Ethiopia Land and Soil Experimental Research (LASER) Study

Basic Information Document

Version 2 - October 2017

Central Statistical Agency
Living Standards Measurement Study, World Bank
World Agroforestry Centre

Acronyms

CC	Crop-Cutting
CSA	Central Statistical Agency of Ethiopia
EA	Enumeration Area
HH	Household
HHID	Household Identification Number
ICRAF	World Agroforestry Centre
LSMS	Living Standards Measurement Study
LSMS-ISA	Living Standards Measurement Study – Integrated Surveys on Agriculture
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-SA 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 Ethiopia Land and Soil Experimental Research (LASER) Study focuses on three of the abovementioned components (land area, soil fertility, and computer-assisted personal interviewing) in addition to the primary agricultural output, crop production.

Measuring land area and soil quality is essential in properly estimating the factors that both promote and hinder agricultural productivity. It is also critical to assess the accuracy of the key output variable, crop production, in order to validate the methodologies used to collect harvest data as well as analyze the impact of various input measurements on yield estimates. By measuring these components using a variety of methods it is possible to identify the implications of using each and move forward with the superior methods in future household surveys. The aim of the LASER study, therefore, is to assess the data quality associated with a number of possible measurement methodologies associated with land area, soil quality, and crop production while piloting the use of each method and assessing the feasibility of implementation in national household surveys.

Methods

Land Area

The area of an agricultural parcel can be measured in a number of ways, including traversing (also known as the compass and rope method), by handheld GPS unit, and by farmer self-reported estimate. Each of these methods possesses unique costs and benefits. While experience suggests that traversing 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 LASER experiment employs all three of the methods listed above: traversing, GPS measurement, and farmer self-reported estimation.

Soil Quality

New, rapid low cost technology for assessing soil characteristics using infrared spectroscopy has made soil fertility evaluations feasible in large studies (Shepherd & Walsh, 2002; 2007). Several methodologies are available for soil analysis. Of interest to this project are two, namely conventional soil analysis (CSA) and spectral soil analysis (SSA). CSA includes traditional wet chemistry methods for soil nutrient extraction and some basic soil physical analyses, such as water holding capacity. SSA includes mid-infrared diffuse reflectance spectroscopy (MIR), laser diffraction particle size analysis (LDPSA), x-ray methods for soil mineralogy (XRD) and total element analysis (TXRF). Although CSA is more expensive and destroys the soil samples analyzed, it provides comprehensive data on different elements and provides a reference method of analysis of soil characteristics. SSA is non-destructive and relatively inexpensive. Conventional soil analysis was conducted on 10% of the soils, while spectral soil analysis was be conducted on all samples. Including a subset of samples that is tested with both methods is critical in order to enable calibration of the soil spectral library and validation of its predictive performance. Soil analysis was executed by the World Agroforestry Centre (ICRAF).

A series of subjective soil quality questions were also asked of the household in order to allow for comparison between subjective and objective methods of soil quality assessment.

Crop Production

Without a sound measurement of crop production itself, yield estimates may still be inaccurate no matter how precise the land area measure. For this reason, the LASER study couples land area and crop-cutting components. Due to the relatively small sample in this experiment, we will focus only on maize yields. If only maize plots were selected, however, the sample could be biased for the land area and soil quality analysis. Therefore, only a subsample of plots are maize and subject to crop-cutting, while the remaining plots were randomly selected irrespective of crop type. Farmer self-reported production and crop-cutting methods were employed. Crop-cutting activities were executed on a 4x4m subplot of each selected pure stand maize field. The 4x4m subplot was then divided into 2x2m quadrants and production was recorded separately for each quadrant. Additionally, all production from the crop-cutting subplot was weighed using both an analog and digital scale.

Survey Instruments

The LASER study consists of three questionnaires: Post-Planting, Crop-Cutting, and Post-Harvest.

