Policy Research Working Paper 10964

Yielding Insights

Machine Learning-Driven Imputations to Filling Agricultural Data Gaps

Ismaël Yacoubou Djima
Marco Tiberti
Talip Kilic

WORLD BANK GROUP
Development Economics
Development Data Group
November 2024

Policy Research Working Paper 10964

Abstract

This paper addresses the challenge of missing crop yield data in large-scale agricultural surveys, where crop-cutting, the most accurate method for yield measurement, is often limited due to cost constraints. Multiple imputation techniques, supported by machine learning models are used to predict missing yield data. This method is validated using survey data from Mali, which includes both crop-cut and self-reported yield information. The analysis covers several crops, providing insights into the importance of different predictors, including farmer-reported yields and geo-spatial variables, and the conditions under which the approach is valid. The findings show that machine learning-based imputations can provide accurate yield estimates, especially for crops with low intercropping rates and higher commercialization. However, survey-to-survey imputations are less accurate than within-survey imputations, suggesting limitations in extrapolating data across different survey rounds. The study contributes valuable insights into improving cost-efficiency in agricultural surveys and the potential of imputation methods.

This paper is a product of the Development Data Group, Development Economics. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at http://www.worldbank.org/prwp. The authors may be contacted at iyacouboudjima@worldbank.org; mtiberti@orldbank.org; and tkilic@worldbank.org.

Yielding Insights: Machine Learning-Driven Imputations to Filling Agricultural Data Gaps*

Ismaël Yacoubou Djima† Marco Tiberti† Talip Kilic†

JEL Codes: C53; C55; C83; Q12.

Key words: Smallholder farming, Agricultural Crop Yields Measurements, Machine learning, Missing Data, Multiple Imputation, Household Surveys

*The authors would like to thank (i) Ksenia Abanokova, the participants of The Ninth International Conference on Agricultural Statistics (ICAS IX) for their comments, (ii) the Statistics Unit of the Ministry of Agriculture in Mali (CPS/SDR) for the successful implementation of the agricultural surveys that this study leverages, (iii) Giulia Ponzini who co-lead the technical support to CPS/SDR for the implementation and the data curation of the surveys data used for this study. 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. 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.

†Living Standards Measurement Study (LSMS), Development Data Group, World Bank.

1 Introduction

Microdata from agricultural surveys is an essential ingredient for empirical analysis of the economic life of smallholder farmers. Particularly in Africa, farm surveys remain the backbone of agricultural statistics (Carletto et al., 2015). However, collecting accurate data in low-income countries through agricultural surveys is challenging due to various factors, such as the complexity and seasonality of agricultural operations, illiteracy among respondents, and unfamiliarity with standard units of measurement. Overcoming these challenges requires the implementation of complex and costly survey operations. Therefore, the cost-effectiveness and design efficiency of farm surveys, as well as the use of statistical methods to overcome the analysis challenges of their data, remain active areas of methodological research to which this paper intends to contribute.

The analysis focuses on crop yields at the plot level. Generally, plot-level crop yield statistics obtained from agricultural surveys are calculated using either farmers' reported harvest weight or enumerators' measured crop-cut harvest weight. Research has shown that self-reported crop yields are prone to non-classical measurement errors and tend to be higher than objectively measured yields obtained through crop cutting (Abay et al., 2019, Desiere and Jolliffe, 2018, Gourlay et al., 2019, Yacoubou Djima and Kilic, 2024). Although enumerator-measured crop-cutting is widely considered the most precise method for quantifying agricultural output (Fermont and Benson, 2011), its implementation is more costly and time-intensive. As a result, this information is often available only for a subset of the survey, leading analysts to rely on statistical methods and imputation techniques to fill these data gaps.

With the recent advances and their off-the-shelves availability, machine learning (ML) algorithms have become go-to tools of analysts to perform prediction exercises so including them as part of an imputation exercise is a natural next step. Economics research has seen the increasing use of ML techniques in recent years, with proven effectiveness. According to Athey (2018), evaluation of ML's early contributions to economics, ML techniques have been particularly successful when applied to prediction-based problems, such as the one addressed in our study. In addition, utilizing ML in agricultural economics research is a logical approach to model critical variables, such as crop yield, which involves a complex combination of factors, including soil quality, weather, input timing, and management choices, which have non-linearities and interactions (Storm et al., 2019). Previous research attempting to estimate crop-yields in the context of smallholder farmers using ML techniques often attempt this exercise combining manually-labeled optical imagery, and ground data collection, including as part of household and farm surveys. Azzari et al. (2021) provide examples of these studies and recommendations on how large-scale household surveys should be conducted to generate the data needed to train models for satellite-based crop type mapping in smallholder farming systems. Our study relies essentially on survey data integrated with geospatial variables to make crop yield predictions at a granular level. Yacoubou Djima and Kilic (2024) attempt to apply ML techniques for the prediction of yields, leveraging survey data from Mali, to validate an alternative approach to estimate the relationship between crop yields and inputs. The authors conduct 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. This approach allows the authors to replicate the relationship between yields and input obtained with crop-cut yields using the predicted yield showcasing the validity of this imputation framework. However, their exercise is conducted with one crop (sorghum) and only considers one agricultural season limiting the external validity of the framework they develop.

Against this background, our analysis attempts to address these limitations. We take advantage of the availability of two consecutive rounds of the nationally representative agriculture survey of Mali collected as part of the Living Standards Measurement Study - Integrated Surveys on Agriculture (LSMS-ISA)[1] project to validate their approach when used in a survey-to-survey imputation framework. Moreover we extend the imputation exercise to more crops: in addition to sorghum, we work with millet, maize, rice, groundnut and cowpea allowing us to generalize the conditions on which the approach is valid. For reasons of computational ease and tractability, we focus on the estimation of the mean value of yields. We go beyond within-survey imputations exercises to look at survey-to-survey imputations. We also explore which set of covariates are more potent for the imputations exercises. Finally we also explore the effect of training sample sizes and strategies. By varying the size of the training sample, we investigate whether reducing the sample size can still yield reliable imputation results, with the aim of balancing cost efficiency and the accuracy of imputation methods.

Our main findings are threefold: (i) farmers' reported yields are good predictors of crop-cut yields but the results point to a greater predictive power of integrated geo-spatial variables; (ii) on average the imputation exercises work better for crops that have low intercropping rate, and are more commercialized, which can be related to the accuracy in standard units of the farmer-reported yields; (iii) in most of the cases, the imputation approach provides accurate results in the within-survey imputation framework and collecting crop-cut measurements on a sub-sample of plots can offer a cost-effective approach while still achieving reliable ML predictions of CC yields, but less so in the survey-to-survey framework especially when statistics are computed at a desegregated level. These results point to important cost savings in survey operations, but also highlight the stringent survey data requirements in terms of quality.

The paper is organized as follows. Section 2 describes data situations in which crop-cut yields are missing and how the data at our disposal allows us to investigate imputation frameworks that are suited to handle the situations. Section 3 describes the survey data. Section 4 presents the empirical approach. Section 5 discusses the results and section 6 concludes.

2 Data situations

Contexts where yields data based on crop-cuts are unavailable at large scale are very common and it has led researchers to develop statistical methods to overcome such scarcity. Table 1 provides an overview of data situations as it pertains to crop-cut yield and the potential

[1]The Living Standards Measurement Study - Integrated Surveys on Agriculture (LSMS-ISA) is a unique system of longitudinal surveys designed to improve the understanding of household and individual welfare, livelihoods and smallholder agriculture in Africa. Partners countries include: Burkina Faso, Ethiopia, Malawi, Mali, Niger, Nigeria, Tanzania, and Uganda. For more information see: https://www.worldbank.org/en/programs/lsms/initiatives/lsms-ISA.

imputation solutions to fill the gap for this measure. We group these situations into two broad categories. In the first category, where the data is "partially missing", the household survey collects plot level information and agriculture practices, and crop-cuts on a subsample usually selected randomly. A within-survey imputation framework is applied in this case. In the second category, the household survey has plot level information on agriculture practices and self-reported crop production and no harvest data based on crop-cutting. In this case an across survey imputation framework is applied. We provide more details for each of these cases in the reminder of this section.

2.1 Partially missing: Within-survey imputation

The situation of partially missing crop-cut data is a common scenario in most farm surveys.[2] In such cases, crop yield measurements are obtained in two ways: through relatively inexpensive, but less accurate, self-reported harvest data, and through more fieldwork-intensive, expensive, but more objective measures acquired via crop-cutting on a subsample. This survey design is adopted for cost-efficiency and has the advantage of providing multiple measures of crop yields for a small proportion of plots in a single survey.

In the case of LSMS-ISA surveys, several country programs[3] have integrated crop-cutting into their survey operations. The proportion of the sample with crop-cuts (CC) varies across programs. Generally, a list of plot-crop combinations is made at the Enumeration Area (EA) level. If the number of these combinations exceeds a minimum threshold, a random sample of plots is selected for crop-cut implementation.[4]

This survey design is conducive to within-survey (within agricultural season) imputation. A key reason for this is that plot-level information and agricultural practices are collected for all plots in the survey, providing a common basis for the covariates used in statistical modeling. In this context, an important question arises: What proportion of crop-cuts is necessary to obtain valid crop yield statistics? This paper addresses this question by varying the proportion of the sample used to train the model.

2.2 Completely missing: Across surveys imputation

In general, the scenarios we consider for completely missing crop-cut data involve living standard surveys with integrated agricultural modules where crop-cut data were not collected, yet there exists a farm survey within the statistical system that does collect plot-level agricultural practice data. To illustrate, consider Mali or Ethiopia, which conduct repeated cross-sectional crop cutting annually. One might question whether it is feasible to conduct crop cutting less frequently—perhaps every other year or every three years—and rely on data imputation to bridge the gaps while reducing costs. This approach is particularly pertinent for these countries, given

[2]Among countries in West Africa conducting annual farm surveys, Burkina Faso with the Enquête Permanente Agricole (EPA) conducts crop-cuts on all plots of the households sampled.

[3]Mali, Ethiopia, Uganda, and Malawi are among the countries with a survey design that includes crop-cut collection on a subsample.

[4]Details of the crop-cut sampling protocol can be found in the survey documentation manual on The World Bank Microdata Catalog: https://microdata.worldbank.org/

their recent experiences with conflict. Imputation techniques could be invaluable in fragile and conflict-affected regions, offering a way to circumvent the challenges and practical difficulties of extensive crop-cutting surveys. These techniques provide a more efficient and cost-effective method for estimating crop yields, enabling timely and reliable data analysis. Moreover, they can address data collection disruptions caused by conflict or political instability. By imputing missing values, we can maintain the continuity of data series, leading to more complete and consistent datasets for analysis. This is crucial for assessing agricultural trends and evaluating interventions in situations where direct data collection is challenging.

Another practical application of survey-to-survey imputation of crop yields is seen in the WAEMU harmonized survey program. Here, national statistics offices in West Africa collect plot-level agricultural information while conducting parallel farm surveys with crop-cut operations either in the same or the previous year. An across-survey imputation model could effectively fill these crop-cut data gaps.

In this study, we focus on evaluating the validity of the imputation procedure by forecasting (and backcasting) mean crop-cut yields in the 2018 (2017) dataset using the 2017 (2018) crop-cut sample as the training sample.

3 Data

We have access to data from two nationally representative surveys of rural areas in Mali, conducted during two successive agricultural campaigns. The Enquête Agricole de Conjoncture Intégrée aux Conditions de Vie des Ménages 2017 (EACI-2017), administered in the 2017-2018 season, covered 8,398 households, while the Enquête Agricole de Conjoncture 2018 (EAC-2018), carried out in the 2018/2019 season, involved 8,225 households. These datasets were collected by the Statistics Unit of the Ministry of Agriculture, in collaboration with the LSMS-ISA team of the World Bank. This collaboration resulted in repeated cross-sectional surveys of households (in the same Enumeration Areas, or EAs) that gathered comprehensive data on household characteristics and living conditions. The surveys, with a focus on agriculture, provided detailed and precise plot-level information. For each edition of the survey, households were visited twice: the same households interviewed on the first visit were re-interviewed on the second visit, and the visits were planned to match the timing of the post-planting (between August and October) and post-harvest (November and February) periods of the 2017/18 and 2018/19 agricultural rainy seasons. Semi-resident enumerators conducted these visits, with an additional visit between the two for crop-cut harvesting. In both survey waves, crop cuts were carried out for all crops grown during the main agricultural season. A third of the plots per crop were randomly selected for crop cutting from a list of all plots, which was stratified by the crops cultivated by the sampled households within an enumeration area.

3.1 Summary statistics and sample counts

In this analysis, we concentrate on six crops that appear in 80% of all observations and for which non-standard unit conversion factors are used: millet, sorghum, rice, maize, cowpea, and groundnut. Table 2 presents the sample details for these crops across each survey wave.

In total, we have approximately 31,032 observations: 21,657 from the 2017 wave and about 9,375 from the 2018 wave.[5] For each crop, we have slightly less than a third of observations with both crop-cut and self-reported production information. This is in line with the crop-cutting protocol of the survey, which stipulates that crop-cuts are conducted randomly on a third of all plot-crop combinations in an Enumeration Area (EA). For a more comprehensive view, detailed tables with summary statistics of yields by crop, cropping method (either pure or intercropping), and whether the plot underwent a crop-cut are included in the appendix, specifically in Tables B.1 and B.2.

To validate the imputation framework, our analysis will focus on the subset of the sample where both Crop Cutting (CC) and Self-Reported (SR) yield data are available. It is crucial to ascertain whether the plots selected for crop cutting are representative of the broader sample. This aspect is examined in Table 3 and Table 4, which present summary statistics for SR yields[6] and plot areas for each crop in the 2017 and 2018 campaigns, respectively.

The results reveal that the mean differences between plots with CC and those without are not statistically significant, with the exception of rice in 2017. Regarding plot area, the non-crop cut sample has larger plots of sorghum and maize in 2017, and larger millet plots in 2018. Despite these disparities, the overall evidence suggests that the randomization process was effectively implemented, and the characteristics of the crop cut plots do not systematically differ from the non-crop cut plots. This finding is particularly relevant for the within-survey imputation exercise, where understanding the nature of data missingness is crucial. The results support the application of the Missing At Random (MAR) framework, indicating that the model is likely trained on a representative sample. Finally, the presence of zeros in the dataset may raise some discussion about the choice of the imputation methods and the analysis results. In our dataset the zeros yields are real zeros as they denote plots that did not produce any outputs. In addition, the share of zeros is quite low. Overall, 5% of the observations have 0 yields. By crop, the share of observations with 0 yields ranges from 2% for millet to 13% for maize. For these reasons we can reasonably argue that our estimates are not biased by the presence of zeros in the dataset.

3.2 Measurement errors in SR yields

In this paper, we capitalize on the unique opportunity provided by having data on harvest from the same plots measured using two distinct methods: crop cuts and self-reported production. Previous studies, such as those by Abay et al. (2019), Desiere and Jolliffe (2018), Gourlay et al. (2019), Yacoubou Djima and Kilic (2024) have explored similar datasets in Ethiopia, Uganda, and Mali, particularly for maize and sorghum, to investigate measurement errors in yield reports. Our descriptive analysis corroborates these findings. We observe that self-reported yields demonstrate the same pattern of correlated measurement errors with plot area, as documented in the literature for all crops examined. Figures 1 and 2 graphically illustrate

[5]For the 2018 wave, we limited our analysis to plots of households that did not cultivate the same crop on multiple plots because self-reported harvest data was collected at the crop level.

