Estimating Crop Yields:
Implications of Construction Decisions

EPAR Technical Brief #358

Pierre Biscaye, David Coomes, Jack Knauer, Josh Merfeld, C. Leigh Anderson & Travis Reynolds

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

September 30, 2017

Crop yield is one of the most commonly used partial factor productivity measures. It is used to estimate the ratio of quantity of crop output, generally measured in kilograms or tons, to a sole input, land area. "Common crop yield", defined as a simple ratio of crop production to area harvested, is a primary indicator of land productivity (or agricultural productivity more generally), but does not take into account any crop area lost prior to harvest (Fermont and Benson 2011).

An alternative to calculating yield based on the plot area harvested is to instead use the area sown with the crop at the time of planting. If area planted is a more accurate measure of land inputs than is area harvested, then yield by area planted will be a more valid proxy for land productivity than yield by area harvested. Recent studies on Tanzania have revealed that farmers harvested less area than they planted 48% of sorghum plots, compared to 43% of maize plots and 33% of rice plots across three waves of nationally-representative household survey data (Anderson et al. 2017, 2013). Yield measures that exclude this "null production area" may significantly overestimate land productivity, particularly for poorer farmers and those with smaller plots.

Ongoing EPAR research explores the policy implications of measuring yield by area planted versus area harvested. In this brief, we consider implications for crop yield estimates of other decisions in how to construct yield measures from household survey microdata. We use data from three waves of the Tanzania National Panel Survey (TNPS) and two waves of the Ethiopia Socioeconomic Survey (ESS), both part of the World Bank's Living Standards Measurement Study-Integrated Surveys on Agriculture (LSMS-ISA).

We look specifically at effects of construction decisions on estimates of maize yields by area planted, as maize is a commonly grown crop in both countries. We consider the effects of four decisions:

1. Whether to report yields separately by the gender(s) of the plot decision-maker(s);
2. Whether to report yields separately for pure-stand and mixed stand (intercropped) plots;
3. Whether to include production from all growing seasons in the numerator, or just from the main growing season (only analyzed for the TNPS, which separately reports on long and short rainy season agricultural production); and
4. Whether to treat outliers by winsorizing (replacing values above the 99th percentile with the 99th percentile value) or by Median Absolute Deviation (MAD) (replacing all values more than two standard deviations above the median with the median)

EPAR uses an innovative student-faculty team model to provide rigorous, applied research and analysis to international development stakeholders. Established in 2008, the EPAR model has since been emulated by other UW schools and programs to further enrich the international development community and enhance student learning.

Please direct comments or questions about this research to Principal Investigators Leigh Anderson and Travis Reynolds at eparinfo@uw.edu.

Results

The decision-maker variable for the Tanzania National Panel Survey (TNPS) is constructed using the answers to the question "Who decided what to plant on this plot in the long rainy season 2008?" For the Ethiopia Socioeconomic Survey (ESS), we use the question "Who in the household makes primary decisions concerning crops to be planted, input use, and the timing of cropping activities on this field?" 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; any larger values are replaced with the 99th percentile. Median Absolute Deviation (MAD) values are constructed by replacing all values more than two standard deviations above the median with the median.

For the TNPS, 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 is included.

A pure stand plot was planted exclusively with maize, while on a mixed stand plot maize was intercropped with another crop. On mixed stand plots, we rescale the area planted measures such that the total area planted to all crops on the plot is equal to the area of the plot. For example if a 2 ha plot is planted with maize on 2 ha and intercropped with beans on the same 2 ha, we assign each crop an area planted of 1 ha so that the total area planted sums to the plot area of 2 ha.

The decision-maker and stand variables are both defined at the plot 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, while some other households also have both mixed and pure stand plots.

All area measures are winsorized at the 99th percentile before yield measures are calculated, and all estimates are computed using area weights. Area weights are calculated by multiplying household weight and area planted (winsorized) for each sub-group. As such, the coefficients for each sub-group are nationally-representative estimates of maize yield by area planted for a given hectare for that sub-group.

