Sampling and Weighting Methodology for Kosovo STEP Employer Survey

David J. Megill
Sampling Consultant, World Bank
April 2016

1. Summary of Sample Design for Kosovo Employer Survey

This section presents a brief summary of the sample design for the Kosovo Employer Survey. The sampling frame is based on a database of firms compiled by the ASK, and a small complementary frame of Serb enterprises in Northern Kosovo that was obtained independently. Two geographic domains were defined for this survey: Prishtina and Other (remaining regions of Kosovo). The sampling frame was stratified by geographic domain and four size strata in reference to the number of employees: 5-9, 10-15, 16-50 and 51+ employees. Table 1 shows the distribution of the firms in the frame by geographic domain and size strata.

Table 1. Distribution of firms in the Kosovo sampling frame for the Kosovo STEP Employer Survey by geographic domain and employment size strata

Geographic domain	5-9 employees	10-15 employees	16-50 employees	51+ employees	Total
Prishtina	824	274	285	154	1,537
Other	1,941	598	507	158	3,204
Total	2,765	872	792	312	4,741

The target sample size for the selection of firms and branches by stratum is shown in Table 2. This sample allocation was based on providing reliable results for each of the tabulation cells, corresponding to the individual strata.

Table 2. Number of target sample firms/branches for Kosovo Employer Survey, by geographic and employment size strata

Geographic domain	5-9 employees	10-15 employees	16-50 employees	51+ employees	Total
Prishtina	60	60	60	60	240
Other	80	60	60	60	260
Total	140	120	120	120	500

This sample size was initially tripled in order to select a reserve of potential replacement firms at the same time. In the first phase the ASK selected this larger sample from their frame. Given the smaller number of firms in the strata with 51 or more employees, all were selected for this first sampling phase: 154 firms for Prishtina and 165 firms for the Other geographic domain. The distribution of the first phase sample selected by the ASK is presented in Table 3.

Table 3. Distribution of larger sample of firms selected by the ASK for the first phase, by geographic and employment size strata, including the reserve list of sample firms to be used for possible replacement

Geographic domain	5-9 employees	10-15 employees	16-50 employees	51+ employees	Total
Prishtina	180	180	180	154	694
Other	240	180	180	165	765
Total	420	360	360	319	1,459

A separate frame of 39 Serb firms in Northern Kosovo was combined with this large sample of firms for the first phase.

In the second phase the target sample of firms/branches specified in Table 2 was selected as a subsample from the combined frame from the first phase. The remaining (non-selected) firms from the first phase were used as a reserve for selecting replacements for sample firms that could not be interviewed. For the second phase the number of sample firms specified in Table 2 for each stratum under 51 employees was selected from the combined first phase frame using random systematic sampling with equal probabilities within the stratum. The firms in the frame for each stratum were sorted in the following order: province, municipality, Activity ID and number of employees. Based on the systematic sampling, this provided implicit stratification to ensure a representative sample geographically and by economic activity.

In the case of the stratum of 51+ employees in the Prishtina and Other geographic domains, within each stratum the subsample of firms was selected systematically with probability proportional to size (PPS), where the measure of size was based on the number of employees. The reason for using PPS selection for this stratum is that the number of employees varies considerably by firm; the largest firm in the frame has 4988 employees. The largest firms were selected with a probability of 1 in the second phase, and some of the firms with more the 1600 employees were allocated 2 to 4 sample branches depending on their size. All of the remaining sample firms were allocated one sample branch each. The original sample for the Kosovo Employer Survey had 490 firms and 500 branches. This sample of firms was identified in a spreadsheet that also specified the number of branches to be selected in each firm.

2. Selection of sample branches

It was necessary to contact each of the 490 sample firms and list all the workplaces (branches), including the headquarters. It is important to record the total number of branches, because this information was needed later for the calculation of the weights. Most of the smaller firms only had one workplace, so it was not necessary to select a branch. In the case of firms with more than one branch, the spreadsheet identifying the sample firms specified how many branches should be selected in each sample firm. There were only 6 self-representing firms with an allocation of 2 or more branches; one branch was selected in each of the remaining 484 sample firms.

In the case of the sample firms that are allocated only one sample branch, the sample branch was selected based on a random number between 1 and the total number of branches. In the case of the six sample firms allocated between 2 and 4 branches, the branches were selected using random systematic sampling.

3. Selection of replacement firms

In the case of an original sample firm that could not be interviewed for any reason, first it was necessary to make a strong effort to complete the interview. Only the supervisor could make a decision to replace an original sample firm. If it was still not possible to complete an interview, it was necessary to select a replacement firm from the reserve list. The replacement firm should always be selected from the same geographic domain and size stratum. When possible the replacement should be selected from the same municipality, and the same or a similar activity.

