Data Curation and Indicator Construction: General Considerations and Principles

EPAR Technical Brief #335 C. Leigh Anderson & Travis Reynolds

Professor C. Leigh Anderson, Principal Investigator
Professor Travis Reynolds, co-Principal Investigator

November 3, 2017

This brief is intended to summarize general considerations and principles for indicator construction, and are drawn from EPAR's choices for treating the data and constructing the selected agricultural development indicators across the three initial LSMS-ISA instruments and two "Baseline" surveys (Ethiopia ACC and India RMS). The Appendix contains an empirical example of how these choices can affect estimates, using Tanzanian maize yields as reported in the TNPS Wave 1.

Data cleaning

General cleaning:

We did not actively clean the data in search of illogical values or inconsistent responses. Our understanding is the baseline collection teams and World Bank LSMS-ISA team go through a rigorous process of data cleaning before sharing the data. However, in the process of creating these variables, if an illogical entry made itself obvious (in a way that made it difficult to merge data files or produced impossible values for the final indicators), we sometimes amended this in the process of indicator construction. *This was rare*.

Examples:

(1) Where both a resident man and woman are listed as married heads of household (i.e., the spouse is categorized as another head), we revised this to categorize the woman as a spouse, not another head.
(2) Where respondents reported the number of animals sold and the value received in a manner that indicates these columns had been accidentally switched (e.g., 150 cows sold for one Ethiopian Birr), these values were switched before estimating livestock income.
(3) In one survey, the multitude of units reported for crops made it possible for a given crop in a household to receive an estimated value that was higher than the reported value of crop sales, even where 100% of that crop was sold. In these odd cases, the value received from sales is considered to be the value of the crop production.

Outliers:

Extreme outliers can influence the average value of an indicator. Outliers are dealt with in a variety of ways including trimming, dropping, replacing with median or mean, multiple imputation, etc. The choice of method to deal with outliers can make an important difference in the result depending on the distribution of the variable (see Appendix for an example of differences results from the choice of outlier treatment when calculating maize yields in Tanzania).

Before computing the summary statistics of final constructed indicators, we identified outliers using the 1st and 99th percentile of the indicator's distribution. We then winsorized values that were either smaller or larger than these thresholds. For example, we applied the value of the 99th percentile to any observations that are larger than this threshold. Depending on what is logical for a given indicator, we winsorize at both the 1st and 99th percentiles or just the 99th percentile (if there is no illogically small value for the variable).

A few variables were winsorized a bit earlier in the indicator construction process. For example, before creating crop yields, area harvested was winsorized at the high and low end (this was done in order to be able to apply correct area weights to the yield observations). Once yield was created, this value (ratio) was winsorized only at the high end (since extremely low yields are not illogical).

Particularly for estimating the average value of a variable, the method chosen for dealing with outliers can make a considerable difference. For this reason, attention should also be given to the median value and the overall distribution of the variable. If a single summary statistic is demanded, we typically prefer the median to the mean.

Weights

Where survey weights are provided (four out of five surveys), these are used in estimating summary statistics for all indicators. Weights are used to ensure that statistics estimated with the sample are unbiased estimates of the population parameters. Thus, weighted statistics can only be reported at geographical levels where the sample is representative of the population. In the India baseline, the sample is representative at the state level and statistics are reported by state. In the LSMS-ISA, the sample is nationally representative but not always regionally representative; thus, we only report the statistics at the national level. No survey weights are available for the Ethiopia baseline, so we implicitly use a weight of "1" for each observation.

For a few indicators, these weights are adjusted so that the summary statistics reflect the average unit of the indicator, rather than the average experience of a household.

Examples:

(1) For land productivity and crop yield, we use area-adjusted weights (hectares * survey weight) so that the mean value reflects the average hectare of land. This follows the World Bank LSMS-ISA team's convention.
(2) For labor productivity, we use labor-adjusted weights (labor-days * survey weight) so that the mean value reflects the average productivity of a typical labor-day.
(3) For milk productivity, we use livestock-adjusted weights (herd-size * survey weight) so that the mean value reflects the milk productivity of an average animal, rather than an average household.

