IFPRI Discussion Paper 02334

April 2025

Parametric and Machine Learning Approaches to Examine Yield Differences Between Control and Treatment Considering Outliers and Statistical Biases

The Case of Insect Resistant/Herbicide Tolerant (IR/HT) Maize in Honduras

José Benjamín Falck-Zepeda

Patricia Zambrano

Arie Sanders

Carlos Rogelio Trabanino

Innovation Policy and Scaling Unit

INTERNATIONAL FOOD POLICY RESEARCH INSTITUTE

The International Food Policy Research Institute (IFPRI), a CGIAR Research Center established in 1975, provides research-based policy solutions to sustainably reduce poverty and end hunger and malnutrition. IFPRI's strategic research aims to foster a climate-resilient and sustainable food supply; promote healthy diets and nutrition for all; build inclusive and efficient markets, trade systems, and food industries; transform agricultural and rural economies; and strengthen institutions and governance. Gender is integrated in all the Institute's work. Partnerships, communications, capacity strengthening, and data and knowledge management are essential components to translate IFPRI's research from action to impact. The Institute's regional and country programs play a critical role in responding to demand for food policy research and in delivering holistic support for country-led development. IFPRI collaborates with partners around the world.

AUTHORS

José Benjamín Falck-Zepeda (j.falck-zepeda@cgiar.org) is a Senior Research Fellow in the Innovation Policy and Scaling (IPS) Unit of the International Food Policy Research Institute (IFPRI), Washington, DC.

Patricia Zambrano (p.zambrano@cgiar.org) is a Senior Program Manager with IFPRI's IPS Unit, Washington, DC.

Arie Sanders (asanders@zamorano.edu) is the Associated Dean Postgraduate Program at the Universidad Zamorano, Tegucigalpa, Honduras.

Carlos Rogelio Trabanino (rtrabanino@zamorano.edu) is an Associate Professor at the Universidad Zamorano, Tegucigalpa, Honduras.

Contents

Tables

Figures

ABSTRACT

Robust impact assessment methods need credible yield, costs, and other production performance parameter estimates. Sample data issues and the realities of producer heterogeneity and markets, including endogeneity, simultaneity, and outliers can affect such parameters. Methods have continued to evolve that may address data issues identified in the earlier literature examining genetically modified (GM) crops impacts especially those of conventional field level surveys. These methods may themselves have limitations, introduce trade-offs, and may not always be successful in addressing such issues. Experimental methods such as randomized control trials have been proposed to address several control-treatment data issues, but these may not be suitable for every situation and issue and may be more expensive and complex than conventional field surveys. Furthermore, experimental methods may induce the unfortunate outcome of crowding-out impact assessors from low- and middle-income countries. The continued search for alternatives that help address conventional survey shortcomings remains critical. Previously, existing assessment methods were applied to the impact assessment of insect resistant and herbicide tolerant maize adoption in Honduras in 2008 and 2012. Results from assessments identified endogeneity issues such as self-selection and simultaneity concurrently with influential outliers. Procedures used to address these issues independently showed trade-offs between addressing endogeneity and outliers. Thus, the need to identify methods that address both issues simultaneously, minimizing as much as possible the impact of method trade-offs, continues. We structured this paper as follows. First, we review the literature to delineate data and assessment issues potentially affecting robust performance indicators such as yields and costs differentials. Second, we discuss and apply four types of approaches that can be used to obtain robust performance estimates for yield and cost differentials including: 1) Robust Instrumental Variables, 2) Instrumental Variable Regressions, and 3) Control/Treatment, and 4) Machine Learning methods that are amenable to robust strategies to deal with outliers including Random Forest and a Stacking regression approach that allows for a number of "base learners" in order to examine the pooled 2008 and 2012 Honduras field surveys. Third, we discuss implications for impact assessment results and implementation limitations especially in low- and middle-income countries. We further discuss and draw some conclusions regarding methodological issues for consideration by impact assessors and stakeholders.

Keywords: Honduras, insect-resistant and herbicide-tolerant maize, yield impact, biases, machine learning, parametric and non-parametric methods.

ACKNOWLEDGMENTS

We thank the Staff and Students at Zamorano University, Honduras for the time and effort to collect, curate, and analyze the original data in which we base this case study paper. This paper was done under the purview of the CGIAR Genome Edited (GEDI) Science Project. The original data was collected with funding from the Templeton Foundation and UC-Davis PIPRA. This paper is the opinion of the authors and not those of IFPRI, Zamorano University, Templeton Foundation, or UC-Davis PIPRA.

ACRONYMS

API = Application Programming Interphase
ATE = Average treatment effect
BMA = Bayesian model averaging
DDML = Double debiased machine learning
GM = Genetically modified
GEd= Genome editing
IV = Instrumental variables
IR/HT = Insect-resistant and herbicide-tolerance
LATE = Local average treatment effects
LMICs = Low- and middle-income countries
ML= Machine learning
OLS = Ordinary least squares
OOE = Out-of-bag error
RANSAC= Random Sample Consensus
RCTs = Randomized control trials
RFR = Random forest regression
RMSE = Root mean square error
RMSPE = Root mean square percent error
2SLS = Second stage least squares
SVM = Support vector machine

1. INTRODUCTION

Economic impact assessments need robust production and performance parameter inputs when comparing control with treatment interventions. This is a critical need as practitioners need to deal with field data limitations including sampling and biases issues, outliers, and other day to day challenges. From the standpoint of economic impact assessment approaches, using robust yield and cost differences as input to economic models is significant, as they define the technology's performance and thus the economic assessment.

Differential control and treatment estimates for yield and costs may be elicited robustly from expert opinion consultations, biosafety and/or agronomic performance field trials, econometric projections from conventional surveys, and quasi-randomized or randomized control trials (RCTs). Obtaining robust indicators can be affected by sample and data issues such as endogeneity, simultaneity and sampling biases, outliers, flexibility, and irreversibility. This and the realities of producer heterogeneity and markets, increase the complexity and difficulty that economic impact assessors face in their work.

As noted in Smale et al. (2009), methods have continued to evolve that can address many of the limitations identified in the earlier literature examining the impact of genetically modified (GM) crops globally. This observation has not changed and is unlikely to change. Economic assessment practitioners know that methods introduce implementation tradeoffs which are important for all statistical approaches, as they convey an idea of their robustness. This is a well understood principle but one that is sometimes dismissed or ignored in practice. In some cases, new and exciting methods rise and fall in part due to the imperfect understanding of the gains and losses in information, accuracy and/or precision in applying such methods, which necessarily imply consideration of method implementation tradeoffs.

The emergence of advanced assessment methods (e.g., case control studies, stepped-wedge designs, randomized control trials, sequential multiple assignments, or micro-managed randomized trials) have provided more tools to examine the potential outcomes from using technology. However, not all

situations and/or research questions lend themselves to the application of such methods. Furthermore, these methods may be expensive, difficult to implement due to their complexity, have long implementation periods, and may crowd out researchers from low- and middle-income countries (LMICs) in their application due to their elevated cost (Kapur, 2020; Ahsanuzzaman and Zilberman 2024). In some cases, properly applied methods may have issues with generalization (see Deaton and Cartwright 2018 for the case of conventional RCTs). It is up to the assessor—taking into account available resources and assessment context—to determine the appropriate method or methods to use in practice, with the understanding that all methods have limitations and trade-offs during implementation.

Understanding these limitations and the complementarities between methods is mandatory for all economic assessment practitioners, who in our opinion must be agnostic and pragmatic by nature about method choice. Finding alternatives that help address the shortcomings of conventional surveys through proper design and using robust econometric methods that may help address such shortcomings is a never-ending story. Such a quest for the "best method" is unlikely to be successful as tradeoffs are necessarily involved in their implementation. This discussion paper's objective is to help address these questions faced by researchers in LMICs, who are dealing with these pragmatic issues in their own context.

To start addressing these issues, this paper is structured as follows. First, we introduce the context where we implement a set of methods to examine yield differences in Honduras for the adoption of a genetically modified insect-resistant and herbicide-tolerant maize. Second, we review the literature to delineate data and assessment issues that can affect obtaining robust performance indicators such as yields and costs. Third, we discuss four approaches that can be used to obtain robust performance estimates including robust instrumental variables, instrumental variable regressions, control/treatment, and machine learning (ML) approaches. The ML approaches include the random forest regression (RFR) and a stacking regression approach that estimates multiple "base learners" simultaneously. Fourth, we discuss results, and policy research implications. We conclude by discussing issues for consideration by relevant stakeholders including decision makers, regulators, producers and consumers as well as practitioners examining economic impact assessments.

2. NATIONAL CONTEXT AND BACKGROUND

Honduras was the first country in Central America to approve the cultivation of a genetically modified insect resistant and herbicide tolerant maize. The country is the first country in the region where genetically modified technology was commercialized in 1998. Falck-Zepeda et al. (2012) and Falck-Zepeda et al. (2015) reported results from surveys and analysis conducted in the country. The genetically modified technology using transgenesis commercialized in Honduras were maize hybrids with stacked insect-resistant and herbicide-tolerance (IR/HT).[1] Macall et al. (2020) estimated an economic surplus model examining actual gains from IR/HT maize adoption in the country. These three papers provide additional context and background for the use of IR/HT maize hybrids in the country. Issues identified in these previous studies for individual year analysis included the existence of influential outliers and endogeneity/biases in the collected data which may be the result of unobserved characteristics and the possibility that early adopter producers were better off than non-adopters.

In this paper, we take advantage of the data collected in surveys conducted in 2008 and 2012. We are not able to assemble both surveys into a quasi-panel dataset due to existing issues related to individual producer identification, difference in geographical locations, and end users included in each survey. However, we do build upon initial efforts to use separately robust regressions and instrumental variables models which were not satisfactory. As indicated in Falck-Zepeda et al. (2012; 2015) available methods at the time were not robust enough to be able to deal with outliers and endogeneity simultaneously. We now have available methods to deal with such issues as well as machine learning approaches, which can help address data collected in field surveys.

Table 1 shows that the pooled 2008 and 2012 dataset included 351 observations, of which 191 were adopters. Average yield for the pooled data was 4.76 tons/ha. Whereas adopters yielded 5.63 tons/ha, non-adopters' yielded only 3.72 tons/ha. Adopters had a smaller standard deviation than non-

[1] Insect resistance was incorporated via genes that expressed a protein derived from the *Bacillus thuringiensis* (Bt) soil bacteria. For the case of the herbicide tolerant trait, developers incorporated tolerance to the herbicide glyphosate. These two and similar traits have been approved by competent authorities and widely used in many countries around the world including the USA, Brazil, Argentina, the Philippines, South Africa, China, and others.

adopters and for the total sample. The smallest yield observed was 0.2 tons/ha, and the maximum yield observed was 9.7 tons/ha. Table 2 shows the descriptive statistics by type of producer and year. Yields were on average larger in 2008 than in 2012. Adopters had larger yields than non-adopters in both years. The standard deviation was slightly lower for non-adopters in 2008 and 2012, but taken together, the overall standard deviation was smaller for adopters.

