Sample Design and Weighting Procedures for Serbia STEP Employer Survey

David J. Megill
Sampling Consultant, World Bank
November 2015

1. Sample Design for Serbia Employer Survey

The sampling frame for the Serbia Employer Survey was based on the business register of all enterprises in Serbia with 5 or more employees, with information on the geographic location, the economic activity and the number of employees. The enterprises in the sampling frame were stratified by region and size in terms of the number of employees. A stratified two-stage sample design was used for the Serbia Employer Survey, with a sample of enterprises selected at the first stage, and branches selected at the second stage.

First it was necessary to allocate the sample by region and employment size based on the distribution of the frame and the sample size needed for each domain. This sample size was tripled in order to select a reserve of potential replacement enterprises at the same time. Then the original sample was selected as a systematic subsample of all the enterprises selected at the initial phase; the remaining sample enterprises were used as a reserve for possible replacements. In the case of strata that do not have triple the sample size in the frame, all of the sample enterprises were selected in the initial phase. Table 1 below shows the target sample size for each stratum; these numbers were multiplied by 3 to obtain the total sample to be selected in the first phase, including the reserves for possible replacement. The strata of enterprises with 101 to 200 employees and 201+ employees were combined into one stratum for 101+ employees. A larger sample is allocated to this stratum since the large enterprises generally have many more branches.

Table 1. Proposed allocation of sample enterprises for Serbia Employer Survey by region and size stratum

	Number of employees				
Region	5-15 employees	16-50 employees	51-100 employees	101+ employees	Total
(1) Beograd	60	60	60	70	250
(2) Vojvodina	60	60	60	70	250
(3) Sumadija and West Serbia	60	60	60	70	250
(4) South and East Serbia	60	60	60	70	250
Total	240	240	240	280	1,000

Within each region by size stratum the enterprises were selected systematically with probability proportional to size (PPS), after sorting the frame by district and activity. The measure of size in this case is the number of employees for each enterprise in the frame. In an initial phase we selected approximately 3 times the number of enterprises specified for each stratum in Table 1. According to the distribution of the Serbia frame of enterprises by region and size stratum, there are only two size strata in the South and East Serbia region that have less than 3 times the number of sample enterprises in the frame: the 51-100 employees and 101+ employees strata. For these two strata, all the enterprises were selected in the initial phase that included reserves for possible replacement.

In the case of enterprises that have more than one branch (location), one branch will be selected with equal probability at the second stage. In this case the number of branches to be interviewed would be the same as the number of sample enterprises. The number of branches is generally correlated with the number of employees, so this sampling strategy should reduce the variability in the weights. In the case of large enterprises that have a measure of size greater than the sampling interval, the number of branches to be selected was determined based on the number of "hits", as explained later.

The enterprises were selected at the first stage systematically with PPS, using the number of employees as the measure of size. The following steps were used for this sample selection:

1. Since the sample enterprises are selected systematically with PPS within each stratum, it is first necessary to calculate the sampling interval, which is equal to the cumulated total number of employees in the frame for the stratum divided by the number of enterprises to be selected in the stratum (the corresponding number in Table 1 multiplied by 3).

2. Any enterprise with a measure of size (number of employees) greater than the sampling interval for the stratum will be selected with certainty (that is, with a probability of 1). These self-representing (SR) enterprises should be separated from frame and included in the original sample to be interviewed.

3. For each SR enterprise, divide the number of employees in the frame by the original sampling interval, and round to the next integer (for example, 1.5 would be rounded to 2) in order to determine the number of "hits". The number of "hits" will correspond to the number of branches to be selected in the SR enterprise.

4. Sum the number of "hits" for all SR enterprises in the stratum. Subtract this number from the total number of sample branches allocated to the stratum to determine the number of non-self-representing (NSR) sample enterprises to be selected in the stratum. For example, in the case of the 101+ employee stratum in Belgrade, if there are 4 SR enterprises with a total of 6 "hits", it would only be necessary to select 204 NSR enterprises in that stratum for the initial phase, and 64 NSR sample enterprises for the target sample in the second phase. In this example we would have a total of 68 sample enterprises and 70 sample branches in the final sample for this stratum.

5. Multiply the number of NSR sample enterprises to be selected in each stratum by 3 to determine the total number of NSR enterprises to be selected for the initial phase, including the reserves for replacement. If this is more than the number of enterprises in the frame for that stratum, it will only be necessary to select the number of sample enterprises specified in Table 1 for that stratum; the remaining enterprises in the frame for that stratum will be used as reserves for possible replacement.

6. Select triple the number of NSR sample enterprises in each stratum systematically with PPS after separating the SR enterprises from the frame. It will be necessary to cumulate the measures of size (number of employees) again (excluding the SR enterprises) and calculate a new sampling interval.

