Annex IV

Description of the weighting process

Weighting in the SEEMIG pilot study[1]

As it was explained before, the target population of the SEEMIG study consists of the following groups:

(1) current LFS-household members who are migrants (aged 15-74);

(2) former LFS-household members of an existing LFS-household who are migrants (aged 15-74)[2];

(3) migrants who are brothers or sisters of a current LFS-household member living in Hungary.

All across the definition household members and migrants are restricted with the age limit 15-74.

The definition has a great impact on both the questionnaire and the weighting scheme. Each of (1)-(3) defines a migrant sub-population, the first two of which are disjoint. As for sampling, reaching sub-population 1 and 2 means direct sampling, while reaching the third one is an indirect sampling method. According to the definition we have three major questionnaire blocks.

Block 1 refers to each current LFS-household member with two important questions, asking

- whether they live abroad (they are migrants)
- how many brothers or sisters they have and how many of them live in Hungary (whether they belong to migrant sub-population 3).

Block 2 refers to former LFS-household members. Current LFS-households are asked

- to list their former household members that are migrants
- and then to provide some further information on them (how many brothers or sisters they have and how many of them live in Hungary, whether they belong to migrant sub-population 3).

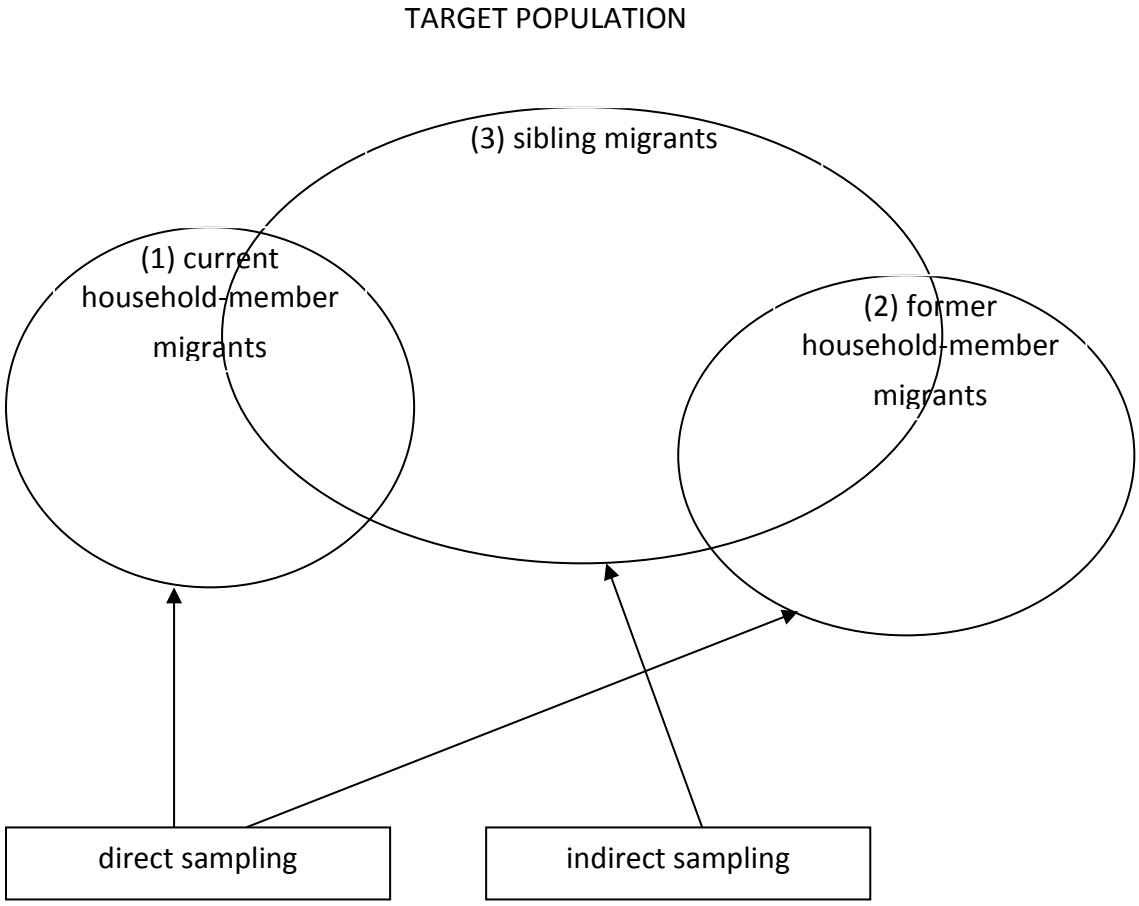
Block 3 refers to the migrant brothers and sisters of the current LFS-household members. Current LFS-households are asked

