IMPROVING HARVEST SURVEYS

LEARNING FROM PULA INPUT INSURANCE

Daniel Mellow Data Specialist, Busara Center for Behavioral Economics

mastercard foundation

MERCY CORPS | AGRIFIN

PROJECT OVERVIEW

Pula provides index-based input insurance products in Kenya, Zambia, Malawi and Nigeria. These products avoid the need for beneficiaries to file claims by automatically replacing inputs in times of drought or widespread crop failure in the beneficiary's agro-ecological zone.

In Nigeria, Pula conducts harvest surveys across the country on a subsample of their customers to determine when the policy is paid.

In 2019, the Busara Center was commissioned to improve the harvest survey in Nigeria by analyzing existing data. Our analysis had three objectives:

BUILD A MODEL to predict yield based on variables that can be known at the beginning of the season, to improve the ability of Pula to forecast payouts.

UNDERSTAND BEST PRACTICES for Pula-insured farmers.

SELECT THE MOST PREDICTIVE VARIABLES from the harvest survey to be included in more targeted data collection efforts.

PULA INPUT INSURANCE

Pula partners together with input suppliers and other smallholder farmer service providers to bundle their insurance products with products that farmers already buy, such as seeds, fertilizer or credit.

Farmers register at the retailer, using the agent's smartphone app and a code embedded in the product they bought.

Payouts are determined by geography, in response to drought or bad harvest, rather than individual circumstances.

The Data

Pula attempted to survey:

3,152

Rice & Wheat Farmers

Survey Period:

2015-19

Yield Estimation:

Crop-cut Method

2,468

Successful Yield*

Estimation

*We classify a survey as 'successful' if the survey team was able to complete the crop cut exercise and measure yield. The main reason for survey non-completion was that the farmer had already harvested.

For analysis purposes:

2,164

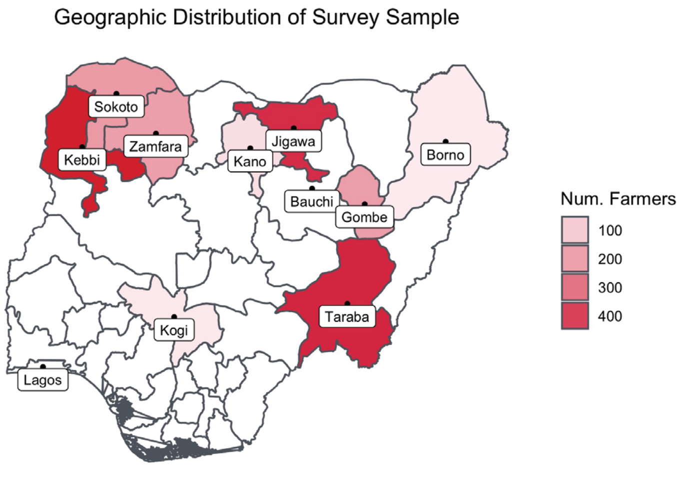
Rice Farmers

Survey Period:

2018

Geographic Distribution Of Survery Sample

Pula farmers are concentrated in irrigated areas in northern Nigeria, many with access to rivers and dams. The median farmer cultivates 2 acres of land. 38% managed to receive fertilizer from ABP, though the majority of those report the supplies arrived too late to be useful.



Lagos added for reference.

The Data

In the analysis we considered:

2,164
Rice Farmers

These farmers were associated with Pula's agribusiness partners

- 933 NIRSAL
- 652 CGAP
- 579 Afex

Data on each farmer include:

- The specific crop-cut measurements
- Precise location
- Practices such as intercropping and irrigation
- Inputs such as fertilizer, pesticides and acerage
- Planting and harvesting time

Variables not available:

- Demographics such as gender and age
- Other farming or income-generating activities
- Weather*
- Historical yield*

* Busara used public data sources to develop local estimates for these measures

PREDICTING YIELD

Approach

Regression models vs. machine learning

Regression models

easy to interpret and allow for inference (determining statistical significance)

Machine learning

are more flexible and generally provide better predictions

Selected method
Existing Limitation:
Machine learning models are “black boxes”, in that the results cannot be interpreted.

Information on input choices is highly predictive of yield

*Requires high compliance in data collection predicting allocation of resources

Ideal method
Existing Limitation:
Depending on the implementation, however, Pula may not have access to this information when needing to predict where to allocate resources.