Table 1 Mean, standard deviation, and frequency for adopter/non-adopter IR/HT maize Honduras 2008 and 2012

	Number of Observations	Mean (tons/ha)	SD	Min	Max
Adopters	191 (45.6%)	5.63	1.44	.98	9.70
Non-Adopters	160 (54.6%)	3.72	1.62	.19	7.36
Total	351 (100%)	4.76	1.80	.19	9.70
Difference between adopters and non-adopters		2.91		0.79	2.36

Table 2 Descriptive statistics by type of producer and year, for conventional (non-adopter) and IR/HT(adopter) maize Honduras

Type		2008	2012	Total
Non-adopter	Mean	4.99	3.18	3.72
	SD	1.26	1.45	1.62
	Frequency	48	112	160
Adopter	Mean	6.03	5.42	5.63
	SD	1.30	1.47	1.44
	Frequency	65	126	191
All producers	Mean	5.59	4.37	4.76
	SD	1.38	1.84	1.80
	Frequency	113	238	351
Difference between adopters and non-adopters	Mean difference	1.04	2.24	1.91

3. THE ECONOMIC AND STATISTICAL ISSUES AND APPROACHES

The research questions

Box 1 introduces a set of 4 research questions which we are pursuing in this and other papers. These research questions are based on the elements of best practice identified in Smale et al. (2009). They represent basic pragmatic questions and issues for those economic assessors which require robust estimates of yield and cost differences which may serve as input to economic models. We begin addressing these research questions in this paper, which can be considered an exploration into the use of advanced econometric and machine learning regression approaches to deal with endogeneity, outliers and other biases in the available datasets. The operating principle of the work done in this paper is to explore and exploit the method landscape, especially machine learning approaches, which may help address dataset limitation while empowering economic impact assessment efforts with access to field collected data.

Box 1. Research questions for field data and datasets

RQ1- Which econometric and statistical methods and machine learning algorithms have been used in the literature for crop yield prediction?

RQ2- Can econometric and statistical approaches and machine learning algorithms be used to address data and sampling issues such as endogeneity and outliers?

RQ3- What are the strategies that can be used with econometric approaches and machine learning algorithms to address data and sampling issues for crop yield difference estimation and prediction?

RQ4- What are challenges in the field of crop yield differences estimation and prediction using econometrics and machine learning algorithms, especially when using field survey data?

Smale et al. (2009) recommended considering biases and endogeneity as early as possible during the research process, starting from designing the research protocol for implementation. Estimation methods such as those described in this paper should not be considered a certain choice to correct for improper sampling design or data collection. In some cases, these methods may not be sufficiently robust to accomplish the task of addressing biases and other issues. These methods may be able to amend and/or

reduce the impact of such experimental design and implementation issues, but there is no certainty that methods will be able to resolve such issues.

Statistical/sampling issues to estimate robust yields differences

Selection bias and endogeneity are two common issues in statistical analysis, particularly in the context of causal inference. Selection bias refers to the situation where a sample is not representative of the population and may arise due to improper and/or non-random sampling strategies, self-selection, simultaneity, or incomplete data. If selection bias is present, estimates will be biased, and thus the real possibility of drawing erroneous conclusions exists. Beyond considering the sampling strategy behind the dataset and the need to ensure that elements of best practice are used in their implementation, practitioners may use propensity scoring matching, inverse probability weighting or instrumental variables type of statistical approaches to address systematic differences between the counterfactual or control and treatment.

Endogeneity arises when an explanatory variable is correlated with the error term in a model. The relationship between variable(s) and error terms can reduce the ability to define causality and its direction. This leads to issues such as simultaneity, omitted variables, sampling/measurement error, and reverse causality. Estimated parameters can be biased and, as with selection bias, can lead to erroneous conclusions, especially those related to causal relationships. Practitioners use instrumental variables, fixed effects or structural models, to control those variables and factors that may be causing endogeneity.

Outliers are unusual and extreme observations in many cases due to measurement errors. Outliers may not be readily explained but do carry information themselves that may be influential. Outliers may exist in independent and dependent variable space or both. In turn, a leverage point is one that has an unusual and sometimes extreme value which is different than the rest of the observations in the sample. Removing an influence point in a dataset will cause a large change in an estimated regression's coefficients. Leverage points may not exert influence if they are close to the estimated line, implying that specific points which are influential need to have leverage.

The importance of these issues is that classical instrumental variables are extremely sensitive to the presence of outliers in the sample analyzed. Outliers are important because they can strongly distort estimated effects of an independent variable into the dependent variable in instrumental variable approaches. Outliers do not have to be even "too leveraged" to be influential within the instrumental variables model application.

In our experience, due to the importance of the presence and influence of outliers (and of statistical biases), the need exists to implement diagnostic techniques to the point of being a standard procedure in an analysis of field surveys. Such diagnostic techniques can help detect outliers and potential biases signaling the need for the use of instrumental variables, robust regressions, and/or related approaches. Practitioners are well advised to note that conventional diagnostic techniques for the detection of outliers need to be applied judiciously. Because they are not based on estimators, conventional diagnostic techniques may not detect unusual outliers that are themselves not sufficiently robust. Furthermore, these approaches fail to address the combined outlier influence in first and second stage IV applications. These issues can be critical considerations in statistical analysis, especially when attempting to draw causal inferences from datasets.

In summary, paying close attention to research design, data collection methods, and appropriate statistical techniques is needed and is the first line of defense to ensure the validity and reliability of statistical conclusions. Understanding the tradeoffs involved with their application is crucial.

The estimation models

Machine learning methods[2] are increasingly used to model yields and other agricultural parameters of interest in output evaluations. Crop yields have been examined using ML, fuzzy logic, and Bayesian approaches by Huang et al. (2017), Garg et al. (2018), Kalairasi and Anbarasi (2022), Chergui

[2] There are two types of machine learning methods: supervised and unsupervised. Supervised models refer to those that are applied where we can identify an outcome and explanatory variables. These can be associated both with regression and classification approaches. Unsupervised models are applied in situations where the datasets consist of explanatory variables but without any being identified as an outcome.

(2022), Abbaszadeh et al., (2022), Brdar et al. (2011), Fernandes et al. (2017), among others. Random forest approaches crop yields as in Fukuda et al. (2013) and Jeong et al. (2016).

ML approaches can help address outliers. As shown in Box 2, assessors can follow a set of strategies to deal with outliers as suggested in the literature. These strategies themselves are elements of best practice while running ML approaches.

Box 2. Dealing with outliers in the ML literature

a. Use robust regression approaches such as Huber, RANSAC (random sample consensus), or Theil-Sen estimator regression.
b. Use ensemble type learners such as random forest or gradient boost.
c. Transform the target variable using log, Box-Cox, or Winsorization
d. Normalize/standardize variables. For example, *scikit-learn* in Python provides libraries including *MinMaxScaler* or *StandardScaler*.
e. Explore regularization techniques/regressions with a penalty including Ridge, Lasso, and Elastic Net
f. Do cross-validation. Using cross-validation, one can compare the performance metrics of the model with and without outliers. Common metrics used for regression models include mean squared error (MSE), root mean squared error (RMSE), and R-squared.

Source: For a very complete of outlier detection and management, see Aguinis et al. (2013).

We use the *pystacked* command in STATA to implement stacked generalization regression in machine learning (Ahrens, Hansen and Schaffer 2023). Stacking is a way of combining predictions from multiple supervised machine learners. These are known as the "base learners" which are used into a final prediction to improve performance (Wolpert, 1992). This is conceptually like the Bayesian model averaging (BMA) approaches which take outcomes from different models to derive a better prediction than any individual *model* as in Huang et al. (2017).

The *pystacked* command is an implementable API in STATA which takes advantage of Python's *scikit-learn* ML algorithms. The currently supported base learners in STATA's *pystacked* command include ordinary least squares (OLS), logistic, lasso/ridge/elastic net, support vector machines, gradient boosted, random forest and neural networks with multiple layers. It is important to note that the *pystacked*

command can be used as a single base learner or can be implemented to identify the best performing base learner among the options chosen.

Random forest regression is an ensemble of predictions from several decision tree regressions. Decision trees are a useful and pragmatic regression and classification tool in ML. This is a non-parametric method which is implemented by recursively portioning datasets through user specified partitioning and stopping criteria, which are usually based on entropy. Entropy approaches describe the information encoded within any dataset (Sepúlveda-Fontaine and Amigó 2024; Shannon 1948). Decision trees are often criticized due to their tendency toward overfitting. Overfitting occurs when the estimated model performs too well in the training dataset by attempting to match the model to the available data points too closely but performs poorly when applied to the testing dataset. This leads to low or poor predictive accuracy, which is known also as generalization accuracy. To reduce this possibility, ML approaches split datasets typically into training and testing data subsets, either in a specified ratio (e.g., 80 percent training vs. 20 percent testing) or through cross validation approaches which help define the most appropriate ratio.

A viable alternative to enhance generalization accuracy is to examine sample subsets and build individual trees which can later be summarized into one meta-decision tree through averaging methods or by selecting the best performing decision tree through voting mechanisms. The random forest approach is an algorithm that builds decision tree ensembles by averaging across multiple individual trees. The algorithm uses boot-strapping—described as bagging in the ML literature—to reduce overfitting and thus enhance generalization accuracy (Breinman 2001; Ho 1995). Bagging is implemented by fitting a decision tree to a bootstrapped sample derived from the overall sample.

We then applied *rforest,* a command in STATA that implements the random forest algorithm (Schonlau and You, 2020). The *rforest* approach considers a set of steps to run. Application of the random forest approach starts with the conventional pre-steps which include cleaning the dataset to eliminate blanks, scaling the data if necessary, splitting the overall dataset into training, and testing datasets. Once the pre-steps are completed, one needs to tune the hyperparameters including: 1) the

number of iterations and 2) the number of variables to ensure avoiding as much as possible overfitting or underfitting.

In Step 1 the need exists to determine the optimal number of iterations by using a cross validation approach and examining the out-of-bag error (OOE) and the root mean square error (RMSE). In our application, the number of iterations was determined to be 150. After finding the optimal number of iterations, we determined the optimal number of variables in Step 2, which is 14 in this case. After optimizing the hyperparameters, a final run using both optimal hyperparameters was done to obtain final estimates and predictions as needed.

