The South Sudan 2014 Enterprise Surveys Data Set

I. Introduction

1. This document provides additional information on the data collected in South Sudan between July 2014 and December 2014 under, 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 135 countries, of which 121 have been surveyed following a 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 South Sudan 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.

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/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

3. Two levels of stratification were used in this country: industry, and region as size was not available in the sampling frame for most contacts. 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 manufacturing industries and two service sectors (retail and other services).

6. Regional stratification for the South Sudan ES was defined in four regions:

- Juba
- Nimule
- Torit
- Yei

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.

8. The international firm of Ipsos was hired to conduct the survey and they partnered with local agency Tango Consult in South Sudan.

9. For the South Sudan ES, a sample frame was built using data from the National Bureau of Statistics as well are municipal commercial registries.

South Sudan, Sample Frame

	Manufacturing	Retail	Other Services
Juba	66	1022	670
Nimule	1	59	105
Torit	2	6	41
Yei	7	223	47

10. The sample design for the South Sudan Enterprise Survey was generated with the aim of obtaining interviews at 720 establishments. Establishments with undefined size were included as part of this sample frame for South Sudan in order to ensure a representative sample. Size information collected during the survey process can then be used to categorize these firms.

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. The local contractor had to screen the contacts by visiting them which resulted in slow fieldwork in many cases.

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:

Achieved sample

	Manufacturing	Retail	Other Services
Juba	78	160	162
Nimule	2	43	58
Torit	1	5	37
Yei	6	150	36

IV. Data Base Structure:

13. The structure of the data base reflects the fact that 2 different versions of the survey instrument were used for all registered establishments. Questionnaires have common questions and respectfully additional manufacturing and services specific questions. The eligible manufacturing industries have been surveyed using the ***Manufacturing*** questionnaire (includes a common set of core variables, plus manufacturing specific questions). Eligible services have been covered using the ***Services*** questionnaire. 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* (some exceptions apply due to comparability reasons). Variable names proceeded by a prefix *"SL"* indicate questions specific to South Sudan, 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 two levels of stratification: industry 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 micro, 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. In the case of South Sudan, this screener was administered face-to-face. 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 variables for 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. 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.

20. 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.

21. Note that the fiscal years vary by firm as there is no standard for all firms in South Sudan. The start and end dates for the fiscal year for each firm can be found in the a20 variables in the dataset

V. Universe Estimates

21. Universe estimates for the number of establishments in each cell in South Sudan were produced for the strict, weak and median eligibility definitions. The estimates were the multiple of the relative eligible proportions.

23. 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.

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

25. 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

26. 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*.

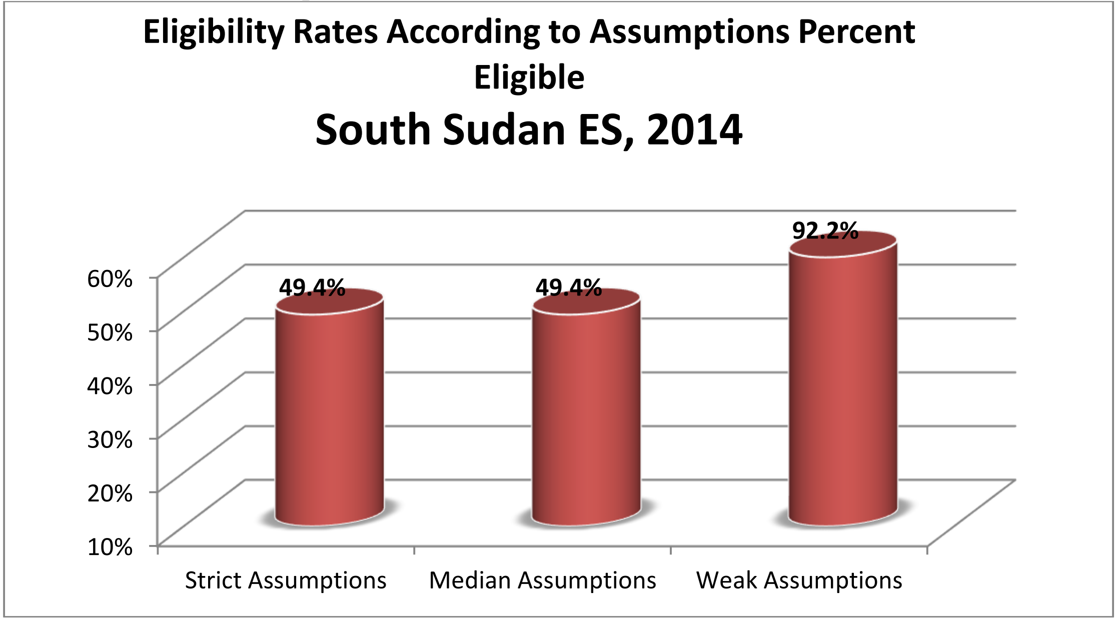
Median eligibility = (Sum of the firms with codes 1,2,3,4,16,10,11, & 13) / Total

