General Household Survey (GHS) - Panel
Farm Area Measurement Validation Study

Basic Information Document

December 2016

National Bureau of Statistics, Nigeria
Living Standards Measurement Study, World Bank

Acronyms

CC	Crop-Cutting
EA	Enumeration Area
GHS-P	General Household Survey - Panel
HH	Household
HHID	Household Identification Number
LSMS	Living Standards Measurement Study
LSMS-ISA	Living Standards Measurement Study – Integrated Surveys on Agriculture
NBS	National Bureau of Statistics, Nigeria
PH	Post-Harvest
PP	Post-Planting

Table of Contents

Background

Accurate and timely crop production statistics are critical to adequate government policy responses and the availability of accurate measures are pivotal to establishing credible performance evaluation systems. However, agricultural statistics are often marred by controversy over methods and overall quality, leading to inertia at best, or entirely incorrect policy actions. Major advances in recent years in technologies and practices offer an opportunity to improve on some of the indicators we commonly use to measure agricultural performance. Considerable efforts were made in the 1960s and 1970s, primarily by the FAO, to build a body of knowledge on agricultural statistics based on sound research which, over the years, has proven invaluable to researchers and practitioners in the field of agriculture. However, little new knowledge has been generated over the past few decades and much of the available methodological outputs are now obsolete in view of the changing structure of the sector, driven by global and local trends in both the agronomics of farming and the environment.

Three decades ago, the lack of information on the measurement and understanding of poverty and the impact of government policies on wellbeing provided the impetus for establishing the Living Standards Measurement Study (LSMS) program at the World Bank. In the course of its lifespan, the LSMS has made a significant contribution in raising the number of developing and transition countries with reliable household survey data for poverty and policy analysis, from 22 in 1990 to over 115 today. Most importantly, the LSMS has contributed to our knowledge on data collection methods, having pioneered, tested and mainstreamed many of the data quality control and household survey design features used today in the majority of household surveys being carried out in developing countries.

The LSMS-ISA, an agriculture-focused project of the LSMS program, and the institutional collaborations on which it is built, provides an ideal platform to support methodological research. The broader LSMS-ISA research agenda is composed of seven primary components: (1) land area measurement, (2) soil fertility, (3) water resources, (4) labor inputs, (5) skill measurement, (6) production of continuous and extended-harvest crops, and (7) computer-assisted personal interviewing for agricultural data. The farm area measurement validation study, conducted on a subsample of the GHS-Panel survey, focuses on the land area measurement component.

The design of the farm area measurement validation study as a subsample of the GHS-Panel was motivated by observed differences between farmer estimates of plot area and GPS measurement. The observed differences between the measures were significantly greater than the differences observed in other LSMS-ISA surveys, and therefore, the validation study was designed in an attempt to explain such large differences (see Dillon et al, 2016 for details).

Methods for Farm Area Measurement

The area of an agricultural parcel can be measured in a number of ways, including by the compass and rope methods (also known as the traversing), by handheld GPS unit, and by farmer self-reported estimate. Each of these methods possesses unique costs and benefits. While experience suggests that compass and rope is time-intensive, it also produces some of the most accurate figures and is therefore often used as the benchmark in comparative exercises (as in

Keita et al. 2010, for example). Farmer self-reported estimates fall on the other end of the spectrum requiring minimal resource expenditures as a trade off for precision. More recently, the availability of affordable and more reliable GPS devices has made GPS-based area measurement a practical alternative that is increasingly being applied in surveys worldwide. Empirical evidence based on nationally representative household surveys comparing GPS-based and self-reported measurement of parcel and plot areas also suggest the existence of systematic errors in self-reported areas (Carletto et al., 2013; Carletto et al., 2015). The GHS-P farm area measurement validation study employs all three of the methods listed above: compass and rope, GPS measurement, and farmer self-reported estimation.