[6]SR yields were adjusted: yields exceeding six times the national norms provided by the Ministry of Agriculture were normalized to the mean yields at the Enumeration Area (EA) level, or to district or regional levels if needed.

this phenomenon. They show the difference in yields (SR-CC) across quintiles of plot area (measured via GPS) for the entire sample. These figures highlight a consistent correlation between the SR-CC difference and the quantity of cultivated area, as noted in the literature: farmers' overestimation is more pronounced on smaller plots and decreases, or even reverses, on larger plots.

The average overestimation of SR yields in comparison to CC yields in Mali is likely influenced by the fact that the average plot area in Mali is larger than in the studies previously mentioned. This larger average plot size means that the underestimation trend observed in other studies begins earlier in the distribution for Mali. Generally, SR yields are overestimated compared to CC yields in the first plot area quintile. However, in each of the subsequent quintiles, SR yields tend to be underestimated. The distinct trend observed for cowpea may be attributed to the fact that cowpeas are typically planted on smaller plots. Despite this, the general trend of decreasing overestimation across land quintiles is still evident, though it is not a linear decrease.

Additionally, the high confidence interval reported for cowpea yields may be attributed to the prevalence of intercropping in its cultivation.[7] Accurately self-reporting harvest quantities in intercropped plots can be particularly challenging. As for groundnut, we observe consistent underestimation across all plots. This finding could be linked to the findings by (Bardasi et al., 2011), which suggests that women's labor tends to be underestimated by men. A plausible explanation for the underestimation observed in our study is that groundnuts are predominantly cultivated by women, while the majority of respondents for the 2017 and 2018 surveys were men, who usually tend to underreport the product harvested by women. According to EACI-17 and EAC-18 data, groundnut plots, which are managed by women at a higher rate than average (18% compared to the overall average of 7%), also have a high rate of proxy respondents. Specifically, 42% of the information about these plots is not provided directly by the manager.

4 Empirical approach

4.1 Validation approach

We have structured our empirical framework to identify the limitations of imputation accuracy and to offer recommendations for capturing the essential data set. This includes the necessary variables, the most effective machine learning (ML) modeling approach, and the minimum sample size for robust statistical analyses of crop yields when crop-cut data is either completely or partially missing.

To determine the required set of variables and the most effective ML modeling approach, we train our model on 75% of the plots with crop-cut data, then make predictions on the remaining 25%. This process is carried out separately in the EACI 2017 and EAC 2018 datasets.

To explore the limitations imposed by survey sampling when using crop-cut data to train an imputation model, we experiment with the ML approach, and with varying sample sizes of

[7] Among the six crops analyzed, cowpea has the highest rate of intercropping, with 27% of the plots cultivated with cowpea being intercropped.

the training dataset. We use proportions of 10, 20, 33, 25, 50, 67, and 75 percent of the crop-cut measures, with the rest of the sample serving as the test set in each scenario. These simulations are performed using the EACI 2017 dataset, which has the largest sample of crop-cut plots.

To assess the robustness of statistical analyses using imputed crop yields, our approach involves several key steps. First, for the within-survey imputation procedure, we train the ML model on one-third of the plots that have crop-cut data and test the model on the remaining two-thirds. This allows us to evaluate the effectiveness of the model in accurately imputing yields within the same survey. Secondly, for the survey-to-survey imputation procedure, we train the machine learning model on the entire crop-cut sample from EACI 2017 and then test its performance on the crop-cut sample from EACI 2018. This step further our understanding of how well the model can generalize and impute data across different survey years. Lastly, we invert the training and test samples to experiment with backcasting. In this process, we train the model on the EACI 2018 crop-cut sample and test it on the EACI 2017 sample. This reverse application provides insights into the model's capability to predict past yields based on more recent data, further testing the limits of the imputation model's robustness.

4.2 Imputation framework

Using machine learning (ML) for the imputation model

In recent years, ML techniques have increasingly been incorporated into economics research, demonstrating significant utility in various applications. As noted by Athey (2018), ML methods have shown remarkable success in predictive problems, much like the one we address in our study. Introducing ML into agricultural economics enables the modeling of key variables, such as crop yields, which are influenced by a complex array of factors including soil quality, weather conditions, timing of inputs, and other management choices, often characterized by non-linearities and interactions.

Moreover, ML algorithms are specifically designed to prevent overfitting to the training dataset. This is achieved through techniques like regularization and the empirical selection of tuning parameters. While overfitting is generally not a primary concern in missing data imputation for within-survey applications, this feature of ML algorithms becomes particularly advantageous in survey-to-survey imputation scenarios. Here, the ability of ML to generalize across different datasets without overfitting is a significant benefit, enhancing the reliability and applicability of the imputation results.

One complication in our setting is that the set of covariates we use has a hierarchical structure: we conduct imputation at the plot level, as each crop, even when intercropped, is modeled separately. However, we include covariates at both the household and EA levels. A potential issue with using ML algorithms in a non-i.i.d. setting is that test performance may be overly optimistic. This occurs when selecting covariates that identify household characteristics or EAs that are good at predicting yields without capturing the idiosyncratic details of plots that may affect yield values. This, in turn, would decrease the predictive performance of the ML model at the plot level. To mitigate this issue, we have interacted geospatially extracted characteristics with plot characteristics that make sense, for example, the date of planting and the amount of rainfall per dekad. This is not done systematically for all variables to keep

the set of covariates manageable from a computational standpoint. Additionally, we split the training data to ensure that algorithms are trained and cross-validated across small clusters when possible. This reduces instances where plots from the same EAs are all clustered in the same training samples.[8] Ngufor et al. (2019) suggests a two-step procedure, which we intend to apply as a robustness check to investigate the improvements it brings in terms of performance.

Outcomes variables

We employ ML methods to impute both yield levels and log yields.[9] In this process, we calibrate the model using Crop Cutting (CC) as the outcome variable. CC is regarded as the 'gold standard' method for measuring yields at the plot level.

Covariates

Our covariates set is made of about 162 variables[10] in the following categories: (1) Self-reported yields[11]; (2) Plot characteristics, excluding the quantity of key plot inputs.[12] This category includes land tenure, soil type, fertility measures, distance from the household, etc.; (3) Climate and rainfall variables, derived from GPS coordinates of the plots; (4) Farmer agricultural practices (such as sowing dates, type of labor) interacted with rainfall measures, along with agricultural and welfare implementation indices; (5) Household characteristics, including composition, demographics of the household head, and welfare measures; (6) Fieldwork variables, encompassing respondent characteristics, dates of visits, and observations of any damage to the crop cut noted by enumerators, survey design variables such as household weights.[13]

Evaluation metrics

To assess the performance of the machine learning models, we utilize the coefficient of determination (R^2) from the regression of Crop Cutting (CC) yields against the ML-predicted yields.

[8]Increasing the number of cross-validation samples is also a way to reduce the decrease in performance, but when we experimented with a higher number of cross-validation samples, the increase in terms of performance was marginal while the computational cost was high.

[9]The logarithm of crop yield is often the preferred variable for regression analyses in empirical research, such as exploring the inverse relationship (IR) between farm size and productivity. Since there is no significant difference between predictions for levels and logs, we have included the results for log predictions in the appendix.

[10]The full list of covariates can be found in the appendix C.

[11]We include self-reported yields in levels for predicting yields and in logarithmic form for predicting log yields.

[12]As analysts often regress yields obtained using ML on the quantity of inputs, we exclude these variables from the ML covariates to ensure that the regression coefficients are not simply a mechanical result of the ML prediction model. We also tested models that included inputs as covariates and found no significant differences in the results.

[13]The inclusion of the weights in the model followed Rubin (1987), who implicitly assumed that the imputer should have access to the variables used to construct any weights and should always include them in the imputation model. However, as pointed out by Quartagno et al. (2019), previous works have shown that following this approach results in an upwardly biased estimate of the variance, although the simulations in these works have shown that this issue is of little practical importance. As a robustness check, we verify whether the inclusion of survey weights has an impact on predictive performance.

While the R^2 does not precisely represent prediction accuracy, it offers an intuitive measure of relative performance, i.e. the proportion of variance in the observed yields that is explained by the predicted yields. The ultimate validity of our proposed imputation framework hinges on the accuracy of the crop yield statistics derived from modeling. Therefore, we examine the difference between the means of the CC yields and those computed using the ML model not only in the test set, but also across the entire sample, both nationally and at disaggregated levels. To put things in perspective, we compare the mean yields obtained through ML with those obtained from self-reported (SR) yields. We also assess whether the ML estimate falls within one standard error or the 95% confidence interval of the observed mean CC estimate, further evaluating the reliability of our imputation approach. We also considered other metrics such as root mean square errors to ensure that it does not impact the conclusion that we make. Our preferred metric remains the adjusted R^2 because it allows us to compare across crops without the average size of yields per crop biasing the conclusions we make.

Multiple imputation

The ML modeling is used in a multiple imputation framework (MI). As such, we generate several imputations, which are different since the cross-validation samples can be drawn differently and thus provide slightly different predictions for each observation. We then follow the MI statistical approach when computing means and standard errors to take into account the variability that comes from the modeling (combining within imputation and between imputation variances). MI, first proposed by Rubin (1987), is a Monte Carlo technique that replaces missing values for a given variable with m $>$ 1 simulated alternatives. MI typically consists of three steps: (i) m imputations (i.e. m complete datasets) are generated based on an imputation model that encompasses a vector of observable covariates that predict the missingness in a given variable, (ii) statistical analysis is performed separately with each of the m complete datasets, and (iii) the results obtained from m complete data analyses are combined into a single set of multiply-imputed parameter estimates and standard errors. The conditions under which valid inferences could be obtained from missing data has been laid out by (Rubin, 1987). Our procedure assumes that data are missing at random (MAR), that is that missing data could be predicted based on observable attributes underlying missingness. While the MAR assumption is not empirically testable, the limits of its tenability could be assessed in our study since the process of missingness is due to the survey design (see discussion in the Data section 3 and precisely Tables 3 and 4). In building the imputation model, the literature advises to include as explanatory variables: (i) the variables appearing in the analysis model that features the multiply imputed variable(s), (ii) the variables that are known to have influenced the occurrence of missing data, and other variables for which the distributions differ between the response and non-response groups, (iii) the variables that explain a considerable amount of variance of the multiply-imputed variable(s) and that help to reduce the uncertainty of the imputations, and (iv) the variables with information on the features of the complex survey design, including stratum and cluster identifiers, and sampling weights. In assessing the validity of our imputed crop-yield, we account for the survey design by weighting our estimations and using design-adjusted standard errors corrected for clustering at the enumeration areas. Since we use a multiple imputation approach, we compute means and standard errors following (Little and

Rubin, 1987) to account for within and between imputation uncertainty. By adhering to these statistical practices, we aim to provide a robust and accurate evaluation of the imputed crop-yield data, ensuring that our findings are reflective of the underlying survey data and the complexities of the imputation process.

5 Results

5.1 Modeling lessons

5.1.1 Top predictors

In this section, we look at the most relevant variables as identified in the random forest algorithm prediction[14] for 50 different simulations. Table 5 shows the top 10 most important variables in the prediction for each crop's yield levels.[15] SR yield variable is an important predictor for crop yield in levels (with the exception of millet), but not as important as for logs predictions (it is the most important predictor of yield in logs only for cowpea). Geospatial variables are consistently among the top predictors for both levels and logs yields. It is reassuring for the machine learning approach to observe that the combinations of farmer practices and different rainfall variables are among the most important predictors for most of the modeled crops (both for logs and levels yield predictions). Indeed, in a context of smallholder farmers like the one in Mali where most of the crop production relies on rainfall, one would expect that yield would be significantly impacted by the complex interaction of timing of the seeding timing and the amount of rain that fall. This underlies the usefulness of collecting detailed agricultural practices information, as well as being able to link the georeferenced data collected with geo-spatial, when conducting a crop yield prediction exercise.

The relevance of SR yield for improving the quality of yield predictions highlights the importance of collecting farmer-reported information on crop production. Therefore, particular attention should be devoted to enhancing the quality of SR yields, for example by reducing the length of the recall period as noted in (Wollburg et al., 2021), and improving the completeness and quality of conversion factors for non-standard measurement units.

To further assess the degree to which the most important variable matter compared to the second one for example, we look at the average value of the importance metric produced by the random forest algorithm for the 50 different simulations we conducted. Figure 3 and Figure 4 exhibit these values for rice yield predictions in levels. Household weights and SR yields appear as one of the top three key variables in the predictions: household weights represent the most important variable for rice yield predictions in levels while SR yields variable is scored as the third most important for rice yield predictions in levels. The plot slope and average annual total rainfall are the second most important variables for levels and logs, respectively. One

[14]For our predictions for log or levels, we rely on the ensemble prediction. But to take a deeper look at the set of covariates, we rely on the covariate importance measure of the Random Forest algorithm, which is the most efficient algorithm for yield predictions for most of the crops when imputing levels, as shown in Table A.1 and it has the lowest error outside the ensemble for the log predictions.

[15]We show the equivalent table for predicting logs in Table B.3.

striking observation from Figure 3 is how much more important SR yields are in comparison to the rest of the covariates when it comes to log predictions. We assess the relative importance of covariates in the accuracy of the prediction in the following subsection.

5.1.2 Relative importance of SR yields and GPS extracted variables

We run the ensemble of methods described in Appendix A for 25 different cross-validation splits to predict CC yields, testing for the inclusion of SR yields and geospatial variables extracted using the EA GPS coordinates[16] among the set of predictors. This approach allows us to evaluate the impact and relevance of these specific variables in the context of our machine learning models and the overall accuracy of the yield predictions.

Table 6[17] shows the impact of including SR yields and/or GPS extracted variables on the quality of the predictions, measured in terms of R^2 for the whole set of crops and individually for each crop. The reference model set up does not include SR or geospatial variables in the set of covariates.

First, including SR yield or GPS extracted variables increases unambiguously the accuracy of yield prediction as the difference in R^2 is for the most part positive and statistically significant. Second, combining SR and GPS provides on average better accuracy than using one or the other set of covariates. Third, there is some heterogeneity in terms of crops. The SR yield variable is more relevant for sorghum and cowpea imputations, although the overall accuracy of prediction of SR yield without including GPS extracted variables is low. GPS extracted variables, both with and without including SR yields, are particularly important for groundnut, sorghum and millet yield predictions. Finally, cowpea yield prediction benefits particularly from the inclusion of both SR yield and GPS extracted variables. The heterogeneity of the impact of the different combinations of covariates for each crop predictions is difficult to interpret especially since the coefficients are in general not statistically different to each other. As such, the main lesson that we retain here is that combining the subjective measure of yields and the objective covariates is always the most beneficial to improve the accuracy of the predictions. This conclusion is also robust to the use of alternative model performance metrics. This is confirmed in Table B.6 which shows how the inclusion of different covariates impacts the performance as measured by the root mean square of the predictions when we model level yields: we find that for all crops including both SR yields and geospatial variables improve the model performance (reduce the root mean square error) the most. The improvement is marginal although significant when SR is included without geospatial variables, highlighting the fact that geospatial variables are the most potent in improving the predictions. Looking at the base root mean square error across crops, it appears that in average the prediction from cowpea yield in average the best prediction, but that is because the average yield for cowpea is smaller than the other crops and as such small deviations from when squared are amplified.

