The Central African Republic 2011 Enterprise Surveys Data Set

I. Introduction

1. This document provides additional information on the data collected in Central African Republic between June 2011 and July 2011 as part of the Africa Enterprise Survey 2011, an initiative of the World Bank.

As part of its strategic goal of building a climate for investment, job creation, and sustainable growth, the World Bank has promoted improving business environments as a key strategy for development, which has led to a systematic effort in collecting enterprise data across countries. The Enterprise Surveys (ES) are an ongoing World Bank project in collecting both objective data based on firms' experiences and enterprises' perception of the environment in which they operate.

The Enterprise Surveys currently cover over 130,000 firms in 125 countries, of which 113 have been surveyed following the standard methodology. This allows for better comparisons across countries and across time. Data are used to create statistically significant business environment indicators that are comparable across countries. The Enterprise Surveys are also used to build a panel of enterprise data that will make it possible to track changes in the business environment over time and allow, for example, impact assessments of reforms.

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 Central African Republic 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 under study), and all public or utilities-sectors.

c. To make sure 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.

---

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

[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

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.

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 E.

4. Industry stratification was designed in the way that follows: the universe was stratified into one manufacturing industry, ne service industry -retail -, and one residual sector as defined in the sampling manual. The manufacturing industry, service industry, and residual sectors had a target each of 120 interviews.

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. This seems to be an appropriate definition of the labor force since seasonal/casual/part-time employment is not a common practice, except in the sectors of construction and agriculture.

6. Regional stratification was defined in two regions (city and the surrounding business area): Bangui and Berberati.

III. Sampling implementation

7. Given the stratified design, sample frames containing a complete and updated list of establishments as well as information on all stratification variables (number of employees, industry, and region) are required to draw the sample. Great efforts were made to obtain the best source for these listings. However, the quality of the sample frames was not optimal and, therefore, some adjustments were needed to correct for the presence of ineligible units. These adjustments are reflected in the weights computation (*see below*).

8. TNS Opinion was hired to implement the Africa 2011 enterprise surveys roll out. In Central African Republic the local subcontractor was HFC Research Associates.

9. The sample frame used for the survey in CAR was Ministry of Commerce in CAR. A copy of that frame was sent to the TNS statistical team in London to select the establishments for interview. Each database contained the following information

- Coverage;
- Up to datedness;

- Availability of detailed stratification variables;
- Location identifiers- address, phone number, email;
- Electronic format availability;
- Contact name(s).

Counts from sample frames are shown below.

Sample Frames

Source: Ministry of Commerce in CAR

Location	Size	Manufacturing	Services	Grand Total
Bangui	Small 5 to 19)	29	190	219
	Medium (20 to 99)	12	33	45
	Large (100+)	3	2	5
Bangui Total		**44**	**225**	**269**
Berberati	Small (5 to 19)	1	11	12
	Medium (20 to 99)		1	1
	Large (100+)			
Berberati Total		**1**	**12**	**13**
Grand Total		**45**	**237**	**282**

10. The enumerated establishments were then used as the frame for the selection of a sample with the aim of obtaining interviews at 150 establishments with 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. In addition, the sample frame contains no telephone/fax numbers so the local contractor had to screen the contacts by visiting them. Due to response rate and ineligibility issues, additional sample had to be extracted by the World Bank in order to obtain enough eligible contacts and meet the sample targets.

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. The percentage of confirmed non-eligible units as a proportion of the total sample issued and contacted for the survey was 23.83% (46 out of 203 establishments)[4].

[4] Based on out of target contacts and impossible to contact establishments

IV. Data Base Structure:

13. The structure of the data base reflects the fact that 2 different versions of the survey instrument were used, i.e. manufacturing and the services questionnaire. Both questionnaires have common questions and respectfully additional manufacturing and services specific questions. Each variation of the questionnaire is identified by the index variable, *a0*.

14. All variables are typically 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*. (Some exceptions apply due to comparability reasons).Variable names proceeded by a prefix *"AF"* indicate questions specific to Africa, 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. In the implementation of the Africa roll 2011 out an experiment was carried in some of the countries to better estimate the effects of the use of show cards in data collection. In some of the sections i.e. innovation the enumerators were trained to alternatively implement the section using either show cards or asking only the questions without showing any cards, please see the variable "ballot".

