# (Brief) Methodological Note

## Background

The Bulgaria Crisis Monitoring Survey (CMS) was a multi-topic panel survey conducted in February 2010, October 2010 and February 2011 aimed at tracking the impact of the economic crisis over time. The CMS collected information on various demographic and socioeconomic characteristics of the households and individual household members, including questions on labor market participation and earnings, access to and receipts from social protection programs, informal safety nets and remittances. The survey focused on how Bulgarian households were affected by and were coping with the economic crisis, and included information on informal employment, reduced spending, postponement of investments, the sale of household assets, and the reliance on formal and informal credit. A national sample of 2,400 households was selected for the main CMS, based on a stratified multi-stage sample design. Given the special need to study the more vulnerable ethnic minority Roma population, an independent "booster sample" of 300 households was selected in settlements and neighborhoods identified as predominantly Roma.

The Bulgarian Longitudinal Inclusive Society Survey (BLISS) was conducted in March and April 2013, as a continuation of the longitudinal CMS, with an additional module about the skills of the adult population in Bulgaria. The BLISS was conducted using the same panel of households as the CMS for the main survey and the "booster" (Roma) survey. Therefore the BLISS will be treated as an additional wave of the CMS in the longitudinal analysis, as well as a cross-sectional survey.

The sample for the CMS is a stratified cluster sample drawn from the listing of EAs from the national census. The results from these surveys are representative of the national population. The frame for the booster survey was developed from key informants and then scientific sampling was conducted. The booster sample is then representative of only the frame from which it was constructed, rather a general population.

The goal of this methodology is to develop a single set of weights that allows data collected from the two frames to be used in the analysis together. This note builds on the work of David Megill, who developed cross sectional weights for the datasets.

## Assumptions

For the purposes of analysis, it is necessary to assume for all waves that the booster sample was constructed as a representative of the population living in Roma neighborhoods. This implies that the population of Roma captured in the main sample and the population of Roma captured in the booster sample are statistically identical. Similarly, it is assumed that the non-Roma captured in the booster sample as they were living in a Roma neighborhood are also statistically identical to the non-Roma in the main survey. It is then possible to combine the two datasets and apply population projections for each ethnicity (developed from the 2011 census) to down-weight / up-weight as necessary to match those totals.

The main sample contains a Roma population of approximately 5 percent, which is well within the confidence interval of 4.5 percent estimated for the same month and year using the census projections. The booster sample contains 59 percent Roma.

**Steps for the individual weights**

The following methodology applies to all the CMS and BLISS cross-sections as well as to the panel dataset at the individual level. As an example, we are going to focus on the construction of the combined weights for the BLISS cross-section.

1. The first step is to calculate the expected population of Roma and non-Roma at the province level for the timing of the fieldwork (March/April 2013 for BLISS) by doing population projections using the 2001 and 2011 census information by province and ethnicity. When the ethnicity information is missing, the individual is considered to be non-Roma.
2. Merge the roster information (including ethnicity) and the two weights files (main sample and booster). Generate new variable which has the relevant weight from each source ("wta_comb").
3. Identify any provinces in the data which have no sub-population members. In the BLISS cross sectional dataset, for example, Kardzhali and Smolyan had no Roma observation but there were Roma recorded in the census. They were combined with Haskovo for the analysis. Similarly Pleven was combined with Vratza, Targovishte with Razgrad, and the two Sofia regions combined. The number of empty regions depends on the dataset and this step must be repeated each time.
4. Aggregate the ethnicity question into "Roma" / "non-Roma", considering the observations with missing ethnicity as non-Roma.
5. Generate sum of the weights by region and ethnicity ("survey"). This is effectively the projected population total from these weights. Note that the region variable includes any province aggregation from step 3.
6. Paste in the code that defines the census population totals from the Excel code generator.
7. Generate an adjustment factor (census/survey) and multiply the existing weights by the adjustment factor ("wta_adj").
8. For verification, "total wta_adj, over(ethnicity)" should yield the following:

```
--------------------------------------------------------------
        Over |      Total   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
wta_adj      |
     nonRoma |    6941724   60009.83       6824086     7059361
        Roma |    317010.1  16548.13       284570.6    349449.6
--------------------------------------------------------------
```

This methodology works with all the CMS and BLISS cross-sections as well as to the panel dataset at the individual level because it generates the sum of the weights internally and the known totals do not change. Note that it is necessary to re-check the empty regions for each new dataset. For example, Sliven also does not have any Roma observations in the panel dataset even though it does in the cross section. The

Excel file needs to be updated with the correctly aggregated totals and then the concatenate commands adjusted. The Stata code is then automatically generated

### Steps for the skills module weights

Unlike the other datasets with information at the individual level, the skills module only covers adults aged 18 to 65 years. Therefore, in order to construct the combined weights for the BLISS skills module, it is necessary to repeat step 1 restricting the population totals to the adult population (18-65). However, the 2001 and 2011 census do not provide information on the adult population totals by province and ethnicity. As a result, in 2001 this information was approximated using the national proportion of individuals aged 20 to 69 over the total population and in 2011 using the percentage of individuals aged 15 to 64 by province.

After doing the population projections using the estimated adult population totals by province and ethnicity, repeat steps 2 to 8 from the previous section.

Finally, a hard correction was applied to the weights, so that both the BLISS full and the skills samples have the same population totals by age group, gender, ethnicity and labor market status[1]. This adjustment assumes that the real population is the one captured in the full BLISS cross-sectional sample.

### Steps for the household weights

In order to construct the weights for all the CMS and BLISS cross-sections as well as for the panel dataset at the household level, one should follow the same procedure explained for the weights construction at the individual level with a slight variation in the first step.

This time, the population totals used in the population projections should be the total number of Roma[2] and non-Roma households at the province level. Given that this information is not available in the 2001 and 2011 census, it was approximated by dividing the total population numbers by the average household size by region[3] and ethnicity from the sample.

After doing the population projections using the estimated total number of households by province and ethnicity, repeat steps 2 to 8 from the previous section.

### Contact Information

Questions, comments or criticisms? Contact Kristen Himelein (khimelein@worldbank.org), Victoria Levin (vlevin@worldbank.org) or Silvia Guallar Artal (sguallarartal@worldbank.org).

[1] In the hard correction, four age categories were considered (18-29, 30-39, 40-49, 50-65), two ethnicity categories (Roma, non-Roma) and four labor market statuses (employed, unemployed, inactive/retired/disable and student).

[2] A household is defined as Roma if at least 50% of its members self-identify themselves as Roma.

[3] Provinces were aggregated at the regional level to ensure a big enough sample size when calculating the average household size. Six regions were considered: North-West, North-Center, North-East, South-West, South-Center, South-East.