[16]The GPS extracted variables used the publicly available coordinates which have been scrambled for privacy purposes. As discussed in Michler et al. (2022), the degree to which spatial anonymization introduces mismeasurement is a function of which remote sensing weather product is used in the analysis. Care must be taken in choosing a remote sensing weather product when looking to integrate it with publicly available survey data. As such, the results might vary with other types of weather products.

[17]See Table B.4 for the results for log yields, which do not differ significantly from the levels.

We also verified the impact of the inclusion of weights in the covariates by testing it on predictions of millet yields in the 2017 data set. We find that in average, the standard errors of the imputed mean from a model with survey weights was only 0.38% larger than in the case of a model without weights. The bias is also slightly larger as we find than in average, the root mean square error is 0.5% when including weights. As these results seem to be of little practical concern, we believe it will not affect the lessons gleaned from the imputation exercises conducted in this work and we leave changes to the covariates sets that could allow for the inclusion of survey weights[18] for future work.

5.1.3 Prediction accuracy by crops

Figure 5 provides a graphical representation of the distribution of the R^2 of the regression of the CC yields on the ML prediction for yields levels predictions,[19] by crop. There appears to be significant heterogeneity in the performance of the ML algorithms across crops: the R^2 for cowpea ranges between 0.1 and 0.42 while that of rice ranges between 0.41 and 0.7. The comparison of these crops is a good case and point for what is the most likely candidate explanation for the disparity in ML prediction performance across crops: the rate of intercropping, which refers to the practice of growing multiple crops in the same field simultaneously. In fact plots with cowpea are the highest intercropping rate (0.26) while rice is not on intercropped plots. Moreover the median performance of the ML algorithm is in general increasing with the rate of intercropping if we focus on sorghum, maize and groundnut. The predictions for millet have a higher performance than those of sorghum or maize even though it has a higher intercropping rate, but we can attribute that to the fact that the number of plots of millet is significantly more important. Another factor that likely explains the heterogeneity in ML predictive performance across crops is the degree of commercialization. Rice and groundnut are also cash crops, hence farmers of these crops use units that have more standardized weights. This also explains why SR yields appears at a higher predictive rank for rice and groundnut than millet or maize for example (see Table 5).

5.1.4 Training set size

This section assesses the minimum share of plot observations that can be sub-sampled for crop-cutting measurements while still obtaining a reliable predicted measure of CC yields. The predictive accuracy, measured in terms of R^2 (coefficient of determination), of the ML prediction is evaluated for different sizes of the training dataset. The training dataset sizes considered are 10, 25, 33, 50, 67, and 75 percent of the CC measures. This evaluation is performed for each crop individually and for the entire set of crops across 50 different simulations.

The results, depicted in Figure 6, show that the R^2 of CC yields on the ML prediction increases as the percentage of the training dataset size grows. However, the marginal gain in accuracy diminishes as the training dataset size increases, particularly for logs imputation.

[18] One way to correct these issues as noted by Quartagno et al. (2019) is to include in the model (i) the weights, (ii) all the domain indicators (i.e., all relevant covariates of the weighted regression), and (iii) their interactions for valid MI inference in general.

[19] See Figure B.1 for results with the log yield predictions.

While there is some heterogeneity in results across different crops, overall, conducting crop cutting on half of the plot observations provides reasonably high accuracy for deriving reliable imputed CC yields, both for levels and logs imputations. Beyond one-third (33 percent) of the sample, the gain in accuracy becomes marginal. Therefore, considering the higher costs associated with crop cutting, the decision regarding the sample size depends on the desired level of R^2 accuracy needed for the specific application, as well as on the main crops of interest. These implications suggest that collecting crop-cut measurements on a sub-sample of plots, particularly around 50 percent of the total sample, can offer a cost-effective approach while still achieving reliable ML predictions of CC yields. This information can guide decision-making regarding the optimal allocation of resources for data collection and the measurement of crop yields.

5.2 Imputation results

5.2.1 within-survey imputation

In this section we assess the performance of the within-survey imputation procedure by examining the difference of the ML imputed mean with the CC mean. We look at the differences in a test set representing 2/3 of the sample for each crops and we then examine the differences at the national and regional level. In summary, the results of the within imputation show that our within-survey imputation procedure reliably impute means in the test sample: as Table 7 shows, the differences are not statistically significant in the test set. Moreover to ensure that the results are not a result of a random draw of the test set, we repeated the operation for 100 different samples. These results are illustrated on Figure 8 which shows that the distributions of means of the CC and ML overlap. In at least 86% of the times for sorghum and over 94% of the times for the other crops, the ML mean is within the 95% confidence interval of the CC mean in the test set. The good performance of the ML prediction of the mean can be assessed computing the means at the national level (see Table 8) for all crops with 1/3 of the crop-cut sample used as training sample. Computing regional yields at this level is more challenging but this is mostly due to the small sample size[20] at this level. In fact when we perform the exercise with 1/2 of the sample, the accuracy of the ML imputation at the regional level increases (see Table B.5) and more results are statistically not distinguishable from the CC means. Moreover, the imputed mean is almost always closer to the CC mean than the SR means. In many cases the difference between ML and CC mean yields is statistically significant even if negligible in terms of magnitude. This is due to the fact that ML means have, by construction, very small standard errors compared to SR means, as shown in Figure 7.

[20]Given that we only have one year difference between the two survey editions, it may be possible to pool data together and consider it as one sample and conduct a within-survey imputation exercise, for crops with low sample size. We leave this for future research considering that this work aimed to consider cases that fit real data situations.

5.2.2 survey-to-survey imputation

The procedure is also tested in a survey-to-survey imputations. CC yields from the 2018 survey sample are used as a test set, with the model trained with the 2017 survey sample. Table 9 summarizes the results of the survey-to-survey imputation exercise, with 100 percent of the 2017 sample used as training sample. We find that except for millet, rice, and groundnut the difference between the mean obtained with the ML model is statistically different from that of the CC means even at the national level. However, even when the difference between the mean of the ML yields and the mean of the CC yields is statistically different, it remains smaller than the difference of the SR in most cases. The statistical difference that we detect may be due to the fact that despite computing the standard errors following the MI procedure, we are getting standard errors that do not account enough for the variability across imputations. This is also visible in Figure 10 which shows the means computed for 100 simulations. Looking at the regional level estimates, we find that in general these differences are even more pronounced.

The relatively low accuracy of the survey-to-survey imputation may be somehow surprising in this framework, given that the 2017 and 2018 surveys are very similar in terms of geographical coverage, representativeness and survey instruments. Nevertheless, there are some important differences in the questionnaires between the two surveys that can explain the relatively low accuracy of the 2017 sample for imputing the 2018 yields (for example, the seeding date, which interacted with rainfall is a top predictor of yields, was not collected in 2018). (Gourlay et al., 2019) find more promising year-to-year yield predictions using CC yields in Uganda. However, their results rely on a panel of households interviewed in two consecutive years, and predictions are even more accurate when plots from the same parcel are sampled. Interestingly, in this framework predicting the logs results in lower accuracy than predicting the levels, and the difference in R^2 level and log yields is positive and significant for all crops. Overall, it seems that the survey-to-survey imputation provides a very low accuracy level for food staple crops, while it is higher for market-oriented crops, in particular for rice and maize (although lower than the within-survey imputation).

6 Conclusion

We assess the validity of imputation of an objective measure of crop yields in farm surveys in situations in which they are partially missing for efficiency design and when they are completely missing in a survey round. Partial and complete missingness of crop-cut yields might be due to resource constraints and logistical limitations in case certain areas are inaccessible or the presence of enumerators must be reduced in terms of time because of security issues.

The data requirements for the validation process reveal several important findings. First, self-reported (SR) yields are considered significant predictors of crop-cut yields. However, the results indicate that integrated geo-spatial variables make a greater contribution to the accuracy of machine learning (ML) predictions. This suggests that factors such as geographical information, climate data, or other spatial variables play a crucial role in accurately predicting crop yields. Furthermore, the accuracy of ML predictions varies across different crops, indicating heterogeneity. This points to the need for reliable non-standard units when dealing with

crops that are not commercially traded and more accurate yield measurement for intercropped plots. Additionally, the measurement of intercropped crop yields is found to be important in improving the accuracy of ML predictions. Finally, our results suggest that collecting crop-cut measurements on a sub-sample of plots, particularly around 33 percent of the total sample, can offer a cost-effective approach while still achieving reliable ML predictions of CC yields. This information can guide decision-making regarding the optimal allocation of resources for data collection and the measurement of crop yields.

In terms of validation, ML yields are found to be valid in the within-survey imputation framework, even when using a subsample of the crop-cut data. This implies that ML predictions can effectively replace missing values within the same survey. However, in the survey-to-survey imputation framework, ML yields are less reliable, although they still generally outperform SR yields in terms of accuracy. This suggests that ML predictions might not be as robust when extrapolating or imputing values across different survey rounds. Moreover, in both the within-survey imputation and survey-to-survey imputation frameworks, the disaggregated (regional) statistics results are not robust. This indicates that obtaining reliable regional-level statistics from the ML predictions might pose challenges and further investigation or refinement of the methodology may be necessary. Further investigation is also needed to improve the predictions of the tails of the distribution of crop yields, as targeting farmers located in these sections of the distribution is of particular importance for agricultural policy.

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.

Athey, S. (2018). *The Impact of Machine Learning on Economics*, pages 507–547. University of Chicago Press.

Azzari, G., Jain, S., Jeffries, G., Kilic, T., and Murray, S. (2021). Understanding the requirements for surveys to support satellite-based crop type mapping: Evidence from sub-saharan africa. *Remote Sensing*, 13(23).

Bardasi, E., Beegle, K., Dillon, A., and Serneels, P. (2011). Do labor statistics depend on how and to whom the questions are asked? *World Bank Economic Review*, 25(3):418–447.

Carletto, C., Jolliffe, D., and Banerjee, R. (2015). From tragedy to renaissance: Improving agricultural data for better policies. *The Journal of Development Studies*, 51(2):133–148.

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.

Fermont, A. and Benson, T. (2011). Estimating yield of food crops grown by smallholder farmers: A review in the uganda context. *International Food Policy Research Institute discussion paper*, 01097.

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.

Little, R. and Rubin, D. (1987). *Statistical Analysis With Missing Data*. Wiley Series in Probability and Statistics. Wiley.

Michler, J. D., Josephson, A., Kilic, T., and Murray, S. (2022). Privacy protection, measurement error, and the integration of remote sensing and socioeconomic survey data. *Journal of Development Economics*, 158:102927.

Ngufor, C., Van Houten, H., Caffo, B. S., Shah, N. D., and McCoy, R. G. (2019). Mixed effect machine learning: A framework for predicting longitudinal change in hemoglobin a1c. *Journal of Biomedical Informatics*, 89:56–67.

Quartagno, M., Carpenter, J. R., and Goldstein, H. (2019). Multiple Imputation with Survey Weights: A Multilevel Approach. *Journal of Survey Statistics and Methodology*, 8(5):965–989.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley.

Storm, H., Baylis, K., and Heckelei, T. (2019). Machine learning in agricultural and applied economics. *European Review of Agricultural Economics*, 47(3):849–892.

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

Yacoubou Djima, I. and Kilic, T. (2024). Attenuating measurement errors in agricultural productivity analysis by combining objective and self-reported survey data. *Journal of Development Economics*, 168:103249.

Tables

Table 1: Data situations

Type	Extent of missing crop-cut data	Typical survey situation	Imputation model
A	Partially missing	(i) LSMS-ISA with SR on all plots, CC on subsample within the same survey	Within survey imputations
		(ii) LSMS-ISA with SR on all plots CC on subsample CC with plot boundaries	Satelitte crop yield maps
B	Completely missing	LSMS-ISA with SR on all plots	Survey to Survey with SR yields

Table 2: Number of observations by crop for each wave

		2017	2018
Millet	1607	5199 (76%)	1615 (24%)
Sorghum	1155	3591 (63%)	2128 (37%)
Rice	4130	2990 (72%)	1140 (28%)
Maize	5777	3668 (63%)	2109 (37%)
Cowpea	1370	918 (67%)	452 (33%)
Groundnut	6401	4530 (71%)	1871 (29%)

Table 3: Weighted summary statistics of covariates by sample type for 2017 campaign.

		Mean	Sd	Min	Median	Max	N
Self-reported Yields (kg/ha)							
	X_{-cc}	864.85	1072.22	0.00	552.19	7111.11	3592
Millet	X_{cc}	830.30	1001.05	0.00	550.00	7142.86	1607
	$X_{-cc} - X_{cc}$	34.549					
	X_{-cc}	654.49	960.22	0.00	394.76	8235.29	2436
Sorghum	X_{cc}	719.22	1050.64	0.00	430.15	8125.00	1155
	$X_{-cc} - X_{cc}$	−64.738					
	X_{-cc}	2208.12	2926.16	0.00	1291.97	20851.06	2291
Rice	X_{cc}	1434.14	1941.22	0.00	791.65	17500.00	699
	$X_{-cc} - X_{cc}$	773.972***					
	X_{-cc}	1175.39	1594.14	0.00	760.43	20930.23	2506
Maize	X_{cc}	1214.71	1686.57	0.00	769.23	20242.42	1162
	$X_{-cc} - X_{cc}$	−39.322					
	X_{-cc}	428.54	555.76	0.00	238.10	2994.01	576
Cowpea	X_{cc}	472.67	552.33	0.00	312.39	2727.27	342
	$X_{-cc} - X_{cc}$	−44.128					
	X_{-cc}	967.94	962.66	0.00	666.67	5386.00	3230
Groundnut	X_{cc}	975.78	970.35	0.00	686.46	5312.50	1300
	$X_{-cc} - X_{cc}$	−7.843					
Plot area (ha)							
	X_{-cc}	2.43	2.51	0.02	1.59	28.86	3592
Millet	X_{cc}	2.35	2.64	0.01	1.55	30.02	1607
	$X_{-cc} - X_{cc}$	0.079					
	X_{-cc}	2.62	2.48	0.02	1.94	17.83	2436
Sorghum	X_{cc}	2.40	2.31	0.03	1.82	17.51	1155
	$X_{-cc} - X_{cc}$	0.216*					
	X_{-cc}	1.00	1.35	0.01	0.58	35.00	2291
Rice	X_{cc}	0.95	1.05	0.00	0.59	9.86	699
	$X_{-cc} - X_{cc}$	0.044					
	X_{-cc}	2.08	2.12	0.01	1.43	16.16	2506
Maize	X_{cc}	1.88	1.89	0.01	1.29	14.04	1162
	$X_{-cc} - X_{cc}$	0.205***					
	X_{-cc}	1.35	1.82	0.01	0.60	13.90	576
Cowpea	X_{cc}	1.30	2.05	0.01	0.63	14.28	342
	$X_{-cc} - X_{cc}$	0.049					
	X_{-cc}	0.97	1.59	0.01	0.46	15.99	3230
Groundnut	X_{cc}	0.90	1.36	0.02	0.46	13.35	1300
	$X_{-cc} - X_{cc}$	0.071					

Notes: Number of observations may differ by variable (missingness) depending on the level (plot or household). †denotes a dummy variable. Difference is assessed with t-statistics based on design-adjusted standard errors corrected for clustering at the enumeration area. Statistical significance levels: *p<0.1; **p<0.05; ***p<0.01.

Table 4: Weighted summary statistics of covariates by sample type for 2018 campaign.