- to list their members' migrant sibling and
- and then to provide some further information on them (questions that help to decide whether they belong to migrant sub-population 1 or 2).

[1] This chapter was prepared and the weighting and estimates were made by Gergely Fraller.

[2] For proper sampling these migrants were enumerated in the household they moved abroad from.

Figure 4 The weighting approach



For each migrant enumerated in the sample in any questionnaire block

- it must be indicated whether they are in the focus of other blocks;
- it must be decided which migrant sub-population they belong to: 1 or 2 or 3 or (1 and 3) or (2 and 3).

The basic idea of weighting can be summarized in a few steps.

The initial weight is the LFS final cross-sectional weight → weight0.

For those that responded questionnaire block 1, weight0 must be corrected for non-response in an ordinary way → weight1. Current household member migrants registered in block 1 represent migrant sub-population 1 with weight1.

For migrants registered in questionnaire block 2 weight0 must be corrected in an ordinary way in three steps (correction for non-response at SEEMIG questionnaire level, at block 2 level and within block 2 level) → weight2. Former household member migrants registered in block 2 represent migrant sub-population 2 with weight2.

Weighting of the sample of sibling migrants is not an ordinary cross-sectional weighting. Because of the indirect sampling, we applied Generalized Weight Share Method (GWSM) → weight3. The sample of sibling migrants with weight3 represents migrant sub-population 3.

Weight1-3 does not mean any proper representation of the whole migrant target population, as the sections of sub-populations 1-3 are represented twice. So the final step is a division of weight1-3 by two for those migrants that fall into any of the intersections.

For a proper weighting it is essential to have precise answers in the questionnaire, especially for GWSM, where we need to know exactly the number of links between the sibling migrant and the LFS-household members living in Hungary. So before weighting we had an exhaustive data editing to do.

Data editing

This is the preparation of the sample data set for weighting.

The majority of data editing refers to checking those households that have sibling migrants, especially those that refused questionnaire block 3. Checking those records it turned out that the questionnaire was not always easy to complete correctly.

Fortunately, the questions in block 1 (how many brothers or sisters they have and how many of them live in Hungary) with the ordinary LFS questions on household status and relations meant a very good basis for editing. The following mistakes could be corrected:

- Duplications could be identified (and ceased): cases where the same migrant was registered in different blocks as different person.
- It could turn out that a current household member is a migrant, although it was not noted in block 1 or block 1 was altogether refused.
- Answers on the number of siblings of migrants and LFS-household members could be harmonised.
- It could turn out that some respondents counted all their siblings irrespective of the age limit.

One of the most important results is that we were able to build up the list of sibling migrants and the number of links for GWSM in those households that refused block 3.[3]

A necessary step was to construct a variable that indicates whether a sibling migrant belongs to sub-population 1 or 2. This question was not asked directly as it was regarded complicated to answer. Instead we used questions on

- whether the migrant is financially dependent on an LFS-household in Hungary;
- who the migrant lived with before migration;
- who the migrant lives abroad with.

As a result of this editing (and imputation), we had a data set where the answers of different blocks are harmonised.

The weighting process

In this chapter we give a description of the weighting procedure, steps (w0)-(w4).

W0 – initial weight

[3] It must be recognised that there were only few cases that remained questionable, but the effect for the estimates is unimportant.

We accepted LFS final cross-sectional weight as the initial weight for weighting the SEEMIG sample. On the one hand it may seem natural. However, there are some drawbacks. SEEMIG is a very special survey where the importance of specific non-response correction is obvious. The SEEMIG migrant battery being only a module of standard LFS, we had to accept and use LFS methodology, which is probably not the best possible solution as for results.