7. In the second phase it will be necessary to select the original sample of enterprises to be included in the survey from all the enterprises selected in the initial phase; the remaining enterprises will be in the reserve for replacements. Since the enterprises are selected systematically with PPS within each stratum in the initial phase, a subsample of enterprises will be selected with equal probability in the second phase to be interviewed. For most strata we will have exactly 3 times the number of NSR sample enterprises in the original sample, so we can systematically assign numbers from 1 to 3 to all the sample NSR enterprises in the initial phase, in the same order in which they were selected. This would identify three systematic subsamples for each stratum. Randomly select an integer from 1 to 3 to determine the original sample of enterprises to be interviewed for each stratum. The remaining two subsamples would be used as reserves for replacement.

8. In the case of the two strata with less than 3 times the number of sample enterprises in the frame, all the sample enterprises are selected in the initial phase. Once the original sample of enterprises is selected systematically with PPS, the remaining enterprises in the frame will be included in the reserve for possible replacements.

9. After each iteration of the systematic PPS selection, it is necessary to separate the SR sample firms (those with a measure of size greater than the sampling interval), and adjust the number of NSR sample firms to be selected accordingly. Then a new sampling interval will be calculated for the selection of the NSR sample firms. The denominator of the sampling interval will be 3 times the number of sample NSR enterprises that still need to be selected to have the specified number of sample branches for the stratum (after subtracting the total number of sample branches in all the SR firms for that stratum). This process will continue until no more firms in the list have a measure of size greater than the interval. Following the first iteration of the systematic PPS selection, any additional SR enterprises that are identified will be allocated one branch each to be selected.

10. For all the sample SR and NSR sample enterprises, it will be necessary to make a list of all the workplaces (branches).

11. At the second stage, select one branch from each sample NSR enterprise with equal probability. The headquarters should be counted as a branch and be listed as one of the possible branches to be selected for interviewing. For each SR enterprise, the number of "hits" will determine the number of sample branches to be selected with equal probability.

Once the original sample of SR and NSR enterprises have been selected for each stratum, it was necessary to contact them to obtain a list of their branches and then randomly select a branch to interview. The total number of branches in each sample enterprise should be recorded since this information will be needed for calculating the weights.

2. Weighting Procedures for Serbia Employer Survey

In order for the sample estimates from the Serbia Employer Survey data to be representative of the population of enterprises, it is necessary to multiply the data by a sampling weight, or expansion factor. The basic weight for each sample branch would be equal to the inverse of its probability of selection.

As described above, a stratified two-stage sample design was used for the Serbia Employer Survey. At the first stage a sample of enterprises was selected in each region by employment size stratum systematically with PPS, based on the number of employees. At this stage some of the enterprises were selected with a probability of 1 because of their size, so they are considered to be self-representing (SR), that is, selected with certainty at the first stage. At the second stage more than one branch can be selected from the large self-representing enterprises with a measure of size that is a multiple of the sampling interval. In the case of the non-self-representing (NSR) sample enterprises, only one branch is selected in each enterprise at the second stage. The weights are specified here separately for the SR and NSR sample enterprises.

For the SR enterprises, the probabilities of selection can be expressed as follows:

$$p_{Shi} = \frac{b_{hi}}{B_{hi}},$$

where:

p_{Shi} = probability of selection for the sample branches in the i-th SR enterprise in stratum (region, employment size) h

b_{hi} = number of sample branches selected and interviewed for the i-th SR enterprise in stratum h

B_{hi} = total number of branches identified in the frame for the i-th SR enterprise in stratum h

In this case the first stage probability of selection is 1, so it does not appear in the formula for the overall probability of selection. The basic weight for the SR sample enterprises is the inverse of this probability of selection, and can be expressed as follows:

$$W_{Shi} = \frac{B_{hi}}{b_{hi}},$$

where:

W_{Shi} = basic weight for the sample branches in the i-th SR enterprise in stratum h

For the NSR sample enterprises, the overall probabilities of selection within each stratum includes components from the first and second sampling stages. This probability can be expressed as follows:

$$p_{Nhi} = \frac{n_h \times E_{hi}}{E_{Nh}} \times \frac{1}{B_{hi}},$$

where:

p_{Nhi} = probability of selection for the sample branch in the i-th sample NSR enterprise in stratum h

n_h = number of NSR sample enterprises selected in stratum h

E_{hi} = number of employees in the frame for the i-th NSR enterprise in stratum h

E_{Nh} = total number of employees in the frame for all the NSR enterprises in stratum h (that is, the cumulated measure of size)

B_{hi} = total number of branches identified in the frame for the i-th NSR enterprise in stratum h

The two components of this probability correspond to the individual sampling stages. The second stage probability is based on the assumption that one branch is selected in each NSR sample enterprise. In the case of enterprises with only one branch, the second stage probability is equal to 1.