Post - Planting Questionnaire	
Cover:	Household Identification & Interview Details
Section 1:	Household Roster
Section 2:	Household Assets
Section 3:	Livestock
Section 4:	Parcel Roster
Section 5:	Field Roster
Section 6:	Field Details
Section 7:	Crop Details
Section 8:	Field Selection
Section 9.1:	In-Field Measurement (selected field #1)
Section 9.2:	In-Field Measurement (selected field #2)

Crop - Cutting Questionnaire	
Crop-Cutting:	Crop-Cutting Activities

Post – Harvest Questionnaire	
Cover:	Household Identification & Interview Details
Section 1:	Household Roster
Section 1.5:	Field Roster
Section 2:	Harvest
Section 3:	Crop Disposition
Section 4:	Harvest Labor

A detailed description of the content of each questionnaire section and the unit of analysis can be found in Table 1 (PP), Table 2 (CC), and Table 3 (PH).

Sample Design

The objectives of this research are multifaceted and include indicators related to soil properties, crop type, and socio-economic characteristics, among others. Because there are multiple indicators, calculating the sample size based on the variance of a single indicator was not the preferred approach. Instead, practical sampling allocation with implicit stratification was used.

Three administrative zones of the Oromia region were selected based primarily on agroecology and geographic diversity. Secondary consideration was made for the availability of local soil research centers that could be used for soil processing. The three selected zones are: East Wellega, West Arsi, and Borena. Using the CSAs Agricultural Sample Survey (AgSS) as the sampling frame, a total of 85 EAs were selected.

Below is the **TOTAL NUMBER** of AgSS EAs in each agroecology:

Zone	# Dega	# Weyna Dega	# Kolla	Total EAs
East Wellega	0	25	15	40
West Arsi	14	25	0	39
Borena	0	17	21	38
	14	67	36	117

The determined **PRACTICAL ALLOCATION** of EAs across administrative and agroecological zones:

Zone	# Dega	# Weyna Dega	# Kolla	Total EAs
East Wellega	0	17	11	28
West Arsi	14	14	0	28
Borena	0	13	16	29
	14	44	27	85

The allocation of EAs was determined by the proportion of EAs across agroecological zones within each administrative zone, with the exception of the allocation within West Arsi, which was split evenly between dega and weyna dega EAs due to the small dega population.

After selecting the EAs, it came to our attention that five of the selected EAs were no longer included in the AgSS survey, and therefore the household listing was unavailable. In Borena, one of the selected EAs was dropped from the AgSS survey without replacement. Two other EAs in Borena (one Kolla, one Weyna Dega) had been replaced in previous years. In East Wellega, two of the selected EAs were dropped from the AgSS survey and replaced. Therefore, we resampled the EAs using simple random sampling within the remaining EAs in the appropriate region and agroecology combination. In East Wellega, the two replacement AgSS EAs were included in the random selection of the replacement EAs for this project.

Within each EA, 12 households were randomly selected from the AgSS household listing completed September 2013. Households which were selected for the AgSS were ineligible for selection in this project. Up to 2 fields were measured per household. First, if any fields contained pure stand maize, one was randomly selected. Then, a second field was randomly selected from the remaining cultivated fields irrespective of crop type. If no fields contained pure stand maize, two fields were randomly selected. Only those fields with pure stand maize were subject to crop cutting.

Two households were not located or refused to participate. Therefore, the total sample size is 1018 households.

Implementation

Training

Fieldwork training was held centrally for all enumerators and team leaders. Training for post-planting and crop-cutting activities was held for approximately three weeks in August 2013 in Debre Zeit, Ethiopia. The Central Statistical Agency of Ethiopia conducted the training, with assistance of the World Bank project manager. The training involved thorough review of the paper questionnaire, training on the Computer Assisted Personal Interviewing (CAPI) program, data management, and theoretical and practical training on the various measurement methods, including the use of GPS, compass and rope, clinometer use, crop-cutting protocols, and soil collection. Training on soil sampling and handling was conducted by a representative from the World Agroforestry Centre.

Training for post-harvest activities was conducted for approximately one week in January 2014 in Hawassa, Ethiopia. Post-Harvest training included thorough review of the paper questionnaire and CAPI program.

In total, 19 enumerators and 5 supervisors were hired. During Post-Planting fieldwork, one enumerator and one supervisor left the project. During Post-Harvest fieldwork, one enumerator was released.

