The Afghanistan 2014 Enterprise Surveys Data Set

I. Introduction

1. This document provides additional information on the data collected in Afghanistan between May 2013 and July 2013. The objective of the Enterprise Survey is to gain an understanding of what firms experience in the private sector.

The Enterprise Surveys, through interviews with firms in the manufacturing and service sectors, capture data covering measures of firm performance, firm structure as well as business perceptions on the biggest obstacles to enterprise growth, and the business environment in general. They are used to create business environment indicators that are comparable across countries.

The report outlines and describes the sampling design of the data, the data set structure as well as additional information that may be useful when using the data, such as information on non-response cases and the appropriate use of the weights.

II. Sampling Structure

2. The sample for Afghanistan was selected using stratified random sampling, following the methodology explained in the *Sampling Manual*[1]. Stratified random sampling[2] was preferred over simple random sampling for several reasons[3]:

a. To obtain unbiased estimates for different subdivisions of the population with some known level of precision.

b. To obtain unbiased estimates for the whole population. The whole population, or universe of the study, is the non-agricultural economy. It comprises: all manufacturing sectors according to the group classification of ISIC Revision 3.1: (group D), construction sector (group F), services sector (groups G and H), and transport, storage, and communications sector (group I). Note that this definition excludes the following sectors: financial intermediation (group J), real estate and renting activities (group K, except sub-sector 72, IT, which was added to the population), and all public or utilities-sectors.

c. To ensure that the final total sample includes establishments from all different sectors and that it is not concentrated in one or two of industries/sizes/regions.

d. To exploit the benefits of stratified sampling where population estimates, in most cases, will be more precise than using a simple random sampling method (i.e., lower standard errors, other things being equal.)

e. Stratification may produce a smaller bound on the error of estimation than would be produced by a simple random sample of the same size. This result is particularly true if measurements within strata are homogeneous.

f. The cost per observation in the survey may be reduced by stratification of the population elements into convenient groupings.

---

[1] The complete text can be found at http://www.enterprisesurveys.org/Methodology

[2] A stratified random sample is one obtained by separating the population elements into non-overlapping groups, called strata, and then selecting a simple random sample from each stratum. (Richard L. Scheaffer; Mendenhall, W.; Lyman, R., "Elementary Survey Sampling", Fifth Edition).

[3] Cochran, W., 1977, pp. 89; Lohr, Sharon, 1999, pp. 95

3. Three levels of stratification were used in this country: industry, establishment size, and region. The original sample design with specific information of the industries and regions chosen is described in Appendix C.

4. Industry stratification was designed in the way that follows: the universe was stratified into manufacturing, retail/wholesale, construction, and other services.

5. Size stratification was defined following the standardized definition for the rollout: small (5 to 19 employees), medium (20 to 99 employees), and large (more than 99 employees). For stratification purposes, the number of employees was defined on the basis of reported permanent full-time workers.

6. Regional stratification was defined in 5 regions (Kabul, Hirat, Kandahar, Mazar, Jalalabad) which include both city and the surrounding business areas.

III. Sampling implementation

7. Given the stratified design, a sample frame containing a complete and updated list of establishments as well as information on all stratification variables (number of employees, industry, and region) is required to draw the sample. Great efforts were made to obtain the best source for these listings. The quality of the sample frame was not optimal due to its being slightly out of date and due to the fact that individuals register their "business" for visa purposes but in fact do not have a business at all.

8. Nielsen India Pvt. Ltd. was hired to implement the Afghanistan 2014 Enterprise Survey. The team comprised of seven supervisors and sixteen interviewers.

9. The sample frame used in Afghanistan was combined from 3 sources. The list of manufacturing firms was obtained from the Afghanistan Investment Support Agency (AISA) and was updated in April 2012. The list for retail firms was generated by the implementing contractor for the 5 cities of fieldwork. The construction/other services list was obtained from the implementing contractor and is largely based on firms registered with AISA. The records obtained from AISA contained the following information

- Business name;
- Business address;
- Business sector classification code;
- Total numbers of male and female employees;
- President and Vice President;
- Phone number and sometimes FAX/email address.

