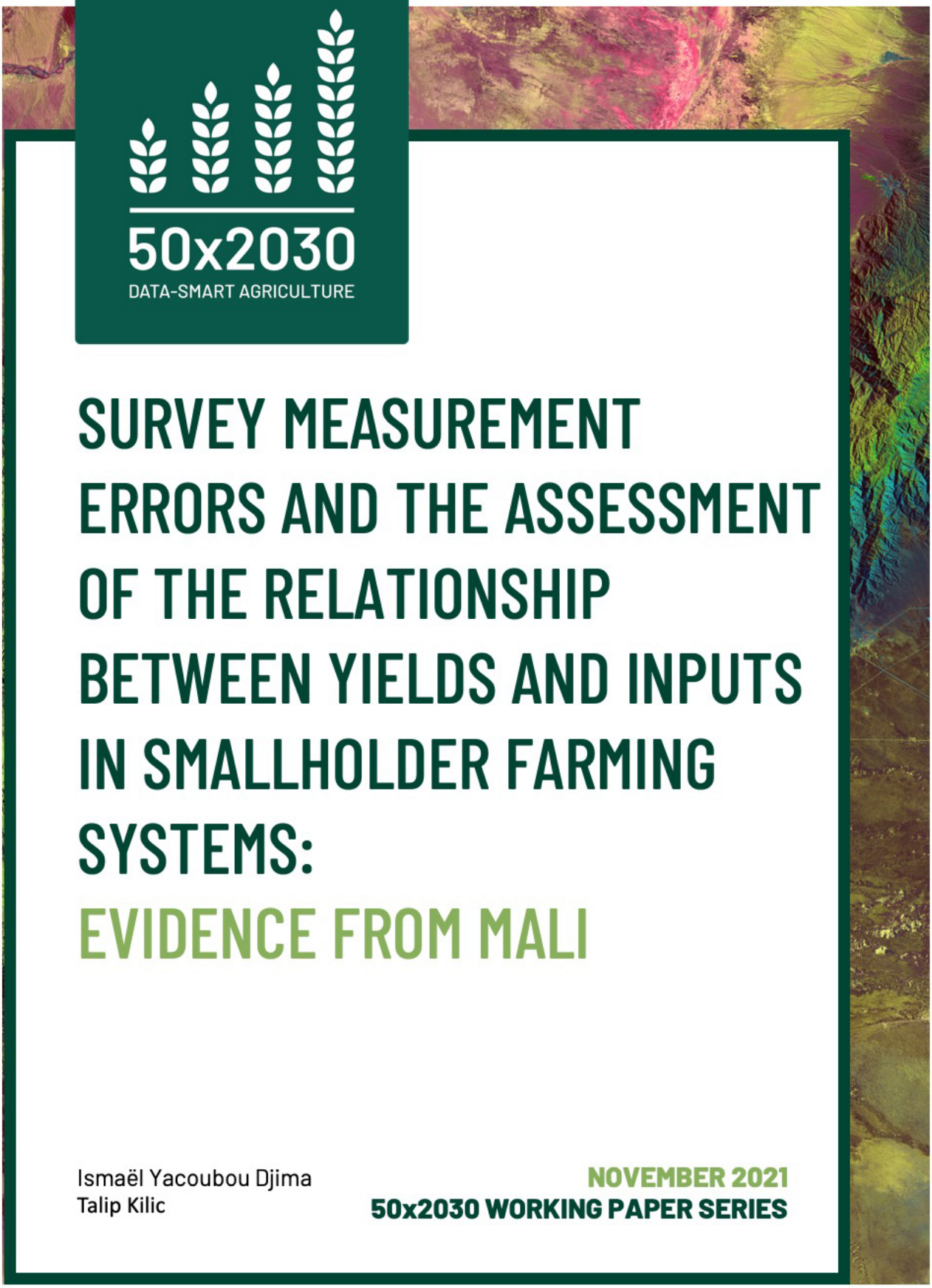


50x2030
DATA-SMART AGRICULTURE

SURVEY MEASUREMENT ERRORS AND THE ASSESSMENT OF THE RELATIONSHIP BETWEEN YIELDS AND INPUTS IN SMALLHOLDER FARMING SYSTEMS:

EVIDENCE FROM MALI

Ismaël Yacoubou Djima
Talip Kilic

NOVEMBER 2021
50x2030 WORKING PAPER SERIES

Survey Measurement Errors and the Assessment of the Relationship between Yields and Inputs in Smallholder Farming Systems: Evidence from Mali*

Ismaël Yacoubou Djima[†‡§]
ismael.yacouboudjima@psemail.eu

Talip Kilic[§]
tkilic@worldbank.org

Abstract

An accurate understanding of how input use affects agricultural productivity in smallholder farming systems is key to designing policies that can improve productivity, food security and living standards in rural areas. Studies examining the relationships between agricultural productivity and inputs typically rely on land productivity measures, such as crop yields, that are informed by self-reported survey data on crop production. This paper leverages unique survey data from Mali to demonstrate that self-reported crop yields, vis-à-vis (objective) crop cut yields, are subject to non-classical measurement error that biases the estimated returns to inputs, including land, labor, fertilizer, and seeds. The analysis validates an alternative approach to estimate the relationship between crop yields and inputs using large-scale surveys, namely a within-survey imputation exercise that derives predicted, otherwise unobserved, objective crop yields that stem from a machine learning model that is estimated with a random sub-sample of plots for which crop cutting and self-reported yields are both available. Using data from a methodological survey experiment and a nationally representative survey conducted in Mali, the analysis demonstrates that it is possible to obtain predicted objective sorghum yields with attenuated non-classical measurement error, resulting in a less-biased assessment of the relationship between yields and agricultural inputs. The discussion expands on the implications of the findings for (i) research on agricultural intensification, and (ii) the design of future surveys in which objective data collection could be limited to a sub-sample to save costs, with the intention to apply the suggested machine learning approach.

JEL Codes: C53, C83, Q12.

Keywords: Agricultural development; Smallholder farming; Agricultural inputs; Crop yields; Measurement error; Crop cutting; Machine learning; Household surveys; Mali; Sub-Saharan Africa.

*The authors would like to thank (i) Karen Macours, Sylvie Lambert, Luc Behaghel, David Rhys Bernard, Sydney Gourlay, and the participants of the World Bank Living Standards Measurement Study (LSMS) Internal Seminar and the Casual Friday Development Seminar at the Paris School of Economics for their comments on the earlier drafts, (ii) ICRISAT-Mali for the successful implementation of the ERIVaS methodological survey experiment, and (iii) Berenger Djoumessi Tiague for his research assistance. We are grateful for the funding from the Mali Mission of the United States Agency for International Development (USAID) for the implementation of the LSMS–Integrated Surveys on Agriculture (LSMS-ISA)-supported surveys in Mali, including ERIVaS. This paper was produced with the financial support from the World Bank LSMS Program (worldbank.org/lsms) and the 50x2030 Initiative to Close the Agricultural Data Gap (50x2030.org), a multi-partner program that seeks to bridge the global agricultural data gap by transforming data systems in 50 countries in Africa, Asia, the Middle East and Latin America by 2030.

[†]Corresponding author.

[‡]PhD Candidate, Paris School of Economics.

[§]Living Standards Measurement Study (LSMS), Development Data Group, World Bank.

1 Introduction

Small farms (with fewer than 2 ha) represent 84% of the estimated 570 million farms worldwide (Lowder et al., 2016). In view of the prevalence of smallholder farming across the developing world and the linkages between agricultural and welfare outcomes, one of the targets of the Sustainable Development Goal of Ending Hunger has been identified as doubling the agricultural productivity and incomes of small-scale food producers. One measure of agricultural productivity is crop yield, which is the amount of crop production harvested per unit of cultivated land. Increasing crop yields has been a long-standing goal of African governments and is key to feeding a growing global population while managing natural resources efficiently (Lobell et al., 2009; van Ittersum et al., 2013; Zhang et al., 2016).

Despite the seemingly simple formula for computing crop yields, and its frequent use in agricultural economics research, obtaining accurate measures of crop yields is underlined by a host of challenges in low- and middle-income countries, mainly due to reliance on household and farm surveys and measurement errors in self-reported information on crop production and cultivated area. Research has demonstrated the effects of these measurement errors on our understanding of the inverse-scale productivity relationship (IR) in agriculture – i.e. the finding that the productivity measures, including crop yields, are, on average, higher on smaller farms and plots vis-à-vis their larger counterparts.

The first strand of research has focused on the denominator in the formula for crop yields: cultivated land area, and has documented systematic discrepancies between GPS-based and self-reported plot areas (Carletto et al., 2013; Carletto et al., 2015; Kilic et al., 2017). The cross-country comparable finding across these studies is that on average, farmers tend to over-report plot areas, and more so at the lower end of the plot area distribution. As a result of using GPS-based plot area measures, as opposed to their self-reported counterparts, research has shown that the IR, though weakened, remained statistically significant.

The second strand of research has focused on the numerator in the formula for crop yields: total crop production, which, in low- and middle-income countries, has traditionally been based on self-reported survey data. There are, however, several complications in eliciting accurate self-reported information on crop production, including (i) potential recall biases, in part as a function of when data are collected vis-à-vis the harvest period (Wollburg et al., 2021), (ii) inclination to round off numbers (as shown by Gourlay et al. (2019) for crop production and by Carletto et al. (2015) for land area), and (iii) reliance on non standard measurement units, such as bunches and carts, and gaps in the availability and quality of conversion factors that are required to express non-standard measurement units in kilogram-equivalent terms and that are ideally differentiated for different conditions in which crop harvest may be reported by farmers (Oseni et al., 2017). Researchers have demonstrated that self-reported crop yields are subject to non-classical measurement errors, and are over-estimated on smaller plots, in comparison to objective measured yields based on crop cutting, i.e. the gold standard, albeit resource- and supervision-intensive, approach to crop yield estimation based on weighing the harvest of a randomly selected sub-section of a farmer's plot. The key cross-country finding of interest is that at the plot-level the IR is strong and statistically significant when self-reported crop yields are used, and that the IR ceases to exist when crop cutting yields are used instead (Desiere and Jolliffe, 2018 and Abay et al., 2019 for Ethiopia; Gourlay

et al., 2019 for Uganda).

Against this background, the first contribution of our paper is to provide empirical evidence from an alternative smallholder farming system in Mali on non-classical measurement errors (NCME) in self-reported (SR) crop yields vis-à-vis their crop cutting (CC) counterparts on the same plots and demonstrate the consequences of NCME in SR yields for the estimated relationships between crop yields and agricultural inputs besides land, specifically seeds and labor. The literature's heavy focus on land is understandable given that it is a primary factor of production. And our analysis fully corroborates the findings of Desiere and Jolliffe (2018), Gourlay et al. (2019), and Abay et al. (2019). However, with rising population densities and low levels of agricultural productivity, especially in Sub-Saharan Africa, more attention is being paid to the effects of measurement errors on the understanding of agricultural intensification (Abay et al., 2021) and of the extent of misallocation of inputs across farms (Gollin and Udry, 2021). In this context, understanding how crop yields relate to the use of other agricultural inputs is of great policy relevance.

To do so, however, it is necessary to ensure that the relationships are correctly assessed and are anchored in high quality survey data. Recognizing the technical and resource constraints that are faced by survey implementing agencies in contexts where data and yield improvements are needed the most, we present and validate an innovative approach to agricultural survey data collection and analysis in a way that requires the implementation of crop cutting only in a sub-sample of households/plots and that leverages machine learning techniques to derive predicted objective measures of crop yields for the sub-sample that is not subject to crop cutting. This constitutes the second contribution of our paper to the literature. In recent years, machine learning (ML) techniques have made their way into economics research and have proved their utility. Athey (2018), as part of an assessment of the early contributions of ML to economics, indicates that ML techniques have been most successful when applied to problems that require predictions as in our case. Furthermore, applying ML in agricultural economics research makes sense to model key variables such as crop yield, which is determined by a complex combination of soil quality, weather, input timing and other management choices, replete with non-linearities and interactions (Storm et al., 2020).

Our analysis showcases (a) the utility of ML techniques for predicting CC yields as a function of SR yields and complementary survey and geospatial data, and (b) the reliability of the resulting predicted CC yields in the analysis of the relationships between crop yields and agricultural inputs. We show analytically and empirically that predicted CC yields, although informed in part by SR yields, have attenuated NCME, which is uncorrelated with the intensity of agricultural input use. Only one-third of the sample is found to be sufficient for implementing crop cutting alongside self-reported survey data collection, such that an ML model can generate predicted CC yields for the rest of the sample in a way that leads to reliable parameter estimates of the relationships between yields and inputs. The findings reveal the importance of using SR yields as a predictor variable, and hence, underscores the need to continue to collect and improve the quality of self-reported survey data on crop production. Our analysis is first conducted with the data from a small-scale methodological survey experiment and is then replicated with the data from a nationally-representative survey to verify the external validity of

our conclusions. Beyond the methodological findings, our analysis has significant policy implications as it suggests that using self-reported yields, the returns of additional inputs per area of land are overstated and as such, it would indicate that agricultural intensification could have limited benefit in the context of the study.

The paper is organized as follows. Section 2 discusses the data and provides the descriptive statistics. Section 3 discusses how the choice of the yield measure affects the characterization of the relationship between land productivity and the agricultural inputs of interest, namely land area, seed quantity and household labor. Section 4 presents the machine learning framework and the results. Section 5 discusses the implications for survey data collection protocols. Section 6 concludes.

2 Data

2.1 Context

Our work is primarily informed by data collected during the 2017/2018 agricultural season as part of a methodological survey: the Enquête sur les Rendements et l'Identification des Variétés du Sorgho (ERIVaS), focused on sorghum yield measurement and variety identification.[1] The fieldwork was conducted in Dioïla district, which is located in the region of Koulikoro, in south central Mali. Sorghum is a major crop in Mali, planted on 26% of plots cultivated by agricultural households, which account for 90% of rural households representing 80% of the population of Mali.[2] Koulikoro is one of the main growing regions of this staple.

ERIVaS was set up to assess the relative accuracy of self-reporting on land areas, sorghum varieties, and sorghum yields vis-à-vis objective measurement methods, namely GPS for land area measurement, DNA fingerprinting for varietal identification, and crop cutting and satellite-based methods for crop yield estimation.[3]

2.2 Fieldwork details

Four 10x10 km areas (henceforth referred to as strata) were geospatially delineated in Dioïla to ensure heterogeneity in terms of relief and terrain type. In each stratum, 150 households were to be selected and the number of households selected in each of the 17 villages located across the four strata was proportional to the total number of households in the villages found during the listing exercise. Given the low incidence of inter-cropping in our study area, only households with pure stand sorghum plots were considered for data collection. In each sampled household, one pure stand sorghum plot was randomly selected among the pure stand sorghum plots cultivated by the household. The result of this sampling exercise was 600 sorghum pure stand plots (one in each) of 600 households distributed in 17

[1] Sorghum yield and varieties identification methodological survey. The survey was designed as part of the World Bank LSMS methodological research agenda on measurement of crop yields in household surveys.

[2] Author's calculation based on the Mali LSMS-ISA 2017 data.

[3] The analysis related to remote sensing measurement of yields and how the method compares to other methods is pursued in (Lobell et al., 2019). We compare part of the results obtained in both studies later in this analysis.

villages located in four strata across Dioïla. ERIVaS was conducted in three visits by semi-resident[4] enumerators.

During the first visit fieldwork, which was conducted in the post-planting period (September-November 2017), each household received a light household questionnaire that elicited basic socioeconomic information, and this was augmented with an agriculture questionnaire that collected detailed information on farm organization, land tenure, crop cultivation and seed use, at either parcel- or parcel-plot-level, depending on the topic. Following the random selection of the pure stand sorghum plot, the enumerator visited the plot location with the farmer, delineated the plot boundaries with the farmer and stored them on a handheld GPS unit, and set-up an 8x8m sub-plot for crop cutting, following the same protocol that was published by Gourlay et al. (2019) and that ensured the random placement of the sub-plot on each sampled plot. Each sub-plot was further divided into four 4x4m quadrants for which harvest was ultimately processed, weighed, and recorded separately.

During the second visit fieldwork, which was conducted immediately after the harvest (December 2017), the enumerators harvested the sub-plots with the help of the farmers and recorded the weight of the crop production immediately after the harvest and after additional drying at the enumerator's residence. During the third visit fieldwork, which was implemented during the post-harvest period (February 2018), each household received an agriculture questionnaire that collected self-reported, parcel-plot-level information on sorghum production and use of agricultural inputs, including fertilizer, household labor, and hired labor. Our analysis is based on 577[5] pure stand sorghum plots, with matching crop cutting and self-reported yield measures, augmented with plot- and household-level survey data and third-party geospatial data that are linked with georeferenced plot locations.

Intensive data quality controls and supervision measures were put in place to ensure that each crop cut sub-plot was placed in accordance with the desired protocol; was not managed by the farmer differently vis-à-vis the rest of the plot; was not harvested by the farmer without the presence and supervision of the enumerator; and was harvested to ensure that production could be weighed and recorded separately for each quadrant, without including plants outside each quadrant.[6] These supervision measures contribute to the substantial additional costs associated with crop cutting and provide further impetus for our efforts to get more out of crop cutting, self-reported survey data and cutting-edge imputation techniques to improve agricultural survey data even when crop cutting may not be implemented at full scale. On the other hand, intensive supervision measures are unlikely to eliminate biases in SR yields[7] and these biases may in turn be correlated with self-reported information on the very agricultural inputs,

[4]One field team covered each stratum and resided in the most amenable village of that stratum throughout the fieldwork

[5]Of the 600 plots selected, we drop 23 from this analysis for the following reasons: 18 crop cut subplots were harvested by farmers in the absence of the enumerators and as such had missing crop cut data, 3 plots had missing self-reported data due to a programming mistake, 1 plot had 100% self-reported loss but 0% enumerator assessed damage, and 1 plot was in the opposite situation: 100% enumerator assessed damage but 0% farmer reported loss.

[6]Appendix C.1 provides details of the fieldwork protocol followed for the crop-cut and the questionnaires fielded are available upon request.

[7]When we compare household production reported by farmers on their selected plots with the non-selected pure sorghum plots, there is no statistically significant difference, indicating that they do not try to be more accurate when reporting the production on the selected plots.

whose relationships with crop yields we are trying to assess.

2.3 Descriptive statistics

Table 1 presents summary statistics of the main plot and household level variables used in the analysis. The first part of the table displays plot characteristics. The variables extracted from the GPS coordinates show that most of the plots are located on similarly flat terrain within a 5 km radius of the household's dwelling. In terms of land tenure, 95% of the plots are reportedly owned by the households cultivating them. The plot managers of these plots are mostly middle-aged men with low literacy levels. The average size of the associated households was 19,[8] and households were headed overwhelmingly by middle-aged men most of whom were illiterate. The average number of cultivated plots is further indication that agriculture is the main activity for the households. Furthermore, the average GPS-based plot area is considerably larger than the available estimates observed in smallholder farming systems elsewhere in Africa. For example, Desiere and Jolliffe (2018), Gourlay et al. (2019), Abay et al. (2019) report average plot sizes ranging from 0.12 to 0.37 hectares in their sample. The incidence of agricultural input use was, on average, 33% for hired labor and 24% for fertilizer.

Table 2 reports estimates for plot-level for sorghum production (kilograms) and yields (kilograms per hectare).[9] The competing figures are based on plot-level self-reported sorghum production versus estimated sorghum production that is tied to the crop harvest for the 8m x 8m crop cut sub-plot and separately, a randomly chosen 4m x 4m quadrant within the sub-plot.[10] All yield measures use the GPS-based plot areas. As we do not find statistically significant differences between crop cut yields as a function of the size of the crop cut, in the rest of the analysis, CC yields are tied to the 8m x 8m sub-plot.

We see that the average difference between SR yields and each of CC yield is statistically significant at the 1% level, and on average, SR yields are 21% higher than their CC counterparts. The distributional differences between SR and CC yields are, however, not statistically significant. On the other hand, although there are statistically significant distributional differences between SR and CC sorghum production estimates, but we find no statistical difference in the means. The fact that we find significant differences between means but not distributions for yield measures and vice versa for production measures can in part be explained by the standard errors, which are relatively larger at the mean in the case of the production measures, making it difficult to detect differences for the point estimates. These differences are picked-up when we conduct the tests of distributional differences.