15. The data has two unique firm 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 codes for the chosen industries for stratification. These codes include most manufacturing industries (15 to 37), other manufacturing (2), retail (52), and (45, 50, 51, 55, 60, 63, 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. However, the phone numbers were unavailable in the sample frame, and thus the enumerators applied the screeners in person. 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. Universe estimates for the number of establishments in each cell in Central African Republic were produced for the strict, weak and median eligibility definitions. The estimates were the multiple of the relative eligible proportions.

24. Appendix B shows the overall estimates of the numbers of establishments in Central African Republic based on the sample frame.

25. 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.

26. Three sets of assumptions on establishment eligibility are used to construct sample adjustments using the status code information.

27. Strict assumption: eligible establishments are only those for which it was possible to directly determine eligibility. The resulting weights are included in the variable *wstrict*.

Strict eligibility = (Sum of the firms with codes 1,2,3,4,&16) / Total

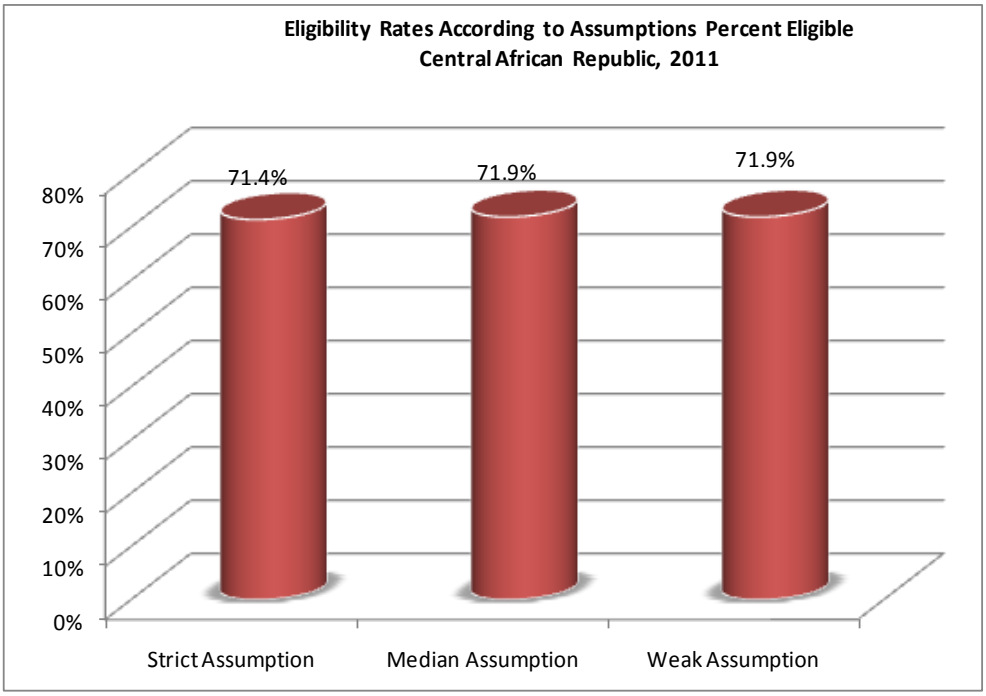
28. Median assumption: eligible establishments are those for which it was possible to directly determine eligibility and those that rejected the screener questionnaire or an answering machine or fax was the only response. The resulting weights are included in the variable *wmedian*.

Median eligibility = (Sum of the firms with codes 1,2,3,4,16,10,11, & 13) / Total

29. Weak assumption: in addition to the establishments included in points a and b, all establishments for which it was not possible to contact or that refused the screening questionnaire are assumed eligible. This definition includes as eligible establishments with dead or out of service phone lines, establishments that never answered the phone, and establishments with incorrect addresses for which it was impossible to find a new address. Under the weak assumption only observed non-eligible units are excluded from universe projections. The resulting weights are included in the variable *wweak*.

Weak eligibility= (Sum of the firms with codes 1,2,3,4,16,91,92,93,10,11,12,&13) / Total

30. The indicators computed for the Enterprise Survey website use the median weights. The following graph shows the different eligibility rates calculated for firms in the sample frame under each set of assumptions.



31. Universe estimates for the number of establishments in each industry-region-size cell in Central African Republic were produced for the strict, weak and median eligibility definitions. Appendix D shows the universe estimates of the numbers of registered establishments that fit the criteria of the Enterprise Surveys.

