Weighting system of the COHORT sample Technical Report

Erika Antal *

February 4, 2016

1 Introduction

The computation of the weights of the first wave concerning the sample of COHORT was based on the same principles that of the weighting system of the SHP samples performed by FORS. The concepts and the methods used for the construction of the weights of the SHP are presented in the related documents. The methods were adapted for the COHORT sample, only the calculation of the sample weights is different. As the sampling design of the two samples are different the computation of the sampling weights differs too. The purpose of this report is to present the details of this first step of the procedures of the construction of weights. Concerning the other methods, see the weighting documentation of the SHP samples (http://forscenter.ch/en/our-surveys/swiss-household-panel/documentationfaq-2/methods/weighting/).

2 Sampling procedure

Every weighting system is based on the sample weights. These are the weights which will be adjusted later for the non-response, shared between household members - if it is necessary - and finally calibrated to known totals of the reference population. So the whole procedure of the calculation of the weights begins by determining these sample weights. The sampling weights are the inverse of the inclusion probabilities which are entirely determined by the sampling design. A sampling design is a probability distribution on the set of the all possible subset of the population of reference. This distribution can be done by determining the probability of being selected for the sample for each unit in the population of reference.

Within the NCCR LIVES, a sample of individuals aged between 15 and 24 years was selected. It was designed to be composed by "secondos " at 2/3.

*Swiss Centre of Expertise in the Social Sciences - Bâtiment Géopolis - CH-1015 Lausanne - Switzerland emails: erika.antal@fors.unil.ch

In order to ensure this allocation in the sample, the sampling procedure was planned by clarifying the main basic concepts and notions such as:

- Population of reference:
  The population of reference consists of all individuals who satisfies the following criteria:
  - Residence in Switzerland (at 01.01.2013) (criterium C1)
  - Year of birth between 1988 and 1997 included (criterium C2)
  - schooled in Switzerland before the 10^{th} birthday (criterium C3)
- Secondos:
  Secondos are young people who have been schooled in Switzerland and whose parents arrived in Switzerland as adults. It would say that persons satisfying the forth criterium (besides the 3 others mentioned above) are considered as "secondos". This forth criterium is:
  - Both parents born abroad and arrived in Switzerland after their 18^{th} birthday (criterium C4)

The sampling procedure is composed by three separated steps.

1. Phase 1: Drawing a stratified simple random sample with unequal inclusion probabilities from the reference population.
2. Phase 2: First detection of secondos
3. Phase 3: Over-representation of secondos by network sampling

In each steps the inclusion probability of each individual can be calculated or efficiently estimated.

2.1 Stratified simple random sample with unequal inclusion probabilities

First, a simple random sample with unequal inclusion probabilities was selected - by the SFSO (OFS) (Swiss Federal Statistical Office) - from the population of reference, that is from the population of Switzerland who potentially satisfy the three first criteria mentioned above. This population was divided earlier in four strata according to two characteristics.

1. Target origin
   - foreigner from Bosnia-Herzegovina or from Croatia or from Italy or from Kosovo or from Macedonia or from Montenegro or from Portugal or from Serbia or from Turkey and holding a permit C or
   - born in Bosnia-Herzegovina or in Croatia or in Italy or in Kosovo or in Macedonia or in Montenegro or in Portugal or in Serbia or in Turkey and swiss or holding a permit C.

2. Target region

    - inhabitant of a municipality affiliated to one of the 30 regions MS: Baden, Basel-Stadt, Biel-Bienne, Glarner Unterland, Glattal-Furttal, Limmattal, Pfannenstiel, Schaffhausen, StGallen, Unteres Baselbiet, Untersee, Werdenberg, Zimmerberg, Zug, Zurich, Genève, Lausanne, Nyon, Vevey, Morges, La Vallé, Aigle, Pays d'Enhaut, Neuchâtel, la-Chaux-de-Fonds, Lugano, Mendrisio, Bellinzona, Locarno, Tre Valli.

The selection of the sample in this first phase was conducted from the Federal Register of Inhabitants with the help of variables present in the register to determine the stratum the individual belongs to.