[8]The household size may appear to be large but it is typical of rural households in the region: based on 153 observations in the Mali LSMS-ISA 2017 data, the mean number of individuals in households with at least one sorghum plot in the district of Dioïla is 17, with a 95% confidence interval of [14.9, 19.6].

[9]These results are robust to the treatment of outliers in the SR variable via winsorizing. See Table A1 in Appendix A.

[10]Total CC-based sorghum production is computed by multiplying the 8x8m CC yield with the GPS-based plot area. The harvest measure for one of the four 4m x 4m quadrants is randomly chosen in each of the 8m x 8m subplots to compute the yield and harvest for the CC 4m x 4m method.

Table 1: Summary Statistics of selected variables in ERIVaS sample

Variable	Mean	Std.Dev	Min	Median	Max	Obs.
Plot Characteristics						
Slope (percent)	1.31	0.70	1.00	1.00	8.00	577
Elevation (m)	335.31	22.21	240.00	333.00	487.00	577
Potential Wetness Index	16.24	4.95	12.00	15.00	36.00	577
Plot has erosion problems†	0.20	0.40	0.00	0.00	1.00	577
Euclidean distance to dwelling (km)	2.11	2.01	0.00	1.66	18.39	577
Self-reported distance to market (km)	7.55	5.40	0.00	7.00	30.00	577
HH owns the plot†	0.95	0.21	0.00	1.00	1.00	577
Plot Manager Characteristics						
Plot manager is female	0.01	0.08	0.00	0.00	1.00	577
Age of plot manager	52.81	14.24	15.00	53.00	97.00	577
Plot manager is literate	0.10	0.30	0.00	0.00	1.00	577
Household Characteristics						
HH size	19.14	11.79	2.00	16.00	60.00	577
Dependency ratio	1.15	0.53	0.00	1.12	3.50	577
Head of HH is female	0.01	0.08	0.00	0.00	1.00	577
Age of head of HH	55.96	14.52	18.00	57.00	97.00	577
Head of HH is literate	0.08	0.28	0.00	0.00	1.00	577
Number of plots cultivated by HH	5.26	2.33	1.00	5.00	18.00	577
Number of Sorghum plots cultivated by HH	1.50	0.81	1.00	1.00	7.00	577
Number of Cotton plots cultivated by HH	1.34	0.82	0.00	1.00	7.00	577
Inputs Usage on plot						
GPS measured area (ha)	1.56	1.25	0.05	1.20	12.02	577
Quantity of Seeds used (kg)	8.30	5.90	0.56	6.56	51.02	577
Household labor on plot (persons-days)	159.30	267.37	8.36	76.40	2711.45	577
Hired labor†	0.33	0.47	0.00	0.00	1.00	577
Hired labor (persons-days)	15.23	180.21	0.00	0.00	4195.02	577
Organic fertilizer†	0.24	0.43	0.00	0.00	1.00	577
Mineral/Chemical fertilizer†	0.30	0.46	0.00	0.00	1.00	577
Mineral/Chemical fertilizer (kg)	25.74	59.85	0.00	0.00	750.00	577
Pesticides†	0.50	0.50	0.00	1.00	1.00	577
Yields (kg/ha)						
Self-reported	636.9	822.3	0.0	400.0	8139.5	577
Crop-cut 8m x 8m	500.7	380.1	0.0	423.4	2472.5	577
Crop-cut 4m x 4m	501.0	402.4	0.0	419.4	2690.0	577
Harvest (kg)						
Self-reported (plots without crop-cut)	738.7	1000.2	0.0	438.5	10000.0	305
Self-reported	837.3	1147.3	0.0	488.7	9773.3	577
Crop-cut 8m x 8m	794.4	959.8	0.0	515.9	9253.2	577
Crop-cut 4m x 4m	782.2	980.0	0.0	463.3	9312.2	577

Notes: Dependent ratio is the ratio of number of members in the household less than 15 and more than 70 years old over the members between 15 and 70 years old. HH refers to household. † denotes a dummy variable.

Table 2: Comparison of plot-level measures of sorghum yields and harvests in ERIVaS

	Yields (kg/ha)	**Harvest (kg)**
Means		
SR	636.9	837.3
CC 8m x 8m	500.7	794.4
CC 4m x 4m	501.0	782.2
Difference in Means		
SR vs CC 8m x 8m	136.2***	42.9
	(33.5)	(45.1)
SR vs CC 4m x 4m	135.9***	55.1
	(33.6)	(47.0)
CC 8m x 8m vs CC 4m x 4m	-0.3	12.2
	(6.1)	(11.0)
Difference in Distribution		
SR vs CC 8m x 8m		***
SR vs CC 4m x 4m	*	***
CC 8m x 8m vs CC 4m x 4m		

Notes: Standard Errors are shown in parentheses. Distributional differences are assessed using the Kolmogorov-Smirnov two-sample test. Statistical significance levels: *p<0.1; **p<0.05; ***p<0.01. Results with winsorizing of top 2.5 percent values of SR are shown in Table A1 in Appendix A

2.4 Correlates of measurement errors in self-reported vs. crop cut measure of yields

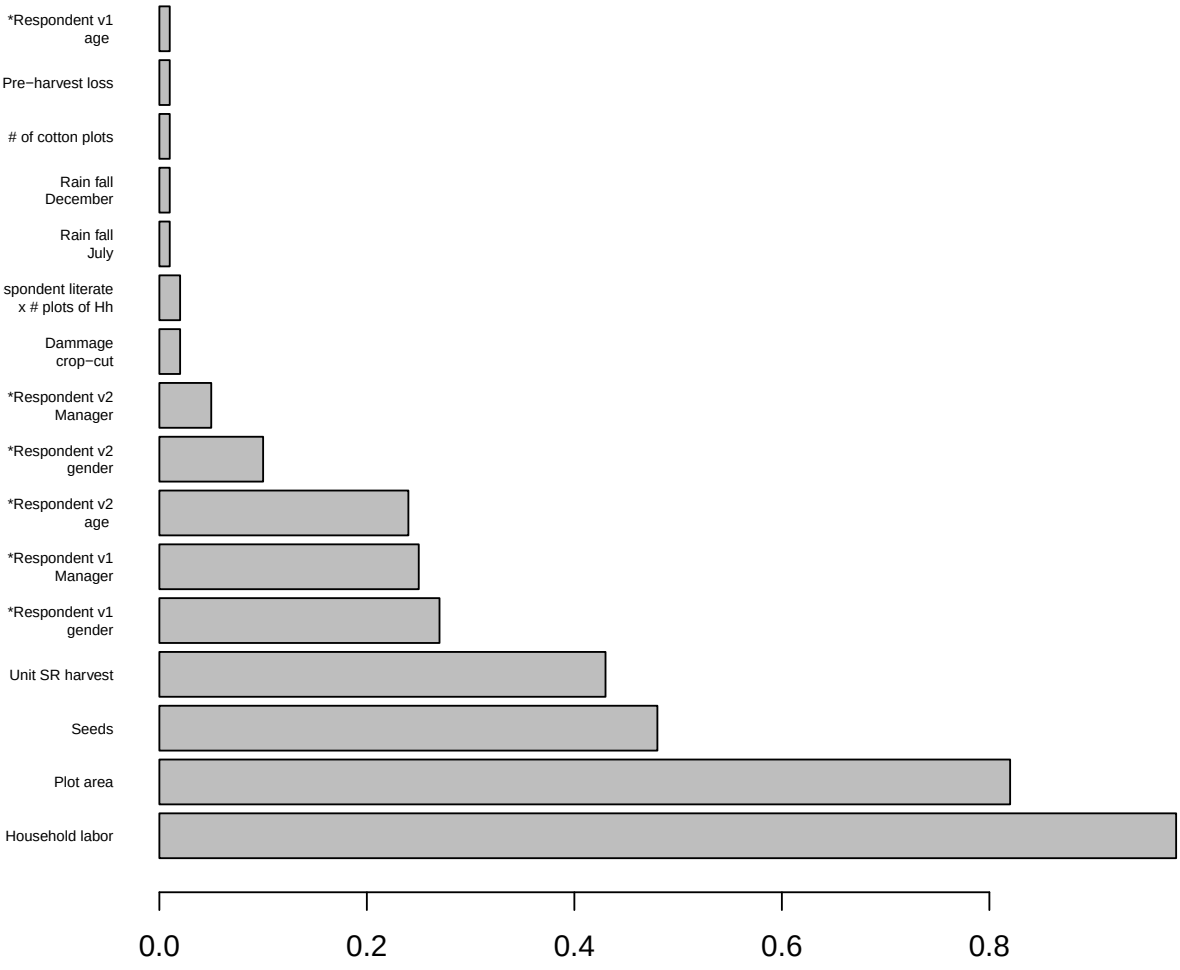
Next, we showcase the drivers of the measurement errors in plot-level SR yields vis-à-vis their CC counterparts. Given the myriad of covariates that can affect both yield measures, we use the ML algorithm LASSO to search through the pool of available covariates[11] and select the variables that are the most predictive of the difference: $[\log \text{SR} - \log \text{CC}]$. Figure 1 shows the variables selected at least once after 100 simulations and their probability of inclusion. The key variables in this list include household labor, plot area, and quantity of seeds. These are the three inputs[12] that we focus on for the rest of the analysis.

To better get a sense of the direction in which the selected covariates impact measurement errors in yields, we regress $[\log \text{SR} - \log \text{CC}]$ on the selected variables. This is akin to how (Abay et al., 2019) and (Gourlay et al., 2019) explore the correlates of measurement errors in production data. While their choice of covariates is based upon the authors' knowledge of the context and may appear arbitrary, our specification is more data driven and similar in spirit to an OLS-post LASSO estimation framework, which has the advantage of having a smaller bias even when the model is mis-specified (Belloni and Chernozhukov, 2013). The regression results are recorded in Table 3. The negative and statistically

[11]The covariates include factors of production that affect yields and field work related variables such as characteristics of respondents interacted with plot, households and enumerators dummies.

[12]It is striking that the dichotomous variables indicating the use of organic and chemical fertilizers do not appear as a key predictor of the difference of SR and CC yields. Per hectare use of chemical fertilizer may have carried a stronger signal but the quality of the data collected on this front was not satisfying so we did not further investigate the question for this input.

Figure 1: The graph shows variables with probability of inclusion greater than 0. 100 simulations are run on a training sample made of 80 percent of the ERIVaS sample. Labor, plot area and quantity of seeds are in log form. * indicates that the variable is interacted with enumerator fixed effect.



significant coefficient on plot area indicates a tendency on the part of farmers to over-estimate their yields on smaller plots. The extent of over-estimation in SR yields is greater on plots with higher intensity of agricultural input use.[13]

It is difficult to assert formally the mechanisms behind the correlates of the overestimation but knowing the context, it is possible to speculate on the potential explanations. For example, the number of cotton plots cultivated by the household is negatively associated with the farmers' tendency to overestimate yields. With the ERIVaS study area being in the Malian cotton belt, farmers with cotton plots have received advice and assistance from agriculture extension agents. They are, therefore, likely to have better grasp of the production potential of their plots and to not overestimate their yields. There are additional findings that may be useful to keep in mind for future survey design. For instance, the extent of overestimation is reduced when the harvest is reported in large non-standard units. This is in part explained by the fact that plots with large harvests are of large size and are subject to underestimation of SR yields. This finding also suggests that allowing for non-standard measurement units in data collection and using conversion factors to express them in kilogram-equivalent terms reduce the discrepancies between SR and CC yields. Another correlate of yield overestimation is the enumerator assessed damage on the crop cut, indicating that farmers underestimate the effect on yields of the various damages that their plots may have been subject to during the growing season. More attention should be paid to the further piloting and potential inclusion of related questions in future surveys. Finally, plot managers (also more likely to be the respondents) who are literate have less tendency to overestimate their production, even with a higher number of plots cultivated, underscoring the need to work with literate respondents.

The analysis of the correlates of $[\log \text{SR} - \log \text{CC}]$ provides evidence that there is NCME in SR yields and NCME is negatively correlated with plot size and positively correlated with seed and labor inputs per hectare. From a policy standpoint, this implies that the returns to input intensification will be overestimated if SR yields are used instead of their CC counterparts. In other words, the systematic measurement errors in SR yields may be driving up the relationships between yields and inputs, and these relationships may in fact be artifacts of the data. We investigate these assertions analytically in the next section.

3 Assessing the relationship between yields and inputs

3.1 Analytical framework

This section attempts to assess how the errors in SR yields, in comparison to CC yields, affect the characterization of the relationships between land productivity and agricultural inputs, namely land area, seeds and household labor. We start out with the analytical framework proposed by Abay et al. (2019). The estimated yield input relationship is:

[13] This assumes that measurement errors in inputs are not correlated with SR too. We come back to this assumption later in the paper.

Table 3: Regression results of errors in yields against ML selected correlates of measurement errors in yields

	[log SR − log CC]	
	OLS	IV
Household labor	0.191***	0.173***
	(0.040)	(0.043)
Plot area	−0.259***	−0.286***
	(0.092)	(0.096)
Seeds	0.148*	0.148*
	(0.087)	(0.082)
Unit SR harvest	−0.121***	−0.122***
	(0.038)	(0.038)
*Respondent v1 : female	1.433***	1.410***
	(0.424)	(0.431)
*Respondent v1 : manager	0.415	0.433
	(0.356)	(0.372)
*Respondent v2 : age	0.012***	0.012***
	(0.004)	(0.004)
*Respondent v2 : female	0.0003	0.0004
	(0.016)	(0.016)
*Respondent v2 : manager	−0.096	−0.095
	(0.525)	(0.534)
Dammage crop-cut	0.169	0.159
	(0.192)	(0.199)
Respondent literate x # plots of Hh	−0.063***	−0.063***
	(0.024)	(0.024)
Rain fall - July	0.005	0.005
	(0.004)	(0.004)
Rain fall - December	0.001	0.003
	(0.029)	(0.029)
Number of cotton plots	−0.121***	−0.132***
	(0.046)	(0.046)
Pre-harvest loss	−0.008***	−0.008***
	(0.002)	(0.002)
*Respondent v1 : age	0.012	0.011
	(0.015)	(0.015)
Enumerator Fixed effect?	Yes	Yes
Observations	577	577
R^2	0.301	0.302
Adjusted R^2	0.256	0.256

Notes: Labor, plot area and quantity of seeds are in log form. In the second column, household labor is instrumentalized with household size. * indicates that the variable is interacted with enumerator fixed effect. v1 indicates post-planting visit and v2 indicates post-harvest visit. Enumerator fixed effect included and errors clustered by enumerator. Significance levels: *p<0.1; **p<0.05; ***p<0.01

$$Y^* - A^* = \beta^*(X^* - A^*) + \epsilon \tag{1}$$

where ϵ is mean zero and uncorrelated with $(X^* - A^*)$ and β^* can be estimated via ordinary least squares (OLS) regression. We consider the following type of non-classical measurement errors parameters:

- δ: correlation of the measurement errors in the yield measure with true value of yield
- λ: correlation of the measurement errors in the yield measure with the true value of input
- α: correlation of the measurement errors in input with the true value of input
- π: correlation of the measurement errors in yield with the measurement errors in the input measures

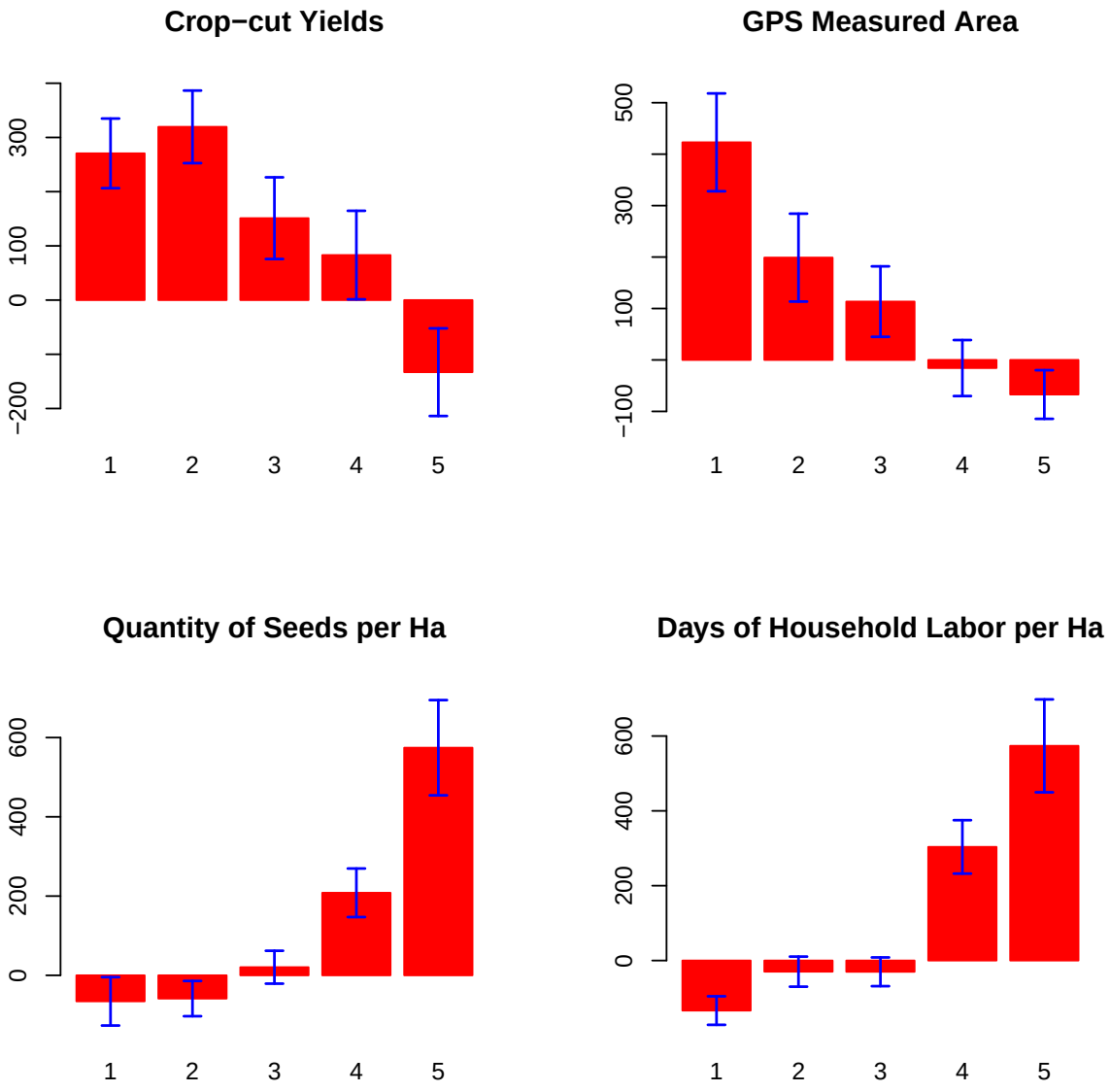
When present, δ appears as a multiplicative attenuation bias. The remaining parameters enter in a general formula for an OLS estimate of β^* as following:

$$\begin{aligned}\beta^{OLS} = &\frac{\beta^*(1+\alpha)\rho x_*^2}{(1+\alpha)^2\rho x_*^2 + \rho\ell^2} - \frac{\alpha(1+\alpha)\rho x_*^2}{(1+\alpha)^2\rho x_*^2 + \rho\ell^2} - \frac{\rho\ell^2}{(1+\alpha)^2\rho x_*^2 + \rho\ell^2} \\ &+ \frac{\lambda\rho x_*^2}{(1+\alpha)^2\rho x_*^2 + \rho\ell^2} + \frac{\pi}{(1+\alpha)^2\rho x_*^2 + \rho\ell^2}\end{aligned} \tag{2}$$

where ρx_*^2 is the variance of X^* and $\rho\ell^2$ the variance of the random component of the measurement error in X. The second and third term appear if and only if the input is present in the dependent and independent variables as in the case of plot area. To characterize the signs of the different parameters, we rely on the results of the regression presented in Table 3 and on Figure 2 which provides graphical evidence of the relationship between the difference of SR and CC yields – taken to be the measurement error in SR yields. If we assume that CC yields are the true value of yields, the difference between SR and CC is negatively correlated with the value of CC yields (see top left panel of Figure 2). This indicates that the sign of δ in our data is negative. The case of the measurement error encapsulated in the parameter δ is that of attenuation bias due to measurement error in the dependent variable.