The basic weight for the NSR sample establishments is the inverse of this probability of selection, and can be expressed as follows:

$$W_{Nhi} = \frac{E_{Nh} \times B_{hi}}{n_h \times E_{hi}},$$

where:

W_{Nhi} = basic weight for the sample branch in the i-th NSR sample enterprise in stratum h

In the case of sample enterprises that cannot be interviewed, the instructions were to select a replacement from reserve pool of enterprises. If it is not possible to replace some sample enterprises, it may be necessary to adjust the basic sampling weights for the corresponding strata. It is also possible that not all the sample branches within a large enterprise can be interviewed, in which case a second-stage adjustment would be made to the weights. In the case of the large self-representing enterprises that cannot be interviewed, it may be necessary to make a special kind of imputation using information from the frame, since each SR enterprise only represents itself. The final adjustment of the weights depends on the particular results of the survey. In the case of the NSR sample enterprises that cannot be replaced, the weight for the other sample NSR enterprises in the same stratum be adjusted as follows:

$$W_{Nhi} = \frac{E_{Nh} \times B_{hi}}{n_h \times E_{hi}} \times \frac{n_h}{n'_h} = \frac{E_{Nh} \times B_{hi}}{n'_h \times E_{hi}},$$

where:

n'_h = number of NSR sample enterprises successfully interviewed in stratum h

Appendix:

Table 1. STEP Employer Survey Report: Overall Summary of Interview Outcome by Strata

Stratum/Number of firms	Target Sample Size	Reserve Sample	Extra Reserve Sample	Distribution of Firms by Result Code and stratum (for all the visits)								
				1. Completed	Ratio to target sample, %	2.Address is not found	3.The organization doesn't exist	4.The organization refused	5.Ineligible. (on size, or status)	6.The respondent refused	7.The respondent is not available during our survey	8. Other
1 Belgrade	250	451	0	261	104,4%	43	9	9	21	147	86	3
2 Vojvodina	250	367	20	266	106,4%	13	13	23	36	250	29	8
3 Sumadija - West	250	440	30	256	102,4%	7	22	18	32	170	25	7
4 South - East	250	465	15	256	102,4%	2	18	2	25	76	23	5
Total	1000	1723	65	1039	103,9%	65	63	52	115	648	166	24
Actually visited firms				48%		3%	3%	2%	5%	30%	8%	1%

Table 2. Distribution of achieved sample by sector and strata

Economic activity by sectors	Code	Belgrade				Other Urban				Share of total, %	Control check: Sample frame
		5-15	16-50	51-100	101+	5-15	16-50	51-100	101+		%

A	**Agriculture, forestry and fishing**	**01**	0	1	0	0	2	9	6	12	2.9%	2,5%
B	**Mining and quarrying**	**02**	3	1	0	0	3	0	0	0	0.7%	0,4%
C	**Manufacturing**	**03**	13	18	8	15	130	82	58	99	40.7%	26,6 %
D	**Electricity, gas, steam and air conditioning supply**	**04**	0	0	0	0	6	0	0	1	0.7%	0,2%
E	**Water supply; sewerage, waste management and remediation activities**	**05**	0	0	1	0	4	2	2	4	1.3%	0,9%
F	**Construction**	**06**	8	4	4	8	22	16	12	13	8.4%	7,9%
G	**Wholesale and retail trade; repair of motor vehicles and motorcycles**	**07**	13	19	14	14	24	39	57	32	20.4%	30,9 %
H	**Transportation and storage**	**08**	9	3	4	4	9	15	13	12	6.6%	6,0%
I	**Accommodation and food service activities**	**09**	0	4	5	2	3	5	8	4	3.0%	3,9%
J	**Information and communication**	**10**	5	3	4	5	4	3	5	1	2.9%	4,0%
K	**Financial and insurance activities**	**11**	7	1	1	1	4	1	1	0	1.5%	1,1%
L	**Real estate activities**	**12**	0	1	0	1	0	0	1	0	0.3%	0,4%
M	**Professional, scientific and technical activities**	**13**	7	6	13	8	0	9	22	2	6.4%	9,7%
N	**Administrative and support service activities**	**14**	9	1	4	3	5	5	0	3	2.9%	2,7%
O	**Public administration and defense; compulsory social security**	**15**	0	0	0	0	0	0	0	0	0.0%	0,0%
P	**Education**	**16**	0	1	1	0	0	0	1	0	0.3%	0,9%
Q	**Human health and social work activities**	**17**	0	0	1	0	0	0	1	1	0.3%	0,5%
R	**Arts, entertainment and recreation**	**18**	0	1	0	0	1	0	1	2	0.5%	0,5%
S	**Other service activities**	**19**	0	0	2	0	0	0	1	0	0.3%	0,9%
T	**Activities of households as employers; undifferentiated goods- and services-producing activities of households for own use**	**20**	0	0	0	0	0	0	0	0	2.9%	0,0%
	Total										**100.0%**	