Model Fit

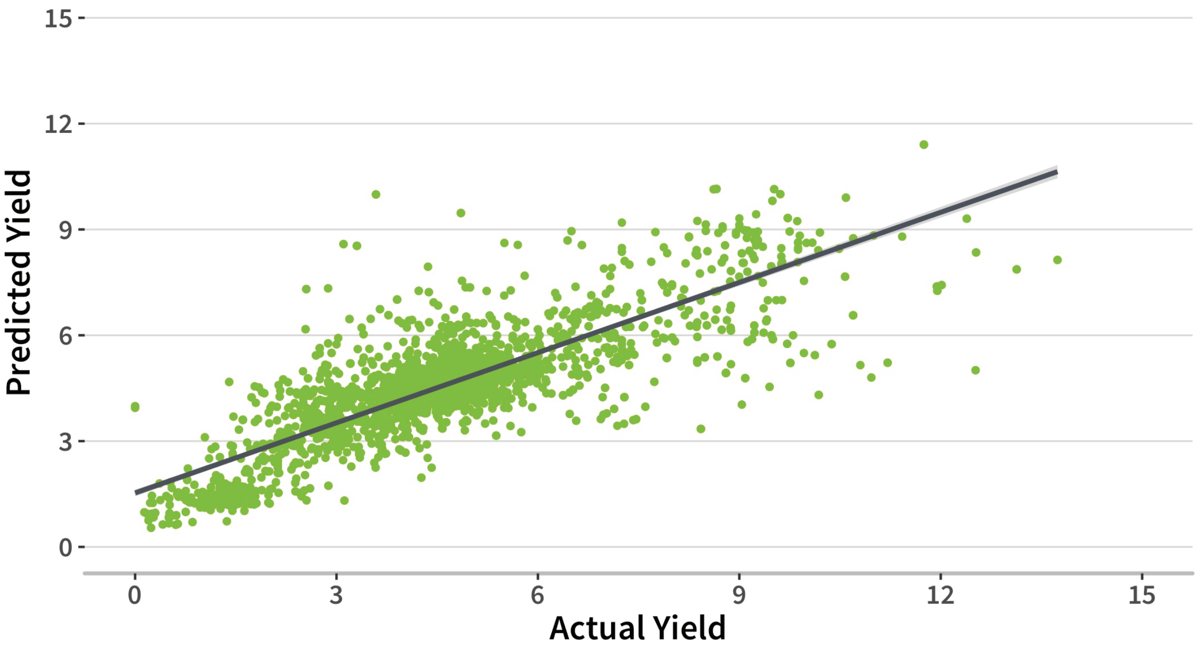
The best performing model was a random forest framework using all the input data.

We calibrated the model using 1,000 trees, a minimum node size of 5 and one third of variables randomly sampled at each split

In this model, cross-validated R-squared was 0.67, meaning 67% of variance is explained by the model, comparable to methods using satellite data. In practice this means individual farmer yield could be predicted, on average, to within a tenth of a standard deviation.

By comparison, a regression with all variables can only explain 29% of variance and a random forest without inputs 49%.

Rice Farmers, 2018



Size of points proportion to size of plot. Yield measured in MT/ha.

Two Methods for Selecting Variables

1

Variable Importance

1. Build a model with all relevant variables.

2. Use metrics of variable importance.

3. The math is complicated, but essentially these methods ask: "how much worse would the model be if this variable was removed?"

2

Forward Feature Selection

1. Build a model with only one variable. Pick the variable that works best by itself.

2. Add that variable to the model. Then try all the remaining variables as the second variable. Add the 2nd best variable to the model.

3. Repeat this process until either (a) adding another variable makes the model worse, or (b) you use all your variables.