Regarding α, we note that the correlation of measurement error in our measure of plot area and true plot size is likely to be negligible because of the accuracy of GPS handheld devices (Carletto et al., 2017) and the large plot size in Mali. For labor, we can rely on the comparisons of recall- (as in our survey) versus diary-based approaches to survey data collection on labor, as reported by Arthi et al. (2018), Gaddis et al. (2020), who present evidence of overestimation in recall-based plot-level labor hours in Tanzania and Ghana, respectively, vis-à-vis diary-based measures. As such, we expect a negative correlation between the measurement error in labor and its true value. Finally, characterizing α is most complicated for quantity of seeds for which we do not have a benchmark for the true value, so we remain agnostic. Turning to λ, for plot area, we use the GPS-based measure, which can be assumed reliably to have measurement errors that are negligibly correlated with the true measure. Top right panel of Figure 2 indicates that the relationship is negative, so we can take λ to be negative for plot area. Characterizing the sign of λ for labor and seed inputs is more difficult since they are based

Figure 2: Self-Reported yield overestimation of crop-cut yield over quintiles of crop-cut yields and key inputs. *Notes:* The y-axis represents the amount by which SR yields overestimate CC yields ($Y = SR - CC$) in kg/ha. The x-axis shows the quintile of the crop-cut yields or of the input per ha.



on farmer-reporting and may be subject to measurement errors that are correlated with measurement errors in yield. However, let's assume at first that this correlation is negligible. Having made this assumption, interpreting Figure 2 indicates that λ is positive for these two inputs.

The bottom left panel of Figure 2 shows that in the first two quintiles of the quantity of seeds per hectare, SR yields are slightly underestimated vis-a-vis CC yields. In the third quintile, there is no statistically significant difference between the two yield measures. In the last two quintiles however, the over- estimation in SR yields is large and clearly statistically different from zero. A similar trend is observed for household labor per hectare (bottom right panel): As the number of days of household labor per hectare increases, there is a larger extent of over-estimation in SR yields. Table 3 includes additional evidence of the positive relationship between the errors in SR yields and the true values of household labor and quantity of seeds. We observe positive and statistically significant signs for labor and seeds, controlling for other drivers of measurement errors in SR yields.

Concerning labor, we note that the respondent literacy interacted with the number of plots - both contributors to the measurement error in recall-based plot-level labor inputs, per Arthi et al. (2018) and Gaddis et al. (2020) is among the covariates listed in Table 3. As such, it can be argued that the regression at least partially controls for measurement errors in household labor and that λ, i.e. the correlation between the measurement error in yields and the true value of labor inputs, is positive. Using an instrumental variable approach, we can further augment the evidence base regarding the expected direction of λ for labor. In the second column of Table 3, we regress the difference in SR and CC yields on household labor days using household size as an instrumental variable for household labor. Household size appears as a good candidate instrument for this exercise. It is directly correlated with household labor inputs, and is easier to measure, and thus, could enable us to partial out the measurement error in household labor. Additionally, the machine learning search for the drivers of measurement error in yields tells us that household size is not a significant contributor to this error. After using household size as an instrumental variable, the coefficient for labor remains positive and large. This finding provides further support to the assumption that measurement errors in yields is positively correlated with the true value of household labor.

Finally, we attempt to characterize π the correlation of the measurement error of yields and the measurement error in inputs. In the case of plot area, considering that the measurement error in GPS-based plot area is negligible, the correlation with the measurement error in yield will also be negligible. For labor, Table 3 provides some sense of the sign of π. We can see that the correlation between household labor and the difference between SR and CC yields is higher using a simple OLS versus an instrumental variable regression. Hence, there seems to be a positive relationship between measurement error in yields and measurement error in labor. At the same time, the coefficients across the two regressions are not statistically significantly different as such π, as it is difficult to find an instrument that allows us to partial out the measurement errors in the quantity of seeds.

Table 4 provides a summary of the conjectures put forth regarding the parameters δ, λ, α, and π. For all inputs, δ is negative so it it exerts a negative bias on β^{OLS} but as noted by Abay et al. (2019), this is a multiplicative bias so it cannot reverse the sign of the relationship. The other parameters

however appear as additive terms. In the case of plot area, the most contributing parameter is λ which is negative. As such, we expect a negative bias, as also documented in previous studies. For labor, λ contributes positively to the bias, π is positive and appears with a positive sign in the formula for $\beta^{0\text{LS}}$; and α but will contribute positively to the bias since it will make the denominator of all terms in Equation 2 smaller, and, therefore, the multiplicative bias in the first term will go up (assuming $|\alpha| < 1$), and the bias from λ and π will grow larger. Hence, the overall bias in the coefficient $\beta^{0\text{LS}}$ for labor is expected to be positive. Finally, for quantity of seeds, we are able to only confidently sign λ, which would contribute positively to the bias, but based on the descriptive evidence, we expect that the final bias will also be positive, similar to labor. We expect that this bias is essentially going to be driven by λ i.e. the correlation of measurement error in yield with the true value of quantity of seeds. In sum, we expect that using SR yields in estimating the relationships between yields and inputs will result in underestimation of the returns to area and an overestimation of the returns to labor and quantity of seeds. The next section provides an empirical assessment of this assertion.

Table 4: Characterization of the NCME parameters in the assessment of yields - input relationship.

Input	δ	λ	α	π	Bias
Area	≤ 0	≤ 0	$\simeq 0$	$\simeq 0$	Negative
Labor	≤ 0	≥ 0	≤ 0	≥ 0	most likely positive
Quantity of seeds	≤ 0	≥ 0	-	-	most likely positive

3.2 Empirical results

This section presents estimates of the yield - inputs relationship. We use simple bivariate regressions as presented in the earlier section. The equation for these regressions can be written as follows:

$$Y_i = \alpha + \beta_1 A_i + \epsilon_i$$

$$Y_i = \alpha + \beta_2 S_i + \epsilon_i$$

$$Y_i = \alpha + \beta_3 L_i + \epsilon_i$$

We also estimate multivariate OLS regressions as is commonly done in the IR literature. Estimating a multivariate regression allows us to alleviate omitted variables bias concerns that could be raised in explaining the differences between the coefficients across the regressions of SR versus CC yields Desiere and Jolliffe (2018). The multivariate model is of the form:

$$Y_i = \alpha + \beta_1 A_i + \beta_2 S_i + \beta_3 L_i + \delta P_i + \sigma O_i + \tau Hi + \mu M_i + \varphi V_i + \epsilon_i$$

In both the bivariate and the multivariate equations above, Y is the logarithmic transformation of the plot-level sorghum yield (kilograms per hectare), based on GPS-based plot area. i denotes plot (recall that in ERIVaS, we select one plot per household), A, S, L are respectively the logarithmic transformation of GPS-based plot area, the logarithmic transportation of the quantity of seeds used per

hectare, the logarithmic transformation of the number of days of household labor per hectare. P is a vector of plot characteristics including inputs other than seeds and household labor, O is a vector of objective measures of plot characteristics (including geospatial variables). H is a vector of household characteristics, M is a vector of plot manager characteristics. V is a vector of dummy variables for villages (village fixed effects) and ϵ_i the error term.

The coefficients of interest are the β's. A negative and statistically significant β_1 would be in support of the inverse scale-productivity relationship at the plot-level. A positive and significant β_2 would indicate that increasing the quantity of seeds per hectare is correlated with higher yields, suggesting that one the potential way to increase yield is to increase the quantity of seeds per hectare. Finally a positive and significant β_3 would indicate that more labor per hectare could potentially increase yields. From a policy perspective, a negative β_1, and positive β_2 and β_3 plead in favor of agriculture intensification to improve yields.

Our empirical analysis confirms our expectations, as presented in the preceding section. Table 5 presents the regressions coefficients of key agricultural inputs, as obtained from the bivariate and multivariate regressions of SR and CC yields. The IR is always confirmed when using SR yields, with a statistically significant negative coefficient associated with plot area. More specifically, it appears that a 1 percent increase in GPS-based plot area results in a .2 percent reduction in sorghum yield. The relationship disappears, however, when using CC yields.

Table 5: Regressions results of SR and CC measures of yields for selected inputs

	Measures of Yields							
	SR	CC	SR	CC	SR	CC	SR	CC
GPS measured plot area	−0.268**	0.138					−0.304***	−0.045
	(0.109)	(0.122)					(0.085)	(0.099)
Quantity of Seeds			0.574***	0.051			0.366***	0.185**
			(0.128)	(0.139)			(0.068)	(0.086)
Days of household Labor					0.417***	0.078**	0.159***	−0.040
					(0.058)	(0.033)	(0.058)	(0.045)
Multivariate Framework?	No	No	No	No	No	No	Yes	Yes
Village Fixed effects?	No	No	No	No	No	No	Yes	Yes
Observations	577	577	577	577	577	577	577	577
R^2	0.033	0.009	0.081	0.001	0.144	0.005	0.443	0.424
Adjusted R^2	0.031	0.007	0.079	−0.001	0.143	0.003	0.391	0.371

Notes: Errors clustered at the enumeration area. Dependent and Independent variables shown here are log transformed. Full multivariate regression results are shown in Table A3 Significance levels: *p<0.1; **p<0.05; ***p<0.01

In sum, assuming that CC yields do not suffer from NCME, the errors in SR yields create spurious relationships between yields and key inputs in the context of smallholder farmers. The policy implications of these results are very concerning if agricultural initiatives are to be motivated by the analysis of these relationships. In fact, examining the full regression results in Table A3, it is striking how sensitive our estimations are to whether we use SR or CC yields. Using SR yields, plots with more intensive input use have higher yields, suggesting that the focus should be on agricultural intensification

and a broader access to or provision of seeds. Although the relation is still positive when using CC measure, the magnitude is considerably different - doubling the seeding rate would increase SR yields by 37 percent while it would only increase CC yields by 19 percent. For household labor inputs, the analysis based on SR yields points to the yield gains associated with increased labor inputs per hectare, suggesting a certain level of misallocation of labor across households. This suggestion, however, does not emerge if the analysis is anchored in CC yields.

These results emphasize the need for adopting objective measurement of production, specifically crop cutting, in household and farm surveys. The adoption of these methods would undoubtedly require additional technical and financial resources, and in the case of national surveys, could present considerable logistical challenges. In these circumstances, it is worth investigating whether crop cutting can be implemented in a sub-sample, with the intention to build an imputation model of CC yields as a function of easier-and-cheaper-to-collect self-reported and geospatial data, and to obtain, in the rest of the sample that is not subject to crop cutting, predicted CC yields that are subject to attenuated NCME and that allow for a less biased assessment of the relationships between yields and inputs. This is what is attempted in the second part of this analysis.

4 Estimating the relationships between yields and inputs using machine learning-based imputed crop cut yields

This section provides an analytical framework on how the production function coefficients associated with agricultural inputs are expected to be modified as a result of using machine learning-based imputed crop cut yields (henceforth referred to as ML yields) vis-à-vis CC and SR yields. This is followed by the presentation of the empirical results from the regressions of SR, CC and ML yields, verifying the predictions stemming from the analytical framework.

4.1 Analytical framework

Recall that we are interested in estimating:

$$Y^* - A^* = \beta^*(X^* - A^*) + \epsilon$$

Where Y^* is the log of production measure and X^* is the log of any explanatory variable of concern (labor, quantity of seeds used) and A^* is area. We do not observe the true value of Y^* nor X^* and A^* but we set aside measurement errors in X and A and focus on the measures of Y^*. We observe in the data Y^{CC}, which is obtained from the enumerator weighted harvest on subplot crop cut and Y^{SR} which is obtained from the farmer self-reported production. We can write both measures in terms of the true Y^* as :

$$Y^{CC} - A^{*CC} = Y^* - A^* + u^{CC} \tag{3}$$

$$Y^{SR} - A^* = Y^* - A^* + u^{SR} \tag{4}$$

Where A^{*CC} is the size of the crop cut subplot. In the case of Y^{CC}, u^{CC} is uncorrelated with X^* (the crop cut harvested weight is on a randomly chosen pre-delimited subplot, so there are no reasons to think that it will have measurement errors correlated with the measure of labor or quantity of seed, area). In that case, the measurement error in Y^{CC} would not contribute to a bias in β^*.
Turning to production obtained through self-reporting Y^{SR}, we documented earlier that u^{SR} is subject to non-classical measurement error. Thus, we take $u^{SR} = \lambda(X^* - A^*) + \zeta$ and expanding and simplifying the equation leads to $\beta^{OLS} = \beta^* + \lambda$.

We introduce $Y^{ML} - A^{*CC}$ which is yield predicted using a machine learning algorithm. For simplicity in derivation, we consider that $Y^{SR} - A^*$ is a linear term and express $Y^{ML} - A^{*CC}$ as:

$$\begin{aligned} Y^{ML} - A^{*CC} &= \gamma(Y^{SR} - A^*) + F(\mathbf{W}) + w \\ &= \gamma(Y^* - A^* + u^{SR}) + F(\mathbf{W}) + w \end{aligned}$$

Where $\mathbf{W}$ is a vector of covariates that is used to predict yield not including X^*. So we end up with:

$$Y^{ML} - A^{*CC} = \gamma(Y^* - A^*) + F(\mathbf{W}) + (\gamma u^{SR} + w) \tag{5}$$

Estimating Equation (1) using $Y^{ML} - A^{*CC}$ and expanding with Equation (5), we get that:

$$\begin{aligned} \beta^{OLS} &= \frac{\text{Cov}(Y^{ML} - A^{*CC}, X^* - A^*)}{\text{Var}(X^* - A^*)} \\ &= \frac{\text{Cov}(\gamma(Y^* - A^*) + F(\mathbf{W}) + (\gamma u^{SR} + w), X^* - A^*)}{\text{Var}(X^* - A^*)} \\ &= \frac{\text{Cov}(\gamma(Y^* - A^*) + F(\mathbf{W}) + (\gamma(\lambda(X^* - A^*) + \zeta) + w), X^* - A^*)}{\text{Var}(X^* - A^*)} \end{aligned} \tag{6}$$

By linearity, we can split the last line of the precedent equation in 4 terms (each on one line in the equation below):

$$\begin{aligned} \beta^{OLS} = &\frac{\text{Cov}(\gamma(Y^* - A^*), X^* - A^*)}{\text{Var}(X^* - A^*)} + \\ &\frac{\text{Cov}(F(\mathbf{W}), X^* - A^*)}{\text{Var}(X^* - A^*)} + \\ &\frac{\text{Cov}(\gamma\lambda(X^* - A^*), X^* - A^*)}{\text{Var}(X^* - A^*)} + \\ &\frac{\text{Cov}(\lambda\zeta + w, X^* - A^*)}{\text{Var}(X^* - A^*)} \end{aligned} \tag{7}$$

The first term simplifies to $\gamma\beta^*$. The second term can be written as $\eta\beta^*$ since $F(\mathbf{W})$ is a predicted measure of Y^* uncorrelated with $(X^* - A^*)$, we can assume that it will have some noise bringing an attenuation bias η in the estimation of the OLS coefficient. The third term simplifies to $\gamma\lambda$. The last term $\text{Cov}(\lambda\zeta + w, X^* - A^*)$ goes to 0 since λ is a constant, and ζ assumed to be uncorrelated to $X^* - A^*$. Similarly, since $\mathbf{W}$ does not include input variables, w is not correlated to $X^* - A^*$. So we end up with:

$$\beta^{OLS} = (\gamma + \eta)\beta^* + \gamma\lambda \tag{8}$$

According to Equation (8), using a predicted measure of crop cut, we will be estimating β^* with an attenuation bias coming from the noise of the ML algorithm which will also contribute to attenuate the bias induced by the SR measure of yield.

4.2 Empirical application

To test empirically whether it is judicious to use machine learning techniques to construct an alternate measure of crop yields based on a sub-sample of plot observations for which both crop cut and self-reported crop production data are available, we use the data from the ERIVaS 2017 experiment, which allows us to compare ML-predicted crop cut yields to their observed counterparts. Using data from a more controlled methodological study also ensures that the crop cut data are of the highest quality. A potential disadvantage, on the other hand, is that because the data are from a localized setting, it may be less heterogeneous, which could make it harder to accurately predict each observation. Our experimental protocol is as follows. First, we randomly select one-third of the observations in each ERIVaS village to construct a pooled data set to train the ML algorithm. The reason why we choose one-third of the observations to be part of the training data set is to mimic the extent of crop cutting conducted in the nationally-representative multi-topic household surveys that have been supported by the World Bank LSMS-ISA program. The remaining two-thirds of the data set serves as our test sample in which we compare the predicted yield measures obtained with ML against those based on farmer self-reporting and crop cutting, with the latter assumed to be closest to the "truth". The statistics tied to this comparative assessment are computed over 100 simulations of training and test sample splits to ensure that are conclusions are not the result of one random realization, although in some parts of this subsection, we may present the results for one realization for better illustration purposes. Second, since our goal is to investigate whether the ML-based predicted crop yields are suitable for the analysis of the relationship between yields and key inputs, we estimate the relationships of interest in the test sample using the most "accurate" modeling set up and compare the resulting regression coefficients to the ones that are estimated with SR and CC yield measures.

4.2.1 Machine learning modeling

We use an ensemble of methods[14] (LASSO, elastic-net, random forest, and gradient boosted trees)— a standard and popular machine learning technique— to generate yields predictions. We test the sensitivity of our estimation to the choice of the dependent variable (CC yields, CC harvest, in log or raw forms), and whether or not the predictors include SR yields. While the full technical details of the machine learning implementation are provided in section B of the Appendix, we discuss here the choice of the covariates considered for ML modeling.

On the whole, we fed the ML algorithms a set of 270 covariates across the following categories: (i) plot characteristics that exclude key plot inputs[15] but that include predictors related to land tenure, soil

[14]The technique is implemented using the R package SuperLearner (Polley et al., 2018).

[15]Since we will regress yields obtained using ML on key inputs, we do not include these variables in the set of ML covariates to ensure that the regression coefficients are not a mere mechanical result of the ML prediction model.

type, soil fertility measures, GPS-based Euclidian distance to the dwelling, (ii) plot-level agroclimatic and rainfall variables that are extracted based on georeferenced plot locations, (iii) plot-level agricultural practices, such as sowing dates and type of labor inputs, interacted with rainfall measures, (iv) household attributes, such as agricultural implement index, wealth index, and a series of household demographic attributes as well as socioeconomic characteristics of the head of household, and (v) other contextual information, including respondent attributes, day of visits to the households, and damage on crop cut sub-plots (as observed by the enumerators).

It is generally difficult to interpret the results of ML modeling, given the complexity of the algorithms, but we can get an insight on how the prediction of yields is made by examining the variables kept by the shrinkage algorithms or the importance of the random trees or forest algorithms. For illustration, Table 6 shows the result of the ML ensemble fit for the first sample split simulation when we fit the log CC and include log SR among the covariates. The fit that minimized the root mean squared error after cross-validation in the training sample was made with a weighted linear combination of a fit by LASSO, random forest and gradient boosting trees. The variables selected by LASSO are the log of SR and the enumerator assessed damage on the crop-cut. This makes sense as we expect SR yield to be a strong (although biased) proxy of CC yield. It is not surprising either that enumerator assessed damage on the crop cut appears as an important predictor since it allows to classify plots for which the harvest was completely lost. While these two variables are also important covariates for the random forest and gradient boosting trees algorithms, we also observe that farmer management practices, the interactions of input timing and amount of rainfall, and farmer access to agricultural equipment appear as top predictors of yields. Hence, the imputation approach allows the analyst to leverage the predictive power of SR yields, which has correlated measurement error, together with the predictive power of other determinants of yields, which suffer from less measurement error, to obtain an imputed measure of CC yields that would therefore have less NCME. We now turn to the assessment of the accuracy of the ML measure of yields.