Survey Instrument

The farm area measurement validation study consists of a single questionnaire instrument, which was administered in an additional visiting following the GHS-P Wave 2 Post-Planting visit (refer to the GHS-P BID for details on the full panel study; www.worldbbank.org/lsms or http://go.worldbank.org/FD4VYBYDX0).

Form F.S. 1	
Section A:	Household Identification & Interview Details
Section B:	Plot Identification
Section C:	Farm Survey Measurements

A detailed description of the content of each questionnaire section and the unit of analysis can be found in Table 1.

Sample Design

The farm area measurement validation study was conducted on a subsample of GHS-Panel households. Agricultural plots and plot IDs can be matched to those in Wave 2 of the GHS-Panel survey (post-planting visit).

Because the plot size plays a significant role in the accuracy of plot area measurement using the various methods, the validation sample was on four plot size strata to ensure we could test the various methods on larger plots, which are much more rare. First, four states were purposefully selected based on safety and past performance in area measurement (Benue, Osun, Oyo, and Kogi). Then, using the second wave of the GHS panel as the sample frame and the GPS measurement of the plot taken in the post-planting visit, every plot was assigned to a plot-size strata (strata 1: <=1000 sq. meters; strata 2: 1000-2500 sq. meters; strata 3: 2500-5000sq. meters; strata 4: >5000 sq. meters). One hundred plots were then randomly selected from each strata. This process yielded the selection of 400 plots (211 households). However, in order to maximize the sample at minimal added cost, we included **all** plots from the selected households, not only the plots that were selected in the first step (totaling 518 plots).

From the 518 selected plots, 23 plots were unable to be measured (5 due to land disputes, 4 due to respondent refusal, 14 for other reasons). Therefore, the total number of plots measured and included in the farm area measurement validation study is 495, coming from a total of 202 households.

The breakdown by state is as follows:

State	Plots	Households
Benue	247	85
Kogi	79	48
Osun	70	32
Oyo	99	37
	494	202

Stratification by plot size in the validation sample results in the unequal probability of plot selection within households from the GHS-Wave 2 sample. Household-level sampling weights were calculated for the validation sample to make them representative of the same household population sampled in Wave 2. Refer to Annex I for details on the construction of the sampling weights.

Implementation

Training

Ten staff members from the Nigeria National Bureau of Statistics headquarters in Abuja carried out the farm area study. Training took place in March 2013 in Abuja. The training involved thorough review of the area measurement form, as well as theoretical and practical training on the various measurement methods, including the use of GPS and compass and rope.

Fieldwork & Data Processing

The ten enumerating field staff (along with assistants) were deployed to the four states in groups. Two staff covered Oyo state, 2 staff covered Osun state, 2 staff covered Kogi state, and 4 staff covered Benue state. Supervisors from NBS and the World Bank LSMS were present for the duration of fieldwork and made roving visits to observe and ensure the proper protocols were followed. Fieldwork began in March 2013 and lasted for approximately 3 weeks.

It is important to note that the GPS measurement methodology employed in the validation study was consistent with that utilized in the GHS-Panel sample. The same GPS devices, the Garmin GPSMAP 62, were used in both studies.[1] The built-in area measurement functionality was used in both the validation exercise and GHS-Panel survey, in which enumerators pace the perimeter

[1] The Garmin GPSMAP 62 has a stated position accuracy of less than 10 meters (http://static.garmin.com/pumac/GPSMAP62_OM_EN.pdf). This particular device is no longer in production but a similar unit, the GPSMAP 64, currently sells for approximately $250.

of the plot (rather than a collect points to construct a polygon post-fieldwork). This method is more user-friendly, leaves less room for measurement error as there are fewer independent data points collected, and is potentially better apt to capture irregularly shaped plots. In both cases, enumerators received supervised practical training on the devices prior to the launch of fieldwork. Given the similarities in training and the consistent measurement technique and hardware, the quality of GPS data is expected to be similar to that collected in the large-scale GHS-Panel survey operations (with potentially more data entry errors in the GHS-Panel given the scale of data collection).