Sub-populations

In many cases, the population for a given indicator will be a sub-population of the overall sample as a result of indicator construction. For example, the yield indicator for a given crop will only apply to farm households, and more specifically to plots in those households where that crop was grown.

For all estimates of the AgDev priority indicators, we restricted the population to rural households in the LSMS-ISA data sets, or all households in the baseline data sets (as these are a sample of rural households). This means that we do not capture the experiences of urban women or urban agriculturalists, for example; if they were included, the variable distributions might shift. This decision can be amended when referring to the data files provided.

It is essential to ensure that subpopulations are correctly dealt with and particularly that the syntax of the code is written in a way that includes all observations in the calculation of the standard errors.

For all indicators, when the sample size of the relevant a sub-population is less than 30, we urge caution in interpreting estimates, as it is likely that the statistical power will be too small to obtain valid estimates.

Data and do file management

We aggregate all code for each instrument into a single master do file. Each do file is structured to make it easier for an outside user to understand the process for constructing a given indicator, and to make it possible to trace indicator construction back to the original raw data. This is intended to facilitate any potential revisions to indicator code, and to reduce the amount of duplication of code across multiple indicator do files.

At the bottom of each data set's master do-file, we created a single data file for all variables that are at the household-level, another data file for individual-level variables, and another for plot-level variables.

Specific indicator construction considerations

Prices

Generally, when we value something for which a price was not observed, we use the median per-unit value at the smallest (most local) geographic unit for which we have at least 10 observations of market prices (or, sometimes, respondent-estimated values). When the country-level median is used, we set no minimum number of observations. The imputed price is specific to a given item-unit combination (for example, a kg of sorghum, a chicken, a basket of fish). These imputed prices are relevant for estimating the value of crop production, the value of livestock (and livestock products) production, the value of fishing income, and crop costs per ha (see below). We are not able to impute values for item-unit combinations with no observed price.

Examples:

For indicator # A15, crop costs per hectare are estimated with explicit costs (observed purchases) and implicit costs (we separately report estimates that only include explicit costs). Where inputs of crop production are applied but were not purchased on the market (for example, family labor), we impute the value based on the local per-unit market price for each item. Again, we use the median per-unit value at the smallest (most local) geographic unit for which we have at least 10 observations of market prices. Thus, land that was used but not rented is valued at the local rental rate; seed/fertilizer inputs that had not been purchased are valued at their local market price; and family/exchange labor is valued at the local per-day agricultural wage. Wherever possible, labor is valued specific to gender and the adult/child distinction.

For indicator # B13, the average daily wage in agriculture is estimated without imputation of unobserved wage. When data are available for multiple seasons, the wage is the simple average across all seasons. For specific instruments (e.g. Ethiopia Baseline) with a small number of observations on hired labor (less than 30) in the other growing season, only the wage for the main growing season is considered.

Currency units

All indicators that are in monetary terms are calculated in the local currency, in U.S. dollars, and in Purchasing Power Parity (PPP) international dollars. These are labeled clearly in the data files.

The exchange rate used for each country reflects the January 1 exchange rate within the period of data collection. For all surveys, this is January 1, 2016. For the Tanzania LSMS-ISA (collected in 2014-15), local

currency values are first inflated to 2016 Tanzania shillings (using the CPI) before we apply the January 1, 2016 exchange rate.

The PPP adjustment rate reflects the 2016 values for each country reported by the World Bank (https://data.worldbank.org/indicator/PA.NUS.PPP). To convert to international dollars, the monetary value in local currency units is divided by the PPP conversion factor.