4.2.2 Accuracy of the ML measure yields

To understand which ML modeling set up gives the most accurate measure of yields, we compare the different ML predictions to CC yields in the test set using the adjusted R^2 and the coefficient of correlation (ρ) between CC yields and the ML yields. To ensure that the assessment is not only the reflection of one random draw of the sample, we conduct this assessment with 100 different sample splits. The results are recorded in the top panel of Table 7. A couple of points emerge from examining this section of the table. First, predicting yields provides better results over predicting harvest. This makes sense since the dependent variable, CC is a method conceived to measure yields in the first place. Second, adding SR among the covariates considerably improves the accuracy of the ML predictions, indicating that it is a useful proxy of CC in our model. Third, it is better to estimate the model with the log-transformed dependent variable even when observed and predicted yields are compared in level terms. Therefore, we conclude that our best modeling set up is to fit the log-transformed CC yields including SR measure in the covariates.

Table 6: Machine Learning algorithm ensemble fit for the first sample split simulation. For illustration, we show the linear fit coefficients or predictive importance of selected variables for each ML algorithms chosen in the ensemble fit.

	LASSO	Random forest	Gradient boosting trees
log SR yield	0.190	35.628	0.132
Enumerator reported damage on the crop-cut	−0.011	22.164	0.0001
Farmer reported percentage of crop loss on plot	0	9.744	0.0001
Agricultural implementation index	0	8.963	0.002
Village dummy	0	5.701	0
Plot acquisition type	0	0.414	0.085
Sowing date 1st half June x rain fall 1st half June	0	0.269	0.131
Sowing date 1st half June x rain fall 2nd half May	0	0.158	0.132
Plot is owned by household	0	0.103	0.131

Notes: The ML model set up used log CC yield as the dependent variable and the covariates included log SR yield. The ML ensemble fit for the first sample split simulation comprised the predictions of LASSO (weight = 0.594), random forest (weight = 0.273), and gradient boosting trees (weight = 0.133). We show the variables selected (with non-zero coefficient) by LASSO and the top 5 most important variables in terms of predictive power for random forest and gradient boosting trees.

To put the accuracy of our ML measure in perspective in context, we also show the R^2 and ρ for SR and crop cut-calibrated satellite yields, which we obtained from Lobell et al. (2019). These results are consigned in the bottom panel of Table 7. Our ML-based imputed CC yields, in comparison to SR yields, explains unambiguously more of the variation of in CC yields, and even more impressive finding is that the R^2 and ρ of the are on-par with that of the CC-calibrated satellite yields.

4.2.3 Empirical assessment of the yields and inputs relationship using predicted yields

We replicate bi-variate and multivariate regressions of yields as a function of key inputs, with competing measures of yields, including SR, CC and ML yields. We use our preferred model specification of ML in the first sample split simulation in the test sample. Since we are working with predicted measures, the standards errors for the regression coefficients are bootstrapped. Further, as a robustness check, we also use the method introduced by Chernozhukov et al. (2018) to compute de-biased regression coefficients and conservative standard errors when the number of control parameters is large as it is the case of the relationships of interest between yields and key inputs.

The regression results are reported in Table 8. We note that in the test sample, the observations that we previously made concerning the difference between the coefficients of the inputs when using SR or CC are still valid. For example, in the multivariate case, the quantity of seeds per hectare is positively correlated with yields but carry a stronger coefficient when the yield measure is SR, as opposed to CC. The regressions based on ML yields yield new findings. The coefficients obtained with ML yields are similar to the ones obtained with CC yields and the coefficient pairs are not statistically distinguishable, irrespective of the agricultural input in question. This is better seen in the bottom panels of Figure

Table 7: Performance of different ML modeling set-ups for predicting log yields in the test set over 100 different training - test sample splits.

		log Yields	
		R^2	ρ
Dependent Variable	Covariates include SR		
log Yields	Yes	0.401	0.633
		[0.392, 0.409]	[0.627, 0.640]
log Yields	No	0.307	0.554
		[0.298, 0.316]	[0.546, 0.562]
log Harvest	Yes	0.242	0.492
		[0.233, 0.251]	[0.483, 0.501]
log Harvest	No	0.106	0.324
		[0.100, 0.113]	[0.313, 0.336]
SR		0.294	0.543
		[0.286, 0.301]	[0.535, 0.550]
Satellite Yields		0.135	0.370
		[0.131, 0.140]	[0.364, 0.376]

Notes: R^2 is the adjusted R^2 of the regression of CC and the corresponding measure of yields. ρ is the correlation between CC and the measure of corresponding measure of yields. We report the average R^2 and ρ in the test set (N = 386) of 100 test sample splits and the 95% confidence interval in brackets.

3 in which we plot the coefficients for area, quantity of seeds per hectare, and labor per hectare with their 95 percent confidence interval, as obtained from the regressions of SR, CC and ML yields. It is apparent from the figure that the coefficients from the regressions of CC and ML yields are statistically equivalent. The coefficients obtained using the method put forth in (Chernozhukov et al., 2018) also provide estimates that are equivalent to the bootstrapped ones (see bottom left panel in magenta).

Furthermore, repeating the estimation of the multivariate yield-inputs regression for the 100 sample split simulations, we find that for all the simulations, the ML yield regression coefficient for plot area is within the 95 percent confidence interval of the comparable coefficient from the CC yield regression. For quantity of seeds per hectare and labor inputs per hectare, the coefficients estimated with ML measure of yield are within the 95 percent confidence interval of the comparable coefficients from the CC yield regression, in 96 and 80 of the simulations, respectively, out of 100. These results are illustrated on the left panel of Figure 4. The red and green kernel density plots represent the distribution of the coefficients when the relationship is estimated with SR and CC yields, respectively, and the blue histogram bars represent the distribution of the comparable coefficient when the relationship is estimated with ML yields. The middle panel of plots show the results of the regressions of ML yields that are obtained without including SR yields among the covariates. Compared to the aforementioned results based on ML yields, the ML yield regression coefficient for labor input per hectare is in fact more frequently within the 95 percent confidence interval of the comparable coefficient from the CC yield regression (compare bottom left and bottom middle density plots).

Under the assumption that the correlation of measurement error of labor with itself (α) and with the measurement error of CC yield (π) are both negligible and that we capture an unbiased coefficient

Table 8: Regression coefficients for key inputs of the yields-inputs relationship in the test set.

	Measures of Yields												
	SR	CC	ML^b	SR	CC	ML^b	SR	CC	ML^b	SR	CC	ML^b	ML^{dml}
Plot area	−0.302**	0.139	−0.023							−0.211**	0.041	−0.069	−0.002
	(0.119)	(0.107)	(0.074)							(0.105)	(0.128)	(0.056)	(0.061)
Quantity Seeds per ha				0.582***	0.075	0.106				0.372***	0.221**	0.106**	0.038
				(0.136)	(0.157)	(0.088)				(0.065)	(0.098)	(0.050)	(0.079)
Days of Hh labor per ha							0.448***	0.095**	0.131***	0.246***	0.013	0.049	0.052
							(0.060)	(0.045)	(0.039)	(0.068)	(0.039)	(0.035)	(0.045)
Village Fixed effect?	No	No	No	No	No	No	No	No	No	Yes	Yes	Yes	Yes
Observations	386	386	386	386	386	386	386	386	386	386	386	386	386

Notes: Standard errors clustered at the village level for SR and CC. ML^b indicates that the beta coefficients were computed via OLS regression and the standard errors for the ML predictions are obtained via bootstrap with 100 replications. ML^{dml} indicates that the ML regression coefficients and the standard errors were computed following the Chernozhukov et al. (2018) double machine learning method. For this estimation, we use LASSO, split the sample in 5 and conduct 100 repetitions. Significance levels: *p<0.1; **p<0.05; ***p<0.01

Figure 3: Distributions and yield-input regression coefficients for different measures of yield in the first sample split of the test set N = 386. *Notes:* Left top panel shows the kernel density of the different measures of yield. The right middle panel shows the scatter plots of CC vs ML yields. The right middle panel shows the scatter plots of CC vs SR yields. The bottom left/right panel shows the bivariate/multivariate yield-input regression coefficients for different measures of yield.

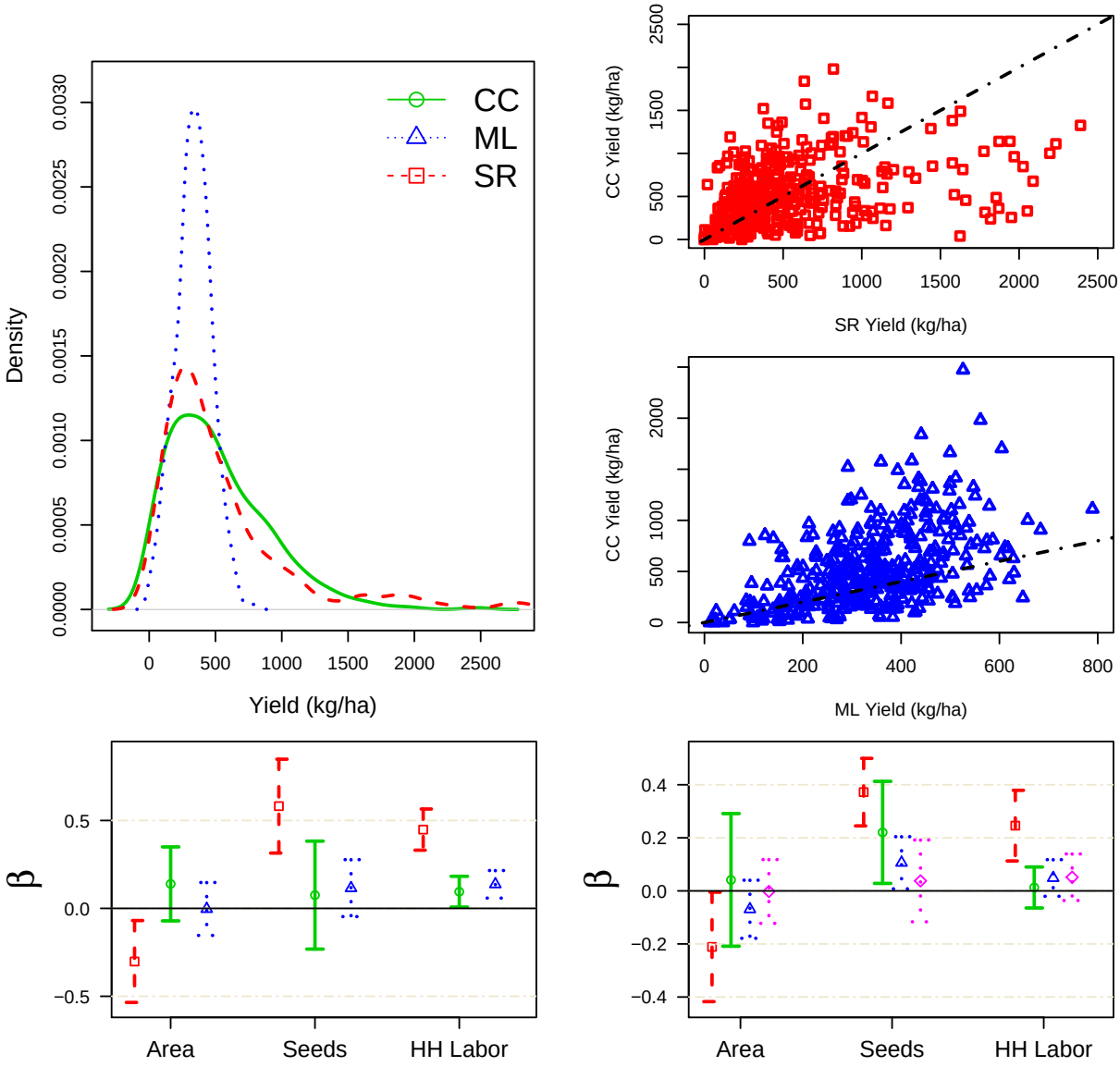
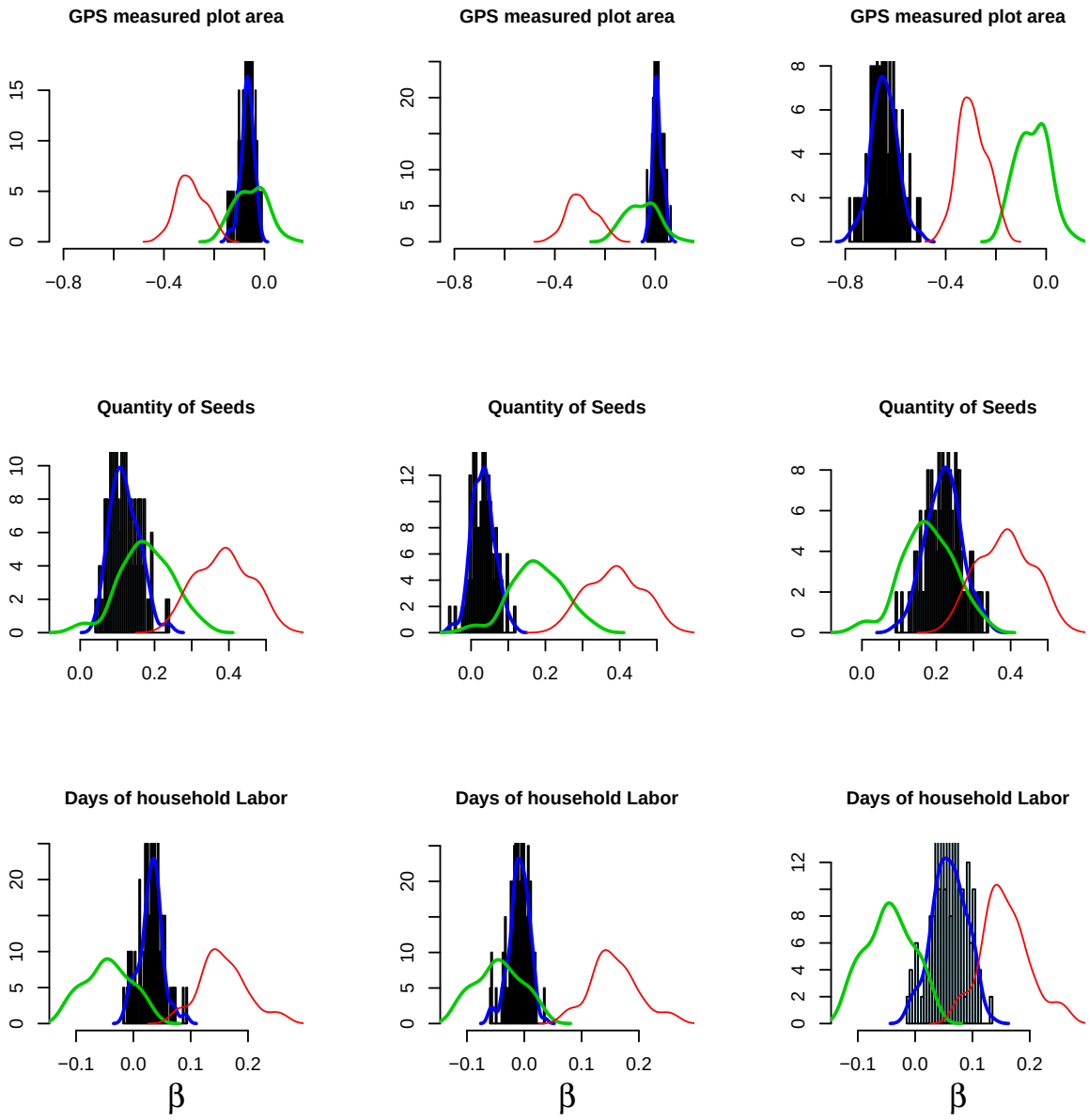


Figure 4: Distribution of yields-inputs regression coefficients in the test for 100 different sample splits. The red and green kernel density plots represent the distribution of the coefficients when the relationship is estimated with SR and CC yields, respectively, and the blue histogram bars represent the distribution of the comparable coefficient when the relationship is estimated with ML yields. The left panel shows the distribution when in the machine learning model, we include SR. The middle panel of plots show the results of the regressions of ML yields that are obtained without including SR yields among the covariates. The right panel shows the results from the alternative regressions of ML yields which are computed with ML-based imputed measures of CC harvest that are in turn standardized by plot area (as opposed to imputing ML yields directly)



for labor with the use of CC yields, these results suggest that if we suspect a high degree of correlation between the measurement error of labor with the true value of labor, it may be better to estimate the yield-input relationship using a predicted measure of CC yields that does not include SR yields among the list of predictors. On the other hand, if λ is not high (as it seems to be the case with quantity of seeds per hectare), it is better to include SR yields among the list of predictors for CC yields. In any case, it is always better to use the ML yields rather than SR yields. Finally, the right panel of Figure 4 shows the results from the alternative regressions of ML yields which are computed with ML-based imputed measures of CC harvest that are in turn standardized by plot area (as opposed to imputing ML yields directly). We can see that the coefficients associated with the key inputs are not estimated correctly - especially for plot area. This confirms that deriving ML-based imputed CC yields, as opposed to CC harvests, is the most appropriate.

To further confirm that the use of ML yields de-biases the estimation of the relationships between yields and inputs, we estimate regression, as presented in Table 3, using the test sample and both SR and ML yields. The results are presented in Table 9. First, inputs do not correlate with overestimation of ML(SR,W) as much as they do with SR. In fact, none of the coefficients are significant. And although the coefficient for quantity of seeds per hectare was positive but not significant under the use of SR yields in the test sample, it is now reversed and is no longer significant. Second, the adjusted R^2 decreases considerably which indicates that the covariates are associated to a lesser extent with the difference in values between CC and ML yields.

4.2.4 Application on the nationally representative survey

To assess the external validity of our findings thus far, we replicate the analysis with data from the Enquête de Conjoncture Intégrée (EACI 2017) [16] a nationally-representative multi-topic household survey, spanning both rural and urban areas of Mali and with detail questionnaire modules on smallholder agricultural activities. The households were interviewed by semi-resident enumerators in three visits, following a fieldwork protocol that was similar to that of ERIVaS. All major crops, including sorghum, were subject to crop cutting during the EACI fieldwork, and the size of the randomly-placed crop cut sub-plot on sorghum plots specifically was 5m x 5m (opposed to 8m x 8m in ERIVaS). One of the objectives of EACI is to derive national- and regional-level estimates of total crop production and yields. On yield measurement, plots, rather than households, are the primarily units of analysis, with direct implications on the sampling of plots for crop cutting purposes. Specifically, in each sampled EACI enumeration area, a complete listing of plots was conducted, and the listed plots were organized into separate subsets in accordance with the cultivated crop (or crop combinations, in the case of intercropped plots). A third of plots were then selected at random from each subset. This means that households with more plots had a greater probability of having a plot selected for crop cutting. In the end, we have farmer self-reported production data for about 24,000 plots of which 7,800 were subject

[16] The EACI 2017 is also the Mali LSMS-ISA 2017. It was built upon the Agricultural Conjuncture Survey which is implemented yearly by the Statistical Unit of the Ministry of Agriculture. For the 2017 edition, living conditions modules were added on a subsample as part of the Mali LSMS-ISA project.

Table 9: Regression results of errors in yields against correlates of measurement errors in yields

	Measurement Errors in Yields	
	log SR- log CC	log ML - log CC
Household labor	0.210***	0.028
	(0.042)	(0.061)
Plot area	−0.283***	−0.169
	(0.106)	(0.104)
Seeds	0.087	−0.123
	(0.105)	(0.131)
Unit SR harvest	−0.076**	−0.003
	(0.034)	(0.035)
*Respondent v1 : female	1.733***	1.903***
	(0.455)	(0.381)
*Respondent v1 : manager	0.519	0.376
	(0.359)	(0.430)
*Respondent v2 : age	0.015***	0.010**
	(0.005)	(0.005)
*Respondent v2 : female	−0.017	−0.0005
	(0.016)	(0.015)
*Respondent v2 : manager	−0.331	−1.266***
	(0.449)	(0.409)
Dammage crop-cut	0.160	−0.567**
	(0.175)	(0.242)
Respondent literate x # plots of Hh	−0.083***	−0.031*
	(0.018)	(0.016)
Rain fall - July	0.003	0.004
	(0.003)	(0.004)
Rain fall - December	0.046	−0.005
	(0.033)	(0.037)
Number of cotton plots	−0.061	−0.058
	(0.069)	(0.044)
Pre-harvest loss	−0.010***	−0.0004
	(0.003)	(0.002)
*Respondent v1 : age	−0.008	−0.007
	(0.016)	(0.015)
Observations	386	386
R^2	0.361	0.150
Adjusted R^2	0.297	0.065

Notes: Estimation conducted in the test sample of the first sample split simulation. Labor, plot area and quantity of seeds are in log form. * indicates that the variable is interacted with enumerator fixed effect. v1 indicates post-planting visit and v2 indicates post-harvest visit. Enumerator fixed effect included and errors clustered by village. Significance levels: *p<0.1; **p<0.05; ***p<0.01

to crop cutting. Focusing on sorghum, EACI collected data on 3,569 plots of which 1,098 were subject to crop cutting.