The area measurement form was completed on paper (one form per plot). Forms were periodically delivered to NBS headquarters in Abuja for data entry. Each staff member had access to a laptop in the field, for use in calculating the compass and rope area and perimeter (refer to enumerator manual for details).

Data

Users of the farm area methodological validation study data are strongly encouraged to familiarize themselves with both the GHS-Panel Wave 2 and area measurement questionnaire instruments, as well as the enumerator manuals prior to analyzing the data. The data files are named according to the questionnaire module numbers, and variable names, whenever possible, reflect the question numbers in the relative modules.

In addition to the data file for each questionnaire section, there is data file of constructed geovariables and plot shape metrics. These supplementary datasets are detailed in Table 2.

Unique Identifiers

Each household has a unique household identifier (*hhid*). Additionally, to identify a specific plot, the *plotid* must also be used. Both the *hhid* and *plotid* variables can be used to merge the area measurement data with the full Wave 2 GHS-Panel data.

Every data file includes the *hhid* and *plotid*. The complete list of data files as well as the unique identification variables are listed in Tables 1 and 2.

Linking with the GHS-Panel

The sample for the farm area measurement validation study is a subsample of the GHS-Panel survey. Households can be matched across both waves of the GHS panel using the *hhid* variable. Because the GHS-Panel is a household panel only, and not a plot panel, plots from the farm area study can only be matched to Wave 2 of the GHS-P (using the *plotid* variable).

Table 1. Form F.S. 1 (Farm Area Measurement)

Section A: Household Identification & Interview Details

Level of Observation: Household - Plot | **Data File:** secta.dta
Unique Identifier: hhid + plotid

Description:
Contains household location variables (state, LGA, sector, enumeration area, RIC code), date and time of interview, and enumerator and supervisor identification. Household level sampling weights are also included.

Key Notes:
All sensitive identifying variables, including the names of the household head and field staff, have been removed to protect the confidentiality of the respondents.

SECTION B: Plot Identification

Level of Observation: Household - Plot | **Data File:** sectb.dta
Unique Identifier: hhid + plotid

Description:
Household identifiers and plot ID are included here, as in Section A. Section B also contains a list of the crops planted on the plot (as identified in the GHS-P Wave 2 Post-Planting Visit).

Key Notes:
N/A

SECTION C: Farm Area Measurements

Level of Observation: Household- Plot | **Data File:** sectc.dta, sectc1.dta, sectc2.dta
Unique Identifier: hhid + plotid

Description:
Plot area measured via farmer estimate, compass and rope, and GPS are included in Section C.

Farmer estimates of plot area are found in *sectc.dta*.

The data file *sectc1.dta* contains the compass and rope measurement details, including the compass bearings and distances of each side. This particular file is at the household-plot-plot side level (thus, the observations are uniquely identified by the following string of variables: hhid plotid q16a q16b).

The area and perimeter measurements taken via compass and rope and GPS are found in the data file *sectc2.dta.* Indicators of tree coverage and weather conditions at the time of measurement are also included in this file.

Key Notes:
N/A

Table 2: Supplementary Data

GEOVARIABLES & FIELD SHAPE METRICS

Level of Observation: Household - Plot | **Data File:** ShapeMetrics_Nigeria_Public
Unique Identifier: hhid + plotid

Description:
Because a primary focus of the farm area study was to understand the factors influencing measurement error in plot area measurement, several indicators of plot shape and measurement duration were constructed from the plot outline saved during the GPS area measurement.

Additional variables were constructed based on the geospatial data available for the plot, such as average percent forest along the plot perimeter.

Key Notes:
The raw GPS data is not released in order to protect the respondent.
A description of the variables is included in Annex II.