Counts from sample frame are shown below.

Sample Frame

Source: Combination of AISA and Nielsen India Pvt. Ltd.

		Manufacturing	Construction	Retail/Wholesale	Other Services	TOTAL
Kabul	Small	296	2123	54	520	2993
	Medium	71	551	2	140	764
	Large	14	141	0	39	194
	Total	381	2815	56	699	3951
Herat	Small	183	106	49	39	377
	Medium	85	40	1	11	137
	Large	8	9	0	1	18
	Total	276	155	50	51	532
Kandahar	Small	48	385	50	61	544
	Medium	32	96	1	12	141
	Large	5	10	0	2	17
	Total	85	491	51	75	702
Balkh (Mazar-e-Sharif)	Small	111	192	50	24	377
	Medium	31	81	0	8	120
	Large	4	10	0	4	18
	Total	146	283	50	36	515
Nangarhar (Jalalabad)	Small	202	345	44	14	605
	Medium	104	85	5	9	203
	Large	8	9	0	0	17
	Total	314	439	49	23	825
GRAND TOTAL		1202	4183	256	884	6525

10. The sample frame, consisting of mostly AISA-register businesses, was used for the selection of a sample with the aim of achieving 360 interviews with establishments of five or more employees.

11. The quality of the frame was assessed at the onset of the project through visits to a random subset of firms and local contractor knowledge. The sample frame was not immune from the typical problems found in establishment surveys: positive rates of non-eligibility, repetition, non-existent units, etc.

12. Given the impact that non-eligible units included in the sample universe may have on the results, adjustments may be needed when computing the appropriate weights for individual observations. Breaking down by stratified industries, the following sample targets were achieved (using a4a and a6a):

Realized Sample

		Manufacturing	Construction	Retail/Wholesale	Other Services	TOTAL
Kabul	Small	27	9	23	26	85
	Medium	7	10	1	5	23
	Large	2	3	0	2	7
	Total	36	22	24	33	115
Herat	Small	14	1	25	5	45
	Medium	6	5	0	1	12
	Large	1	1	0	0	2
	Total	21	7	25	6	59
Kandahar	Small	16	5	18	2	41
	Medium	7	9	1	1	18
	Large	0	0	0	0	0
	Total	23	14	19	3	59
Balkh (Mazar-e-Sharif)	Small	13	4	20	5	42
	Medium	6	9	0	3	18
	Large	1	0	0	1	2
	Total	20	13	20	9	62
Nangarhar (Jalalabad)	Small	11	6	27	1	45
	Medium	5	3	2	3	13
	Large	3	0	0	0	3
	Total	19	9	29	4	61
GRAND TOTAL		119	65	117	55	356

IV. Data Base Structure:

13. The structure of the data base reflects the fact that 2 different versions of the questionnaire were used for 3 categories of businesses (manufacturing, retail, and other services/non-retail). The Manufacturing Questionnaire includes all common questions asked to all establishments and some specific questions relevant to manufacturing firms. The Services Questionnaire, administered to retail and other services/non-retail establishments, includes all common questions asked to all establishments and some specific questions relevant retail and other services firms. Each variation of the questionnaire is identified by the index variable, *a0*.

14. All variables are named using, first, the letter of each section and, second, the number of the variable within the section, i.e. *a1* denotes section *A*, question *1*. Variable names proceeded by a prefix *"SAR"* or *"AFG"* indicate questions specific to the South Asia region or Afghanistan only, therefore, they may not be found in the implementation of the rollout in other countries. All other suffixed variables are global and are present in all country surveys over the world. All variables are numeric with the exception of those variables with an "x" at the end of their names. The suffix "x" denotes that the variable is alpha-numeric.

15. There are 2 establishment identifiers, *idstd* and *id*. The first is a global unique identifier. The second is a country unique identifier. The variables *a2* (sampling region), *a6a* (sampling establishment's size), and *a4a* (sampling sector) contain the

establishment's classification into the strata chosen for each country using information from the sample frame. The strata were defined according to the guidelines described above.