We start this part of the analysis by comparing the data from ERIVaS and EACI. This provides clues regarding the potential modifications, if any, in the parameter estimates for the relationships between yields and inputs when we extend our analysis to a nationally-representative survey. Table 10 shows summary statistics for pure stand sorghum plots on in ERIVaS and EACI samples.[17] We see that the measures of yields and inputs are statistically different between the two surveys. The mean value in EACI is often larger, and as can be observed on Figure A2 in the Appendix, the amplitude is often greater for the inputs. The latter finding reflects the fact that EACI covers the entire country and captures more heterogenous farming practices and outcomes. This may in turn be beneficial for the use of ML techniques to derive imputed CC yields, since these algorithms work best with more observations and are more adapted to handle potential heterogeneity across units of observations.

Table 10: Sample summary statistics of ERIVaS and EACI 2017.

Variable	Sample	Mean	Std.Dev	Min	Median	Max	Obs.
SR Yields	ERIVaS	636.86	822.26	0.00	400.00	8139.45	577
	EACI 2017	536.70	874.18	0.00	317.89	9523.81	3237
		***					
CC Yields	ERIVaS	500.68	380.06	0.00	423.44	2472.50	577
	EACI 2017	760.19	482.30	0.00	720.00	3200.00	980
		***					
GPS measured area (ha)	ERIVaS	1.56	1.25	0.05	1.20	12.02	577
	EACI 2017	2.65	2.49	0.02	1.94	17.83	3237
		***					
Quantity of Seeds used (kg/ha)	ERIVaS	8.30	5.90	0.56	6.56	51.02	577
	EACI 2017	13.38	42.57	0.00	5.57	1300.00	3237
		***					
Household labor on plot (persons-days/ha)	ERIVaS	159.30	267.37	8.36	76.40	2711.45	577
	EACI 2017	69.03	146.20	0.54	34.04	2700.00	3237
		***					

Notes: To ensure comparability of the variables, the EACI 2017 sample is limited to pure stand plots. The stars. below the sample mean values of the two surveys denote the statistical significance level for a t-test comparison of the means in the two samples. Significance levels: *p<0.1; **p<0.05; ***p<0.01

In terms of the non-classical measurement errors that were uncovered in the ERIVaS data, similar patterns emerge from the EACI sample. Table 11 shows the means of alternative yield and harvest measures for pure stand and intercropped sorghum plots. Similar to ERIVaS, the mean differences between SR and CC measures are statistically significant. However, in the nationally-representative sample, the sign of the difference is opposite and we find that the mean SR yields are underestimated vis-à-vis CC yields (i.e., SR - CC < 0). Figure 5 allows us to better understand why this is the case by showcasing the mean difference between SR and CC yields across quintiles of plot area, quantity of seeds per hectare and days of household labor per hectare. SR yields are overestimated vis-à-vis CC yields in the first plot area quintile, while, in each of the remaining quintiles, SR yields are underestimated. On

[17] An expanded set of descriptive statistics computed with both samples can be found in Table A4.

the other hand, SR yields are underestimated vis-à-vis CC yields in each of the first four quintiles of quantity of seeds per hectare and days of household labor per hectare and are conversely overestimated in the top quintile.

Table 11: Comparison of plot-level measures of sorghum yield and harvest in EACI 2017

	Yields (Kg/Ha)			**Harvest (Kg)**		
	All	Pure-stand	Intercropped	All	Pure-stand	Intercropped
Means						
SR	574.6	561.1	687.1	970.6	950.8	1135.1
	(27.2)	(28.0)	(101.0)	(43.6)	(44.1)	(174.3)
CC	757.0	760.2	730.5	2037.1	2052.9	1906.5
	(14.5)	(15.4)	(42.5)	(85.9)	(91.9)	(237.6)
Difference in Means						
SR vs CC	-182.4***	-199.1***	-43.4	-1066.5***	-1102.0***	-771.4***
	(29.1)	(30.3)	(99.5)	(77.9)	(84.1)	(193.9)
Difference in Distribution						
SR vs CC	***	***	***	***	***	***

Notes: Distributional differences are assessed using the Kolmogorov-Smirnov two-sample test
Significance levels: *p<0.1; **p<0.05; ***p<0.01

The differences in the graphs based on ERIVaS versus EACI samples are due, in part at least, to the differences in the distributions of the input variables. For example, the mean SR-CC difference in the third quintile of plot area in EACI is equivalent to the mean difference in the fourth quintile in ERIVaS, so the underestimation trend that we observe in ERIVaS starts earlier in the distribution for EACI 2017. However, the direction of the correlation between the SR-CC difference and the quantity of inputs per hectare are the same in both ERIVaS and EACI samples. As such, it is not surprising to see in Table 12 that the coefficients of the bivariate and multivariate regressions of yields and inputs change significantly based on whether we use SR versus CC yields in the EACI data, but that the changes are fully consistent with the findings based on the data from the methodological experiment.

Furthermore, while we do not have concerns regarding measurement errors in CC yields in the ERIVaS data set due to (i) strict supervision of activities related to crop cut sub-plot placement, harvesting and weighing, and (ii) comprehensive quality control measures applied to the incoming data throughout the fieldwork, there is reason to believe that the EACI crop cutting operations were not implemented as successfully. Figure 6 shows a scatterplot of SR and CC yields in the EACI data set and there is evidence of rounding errors in CC yields. Assuming that these errors are not correlated with the measures of inputs,[18] this would mostly create an attenuation bias in the coefficients for the respective inputs in the regressions of CC yields and could, therefore, explain in part the zero coefficients in Table 12. This finding underscores the need of quality (but still scalable) digital measurement devices when objective measurement is the goal.[19]

[18]We created a variable indicating whether an observation was rounded and it appears not to be associated with the value of the key inputs we study here. A close look at this issue is left to future work.

[19]In fact, discussion with CPS/SDR confirmed that during this round, the survey enumerators relied on scales which lacked accuracy and were scheduled to be replaced.

Figure 5: Self-Reported yield overestimation of crop-cut yield over quintiles of crop-cut yields and key inputs. *Notes:* The y-axis represents the amount by which SR yields overestimate CC yields ($Y = SR - CC$) in kg/ha. The x-axis shows the quintile of the crop-cut yields or of the input per ha. We are restricted to the plots with crop-cut

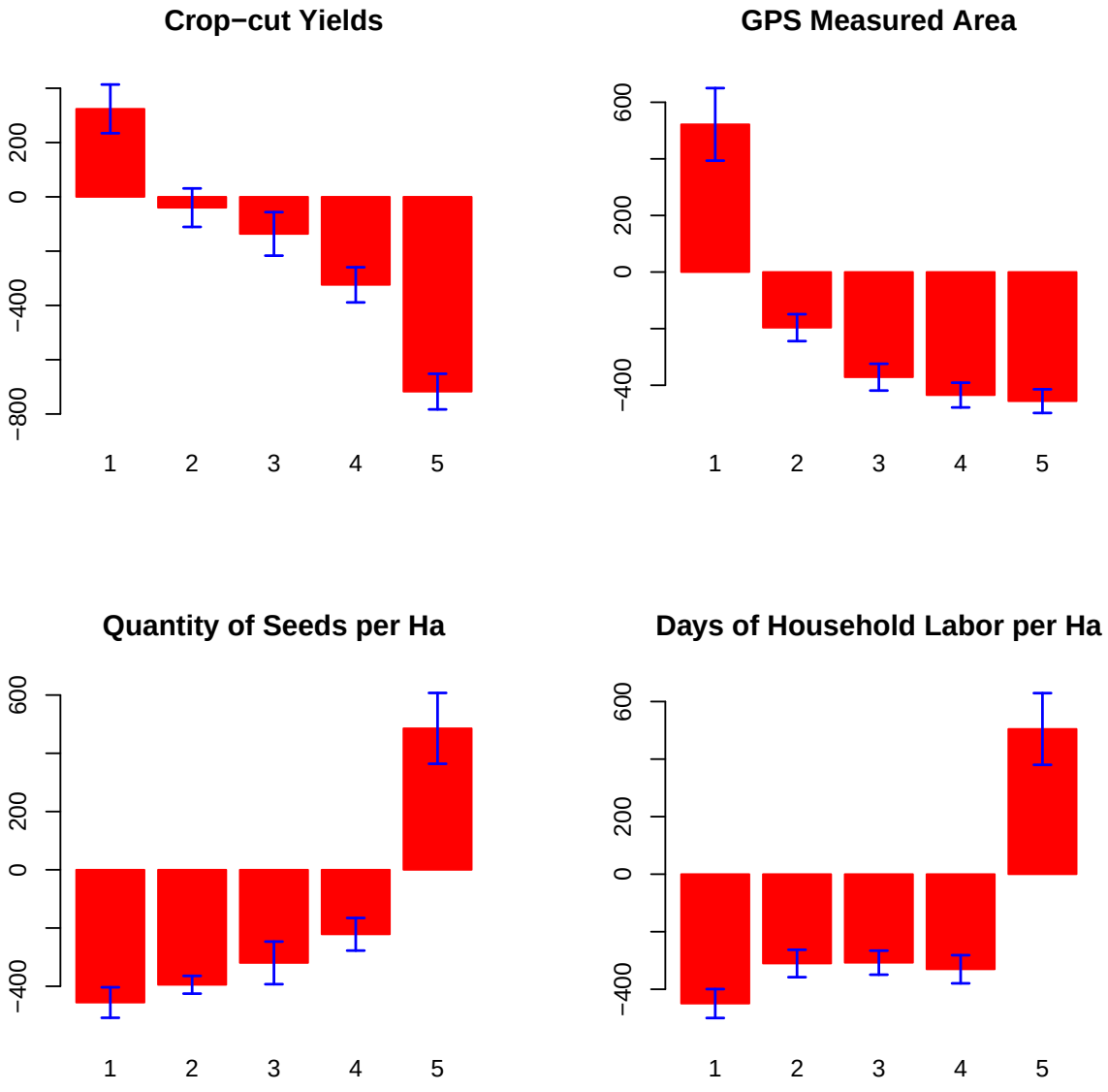
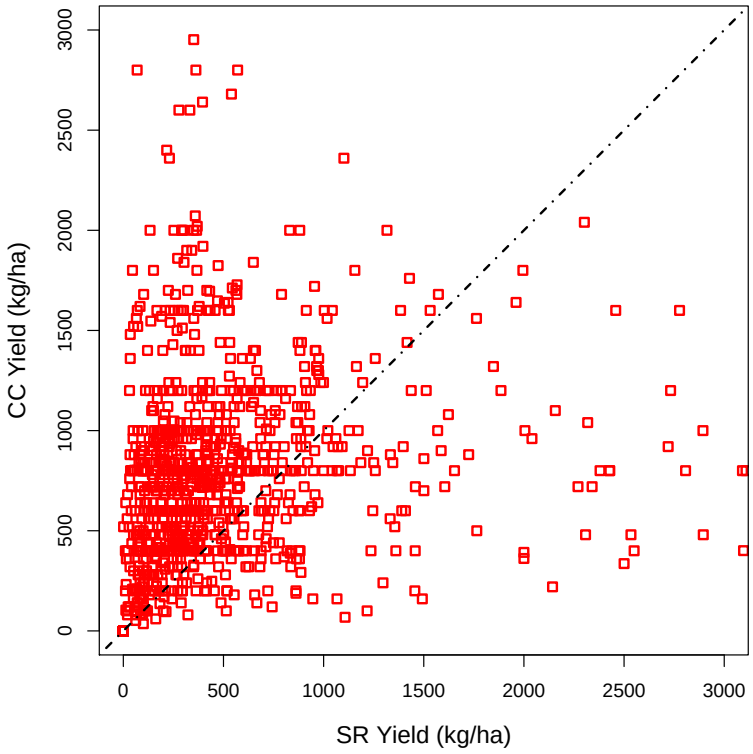


Figure 6: Distribution of SR vs CC yield (EACI sample)



Nevertheless, we replicate the ML approach, as outlined in section 4.2.1, with the nationally-representative EACI sample. Figure 7 summarizes the results for the first sample split out of the 100 replications. The right top-panel shows that ML yields 40 percent of the variation in CC yields, and the level of explained variation is in fact higher than the comparable result based on the ERIVaS sample. Even though we continue to see (in the left-top panel) a higher concentration of values around the mean for ML yields, there is a better tracking of the large values of CC yields. The coefficients for the key inputs in the regressions of ML yields are also comparable to those obtained from the regressions of CC yields, as shown in the bottom two plots of Figure 7. While the size and heterogeneity of the EACI sample are among the factors that improve the predictive accuracy of the ML model vis-à-vis the results achieved with the ERIVaS sample, the EACI rounding errors in CC yields may also be creating a scenario that ML tools are better equipped to accommodate.

To verify whether we can expect the same results from the ML approach done with ERIVaS (selecting randomly a third of the plots) using the EACI 2017 and make inferences to the population of sorghum plots in Mali, we compare in Table A5, plots and households characteristics of plots selected for crop cutting with those that were not. The statistically significant differences appear to be seldom providing confidence that the selection of plot for CC is analogous to the simulation conducted with the data from the experience.

Finally, Table 12 reports the results from the yield regressions estimated with ML, and separately SR, yields using the entire EACI sample, and the results from the yield regressions estimated with CC yields using the sub-sample of EACI plots that were subject to crop cutting. The coefficient for plot area is negative but not statistically significant. This may be because the mean difference in plot area between the crop cut plots and the rest of the sample was statistically significant. The plots that were not subject to crop cutting were, on average, larger, and in view of the evidence regarding the underestimation in SR yields on larger plots, the IR is weakened when the regression is estimated using the entire sample. And while the coefficients for key inputs are consistently positive and statistically significant in the regressions of SR yields, the comparable coefficients in the regressions of ML yields are not statistically significant and are consistent with those estimated in the regressions of CC yields. These results are fully consistent with the findings based on the ERIVaS data set and raise flags concerning the analyses conducted with SR yields.

5 Implications for crop yield measurement in household and farm surveys

Having demonstrated the promise of the use of ML to derive imputed measures of CC yields, this section provides a few recommendations on what data to collect and how to collect it in an efficient way.

First, given the importance of using SR yields to improve the quality of the ML-based imputed CC yield predictions, survey practitioners should continue to collect farmer-reported information on crop production, and reduce NCME in the resulting data, including by reducing length of farmer recall and improving the completeness and quality of conversion factors for non-standard measurement units,

Figure 7: Distributions and yield-input regression coefficients for different measures of yield in one sample split N = 732. *Notes:* Left top panel shows the kernel density of the different measures of yield. The right middle panel shows the scatter plots of CC vs ML yields. The bottom left/right panel shows the bivariate/multivariate yield-input regression coefficients for different measures of yield.

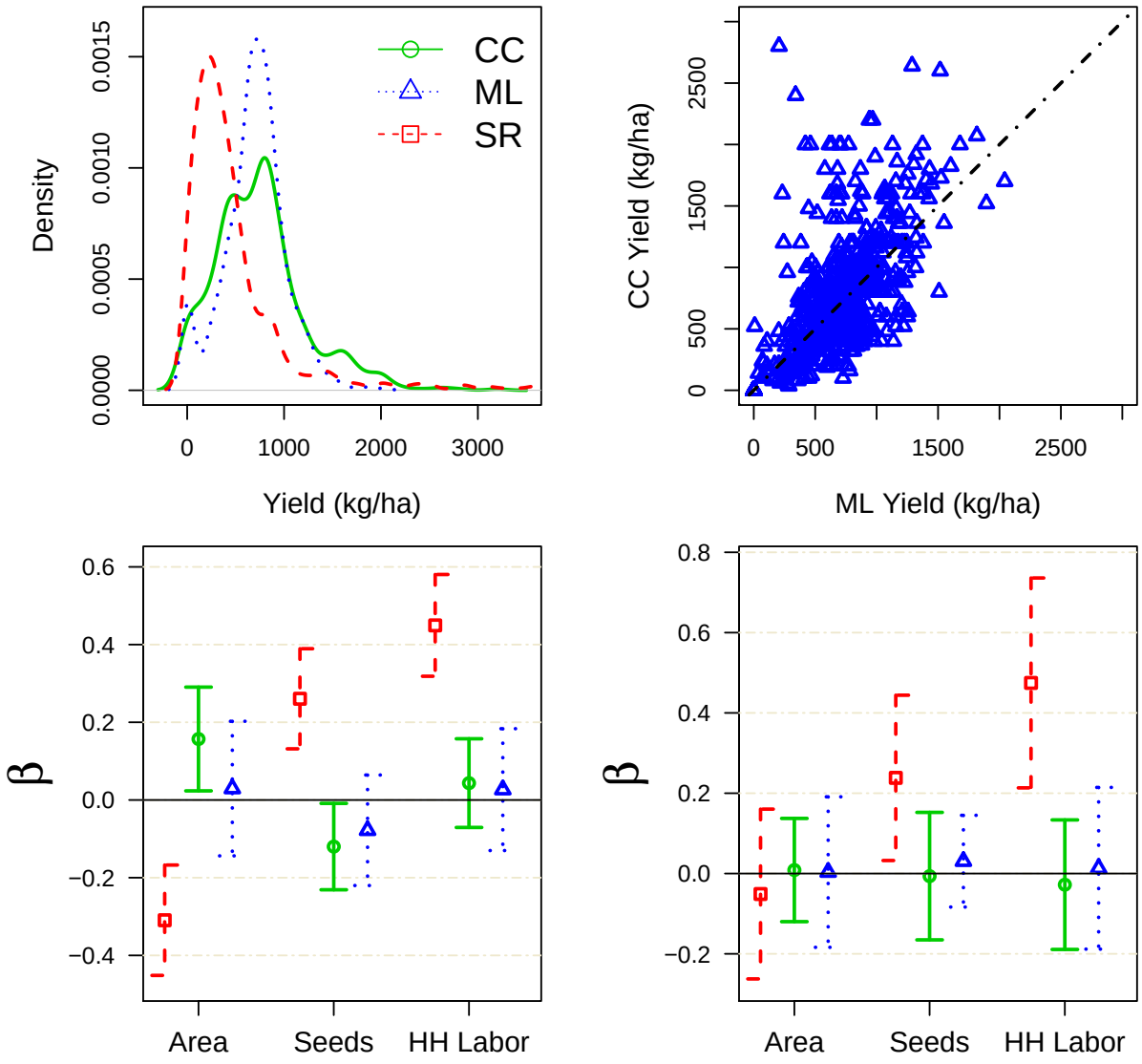


Table 12: Regression coefficients for key inputs of the yields-inputs relationship in in nationally representative sample.