Annex I. Construction of Household Sample Weights

For the Nigeria Panel Survey plot area measurement study, the sampling frame was based on all plots for the Panel sample households in the following four states: Benue, Kogi, Osun and Oyo. Since the measurement error varies by the size of the plot, the frame was stratified by four plot size categories, identified in Table 1. A sample of 100 plots was selected within each of these strata across the four states. Table 1 shows the distribution of the plots in the frame and the sample by stratum, and the corresponding probabilities of selection.

Table 1. Distribution of plots in the frame and the sample for the Panel Survey households in four states by stratum, with original probabilities of selection

Strata	No. plots in frame	No. plots in original sample	Probability	Probability of not being selected
(1) 1 to 1,000 square meters	151	100	0.6623	0.3377
(2) 1,001 to 2,500 square meters	169	100	0.5917	0.4083
(3) 2,501 to 5,000 square meters	142	100	0.7042	0.2958
(4) 5,001 square meters	193	100	0.5181	0.4819
Total	655	400		

Following the selection of the 400 sample plots, it was decided to include in the sample for the plot area measurement study all the plots in each household with at least one plot selected in the original sample. The final sample had a total of 494 plots in 210 sample households. The final weight for these plots would be equal to the Panel Survey household weight for Wave 2 times the inverse of the subsampling rate for the plots, and can be expressed as follows:

$$W_{Pi} = W_{Si} \times \frac{1}{p_{2i}},$$

where:

W_{Pi} = final weight for the sample plots in the area measurement study

W_{Si} = final weight for the households in the Panel Survey for Wave 2

p_{2i} = probability of the household being selected from the Panel Survey frame based on the probabilities of all of its plots

The probability p_{2i} of the Panel Survey household being selected for this study is based on the probability that at least one of its plots is selected. Given the complexity of calculating this probability, it is simpler to determine the probability that none of the plots is selected, and then subtract this from 1 to determine the probability that one or more of the plots of the household are selected. The probability that a particular plot is not selected is also shown in Table 1. Assuming that the original probabilities of selection of the individual plots are independent, the probability that no plot in the household is selected can be calculated by multiplying the probabilities of the individual plots not being selected. Therefore the probability p_{2i} can be calculated as follows:

$$p_{2i} = 1 - \left[(0.3377)^{n_{1i}} \times (0.4083)^{n_{2i}} \times (0.2958)^{n_{3i}} \times (0.4819)^{n_{4i}}\right],$$

where:

n_{hi} = number of plots in stratum h within the i-th sample household

It can be seen in this expression for the probability p_{2i} that the more plots that the household has, the higher its probability of selection, as expected. That is, in the case of a household with many plots, the probability that it is not selected is very small. However, no household has a probability of 1 of being selected.

All the plots in each sample household have the same weight. The sample weights variable (*hh_weight*) can be found in the *secta.dta* data file.