Table 1. Maize Yield by Area Planted – TNPS Wave 1 (2008/2009)

		Gender of Decision-Maker			Type of Stand	
	(1)	(2)	(3)	(4)	(5)	(6)
	All	Male	Female	Mixed	Pure	Mixed
	b/se	b/se	b/se	b/se	b/se	b/se
LRS and SRS						
Yield (kg/ha) - Winsorized	845	910	920	936	1001	788
	(39)	(74)	(112)	(53)	(60)	(53)
Yield (kg/ha) - MAD	862	916	749	961	1009	811
	(43)	(77)	(46)	(63)	(62)	(63)
Households	1462	421	304	559	846	560
LRS						
Yield (kg/ha) - Winsorized	817	828	831	894	921	760
	(40)	(66)	(102)	(51)	(53)	(52)
Yield (kg/ha) - MAD	836	833	654	921	935	779
	(44)	(69)	(40)	(62)	(58)	(60)
Households	1304	439	317	572	872	581

Area weights are calculated by multiplying household weight and area planted (winsorized) for each sub-group. Standard errors are in parentheses.

Table 2. Maize Yield by Area Planted – TNPS Wave 2 (2010/2011)

		Gender of Decision-Maker			Type of Stand	
	(1)	(2)	(3)	(4)	(5)	(6)
	All	Male	Female	Mixed	Pure	Mixed
	b/se	b/se	b/se	b/se	b/se	b/se
LRS and SRS						
Yield (kg/ha) - Winsorized	859	829	759	900	991	708
	(39)	(56)	(62)	(52)	(50)	(46)
Yield (kg/ha) - MAD	777	797	707	791	927	659
	(28)	(54)	(46)	(38)	(45)	(36)
Households	1785	441	406	961	1222	809
LRS						
Yield (kg/ha) - Winsorized	822	831	757	867	951	690
	(37)	(56)	(53)	(52)	(45)	(51)
Yield (kg/ha) - MAD	756	792	734	773	893	638
	(28)	(52)	(50)	(37)	(40)	(39)
Households	1524	348	344	856	1048	684

Area weights are calculated by multiplying household weight and area planted (winsorized) for each sub-group. Standard errors are in parentheses.

Table 3. Maize Yield by Area Planted – TNPS Wave 3 (2012/2013)

		Gender of Decision-Maker			Type of Stand	
	(1)	(2)	(3)	(4)	(5)	(6)
	All	Male	Female	Mixed	Pure	Mixed
	b/se	b/se	b/se	b/se	b/se	b/se
LRS and SRS						
Yield (kg/ha) - Winsorized	825	826	729	860	860	788
	(36)	(56)	(47)	(50)	(41)	(55)
Yield (kg/ha) - MAD	790	816	678	825	811	756
	(32)	(56)	(46)	(46)	(37)	(51)
Households	2243	607	525	1141	1536	997
LRS						
Yield (kg/ha) - Winsorized	836	806	744	912	880	811
	(44)	(56)	(56)	(68)	(44)	(79)
Yield (kg/ha) - MAD	778	798	675	807	811	743
	(33)	(56)	(42)	(46)	(38)	(51)
Households	1937	518	441	999	1297	877

Area weights are calculated by multiplying household weight and area planted (winsorized) for each sub-group. Standard errors are in parentheses.

Table 4. Maize Yield by Area Planted – ESS Wave 2 (2013/2014)

		Gender of Decision-Maker			Type of Stand	
	(1)	(2)	(3)	(4)	(5)	(6)
	All	Male	Female	Mixed	Pure	Mixed
	b/se	b/se	b/se	b/se	b/se	b/se
Yield (kg/ha) - Winsorized	2630	3882	2807	2449	2708	2452
	(453)	(1694)	(842)	(319)	(560)	(544)
Yield (kg/ha) - MAD	2181	1966	2052	2321	2543	2174
	(378)	(348)	(364)	(273)	(539)	(320)
Households	1541	377	210	976	1079	736

Area weights are calculated by multiplying household weight and area planted (winsorized) for each sub-group. Standard errors are in parentheses. The ESS does not report on agricultural production in different seasons separately.