- Municipality (characteristic 51)
- Nationality (characteristic 41)
- Permits (characteristic 42)
- Place of birth (characteristic 43)

2.1.1 Computation of the inclusion probabilities

The table 1 shows the estimated number of persons in the reference population by strata:

	Target origin	Other origin	
Target region	45'832	315'591	361'423
Other region	51'912	489'510	541'422
	97'744	805'101	902'845

The table 2 presents the structure of the selected sample (with the reserve) by strata:

	Target origin	Other origin	
Target region	1'445	857	2'298
Other region	927	775	1'702
	2'372	1'628	4'000

By knowing the number of persons in the population of reference and in the sample, the inclusion probabilities can be calculated for each strata. These inclusion probabilities of the first phase are presented in the table 3.

	Target origin	Other origin	
Target region	0.0315281899	0.0027028654	0.0063582008
Other region	0.0178571429	0.0015832159	0.0031435738
	0.0242674742	0.0020221065	0.0044304393

2.2 First detection of secondos

In order to have the most target persons possible in the final sample, this first sample of 4000 persons was divided randomly in 2 groups. The first group - "keep only secondos" - was composed of 2800 persons, the second - "keep everyone", was composed of 1200 persons. In the first group only the potential secondos while in the second, each person were kept for the sample.

2.2.1 Computation of the inclusion probabilities

As the division was done at random, the inclusion probabilities of the second phase can be easily calculated. A person considered as secondos was kept in the sample with a probability 1, and a person considered as not secondos with a probability 0.3 corresponding to the proportion of the second group ("keep everyone") in the whole sample. This first screening refines the inclusion probabilities after the first two phases, and can be calculated separately for the assumed secondos and for the assumed not secondos. In fact, it changes only the probabilities of the persons assumed not to be secondos. [1]

2.3 Network sampling

In order to calculate precisely the inclusion probabilities, every possibility to be selected in the sample shall be taken into account. At the third phase of the sampling design, persons in the network of the person selected (and answered) are also selected for the sample. So, a particular person can be selected for the sample:

1. if he/she is selected at the first phase **and**
2. at the screening he/she is selected (second phase) **or**
3. if he/she is in the network of a person who is selected (and answered) at the two first phases and he/she is chosen (third phase) .

In this section, the third phase of the sampling design is explained in details. The network design itself also composed by 2 almost identical steps. The persons selected (and answered) at the two first phases are asked to fill out a questionnaire concerning persons in their network. Each network members are likely to be chosen. The selection is done by applying a sampling procedure of a random sampling design with unequal inclusion probabilities and fixed size. Once member/s are selected from the network a second iteration of this design is applied. It would say, when the person is selected from the network of a person selected at the two first phases, members of his network will also be selected, but the chain stops after. These latest members will be asked to fill out the network questionnaire, but any additional persons will be selected for an extra iteration. As the two iterations of the network designs differ only in their parameters and

[1] Because of the too many different possible values of the inclusion probabilities, tables containing the inclusion probabilities will not be presented from now on.

not in their sampling design or procedure, the details concerning the computation of the inclusion probabilities explained here upon are valid for both of the iterations.

The criteria for being selected by the network sampling are:

1. He/she is not selected at the two first phases.
2. He/she has at least one person that was selected at the first two phases and he/she was mentioned in the network of this person.
3. He/she is selected from the network of this person.

Each event mentioned above has its own probability that can be calculated. As for being selected at the third phase, each of these events need to be occurred, thus in order to compute the final probability of the third phase the product of these probabilities has to be taken. The following subsections discuss in detail the computation of the probabilities of these events separately.

2.3.1 Being not selected at the two first phases (not selected at earlier stages)

The probability of being not selected at the two first phases is quite easy to calculate. It is simply the complementary of the event "being selected". It would say $1 - \pi_{p1p2}$, where π_{p1p2} is the probability of being selected at the two first phases.

2.3.2 Having at least one person selected at the first two phases and being mentioned in the network of this person

First of all, this event is only relevant if the "symmetry" of the links between persons supposed to be satisfied. The symmetry of the links means that if person "A" mentions person "B" in his network, he has to be mentioned in the network of person "B" too. If the links are not symmetric, even if the person mentions somebody in his network but this person does not mention him in his network, the person can not be selected. Considering that not every person mentioned in the network of somebody is selected, or even if he is selected, do not always respond, this condition is difficult to verify. However, it is quite realistic to assume that if person "A" knows person "B", person "B" also knows person "A". In addition the characteristics of persons to be cited in his "network" were defined based on objective criteria. The third reason to reliably suppose that this condition is satisfied is that the symmetry of the links was verified in the known networks of the sample members.