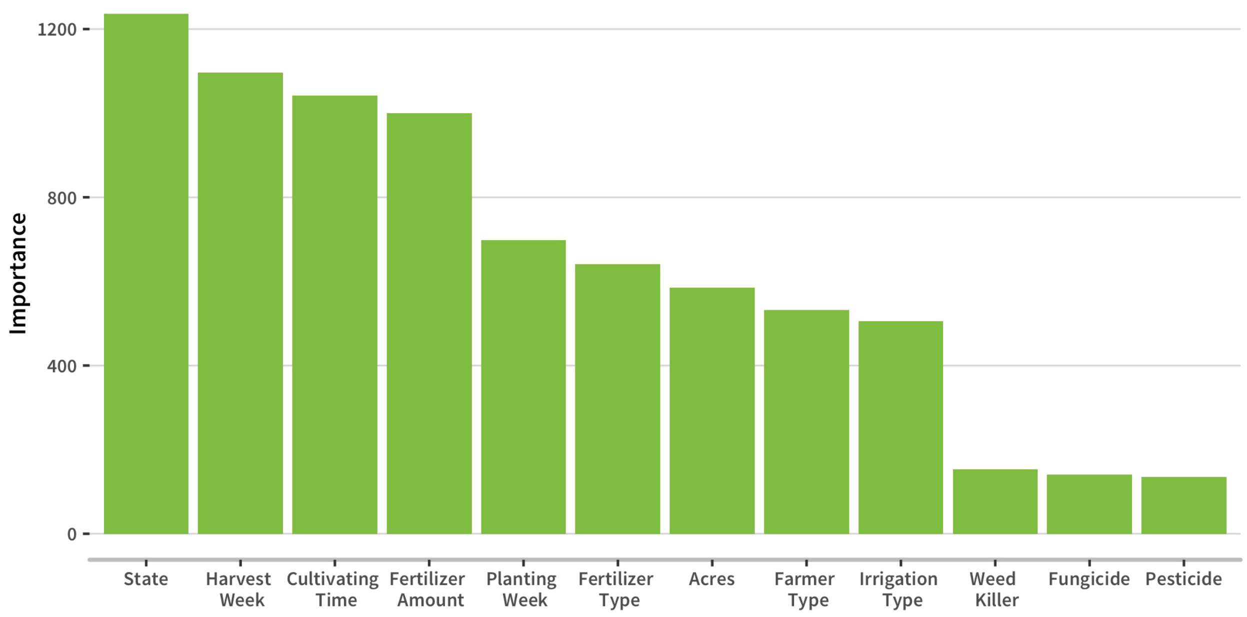
Variable Importance

Location (state) is the most important variable in all frameworks, explaining about 15% of variation.

We find that when a farmer plants and harvests are crucial determinants of yield, as is the amount of fertilizer. These relationships are difficult to find in a linear model (regression), but the random forest can find the right places to split the variable.

The other inputs appear relatively unimportant.