Fieldwork

Field staff were divided into five teams each with 3 or 4 enumerators, one supervisor and one driver. Initially, two teams were sent to East Wellega zone, two teams to Borena zone, and one team to West Arsi zone. Each team was provided with one vehicle for the duration of fieldwork. Teams moved from EA to EA interviewing households and conducting the measurements. Within the EAs, the enumerators separated and interviewed different households. Each enumerator was able to hire one local guide in each EA to assist with the measurements, particularly for crop-cutting.

Each household was visited more than one time. Depending on the area and the timing of harvest, each house was visited up to 4 times. The visits were as follows:

The first visit took place in September or October, after the crops had been planted. During this visit, the enumerator administered the post-planting questionnaire, conducted the in-field measurements (including GPS, compass and rope, and clinometer), collected soil samples, and set the area for crop-cutting. It was critical that the household was visited *before* the maize harvest in order to properly prepare for crop-cutting. Soils samples were collected during the ost-planting visit and delivered to the local processing lab within 5-10 days of collection.

The team returned to the household for the second visit when the maize was ready to be harvested. The kebele officer alerted the supervisor when the field was ready. At this time, the enumerator harvested the crop from the crop-cutting area and took the fresh weight. The enumerator then dried the maize until the weight was steady from one day to the next and then took the dry weight and returned the crop to the household (or returned the crop during the final household visit). The timing of the crop-cutting visits was different for every EA.

The last visit took place in January or February. This visit occurred after all or most of the crops had been harvested. During this visit the enumerator administered the post-harvest questionnaire.

The general timeline was as follows:

Post-Planting Fieldwork:	September 12 – November 15, 2013
Crop-Cutting Activities:	October - December 2013
Post-Harvest Training:	January 8-10, 2014
Post-Harvest Fieldwork:	January 13 – February 7, 2014

Data Processing & Management

Data collection for LASER was completed via CAPI. Each enumerator and supervisor had a personal laptop computer equipped with the CSPro based CAPI application for the Post-Planting, Crop-Cutting, and Post-Harvest questionnaires. Each team was provided with a flash drive, to share data from enumerator to supervisor, and a wireless router, to share consolidated team data with the World Bank project manager. Supervisors were instructed to share data at the close of EA, and only after reviewing all completed questionnaires.

Data review and cleaning took place via supervisor review, periodic error reports generated by the World Bank project manager, unplanned CSA supervisor household visits to cross-check responses, and ultimately data review and standard checks (possible value ranges, outliers, etc.).

LASER Data

Users of the LASER data are strongly encouraged to familiarize themselves with the questionnaire instruments and enumerator manual 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. Note that the questionnaire makes use of skip patterns to maximize interview efficiency. It is necessary to keep these skip patterns in mind to properly interpret the data. Skip patterns are indicated in the questionnaire with a black triangle followed by the number of the question to which the enumerator should skip. Refer to the enumerator manual for more detail explanation and examples.

In addition to the data file for each questionnaire module, there is data file with results from the soil analysis and constructed geovariables and plot shape metrics. These supplementary datasets are detailed in Table 4.

Unique Identifiers

Each household was assigned a unique four-digit household identification number (HHID). The first two digits of the HHID correspond to the cluster, and the second two digits range from 01-12. Every data file includes the HHID. Data files that are at the parcel level include the HHID and a parcel ID. Data files at the field level include the HHID, parcel ID, and field ID. It is necessary to use all three of these variables when merging field level files. The complete list of data files as well as the unique identification variables are listed in Tables 1-4.

Table 1. Post-Planting Questionnaire

COVER PAGE: Household Identification & Interview Details

Level of Observation: Household **Data File:** COVER_PAGE
Unique Identifier: hhid

Description:
Contains household location variables, date and time of interview, and enumerator and supervisor identification. Also included are the agroecological zone of the enumeration area, as identified by the CSA in the sampling stage, and the type of tablet used for the alternative GPS measurement option (see Section 9 for details).

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

SECTION 1: Household Roster

Level of Observation: Individual **Data File:** PP1_HH_ROSTER
Unique Identifier: hhid, individual_id

Description:
Roster of household members and their individual characteristics, including sex, age, household membership status, education variables, general occupation, religion and marital status.

Key Notes:
The individual name has been removed to protect the respondent.

SECTION 2: Household Assets

Level of Observation: Household **Data File:** PP2_ASSETS
Unique Identifier: hhid, item_name

Description:
Quantity of key items owned by the household and estimated current market value (per item) in birr.