		Mean	Sd	Min	Median	Max	N
Self-reported Yields (kg/ha)							
	X_{-cc}	791.81	639.81	0.00	702.76	7052.24	1096
Millet	X_{cc}	829.18	651.81	0.00	695.98	6002.40	519
	$X_{-cc} - X_{cc}$	−37.368					
	X_{-cc}	933.42	950.33	0.00	725.57	8100.00	1435
Sorghum	X_{cc}	888.75	896.81	0.00	700.98	8130.08	693
	$X_{-cc} - X_{cc}$	44.664					
	X_{-cc}	3205.33	3475.40	0.00	2100.00	21000.00	899
Rice	X_{cc}	2963.22	3459.49	0.00	2035.28	19983.21	241
	$X_{-cc} - X_{cc}$	242.103					
	X_{-cc}	1845.10	1665.25	0.00	1443.02	18000.00	1465
Maize	X_{cc}	1890.11	1792.05	0.00	1428.57	18646.90	644
	$X_{-cc} - X_{cc}$	−45.007					
	X_{-cc}	535.53	454.56	0.00	427.38	2711.86	265
Cowpea	X_{cc}	553.83	484.77	0.00	434.66	2500.00	187
	$X_{-cc} - X_{cc}$	−18.296					
	X_{-cc}	1053.99	918.97	0.00	769.23	5378.57	1317
Groundnut	X_{cc}	1020.61	806.37	0.00	769.23	4911.59	554
	$X_{-cc} - X_{cc}$	33.379					
Plot area (ha)							
	X_{-cc}	2.52	2.85	0.05	1.84	44.09	1096
Millet	X_{cc}	2.16	2.18	0.07	1.50	21.01	519
	$X_{-cc} - X_{cc}$	0.366***					
	X_{-cc}	1.79	2.26	0.02	1.18	52.05	1435
Sorghum	X_{cc}	1.75	1.65	0.03	1.31	13.61	693
	$X_{-cc} - X_{cc}$	0.038					
	X_{-cc}	1.21	1.59	0.01	0.76	30.00	899
Rice	X_{cc}	1.51	1.91	0.01	0.63	10.00	241
	$X_{-cc} - X_{cc}$	−0.299					
	X_{-cc}	1.53	1.60	0.01	1.07	22.97	1465
Maize	X_{cc}	1.48	1.59	0.02	1.00	17.80	644
	$X_{-cc} - X_{cc}$	0.051					
	X_{-cc}	0.74	1.08	0.00	0.48	8.01	265
Cowpea	X_{cc}	0.84	1.72	0.02	0.40	14.63	187
	$X_{-cc} - X_{cc}$	−0.097					
	X_{-cc}	0.72	0.90	0.00	0.49	14.18	1317
Groundnut	X_{cc}	0.77	1.08	0.03	0.48	9.48	554
	$X_{-cc} - X_{cc}$	−0.046					

Notes: Number of observations may differ by variable (missingness) depending on the level (plot or household). †denotes a dummy variable. Difference is assessed with t-statistics based on design-adjusted standard errors corrected for clustering at the enumeration area. Statistical significance levels: *p<0.1; **p<0.05; ***p<0.01.

Table 5: Random forest importance for the top 10 most important variables for ML predictions (using levels) for all crops using 3/4 sample of the 2017 wave.

Millet	**Sorghum**
Seeding date x wettest quarter avg	Seeding date x wettest quarter avg
Avg total rainfall in wettest quarter	Avg total rainfall in wettest quarter
Avg annual total rainfall	Avg annual total rainfall
HH Distance to admin center of District of Residence	SR Yield
HH Distance to Nearest Population Center	Annual Precipitation
Annual Precipitation	HH Distance to admin center of District of Residence
Household weights	Seeding date x wettest quarter
Seeding date x rain fall planting season	Seeding date wettest quarter average start
HH Distance to Nearest Border Crossing	Slope
Long-term max dekadal NDVI value in primary growing season	HH Distance to Nearest Population Center

Rice	**Maize**
Household weights	Elevation
Slope	Seeding date x rain fall planting season
SR Yield	Rain fall during planting season
HH Distance to Nearest Border Crossing	Long-term max dekadal NDVI value in primary growing season
Long-term max dekadal NDVI value in primary growing season	Annual Mean Temperature
Excess salts	Long-term avg NDVI value in primary growing season
Long-term avg NDVI value in primary growing season	Avg annual total rainfall
Seeding date x rain fall planting season	HH Distance to admin center of District of Residence
Elevation	Seeding date x wettest quarter avg
Annual Precipitation	SR Yield

Cowpea	**Groundnut**
SR Yield	Annual Precipitation
HH Distance to Nearest Population Center	Long-term avg NDVI value in primary growing season
HH Distance to admin center of District of Residence	SR Yield
Slope	Seeding date x rain fall planting season
HH Distance to Nearest Border Crossing	Avg annual total rainfall
Household weights	HH Distance to admin center of District of Residence
Crop plot percentage	Rain fall during planting season
of males age 41-70	HH Distance to Nearest Population Center
Long-term avg NDVI value in primary growing season	Seeding date x wettest quarter avg
of males age 0-5	Household weights

Table 6: Difference in terms of R^2 of the levels CC regressed on the ML predictions for different combinations of covariates

	Levels						
	All crops	Millet	Sorghum	Rice	Maize	Cowpea	Groundnut
SR in W, GPS not in W	0.052***	0.047***	0.095***	0.033	0.012	0.072***	0.055***
	(0.014)	(0.014)	(0.018)	(0.032)	(0.016)	(0.024)	(0.012)
SR not in W, GPS in W	0.184***	0.221***	0.234***	0.161***	0.106***	0.103***	0.280***
	(0.014)	(0.014)	(0.018)	(0.032)	(0.016)	(0.024)	(0.012)
SR,GPS in W	0.206***	0.223***	0.248***	0.196***	0.112***	0.167***	0.290***
	(0.014)	(0.014)	(0.018)	(0.032)	(0.016)	(0.024)	(0.012)
Constant	0.277***	0.277***	0.132***	0.342***	0.427***	0.177***	0.309***
	(0.010)	(0.010)	(0.013)	(0.023)	(0.011)	(0.017)	(0.008)
Obs	600	100	100	100	100	100	100
Adj. R^2	0.350	0.797	0.716	0.334	0.448	0.330	0.905

Notes: The reference group is covariate without SR and GPS extracted. The constant shows the mean average R^2 for the reference group Significance levels: *p<0.1; **p<0.05; ***p<0.01

Table 7: CC mean yields and differences with ML and SR yields in the test set by crops and by survey year.

	Millet	Sorghum	Rice	Maize	Cowpea	Groundnut
	2017					
CC	936.7	1049.0	2576.1	1685.0	332.5	1841.2
ML - CC	-7.1	22.6	-111.6	-0.4	-7.7	185.2
SR - CC	-101.0	-284.6***	-1164.8***	-429.6***	203.9*	-909.0***
N	1072	773	465	775	228	869
	2018					
CC	931.0	1182.1	3353.0	1914.1	577.4	1997.2
ML - CC	6.1	-61.7*	829.2**	-47.6	-121.0**	178.0*
SR - CC	-97.2	-296.6***	-427.4	6.0	-5.8	-928.0***
N	346	458	160	428	123	371

Notes: The training set represent 1/3 of the total sample size. The mean CC yield and differences with ML and SR yields are computed using survey design weights. Statistical significance are assessed using standard errors clustered at the enumeration area level. In the case of the ML, the standard errors used for statistical significance tests account for the between imputations variance computed with 100 simulations. Significance levels: *p<0.1; **p<0.05; ***p<0.01

Table 8: CC mean yields and differences with ML and SR yields by crops, in the overall sample and by region for 2017.

	Millet	Sorghum	Rice	Maize	Cowpea	Groundnut
			National			
CC	934.3	1045.5	2542.8	1684.7	320.5	1866.3
ML - CC	2.4	-6.4	95.5	-28.8	-6.8	108.3
SR - CC	-104.0*	-326.2***	-1108.7***	-470.0***	152.2	-890.5***
N	1607	1155	699	1162	342	1300
			Kayes			
CC	548.0	754.7	4114.4	774.9	216.3	2090.8
ML - CC	198.8***	155.2***	22.4	336.5***	62.7*	47.1
SR - CC	619.1	-196.2**	-3098.6*	128.1	113.8	-753.3***
N	64	327	26	284	47	594
			Koulikoro			
CC	1024.3	1057.1	1523.0	1650.5	288.6	1398.4
ML - CC	-44.6*	-4.9	267.4*	24.2	-5.8	510.0**
SR - CC	-400.2***	-483.5***	-330.5	-712.5***	342.3*	-696.8**
N	276	305	110	247	58	235
			Sikasso			
CC	832.8	1079.5	2299.3	2141.8	330.6	1518.6
ML - CC	-97.8***	33.4	98.8	-222.3***	-7.6	255.4***
SR - CC	-121.2	-184.1	-1473.5***	-548.2***	278.3***	-640.7***
N	166	222	156	534	79	239
			Segou			
CC	1132.3	1438.9	5786.8	2196.4	433.6	2566.8
ML - CC	-103.3***	-212.5***	-526.6	-256.6**	-67.8***	-559.9***
SR - CC	-142.5	-527.4***	-4268.2**	-1144.7***	14.0	-1837.8***
N	485	243	40	93	68	168
			Mopti			
CC	862.0	839.3	1760.5	768.3	315.9	1711.1
ML - CC	71.2***	-25.5	585.1*	-0.0	2.9	108.7
SR - CC	-54.2	-37.0	-376.4	-108.7	-11.4	-1142.2***
N	594	54	154	4	88	63

Notes: The training set represent 1/3 of the total sample size. The mean CC yield and differences with ML and SR yields are computed using survey design weights. Statistical significance are assessed using standard errors clustered at the enumeration area level. In the case of the ML, the standard errors used for statistical significance tests account for the between imputations variance computed with 10 simulations. Significance levels: *p<0.1; **p<0.05; ***p<0.01

Table 9: Survey to survey imputation results: CC mean yields and differences with ML and SR yields by crops, in the overall sample and by region for 2018..

	Millet	Sorghum	Rice	Maize	Cowpea	Groundnut
			National			
CC	928.2	1167.1	3661.7	1898.4	534.6	2033.8
ML - CC	-54.7	-151.7***	-297.3	-229.1***	-122.4**	103.3
SR - CC	-99.0	-278.4***	-698.4	-8.2	19.3	-1013.2***
N	519	693	241	644	187	554
			Kayes			
CC	949.1	1129.8	8293.7	1201.2	580.8	2603.2
ML - CC	-282.0*	-282.4***	-2823.2	-245.7***	-218.5**	-436.6***
SR - CC	-18.6	-301.6***	-5520.2**	320.0	175.4	-1384.6***
N	40	257	9	206	31	179
			Koulikoro			
CC	1035.8	1214.7	2360.7	1931.8	746.4	1717.9
ML - CC	-110.9	-154.6**	905.8	-215.2	-249.6**	204.4
SR - CC	-358.5***	-260.0*	-713.7	-42.2	-145.4	-862.3***
N	154	163	37	158	44	138
			Sikasso			
CC	1018.4	1246.2	2881.2	2383.5	473.5	1667.7
ML - CC	-133.8*	-222.0*	-330.9	-212.7**	-110.2	188.0
SR - CC	-16.9	-338.0***	-1152.0*	-95.8	-47.8	-351.1
N	107	109	53	225	41	94
			Segou			
CC	1106.4	1308.4	7329.0	2203.7	381.0	2116.1
ML - CC	-52.8	-7.9	-2665.6**	-353.6**	56.6	493.1***
SR - CC	-285.8***	-506.9***	-4928.6***	-686.6***	145.7*	-1330.5***
N	110	123	23	53	42	122
			Mopti			
CC	639.7	780.8	2923.3	786.7	460.1	1469.8
ML - CC	195.1*	77.4	-235.0	289.2	-90.6	93.9
SR - CC	227.7*	317.8	955.7	459.4	70.8	-525.3*
N	70	21	40	2	19	21

Notes: The training set is the EACI 2017. The mean CC yield and differences with ML and SR yields are computed using survey design weights. Statistical significance are assessed using standard errors clustered at the enumeration area level. In the case of the ML, the standard errors used for statistical significance tests account for the between imputations variance computed with 100 simulations. Significance levels: *p<0.1; **p<0.05; ***p<0.01

Figures

Figure 1: SR over area quintile for 2017 (left) and 2018 (right) wave for Millet (top), Sorghum (middle), rice (bottom)

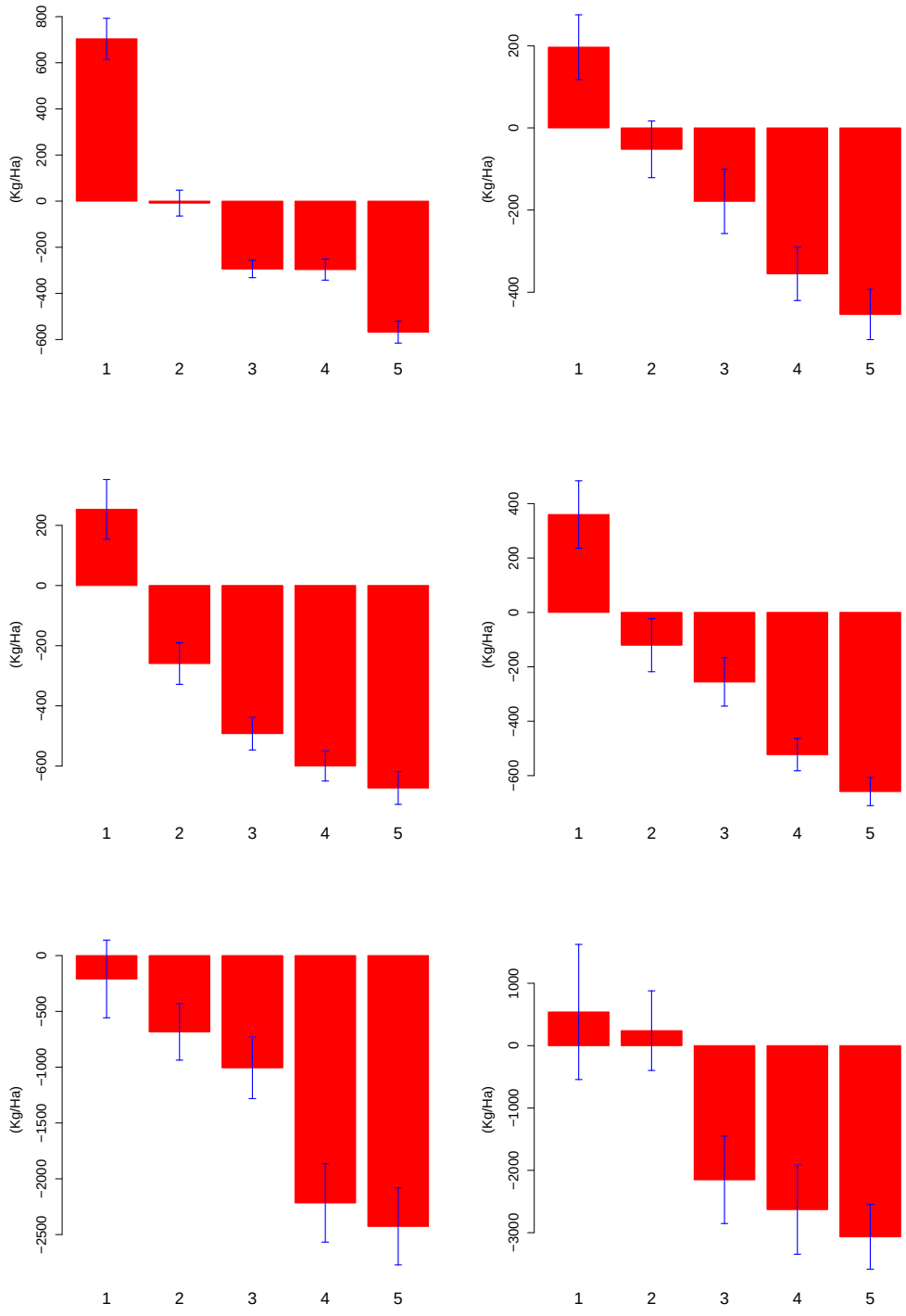


Figure 2: SR over area quintile for 2017 (left) and 2018 (right) wave for Maize (top), Cowpea (middle), Groundnut (bottom)