The probability of having at least one person selected at the first two phases in his network can be determined by calculating also the complementary probability. It would say taking the complementary of the probability of having no one

selected in the network the desired probability can be determined.

$$p_{alops} = 1 - p_{nos}$$

where p_{alops} is the probability of having at least one person selected at the first two phases in its network and p_{nos} is the probability of having no one selected in its network. The computation of the probability of being not selected at the two first phases has already described above. This method can be adopted for each person in the network and finally taking the product of these probabilities the probability of having no one selected in the network can be computed.

$$p_{nos} = \prod_{k=1}^{n_s} (1 - \pi^k_{p1p2})$$

where π^k_{p1p2} is the probability of being selected at the first two phases of persons k and n_s is the number of persons in the network.

2.3.3 Being selected from the network of a person

As the network of a person secondos and that of a person not secondos have not necessarily the same structure, the probabilities of being selected from their networks are not necessarily the same. This is the reason why the formula mentioned above has to be computed separately for the network-members secondos and for the network-members not secondos. The properties of the unknown networks concerning its length (the number of persons) and its composition (the number of secondos and that of not secondos) need to be estimated. Theses parameters were calculated using the properties of the networks of peoples who participated in the survey. Evidently it is done separately for the secondos and for the not secondos. Knowing the properties of the network of the persons in the network of somebody permits us to calculate the probability of being selected from the network of a friend secondos, and from that of not secondos. As at the phase of selection from the network, the secondos and the not secondos persons have different probabilities to be selected, the properties of the random sampling design with unequal inclusion probabilities and fixed size can be applied. In order to compute the inclusion probabilities several parameters need to be defined. They are

- The number of persons in the "population", from where the persons are selected.
  - It is the number of persons in the network of the person, the length of the network.
- The number of persons to be selected.
  - It was different for the two steps of the network design. It was arbitrary set at 4 for the first and 2 for the second phase.

- The ratio of the different probabilities.
  - It is arbitrary set to 4:1. It would say that the inclusion probability of a person secondos was set 4 times larger than the inclusion probability of a person not secondos.

As the length and the composition of the networks are known or can be estimated for the unknown networks and the two others parameters are set the inclusion probabilities can be computed.

Once the probability of having at least one person selected at the first two phases in its network and the inclusion probabilities of the network selections are computed, the probability of being selected via this network design can be determined. The sum of these probability is the whole probability of being selected for the sample.

3 Determination of characteristics

As it is mentioned above the inclusion probabilities depend on several characteristics. In addition, there is also a lot of sources from where these characteristics can be determined. This section describes how these parameters were set.
The characteristics to be determined:

- Residence in Switzerland (at 01.01.2013) (criterium C1)
- Year of birth between 1988 and 1997 included (criterium C2)
- schooled in Switzerland before the 10^{th} birthday (criterium C3)
- Both parents born abroad and arrived in Switzerland after their 18^{th} birthday (criterium C4)
- Secondos
- Target origin
- Target region
- Number of contacts
  - Number of secondos contacts
  - Number of not secondos contacts
- Age

The source of the information:

- "Cohorte2013_ContactFile_20140811_def.dta"
  - This is the contact-file. It contains information about the interviewer, the addresses and the calls. The most important variables kept:

    - identification variables,
    - identification variables of the interviewer,
    - canton, region, postal codes,
    - date, number, duration of the call,
    - variables about the verdict concerning the different questionnaires (if it is filled out or not during the interview),
    - in the case of refusal, variable about the reason of the it.

- "Cohorte2013_Grid_and_HH_20141125.dta"
  - This file contains information about the household and the answers of the household questionnaire. The most important variables kept:
    - identification variables,
    - age, birth date, sex,
    - canton, region, postal codes,
    - civil status, education, employment, occupation, language, permit (visa), nationality,
    - type of household, number of household members, number of adults, number of kids,
    - target origin, target region,
    - biographic, grid and household, network,
    - variables "scores" for defining the status concerning the criteria C1-C4 mentioned above, that is:
      - the year of the birth,
      - if the person were schooled in Switzerland before its 10^{th} birthday or not,
      - if the parents born abroad and arrived in Switzerland after their 18^{th} birthday or not.