Plot sizes

Across all surveys, plot sizes are converted to hectares. In surveys where plots were never measured by GPS, we rely entirely on the respondents' estimates. In surveys where plots were sometimes measured by GPS, we use the measured values where available and the respondent estimates for unmeasured plots. In the Tanzania LSMS-ISA, where all plot areas are reported in acres, we refer to the farmer estimates where measures are missing. In the Ethiopia LSMS-ISA, where the units for plot areas are quite diverse, we use conversion factors provided with the data set when available and estimated the size of units missing from the conversion file by referring to the local median per-unit measured area to estimate the area of plots that were not themselves measured. In the Nigeria LSMS-ISA, where the units for plot areas are not as diverse, we use the respondent's estimates when the unit is acres or hectares, and apply a conversion factor provided with the data set when the unit is "heaps" or "ridges".

An alternative approach to measuring plot area for instruments where a sufficient number of GPS-measured and respondent-estimate pairs of area values are available would be to apply multiple imputation techniques.

Plot decision makers and gender productivity gap

Across all surveys, we categorized plot decision maker as "male-only" if all decision makers listed are male, as "female-only" if all decision makers listed are female, and as "mixed" if there both male and female among the decision makers. To estimate the gender productivity gap, we compare the average productivity of male-only managed plots to the average productivity of female-only managed plots.

Appendix: Tanzania TNPS Wave 1 example for understanding the implications of indicator construction choices

What follows is an example for Tanzania TNPS Wave 1 that replicates a few different indicator construction choices to get a sense of how important, or not, some of these choices might be. This set of estimates illustrates the difference between including one or two seasons, sex of plot decision maker, mixed and pure stand, and trimming choice:

- including or excluding the SRS (small difference)

- sex of plot decision maker (medium difference)

- type of stand (big difference)

- type of stand x sex of plot decision maker (big difference)

- trimming choice (winsorizing v MAD) (small difference)

- trimming choice (winsorizing v MAD) x sex of plot decision-maker (big difference, though only for Wave 1)

These differences are specific to this country/wave/crop and "small" and "big" are our non significance-tested views, but seem somewhat self-evident in the attached.

In the trade-off between two goals - (country) comparability and contextual validity - our priors are that the potential for misrepresenting women and the smallest smallholders rises with the comparability goal and the likely necessity of a blunter approach, with more consequences for sub-populations with fewer observations (and therefore all else equal, larger standard errors) and observations more likely to be in the tails. Ditto for small n commodities, nutritional indicators that vary a lot across country and instrument, etc.

Table 1: Maize Yields – Tanzania TNPS Wave 1 (2008/2009)

		Gender of Decision-Maker			Type of Stand	
	(1) All b/se	(2) Male b/se	(3) Female b/se	(4) Mixed b/se	(5) Pure b/se	(6) Mixed b/se
LRS + SRS						
Yield (kg/ha) - Winsorized	837	870	839	899	930	784
	(39)	(67)	(91)	(49)	(50)	(52)
Yield (kg/ha) - MAD	837	865	674	898	926	779
	(39)	(65)	(38)	(49)	(49)	(51)
Plots	2378	716	512	994	1446	902
Households	1457	472	354	626	975	640
LRS						
Yield (kg/ha) - Winsorized	815	828	840	894	921	757
	(39)	(66)	(102)	(51)	(53)	(51)
Yield (kg/ha) - MAD	812	822	662	893	920	752
	(39)	(64)	(40)	(51)	(53)	(51)
Plots	1909	623	423	851	1198	711
Households	1303	439	316	572	872	580

Table 1 and 2 notes: The decision-maker variable is constructed using the answers to the question "Who decided what to plant on this plot in the long rainy season 2008?" The decision-maker is coded as male only if all listed decision-makers are male. The variable is coded as female only if all listed decision-makers are female. The variable is otherwise coded as mixed.

Winsorized values are replaced at the 99th percentile. MAD values are constructed by replacing all values more than two standard deviations above the median with the median.

In the top panel statistics include both the long and short rainy seasons, with output aggregated over both seasons but only the largest area planted used as area. In other words, if a household plants one hectare of maize in the long rainy season but two hectares of maize in the short rainy season, two hectares is used as the area for construction of total yield. In the bottom panel, only the long rainy season area planted and output is included.