Table 5. Maize Yield by Area Planted – ESS Wave 3 (2015/2016)

		Gender of Decision-Maker			Type of Stand	
	(1)	(2)	(3)	(4)	(5)	(6)
	All	Male	Female	Mixed	Pure	Mixed
	b/se	b/se	b/se	b/se	b/se	b/se
Yield (kg/ha) - Winsorized	2130	2259	2260	2206	2080	2330
	(238)	(283)	(222)	(338)	(289)	(248)
Yield (kg/ha) - MAD	2426	2066	2230	2586	2044	2590
	(335)	(244)	(225)	(476)	(292)	(392)
Households	1565	392	218	993	1168	687

Area weights are calculated by multiplying household weight and area planted (winsorized) for each sub-group. Standard errors are in parentheses. The ESS does not report on agricultural production in different seasons separately.

Discussion

In Tanzania, we observe large differences between pure stand and mixed stand maize yields, with pure stand yields consistently higher. Yields for the LRS and SRS combined are, not surprisingly, consistently larger than yields for the LRS alone, since the former includes production from both seasons but only counts area planted once. Yield estimates are highest on plots managed by at least one male and female, compared to plots managed solely by men or solely by women. In Wave 1, estimates where outliers were treated by MAD are consistently larger than estimates where outliers were winsorized, with the exception of yields on female-only managed plots, which are much lower. In Waves 2 and 3, by contrast, the winsorized estimates are consistently larger, and we observe the lower yields on female-only managed plots for both the winsorized and MAD estimates.

We also observe larger estimates of maize yield for pure stand plots and for plots where outliers are winsorized in Ethiopia Wave 2. For Wave 3, however, yield estimates are generally greater for mixed stand plots and for plots where outliers are treated by MAD. We do not observe the same patterns in yield by gender of the plot decision-maker as in Tanzania, and the patterns differ depending on the method for dealing with outliers. In both waves, winsorized yield estimates indicate that plots managed solely by men have the highest yields, followed by plots managed solely by women. Estimates treated by MAD, however, indicate that plots with mixed gender managers have the highest yields, and that plots managed solely by men have the lowest yields.

These estimates help to illustrate some of the potential effects of decisions of how to construct and report yield estimates. Other potential decisions that may affect estimates include how plot area is calculated (e.g., what estimate to use if GPS measures of area are not available), how to calculate area planted by crop on intercropped plots, how to treat plots where the harvest was not complete at the time of the survey, how to estimate area planted for permanent or fruit crops, and how to deal with quantity harvested estimates in non-standard units. Interpretations of crop yield estimates should therefore pay particular attention to the decisions made in arriving at those estimates and the directions of bias potentially introduced by those decisions.

References

Anderson, C.L., Slakie, E., Reynolds, T., & Gugerty, M.K. (2013). Do Common Yield Measures Misrepresent Productivity Among Smallholder Farmers? A Plot-Level Analysis of Rice Yields in Tanzania. Evans School Policy Analysis & Research Group (EPAR) Brief No. 252. https://evans.uw.edu/sites/default/files/EPAR_UW_Request_252_TZNPS_Productivity_Measurement_Rice_010814.pdf

Anderson, C.L., Reynolds, T., & Biscaye, P. (2017). Proxy errors with policy consequences: How common yield measures bias estimates of agricultural productivity in Tanzania. Evans School Policy Analysis & Research Group (EPAR) Technical Report #303. Working paper available from authors upon request.

Fermont, A. and Benson, T. (2011). *Estimating yield of food crops grown by smallholder farmers*. Washington, D.C.: International Food Policy Research Institute.