27. 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

28. 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.



29. Universe estimates for the number of establishments in each industry-region-size cell in South Sudan 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.

30. 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

31. 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]

32. 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

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

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[5], 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.

VII. Appropriate use of the weights

33. 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.

34. 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.)[6]

35. 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.[7] 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

36. 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.

37. 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).

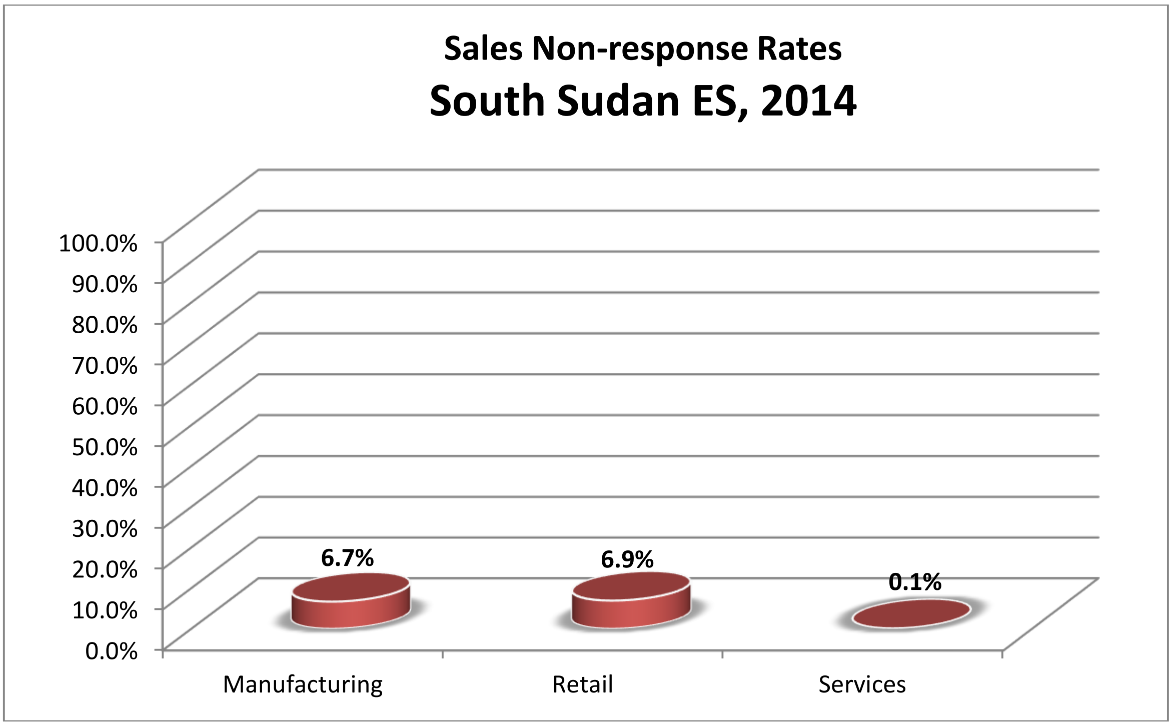
b- Establishments with incomplete information were re-contacted in order to complete this information, whenever necessary. However, there were clear cases of low

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

[6] 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.