Instrumental Variables (IV) are a statistical and econometric method used to address endogeneity and measurement issues. The IV methods help address causal effects by considering selection bias, self-selection, or confounding effects or variables which are unmeasured. Conceptually, the IV methods use an instrumental variable which is correlated with the predictor variables but uncorrelated with the response variable.

Importantly, IV methods are typically biased and are less precise than OLS; thus, their finite-sample properties are also typically biased. This implies that support for IV methods is asymptotic and has small sample issues which may be problematic. As a result, using a structured and diligent process to determine whether instruments are robust and how well these methods address issues is a critical part of the implementation process. A typical estimation approach is the two-stage least squares, where a regression model uses the instrumental variable as a predictor and estimates its predicted values in the first stage. In the second stage, a second regression is run using the predicted values as a predictor.

We make use of IV approaches in STATA, including *ivreg2* (Baum *et al. 2007)*. We use an approach to simultaneously address robustness, outliers, and bias/endogeneity using *robivreg*. We also used more sophisticated IV approaches including those using heteroskedasticity *ivreg2h* and instrumental variable quantile regression *ivqregress.*

ivreg2 implements the classical instrument variable model approach in STATA. The command allows the possibility of implementing two-stage least squares, generalized method of moments, limited

information maximum likelihood, and k-class estimators. The command allows us to define included exogenous regressors or instruments and exogenous regressor excluded from the regression. The command allows robust heteroskedasticity and autocorrelation variance estimates but also includes a set of diagnostic tests and statistics to help identify the appropriateness of instruments and the robustness of the approach.

robivreg is a robust IV command in STATA (Cohen, Freue, Ortiz-Molina, and Zamar 2011; Desbordes and Verardi 2012) that addresses robustness and endogeneity issues simultaneously. This approach has an enhanced estimator compared to other approaches, as it uses a weighting scheme that allows the IV process of computing usual identification and overidentifying tests as well as a generalized Hausman test for outlier presence.

We explored two additional IV alternatives: the instrumental variable quantile regression and the instrumental variable regression using heteroskedasticity-based instruments. We present the preliminary results from this exercise, although these assessments will require much more work for complete implementation. We intend to apply these two approaches in more detail in future papers. *Ivreg2h* is a STATA command that implements the IV approach which provides estimations using heteroskedastic-based instruments. This approach contemplates identifying structural parameters and the use of regressors uncorrelated with heteroskedastic errors. It is important to note that a higher degree of error heteroskedasticity is associated with a higher correlation of instruments to the included endogenous variables in the first stage regression. This technique may be used when there are no external instruments available or to enhance the capacity of external instruments to augment the efficiency of IV estimators. This includes those situations with binary and endogenous regressors. This includes those where outcome and treatment regressors may be binary.

ivqreg2 command module in STATA estimates the structural quantile functions defined by Chernozhukov and Hansen (2008) using the method of Machado and Santos Silva (2018a and 2018b). Notably, if no instruments are specified during approach coding, *ivqreg2* estimates the regression

quantiles imposing the restriction that quantiles do not cross (see He 1997). In turn, *ivqregress* uses the smoothed IV approach proposed in Kaplan and Sun (2017).

We estimated control/treatment approach using *etregress* in STATA (StataCorp 2023). This model can estimate an average treatment effect (ATE) and other parameters of a linear regression model augmented with an endogenous binary-treatment variable. A consistent estimator, *etregress* is a two-step full maximum likelihood estimator approach. In addition to the ATE, *etregress* can be used to estimate the ATE when the outcome is not conditionally independent of the treatment.

The data sources

We make use of two distinct datasets seeking to describe the performance of the stacked IR/HT maize hybrids in Honduras in 2008 and 2012. Zamorano University and IFPRI collected both datasets. Although attempts were made to build up a quasi-panel, differences in the number of respondents, instrument questions, and geographical diversity, as well as difficulties in correlating the identity of the respondents, precluded us from doing so. In a previous analysis, we used IV and robust models independently to deal with simultaneous issues in the dataset, but the methods to address them at the same time were either not readily available or not robust enough at the time (Falck-Zepeda *et al.* 2012; Falck-Zepeda 2015).

Figure 1 Outliers by type and year, IR/HT maize in Honduras, 2008 and 2012

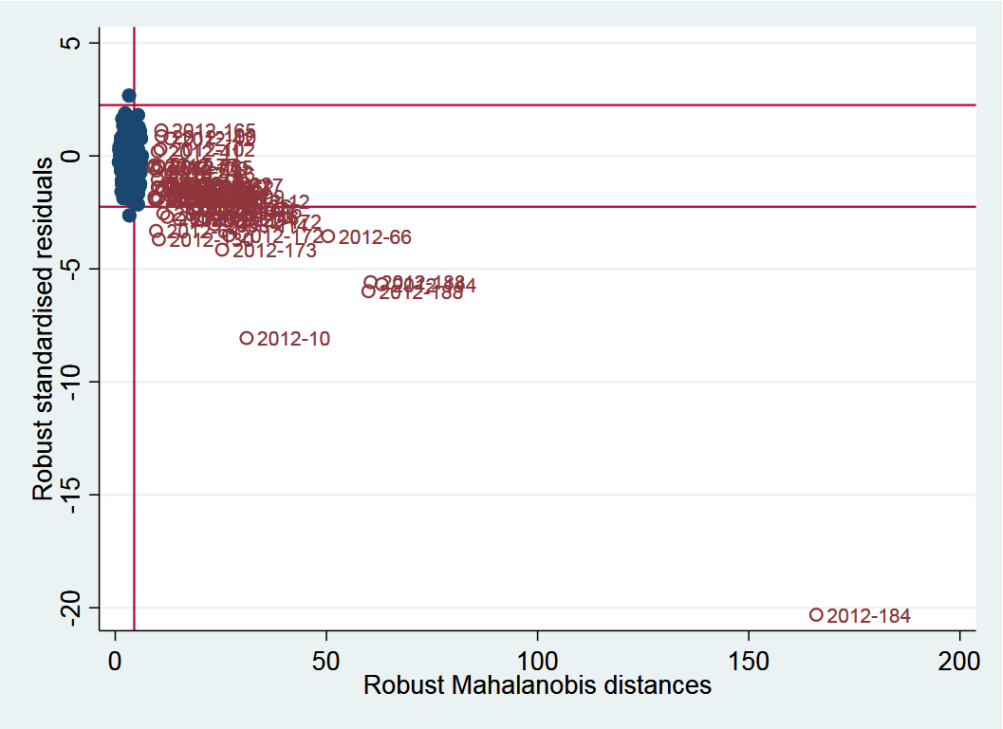

Figure 2 Outliers by individual observation, IR/HT maize in Honduras, 2008 and 2012

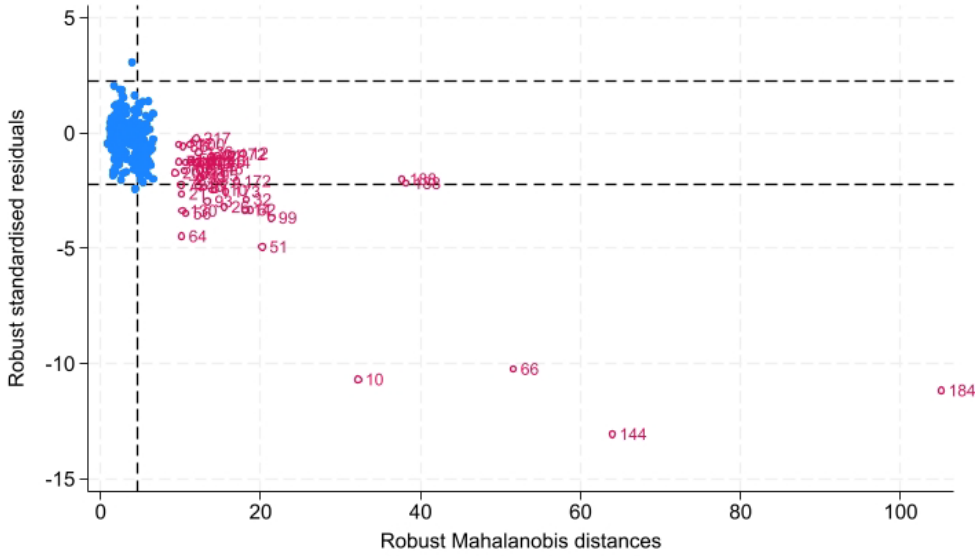

Table 3 Variables and definitions used in estimations

Variable	Definition
RENDMAIZ	Maize yields (tons/ha)
DPRODGEN	Dummy GM maize adoption (0 = non-adopter, 1 = adopter)
EXPERIENCE	Years of experience planting IR/HT maize
DYEAR	Dummy year survey conducted (0 = 2008, 1 = 2012)
TOTALAREA	Total area planted to maize
DIRRIG	Dummy irrigation used (0 = No, 1 = Yes)
DCREDIT	Dummy access to credit (0 = No, 1 = Yes)
DTECHASSIST	Dummy access to technical assistance (0 = No, 1 = Yes)
CULTIVATION	
INSECTICIDE	Insecticide applications (kg/ha)
HERBICIDE	Herbicide applications (kg/ha)
FERTILIZER	Fertilizer applied (kg/ha)
Constant	Constant

4. RESULTS

Figure 3 describes the in-sample predictions for individual base learners and for the stacked version that results from averaging across base learners in the model. As seen in Figure 3, we used OLS, lasso regression with cross validation, random forest, support vector machine (SVM), and gradient boost base learner regression to derive the stacking version. In both predictions, lasso regression is weighted more heavily, followed by random forest, gradient boost, and SVM base learners. OLS has either a very small weight or no weight.