In order to facilitate the selection of replacement firms, the full frame of sample and replacement firms was listed in another spreadsheet. The firms within each stratum in this frame appeared in the same order used for the systematic sample selection. Within each stratum the firms were sorted hierarchically by province and municipality codes, Activity ID and number of employees. The sample firms were identified by a Sample ID code which was used for operational control throughout the survey process. Another code that was introduced for operational control is the Frame ID, from 1 to 1459, assigned to all the firms in the ordered frame.

When one of the original sample firms had to be replaced, the following procedures were used to select the replacement firm:

1. In the spreadsheet with the ordered frame, find the Sample ID of the sample firm that is being replaced.

2. Go down the list to the next non-selected (reserve) firm in the ordered frame. If it is in the same municipality and the same (or similar) activity, select this firm as the replacement. If the municipality changes, select the reserve firm found prior to the sample firm being replaced, if it is in the same municipality. Then mark the selected replacement with an R in a new column, followed by the Sample ID of the firm being replaced, to indicate that it has already been selected as a replacement, so it will not be selected again. Within each stratum the firms in the frame were ordered by province,

municipality and activity, so it is only necessary to follow the ordering of the list. If the activity of the replacement does not match the original, this is not a problem, as some activities may have few firms in the frame.

3. The first replacements will be easier to find, because all the reserve firms will be available. However, over time as more replacements are selected it may be necessary to select a replacement that is less close in the list to the original sample firm being replaced. When all the reserve firms in a municipality are used as replacements, it will be necessary to select from the next municipality. If all the reserve firms in a stratum are used as replacements, this means that the response rate is extremely low, and the survey manager will need to be contacted for advice.

4. It is very important to record the relevant information for each replacement. A special control form should be developed for this purpose. For each original sample firm being replaced, it is necessary to record the Sample ID and Frame ID codes, as well as the Frame ID code of the replacement firm. The reason for the original replacement should be identified (firm refused, not found, etc.). If there are multiple replacements for the same original sample firm, there should be additional columns for the Frame IDs of the second and third replacements.

4. Procedures for calculating the weights

In order for the sample estimates from the Kosovo STEP Employer Survey to be representative of the firms and branches in the frame, 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 (calculated by multiplying the probabilities at each sampling stage). Although the sample firms were selected in two different phases (with a larger sample selected during the first phase that included the reserves for possible replacement), the probabilities are the same as if the firms had been selected directly from the frame in one sampling stage. As described in the section on the sample design, the probabilities are different for the strata of firms with less than 51 employees and the strata of firms with 51 or more employees, so the weighting procedures are described separately here.

4.1. Weights for strata of firms with less than 51 employees

In the case of the employment size strata with less than 51 employees, the sample firms were selected with equal probability within each stratum. Therefore the probability of the sample branch can be expressed as follows:

$$p_{hi} = \frac{f_h}{F_h} \times \frac{1}{B_{hi}},$$

where:

p_{hi} = probability of selection of the branch in the i-th sample firm of stratum h (in the case of the strata of firms with less than 51 employees)

f_h = number of sample firms selected and successfully interviewed in stratum h, including sample replacements

F_h = total number of firms in the sampling frame for stratum h

B_{hi} = total number of branches in the i-th sample firm in stratum h

This probability is based on the selection of one branch for each of the sample firms that have more than one branch (location). In the case of firms with only one branch, the second component of the probability would be equal to 1. The sample firms with more than one branch selected are all in the stratum of firms with 51 or more employees.

The basic weight for each sample firm/branch in the strata of firms with less than 51 employees is calculated as the inverse of this probability of selection. Based on the previous expression for the probability, the branch weight can be expressed as follows:

$$W_{hi} = \frac{F_h}{f_h} \times B_{hi} \, ,$$

where:

W_{hi} = weight for the sample branch in the i-th sample firm of stratum h

4.2. Weights for strata of firms with 51 or more employees

As described in the section on the sample design, the firms with 51 or more employees were selected systematically with PPS within each stratum, where the measure of size was based on the number of employees. Based on this sampling procedure, the firms with more than 1,600 employees were selected in the sample with certainty, and more than one branch was selected for a few of the largest firms. In the case of the sample firms that were not selected with certainty in the strata of firms with 51 or more employees, the probability of selection for the sample branch can be expressed as follows:

$$p_{hi} = \frac{f_h \times E_{hi}}{E_h} \times \frac{1}{B_{hi}} \, ,$$

where:

p_{hi} = probability of selection of the branch in the i-th sample non-certainty firm of stratum h (in the case of the strata of firms with 51 or more employees)

f_h = number of sample non-certainty firms selected and successfully interviewed in stratum h, including sample replacements