[7] 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.

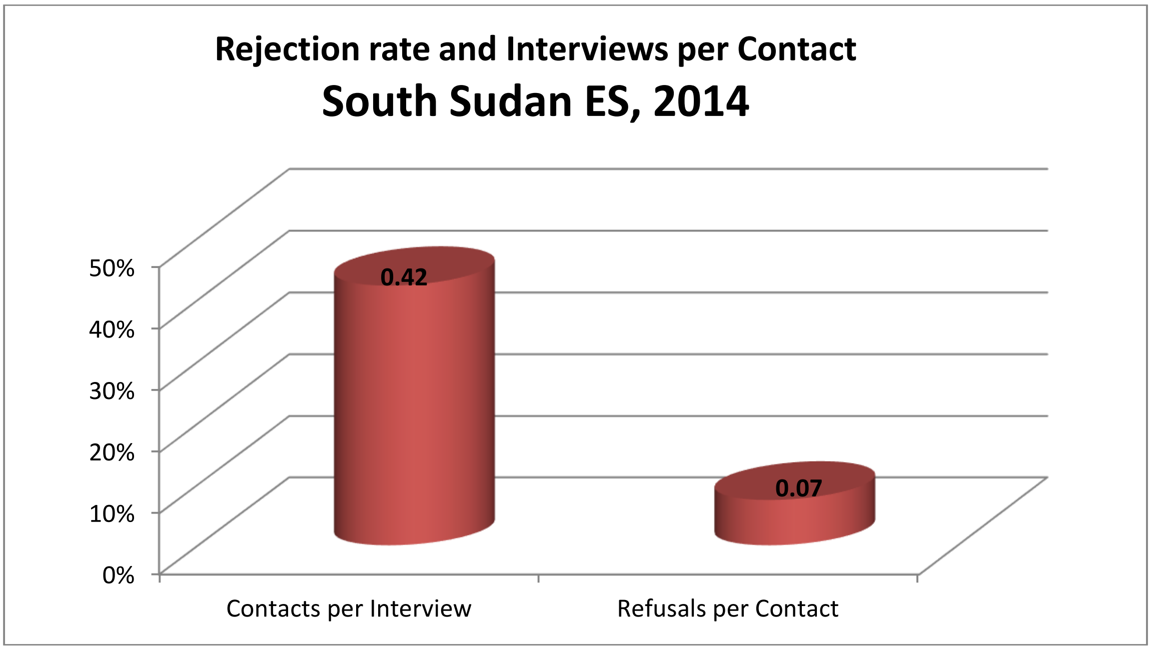
response. The following graph shows non-response rates for the sales variable, *d2,* by sector. Please, note that the coding utilized in this dataset does not allow us to differentiate between “Don’t know” and “refuse to answer”, thus the non-response in the chart below for both enterprise surveys (ES) reflect both categories (DKs and NAs).



38. 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.

39. As the following graph shows, the number of interviews per contacted establishments was 0.42[8]. 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 0.07.

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



40. 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 South Sudan. 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:

	South Sudan
Sample Target	**720**
Complete interviews **(Total)**	738
Incomplete interviews	1
Eligible in process	0
Refusals	130
Out of target	132
Impossible to contact	663
Ineligible - coop.	7
Refusal to the Screener	0
Total	**1672**

Response rate	**85%**
Out of target + impossible to contact	**48%**
Impossible to contact	**40%**

