas longitudinal weights refer to the population at the time of the first wave, transitional factors enable to make reference to an alternative starting point and to develop therefore a "custom made" longitudinal sample. Another advantage of individual transitional factors is that they allow to consider also the non-OSM for analysis. The transitional factors are intended to correct the effects of non-response from the previous year. They are not a weight but a factor of multiplication to apply to the cross-sectional weights of the previous year. As such, there is no transitional factor for 1999 nor for 2004 or 2013. The transitional factors are designed to construct a two-wave sample. In order to construct the longitudinal weight, the waves of interest at time t, $t+1$, ..., $t+k$ are taken. Starting with the cross-sectional weight at time t, the corresponding value is multiplied by the transitional factors of the subsequent waves $t+1$, ..., $t+k$ and thus the following longitudinal weight for the period t to $t+k$ can be computed:

$$\textit{longitudinal}_{t,t+k} = \textit{cross-sectional } w_t \cdot \textit{cross-sectional } f_{t+1} \cdot \ldots \cdot \textit{cross-sectional } f_{t+k},$$

where *cross-sectional* w_t denotes the cross-sectional weight at time t and *cross-sectional* f represent the subsequent transitional factors. The factors are determined in two steps (Weaver, 2009). The first step uses the segmentation process to model the (non-)response to the grid at wave t given response at wave $t-1$. In the second step, the response to the individual questionnaire in wave t is modeled given response to the grid at the same wave, again using the segmentation process. The transitional factors show, however, some weak points, especially if many waves are strung together. Because of this, to use of more then three consecutive waves is not recommend (Weaver, 2009).

References

Deville, J.-C. and Lavallée, P. (2006). *Indirect sampling: The foundations of the generalized weight share method*, volume 32(2).

Deville, J.-C. and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87(418):376–382.

Graf, E. (2009). Pondération du panel suisse de ménage psm i vague 9 psmii vague 4 psm i et psm ii combinés. description détaillée. Technical report, Office fédéral de statistique, Neuchâtel.

Kalton, G. and Brick, J. (1995). Weighting schemes for household panel surveys. *Survey Methodology*, 21(1):33–44.

Kalton, G. and Brick, J. (2000). Weighting in household panel surveys. *Researching social and economic change. The uses of household panel studies*, pages 96–112.

Kass, G. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*, 29:119–127.

Latouche, M., Dufour, J., and Merkouris, T. (2000). Cross-sectional weighting for the slid: Combining two or more panels. Technical report, Statistics Canada.

Lavallée, P. (2002). *Le sondage indirect ou la méthode généralisée du partage des poids*. Ellipses, Paris.

Merkouris, T. (2001). Estimation transversale dans le cas des enquêtes auprès des ménages à panels multiples. *Techniques d'enquètes*, 27:189–200.

Naud, J.-F. (2004). Pondération longitudinale avec panels combinés. Technical report, Statistics Canada.

Weaver, B. (2009). Swiss household panel user guide (1999-2008), wave 10. Technical report, Swiss foundation for research in social sciences.