- "Cohorte2013_Bio_20141125.dta"
  - This file is the biographic file and contains retrospective information about the person and his life's events. The most important variables kept:
    - identification variables,
    - age, birth date, sex,
    - region,
    - permit (visa), nationality,
    - residence,
    - target origin, target region,
    - network,
    - variables for defining the status concerning the criterium C1-C3 mentioned above, that is the age and the year of arriving in Switzerland.

- "Cohorte2013_Network_20141125.dta"
  - This is the network file, containing information about the network of the person. I.e. proxy information about the network members of the person. The most important variables kept:
    * identification variables,
    * age, birth date, sex,
    * region,
    * permit (visa), nationality,
    * residence,
    * target origin, target region,
    * bio,
    * variables for defining the status concerning the criterium C1-C3 mentioned above, that is the age and the year of arriving in Switzerland.
- "Cohorte2013_NonResponseSurvey_20140813.dta"
  - This file is the non response file. It contains some information about persons that was selected for the survey but did not participated in, but filled out the non-response or the refusal questionnaire. The most important variables kept:
    * identification variables,
    * age, birth date, sex,
    * region,
    * civil status, education, occupation, language, nationality,
    * number of household members,
    * type of questionnaire filled out (non-response or refusal),
    * variables identifying the reasons of non participation of the survey.
- "neededMISID.dta"
  - This file is the file received from the statistical office (OFS) containing minimal information about the persons selected at the first stage (seeds). The variables kept:
    * identification variables,
    * target origin, target region.
- "02_table_complete_ofs_poste_2014.dta"
  - This file contains information about the addresses. The variables kept:
    * numbers of OFS,
    * names of OFS,

        * postal codes.

- "Liste_de_communes_cibles.dta"
  - This file contains information of the target municipalities (OFS). The variables kept:
    * numbers of OFS,
    * names of OFS,
    * postal codes.

For certain characteristics there are several files from where they can be determined. The computation of these parameters will be explained per files right below.

- "Cohorte2013_ContactFile_20140811_def.dta"
  - This file together with the file "02_table_complete_ofs_poste_2014.dta" and "Liste_de_communes_cibles.dta" are used essentially to determined the "target postal code" wich will be of help to fix the variable "target region". As the file "Liste_de_communes_cibles.dta" contains onle the postal codes of the target region, merging it with the file "02_table_complete_ofs_poste_2014.dta" the target postal codes can be determined.
  -

- "Cohorte2013_Grid_and_HH_20141125.dta"
  - C1
    * The variables postal code, canton and the year of arriving to Switzerland were used.
  - C2
    * The variable birth year and the variable scr1a which also indicate the year of birth were used.
  - C3
    * The variable year of arriving in Switzerland, the variable age of arriving in Switzerland and the variable birth year were used
  - C4
    * The variables scr3, scr4, scr5, scr6, the variables indicating if the parents born abroad and arrived in Switzerland after their 18^{th} birthday or not.
  - Secondos
    * The variable secondos were set using the variables C1-C4. More precisely if the criteria C1-C3 are satisfied the person is eligible. A person eligible is considered secondos if the criterium C4 is satisfied. If not the person eligible is considered as not secondos.

- Target origin
  - First the categories (mentioned above) of a person of target origin were determined. Using the country codes of the birthplace, the nationality (first, second or third) and the visa of the person, the age, the age and the year of arriving to Switzerland these categories can be valued. The variable target origin was set at 1 if at least one of the categories is satisfied and 0 if none of them. The computation was needed only for the non-seeds.
- Target region
  - The computation was needed also only for the non-seeds. Merging the file with the file containing the target postal codes the variable target region was easily settable.
- Age
  - The file contains explicitly the information about the age. (variable age)
- "Cohorte2013_Bio_20141125.dta"
  - C1
    - The variable residence were used.
  - C2
    - The variable birth were used.
  - C3
    - The variables residence, birth date and year of arriving in Switzerland were used.
  - Target origin
    - First the categories (mentioned above) of a person of target origin were determined. Using the country codes of the nationality (first, second or third) and the visa of the person, the age, the age and the year of arriving to Switzerland and the variables about the residences of the person, these categories can be valued. The variable target origin was set at 1 if at least one of the categories is satisfied and 0 if none of them. The computation was needed only for the non-seeds.
  - Age
    - The variable birth date was used.
- "Cohorte2013_Network_20141125.dta"
  - C1
    - The variable coding the answers of the question about the region of Switzerland where the network member lives was used.
  - C2
    - The variable age was used.