Theoretically, it is possible to make a cross-sectional weighting for SEEMIG-LFS which is not the same as LFS weighting. At the planning stage we were to use census 2011 results and build up some auxiliary variables on migrants into the calibration of SEEMIG-LFS sample Due to some technical problems, this could not be realised, this is why we used LFS final cross-sectional weight as initial weights (weight0).

W1 – weighting for current household member migrants

There are 785 individuals in the LFS sample who refused SEEMIG block 1, which is a very small proportion of the 51 thousand LFS sample (aged 15-74). The majority of non-response of this type occurred at household level. The first correction of the initial weight weight0 was to correct this kind of non-response.

As we have a couple of information of non-respondents from the LFS questionnaire, we applied logistic regression model to get the probability of answering SEEMIG questionnaire, where we used individual and household level explanatory variables.

Then the initial weight, weight0 was divided by this probability, resulting in weight1, which is the weight for block1 respondents. This weight is directly applicable for the current LFS-household member migrants representing sub-population 1.

W2 – weighting for former household member migrants

To reach the sample of former household member migrants of the LFS sample, we came through non-response at three stages. First we had 356 households that refused SEEMIG as a whole (the major part of non-response mentioned in the previous section). Then we had 230 households that refused questionnaire block 2. Finally, we had 115 non-respondent individuals in households that listed their former household members who are migrants, but gave no further information on them.

Weight-adjustment for each of these three stages is a simple non-response correction, similar to the one mentioned previously. The main question here is the second-stage non-response. Although those 230 households are very suspicious to be involved in migration because of refusal, we found no statistical evidence for this, so it caused only a slight weight correction. Finally, we got weight2 for our 583 former household member migrants with which they represent sub-population 2.

W3 – weighting for sibling migrants, GWSM

Thanks to data editing and imputation we had non-response for sibling migrants at only two stages (see chapter 2). First we had 356 households that refused SEEMIG as a whole (the major part of non-response mentioned in the previous section). Weight-adjustment for this stage is a simple non-response correction, the same as the one mentioned previously. In the respondent households we had 1626 sibling migrants[4] registered. For these migrants we applied GWSM, taken into account the number of links between the migrant and the LFS

[4] This figure includes those belonging both to the group of sibling-migrants and to the group of (former) household-member migrants.

sample and the LFS target population (for the sake of simplicity: number of siblings of a migrant in the LFS household and in Hungary).

This GWSM-weight had to be adjusted as we only had 946 respondents of the 1626 sibling migrants. This was done with modelling response probability with logistic regression.

Finally, we got weight3 for our 946 sibling migrants with which they represent sub-population 3.

W4 – final weights for migrants

Sections of sub-populations 1-3 are represented twice, so the final step was a correction for this. In each sub-sample (of current and former household member migrants and sibling migrants) we identified individuals that fall into the sections of sub-populations1-3.

For migrants that belong to sub-samples 1 and 3 at the same time, the final weight is the average of weight1 and weight3.

For migrants that belong to sub-samples 2 and 3 at the same time, the final weight is the average of weight2 and weight3.

For migrants that belong to sub-sample 1 and are sibling migrants at the same time, the final weight is half of weight1.

For migrants that belong to sub-sample 2 and are sibling migrants at the same time, the final weight is half of weight2.

For migrants that belong to sub-sample 3 and are current or former household migrants at the same time, the final weight is half of weight3.

For the rest of the migrants the final weight is equal to weight1-3, depending on which subsample they belong to.

Weighting process in Serbia

Weighting process initially correspond to estimates for two-stage stratified sampling design (PPSWR at the first stage and SRSWOR at the second stage) within six rotation groups as representative sub-samples. The initial weight for each household is a product of inverse inclusion probability at each stage and correction for non-response and was used as a weight in the SEEMIG sample.

Estimates for population (individual level) were obtained to correspond to current demographics projections, as a post-stratification, according to age (below 14, 15 and more), and regions and was used only for LFS data.