16. There are three levels of stratification: industry, size and region. Different combinations of these variables generate the strata cells for each industry/region/size combination. A distinction should be made between the variable a4a and d1a2 (industry expressed as ISIC rev. 3.1 code). The former gives the establishment's classification into one of the chosen industry-strata, whereas the latter gives the actual establishment's industry classification (four digit code) in the sample frame.

17. All of the following variables contain information from the sampling frame. They may not coincide with the reality of individual establishments as sample frames may contain inaccurate information. The variables containing the sample frame information are included in the data set for researchers who may want to further investigate statistical features of the survey and the effect of the survey design on their results.

-*a2* is the variable describing sampling regions

-*a6a*: coded using the same standard for small, medium, and large establishments as defined above. The code -*9* was used to indicate units for which size was undetermined in the sample frame.

-*a4a*: coded using ISIC Rev 3.1 codes for the chosen industries for stratification. These codes include most manufacturing industries (15 to 37), retail (52), and (45, 50, 51, 55, 60-64, 72) for other services.

18. The surveys were implemented following a 2 stage procedure. Typically first a screener questionnaire is applied over the phone to determine eligibility and to make appointments. Then a face-to-face interview takes place with the Manager/Owner/Director of each establishment. The variables *a4b* and *a6b* contain the industry and size of the establishment from the screener questionnaire. Variables *a8* to *a11* contain additional information and were also collected in the screening phase.

19. Note that there are additional variables for location (*a3x*) and size (*l1*, *l6* and *l8*) that reflect more accurately the reality of each establishment. Advanced users are advised to use these variables for analytical purposes.

20. Variable *a3x* indicates the actual location of the establishment. There may be divergences between the location in the sampling frame and the actual location, as establishments may be listed in one place but the actual physical location is in another place.

21. Variables *l1*, *l6* and *l8* were designed to obtain a more accurate measure of employment accounting for permanent and temporary employment. Special efforts were made to make sure that this information was not missing for most establishments.

22. Variables *a17x* gives interviewer comments, including problems that occurred during an interview and extraordinary circumstances which could influence results. Please note that sometimes this variable is removed due to privacy issues.

V. Universe Estimates

23. Appendix A shows the overall estimates of the numbers of establishments in Afghanistan based on data from the Central Statistics Organization Afghanistan (CSO).

24. For some establishments where contact was not successfully completed during the screening process (because the firm has moved and it is not possible to locate the new location, for example), it is not possible to directly determine eligibility. Thus, different assumptions about the eligibility of establishments result in different adjustments to the universe cells and thus different sampling weights.

25. Universe estimates for the number of establishments in each industry-region-size cell in Afghanistan were created based on the findings from the Integrated Business Enterprise Survey 2009 conducted by the CSO (Tables 2.2 and 2.10). In addition to this 2009 report, population totals from the CSO website were used to apportion the number of establishments across the 5 subnational locations.

26. Once an estimate of the universe cell projection was made, base weights were computed by dividing the universe number over the achieved number of interviews in each cell.

VI. Weights

27. Since the sampling design was stratified and employed differential sampling, individual observations should be properly weighted when making inferences about the population. Under stratified random sampling, unweighted estimates are biased unless sample sizes are proportional to the size of each stratum. With stratification the probability of selection of each unit is, in general, not the same. Consequently, individual observations must be weighted by the inverse of their probability of selection (probability weights or *pw* in Stata).[4]

28. Appendix B shows the cell weights for registered establishments in Afghanistan.

VII. Appropriate use of the weights

29. Under stratified random sampling weights should be used when making inferences about the population. Any estimate or indicator that aims at describing some feature of the population should take into account that individual observations may not represent equal shares of the population.