Eligibles	1.Elegible establishment (Correct name and address)	846
	2. Eligible establishment (Different name but same address - the new firm/establishment bought the original firm/establishment)	1
	3. Eligible establishment (Different name but same address - the firm/establishment changed its name)	25
	4. Eligible establishment (Wrong address - the firm/establishment has changed address and the address could be found)	12
Ineligibles	5. The establishment has less than 5 permanent full time employees	69
	6. The firm discontinued businesses	36
	7. Not a business: private household	10
	8. Ineligible activity: education, agriculture, finances, governments…	17
Unobtainable	91. No reply *(after having called in different days of the week and in different business hours)*	19
	92. Line out of order	197
	93. No tone	3
	94. Phone number does not exist	131
	10. Answering machine	0
	11. Fax line - data line	0
	12. Wrong address/ moved away and could not get the new references	313
	13. Refuses to answer the screener	0
	14. In process *(the establishment is being called/ is being contacted - previous to ask the screener)*	**7**
	151. Out of target - outside the covered regions, firm moved abroad	7
	152. Out of target - firm moved abroad	0
	153. Out of target - Not registered with SAT	0
	Total	**1693**

Appendix C Weights

Strict Weights

	Manufacturing	Retail	Other Services
Juba	1.00	2.69	1.97
Nimule	1.00	1.00	1.16
Torit	2.45	1.00	1.00
Yei	1.25	1.03	1.03

Median Weights

	Manufacturing	Retail	Other Services
Juba	1.00	2.69	1.97
Nimule	1.00	1.00	1.16
Torit	2.45	1.00	1.00
Yei	1.25	1.03	1.03

Weak Weights

	Manufacturing	Retail	Other Services
Juba	1.00	5.85	3.87
Nimule	1.00	1.34	1.81
Torit	2.05	1.16	1.09
Yei	1.01	1.20	1.09

Appendix D

Strict Universe Estimates South Sudan

	Manufacturing	Retail	Other Services
Juba	78	431	319
Nimule	2	43	67
Torit	2	5	37
Yei	7	156	37

Median Universe Estimates South Sudan

	Manufacturing	Retail	Other Services
Juba	78	431	319
Nimule	2	43	67
Torit	2	5	37
Yei	7	156	37

Weak Universe Estimates

	Manufacturing	Retail	Other Services
Juba	78	935	627
Nimule	2	58	105
Torit	2	6	40
Yei	6	182	39

Appendix E

Original Sample Design, South Sudan:

	MANUFACTURING	RETAIL	OTHER SERVICES	Grand Total
Juba	66	150	157	373
Nimule	1	43	65	109
Torit	2	6	35	43
Yei	7	150	38	195
Grand Total	**76**	**349**	**295**	**720**

Appendix F

Local Agency team involved in the study:

Local Agency	Tango Consult
Name of Project Manager	Peter Edopu
Name and position of other key persons of the project: Local Survey Implementation Team and corresponding supervisor and enumerator codes:	Data processing: Cynthia Winfred 43 enumerators, 9 supervisors

Sample Frame:

Characteristics of sample frame used	Variables: name of establishment, address, sector, region, telephone number (for around half of records)
Year:	2012, 2013
Comments on the quality of sample frame:	Elements of the frame were of relatively high quality, as the country's statistics office had recently conducted a census after indepedence. However, following conflict between the Dinka and Nuer, large parts of the frame quickly became out of date, meaning more face-to-face verification was required, delaying the close of fieldwork.
Year and organism who conducted the last economic census	National Bureau of Statistics
Other sources for companies statistics	Town commercial registeries, local agency free-find

Sample:

Comments/ problems on sectors and regions selected in the sample	Region selection was limited by the on-going conflict in the Upper Nile Region.
Comments on the response rate	Response rates were relatively high (73%). The survey benefitted from taking a place in a relatively under-researched environment as well as strong field force management from the local contractor.
Comments on the sample design:	Sample design was slanted towards Juba (with over half of interviews slated to come from the city). This was primarily driven by lack of sample in the smaller cities.

Other comments:	None

Fieldwork and country situation:

Date of Fieldwork	July 2014-December 2014
Locations	Juba, Nimule, Torit, Yei
Interview number	739
Problems found during fieldwork	Fieldwork ran relatively smoothly. Rainy season and outbreaks of localised conflict were the biggest challenges resulting in a late start.
Other observations:	PAPI used for data collection