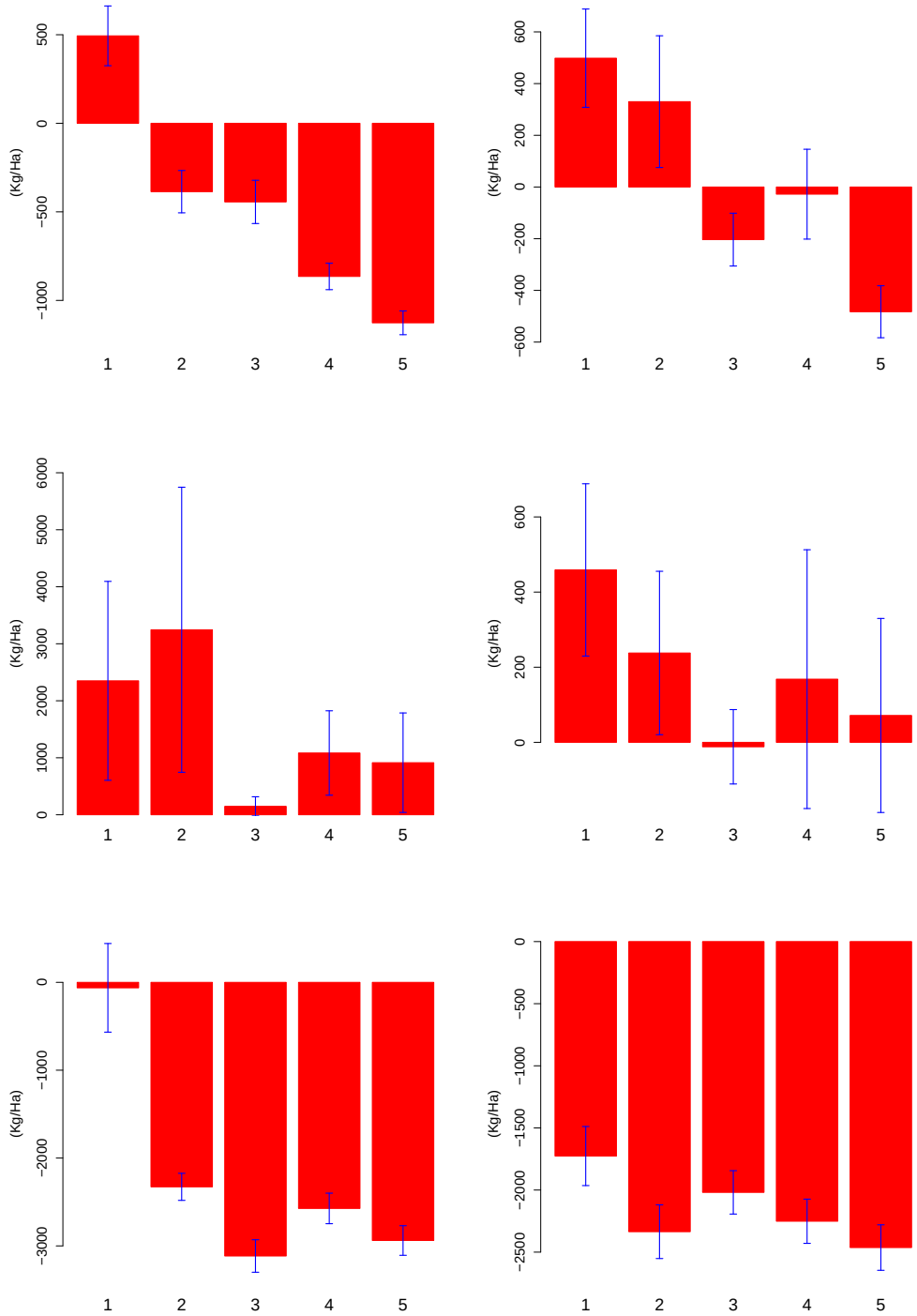


Figure 3: Random forest importance for the top 15 most important variables for ML prediction of levels of Rice. *Notes:* 3/4 of the crop-cut sample is used. 5 validation cross-fold splits are used.

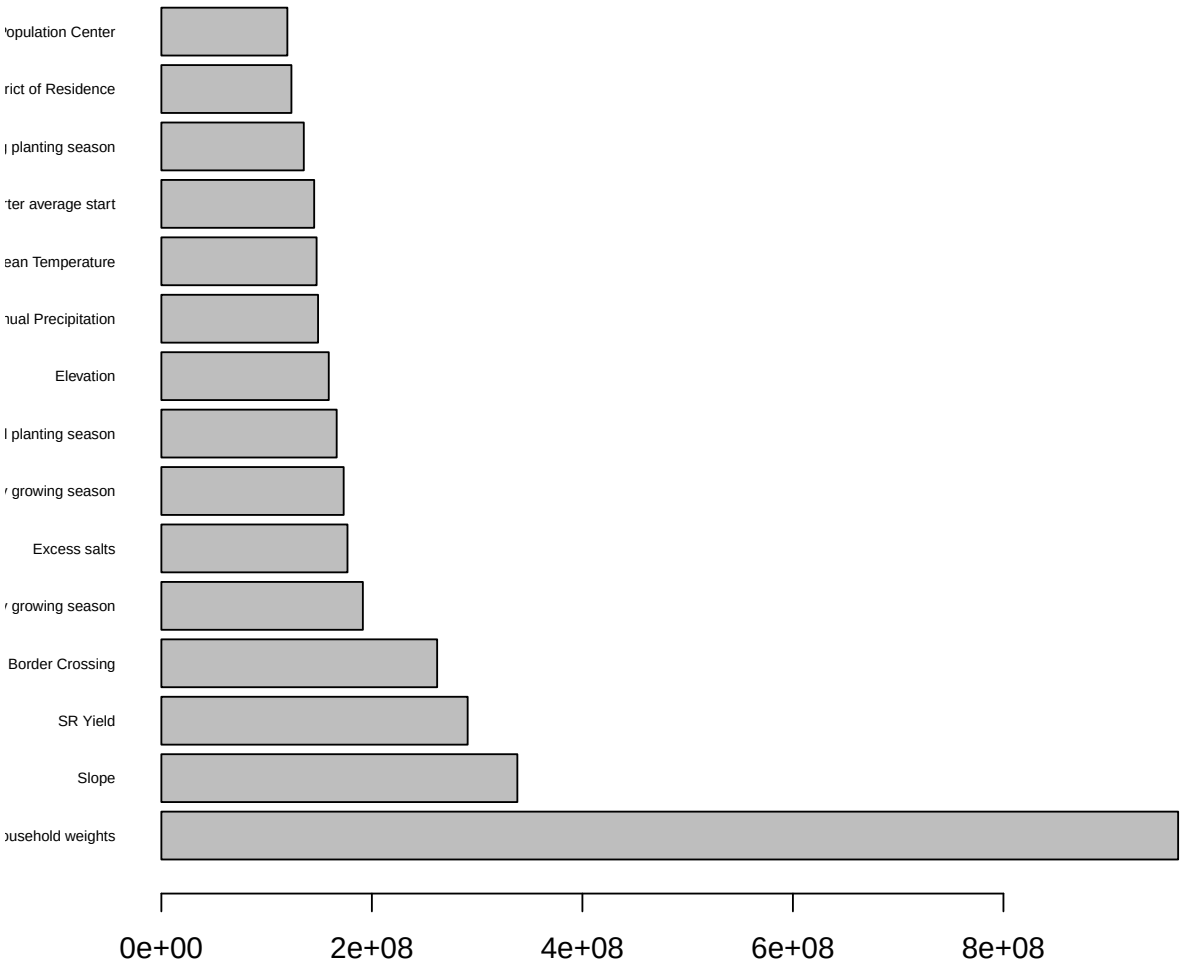


Figure 4: Random forest importance for the top 15 most important variables for ML prediction of logs of Rice. *Notes:* 3/4 of the crop-cut sample is used. 5 validation cross-fold splits are used.

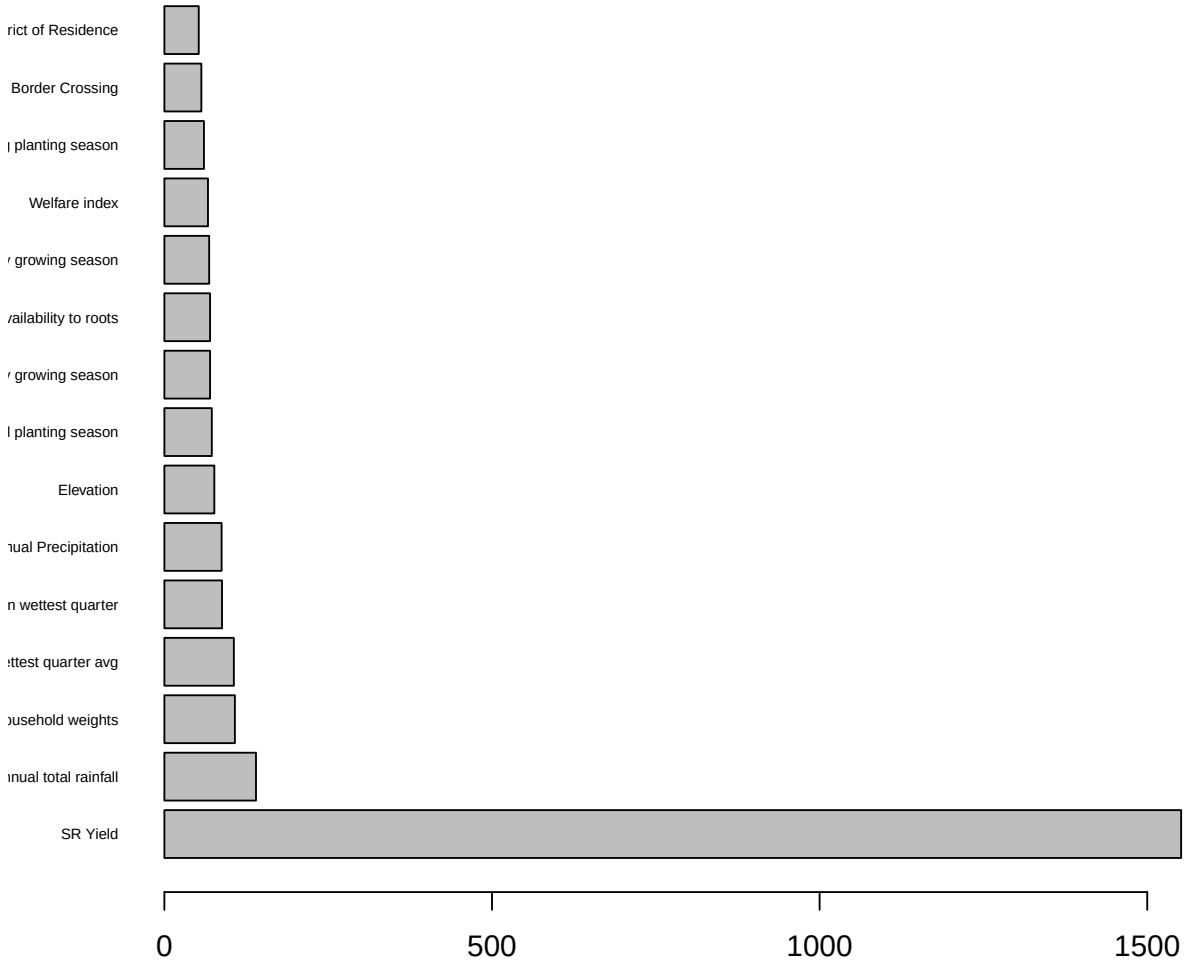


Figure 5: R^2 of CC regressed again the ML prediction for each crop. *Notes:* The model predicts yields levels. The training sample is 3/4 of the 2017 sample.

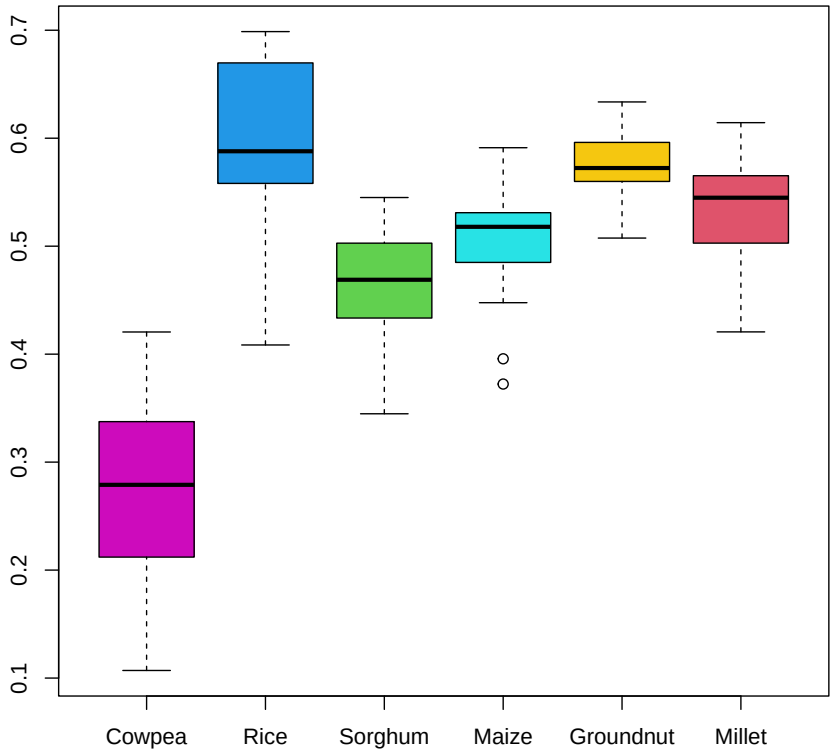


Figure 6: R^2 of CC regressed again the ML prediction for each crop and overall. *Notes:* The black curve shows the overall average. The model includes both SR and GPS extracted vars. By increasing order of sample size: Purple is for cowpea, blue for rice, green for sorghum, sky blue for maize, yellow for groundnut and red for millet.