E_{hi} = number of employees in the frame (measure of size) for the i-th sample non-certainty firm in stratum h

E_h = total number of employees in all the non-certainty firms in the frame for stratum h (cumulated measure of size)

B_{hi} = total number of branches in the i-th sample firm in stratum h

The weight for the branches of these non-certainty sample firms in the strata of firms with 51 or more employees would be calculated as the inverse of this probability, and can be expressed as follows:

$$W_{hi} = \frac{E_h \times B_{hi}}{f_h \times E_{hi}},$$

where:

W_{hi} = weight of the sample branch in the i-th sample non-certainty firm of stratum h (in the case of the strata of firms with 51 or more employees)

For the large firms that were selected with certainty (that is, a probability of 1), the probability of selection for the sample branches would simply be the following:

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

where:

p_{hi} = probability of selection of the branches in the i-th sample certainty firm of stratum h

b_{hi} = number of sample branches selected and interviewed for the i-th sample certainty firm of stratum h, including replacements.

A spreadsheet with the final list of sample firms and branches with completed interviews was compiled with all the information from the sampling frame, as well as the total number of branches and selected branches for each firm. This spreadsheet was used for the calculation of the final weights using the formulas specified above. Some adjustments to the sample had to be made in the case where a particular large certainty firm had to be replaced, and the number of branches in the replacement firm was less than the number of branches that was specified to be selected. In a few cases it was necessary to select more than one replacement firm to select the specified number of branches. In the end the final data set had a total of 500 branches selected in 494 sample firms.

It should be pointed out that the sum of the weights of all the sample branches corresponds to the weighted estimate of the total number of branches in the frame, not the total number of firms.

Appendix:

Table No. 1 represents the final results. 351 firms were interviewed from the target sampleand 149 firms from the reserve sample leading to a total of 500 interviewed firms which is in harmony with the targeted sample.

Final Results		
Target Sample	**Reserve Sample**	**Total**
351	149	**500**

Table 1 - final results

Table No. 2 represents the final results as per geographic domain. The division is close to the targeted division per geographic domain, and the changes are a result of the randomization of the firms with more than one branch and to the reallocation of the offices of some firms.

Geographic domain		**Frequency**	**Percent**
Valid	Prishtina	217	43.4
	Other	283	56.6
	Total	500	100.0

Table 2 - final results per geographic domain

Table 3 table gives information about the total number of firms contacted and the results of their visits.

Visit Outcome	**Frequency**	**Percent**
1. No contact	2	0.3
2. Refuses to participate in the survey – refuses the interviewer	90	12.9
3. Refuses to participate in the survey – refuses the coordinator	14	2.0
4. Firm stopped working	23	3.3
5. Firm is in bankruptcy	3	0.4
6. Firm has been blocked for more than two months	4	0.6
7. Wrong address/ they moved away and it was not possible to get new data	25	3.6
8. Established contact, but appropriate person was not available	10	1.4
9. It doesn’t fit the target group – it has less than 5 employees	4	0.6
11. Firm was given to another coordinator, since the unit to be interviewed is positioned there	1	0.1
12. Scheduled interview	1	0.1
14. Completed firm	511	73.5

15. Other, please specify	7	1.0
Total	**695**	**100%**

Table 3 - visit outcome

Table no. 4 presents data related to the economic activity by sectors.

Economic activity by sector	**Code**	**Frequency**	**Percent**
Agriculture, forestry and fishing	A	4	0.8
Mining and quarrying	B	2	0.4
Manufacturing	C	61	12.2
Electricity, gas, steam and air conditioning supply	D	10	2.0
Water supply; sewerage, waste management and remediation activities	E	17	3.4
Construction	F	68	13.6
Wholesale and retail trade; repair of motor vehicles and motorcycles	G	90	18.0
Transportation and storage	H	23	4.6
Accommodation and food service activities	I	29	5.8
Information and communication	J	14	2.8
Financial and insurance activities	K	10	2.0
Real Estate activities	L	0	0.0
Professional, scientific and technical activities	M	4	0.8
Administrative and support service activities	N	5	1.0
Public administration and defense; compulsory social security	O	3	0.6
Education	P	6	1.2
Human health and social work activities	Q	16	3.2
Arts, entertainment and recreation	R	8	1.6
Other service activities	S	124	24.8
Activities of households as employers; undifferentiated goods	T	6	1.2
Total		**500**	**100%**

Table 4 –economic activity by sector