30. However, there is some discussion as to the use of weights in regressions (see Deaton, 1997, pp.67; Lohr, 1999, chapter 11, Cochran, 1953, pp.150). There is not a strong large sample econometric argument in favor of using weighted estimation for a common population coefficient if the underlying model varies per stratum (stratum-

4 This is equivalent to the weighted average of the estimates for each stratum, with weights equal to the population shares of each stratum.

specific coefficient): both simple OLS and weighted OLS are inconsistent under regular conditions. However, weighted OLS has the advantage of providing an estimate that is independent of the sample design. This latter point may be quite relevant for the Enterprise Surveys as in most cases the objective is not only to obtain model-unbiased estimates but also design-unbiased estimates (see also Cochran, 1977, pp 200 who favors the used of weighted OLS for a common population coefficient.)[5]

31. From a more general approach, if the regressions are descriptive of the population then weights should be used. The estimated model can be thought of Banglas the relationship that would be expected if the whole population were observed.[6] If the models are developed as structural relationships or behavioral models that may vary for different parts of the population, then, there is no reason to use weights.

VIII. Non-response

32. Survey non-response must be differentiated from item non-response. The former refers to refusals to participate in the survey altogether whereas the latter refers to the refusals to answer some specific questions. Enterprise Surveys suffer from both problems and different strategies were used to address these issues.

33. Item non-response was addressed by two strategies:

a- For sensitive questions that may generate negative reactions from the respondent, such as corruption or tax evasion, enumerators were instructed to collect the refusal to respond as a different option from don't know (-8).

b- Establishments with incomplete information were re-contacted in order to complete this information, whenever necessary. However, there were clear cases of low response. The following shows non-response rates for the sales variable, *d2*, by sector: 63% for manufacturing, 80% for retail, and 77% for other services. Please, note that the coding utilized in this dataset does not allow us to differentiate between "Don't know" and "refuse to answer", thus the non-response rates reflect both categories (DKs and NAs).

34. Survey non-response was addressed by maximizing efforts to contact establishments that were initially selected for interview. Attempts were made to contact the establishment for interview at different times/days of the week before a replacement establishment (with similar strata characteristics) was suggested for interview. Survey non-response did occur but substitutions were made in order to potentially achieve strata-specific goals. Further research is needed on survey non-response in the Enterprise Surveys regarding potential introduction of bias.

5 Note that weighted OLS in Stata using the command regress with the option of weights will estimate wrong standard errors. Using the Stata survey specific commands *svy* will provide appropriate standard errors.

6 The use of weights in most model-assisted estimations using survey data is strongly recommended by the statisticians specialized on survey methodology of the JPSM of the University of Michigan and the University of Maryland.

35. Details on the rejection rate, eligibility rate, and item non-response are available at the strata level. This report summarizes these numbers to alert researchers of these issues when using the data and when making inferences. Item non-response, selection bias, and faulty sampling frames are not unique to Afghanistan. All Enterprise Surveys suffer from these shortcomings, but in very few cases they have been made explicit.

36. Appendix D provides observations and experiences from the implementing contractor during the survey fieldwork in Afghanistan.

References:

Cochran, William G., Sampling Techniques, 1977.

Deaton, Angus, The Analysis of Household Surveys, 1998.

Levy, Paul S. and Stanley Lemeshow, Sampling of Populations: Methods and Applications, 1999.

Lohr, Sharon L. Samping: Design and Techniques, 1999.

Scheaffer, Richard L.; Mendenhall, W.; Lyman, R., Elementary Survey Sampling, Fifth Edition, 1996.

Appendix A

Universe Estimates, Afghanistan:

		Manufacturing	Construction	Retail/Wholesale	Other Services	TOTAL
Kabul	Small	1326	221	1372	830	3748
	Medium	22	27	2	16	67
	Large	16	3	0	2	20
	Total	1364	250	1374	848	3836
Herat	Small	623	104	644	390	1761
	Medium	10	13	0	8	31
	Large	7	1	0	0	9
	Total	641	118	644	398	1800
Kandahar	Small	401	67	415	251	1135
	Medium	7	9	1	5	21
	Large	0	0	0	0	0
	Total	408	76	416	256	1156
Balkh (Mazar-e-Sharif)	Small	434	72	449	272	1228
	Medium	7	9	0	5	21
	Large	5	0	0	1	6
	Total	447	81	449	278	1255
Nangarhar (Jalalabad)	Small	505	84	522	316	1426
	Medium	8	10	2	6	27
	Large	6	0	0	0	6
	Total	519	94	524	322	1459
GRAND TOTAL		3378	619	3408	2102	9507