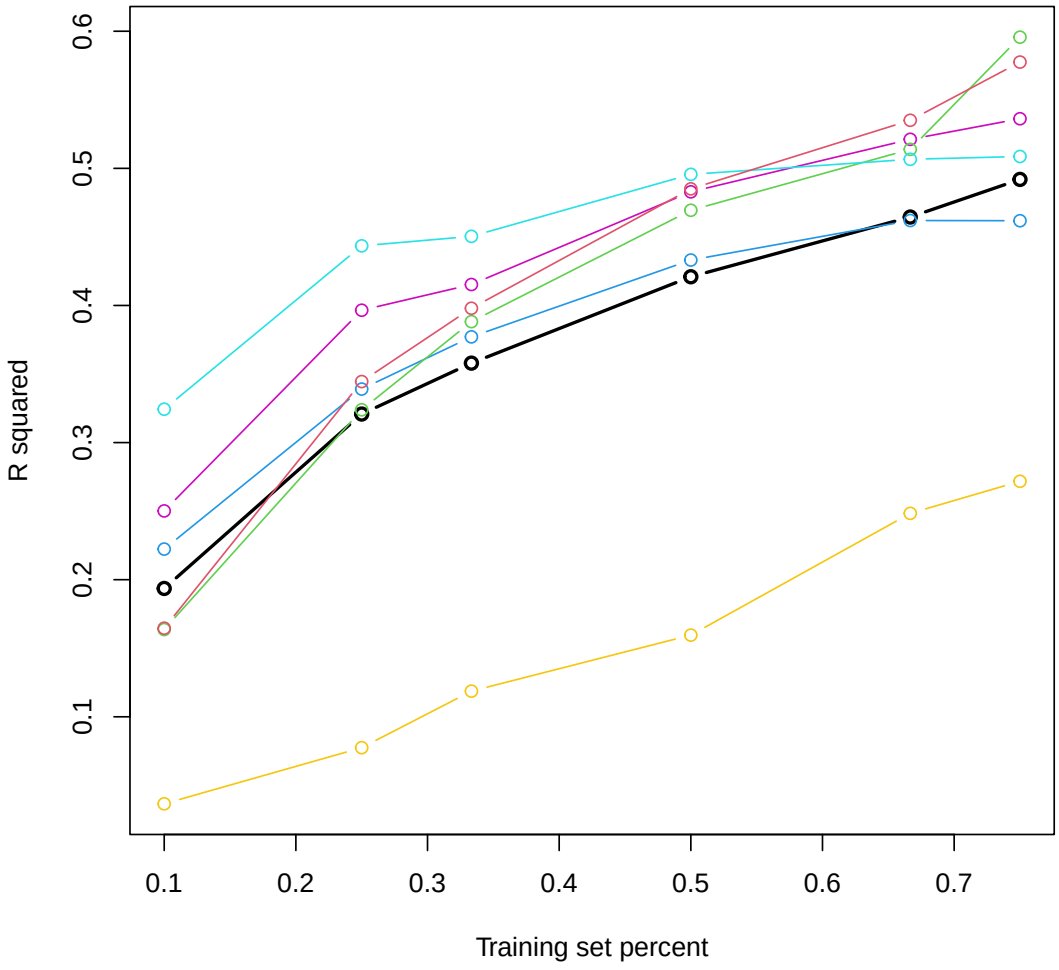


Figure 7: crop-cut (red), machine-learning (green) and, self-reported (blue) yield means at the national and regional levels in 2017.

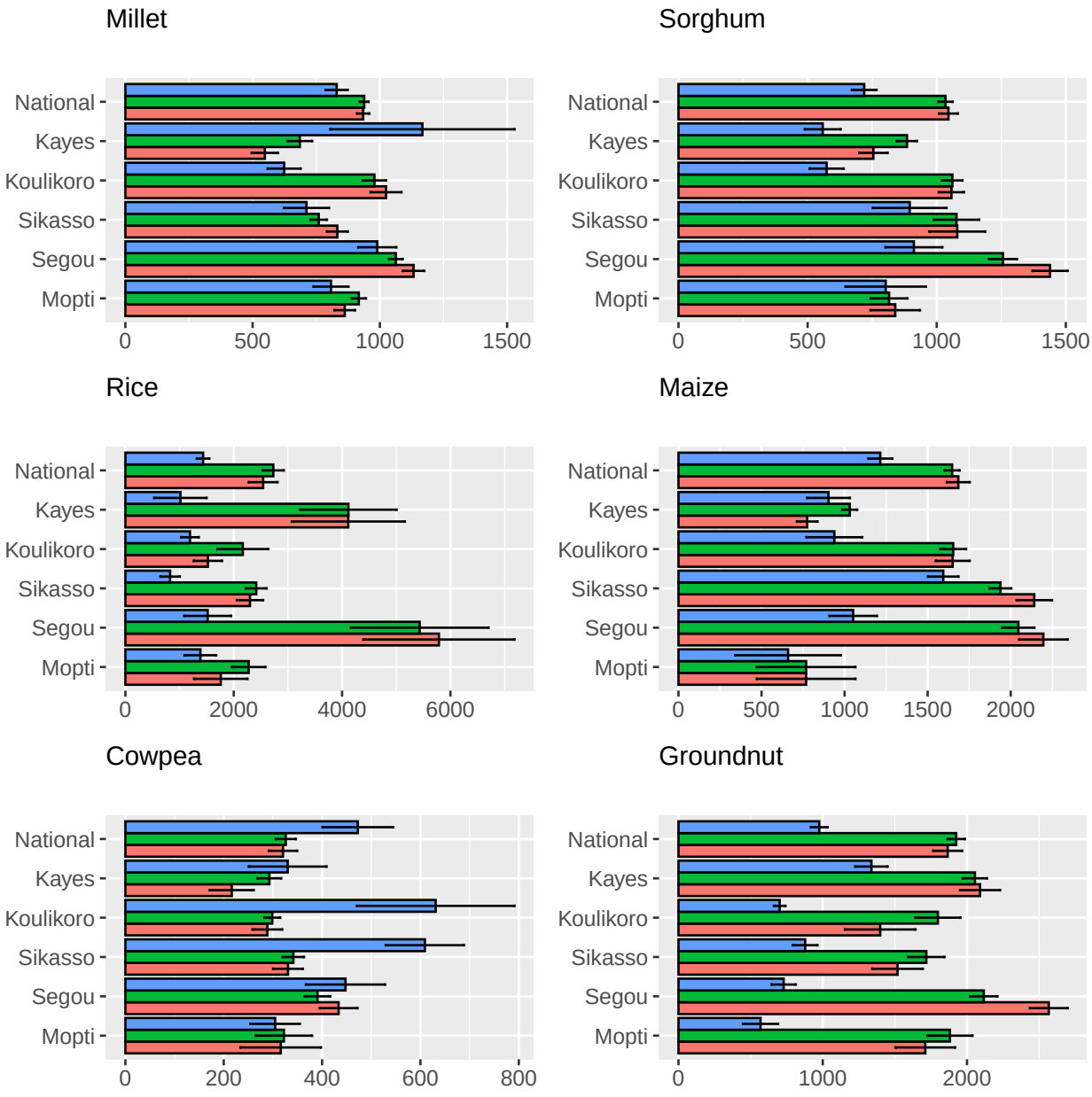


Figure 8: Distribution of the means of the CC (red) and ML (blue) in the test sample for 100 simulations

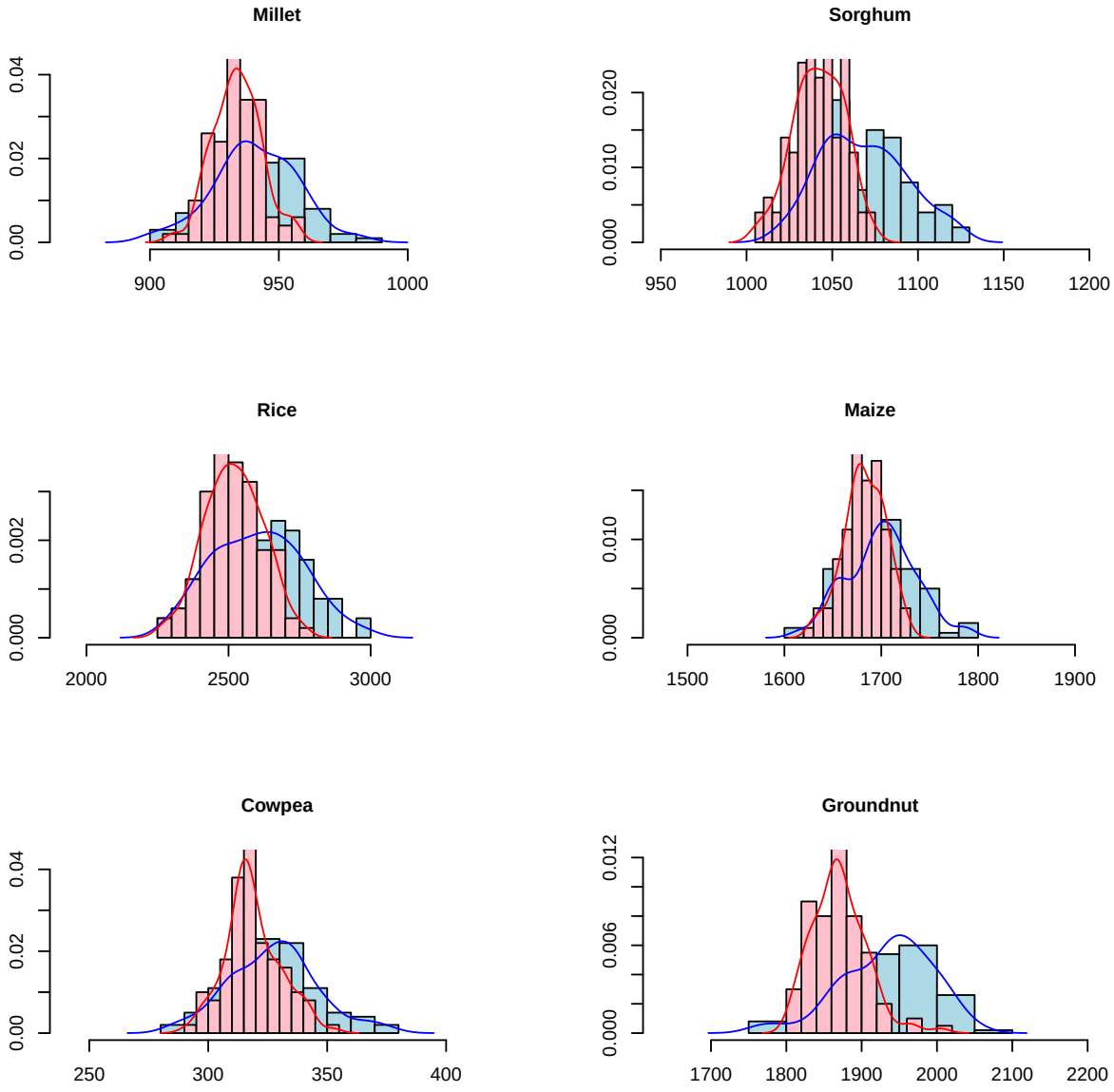


Figure 9: Survey to survey imputation results: crop-cut (red), machine-learning (green) and, self-reported (blue) yield means at the national and regional levels in 2018.

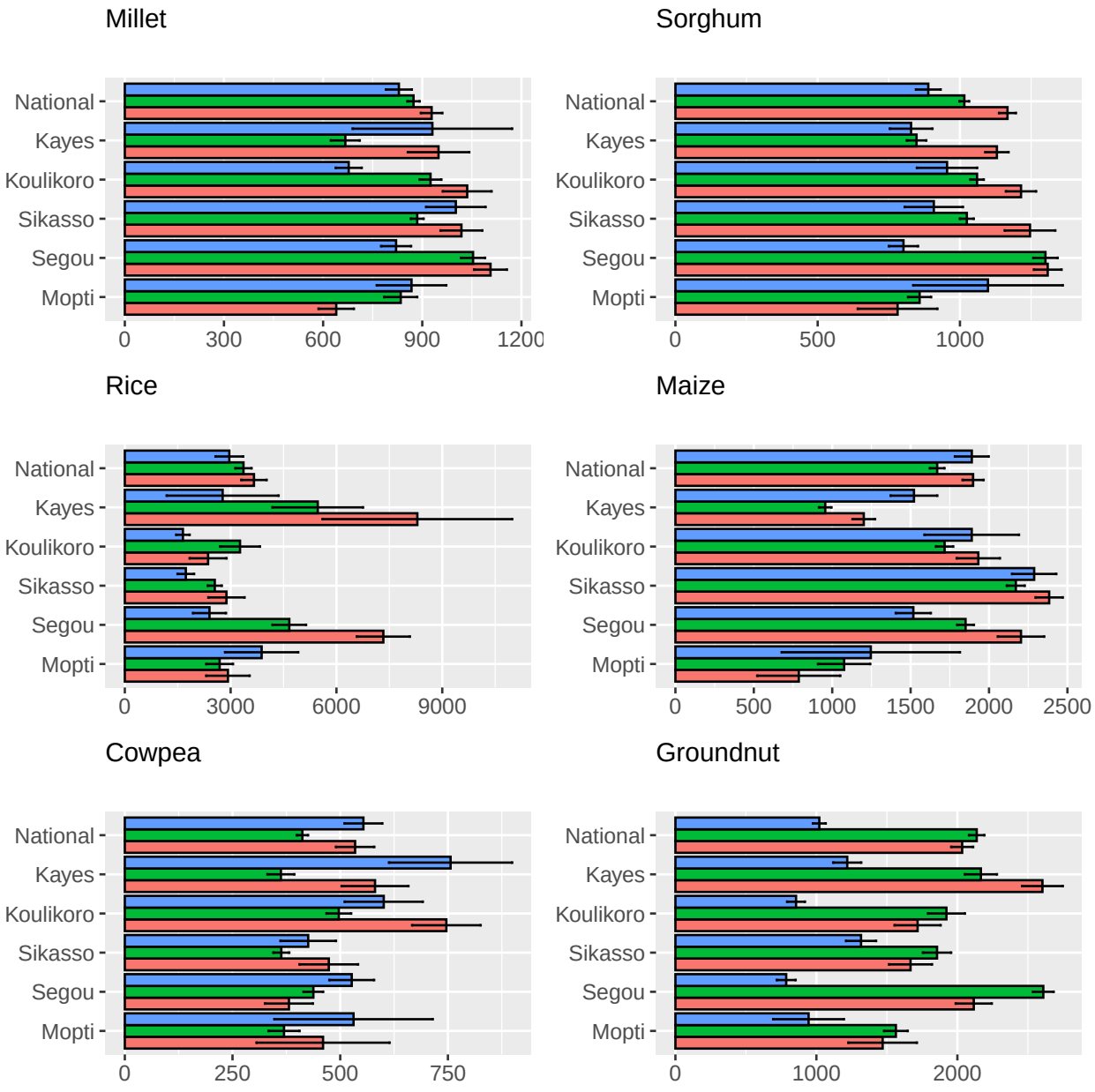
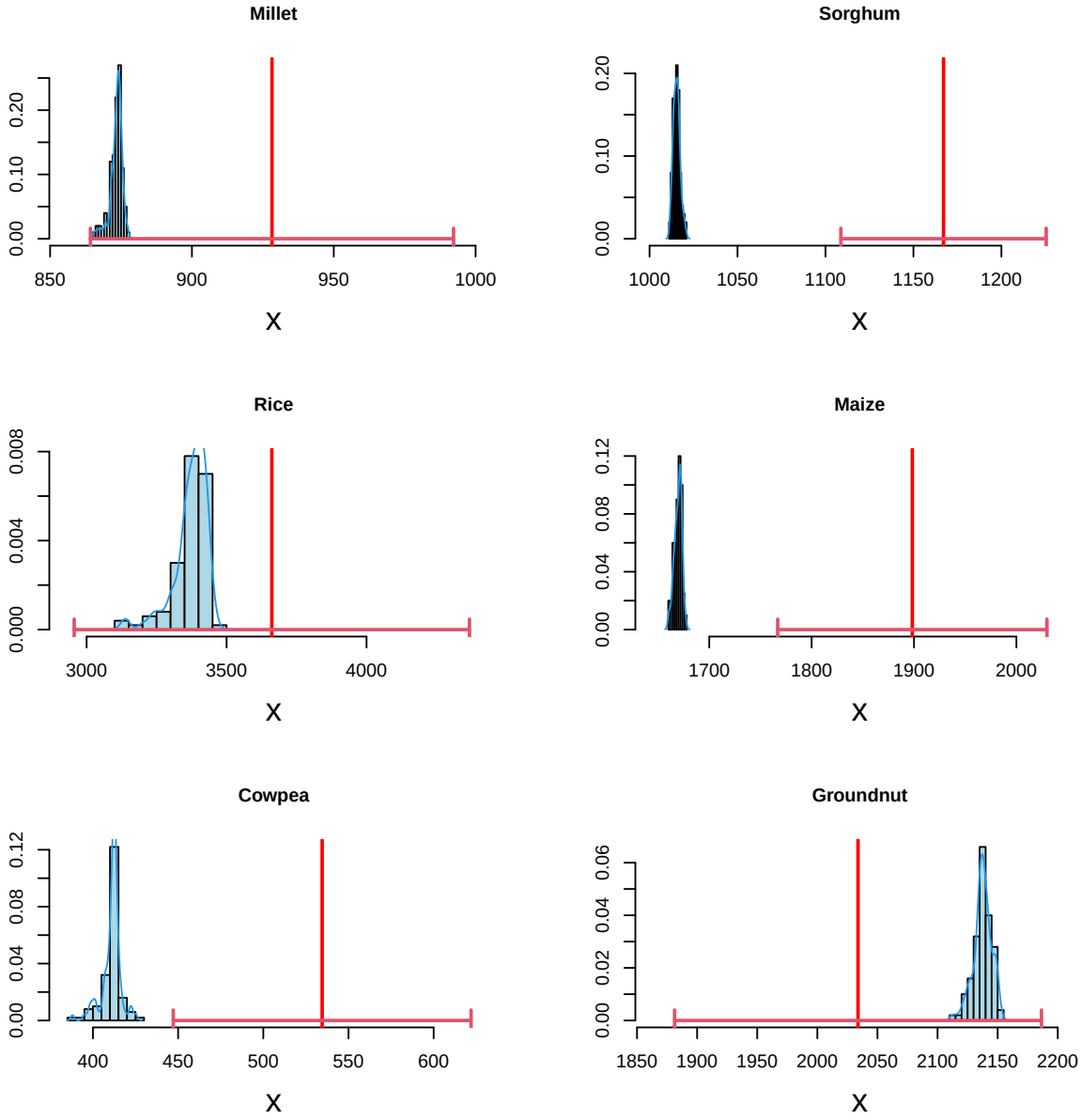