Key Notes:
N/A

SECTION 3: Livestock

Level of Observation: Household
Unique Identifier: hhid, livestock

Data File: PP3_LIVESTOCK

Description:
Quantity of different types of livestock owned by the household at the time of the interview.

Key Notes:
N/A

SECTION 4: Parcel Roster

Level of Observation: Parcel
Unique Identifier: hhid, parcel_id

Data File: PP4_PARCEL_ROSTER

Description:
Lists all parcels of land owned or cultivated by the household. Collects information on property rights, acquisition details, and distance from the parcel to the household, nearest road, and nearest market.

Key Notes:
A parcel may contain one or more fields. A parcel is defined as a contiguous piece of land with identical (uniform) tenure and physical characteristics. It is entirely surrounded by land with other tenure and/or physical characteristics or infrastructure e.g. water, a road, forest, etc.

SECTION 5: Field Roster

Level of Observation: Parcel-Field
Unique Identifier: hhid, parcel_id, field_id

Data File: PP5_FIELD_ROSTER

Description:
Lists all fields for all parcels owned or cultivated by the household. Collects information on estimates of field area, use of the field, decision-making on the field, and agricultural practices (including crop rotation, irrigation, and fallow periods). There is also a question on whether maize was planted on the field. This will serve as a filter section for the crop-cutting selection.

Key Notes:
A field is defined as a contiguous piece of land within a parcel on which a specific crop or a crop mixture is grown. A parcel may be made up of one or more fields.

SECTION 6: Field Details

Level of Observation: Parcel-Field
Data File: PP6_FIELD_DETAILS
Unique Identifier: hhid, parcel_id, field_id

Description:
This section collects more detailed data on all *cultivated* fields listed in the field roster. Data collected includes subjective assessment of field slope and multiple soil quality indicators, as well as fertilizer use and pre-harvest labor.

Key Notes:
N/A

SECTION 7: Crop Details

Level of Observation: Field-Crop
Data File: PP7_CROP_DETAILS
Unique Identifier: hhid, parcel_id, field_id, cropcode

Description:
This sections includes questions on the crops planted on each field, cropping patterns, pesticide/herbicide use, and type and quantity of seed used (where applicable).

Key Notes:
This section includes both annual/field crops and permanent/root crops.

SECTION 8: Field Selection

Level of Observation: Field
Data File: PP8_FIELD_SELECTION
Unique Identifier: hhid, parcel_id, field_id

Description:
This section is used to select 2 fields for detailed testing. The fields selected here are measured with GPS, compass and rope, and clinometers. Soil samples were also collected from these two fields. Crop-cutting was conducted on pure stand maize fields only.

Two fields are selected randomly. The first field is selected only from the pure stand maize fields. The second field is selected from all remaining fields. If the household does not have any pure stand maize fields, two fields are selected from the full list of cultivated fields. You will need Random Number Table #1 in order to complete this section.

Key Notes:
The CAPI program automatically completed questions 1-5. The enumerator then used the field selection protocol (see enumerator manual, page 20-23) and a household-specific random number table to complete the selection of up two fields per households.

SECTION 9.1: In-Field Measurement (Field #1)

Level of Observation: Field
Unique Identifier: hhid, parcel_id, field_id

Data File: PP9_1_IN_FIELD_MEASUREMENT

Description:
Section 9 is completed for both of the field selected in Section 8. Section 9.1 is for the first field, and 9.2 is for the second field. This section is captures the in-field activities, including details on compass and rope area measurement, GPS area measurement, GPS coordinates, slope measurement with clinometer, demarcation of the crop-cutting subplot, and soil sample collection. Refer to the enumerator manual for detailed instructions for each of the activities mentioned above.

Key Notes:
Questions 7-9 are captured in a separate data file (see below).

GPS coordinates and area measurement were conducted with Garmin eTrex 30 devices. The compass/clinometer used was the Suunto MC-2G.

GPS coordinates and field outlines are not released in the public data to protect the respondent. A selection of variables derived from the field location and outline were derived internally and are made available in the supplementary data (see Table 4).