	Measures of Yields											
	SR	CC	ML^b	SR	CC	ML^b	SR	CC	ML^b	SR	CC	ML^b
Plot area	−0.194*	0.056	0.219***							−0.173***	0.003	0.089
	(0.099)	(0.048)	(0.063)							(0.053)	(0.050)	(0.063)
Quantity Seeds per ha				0.148*	−0.069*	−0.218***				0.151***	0.055*	0.012
				(0.082)	(0.039)	(0.047)				(0.035)	(0.031)	(0.033)
Days of Hh labor per ha							0.396***	0.006	−0.112*	0.143**	−0.050	−0.038
							(0.096)	(0.045)	(0.059)	(0.058)	(0.056)	(0.056)
EA Fixed Effect?	No	No	No	No	No	No	No	No	No	Yes	Yes	Yes
Observations	3,569	1,098	3,569	3,569	1,098	3,569	3,569	1,098	3,569	3,569	1,098	3,569

Notes: Standard errors clustered at the enumeration area level for SR and CC. ML^b indicates that the beta coefficients were computed via OLS regression and the standard errors for the ML predictions are obtained via bootstrap with 100 replications. Significance levels: *p<0.1; **p<0.05; ***p<0.01

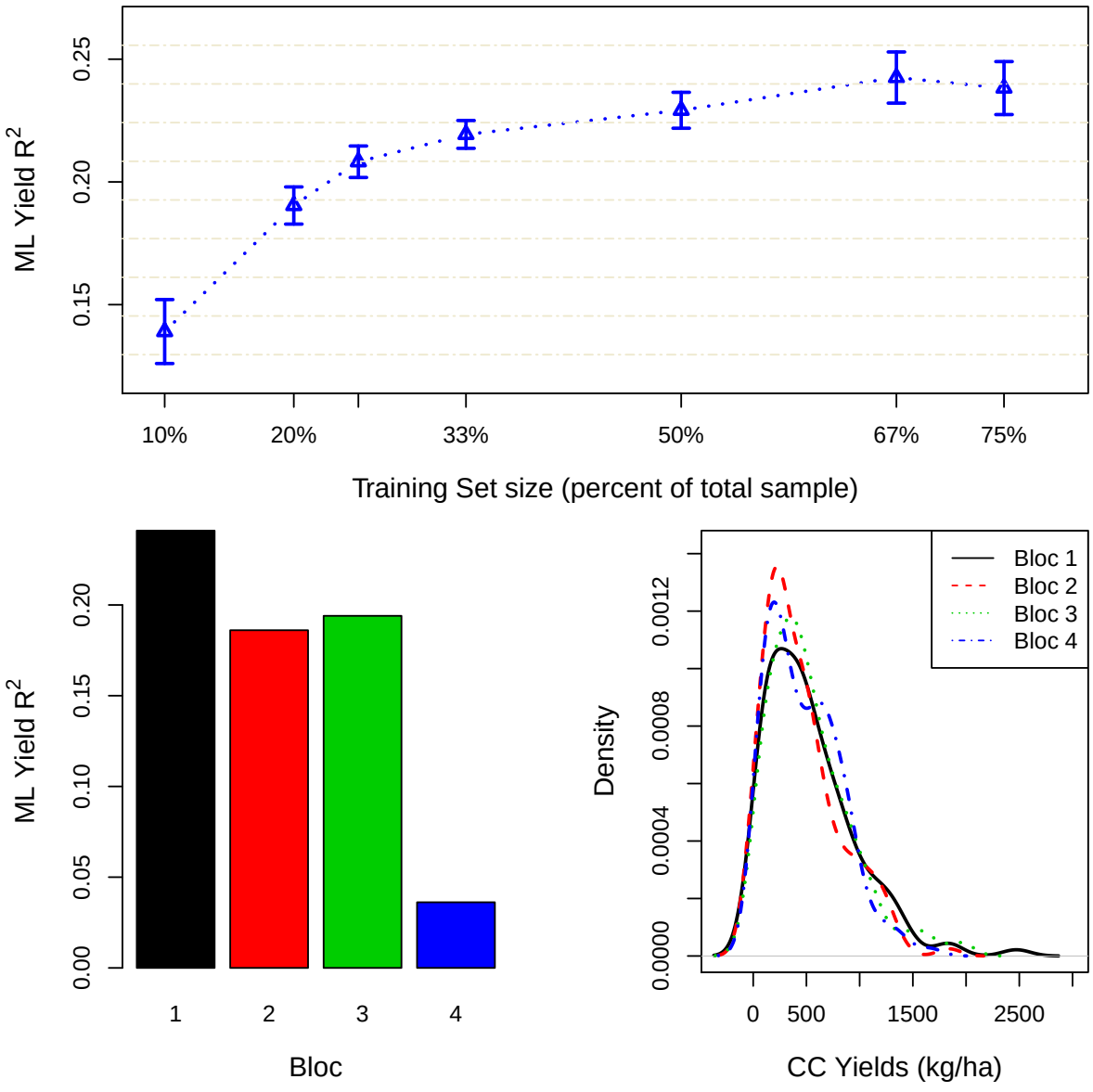
among other measures.

Second, in view of the complex logistics, supervision requirements, and ultimately higher costs associated with crop cutting, the question is whether crop cutting can be implemented on a sub-sample basis and if so, the minimum share of plot observations that should be subject to crop cutting in a way that allows us to estimate an imputation model that provides reliable measures of imputed CC yields. We assessed the predictive accuracy of our preferred ML specification with different sizes of the training data set - experimenting with 10, 20, 25, 50, 67, and 75 percent of the crop cut measures, while leaving the remaining sample in each scenario to be the test sample. It appears (see top panel of Figure 8) that conducting crop cutting on 50 percent of the plot observations may be sufficient for deriving reliable imputed CC yields. After the 50 percent mark, the gains in accuracy appear to be marginal.

Third, a related question is whether it is necessary to randomly designate a share of plot observations for crop cutting in all primary sampling units or whether reliable predictions can be obtained by working with a subsample of primary sampling units and conducting crop cutting on all plots in the designated areas. Since ML algorithms can perform well out of sample, logistics and supervision of crop cutting can potentially be simplified by limiting our footprint to fewer areas in a given country. To answer this question, we create an alternative training data set composed only of crop cut measures obtained in one randomly chosen ERIVaS stratum – as opposed to working with one-third of plots randomly chosen across all strata. The test sample in this case are composed of plot observations from the remaining ERIVaS strata. The bar plot in the bottom left panel of Figure 8 shows that the choice of the stratum has an important effect on the accuracy of the ML measure. The choice of the first stratum as the source of the training data results in an the R^2 value of 0.24 from the bivariate regression of observed CC yield against imputed CC yield in the test sample. The comparable R^2 estimate is 0.04 while working with the fourth stratum as the source of the training data. The reason behind these differences in performance might be the fact that the first stratum was the most heterogeneous unit in terms of CC yields: the kernel density of CC yields is more spread out within the first stratum vis-à-vis the rest of the strata (see bottom right panel on Figure 8).

So, in view of national surveys that may consider implementing crop cutting on a sub-sample basis and deriving imputed CC yields for the rest of the sample, the findings imply that implementing crop cutting in a way that is confined to a particular region may only result in reliable predictions if the chosen region is sufficiently heterogenous in terms of crop yields.

Figure 8: Sampling strategies for optimal ML predictions. *Notes:* The top panel shows the R^2 of theregression of CC on ML as a function of the training set size. We repeated the operation 50 times foreach sample size, so each point is the mean of the R^2 and the bars are the standard errors. The bottom panel illustrates how sampling unit choice afffects the out of sample R^2. The bottom right panel shows the kernel density of the crop-cut yields in each sampling unit. For all simulations, we use the preferred ML model set up specification: we fit the model using log CC and include SR in the covariates.



6 Concluding remarks

This paper attempts to increase the reliability of assessments of yield-input relationships when yields based on self-reported survey data on crop production are subject to non-classical measurement errors. We show analytically and empirically that it is possible to construct imputed measures of objective crop cut yields, using a machine learning model and a sub-sample of plot observations with both crop cut and self-reported yield measures in a household survey. The imputed crop cut yields are shown to be less biased vis-à-vis self-reported yields, and the errors in the imputed measures are uncorrelated with input use per hectare. As such, the regressions of imputed crop cut yields replicate the true relationships between yields and inputs that are estimated with observed crop cut yields. These conclusions are robust to whether we use the data from a small-scale methodology study conducted in southwestern Mali or a nationally-representative survey.

From a data collection point of view, these findings carry the implications that more objective methods of collecting crop production data need to be integrated in the protocol of household surveys. We rely on two household survey data sets from Mali with the rare characteristic of having both self-reported and crop cut data. The first data set from a methodological experiment enables us to empirically show that it is possible to combine self-reported yields data to judiciously predict crop cut yields and, thus build a measure of yields that appears to be fit for a correct assessment of the relationship between yields and inputs with a random sample representing one-third of the entire sample. The second data set, a nationally representative survey, enabled us to validate the approach but also served as a stark reminder that even when a data collection method is theoretically objective, execution is key to ensure that meaningful inference can be made based on it. Although these results are promising, further work considering other crops should be pursued to improve the machine learning predictions, evaluate the limit of the inferences that can be carried out using this approach, and better understand its potential bias reducing properties.

Finally, our analysis expands on the body of work related to non-classical measurement errors and its impact on agricultural policy analysis in a smallholder context. We show that measurement errors in self-reported yields are correlated with major factors of production in smallholder farming systems. On land; the factor that has been the focus of the recent literature regarding the study of the inverse scale-productivity relationship (Desiere and Jolliffe, 2018; Abay et al.,2019; Gourlay et al.,2019), our results corroborate the existing evidence that IR disappears and is even reversed when the relationship is assessed with crop cut yield as opposed to self-reported yields. We also examine the relationships between yields and seeds per hectare and yields and household labor per hectare. These relationships have serious implications for agricultural policy especially as it pertains to the question of agricultural intensification. A positive relationship between yields and quantity of seeds suggests that farmers could improve their yields by applying more seeds and thus would provide support for making seeds more accessible to farmers. A positive relationship between yield and household labor indicates that plot-level land productivity could be increased with more labor, suggesting that labor is not correctly allocated across households. But similar to what is observed in the case of plot area, we document that measurement errors in self-reported yields are correlated with the quantity of seeds and household

labor used on the plot and that these errors lead to spurious relationships between yield and inputs use per hectare. We find that the positive relationship between yield and quantity of seeds is weakened or disappears, and the positive relationship between yield and household labor ceases to exist. Although our findings cannot be interpreted as causal and our measures of quantity of seed and labor need to be improved, these pieces of descriptive evidence suggest that intensification may not necessarily be the approach to follow to improve yields for smallholder farmers and that further research is needed in the domain.

References

Abay, K. A., Abate, G. T., Barrett, C. B., and Bernard, T. (2019). Correlated non-classical measurement errors, 'second best' policy inference, and the inverse size-productivity relationship in agriculture. *Journal of Development Economics*, 139:171 – 184.

Abay, K. A., Bevis, L. a. E. M., and Barrett, C. B. (2021). Measurement error mechanisms matter: Agricultural intensification with farmer misperceptions and misreporting. *American Journal of Agricultural Economics*, 103(2):498–522.

Arthi, V., Beegle, K., Weerdt, J. D., and Palacios-Lopez, A. (2018). Not your average job: Measuring farm labor in tanzania. *Journal of Development Economics*, 130:160 – 172.

Athey, S. (2018). The Impact of Machine Learning on Economics. In *The Economics of Artificial Intelligence: An Agenda*, NBER Chapters. National Bureau of Economic Research, Inc.

Belloni, A. and Chernozhukov, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2):521–547.

Carletto, C., Gourlay, S., Murray, S., and Zezza, A. (2017). Cheaper, faster, and more than good enough: Is gps the new gold standard in land area measurement? *Survey Research Methods*, 11(3):235–265.

Carletto, C., Gourlay, S., and Winters, P. (2015). From guesstimates to gpstimates: Land area measurement and implications for agricultural analysis. *Journal of African Economies*, 24(5):593–628.

Carletto, C., Savastano, S., and 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.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68.

Desiere, S. and Jolliffe, D. (2018). Land productivity and plot size: Is measurement error driving the inverse relationship? *Journal of Development Economics*, 130:84 – 98.

Gaddis, I., Oseni, G., Palacios-Lopez, A., and Pieters, J. (2020). Measuring Farm Labor: Survey Experimental Evidence from Ghana. *The World Bank Economic Review*, 35(3):604–634.

Gollin, D. and Udry, C. (2021). Heterogeneity, Measurement Error, and Misallocation: Evidence from African Agriculture. *Journal of Political Economy*, 129(1):1–80.

Gourlay, S., Kilic, T., and Lobell, D. B. (2019). A new spin on an old debate: Errors in farmer-reported production and their implications for inverse scale - productivity relationship in uganda. *Journal of Development Economics*, 141:102376.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2014). *An Introduction to Statistical Learning: With Applications in R*. Springer Publishing Company, Incorporated.

Kilic, T., Zezza, A., Carletto, C., and Savastano, S. (2017). Missing(ness) in Action: Selectivity Bias in GPS-Based Land Area Measurements. *World Development*, 92(C):143–157.

Lobell, D. B., Cassman, K. G., and Field, C. B. (2009). Crop yield gaps: Their importance, magnitudes, and causes. *Annual Review of Environment and Resources*, 34(1):179–204.

Lobell, D. B., Di Tommaso, S., You, C., Yacoubou Djima, I., Burke, M., and Kilic, T. (2019). Sight for sorghums: Comparisons of satellite- and ground-based sorghum yield estimates in mali. *Remote Sensing*, 12(1):100.

Lowder, S. K., Skoet, J., and Raney, T. (2016). The number, size, and distribution of farms, smallholder farms, and family farms worldwide. *World Development*, 87:16–29.

Oseni, G., Durazo, J., and McGee, K. (2017). The use of non-standard units for the collection of food quantity, a guidebook for improving the measurement of food consumption and agricultural production in living standards surveys. Technical guidebook, The World Bank.

Polley, E., LeDell, E., Kennedy, C., and van der Laan, M. (2018). *SuperLearner: Super Learner Prediction*. R package version 2.0-24.

Storm, H., Baylis, K., and Heckelei, T. (2020). Machine learning in agricultural and applied economics. *European Review of Agricultural Economics*, 47(3):849–892.

van Ittersum, M. K., Cassman, K. G., Grassini, P., Wolf, J., Tittonell, P., and Hochman, Z. (2013). Yield gap analysis with local to global relevance?a review. *Field Crops Research*, 143:4 – 17. Crop Yield Gap Analysis ? Rationale, Methods and Applications.

Wollburg, P., Tiberti, M., and Zezza, A. (2021). Recall length and measurement error in agricultural surveys. *Food Policy*, 100:102003.

Zhang, W., Cao, G., Li, X., Zhang, H., Wang, C., Liu, Q., Chen, X., Cui, Z., Shen, J., Jiang, R., Mi, G., Miao, Y., Zhang, F., and Dou, Z. (2016). Closing yield gaps in china by empowering smallholder farmers. *Nature*, 537:671 EP –.

Appendix A Additional tables, full regression results, and figures

Table A1: Comparison of plot-level measures of sorghum yields and harvests in ERIVaS with outliers treatment

	Yields (kg/ha)	**Harvest (kg)**
Means		
SR	593.5	802.3
	(24.9)	(43.2)
CC 8m x 8m	500.7	794.4
	(15.8)	(40.0)
CC 4m x 4m	501.0	782.2
	(16.8)	(40.8)
Difference in Means		
SR vs CC 8m x 8m	92.8***	7.8
	(24.2)	(40.1)
SR vs CC 4m x 4m	92.5***	20.0
	(24.6)	(42.3)
CC 8m x 8m vs CC 4m x 4m	-0.3	12.2
	(6.1)	(11.0)
Difference in Distribution		
SR vs CC 8m x 8m		***
SR vs CC 4m x 4m	*	***
CC 8m x 8m vs CC 4m x 4m		

Notes: Standard Errors are shown in parentheses. Distributional differences are assessed using the Kolmogorov-Smirnov two-sample test. Statistical significance levels: *p<0.1; **p<0.05; ***p<0.01

Table A2: Regression results of SR yields overestimation of CC yields

	$\text{Log}\left(\frac{SR}{CC}\right)$	
Log of GPS measured plot area	-0.174^{**}	-0.231^{*}
	(0.088)	(0.119)
Log of quantity of seeds per Ha	0.260^{***}	0.272^{***}
	(0.075)	(0.074)
Log of total household labor per Ha	0.195^{***}	0.141^{***}
	(0.064)	(0.054)
Household did not use external labor	-0.019	0.077
	(0.154)	(0.143)
Log of total external labor per Ha	-0.023	-0.002
	(0.053)	(0.047)
Plot received organic fertilizer	0.092	0.097
	(0.128)	(0.128)
Plot received mineral/chemical fertilizer	0.151^{*}	0.157^{**}
	(0.078)	(0.077)
Plot received pesticides	0.018	0.033
	(0.072)	(0.074)
Plot is owned by household	0.213	0.153
	(0.199)	(0.170)
Plot was left fallowed during the past 10 years	0.278^{*}	0.258
	(0.167)	(0.181)
Farmer reported distance to the market	-0.007	-0.012^{**}
	(0.005)	(0.005)
Agriculture Equipment Index	-0.084^{*}	-0.113^{**}
	(0.048)	(0.045)
Number of plots cultivated by HH	-0.024	-0.023
	(0.018)	(0.022)
Number of sorghum plots cultivated by farmers	0.037	0.022
	(0.064)	(0.064)
Number of cotton plots cultivated by household	-0.138^{***}	-0.117^{**}
	(0.049)	(0.047)
Log of household size	0.074	0.082
	(0.099)	(0.099)
Household dependency ratio	-0.025	-0.051
	(0.029)	(0.037)
Polygamous household	-0.048	-0.041
	(0.120)	(0.108)
Welfare index	0.074^{**}	0.104^{***}
	(0.030)	(0.035)
Head of household is female	-0.166	-0.026
	(0.163)	(0.239)
Log of age of household head	0.125	0.208
	(0.162)	(0.198)
Household head is literate	0.324^{**}	0.318^{***}
	(0.128)	(0.120)

Table A2: Regression results of SR yields overestimation of CC yields - cont.

	$\text{Log}\left(\frac{SR}{CC}\right)$	
Log of age of plot manager	−0.021	−0.079
	(0.104)	(0.109)
Manager is literate	−0.409***	−0.398***
	(0.140)	(0.145)
SR in sheafs	0.069	0.120
	(0.167)	(0.174)
SR in Bags	−0.086	0.038
	(0.094)	(0.095)
SR in donkey carts	−0.452***	−0.392***
	(0.102)	(0.126)
SR in ox carts	−0.268*	−0.362**
	(0.140)	(0.175)
Enumerator assessed percent of dammage on plot	0.009***	0.008***
	(0.002)	(0.002)
Days between CC harvest and Post Harvest interview	−0.015*	−0.018***
	(0.008)	(0.007)
Respondent was plot manager during post-planting interview	0.004	0.022
	(0.094)	(0.095)
Respondent was plot manager during post-parvest interview	0.045	0.075
	(0.076)	(0.065)
CC Yield is 0	−1.168***	−1.066***
	(0.242)	(0.278)
SR yield is 0	−0.360	−0.409
	(0.328)	(0.352)
Enumerator Fixed effect?	No	Yes
Observations	577	577
R^2	0.251	0.305
Adjusted R^2	0.204	0.234
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table A3: Full Multivariate Regressions results of SR and CC measures of yields for selected inputs

	Measures of Yields			
	SR	SR	CC	CC
GPS measured plot area	−0.252***	−0.304***	−0.013	−0.045
	(0.065)	(0.085)	(0.083)	(0.099)
Quantity of Seeds	0.352***	0.366***	0.159**	0.185**
	(0.075)	(0.068)	(0.077)	(0.086)
Days of household Labor	0.185***	0.159***	−0.028	−0.040
	(0.058)	(0.058)	(0.035)	(0.045)
Hired labor used on plot	−0.268	−0.248	−0.255**	−0.246*
	(0.181)	(0.203)	(0.122)	(0.128)
Days of hired labor	−0.100	−0.087	−0.073	−0.082*
	(0.066)	(0.077)	(0.046)	(0.048)
Organic fertilizer used on plot	−0.006	−0.017	−0.068	−0.078
	(0.101)	(0.099)	(0.084)	(0.094)
Mineral/Chemical fertilizer used on plot	0.151**	0.121	−0.006	−0.031
	(0.074)	(0.086)	(0.066)	(0.074)
Pesticides used in plot	0.160*	0.142*	0.080	0.086
	(0.082)	(0.083)	(0.076)	(0.078)
Log Distance to dwelling	−0.084***	−0.072**	−0.039	−0.015
	(0.025)	(0.032)	(0.040)	(0.059)
Slope of the plot (GPS)	−0.133**	−0.068	−0.064	−0.041
	(0.059)	(0.047)	(0.051)	(0.069)
Potential wetness index	0.006	0.009	0.007	0.007
	(0.006)	(0.007)	(0.006)	(0.007)
Plot is owned by HH	0.104	0.224	−0.373**	−0.308*
	(0.166)	(0.160)	(0.171)	(0.158)
Plot was inherited	−0.067	−0.101	0.060	0.062
	(0.088)	(0.078)	(0.082)	(0.087)
Plot has erosion	−0.149	−0.151	−0.045	−0.036
	(0.121)	(0.134)	(0.098)	(0.108)
Plot Protected against erosion	0.502***	0.501**	0.224	0.254
	(0.182)	(0.221)	(0.210)	(0.222)
Agriculture implementation index	0.025	0.003	0.139***	0.142***
	(0.049)	(0.040)	(0.041)	(0.043)
Plot was not plowed	0.056	0.043	0.044	0.037
	(0.054)	(0.048)	(0.069)	(0.067)
Plot fallowed past decade	−0.313***	−0.258**	−0.641***	−0.572***
	(0.083)	(0.124)	(0.157)	(0.150)
Number of sorghum plots	0.032	0.025	0.139***	0.141***
	(0.073)	(0.067)	(0.032)	(0.031)
Number of cotton plots	−0.010	−0.065*	−0.032	−0.039
	(0.054)	(0.036)	(0.063)	(0.068)
Number of plots	0.186*	0.194**	0.121	0.105
	(0.102)	(0.089)	(0.079)	(0.084)

Notes: Errors clustered at the village level. Dependent and inputs Independent variables shown here are log transformed. Significance levels: *p<0.1; **p<0.05; ***p<0.01

Table A3: Full Multivariate Regressions results of SR and CC measures of yields for selected inputs - cont.