Figure 10: Distribution of the means of the ML (blue) in the test sample (2018) for 100 simulations. The red vertical line shows the mean CC and the red horizontal line shows the 95% confidence interval in the 2018 sample



Appendix A Machine learning modeling

We use an ensemble of methods (super learner, LASSO, elastic-net, random forest, ridge, and gradient boosted random forest) implemented using the R package SuperLearner. Ensemble methods are standard and popular machine learning techniques to generate predictions. Figure A.1 illustrates how the algorithm works. We looked more closely at which algorithms were getting the most best predictions. Figure A.2 shows the risk (error) and the standard error (it is the average risk over the 5 cross validation splits) for the run of the 2017 rice logs. Table A.1 gives the summary for a run for all crops. In most of the cases, for the levels, the random forest have the best predictions but for the logs, combining the predictions predictions is the best way to go. Overall random forest seems to be the most efficient algorithm which is a good information if computational efficiency is a concern for this exercise.

Tables

Table A.1: Cross validation risk for the different allgorithms for all crops using 3/4 sample of the 2017 wave.

Algorithm	Millet	Sorghum	Rice	Maize	Cowpea	Groundnut
Levels						
Super Learner	231139	304845	14520027	931240	**2192122**	108670
Lasso	256340	403974	14725440	968075	2972522	119242
Elastic Net	255117	398527	**14278840**	959681	3011342	115734
Ridge	255245	394622	15281217	955227	2784184	114111
Random Forest (RF)	**229718**	**302959**	14625590	**898914**	2195106	**104182**
Boosted RF	278353	364371	16859903	1022238	2593876	120141
Logs						
Super Learner	**1.034**	2.730	**4.928**	**2.445**	**4.809**	3.254
Lasso	1.117	3.260	5.143	2.533	4.849	3.368
Elastic Net	1.115	3.258	5.104	2.486	4.816	3.368
Ridge	1.124	3.378	5.800	2.498	5.002	3.656
Random Forest (RF)	1.133	**2.632**	5.118	2.513	5.059	**3.154**
Boosted RF	1.484	3.091	7.140	4.101	6.015	3.707

Figures

Figure A.1: Illustration of the SuperLearner algorithm

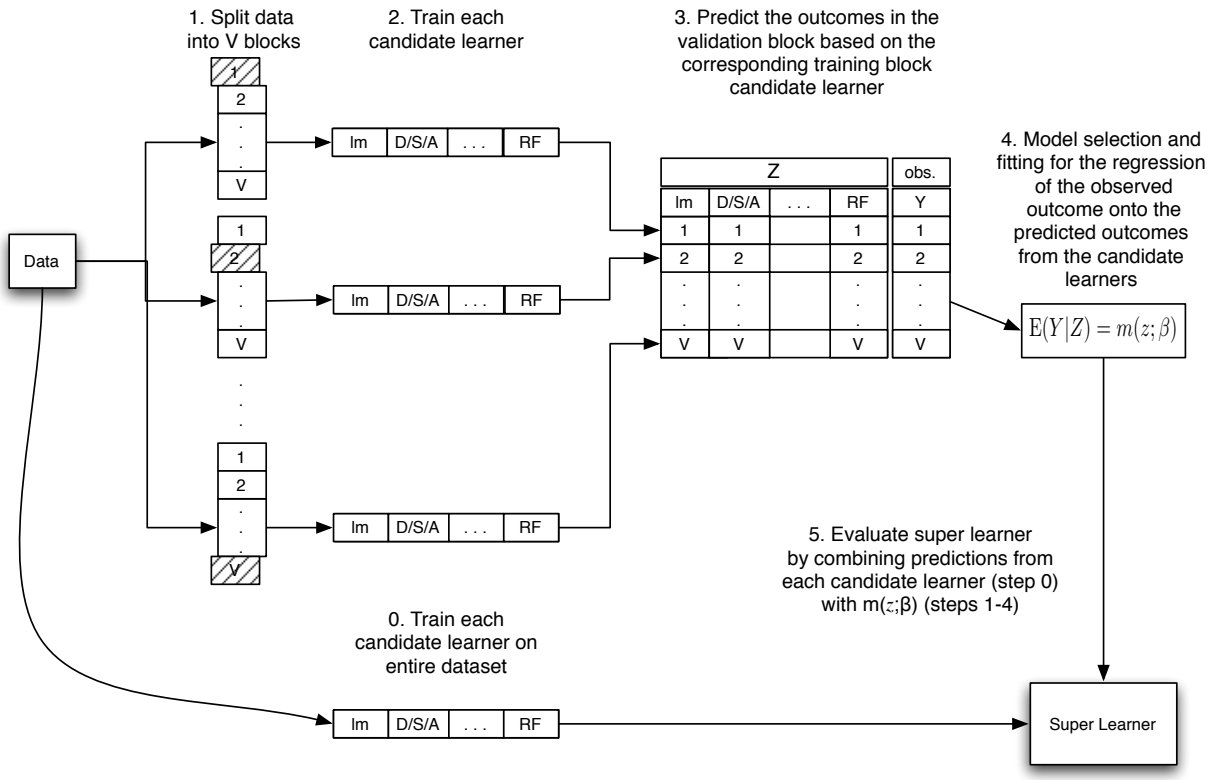
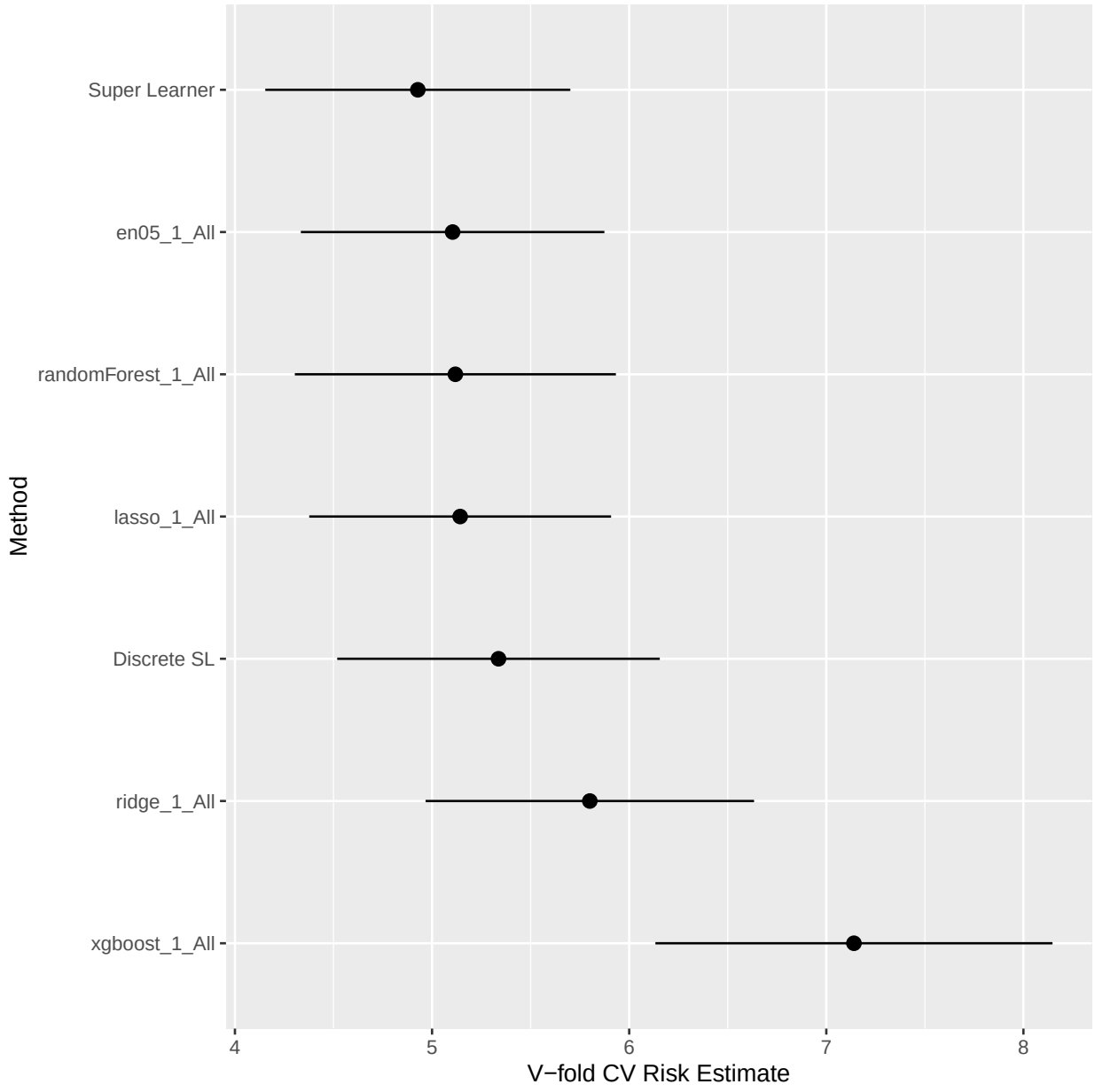


Figure A.2: Cross validation risk (in terms of rmse) for each algorithms and the ensemble. *Notes:* 3/4 of the crop-cut sample is used. 5 validation cross-fold splits are used. The illustration is for logs of rice



Appendix B Additional Tables and Figures

Tables

Table B.1: Weighted summary statistics of yields by methods for 2017 campaign.

		Mean	Sd	Min	Median	Max	Observations
	Millet						
CC	All	934	453	0	935	2618	1607
	Pure Stand	946	470	0	935	2618	1409
	Intercropped	860	316	0	841	2471	198
SR	All	854	1050	0	551	7143	5199
	Pure Stand	829	1028	0	538	7143	4544
	Intercropped	1011	1170	0	645	7111	655
	Sorghum						
CC	All	1045	629	0	1003	3960	1155
	Pure Stand	1057	636	0	1003	3960	1037
	Intercropped	965	571	0	1003	3454	118
SR	All	675	991	0	413	8235	3591
	Pure Stand	664	989	0	395	8235	3251
	Intercropped	768	1001	0	466	7407	340
	Rice						
CC	Pure Stand	2543	3384	0	1324	20175	699
SR	Pure Stand	2036	2757	0	1143	20851	2990
	Maize						
CC	All	1685	992	0	1583	4916	1162
	Pure Stand	1693	1002	0	1566	4916	1133
	Intercropped	1440	612	67	1633	2366	29
SR	All	1188	1625	0	766	20930	3668
	Pure Stand	1160	1587	0	747	20930	3553
	Intercropped	1981	2331	0	1250	20000	115
	Cowpea						
CC	All	320	288	0	285	2081	342
	Pure Stand	387	298	0	376	2081	255
	Intercropped	177	200	0	142	1189	87
SR	All	446	555	0	256	2994	918
	Pure Stand	430	542	0	246	2994	656
	Intercropped	481	581	0	280	2950	262
	Groundnut						
CC	All	1866	1414	0	1650	5374	1300
	Pure Stand	1887	1413	0	1753	5374	1277
	Intercropped	1244	1298	0	707	4242	23
SR	All	970	965	0	667	5386	4530
	Pure Stand	971	965	0	669	5386	4447
	Intercropped	956	951	0	621	4132	83

Table B.2: Weighted summary statistics of yields by methods for 2018 campaign.

		Mean	Sd	Min	Median	Max	Observations
Millet							
CC	All	928	486	0	902	3880	519
	Pure Stand	929	493	0	904	3880	481
	Intercropped	919	415	0	880	2707	38
SR	All	804	644	0	700	7052	1615
	Pure Stand	803	650	0	709	7052	1488
	Intercropped	808	584	1	677	5462	127
Sorghum							
CC	All	1167	522	0	1075	3024	693
	Pure Stand	1181	528	0	1091	3024	638
	Intercropped	1022	424	0	1046	2150	55
SR	All	919	934	0	721	8130	2128
	Pure Stand	929	951	0	742	8130	1966
	Intercropped	800	691	44	671	7000	162
Rice							
CC	Pure Stand	3662	3531	0	2207	19564	241
SR	Pure Stand	3149	3473	0	2064	21000	1140
Maize							
CC	All	1898	942	0	1804	4766	644
	Pure Stand	1887	950	0	1747	4766	627
	Intercropped	2201	616	267	2077	3999	17
SR	All	1859	1705	0	1442	18647	2109
	Pure Stand	1848	1704	0	1439	18647	2068
	Intercropped	2337	1662	353	1759	8400	41
Cowpea							
CC	All	535	369	0	510	1491	187
	Pure Stand	587	357	0	595	1491	166
	Intercropped	185	233	0	135	892	21
SR	All	544	468	0	432	2712	452
	Pure Stand	564	464	0	449	2712	411
	Intercropped	384	466	0	182	2000	41
Groundnut							
CC	All	2034	1224	0	1943	5303	554
	Pure Stand	1999	1219	0	1904	5303	538
	Intercropped	2676	1133	943	2166	5206	16
SR	All	1044	887	0	769	5379	1871
	Pure Stand	1038	889	0	769	5379	1842
	Intercropped	1257	798	0	1207	4912	29

Table B.3: Random forest importance for the top 10 most important variables for ML predictions (using logs) for all crops using 3/4 sample of the 2017 wave.