Figure 3 In-sample predictions for Honduras IR/HT maize 2008 and 2012 using *pystacked*

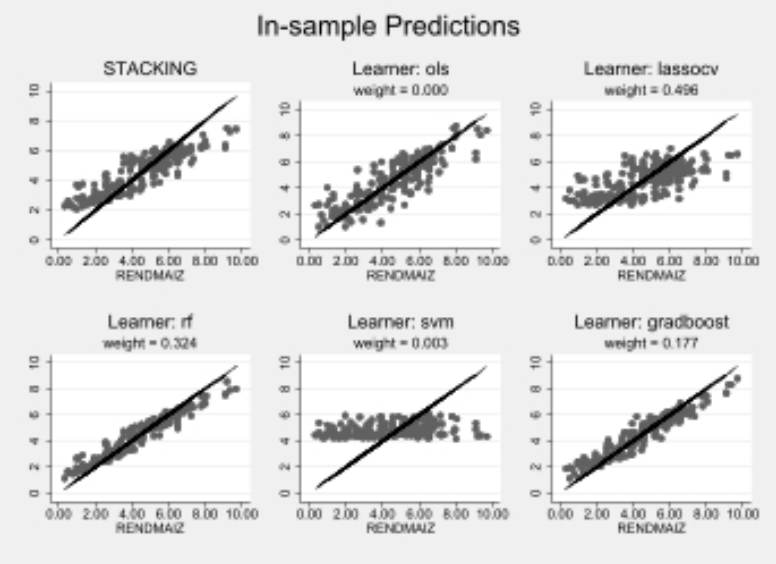

Figure 4 introduces the out-of-sample and in-sample predictions for all the base learners runs using *pystacked* in STATA. In this case, we used the option command to select the best base learner within the ensemble of learners. From the out-of-sample predictions, we can see that random forest is the best learner amongst the base learners used in the run; gradient boost seems to be a second-best. This is not entirely surprising given that the ML literature seems to indicate some advantage of random forest and gradient boost models—as ensemble tree models—to deal with outliers and other data issues (Gómez-Méndez and Joly 2023; Roy and Larocque 2012.

Figure 4 Out-of-sample predictions for IR/HT maize in Honduras, 2008 and 2012 using *pystacked*

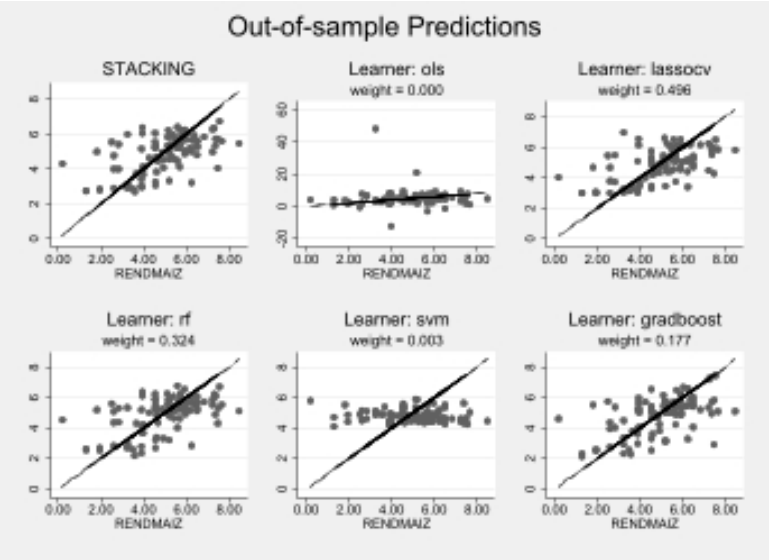

Figure 5 In-sample predictions for the best learner IR/HT maize in Honduras, pooled 2008 and 2012 using *pystacked*

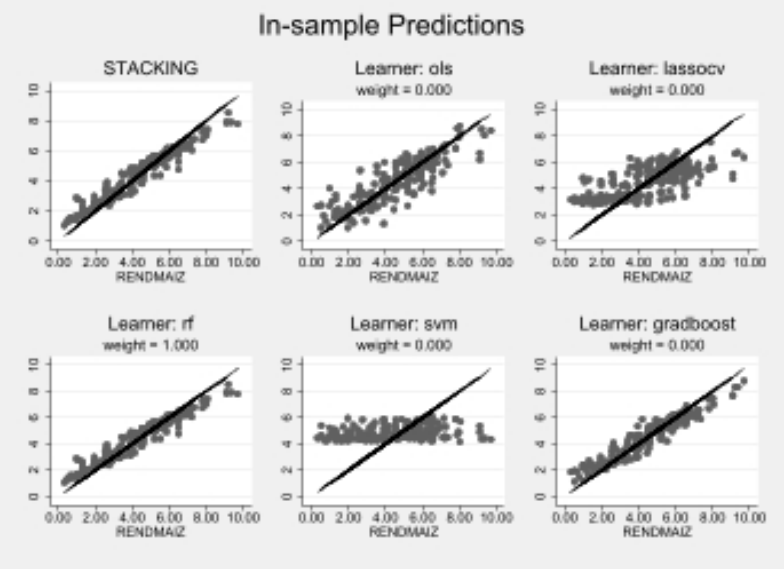

Figure 6 Out-of-sample predictions for the best learner IR/HT maize in Honduras, pooled 2008 and 2012 using *pystacked*

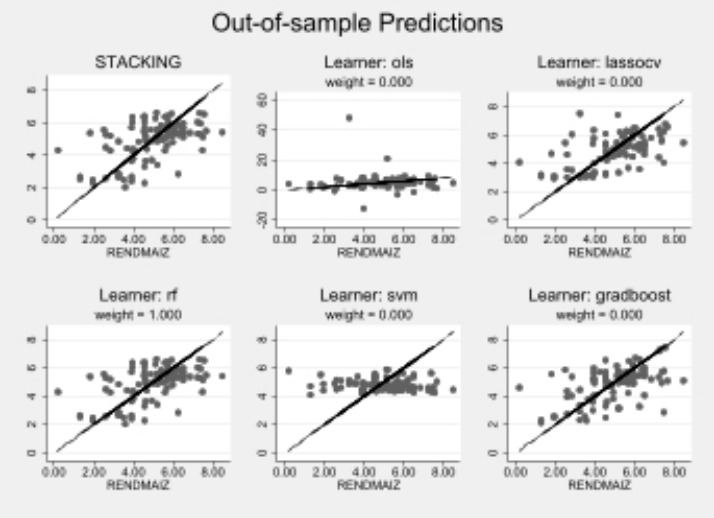

Figure 7 and 8 introduces the OOBE for tuning the number of iterations and hyperparameters variables. The first stage run of *rforest* helps determine that the number of variables at each randomly selected split was 15. We took this number and ran the regression to determine the optimal number of iterations, which was 150. Both numbers were used in a final run.

Figure 7 Out-of-bag errors (OOBE) and validation RMSE with hyperparameter optimization using *rforest* for IR/HT maize in Honduras, pooled 2008 and 2012

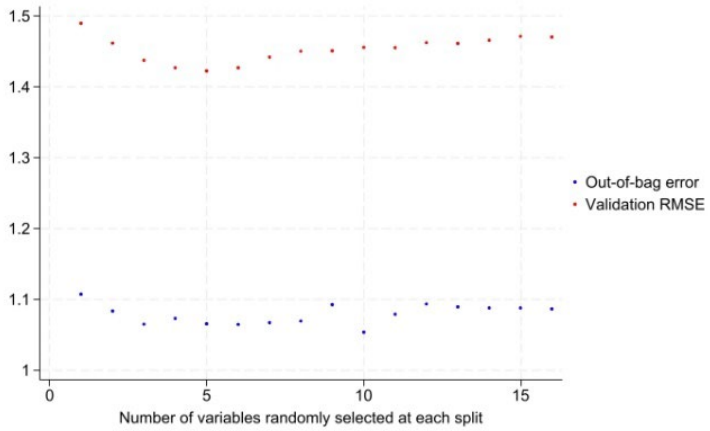

Figure 8 Out-of-bag errors (OOBE) for iterations and hyperparameters optimization using *rforest* in STATA for IR/HT maize in Honduras, pooled 2008 and 2012

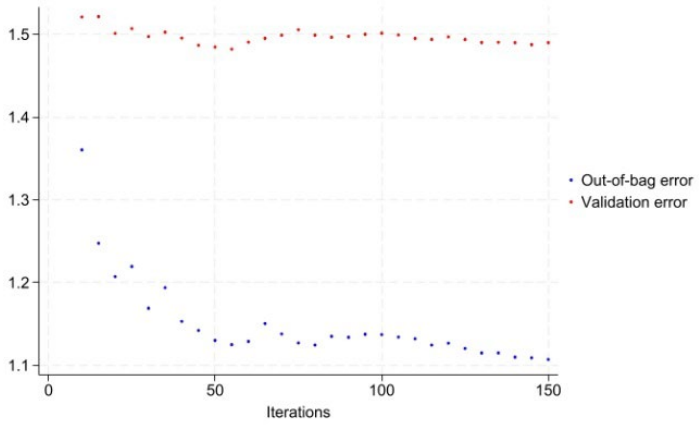

Figure 9 helps visualize the predicted values for adopters and non-adopters of the IR/HT maize in Honduras. Adopters have a higher predicted mean than non-adopters, at 5.57 tons/ha (standard deviation of 0.704) and 3.53 tons/ha (standard deviation of 1.239), respectively.

Figure 9 Predicted values for adopters and non-adopters of IR/HT maize in Honduras, pooled 2008-2012 datasets using *rforest*

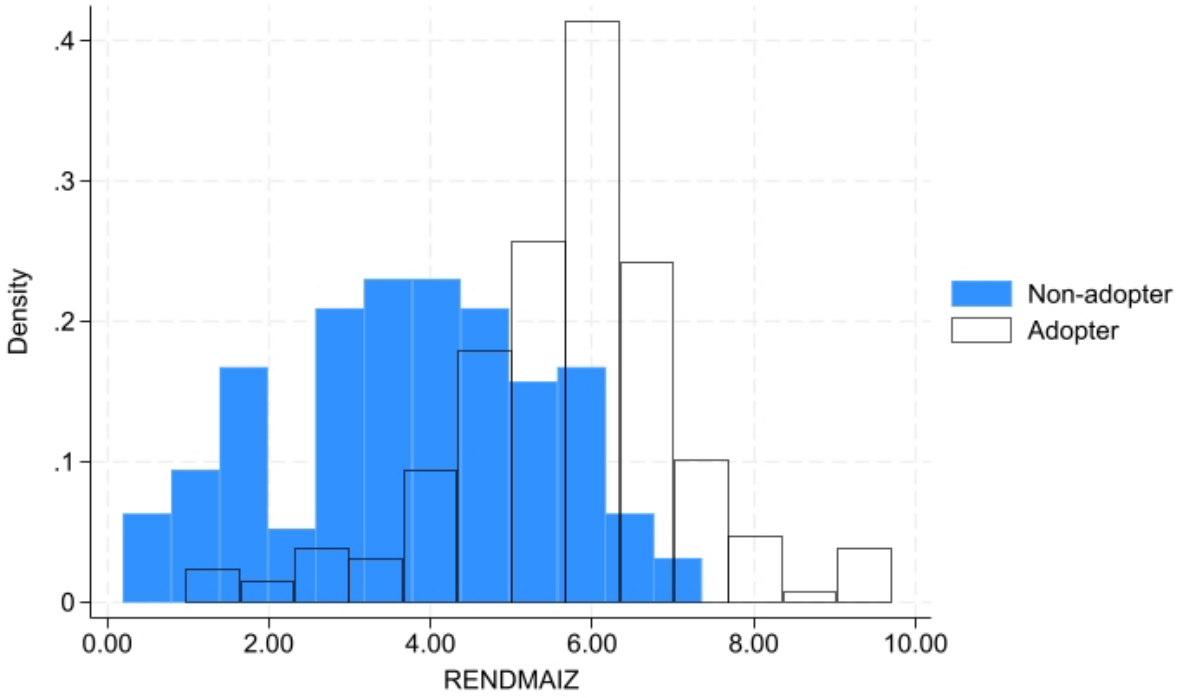

Table 4 introduces the results from the robust IV regression using *robivreg* in STATA. Results in Table 4 are for the second stage least squares (2SLS) regression. The coefficient for adopters of IR/HT maize is significant at the 1-percent level. Adopters of the IR/HT maize obtain a yield difference of 2.326 tons/ha. Other significant relevant variables include insecticide and herbicide use, as well as the difference between 2008 and 2012.