32. Once an accurate estimate of the universe cell projection was made, weights for the probability of selection were computed using the number of completed interviews for each cell.

VI. Weights

33. 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).[5]

34. Special care was given to the correct computation of the weights. It was imperative to accurately adjust the totals within each region/industry/size stratum to account for the presence of ineligible units (the firm discontinued businesses or was unattainable, education or government establishments, establishments with less than 5 employees, no reply after having called in different days of the week and in different business hours, no tone in the phone line, answering machine, fax line[6], wrong address or moved away and could not get the new references) The information required for the adjustment was collected in the first stage of the implementation: the screening process. Using this information, each stratum cell of the universe was scaled down by the observed proportion of ineligible units within the cell. Once an accurate estimate of the universe cell (projections) was available, weights were computed using the number of completed interviews.

35. Appendix C shows the cell weights for registered establishments in Central African Republic.

VII. Appropriate use of the weights

36. 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.

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

[6] For the surveys that implemented a screener over the phone.

37. 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 strong large sample econometric argument in favor of using weighted estimation for a common population coefficient if the underlying model varies per stratum (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.)[7]

38. 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 as the relationship that would be expected if the whole population were observed.[8] 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

39. 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.

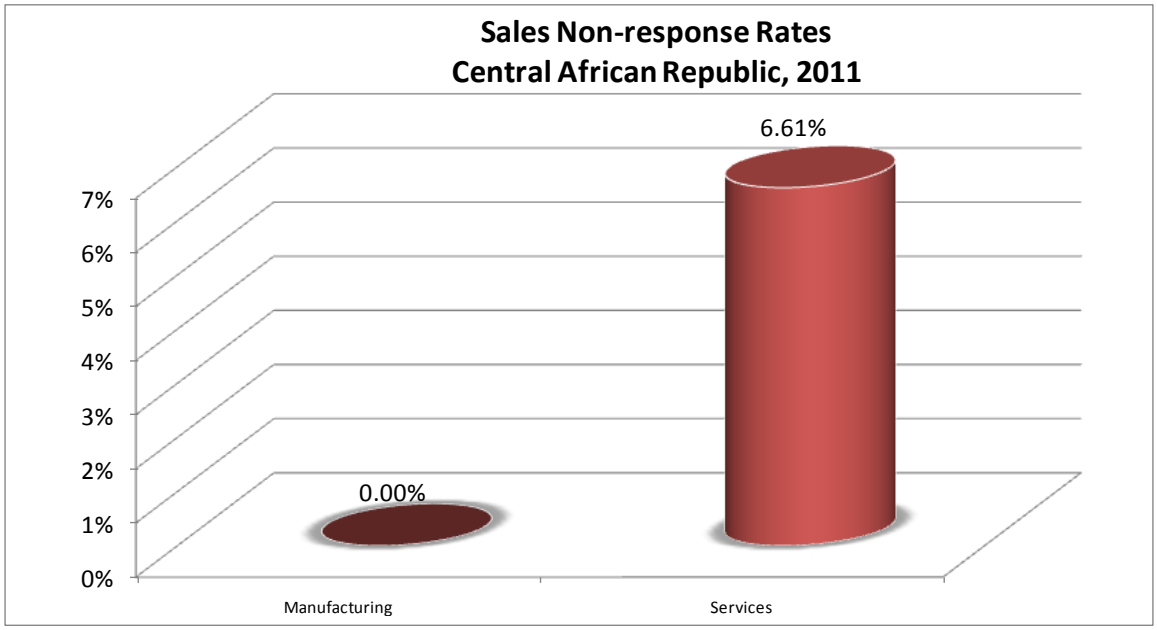
40. Item non-response was addressed by two strategies:

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 (-7).

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 graph shows non-response rates for the sales variable, *d2*, by sector. 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 in the chart below reflects both categories (DKs and NAs).

[7] 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.