Appendix B

Median Cell Weights Afghanistan

		Manufacturing	Construction	Retail/Wholesale	Other Services
Kabul	Small	49.11	24.56	59.65	31.92
	Medium	3.14	2.70	2.00	3.20
	Large	8.00	1.00	0.00	1.00
Herat	Small	44.50	104.00	25.76	78.00
	Medium	1.67	2.60	0.00	8.00
	Large	7.00	1.00	0.00	0.00
Kandahar	Small	25.06	13.40	23.06	125.50
	Medium	1.00	1.00	1.00	5.00
	Large	0.00	0.00	0.00	0.00
Balkh (Mazar-e-Sharif)	Small	33.38	18.00	22.45	54.40
	Medium	1.17	1.00	0.00	1.67
	Large	5.00	0.00	0.00	1.00
Nangarhar (Jalalabad)	Small	45.91	14.00	19.33	316.00
	Medium	1.60	3.33	1.00	2.00
	Large	2.00	0.00	0.00	0.00

Appendix C

Original Sample Design, Afghanistan:

		Manufacturing	Construction	Retail/Wholesale	Other Services	TOTAL
Kabul	Small	25	8	25	25	
	Medium	8	11	2	5	
	Large	6	3	0	2	
	Total	39	22	27	32	120
Herat	Small	15	1	25	5	
	Medium	3	6	1	1	
	Large	2	1	0	0	
	Total	20	8	26	6	60
Kandahar	Small	10	1	20	3	
	Medium	6	8	1	4	
	Large	5	1	0	1	
	Total	21	10	21	8	60
Balkh (Mazar-e-Sharif)	Small	11	1	22	3	
	Medium	5	9	0	3	
	Large	4	1	0	1	
	Total	20	11	22	7	60
Nangarhar (Jalalabad)	Small	13	1	23	4	
	Medium	4	7	1	3	
	Large	3	1	0	0	
	Total	20	9	24	7	60
GRAND TOTAL		120	60	120	60	360

Appendix D

Challenges and Difficulties during Survey Fieldwork (Notes from the Project Supervisor based in Kabul)

1: A number of addresses of the firms were really difficult to find, we had to make a number of phone calls to get the correct address.

2: Some of the respondents were suspicious of the survey, despite our efforts of sharing the country profile and outcomes of previous ES studies; however they were still reluctant and were not willing to be interviewed.

3: There were occasions while making phone calls to different firms and confirming their participation in the survey they were abusive and their language was very offensive.

4: Some of the times after our interviewers would go for an appointment to conduct an interview, the length and volume of questionnaire would scare the respondent and they would make excuses to suspend the interview and get rid of us.

5: A number of firms were suspicious of the interview and during the interviews they would skip, ignore or respond IDK (I don't know) to most of the important and essential questions.

6: As the overall pessimism in the country a good number of the firms were skeptical of the survey and they firmly believed that this survey was a waste of money and time. According to them often the outcomes of such surveys are not clear or in some cases there is not result of such studies at all.

7: Some of the companies that our interviewers were interviewing, during different phases they had expectations from us, financial and others. They often asked that the World Bank should support them and give them projects to be implemented through them.

8: Erroneous of contacts and addresses was one major problem, which contributed handsomely to the length of the project as well.

9: As a matter of fact a large number of our respondents in different companies were people with no literacy, and the requirement of the survey being of a very statistical nature made this even troublesome, as in some cases the respondents did not have the right number and figure and they were giving estimates.

10: Lack of commitment from respondents sticking to their promise to be available on the date and time of interview. This was one of the big challenges we came across in every province and with most of the respondents. On average we had to make three phone calls to make sure the respondent was firmed and available for the interview.

11: In some of the cases the respondents did not even talk to us, because they thought we were from income tax entity or from the ministry of economy to check rules and regulations.

12: Towards the end of the interview the interviewees were getting bored and tired of so many questions.