	Measures of Yields			
	SR	SR	CC	CC
Log Household size	0.038	0.016	0.056	0.070
	(0.057)	(0.065)	(0.058)	(0.056)
Dependency ratio	0.004	0.020	−0.077*	−0.065*
	(0.043)	(0.039)	(0.041)	(0.038)
Welfare Index	−0.026	−0.019	−0.010	−0.012
	(0.026)	(0.022)	(0.019)	(0.018)
Log age of Head	0.034	−0.028	−0.289	−0.271
	(0.124)	(0.130)	(0.188)	(0.195)
Head is literate	0.205	0.020	−0.664**	−0.709*
	(0.209)	(0.213)	(0.331)	(0.377)
Log age of Manager	−0.077	−0.074	0.179	0.170
	(0.130)	(0.117)	(0.165)	(0.172)
Manager is literate	−0.360*	−0.156	0.527*	0.557*
	(0.189)	(0.234)	(0.299)	(0.324)
Dammage on plot	−0.021***	−0.019***	−0.029***	−0.028***
	(0.004)	(0.003)	(0.004)	(0.003)
Sowing Early	0.084	0.029	0.059	0.022
	(0.126)	(0.141)	(0.092)	(0.107)
Sowing Late	−0.644	−0.549	−0.665*	−0.586*
	(0.504)	(0.472)	(0.360)	(0.333)
Log Rain first half August	2.530***	1.781	1.737***	1.730***
	(0.628)	(1.421)	(0.532)	(0.658)
Log Total amount of rain	−0.181	−0.440	−2.181***	−0.955
	(1.303)	(2.466)	(0.676)	(1.443)
Village Fixed effects?	No	Yes	No	Yes
Observations	577	577	577	577
R^2	0.406	0.443	0.407	0.424
Adjusted R^2	0.370	0.391	0.371	0.371

Notes: Errors clustered at the village level. Dependent and inputs Independent variables shown here are log transformed. Significance levels: *p<0.1; **p<0.05; ***p<0.01

Table A4: Sample summary statistics of ERIVaS and EACI 2017.

Variable	Sample	Mean	Std.Dev	Min	Median	Max	Obs.
SR Yields	ERIVaS	636.86	822.26	0.00	400.00	8139.45	577
	EACI 2017	536.70	874.18	0.00	317.89	9523.81	3237
		***					
CC Yields	ERIVaS	500.68	380.06	0.00	423.44	2472.50	577
	EACI 2017	760.19	482.30	0.00	720.00	3200.00	980
		***					
GPS measured area (ha)	ERIVaS	1.56	1.25	0.05	1.20	12.02	577
	EACI 2017	2.65	2.49	0.02	1.94	17.83	3237
		***					
Quantity of Seeds used (kg/Ha)	ERIVaS	8.30	5.90	0.56	6.56	51.02	577
	EACI 2017	13.38	42.57	0.00	5.57	1300.00	3237
		***					
Household labor on plot (persons-days/ha)	ERIVaS	159.30	267.37	8.36	76.40	2711.45	577
	EACI 2017	69.03	146.20	0.54	34.04	2700.00	3237
		***					
Hired labor (proportion of plots)	ERIVaS	0.33	0.47	0.00	0.00	1.00	577
	EACI 2017	0.45	0.50	0.00	0.00	1.00	3237
		***					
Hired labor (persons-days/ha)	ERIVaS	15.23	180.21	0.00	0.00	4195.02	577
	EACI 2017	6.60	69.39	0.00	0.00	3178.39	3237
Organic fertilizer	ERIVaS	0.24	0.43	0.00	0.00	1.00	577
	EACI 2017	0.46	0.50	0.00	0.00	1.00	3237
		***					
Mineral/Chemical fertilizer (proportion of plots)	ERIVaS	0.30	0.46	0.00	0.00	1.00	577
	EACI 2017	0.19	0.39	0.00	0.00	1.00	3237
		***					
Pesticides (proportion of plots)	ERIVaS	0.50	0.50	0.00	1.00	1.00	577
	EACI 2017	0.13	0.33	0.00	0.00	1.00	3237
		***					
Household Size	ERIVaS	19.14	11.79	2.00	16.00	60.00	577
	EACI 2017	12.55	8.36	1.00	11.00	74.00	3237
		***					
HH head is Female	ERIVaS	0.01	0.08	0.00	0.00	1.00	577
	EACI 2017	0.01	0.10	0.00	0.00	1.00	3237
HH head is literate	ERIVaS	0.08	0.28	0.00	0.00	1.00	577
	EACI 2017	0.20	0.40	0.00	0.00	1.00	3237
		***					

Notes: To ensure comparability of the variables, the EACI 2017 sample is limited to pure stand plots. The stars below the sample mean values of the two surveys denotes the statistical significance of the difference between the two means: $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Table A5: Plots and inputs usage in EACI 17 for CC and not CC sample

Variable	Entire Sample	No CC	CC	$X_{-cc} - X_{cc}$
Yields				
Self-reported Yield	563.08	546.03	600.94	−54.90
Plots Characteristics				
GPS measured area (Ha)	2.53	2.60	2.37	0.22*
Quantity of Seeds used (Kg/Ha)	13.45	13.84	12.59	1.25
Household labor on plot (persons-days / Ha)	75.75	73.45	80.86	−7.40
Hired labor (proportion of plots)	0.42	0.43	0.40	0.03
Hired labor (persons-days / Ha)	6.75	7.97	4.04	3.93*
Organic fertilizer (proportion of plots)	0.45	0.45	0.46	−0.01
Mineral/Chemical fertilizer (proportion of plots)	0.21	0.20	0.21	−0.01
Pesticides (proportion of plots)	0.13	0.13	0.13	−0.00
Agricultural Practices				
Plot is intercropped	0.10	0.09	0.12	−0.03***
Hitched plow	0.62	0.62	0.60	0.02
Plot was fallowed in the past 10 years	0.20	0.19	0.22	−0.02
Household Characteristics				
Household size	12.80	13.06	12.22	0.84
Head of HH is female	0.01	0.01	0.01	−0.00
Age of head of HH	54.47	54.59	54.20	0.40
Head of HH is literate	0.21	0.21	0.22	−0.00
Welfare Index	-0.26	-0.25	-0.28	0.03
N	3569	2471	1098	

Note: Statistical significance of the difference is based on T-statistics computed using design-adjusted standard errors corrected for clustering at the enumeration area.
*p<0.1; **p<0.05; ***p<0.01

Figure A1: Kernel densities of yields (left panel) and harvests (right panel) for alternative measures of production.

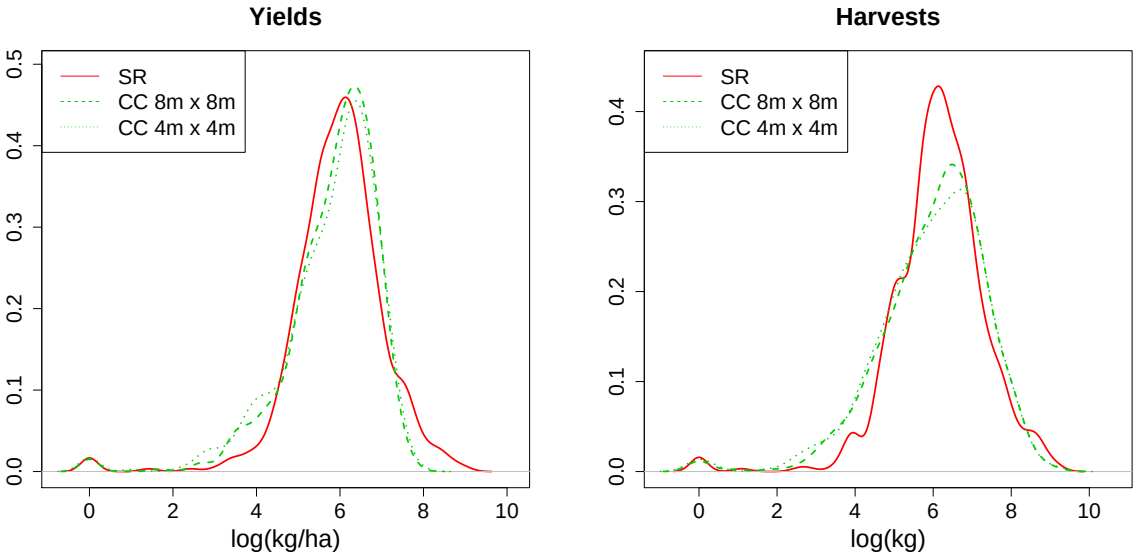


Table A6: Yields - inputs relationship, plots with both self-reported and crop cut in EACI 2017. Full regression results

	Yield Measures	
	SR	CC
GPS measured plot area	−0.170**	0.087**
	(0.086)	(0.042)
Quantity of Seeds	0.262***	−0.008
	(0.069)	(0.039)
Household Labor	0.293***	0.067
	(0.097)	(0.052)
External labor per ha	−0.004	0.011
	(0.060)	(0.034)
Organic Fertilizer use	0.103	−0.028
	(0.092)	(0.071)
Mineral/Chemical Fertilizer use	0.057	0.007
	(0.080)	(0.060)
Pesticides use	−0.143	0.162*
	(0.173)	(0.096)
Plot is intercropped	−0.334	−0.301***
	(0.237)	(0.101)
Plot was inherited	−0.027	−0.076
	(0.160)	(0.103)
Plot was fallowed during the past 10 years	0.033	0.015
	(0.164)	(0.078)
Plot relief^{SR} = plain	0.009	0.018
	(0.129)	(0.061)
Plot soil^{SR} = sandy	0.059	−0.193
	(0.296)	(0.128)
Plot soil^{SR} = clayey	−0.037	−0.177
	(0.302)	(0.118)
Plot soil^{SR} Quality = Good	0.033	−0.023
	(0.113)	(0.066)
Number of structures against erosion	0.175	0.048
	(0.152)	(0.099)
Multivariate Framework?	Yes	Yes
Enumeration area Fixed Effects?	Yes	Yes
Observations	1,098	1,098
R^2	0.925	0.958
Adjusted R^2	0.871	0.928

Notes: Errors clustered at the enumeration area. Significance levels: *p<0.1; **p<0.05; ***p<0.01

Table A6: Yields - inputs relationship, plots with both self-reported and crop cut in EACI 2017. Full regression results -cont.

	Yield Measures	
	SR	CC
Plot manager is female	−0.164	0.285**
	(0.335)	(0.129)
Log age of manager	0.834***	−0.091
	(0.249)	(0.093)
Manager is literate	0.686**	0.030
	(0.345)	(0.097)
Agriculture Equipment index	0.157	0.084
	(0.139)	(0.093)
Number of plot of HH	−0.025	0.007
	(0.018)	(0.010)
HH head is female	−0.037	−0.576***
	(0.473)	(0.215)
Log size of HH head	0.085	−0.045
	(0.086)	(0.056)
Log age of HH head	−0.814***	0.238**
	(0.232)	(0.118)
HH head is literate	−0.724**	0.041
	(0.336)	(0.103)
Welfare index	0.002	−0.040
	(0.091)	(0.049)
Total amount of rain in 2017	−0.019***	−0.012***
	(0.007)	(0.003)
Start of the wettest quarter	0.067***	−0.004
	(0.011)	(0.009)
Distance EA to population center	−0.115***	−0.004
	(0.014)	(0.013)
Crop-cut has dammage	−3.853***	−5.888***
	(0.451)	(0.365)
Percent of pre-harvest losses	−0.010***	−0.003
	(0.004)	(0.002)
Multivariate Framework?	Yes	Yes
Enumeration area Fixed Effects?	Yes	Yes
Observations	1,098	1,098
R^2	0.925	0.958
Adjusted R^2	0.871	0.928

Notes: Errors clustered at the enumeration area. Significance levels: *p<0.1; **p<0.05; ***p<0.01

Table A7: Yields - inputs relationship, plots with self-reported, crop cut and ML in EACI 2017. Full regression results

	Yield Measures		
	SR	CC	ML
GPS measured plot area	−0.086	0.087**	0.083
	(0.060)	(0.042)	(0.052)
Quantity of Seeds	0.162***	−0.008	0.007
	(0.043)	(0.039)	(0.028)
Household Labor	0.323***	0.067	0.026
	(0.083)	(0.052)	(0.065)
External labor per ha	0.086**	0.011	0.004
	(0.040)	(0.034)	(0.037)
Organic Fertilizer use	0.175**	−0.028	0.060
	(0.068)	(0.071)	(0.053)
Mineral/Chemical Fertilizer use	0.119*	0.007	0.059
	(0.063)	(0.060)	(0.057)
Pesticides use	0.058	0.162*	0.013
	(0.083)	(0.096)	(0.064)
Plot is intercropped	−0.378***	−0.301***	−0.162**
	(0.143)	(0.101)	(0.072)
Plot was inherited	0.049	−0.076	−0.020
	(0.082)	(0.103)	(0.082)
Plot was fallowed during the past 10 years	−0.117	0.015	−0.118
	(0.125)	(0.078)	(0.111)
Plot relief^{SR} = plain	−0.052	0.018	−0.089
	(0.105)	(0.061)	(0.078)
Plot soil^{SR} = sandy	0.093	−0.193	−0.095
	(0.162)	(0.128)	(0.102)
Plot soil^{SR} = clayey	0.100	−0.177	−0.133
	(0.163)	(0.118)	(0.100)
Plot soil^{SR} Quality = Good	0.170**	−0.023	0.169*
	(0.072)	(0.066)	(0.095)
Number of structures against erosion	0.083	0.048	0.159
	(0.090)	(0.099)	(0.097)
Multivariate Framework?	Yes	Yes	Yes
Enumeration area Fixed Effects?	Yes	Yes	Yes
Observations	3,569	1,098	3,569
R^2	0.755	0.958	0.769
Adjusted R^2	0.714	0.928	0.729

Notes: Errors clustered at the enumeration area. Significance levels: *p<0.1; **p<0.05; ***p<0.01

Table A7: Yields - inputs relationship, plots with self-reported, crop cut and ML in EACI 2017. Full regression results -cont.

	Yield Measures		
	SR	CC	ML
Plot manager is female	−0.366	0.285**	−0.051
	(0.264)	(0.129)	(0.139)
Log age of manager	0.413	−0.091	0.225
	(0.266)	(0.093)	(0.224)
Manager is literate	0.278**	0.030	0.261**
	(0.135)	(0.097)	(0.114)
Agriculture Equipment index	0.176	0.084	0.081
	(0.123)	(0.093)	(0.093)
Number of plot of HH	−0.011	0.007	−0.011
	(0.012)	(0.010)	(0.013)
HH head is female	0.239	−0.576***	−0.204
	(0.306)	(0.215)	(0.209)
Log size of HH head	−0.023	−0.045	−0.066
	(0.062)	(0.056)	(0.059)
Log age of HH head	−0.382	0.238**	−0.217
	(0.239)	(0.118)	(0.217)
HH head is literate	−0.321**	0.041	−0.354***
	(0.133)	(0.103)	(0.115)
Welfare index	−0.070	−0.040	−0.101*
	(0.062)	(0.049)	(0.060)
Total amount of rain in 2017	−1.723***	−0.012***	−0.192*
	(0.127)	(0.003)	(0.107)
Start of the wettest quarter	0.490***	−0.004	0.034
	(0.025)	(0.009)	(0.021)
Distance EA to population center	−1.492***	−0.004	−0.126*
	(0.079)	(0.013)	(0.068)
Crop-cut has dammage		−5.888***	
		(0.365)	
Percent of pre-harvest losses	−0.032	−0.003	−0.032***
	(0.004)	(0.002)	(0.005)
Multivariate Framework?	Yes	Yes	Yes
Enumeration area Fixed Effects?	Yes	Yes	Yes
Observations	3,569	1,098	3,569
R^2	0.755	0.958	0.769
Adjusted R^2	0.714	0.928	0.729

Notes: Errors clustered at the enumeration area. Significance levels: *p<0.1; **p<0.05; ***p<0.01