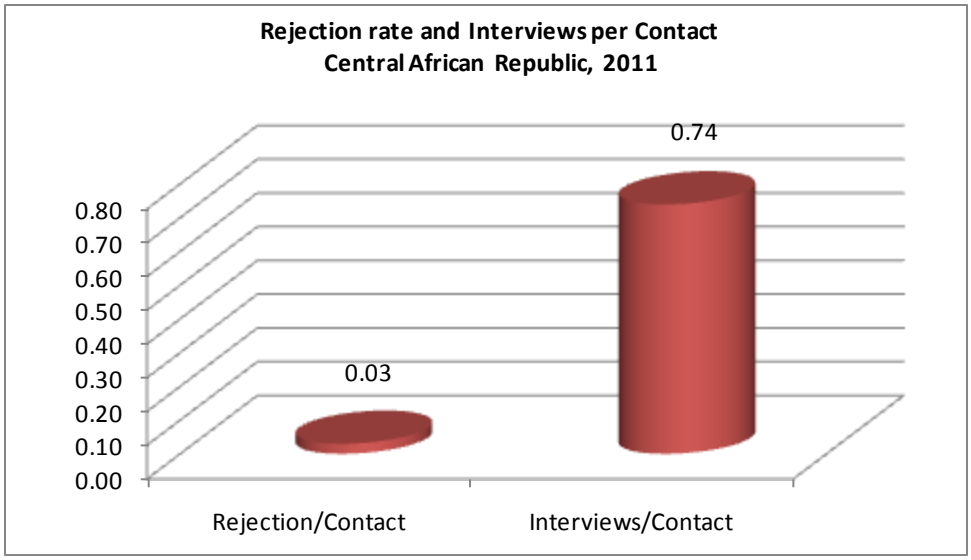
[8] The use weights in most model-assisted estimations using survey data is strongly recommended by the statisticians specialized on survey methodology of the JPSM of the University of Michigan and the University of Maryland.



41. 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.

42. As the following graph shows, the number of contacted establishments per realized interview was 0.85[9]. This number is the result of two factors: explicit refusals to participate in the survey, as reflected by the rate of rejection (which includes rejections of the screener and the main survey) and the quality of the sample frame, as represented by the presence of ineligible units. The number of rejections per contact was 0.03.

[9] The estimate is based on the total no. of firms contacted including ineligible establishments.



43. Details on the rejection rate, eligibility rate, and item non-response are available at the level strata. 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 Central African Republic. All enterprise surveys suffer from these shortcomings, but in very few cases they have been made explicit.

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

Status Codes Fresh:

	ELIGIBLES	
Eligible	1.Elegible establishment (Correct name and address)	115
	2. Elegible establishment (Different name but same address - the new firm/establishment bought the original firm/establishment)	0
	3. Elegible establishment (Different name but same address - the firm/establishment changed its name)	30
	4. Elegible establishment (Wrong address - the firm/establishmen has changed address and the address could be found)	0
	16. Panel firm - now less than five employees	0
Ineligible	5. The establishment has less than 5 permanent full time employees	0
	6. The firm discontinued businesses	9
	7. Not a business: private household	0
	8. Ineligible activity: education, agriculture, finances, governments…	5
	151. Out of target - outside the covered regions, firm moved abroad	1
	152. Out of target - firm moved abroad	0
	153. Impossible to find	31
Unobtainable	91. No reply *(after having called in different days of the week and in different business hours)*	0
	92. Line out of order	0
	93. No tone	0
	94. Phone number does not exist	0
	10. Answering machine	0
	11. Fax line - data line	0
	12. Wrong address/ moved away and could not get the new references	0
	13. Refuses to answer the screener	1
	14. In process *(the establishment is being called/ is being contacted - previous to ask the screener)*	**1**
	Total	203

Response Outcomes Fresh:

Complete interviews *(Total)*	150
Incomplete interviews	0
Eligible in process	0
Refusals	5
Out of target	14
Impossible to contact	0
Ineligible - coop.	32
Refusal to the Screener	1
Total	202

Appendix B

Universe Estimates, Central African Republic:

Source: Ministry of Commerce

Location	Size	MANUFACTURING	SERVICE	Grand Total
Bangui	5 to19	29	190	219
	20 to 99	12	33	45
	>100	3	2	5
	(unknown size)	1	1	2
Bangui Total		**45**	**226**	**271**
Berberati	5 to19	1	11	12
	20 to 99		1	1
	>100			
Berberati Total		**1**	**12**	**13**
Grand Total		**46**	**238**	**284**

Appendix C

Strict Cell Weights Central African Republic:

Location	Size	Manufacturing	Services
Bangui	5 to 19	1.02	1.62
	20 to 99	1.08	1.41
	100+	1.00	1.98
Berberati	5 to 19		1.00
	20 to 99		
	100+		

Median Cell Weights Central African Republic:

Location	Size	Manufacturing	Services
Bangui	5 to 19	1.02	1.64
	20 to 99	1.07	1.41
	100+	1.00	1.99
Berberati	5 to 19		1.00
	20 to 99		
	100+		

Weak Cell Weights Central African Republic:

Location	Size	Manufacturing	Services
Bangui	5 to 19	1.02	1.64
	20 to 99	1.07	1.41
	100+	1.00	1.99
Berberati	5 to 19		1.00
	20 to 99		
	100+		

Appendix D

Strict Universe Estimates

Location	Size	Manufacturing	Services	Grand Total
Bangui	5 to 19	22	139	162
	20 to 99	11	28	39
	100+	3	2	5
Bangui Total		36	169	205
Berberati	5 to 19		8	8
	20 to 99			
	100+			
BerberatiTotal			8	8
Grand Total		36	177	213

Weak Universe Estimates

Location	Size	Manufacturing	Services	Grand Total
Bangui	5 to 19	22	141	163
	20 to 99	11	28	39
	100+	3	2	5
Bangui Total		36	171	207
Berberati	5 to 19		8	8
	20 to 99			
	100+			
BerberatiTotal			8	8
Grand Total		36	179	215

Median Universe Estimates

Location	Size	Manufacturing	Services	Grand Total
Bangui	5 to 19	22	141	163
	20 to 99	11	28	39
	100+	3	2	5
Bangui Total		36	171	207
Berberati	5 to 19		8	8
	20 to 99			
	100+			
Berberati Total			8	8
Grand Total		36	179	215

Appendix E

Original Sample Design, Central African Republic:

Location	Size	Manufacturing	Services	Grand Total
Bangui	5 to 19	29	76	105
	20 to 99	12	20	32
	100+	3	2	5
Bangui Total		44	98	142
Berberati	5 to 19	1	6	7
	20 to 99		1	1
	100+			
BerberatiTotal		1	7	8
Grand Total		45	105	150

Completed Interviews, Central African Republic:

Location	Size	Manufacturing	Services	Grand Total
Bangui	5 to 19	22	86	108
	20 to 99	10	20	30
	100+	3	1	4
Bangui Total		35	107	142
Berberati	5 to 19		8	8
	20 to 99			
	100+			
BerberatiTotal			8	8
Grand Total		35	115	150

Appendix F

Local Agency team involved in the study:

Local Agency	Name: Consumer Opinion Country: Cameroon Activities since: 2009
Enumerators involved:	Enumerators: 18 Recruiters: 3
Other staff involved:	Fieldwork Coordinators: 3

Sample Frame:

Characteristic of sample frame used:	Fresh sample Registered companies operating in CAR
Source:	Ministry of Commerce (corrected using Yellow pages)
Year:	2011
Comments on the quality of sample frame:	The sampling frame had some flaws, particularly regarding the address information and/or activities (products) of the elements.

Sectors included in the Sample:

Original Sectors	The manufacturing sector comprises all manufacturing establishments as mentioned in group D. The service sector includes Group F (construction), Groups G without 52 and H (services), Group I (transport, storage, and communications) and subsector 72 from Group K

Fieldwork:

Date of Fieldwork	9th of June 2011 to 13th of July 2011
Country	Central African Republic
Problems found during fieldwork:	Enumerators access to Berberati was difficult due to poor road conditions. Recruitment of large firms was relatively more difficult compared to small and medium ones.
Other observations:	The sample composition includes mostly services firms as there are relatively few manufacturing establishments in CAR.

Questionnaires:

Problems for the understanding of questions (write question number)	Some respondents had difficulties understanding one of the following questions: C3, C12 , C19, K7 and AFJ8.

Appendix G: Subcontractors

	Country	Local team	Survey coordinator name	Since	# F2F int.	Experience F2F int. (years)
Kenya local coordination team	Kenya	TNS Research International	Liston Njoroge	1972	360	10
Indicator Survey	Central African Republic	University of Bangui-Head of geosciences.	Dr Moloto Gaetan Roch	2008	50	2
Enterprise Survey	Angola	HFC	Henrique Freitas	2006	25	3
	Botswana	PROBE	Florence J A Onyango	2001	300	10
	Democratic Republic of Congo	University of Kinshasa	Prof Mukoko Samba	2000	25	5
	Mali	PSI	Dominic Kpanja	2005	45	3