Table 4 Robust instrumental variables regression (*robivreg*), 2SLS regression

RENDMAIZ	**Coefficient**	**SE**	**t**	**P>t**	**95% confidence interval**	
DPRODGEN	2.326***	0.327	7.120	0.000	1.680	2.972
EXPERIENCE	-0.020**	0.008	-2.460	0.015	-0.036	-0.004
TOTALAREA	0.008*	0.004	1.740	0.084	-0.001	0.017
CULTIVATION	-0.002	0.002	-1.460	0.147	-0.005	0.001
INSECTICIDE	0.030***	0.008	3.540	0.001	0.013	0.046
HERBICIDE	0.020***	0.005	4.040	0.000	0.010	0.030
FERTILIZER	0.003	0.002	1.380	0.171	-0.001	0.006
DYEAR	-1.053***	0.319	-3.300	0.001	-1.683	-0.423
DIRRIG	0.001	0.325	0.000	0.997	-0.642	0.644
DCREDIT	-0.168	0.210	-0.800	0.426	-0.583	0.247
DTECHASSIST	-0.098	0.304	-0.320	0.748	-0.699	0.503
Constant	3.505***	0.458	7.660	0.000	2.600	4.409

Number of observations = 162, F (11,150) = 25.30, Prob > F = 0.0000
Total (centered) SS = 493.64, Centered R2 = 0.63, Total (uncentered) SS = 4285.93, Uncentered R2 = 0.95
Residual SS = 182.874014, Root MSE = 1.104

Notes: 1) ***= significant at the 1-percent level, **=significant at the 5-percent level, *=significant at the 10-percent level. 2) Estimates efficient and statistics consistent with homoskedasticity only. 3) SE= Standard Error.

In turn, Table 5 implements the smoothed IV quantile regression using *ivregress* in STATA. We followed the conventional steps for robustness in an instrumental variable estimation. Reported results are for the median regression; other quantiles results are available. The yield difference between adopters and non-adopters is 1.82 tons/ha. This regression approach controls for endogeneity, and instruments used are fairly robust to address endogeneity.

Table 5 Instrumental variable smoothed quantile regression, median regression (*ivqregress*)

Yield Maize	Robust Coefficient	SE	z	P>z	95% confidence interval	
DPRODGEN	1.82***	0.32	5.77	0.00	1.20	2.44
EXPERIENCE	-0.01	0.01	-0.75	0.45	-0.02	0.01
DYEAR	-0.98***	0.24	-4.16	0.00	-1.45	-0.52
TOTALAREA	0.00	0.00	0.05	0.96	0.00	0.00
DIRRIGATION	0.42	0.39	1.09	0.28	-0.34	1.18
DCREDIT	0.16	0.21	0.76	0.45	-0.25	0.58
DTECHASSIST	0.26	0.43	0.60	0.55	-0.58	1.10
CULTIVATION	0.00	0.00	-0.67	0.51	0.00	0.00
INSECTICIDE	0.00	0.00	1.46	0.14	0.00	0.01
HERBICIDE	0.00	0.00	0.72	0.47	-0.01	0.01
Constant	4.43***	0.32	13.62	0.00	3.79	5.06

Number of observations = 351, Wald $chi^2(90) = 547.93$, Prob > $chi^2 = 0.0000$
Notes: 1) ***= significant at the 1-percent level, **=significant at the 5-percent level, *=significant at the 10-percent level. 2) SE= Standard Error.

Figure 10 maps the coefficients for the dummy variable for IR/HT maize adoption (DPRODGEN) for each quantile. The red line is the estimate for the conventional IV regression. We can observe a variation across quantiles, although these variations fall within the 95 percent pointwise confidence interval.

Figure 10 Predicted and conventional instrumental variables using *ivqregress* by quantile for IR/HT maize in Honduras, 2008 and 2012

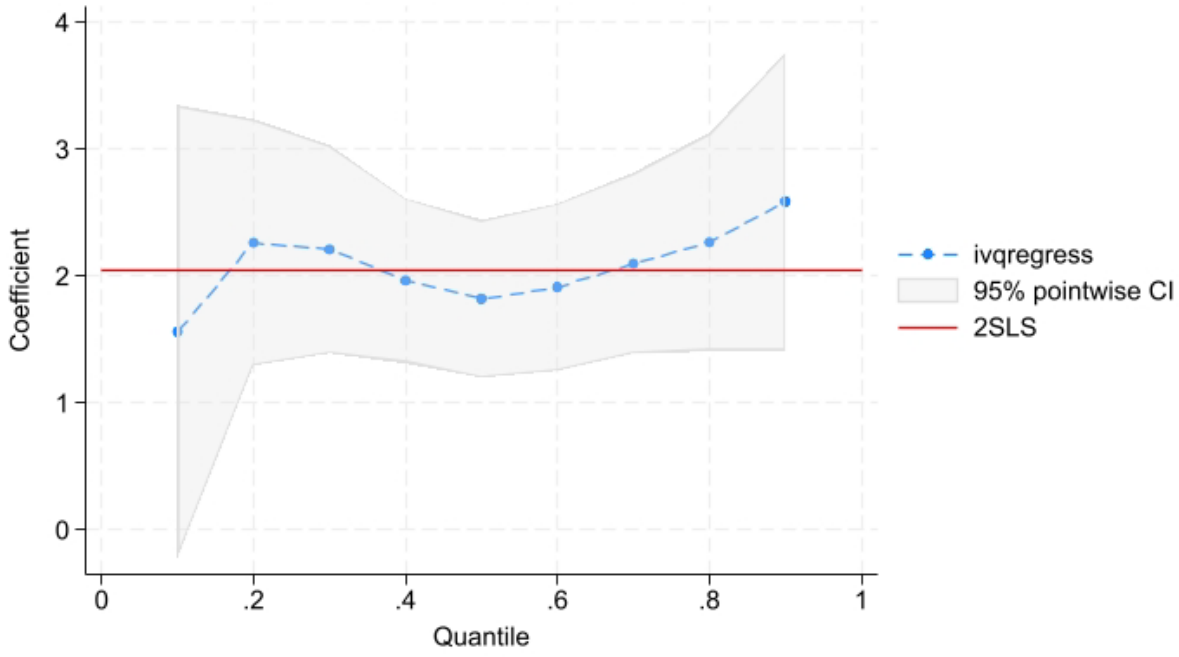

Table 6 introduces results for a linear regression model that utilizes endogenous binary treatment variables. This approach uses the routine *etregress* in STATA. This control-treatment approach shows that IR/HT maize adopters have a 2.133 tons/ha yield difference compared to non-adopters (conventional maize). Other relevant and significant variables include insecticide and fertilizer use, as well as the yield difference between 2008 and 2012.

Table 6 Linear regression with endogenous binary treatment variable using *etregress*

Variable	Coefficient	Robust SE	z	P>z	95% confidence interval	
RENDMAIZ						
EXPERIENCE	-0.009	0.006	-1.5	0.134	-0.020	0.003
DYEAR	-1.670***	0.228	-7.33	0.000	-2.117	-1.224
TOTALAREA	0.000	0.000	1.08	0.279	0.000	0.001
DIRRIG	0.275	0.276	1	0.319	-0.267	0.817
DCREDIT	0.047	0.166	0.28	0.778	-0.278	0.372
DTECHASSIST	-0.014	0.243	-0.06	0.956	-0.489	0.462
CULTIVATION	0.001	0.001	1.05	0.292	-0.001	0.002
INSECTICIDE	0.003***	0.001	2.7	0.007	0.001	0.005
HERBICIDE	0.003	0.002	1.29	0.196	-0.002	0.007
FERTILIZER	0.002***	0.001	3.12	0.002	0.001	0.003
DPRODGEN	2.133***	0.266	8.02	0.000	1.612	2.654
Constant	3.932***	0.294	13.38	0.000	3.357	4.508
DPRODGEN						
SEEDTOTALCOST	9.336***	1.557	5.99	0.000	6.284	12.389
SPriceUSMT	0.004***	0.001	4.67	0.000	0.003	0.006
Constant	-3.055***	0.377	-8.1	0.000	-3.794	-2.316
/ath Rho	-0.369**	0.143	-2.57	0.010	-0.650	-0.088
/lnsigma	0.343***	0.045	7.59	0.000	0.255	0.432
Rho	-0.353	0.126			-0.572	-0.088
Sigma	1.410	0.064			1.290	1.540
Lambda	-0.497	0.188			-0.867	-0.128

Number of observations = 351, Estimator: Maximum likelihood, Wald $chi^2(11) = 156.17$, Log pseudolikelihood = -788.71604, Prob > $chi^2 = 0.0000$

Notes: 1) ***= significant at the 1-percent level, **=significant at the 5-percent level, *=significant at the 10-percent level. 2) Wald test of independent equations (rho = 0): $chi^2(1) = 6.62$, Prob > $chi^2 = 0.0101$. 3) SE= Standard Error.

Table 7 introduces results of the instrumental variables with heterokedastic efficient estimates and robust statistics using *ivreg2h* in STATA. The yield advantage of IR/HT maize was 1.72 tons/ha. There is

a significant difference between years included in the analysis, with 2012 having a lower tield on average. Fertilizer use is significant but with a small effect on yield differences.