Annex II. Geovariables and Field Shape Metrics

Variable Name	Description	data input format	notes
hhid	household id	gpx	
plotid	plot id	gpx	
date	date of gps data collection	gpx	
start_time_gpx	time of start of perimeter walk (time format h:mm:ss AM/PM)	gpx	
end_time	time of end of perimeter walk (time format h:mm:ss AM/PM)	gpx	
time_mm_ss_gpx	total time of perimeter walk (time format mm:ss)	gpx	
time_dmin_gpx	total time of perimeter walk, units are minutes	gpx	
num_vert	number of vertices defining perimeter	gpx	
walk_speed_gpx	derived using time stamp on first and last vertex and perimeter length. Units are meters per minute	calculation	
vert_density	derived using number of gpx vertices and perimeter length. Units are meters per vertex.	calculation	
perimeter_gpx	length of perimeter. Units are meters	shapefile - polygon	1
area_gpx	area of plot. Units are acres	shapefile - polygon	1; Units are acres
altitude_gpx	Altitude measured by handheld GPS (meters)	gpx	
proximity	Proximity Index: average Euclidean distance from all interior points to the centroid (center of gravity)	shapefile - polygon	2
nproximity	Proximity Index normalized using circle of equal area (reduces to measure of compactness, removes effect of shape)	shapefile - polygon	2
depth	Depth Index: average distance from the shape's interior points to the nearest point on the perimeter	shapefile - polygon	2
ndepth	Depth Index normalized using circle of equal area (reduces to measure of compactness, removes effect of shape)	shapefile - polygon	2
girth	Girth Index: radius of the largest circle that can be inscribed in the shape	shapefile - polygon	2
ngirth	Girth Index normalized using circle of equal area	shapefile - polygon	2
range	The Range Index: diameter of the smallest circle that fully circumscribes the polygon	shapefile - polygon	2
nrange	Range Index normalized using circle of equal area	shapefile - polygon	2
detour	Detour Index: perimeter of comvex hull	shapefile - polygon	2
ndetour	Detour Index normalized using circle of equal area	shapefile - polygon	2
for2010_avg	average percent forest along perimeter, extracted by vertex	shapefile - point	3
for2010_max	max percent forest along perimeter, extracted by vertex	shapefile - point	3
ag_pct	landscape-level percent cultivated area (1km resolution), extracted by plot centerpoint	shapefile - point	4
dem	elevation of plot centerpoint, extracted by plot centerpoint. Units are meters	shapefile - point	5
slp	slope at plot centerpoint, extracted by plot centerpoint. Units are percent	shapefile - point	5
rat	surface area ratio (surface area divided by planimetric area), extracted by plot centerpoint	shapefile - point	5

zs_dem_mean	average elevation over plot area, may be missing for very small plots. Units are meters	raster	5
zs_rat_mean	average surface area ratio over plot area, may be missing for very small plots. Unitless	raster	5
zs_slp_mean	average slope over plot area, may be missing for very small plots. Units are meters	raster	5

Notes:

1 - Derived in arcGIS using Transverse Mercator projection (CM=38.0, LO=0,0)

2 - Shape metrics tool downloaded from Univ of Connecticut Center for Land Use Education and Research http://clear.uconn.edu/tools/Shape_Metrics/download.htm

3 - High-resolution forest cover 2012 (derived from 2000 base year and total change). (M. C. Hansen et al.) downloaded from http://earthenginepartners.appspot.com/science-2013-global-forest.

4 - Fritz et al. Global Ag Hybrid

5 - ASTER GDEM tiles downloaded from http://gdem.ersdac.jspacesystems.or.jp, elevation derivatives generated using DEM Surface Tools for ArcGIS (2012, J. Jenness)

References

Carletto, G., Gourlay, S., & Winters, P. (2015). From Guesstimates to GPStimates: Land Area Measurement and Implications for Agricultural Analysis. *Journal of African Economies*, 24 (5), 593-628.

Carletto, G., Savastano, S., & Zezza, A. (2013). Fact or artifact: The impact of measurement errors on the farm size–productivity relationship, *Journal of Development Economics*, 103(C), 254-261.

Dillon, Andrew S.; Gourlay, Sydney; Mcgee, Kevin Robert; Oseni, Gbemisola O.. 2016. *Land measurement bias and its empirical implications : evidence from a validation exercise*. Policy Research working paper; no. WPS 7597. Washington, D.C. : World Bank Group. http://documents.worldbank.org/curated/en/258851468197930768/Land-measurement-bias-and-its-empirical-implications-evidence-from-a-validation-exercise

Jenness, J., 2012. DEM Surface Tools. Jenness Enterprises. Available from: http://www.jennessent.com/arcgis/surface_area.htm.

Keita, N., Carfagna, E., and Mu'Ammar, G. (2010). "Issues and Guidelines for the Emerging Use of GPS and PDAs in Agricultural Statistics in Developing Countries." The Fifth International Conference on Agricultural Statistics; Kampala, Uganda.