    * C3
        · The variable concerning the country of birth of the person was applied.
    * C4
        · The question "If at least one of the parents of "member PX" grew up in Switzerland ?" with the answers "1: sure yes", "2: sure non" and "3: not sure" was used. The variable C4 was set at 0 if the answers for this question was 1, otherwise (response 2 or 3) it was set at 1.
    * Secondos
        · The variable secondos were set using the variables C1-C4. More precisely if the criteria C1-C3 are satisfied the person is eligible. A person eligible is considered secondos if the criterium C4 is satisfied. If not the person eligible is considered as not secondos.
    * Target origin
        · First the variables concerning the nationality of the member were used, then the created variable secondo was also used.
    * Target region
        · The variable concerning the residence of the member, coded by the MIS numbers was used.
    * Number of contacts, number of secondos contacts and of not secondos contacts
        · The file contains explicitly the information about the number of contacts.

- "Cohorte2013_NonResponseSurvey_20140813.dta"
    - the non-response file was only used for the adjustment for the non response. These characteristics determine the inclusion probabilities so the sampling weights and the method of adjustment for the non response is applied for calculating the initial weights. So this file is not used for computing the inclusion probabilities.

- "neededMISID.dta"
    - Target origin
        * The file contains explicitly the information about the target origin. (only for the seeds)
    - Target region
        * The file contains explicitly the information about the target region. (only for the seeds)

After determining these characteristics from each file separately the information had to be confronted. The hierarchies between the files concerning the determination of these parameters was established as follow:

- the merged file of "Cohorte2013_ContactFile_20140811_def.dta", "needed-MISID.dta", "02_table_complete_ofs_poste_2014.dta" and "Liste_de_communes_cibles.dta",
- "Cohorte2013_Grid_and_HH_20141125.dta",
- "Cohorte2013_Bio_20141125.dta",
- "Cohorte2013_Network_20141125.dta"

It is needed to precise that for computing the inclusion probability the status if the person is considered as secondo or not, the proxi information is not corrected. The reason of it is that at the moment of the selection from the the network of somebody else the self reported information is not yet known. Contrarily, when the structure of the network (number of secondos and not secondos) the proxy information is corrected if it is needed with the self reported information (if available).

3.1 Estimation of network characteristics

The structure of the network is essential for the computation of the inclusion probabilities. Anyhow it is not always known, thus estimation is needed. As it was mentioned above, characteristics are determined applying the different files filled out by the surveyed person. The length of the network, the number of secondos and not secondos in it can be determined for the persons who filled out the network file. For the others it has to be estimated. As the probabilities of being selected are not the same for a person considered secondo and for a person considered as not secondo, this parameters have to be estimated separately.

There was no difference regarding the average length of a network of a person considered secondos and that of a person considered non secondos. It was estimated 7. A slight difference concerning the structure was however found. The average number of the persons considered secondos in the network of a person considered secondos was 4 and that of the persons considered not secondos was 3. For a person considered not secondos it is the inverse. 4 persons considered not secondos and consequently 3 persons secondos was estimated for an unknown network of a person considered not secondos.

4 Computation of the estimated sample weights and the initial weights

The sampling weights are simply the inverse of the inclusion probabilities, the probabilities of being selected for the sample. The sampling weights are calculated to offset the effect of not having the entire population but only a part of it. It is calculated for every person selected in the sample, but not every person selected reply to the survey. In practice there is always non-response. To compensate this effect, the effect of the non-response the weights of the respondent

are adjusted. The adjusted sample weights are commonly called initial weights. The adjustment was realized using non-response questionnaire and applying the method of "CHAID". This is the same method that FORS apply for the surveys of Swiss Household Panel. The detailed description of the method can be found on the website of FORS.

5 The subsequent steps

Once the initial weights calculated, the various stages of calculation, i.e. adjustments for non-response to the individual questionnaire at the individual level, or to the household questionnaire at the household level, or the weight-sharing (not applicable in the first year) or the marginal calibration were effectuated by FORS, in the same way as for the survey of the SHP, evidently applying necessary minor adaptations for the Cohort sample. The final weights to use for analysis are the ones named "wp13tcs", "wh13tcs" and "wn13tcs" depending on the level (personal, household, network) the data are used. For the details of these methods see the website of FORS (http://forscenter.ch/en/our-surveys/swiss-household-panel/documentationfaq-2/methods/weighting/).