Table 7 Instrumental variables using heteroscedasticity efficient estimates and robust statistics (*ivreg2h*)

RENDMAIZ	Coefficient	Robust SE	T	P>t	95% confidence interval	
DPRODGEN	1.702***	0.562	3.030	0.003	0.596	2.808
EXPERIENCE	-0.012***	0.005	-2.260	0.024	-0.023	-0.002
DYEAR	-1.362***	0.236	-5.770	0.000	-1.826	-0.898
TOTALAREA	0.001	0.000	1.050	0.295	0.000	0.001
DIRRIG	0.374	0.284	1.320	0.189	-0.185	0.933
DCREDIT	0.059	0.171	0.350	0.728	-0.277	0.396
DTECHASSIST	-0.106	0.258	-0.410	0.682	-0.614	0.402
CULTIVATION	0.000	0.001	0.130	0.900	-0.002	0.002
INSECTICIDE	0.003**	0.001	2.570	0.010	0.001	0.005
HERBICIDE	0.003	0.002	1.130	0.258	-0.002	0.007
FERTILIZER	0.002***	0.001	2.730	0.007	0.001	0.003
Constant	4.214***	0.292	14.410	0.000	3.638	4.789

Number of observations = 351, F(11, 339) = 15.99, Prob > F= 0
Total (centered) SS = 1129.918631, Centered R^2 = 0.4095
Total (uncentered) SS = 9079.344828, Uncentered R2 = 0.9265
Residual SS = 667.2723254, Root MSE = 1.403
Notes: 1) ***= significant at the 1-percent level, **=significant at the 5-percent level, *=significant at the 10-percent level. 2) SE= Standard Error.

Table 8 summarizes results from the models included in this case study. There are significant differences between different models reflecting the influence of outliers and/or endogeneity in these estimates. The yield difference range between control and treatment was 1.03 and 2.33 tons/ha across methods. Quantile regression reflected a variation across quantiles of 30 percent between 1.03 and 2.59 tons/ha, with a tendency of lower yields associated with lower quantiles. This result is not unexpected.

Nevertheless, this work in progress will require more in depth analysis for some of the wrappers and libraries behind these models. For example, the stacking approach reports root mean square percent error (RMSPE) which needs to be converted to RMSE so that it becomes comparable to other approaches. Although we ran the ML learning with cross validation, we still need to fine tune hyperparameters in the individual learner models perform better.

Table 8 Summary maize yield results from models

Model	Coefficient	RMSE	R^2
Robust Instrumental Variables (*robivreg*)	2.33	1.10	0.63
Instrumental variables robust to heteroscedasticity (*ivreg2h*)	1.70	1.40	0.41
Quantile regression (*qreg2*)	1.22 (1.03 – 1.3)	n.a.	0.44
Smoothed IV quantile regression (*ivqregress*)	1.82 (1.56 – 2.59)	n.a.	n.a.
ML Stacking regression (*pystacked*)	1.99	1.33	n.a.
ML Random Forest (*rforest*)	2.04	1.46	n.a.
Endogenous Binary Treatment model (*etregress*)	2.13	n.a.	n.a.

Notes: 1) Reported results for the Quantile regression (*qreg2*) and Smoothed IV quantile regression (*ivregress)* are for the median regression. 2) For the Smooth IV regression *(ivregress*) reported values varied from a lower value of 1.56 to the higher value of 2.59 for the 10 and 90th percentile, respectively. 3) n.a.= not available, RMSE= Root Mean Square Error.

Table 9 summarizes the significance results for each of the variables in the estimated parametric models. Only the outcome variable (DPRODGEN) and the constant are highly significant across all parametric models. Significant explanatory variables included experience and year when the survey was conducted. Other input use variables such as insecticide, herbicide, or fertilizer use were significant in some models but not others.

Table 9 Summary significance results from parametric models

RENDMAIZ	ivreg2h	robivreg	ivqregress	etregress
DPRODGEN	***	***	***	***
EXPERIENCE	***	**	n.s.	n.s.
DYEAR	***	***	n.s.	***
TOTALAREA	n.s.	*	n.s.	n.s.
DIRRIG	n.s.	n.s.	n.s.	n.s.
DCREDIT	n.s.	n.s	n.s.	n.s.
DTECHASSIST	n.s.	n.s	n.s.	
CULTIVATION	n.s.	n.s.	n.s.	n.s.
INSECTICIDE	**	***	n.s.	***
HERBICIDE	n.s.	***	n.s.	n.s.
FERTILIZER	***	n.s.	n.s.	***
Constant	***	***	***	***

Note: 1) ***= significant at the 1-percent level, **=significant at the 5-percent level, *=significant at the 10-percent level.

5. DISCUSSION AND POLICY RESEARCH IMPLICATIONS

Dealing with biases and outliers for field data analysis is not easy as there are multiple issues and tradeoffs involved with estimation. As seen from the results presented in this paper and other experiences in the literature, models or approaches have specific abilities to address field data issues such as dealing with biases and outliers (Smale et al. 2009). Furthermore, there are tradeoffs involved with their use. There is simply no substitute for careful statistical and experimental design efforts prior to field work implementation. However, even if careful efforts are undertaken, facing the realities of field data in social sciences of heterogeneity, biases, and outliers are still the norm rather than the exception.

As indicated earlier in this discussion paper, newer approaches such as randomized and quasi-randomized approaches can help address some of these issues. These approaches may be expensive, not applicable to all research questions, and insufficient to address heterogeneity and other field data issues. This leaves practitioners in LMICs with the reality of using as many tools and approaches available to them, judiciously using those that address and embrace the tradeoffs involved with multiple and often competing issues such as statistical biases and outliers. The task at hand for LMICs practitioners is to understand the gains and losses involved with the different models and approaches available and to continue updating their toolbox with more sophisticated and/or novel approaches which may help with their estimation issues in practice. Here, we contrasted a set of four approaches which can be used to address field data issues.

ML algorithms are quite flexible in incorporating non-linearities, and in some cases, allow complex problems to be examined, but these approaches should be used prudently (Peet et al. 2022; Auret and Aldrich 2012). ML algorithms—as any black-box approach—may not be sufficient to help inform proper decision making as they may not be able to clearly define relationships between features and targets or outcomes (McAlexander and Mentch 2020).

Application of an approach that addresses outliers and statistical biases simultaneously, such as robust instrumental variables, can help balance both issues because IV models are sensitive to outliers and

biases. Application of a quantile regression approach may help identify significant differences between different segments of a population. In the case of LMIC agriculture, heterogeneity is the norm, rather than the exception. Trying to understand technology performance within the different segments of the producer population, especially of those lower quantiles which are usually the target population for the public sector in LMICs, continues to be a major challenge.

The lesson learned from the estimations done in this paper and in other quantitative exercises (Rita et al. 2023; Du and Feng 2025) is that parametric and non-parametric approaches can be and are complementary in dealing with multiple challenges. Challenges such as dealing with collected field data, even when pursuing state-of-the-art sampling, experimental design or random approaches, and other field collection strategies. At a minimum, method triangulation continues to be an element of best practice identified in Smale et al. (2009).

When dealing with the specific issues of endogeneity and outliers, we propose an integrated approach for an estimation strategy as follows:

1. Check for sampling and other statistical biases and issues (simultaneity, self-selection). If present and relevant, use Two stage/Heckman-based approaches.
2. Check for the presence of outliers and their influences. If present and relevant, use Robust Regression approaches.
3. If both present (endogeneity and outliers):
    a. Explore using robust IV approaches. It is important to note that there are tradeoffs involved with correcting for biases and outliers simultaneously.
    b. Explore using machine learning approaches, which may help to "smooth" influence of both especially in the case of random forest and SVM algorithms due to the use of multiple branches or ensembles. As described in this paper with the *pystacked* routine, using a "stacking" approach can help identify the best learner model for further exploration. This allows a more detailed tuning of hyperparameters, accomplished by

using individual learner models (e.g., SVM, random forest, lasso, or neural networks) in STATA and other statistical packages available in Python and R.

4. Explore cross-method comparisons to examine robustness.

One central question of interest to economic practitioners dealing with datasets is how ML can be used for causal inference (see multiple challenges in Box 3). The need exists to deal with local average treatment effects (LATE) issues, weak identification, and high-dimensional covariates. The LATE parameter measures causal effects of treatment for specific groups within a population. In many cases, they rely on instrumental variables approaches to deal with endogeneity. Weak identification relates to cases where there is weak correlation between instruments and endogenous variables. In these cases, IV estimation becomes imprecise, leading to unreliable statistical tests and confidence intervals. In turn, high dimensional covariates relate to big data, where datasets may be quite large. Inference methods become problematic.

ML literature has explored at least two major approaches (Leist et al. 2022; Brand, Zhou, and Xie 2023; Cui and Athey 2022). One explores treatment effect heterogeneity. This approach includes the use of causal forests and generalized random forest algorithms. The second pursues robust inference in the presence of high-dimensional controls and/or instruments. These are methods with uniformly asymptotic sample sizes. These approaches may be robust even with weak identification and high dimensionality.

Ongoing research in the ML literature explores, for example, the use of double debiased machine learning (DDML) approaches for causal inference (Ahrens et al. 2024; Chernozhukov 2018). This is one approach we are exploring for implementation. Employing DDML to estimate nuisance parameters, where we can infer the high-dimensional LATE. DDML algorithms may include inverting proposed statistics to derive confidence intervals and other ML approaches to overcome regularization bias and overfitting within the high-dimensional model.

Building on ML and even Big Data approaches have been facilitated by the emergence of STATA capabilities to run Python and R within do-files. This capability also allows the possibility of

running wrapper and other commands such as *pystacked*, which integrates STATA with Python and R, with existing estimation libraries along with graphing, data management, and other appealing approaches such as DDML (which uses *pystacked* as the "base model"). For example, using this capability allows using *scikit-learn, Pandas, XGBoost*, and other libraries in Python within STATA. This is a quite useful capability and tool which opens many possibilities and some issues, including tracing what is being done where in those programs and environments.