Millet	**Sorghum**
SR Yield	SR Yield
HH Distance to admin center of District of Residence	Mean Temperature of Wettest Quarter
Seeding date x wettest quarter avg	Seeding date x wettest quarter avg
Seeding date x rain fall planting season	Avg total rainfall in wettest quarter
Seeding date x wettest quarter	Avg annual total rainfall
Avg total rainfall in wettest quarter	Seeding date
Rain fall during planting season	Seeding date x rain fall planting season
Rainfall during wettest quarter	Rain fall during planting season
HH Distance to Nearest Population Center	HH Distance to admin center of District of Residence
Avg annual total rainfall	Welfare index
Rice	**Maize**
SR Yield	SR Yield
Avg annual total rainfall	Elevation
Household weights	Rain fall during planting season
Seeding date x wettest quarter avg	Seeding date x rain fall planting season
Avg total rainfall in wettest quarter	Mean Temperature of Wettest Quarter
Annual Precipitation	Annual Mean Temperature
Elevation	Avg annual total rainfall
Seeding date x rain fall planting season	Long-term max dekadal NDVI value in primary growing season
Long-term max dekadal NDVI value in primary growing season	HH Distance to Nearest Border Crossing
Oxygen availability to roots	Long-term avg NDVI value in primary growing season
Cowpea	**Groundnut**
SR Yield	SR Yield
Seeding date x wettest quarter avg	HH Distance to Nearest Population Center
Avg total rainfall in wettest quarter	Seeding date x wettest quarter avg
Avg annual total rainfall	Long-term avg NDVI value in primary growing season
Mean Temperature of Wettest Quarter	Long-term max dekadal NDVI value in primary growing season
HH Distance to Nearest Population Center	Seeding date x rain fall planting season
Crop plot percentage	HH Distance to admin center of District of Residence
HH Distance to admin center of District of Residence	Avg total rainfall in wettest quarter
Seeding date	Avg annual total rainfall
Long-term avg NDVI value in primary growing season	Rain fall during planting season

Table B.4: Difference in terms of R2 of the logs CC regressed on the ML predictions for different combinations of covariates

	Logs						
	All crops	Millet	Sorghum	Rice	Maize	Cowpea	Groundnut
SR in W, GPS not in W	0.205***	0.191***	0.160***	0.213***	0.138***	0.371***	0.160***
	(0.015)	(0.026)	(0.022)	(0.023)	(0.015)	(0.032)	(0.018)
SR not in W, GPS in W	0.142***	0.220***	0.184***	0.073***	0.094***	0.167***	0.113***
	(0.015)	(0.026)	(0.022)	(0.023)	(0.015)	(0.032)	(0.018)
SR,GPS in W	0.275***	0.279***	0.258***	0.237***	0.177***	0.465***	0.233***
	(0.015)	(0.026)	(0.022)	(0.023)	(0.015)	(0.032)	(0.018)
Constant	0.166***	0.136***	0.133***	0.381***	0.215***	0.060***	0.071***
	(0.011)	(0.018)	(0.015)	(0.017)	(0.010)	(0.022)	(0.012)
Obs	600	100	100	100	100	100	100
Adj. R^2	0.370	0.565	0.597	0.580	0.620	0.721	0.647

Notes: The reference group is covariate without SR and GPS extracted. The constant shows the mean average R2 for the reference group Significance levels: *p<0.1; **p<0.05; ***p<0.01

Table B.5: CC mean yields and differences with ML and SR yields by crops, in the overall sample and by region for 2017.

	Millet	Sorghum	Rice	Maize	Cowpea	Groundnut
			National			
CC	934.3	1045.5	2542.8	1684.7	320.5	1866.3
ML - CC	4.3	-12.0	180.4	-36.4	5.9	59.6
SR - CC	-104.0*	-326.2***	-1108.7***	-470.0***	152.2	-890.5***
N	1607	1155	699	1162	342	1300
			Kayes			
CC	548.0	754.7	4114.4	774.9	216.3	2090.8
ML - CC	137.3***	129.7***	-18.7	258.1***	77.1*	-36.6
SR - CC	619.1	-196.2**	-3098.6*	128.1	113.8	-753.3***
N	64	327	26	284	47	594
			Koulikoro			
CC	1024.3	1057.1	1523.0	1650.5	288.6	1398.4
ML - CC	-46.6*	2.5	633.8	4.7	10.0	400.3***
SR - CC	-400.2***	-483.5***	-330.5	-712.5***	342.3*	-696.8**
N	276	305	110	247	58	235
			Sikasso			
CC	832.8	1079.5	2299.3	2141.8	330.6	1518.6
ML - CC	-74.9**	-3.3	104.8	-202.6***	11.0	200.8*
SR - CC	-121.2	-184.1	-1473.5***	-548.2***	278.3***	-640.7***
N	166	222	156	534	79	239
			Segou			
CC	1132.3	1438.9	5786.8	2196.4	433.6	2566.8
ML - CC	-69.4***	-183.5***	-354.0	-151.0*	-42.6**	-449.1***
SR - CC	-142.5	-527.4***	-4268.2**	-1144.7***	14.0	-1837.8***
N	485	243	40	93	68	168
			Mopti			
CC	862.0	839.3	1760.5	768.3	315.9	1711.1
ML - CC	56.1***	-24.4	515.1*	-0.0	7.0	175.5
SR - CC	-54.2	-37.0	-376.4	-108.7	-11.4	-1142.2***
N	594	54	154	4	88	63

Notes: The training set represent 1/2 of the total sample size. The mean CC yield and differences with ML and SR yields are computed using survey design weights. Statistical significance are assessed using standard errors clustered at the enumeration area level. In the case of the ML, the standard errors used for statistical significance tests account for the between imputations variance computed with 5 simulations. Significance levels: *p<0.1; **p<0.05; ***p<0.01

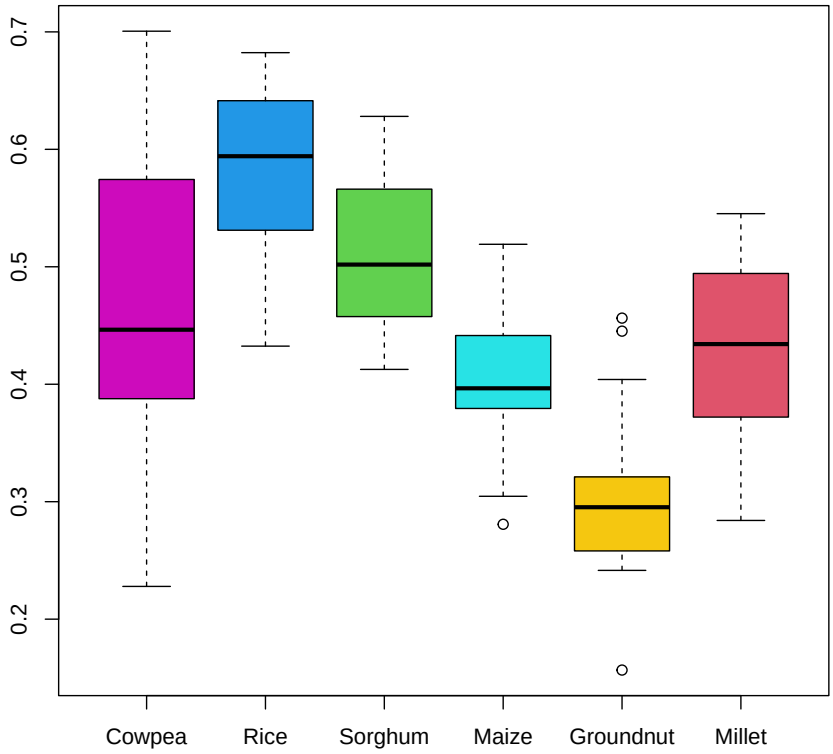
Table B.6: Difference in terms of Root mean square of the levels CC regressed on the ML predictions for different combinations of covariates

	Levels						
	All	Millet	Sorghum	Rice	Maize	Cowpea	Groundnut
SR in W, GPS not in W	−33.674	−4.744	−32.995***	−84.852**	−22.502**	−7.643	−49.308***
	(74.359)	(4.017)	(5.468)	(40.250)	(8.997)	(4.975)	(12.640)
SR not in W, GPS in W	−186.267**	−74.107***	−105.042***	−523.009***	−122.527***	−16.146***	−276.770***
	(74.359)	(4.017)	(5.468)	(40.250)	(8.997)	(4.975)	(12.640)
SR,GPS in W	−196.977***	−73.130***	−109.678***	−560.672***	−124.986***	−22.350***	−291.047***
	(74.359)	(4.017)	(5.468)	(40.250)	(8.997)	(4.975)	(12.640)
Constant	743.780***	306.806***	435.578***	1,920.890***	589.044***	222.296***	988.066***
	(52.579)	(2.841)	(3.866)	(28.461)	(6.362)	(3.518)	(8.938)
Observations	360	60	60	60	60	60	60
Adj. R^2	0.023	0.914	0.909	0.840	0.843	0.255	0.935

Notes: The reference group is covariate without SR and GPS extracted. The constant shows the mean average root mean square for the reference group Significance levels: *p<0.1; **p<0.05; ***p<0.01

Figures

Figure B.1: R^2 of CC regressed again the ML prediction for each crop. *Notes:* The model predicts logged yields. The training sample is 3/4 of the 2017 sample.



Appendix C List of Covariates

- **Plot characteristics (plot level)**
  - Plot area (ha), Plot area (ha), Plot area (ha) - GPS, Plot area (ha) - SR, Fallowed in past 10 years†, Plot owned w. title†, Plot owned w.o title†, Plot rented for free†, Plot rented for fee†, Plot on low-land, Plot on high-land, Plot on shallow land, Plot on low slope, Plot on high slope, Plot in oasis.
- **Plot input (plot level)**
  - Local variety†, Local variety†, Improved variety (1st use)†, Improved variety (2nd year)†, Improved variety (3rd year)†, Improved variety (unknown year)†, Seed quantity (kg/ha), Organic fertilizer use†, Chem. Fert. NPK (kg/ha), Chem. Fert. DAP (kg/ha), Chem. Fert. Urea (kg/ha), Chem. Fert. Other (kg/ha), Insecticides use†, Fongicides use†, Herbicides use†, Other pesticides†, Water: rain, Water: river recession, Water: shallow fields, Water: uncont. irrigation, Water: irrigation cont., Water: irrigation subm., Water: irrigation shallow fields.
- **Crop system (plot level)**
  - Pure stand plot†, Pure stand plot†, Crop proportion rate†.
- **Agriculture labor (plot level)**
  - Post-planting: Hh Days of labor (days/ha), Post-planting: Hh Days of labor (days/ha), Post-planting: Exch Days of labor (days/ha), Post-planting: Paid Days of labor (days/ha), Post-harvest: Hh Days of labor (days/ha), Post-harvest: Exch Days of labor (days/ha), Post-harvest: Paid Days of labor (days/ha), No plowing†, Manual plowing†, Animal plow†, Manual/animal plow†, Machine plow†, Manual/machine plow†, Animal/machine plow†.
- **Household characteristics (household level)**
  - Female 0-5, Female 0-5, Female 6-15, Female 15-40, Female 41-70, Female 71+, Male 0-5, Male 6-15, Male 15-40, Male 41-70, Male 71+, Dependency ratio, Prop. hh members literate†, Prop. hh members w. schooling†, Head female†, Head age, Head can read†, Head can write†, Head single†, Head monogam†, Head polygam†, Head separated†, Head divorced†, Head widowed†, Head ethnicity: Bambara†, Head ethnicity: Sarakole†, Head ethnicity: Kassonke†, Head ethnicity: Senoufo†, Head ethnicity: Dogon†.
- **Plot manager characteristics (plot level)**
  - Manager female†, Manager female†, Manager age, Manager can read†, Manager can write†, Manager ethnicity: Bambara†, Manager ethnicity: Sarakole†, Manager ethnicity: Kassonke†, Manager ethnicity: Senoufo†, Manager ethnicity: Dogon†, Manager ethnicity: Tamacheq†, Manager ethnicity: Bobo†, Manager ethnicity: Bozo†, Manager ethnicity: Other†, Manager ethnicity: Foreign†.
- **Post planting respondent (plot level)**

- PP. resp. female†, PP. resp. female†, PP. resp. age, PP. resp. can read†, PP. resp. can write†, PP. resp. ethnicity: Bambara†, PP. resp. ethnicity: Sarakole†, PP. resp. ethnicity: Kassonke†, PP. resp. ethnicity: Senoufo†, PP. resp. ethnicity: Dogon†, PP. resp. ethnicity: Tamacheq†, PP. resp. ethnicity: Bobo†, PP. resp. ethnicity: Bozo†, PP. resp. ethnicity: Other†, PP. resp. ethnicity: Foreign†, PP. resp is manager†.

- **Post harvest respondent (plot level)**
  - PH. resp. female†, PH. resp. female†, PH. resp. age, PH. resp. can read†, PH. resp. can write†, PH. resp. ethnicity: Bambara†, PH. resp. ethnicity: Sarakole†, PH. resp. ethnicity: Kassonke†, PH. resp. ethnicity: Senoufo†, PH. resp. ethnicity: Dogon†, PH. resp. ethnicity: Tamacheq†, PH. resp. ethnicity: Bobo†, PH. resp. ethnicity: Bozo†, PH. resp. ethnicity: Other†, PH. resp. ethnicity: Foreign†, PH. resp is manager†, Ph. resp. is PP resp..

- **Indexes (household level)**
  - Welfare index, Welfare index, Ag. equip index.

- **Geovars (ea level)**
  - HH Distance in (KMs) to Nearest Population Center with +20,000, HH Distance in (KMs) to Nearest Population Center with +20,000, HH Distance in (KMs) to Nearest Border Crossing, HH Distance in (KMs) to Nearest Road, HH Distance in (KMs) to the Boma of Current District of Residence, Annual Mean Temperature (degC * 10), Mean Temperature of Wettest Quarter (degC * 10), Long-term average NDVI value in primary growing season (highest quarter), Long-term maximum dekadal NDVI value in primary growing season (highest quarter), Agro-ecological Zones, Nutrient availability, Nutrient retention capacity, Rooting conditions, Oxygen availability to roots, Excess salts, Toxicity, Workability (constraining field management), Annual Precipitation (mm), Long term Avg total rainfall in wettest quarter(mm), Long term Avg annual total rainfall(mm), Slope (percent), Elevation (m), Potential Wetness Index, Avg start of wettest quarter in dekads 1-36, where first dekad of year =1.

- **Agricultural year rain fall (ea level)**[21]
  - Rain fall amount from May to August (planting season), Rain fall amount from May to August (planting season), Rain fall amount during wettest quarter, Start of the wettest quarter.

[21] Also interacted with plot level variables