Figure A2: Distributions of key variables in ERIVaS and EACI 2017

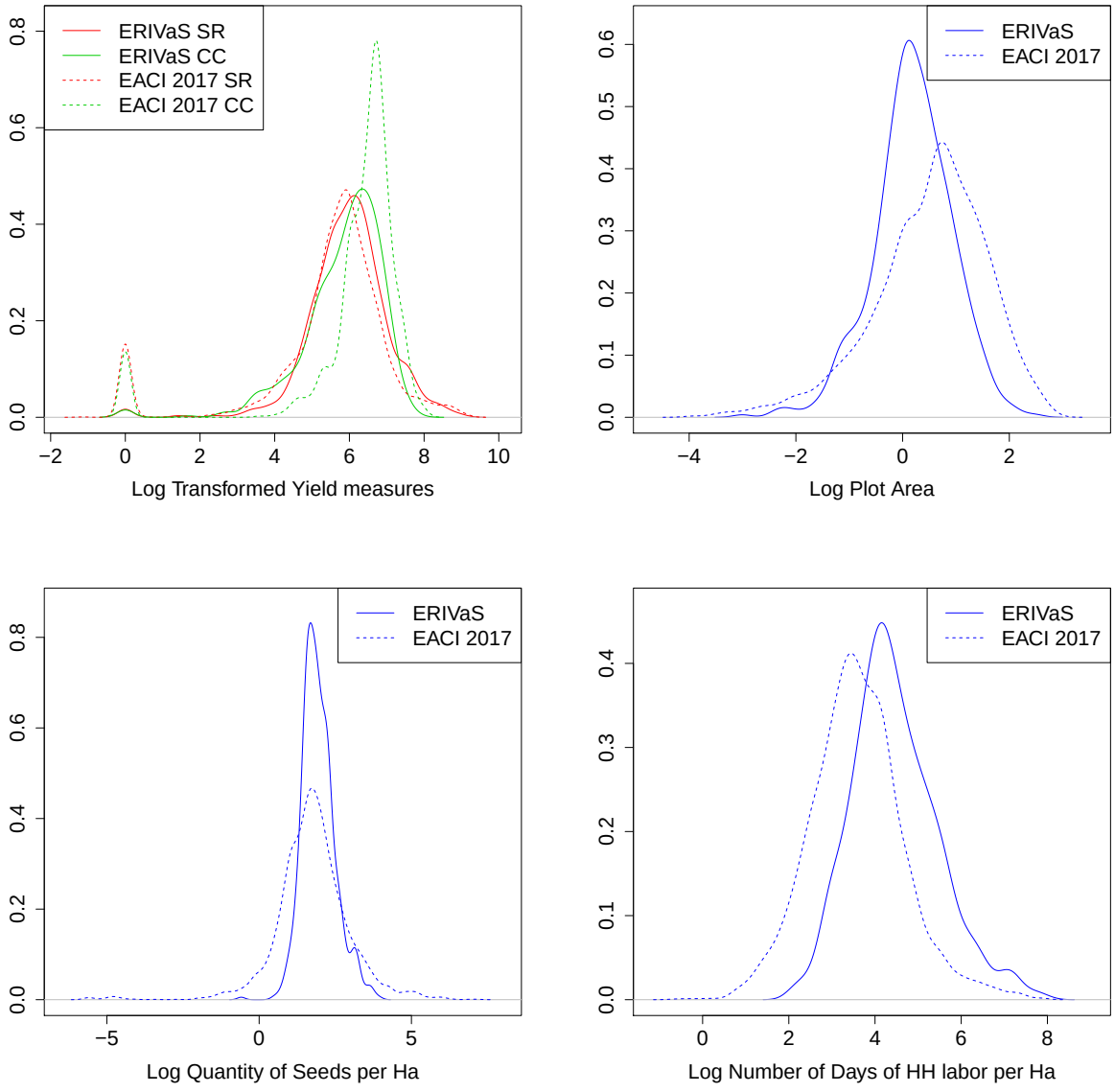
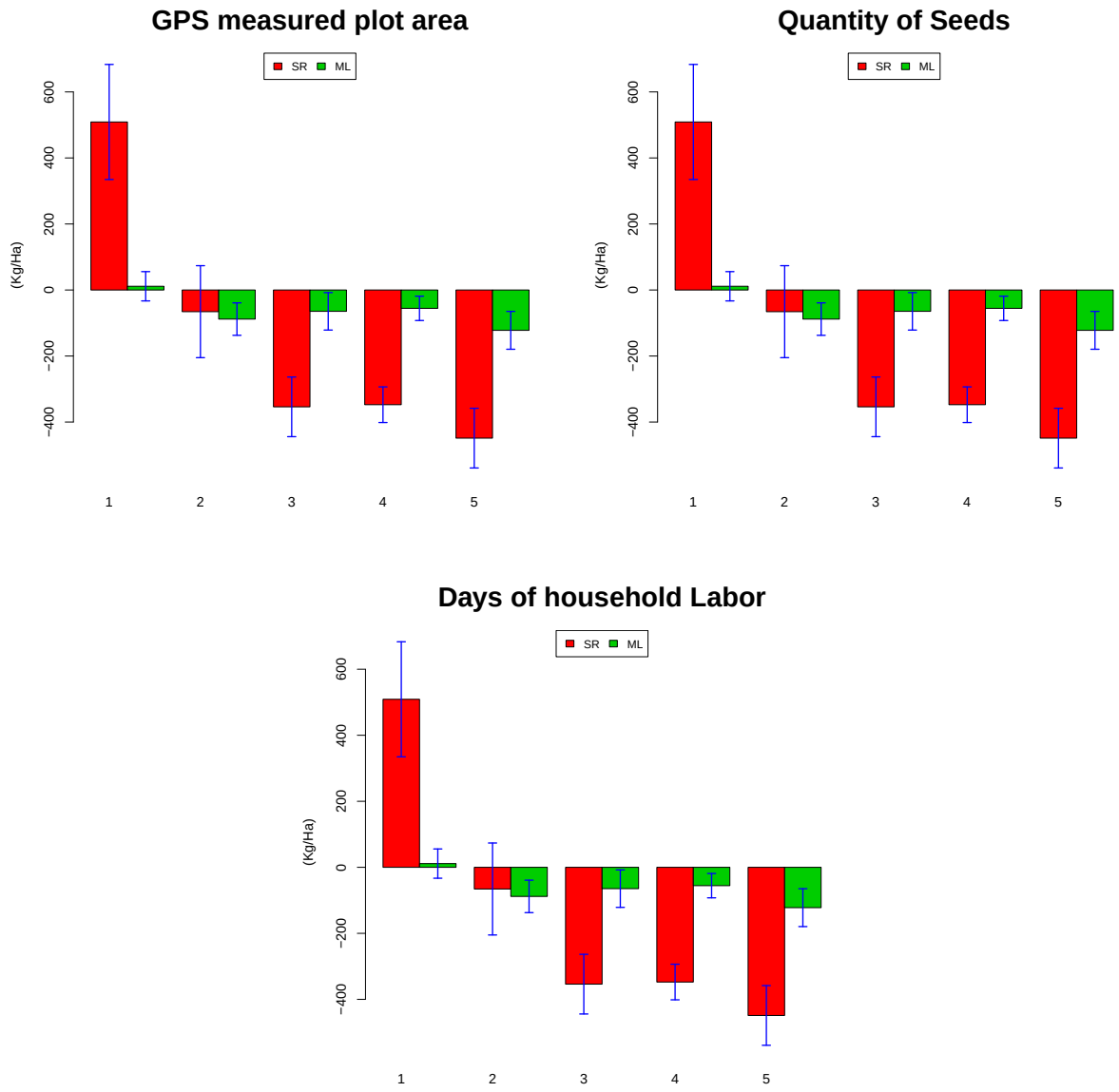


Figure A3: SR and ML over-estimation of CC yield over Quintile of Inputs



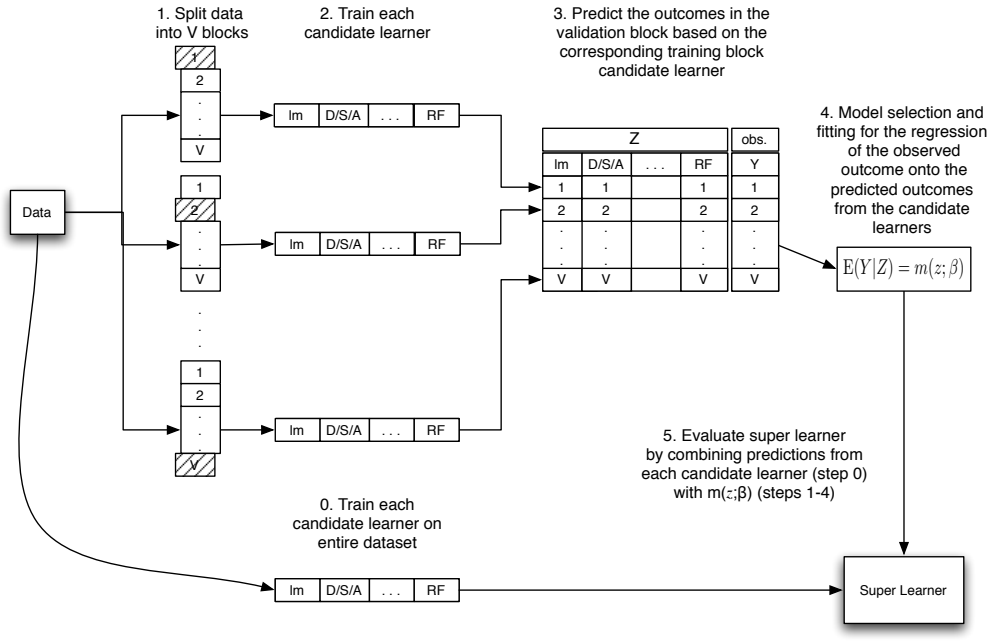
Appendix B Details of the estimation procedure of the yields using machine learning

In using ML to predict outcomes variables such as income, or poverty, the literature advises to train a model following key steps:

- Exclude a test set
- Estimate parameters of models on the training set
- Use cross-validation on the training set to estimate the errors of these models
- Pick the best model or use ensemble methods to optimally stack models together
- Estimate the error on the test set (it is important not to change the model at this point to get a better fit at the risk of over fitting it to the test set)

We use the R package SuperLearner (Polley et al., 2018) which provides a simple way to implement to combine a wide library of machine learning algorithms to train and test an ML model. Ensemble methods are often used to choose the best algorithm using cross validation and ensemble methods. Figure B1 illustrates the procedure when using the R package SuperLearner. [20]. The data is split

Figure B1: Illustration of the SuperLearner application



Source: UC Berkeley ML course

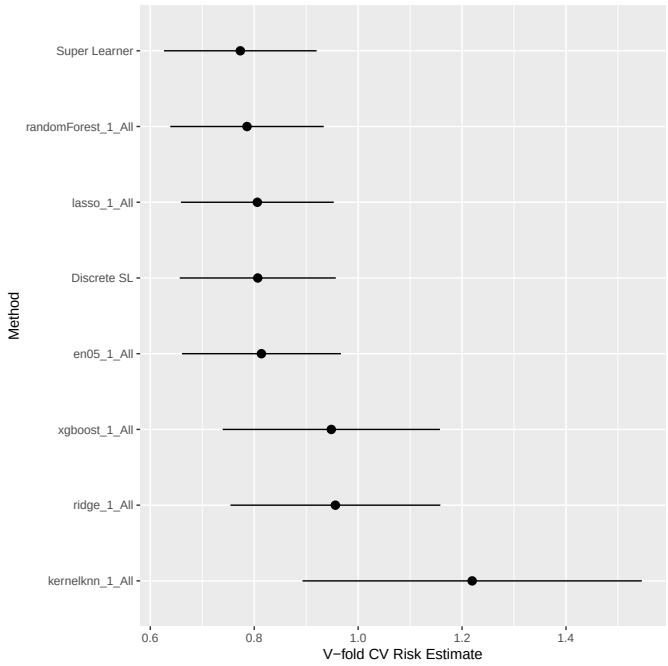
in cross validation blocks (randomly formed mutually exclusive sub-samples of the training data set) leaving one out for validation which is also done on the entire training data set. Then each candidate

[20]For more on the SuperLearner package see: Polley et al. (2018)

learner (ML algorithm) is trained in the cross validation blocks and predictions are made based on the corresponding training blocks in the validation sample. The next step consists in model selection and fitting for the regression of the observed outcome onto predicted outcomes (the observed outcome is regressed on predicted outcomes). Finally, the SuperLearner performance is evaluated by combining the predictions from each ML algorithm. We follow this procedure in the ERIVaS sample. We start our simulation exercise with a training set representing one-third of the total sample (this is the percentage of plots that typically received crop cut subplots in farm surveys which have adopted crop cut methodology as part of the World Bank LSMS-ISA program).

Figure B2 shows the cross validation risk estimate (a measure of the mean square error of the model) using 5 cross validation splits (5 randomly selected sub-samples of the training set on which the ML algorithms are trained independently). We settled on 5 cross validation splits after experimenting with different number of splits. This choice is also supported by the literature: (James et al., 2014) note that "there is a bias-variance trade off associated with the choice of the number of splits and typically, given these considerations, one performs cross-validation using 5 or 10 splits, as these values have been shown empirically to yield test error rate estimates that suffer neither from excessively high bias nor from very high variance." Given the low number of observations in our sample, it is also understandable that we would have to choose 5 splits instead of 10 splits in which case we would have too small a sample to fit the prediction.

Figure B2: Cross Validation errors for each ML algorithms based on 5 fold cross validation (test set size is 33% of entire sample). The figure shows each ML algorithms used by the SuperLearner and the measure of the mean square errors in the validation samples. The ML model with the lowest error (risk) is a combination of the ML algorithms.



In the end, the risk is minimized for the ensemble (the package finds optimal combination of ML

algorithms predictions that minimize mean squared errors) with a 5-fold cross validation splits. The linear combination of learners that results in the minimal risk is recorded in Table B1. The algorithms chosen by the SuperLearner by order of importance are an equally weighted linear combination of Lasso and Ridge, Lasso, a boosted Random Forest, and Random Forest.

Table B1: Optimal combination of learners determined by the SuperLearner Package. The SuperLearner Package fit a prediction with all the ML algorithms that are available and provide a linear combination of the fitted predictions that resulted in the lowest mean square error. Here we show the combination when the training set is a third of the sample. Note that all the algorithms are regression algorithms.

Algorithms	Coefficient
Lasso	0.135
Linear combination Lasso and Ridge	0.681
Ridge	0.000
Random Forest	0.053
Kernel Regression	0.000
Boosted Random Forest	0.131

Appendix C Field work protocol inputs

C.1 Field work protocol for crop cutting

The following is a translated (from French) excerpt from the ERIVaS field staff manual. In ERIVaS crop cut sub-plots of 8x8m subdivided in four 4x4m quadrants were set up.

There are three aspects to this exercise – the first is conducted with the post-planting questionnaire and the last two are conducted at the time of harvest. The first aspect is the selection of a random 8m x 4m crop cutting subplot within the plot. Using the rope, the 4mx4m subplot will be divided into four 4mx4m quadrants. The 8x8m subplot will be selected using random number tables. The materials that you will need for use in this exercise are:

- **Section H (crop cutting Questionnaire)** will be completed as part of the post-planting visit, when the crop cutting subplots are demarcated.
- **Compass**
- **Pre-measured 8m x 8m PVC pipe:** A set of PVC pipes that are pre-measured to create 8x8 meter square will be provided to each enumerator to ensure the crop cut area is precisely 8x8 meters.
- **Pre-measured 4m x 4m PVC pipe**
- **Sticks (8) for Area Demarcation**
- **Measuring Tape:** This is a distance-measuring instrument marked in metric-units (segments), which will be used to determine the location of the areas in the plot.
- **Rope** (32+ meters per household)

Procedure for Crop Cutting

We will be conducting crop cutting on a 8m x 8m subplot. However, we will divide the 4mx4m area into four 2mx2m squares (also called quadrants). Therefore, there will be a total of FOUR 4m x 4m quadrants. The harvest of each quadrant will be recorded separately. Here, we describe in further detail each of the four main aspects to the crop cutting exercise.

You will first construct the 4m x 4m subplot by following steps 1 and 2 below.

1 **Crop Cutting Area Selection:**

   a. Use Random Number Table # 1 to identify the corner from which you will start. Use the first number in the random number table that matches one of the corners of the plot. The corner in which you started the area measurement, the northwest corner, is corner# 1. Corner # 2 is the next corner of the plot, moving around the plot clockwise.

   b. Measure the distance of the two sides along the selected corner with the measuring tape. Identify which is the longer side and which is the shorter side.

   c. Take the bearing from the start corner down the shorter side. Note this in your notebook.

      * Open the mirror of the compass and hold the compass at the level of your eyes such that you can read the image of the capsule in the mirror

- Align the objective with the short side direction
- Without moving the compass turn the capsule such that the 0 degree mark is in the direction of the short side
- The bearing is the angle indicated by the North mark. .

d. Use the Random Number Table #2 provided for this household. The first number should be the number of meters that you will walk along the length of the longer side of the plot. If the first number is larger than the length of the side, choose the next random number (and so on, until you find a number that is less than the length of the side). For example, if the length of the longer side is 25 meters and the first random number in the list is 28, move on to the next number.

e. Beginning at your starting point and continuing along the longer side of the plot, walk the number of meters indicated by your random number.

f. Turn into the plot so that your bearing is the same as the bearing you measured down the shorter side of the plot. This means you will be entering the plot parallel to the shorter side. Choose the next random number from Random Number Table #2 that is shorter than the length of the shorter side and walk the number of meters indicated by this second random number. You should be walking in a direction that is parallel to the shorter edge of the plot. Walk in a straight line. Try not to veer to the right or left to avoid shrubs or wet spots.

g. The corner of the crop cutting subplot is located where your foot lands on the last step: this is point A.

C.2 Farmer Self-Reported Production Estimates

Table C1: Village level conversion factors used for self-reported production

Unit	Mean	Std.De	Frequency
kg	NA	NA	136
Sheaf	0.898	0.110	144
Bag-50kgs	NA	NA	5
Bag-100kgs	97.243	9.887	103
Donkey-cart	178.543	47.512	157
Ox-cart	153.524	17.931	29

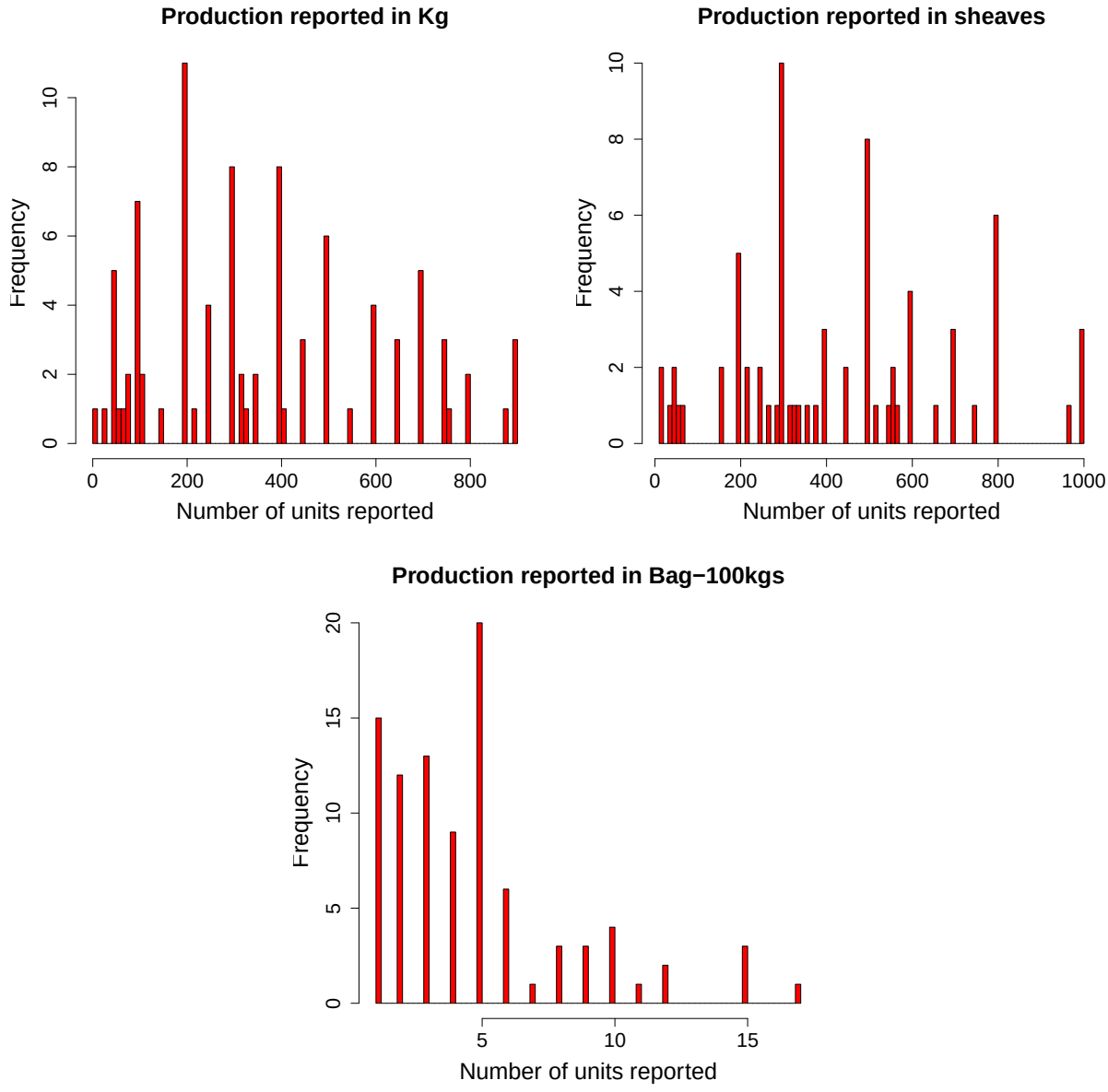
Notes: We show the conversion factors adjusted for shelling losses.
Bag-50kgs were declared in grains conditions so there's no CF adjusted for shelling.

Table C2: Descriptives by heaping status

Unit	Heaped	Not Heaped	Differences in Means
Plot area GPS (ha)	1.259	1.666	***
$SR - CC$	188.347	118.188	
SR/CC	2.283	1.970	
Proportion	0.256	0.744	

Notes: Production tagged as heaped if self-reported as multiple of hundreds when less than a ton and reported in kgs sheaves or if self-reported as multiple of 500 hundreds when more than a ton. and reported in kgs or sheaves or if self-reported 1, 2, 3, 4, 5, or 10 Bag-100kgs
*p<0.1; **p<0.05; ***p<0.01

Figure C1: Histograms showing distribution of self-reported production by units



CONTACT US AT

info@50x2030.org

www.50x2030.org