Box 3. Describing causal inference challenges in machine learning - The Artificial Intelligence version

1. Confounding variables: Identifying and accounting for confounding variables is crucial. These are factors that affect both the treatment and outcome, leading to biased estimates if not properly addressed.
2. Selection bias: When the sample data is not representative of the entire population, causal inferences can be misleading. Proper sampling techniques are essential to mitigate this bias.
3. Temporal Order: Establishing the correct temporal order between cause and effect is challenging. Sometimes, the cause and effect may occur simultaneously or in a complex sequence.
4. Measurement error: Accurate measurement of variables is critical. Measurement errors can introduce noise and affect causal estimates.
5. Endogeneity: Endogenous variables are influenced by other variables within the system. Untangling these relationships requires careful modeling.
6. Non-compliance: In experiments, participants may not adhere to assigned treatments. Handling non-compliance is essential for valid causal conclusions.
7. Heterogeneity: Effects may vary across different subgroups. Accounting for heterogeneity ensures robust causal inference.
8. Generalizability and external validity: Causal findings from one context may not apply universally. Understanding the external validity of causal relationships is essential.
9. Ethical challenges: Conducting experiments for causal inference may raise ethical concerns, especially when interventions involve risks.
10. Causal complexity – Causality vs correlation: Real-world systems are often intricate, with multiple causal pathways. Simplifying assumptions may not capture all nuances. Distinguishing between correlation and causation is essential for accurate causal inference. Failing to do so can lead to incorrect conclusions.
11. *Data quality and availability*: Obtaining high-quality data that captures relevant variables and accurately represents the underlying population is crucial for causal inference. Additionally, causal inference often requires longitudinal data, which may be expensive or difficult to obtain.
12. *Model complexity*: As models become more complex, interpreting and understanding causal relationships becomes more challenging. Complex models may capture spurious correlations or overfit to noise in the data, leading to inaccurate causal inferences.

Addressing these challenges often requires interdisciplinary approaches that combine expertise from statistics, machine learning, econometrics, and domain-specific knowledge. Moreover, ongoing research is essential to develop new methodologies and techniques that can improve the accuracy and reliability of causal inference in machine learning.

Source: Slightly edited and combined content from content generated by ChatGPT 3.5 and Bing's Co-Pilot

6. CONCLUDING COMMENTS

The quest for quantitative methods that help address field data gaps and limitations is ongoing and will need to continue for the foreseeable future, as field surveys remain viable options in many LMICs. In this case study, we explored four distinct approaches which can help economic impact assessors deal with field data issues including outliers and bias/endogeneity. This includes instrumental variables, robust instrumental variables, control/treatment, and machine learning models. Using one model over another will depend upon the presence of influential outliers, biases and endogeneity, and other sampling and experimental design issues. All models discussed here, with possibly the exception of conventional IV models, still need further refinement and more experience and expertise in their use.

The integration of Python (and the programming language R) into statistical software such as STATA has opened many avenues that can take advantage of cross environment and computer packages libraries and resources. This approach can be quite valuable in the economic assessor toolkit but also needs more work especially when defining limitations and the black-box nature inherent to ML models as well as the integration of models as highlighted in the text. This case study is a first step into the process of gaining more experience with these models and the model triangulation and interoperability that has the potential to drive more robust results. This approach, however, does not and should not distract from starting field work based on elements of best practice from project and experiment design to implementation, to postmortem data gaps.

APPENDIX

The evaluation procedures for ML models are often a complex procedure that contemplates a set of steps. Typically, such analysis considers: 1) choosing a validation method and/or strategy before running models; 2) selecting proper evaluation metrics, which may require combining with subject expertise; 3) documenting experiments and runs; and 4) comparing experiments/runs and choosing the best alternative. As with other quantitative approaches, the likelihood that no best model exists is almost certain. In this case, the second-best alternative is to choose a model that is good enough for in-sample and out-of-sample runs.

1. Choosing a model validation strategy

Resampling methods

Resampling methods are simple techniques for rearranging data samples to inspect whether the model performs well on data samples that it has not been trained on. In other words, resampling helps us understand whether the model can be generalized well.

Types of sampling approaches

Type	Approach	Notes
Random split	Random percentage of data into training, testing, and validation sets	• Higher chance that the original population is represented properly by sample • Reduces chance of a biased sample
Time-based split	Choose a split based on relevant *seasonality* and/or timeline	• Multiple data points may not be mutually independent over time • Major differences before and after split may make ML models unable to learn
k-fold cross-validation	Randomly shuffling the dataset and splitting the dataset into *k* groups. Each *k* group is considered a test set, and the remaining groups are aggregated into training sets. The model runs on the test group repeatedly with *k* different results from *k* different test groups.	• The best model chosen across the *k* different results
Stratified *k*-fold cross-validation	A modified *k*-fold approach considering values of the target variable into account. If the target variable is a categorical variable with *n* classes, the stratified *k*-fold approach ensures each test gets the same ratio for each of the two classes compared to the training set.	• The model is more accurate and less likely to be biased towards one or the other strata
Bootstrapping	This is a random-splitting method implemented via random sampling. This implies selecting a sample size (typically the size of the original dataset). Sample data points are selected randomly and added to the bootstrap sample. After the addition process has been completed,	• The model is trained on the bootstrap sample and evaluated on those data points that were not included in the bootstrapped sample.

	the sample is reinserted into the original sample. The process is repeated *N* times, where *N* is the sample size. This is a type of sampling approach using replacement as a basis for the procedure. A sample data point may be seen multiple times.	

2. Choosing model evaluation metrics

There are several evaluation metrics that can be used to make a comparison between ML and parametric models. Here, we describe the most used metrics in the ML and parametric model literature.

- Mean square error (MSE): The difference between the actual and predicted values as a measure of error. The error was squared, and the average was calculated for all errors. The drawback of this metric is that it is significantly sensitive to outliers, even when predictions may be well fit.
- Root mean square error (RMSE): The square root of the MSE. Scale down errors closer to actual values, facilitating interpretation. This is one of the most common metrics used to compare ML models.
- Mean absolute error (MAE): Mean of the absolute error values. Error values are defined as the actual minus the predicted values. MAE significantly reduces the penalty posed by outliers by not considering square values compared to MSE.
- Root mean squared log error (RMSLE): This estimation procedure uses the same procedure as RMSE, except that the actual and predicted values are expressed as log values. The effect of outliers is reduced by reducing the higher error rates when the log is applied.
- R-squared (R^2): Defined as the proportion of variance of target variables that can be explained using the use of independent variables. A widely used metric with important features. R^2 increases with the addition of predictors. One balancing issue related to ML is the ability to identify whether a model performs better with fewer predictors.
- Adjusted R-squared (A-R^2): This is an adjustment to standard R^2, where the score is penalized by the addition of more features.

3. Trade-offs with ML model selection

ML approaches take advantage of statistics and computational power through algorithms or models to define relationships and outcomes. Results from ML can be quite powerful in terms of enhancing accuracy; however, they may not be completely accurate. The most common expressions describing accuracy deviations are model bias and variance.

High bias occurs when the ML model pays too little attention to the training dataset because it strictly follows and is ruled by assumptions inbuilt into the estimating algorithm. When the ML model is exposed to the test dataset, the actual variables are typically nonlinearly related to the predictors or independent variables. High bias leads to underfitting.

High variance occurs when models concentrate too much on the training data by attempting to provide too accurate estimation of the relationship to existing data points in the training dataset. When high variance is present, the model will not generalize well to the test dataset or to any dataset that the model has not been exposed to before. This is called overfitting. When tracing the error level for both bias and variance, we found that they are inversely related to each other. To find an optimal model, one can trace both bias and variance and define the point at which they intersect–the point at which the total error is lowest. In practice, the efficiency and appropriateness of a model can be increased by iteratively fine-tuning the input parameters that are fed to the model function, which are known as hyperparameters. This evaluation uses appropriate metrics such as RMSE.

The proper procedure to achieve model training, build-up, and testing is to use learning curves. These are typically expressed as gains in terms of the improvement of the ML model over time. This is a measure of model learning. Two of the predominantly used learning curves are those used for model training and validation. These two learning curves are useful for defining how well a model generalizes or if it is not performing well. These are tradeoffs between bias and variance and, therefore, between overfitting and underfitting.

REFERENCES

Abbaszadeh, P., K. Gavahi, A. Alipour, P. Deb, and H. Moradkhani. 2022. "Bayesian Multi-Modeling of Deep Neural Nets for Probabilistic Crop Yield Prediction." *Agricultural and Forest Meteorology 314: 108773*. https://doi.org/https://doi.org/10.1016/j.agrformet.2021.108773.

Aguinis, H., Gottfredson, R. K., & Joo, H. 2013. "Best-Practice Recommendations for Defining, Identifying, and Handling Outliers." *Organizational Research Methods, 16(2):270-301.* https://doi.org/10.1177/1094428112470848

Ahrens, A., C. B. Hansen, M. E. Schaffer, and T. Wiemann. 2024. ddml: Double/debiased machine learning in Stata. The Stata Journal, 24(1):3–45.

Ahrens, A., Hansen, C. B., and M.E. Schaffer. 2023. "pystacked: Stacking generalization and machine learning in Stata." *The Stata Journal, 23(4):909-931*. https://doi.org/10.1177/1536867X231212426

Ahsanuzzaman, H. H. and D. Zilberman. 2024. "Complementarity of field studies and RCTs: evidence from Bt eggplant in Bangladesh." *European Review of Agricultural Economics*, jbae003, https://doi.org/10.1093/erae/jbae003.

Auret, L. and C. Aldrich. 2012. Interpretation of nonlinear relationships between process variables by use of random forests, Minerals Engineering, 35: 27-42. https://doi.org/10.1016/j.mineng.2012.05.008.

Baum, C. F. and M.E. Schaffer. 2012. ivreg2h: Stata module to perform instrumental variables estimation using heteroskedasticity-based instruments." *Statistical Software Components S457555*, Boston College Department of Economics, revised 26 Sep 2024. http://ideas.repec.org/c/boc/bocode/s457555.html

Baum, C. F., M. E. Schaffer, and S. Stillman. 2007. "Enhanced routines for instrumental variables/generalized method of moments estimation and testing." *Stata Journal 7: 465–506.*

Brdar, S., D. Ćulibrk, B. Marinković, J. Crnobaracy and V. Crnojević. 2011. Support Vector Machines with Features Contribution Analysis for Agricultural Yield Prediction. Conference paper presented at the Second International Workshop on Sensing Technologies in Agriculture, Forestry and Environment (EcoSense 2011), Belgrade, Serbia. Conference Paper downloaded from https://www.researchgate.net/publication/301216536

Brand, J.E., X. Zhou, and Y. Xie. 2023. Recent Developments in Causal Inference and Machine Learning. *Annual Review of Sociology*, 49:81-110. https://doi.org/10.1146/annurev-soc-030420-015345

Breiman L. 2001. "Random forests." *Machine Learning, 45(1):5–32*.