SECTION 9.1: Compass & Rope (Field #1)

Level of Observation: CR Measurement
Unique Identifier: hhid, parcel_id, field_id, pp91_7a, pp91_7b

Data File: PP9_1_CR

Description:
This data file contains questions 7-9 from Section 9. This captures the front-bearing, back-bearing, and distance between every two consecutive corners of the field, as recorded in the compass and rope area measurement.

Key Notes:
Refer to the enumerator manual for full instructions on completing the compass and rope measurement.

If the closing error was greater than 5%, the enumerator was instructed to repeat the measurement until they achieved a closer error lower than that threshold. The bearings and distances in this section reflect the final measurement.

SECTION 9.2: In-Field Measurement (Field #2)

Level of Observation: Field
Unique Identifier: hhid, parcel_id, field_id

Data File: PP9_2_IN_FIELD_MEASUREMENT

Description:
Same as Section 9.1, repeated for second selected field.

Key Notes:
N/A

SECTION 9.2: Compass & Rope (Field #2)

Level of Observation: CR Measurement
Unique Identifier: hhid, parcel_id, field_id, pp92_7a, pp92_7b

Data File: PP9_2_CR

Description:
Same as Section 9.1, repeated for second selected field.

Key Notes:
N/A

Table 2: Crop-Cutting Questionnaire

CROP-CUTTING FORM: Crop-Cutting Activities

Level of Observation: Field
Unique Identifier: hhid, parcel_id, field_id

Data File: Crop_cutting

Description:
The crop-cutting questionnaire includes the cover details from the post-planting questionnaire, the date of crop-cutting, damage or premature harvesting in the crop-cutting subplots, weight of maize at the time of harvest for each 2x2 quadrant, and weight of maize for each 2x2 quadrant after drying.

Key Notes:
Teams were instructed to dry the maize until the weight was consistent from one day to the next (approximately 15 days, depending on moisture content of maize at time of harvest).

Table 3: Post-Harvest Questionnaire

COVER PAGE: Household Identification & Interview Details

Level of Observation: Household **Data File:** PH_COVER_PAGE
Unique Identifier: hhid

Description:
Contains household location variables, date and time of interview, and enumerator and supervisor identification.

Key Notes:
N/A

SECTION 1: Household Roster

Level of Observation: Individual **Data File:** PH1_HOUSEHOLD_MEMBER_R
Unique Identifier: hhid, individual_id

Description:
This section captures the name, age, sex, and ID number of each household member.

Key Notes:
This information is automatically populated from the post-planting household roster. New members are added.

SECTION 1.5: Field Roster

Level of Observation: Field **Data File:** PH2_1_FIELD_ROSTER
Unique Identifier: hhid, parcel_id, field_id

Description:
Lists the fields and the respective cultivation status.

Key Notes:
This section is automatically populated from the Post-Planting field roster and was used only as a reference and to ease CAPI implementation.

SECTION 2: Harvest

Level of Observation: Field - Crop

Unique Identifier: hhid, parcel_id, field_id, cropcode

Data File: PH2_HARVEST

Description:
This section captures the estimated quantity of production and damage for all crops on all cultivated fields.

Key Notes:
The parcel ID, field ID, and cropcode were automatically populated from the crop details section of the post-planting questionnaire.

The scope of the questions is limited to crops that were planted at the time of the initial interview. If the household planted additional crops or acquired more fields since the post-planting interview, that information was not recorded.

SECTION 3: Crop Disposition

Level of Observation: Crop

Unique Identifier: hhid, cropcode

Data File: PH3_CROP_DISPOSITION

Description:
This section collects information on how the crops were used and crop sales, if any.

Key Notes:
As with the section above, questions are only asked of the crops that were in the field at the time of the post-planting interview.

SECTION 4: Harvest Labor

Level of Observation: Field

Unique Identifier: hhid, parcel_id, field_id

Data File: PH4_HARVEST_LABOR

Description:
This section asks detailed questions about the harvest labor used on each cultivated field.

Key Notes:
N/A

Table 4: Supplementary Data

SOIL ANALYSIS RESULTS

Level of Observation: Field – Sample Depth
Unique Identifier: hhid, parcel_id, field_id, depth, cc

Data File: LASER_SoilResults_V2_Ensemble*

Description:
Soil samples were collected from each of the selected fields. From each field, a top-soil sample (0-20cm depth) and a sub-soil sample (20-50cm depth) were analyzed. Plots with crop-cutting had a third sample tested, from the 4x4m crop-cutting subplot. Samples from the crop-cutting subplot are identified by the “cc” variable (cc=1 if crop-cutting sample).