The decision-maker variables are defined at the plot level but the stand variables are defined at the plot-crop level. As such, some households are represented multiple times across variables. For example, some households have plots under both male decision-makers and female decision-makers. Other households also have both mixed and pure stand plots.

Area weights are used; constructed by multiplying the household weight by the area planted. These weights are constructed separately for each subsample.

Table 2: LRS+SRS Maize Yields – Tanzania TNPS Wave 1 (2008/2009)

Decision Maker:	Pure Stand			Mixed Stand		
	Male b/se	Female b/se	Mixed b/se	Male b/se	Female b/se	Mixed b/se
Yield (kg/ha) - Winsorized	1004	847	1008	804	960	828
	(96)	(97)	(70)	(77)	(197)	(67)
Yield (kg/ha) - MAD	870	706	1008	800	692	704
	(62)	(47)	(69)	(78)	(61)	(48)
Households	311	244	410	210	135	290

Estimates differ from Table 1 because of different overlapping of plots and area weights which are constructed separately for each subsample.

The green shaded cells in Table 1 highlight the difference between including both short and long rainy seasons compared to just reporting the long rains. In the case of Tanzania the difference is only around 5%. It may be more (and more gender differentiated) in other countries.

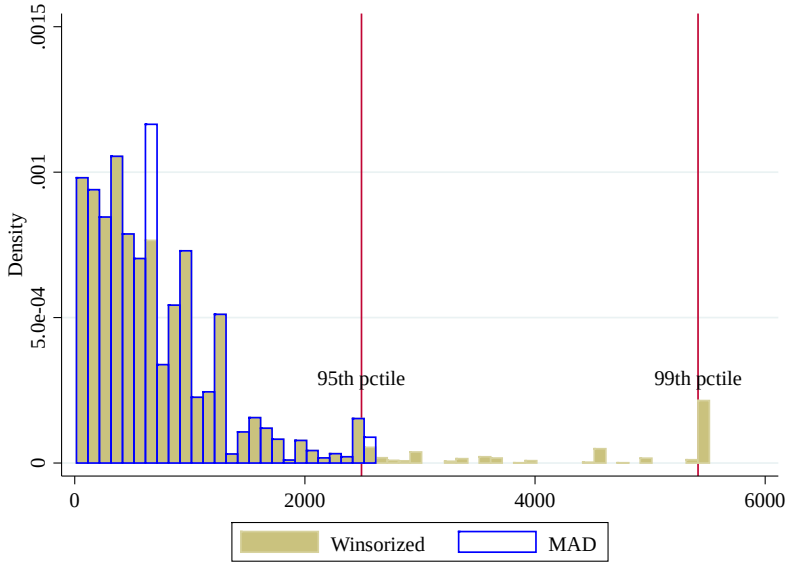
Blue shading in Table 1 illustrates the large difference in estimated yield between pure and mixed stand plots, regardless of trimming choice. Mixed stand yield estimates are about 17% smaller than pure stand yield estimates.

Blue shading in Table 2 illustrates the different story when pure and mixed plots are differentiated by gender. Sex-disaggregated, men's mixed stand plot yield is 20% lower than pure stand yield and women's mixed stand yield is 12% higher than pure stand (and higher than men's mixed stand).

Yellow shading in both tables shows the difference depending on the trimming choice. MAD trims 2 standard deviations from the median and replaces at the median, while Winsorizing trims and replaces at the tails, so the more skewed the data, the more the two will differ both because the trimming varies the more the median and mean do not coincide, and because replacing at the center rather than the tail has more implications.

For example, in Table 1 female yield estimates are lowered considerably more than men's when moving from Winsorizing to MAD. The histograms below indicate why. For female plots, when Winsorizing, all the higher yields to the right of the 99th percentile are trimmed and replaced at the 99th percentile (so they remain as above median observations, just not as extreme). With MAD, all the observations more than two standard deviations from the median (those not in blue) are trimmed and replaced at the median (the white central bar). Male plots have a smaller right-hand tail, so are less sensitive to the trimming choice.

Female Plots – Wave 1



Male Plots – Wave 1