Chergui, Nabila. 2022. "Durum Wheat Yield Forecasting Using Machine Learning." *Artificial Intelligence in Agriculture 6: 156–66.* https://doi.org/https://doi.org/10.1016/j.aiia.2022.09.003.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, J. Robins. 2018. Double/debiased machine learning for treatment and structural parameters, *The Econometrics Journal,* 21(1): C1–C68, https://doi.org/10.1111/ectj.12097

Chernozhukov, V. and C. Hansen. 2008. "Instrumental Variable Quantile Regression: A Robust Inference Approach," *Journal of Econometrics,142:379-398.*

Cohen Freue, G. V., H. Ortiz-Molina, and R. H. Zamar. 2013. A Natural Robustification of the Ordinary Instrumental Variables Estimator, *Biometrics*, Volume 69, Issue 3, September 2013, Pages 641–650. https://doi.org/10.1111/biom.12043

Cui, P. and S. Athey.2022. Stable learning establishes some common ground between causal inference and machine learning. *Nature Machine Intelligence,* 4, 110–115 (2022). https://doi.org/10.1038/s42256-022-00445-z

Deaton, A. and N. Cartwright. 2018. Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine, 210:2-21*.

Desbordes, R. and V. Verardi. 2012. A robust instrumental-variables estimator. *The Stata Journal, 12(2):169–181.*

Du, Z., Feng, H., and J. Arbuckle. 2025. Exploring the complementarity between traditional econometric methods and machine learning – an application to adoption and disadoption of conservation practices. *Applied Economics*, 1–16. https://doi.org/10.1080/00036846.2025.2462792

Falck-Zepeda, J.B., A. Sanders, C. R. Trabanino, R. Batallas-Huacon. 2012. "Caught between Scylla and Charybdis:impact estimation issues from the early adoption of GM maize in Honduras." *AgBioForum, 15(2):138–151.*

Falck-Zepeda, J. B., P. Zambrano, D. McLean, A. Sanders, M.M. Roca, C. Chi-Ham, and A. Bennett. 2015. Honduras and Bt/HT maize – a small country model for GM crop adoption? In Analyses: Africa's future ... can biosciences contribute? Mitton, Patrick; Bennett, David (Eds.). Chapter 11. Pp. 106-118.

FAOSTAT. 2024. Food and Agriculture Organization of the United Nations,Statistics Division. https://www.fao.org/ home/en/

Fernandes, J. L., N. F. Favilla Ebecken, and J. C. Dalla Mora Esquerdo. 2017. "Sugarcane Yield Prediction in Brazil Using NDVI Time Series and Neural Networks Ensemble." *International Journal of Remote Sensing 38 (16):4631–44*. https://doi.org/10.1080/01431161.2017.1325531.

Fukuda, S., W. Spreer, E. Yasunaga, K. Yuge, V. Sardsud, and J. Muller. "Random Forests modelling for the estimation of mango (Mangifera indica L. cv. ChokAnan) fruit yields under different irrigation regimes." *Agricultural Water Management, 116:142–50.*

Garg, B., S. Aggarwal, and J. Sokhal. 2018. "Crop Yield Forecasting Using Fuzzy Logic and Regression Model." *Computers & Electrical Engineering 67:383–403*. https://doi.org/https://doi.org/10.1016/j.compeleceng.2017.11.015.

Gómez-Méndez, I., and Joly, E. 2023. Regression with missing data, a comparison study of techniques based on random forests. *Journal of Statistical Computation and Simulation*, 93(12), 1924–1949. https://doi.org/10.1080/00949655.2022.2163646

He, X. 1997. "Quantile Curves Without Crossing," *The American Statistician, 51:186-192*.

Ho T. K. 1995. Random decision forests. In Proceedings of 3rd International Conference on Document Analysis and Recognition, 278–282. Piscataway, NJ: IEEE. https://ieeexplore.ieee.org/abstract/document/598994

Huang, X., G. Huang, C. Yu, S. Ni, and L. Yu. 2017. "A Multiple Crop Model Ensemble for Improving Broad-Scale Yield Prediction Using Bayesian Model Averaging." *Field Crops Research 211:114–24*. https://doi.org/https://doi.org/10.1016/j.fcr.2017.06.011.

Jeong, J. H., J. P. Resop, N. D. Mueller, D. H. Fleisher, K. Yun, E. E. Butler, D. J. Timlin, et al. 2016. "Random Forests for Global and Regional Crop Yield Predictions." *PLOSone 11 (6): e0156571*. https://doi.org/10.1371/journal.pone.0156571.

Kalaiarasi, E, and A Anbarasi. 2022. "Multi-Parametric Multiple Kernel Deep Neural Network for Crop Yield Prediction." Materials Today: Proceedings 62: 4635–42. https://doi.org/https://doi.org/10.1016/j.matpr.2022.03.115.

Kaplan, D. M. 2022. "Smoothed instrumental variables quantile regression." *Stata Journal 22: 379–403*.

Kaplan, D. M., and Y. Sun. 2017. "Smoothed estimating equations for instrumental variables quantile regression." *Econometric Theory 33:105–157*. https://doi.org/10.1017/S0266466615000407.

Kapur, D. 2020. "Poverty, power and RCTs." *World Development, 127:104811.*

Klompenburg, Thomas van, Ayalew Kassahun, and Cagatay Catal. 2020. "Crop Yield Prediction Using Machine Learning: A Systematic Literature Review." *Computers and Electronics in Agriculture* 177: 105709. https://doi.org/https://doi.org/10.1016/j.compag.2020.105709.

Lewbel, A, 2012. "Using Heteroscedasticity to Identify and Estimate Mismeasured and Endogenous Regressor Models." *Journal of Business and Economic Statistics, 30(1):67-80.* http://fmwww.bc.edu/EC-P/wp587.pdf.

Lewbel, A, 2016. Identification and Estimation Using Heteroscedasticity Without Instruments: The Binary Endogenous Regressor Case. Boston College Economics Working Paper 927. http://fmwww.bc.edu/EC-P/wp927.pdf

Lewbel, A, 2018. “Identification and Estimation Using Heteroscedasticity Without Instruments: The Binary Endogenous Regressor Case.” *Economics Letters, 165:10-12.*

Leist, A. K., M. Klee, J. H. Kim, D. H. Rehkopf, S. P. A. Bordas, G. Muniz-Terrera, and S. Wade. 2022. Mapping of machine learning approaches for description, prediction, and causal inference in the social and health sciences. Science Advances, 8,eabk1942.https:// doi.org/10.1126/sciadv.abk1942

Macall, D.M., C. R. Trabanino, A. H. Soto, et al. Genetically modified maize impacts in Honduras: production and social issues. 2020. *Transgenic Research, 29:575–586.* https://doi.org/10.1007/s11248-020-00221-y

Machado, J.A.F. and J. M. C. Santos Silva. 2018a. “Quantiles via Moments.” *Journal of Econometrics, 213(1):145-173*.

Machado, J. A. F. & J. M. C. Santos Silva, 2018b. "IVQREG2: Stata module to provide structural quantile function estimation," Statistical Software Components S458571, Boston College Department of Economics, revised 06 Mar 2023.

McAlexander, R. J., and L. Mentch. 2020. Predictive inference with random forests: A new perspective on classical analyses. *Research & Politics*, 7(1). https://doi.org/10.1177/2053168020905487

Pavani, S. and A. S. Beulet P. “Improved Precision Crop Yield Prediction Using Weighted-Feature Hybrid SVM: Analysis of ML Algorithms.” *IETE Journal of Research, 70(3)2628-2640*. https://doi.org/10.1080/03772063.2023.2192000

Peet, E.D., B. G. Vegetabile, M. Cefalu, J. D. Pane, C. L. Damberg. 2022. Machine Learning in Public Policy The Perils and the Promise of Interpretability. Santa Monica, California, Rand Corporation. Published November 2022. https://www.rand.org/pubs/perspectives/PEA828-1.html

Rita, R., V. Marques, D. Bárbara, I. Chaves, P. Macedo, V. Moutinho, and M. Pereira. 2023. Crossing non-parametric and parametric techniques for measuring the efficiency: Evidence from 65 European electricity Distribution System Operators, *Energy*, 283:128511. https://doi.org/10.1016/j.energy.2023.128511

Rousseeuw, P. J. 1985. Multivariate estimation with high breakdown point. In Mathematical Statistics and Applications, ed. W. Grossmann, G. Pflug, I. Vincze, and W. Wertz, vol. 8, 283–297. Dordrecht: Reidel.

Roy, M. H., and D. Larocque. 2012. Robustness of random forests for regression. *Journal of Nonparametric Statistics*, 24(4), 993–1006. https://doi.org/10.1080/10485252.2012.715161

Sanglestsawai, S., R. M. Rejesus, J. M. Yorobe. 2014. “Do lower yielding farmers benefit from Bt corn? Evidence from instrumental variable quantile regressions.” *Food Policy, Volume 44:285-296*. https://doi.org/10.1016/j.foodpol.2013.09.011.

Schonlau, M. and R.Y. Zou. 2020. “The random forest algorithm for statistical learning.” *The Stata Journal, 20(1):3–29.*

Sepúlveda-Fontaine, S. A., and J.M. Amigó. 2024. Applications of Entropy in Data Analysis and Machine Learning: A Review. *Entropy*, 26(12), 1126. https://doi.org/10.3390/e26121126

Shannon, C.E. 1948. A mathematical theory of communication. *Bell System Technical Journal*, 27, 379–423.

Smale, M., P. Zambrano, G. P. Gruère, J. B. Falck-Zepeda, I. Matuschke, D. Horna, L. Nagarajan, I. Yerramareddy, H. Jones. 2009. Measuring the economic impacts of transgenic crops in developing agriculture during the first decade: Approaches, findings, and future directions. Food Policy Review 10. Washington, D.C. International Food Policy Research Institute (IFPRI). http://www.ifpri.org/sites/default/files/publications/pv10.pdf http://dx.doi.org/10.2499/0896295117FPRev10

StataCorp. 2023. Stata: Release 18. Statistical Software. College Station, TX: StataCorp LLC.

Verardi, V., and C. Croux. 2009. Robust regression in Stata. Stata Journal 9: 439–453.

Verardi, V. and C. Croux. 2010. Software Update: st0173 1: Robust regression in Stata. Stata Journal 10: 313

Wolpert, D. H. 1992. Stacked generalization. *Neural Networks* 5 (2): 241-259. https://doi.org/10.1016/S0893-6080(05)80023-1.

Zheng, Y., A. Gracia, L. Hu. 2023. “Predicting Foodborne Disease Outbreaks with Food Safety Certifications: Econometric and Machine Learning Analyses.” *Journal of Food Protection. 86:100136.* https://doi.org/10.1016/j.jfp.2023.100136

ALL IFPRI DISCUSSION PAPERS

All discussion papers are available **here**

They can be downloaded free of charge

INTERNATIONAL FOOD POLICY RESEARCH INSTITUTE
www.ifpri.org

IFPRI HEADQUARTERS
1201 Eye Street, NW
Washington, DC 20005 USA
Tel.: +1-202-862-5600
Fax: +1-202-862-5606
Email: ifpri@cgiar.org