Key Notes:
ICRAF tested 10% of the samples using conventional testing and employed spectral analysis on all samples. A set of indicators was then predicted from the 10% subsample to the full sample. A description of the variables is included in Annex I.

*This is the second version of this file. It was updated in October 2017 to reflect ICRAF’s improved prediction model, known as the ensemble model. The first version of this file included data predicted purely with the Random Forest model.

GEOVARIABLES & FIELD SHAPE METRICS

Level of Observation: Field
Unique Identifier: hhid, parcel_id, field_id

Data File: ShapeMetrics_LASER_Public

Description:
Because a primary focus of the LASER 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. Soil Analysis Data

The table below describes the variables found in the *LASER_SoilResults_V2_Ensemble* data file. This dataset is the result of conventional and spectral soil analysis conducted by ICRAF in Nairobi, Kenya. The predictive power of the ensemble method (the predictive model used) for each variable is also included below (R-squared).

Variables	Units	Variable Description	R-Squared
depth	-	Topsoil=1, Subsoil=2	-
cc	-	=1 if sample is from crop-cutting subplot	-
tc	% by weight	Total Carbon	0.92
oc	% by weight	Organic Carbon	0.9
tn	% by weight	Total Nitrogen	0.87
on	% by weight	Organic Nitrogen	0.94
p*	mg/kg	Total Phosphorus	0.91
k	mg/kg	Total Potassium	0.85
exk	cmolc/kg	Exchangeable potassium concentration by Mehlich 3 extraction	0.81
exca	cmolc/kg	Exchangeable calcium concentration by Mehlich 3 extraction	0.9
mg	mg/Kg	Total Magnesium	0.69
exmg	cmolc/kg	Exchangeable magnesium concentration by Mehlich 3 extraction	0.89
s	mg/kg	Total Sulphur	0.96
m3s	mg/kg	Sulphur by Mehlich 3 extraction	0.79
cl	mg/kg	Total Chlorine	0.92
mcfe	mg/kg	Iron by Mehlich 3 extraction	0.92
m3b	mg/kg	Boron by Mehlich 3 extraction	0.86
m3mn	mg/kg	Manganese by Mehlich 3 extraction	0.85
zn	mg/kg	Total Zinc	0.93
m3zn	mg/kg	Zinc by Mehlich 3 extraction	0.76
cu	mg/Kg	Total Copper	0.85
m3cu	mg/kg	Copper by Mehlich 3 extraction	0.83
ni	mg/Kg	Total Nickel	0.8
na	mg/kg	Total Sodium	0.56
exna	cmolc/kg	Exchangeable sodium concentration by Mehlich 3 extraction	0.81
al	mg/kg	Total Aluminium	0.94
m3al	Mg/kg	Aluminium by Mehlich 3 extraction	0.93
ecd*	dS/m	Soil electrical conductivity	0.82
exac	cmolc/kg	Exchangeable Acidity	0.62
exbas	cmolc/kg	Sum of Mehlich-3 bases (Ca, Mg, K, Na)	0.9
psi	-	Phosphorus Sorption Index	0.88
ph	-	pH	0.95
clay	% by volume	Clay	0.91
silt	% by volume	Silt	0.9
sand	% by volume	Sand	0.89

* Variables "p" and "ecd" are included as string variables. There are 2 observations of "ecd" and 256 observations of "p" which are reported as "NA". These values are not available as the prediction resulted in a negative value.

Annex II. Geovariables and Field Shape Metrics

Variable Name	Description	data input format	notes
hhid	household id	gpx	
parcel_id	parcel id	gpx	
field_id	field 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
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.

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.

Shepherd, K. and M. G. Walsh (2002). "Development of reflectance spectral libraries for characterization of soil properties." Social Science Society of America Journal 66 (3), 988-998.

Shepherd, K. and M. G. Walsh (2007). "Infrared spectroscopy-enabling an evidence-based diagnostic surveillance approach to agricultural and environmental management in developing countries." Journal of Near Infrared Spectroscopy 15, 1-19.