Variable Importance Random Forest Model

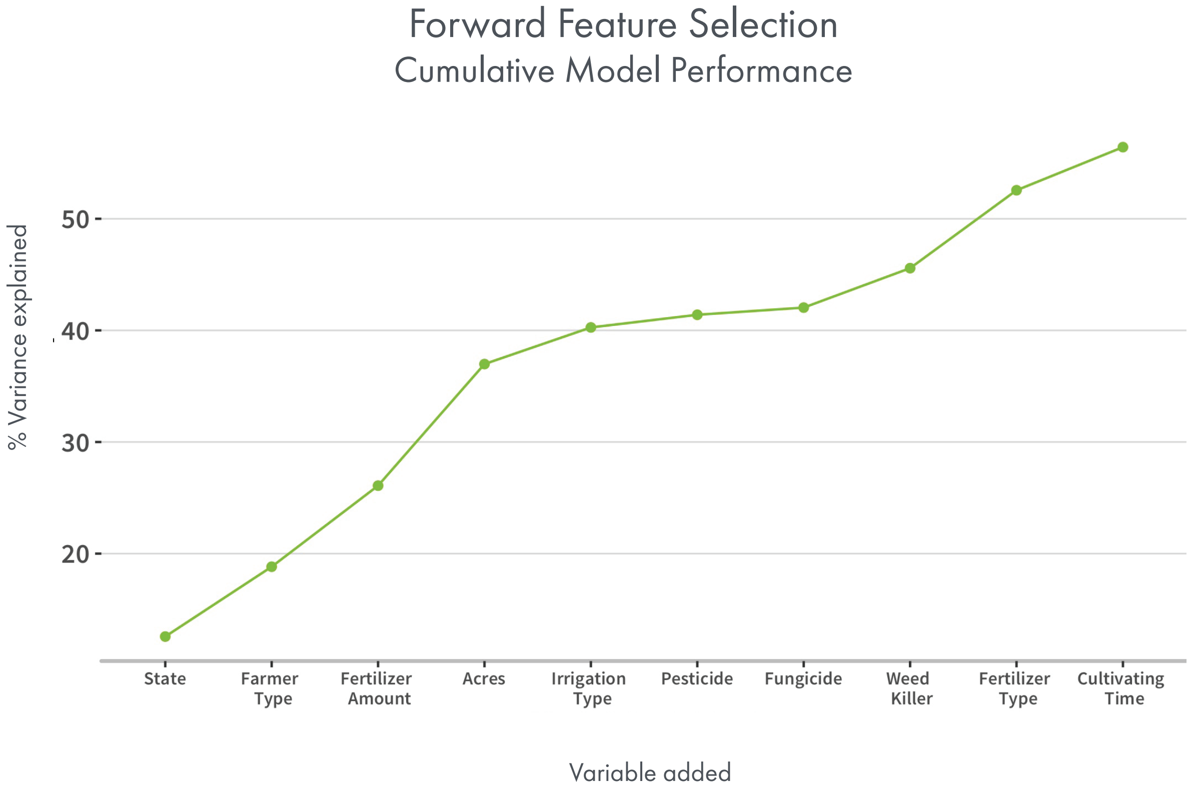


Importance measured in mean increase in node impurity from removing the variable in question

Forward Feature Selection

We use the FFS algorithm to progressively build a model with 10 predictor variables.

State is still the most important, but this method designates farmer characteristics and inputs as the next variables. Timing variables are added later, with planting week not among the 10 most important variables.



Supplementing Survey Data With Weather and Historical Yield

Model	Random forest without inputs	Plus inputs	Plus historical yield estimates	Plus weather
Fit	49%	67.9%	68.0%	76%
Important Variables	Location	Location, planting, harvest timing, fertilizer amount	Planting and harvest timing, fertilizer type and amount	Rainfall, fertilizer amount and type, harvest timing
Notes	Model does not include any information that would not be known prior to planting	Model uses the entire harvest survey	Includes data on the average rice yield in each State from public sources for the past five years	Includes data on temperature and humidity across the cultivation period for each farmer

Adding individual farmer decisions to the model greatly enhances the predictive power, suggesting that there is value to conducting mid-season or planting-time surveys to accurately estimate risk.

The model would have benefited from having data on farmer yield in previous seasons. Since those data are not available, we utilize publicly available data on local yield averages as a proxy. However, this does not add much value to the model. This likely because, while yields fluctuate from year to year, likely due to weather patterns, the relationships between States remain largely constant over time.

Kano and Taraba, for example, had very similar average yields in 2015, at 3.0 MT/ha and 3.1 MT/ha, respectively.

The following year, both suffered a large fall in average yield, but the drop was almost exactly the same in both states, to 2.1 MT/ha and 2.0 MT/ha, respectively. This example illustrates the general principle that yields move in tandem, and why we should not expect gathering more historical data to give us more predictive power. The value of historical yield data collected at the individual level may be much greater, but we did not have access to this data.

Unsurprisingly, microdata on weather in a particular season adds a great deal of predictive power. In the final model, aggregate rainfall between planting and harvest is by far the most influential variable. We find that even seemingly inconsequential weather characteristics are more predictive than historical yield; the lowest humidity reading on the day of planting is more than 3 times as important as the average statewide yield in the previous year (2017), for example.

FINDINGS FOR BEST PRACTICES

BACKGROUND

Pula also provides agricultural advice to its insured farmers. With the use of available data, Busara sought to identify agricultural practice patterns.

Approach: Busara combined data summaries with regression analysis to understand practices and farmer attributes that are associated with yield.

Regression models

easy to interpret and allow for inference (determining statistical significance)

Evidence on Best Practices

Results at a glance:

Fungicide is associated with **0.39 MT/ha, or 9%,** lower yield. This is due to the fact that farmers use it reactively rather than for prevention.

NIRSAL farmers overperform when inputs are taken into account, while Afex and CGAP farmers fall behind.

Farm size plays less of an important role once inputs and practices are taken into account.

Farmers who receive free fertilizer from government programs or NGOs perform lower than their peers.

Implications:

Fungicide purchases could be used as an early warning indicator.

Since we cannot infer causality from survey data, experimental research is needed to follow up effects of different fertilizer sources and farmer group programs.

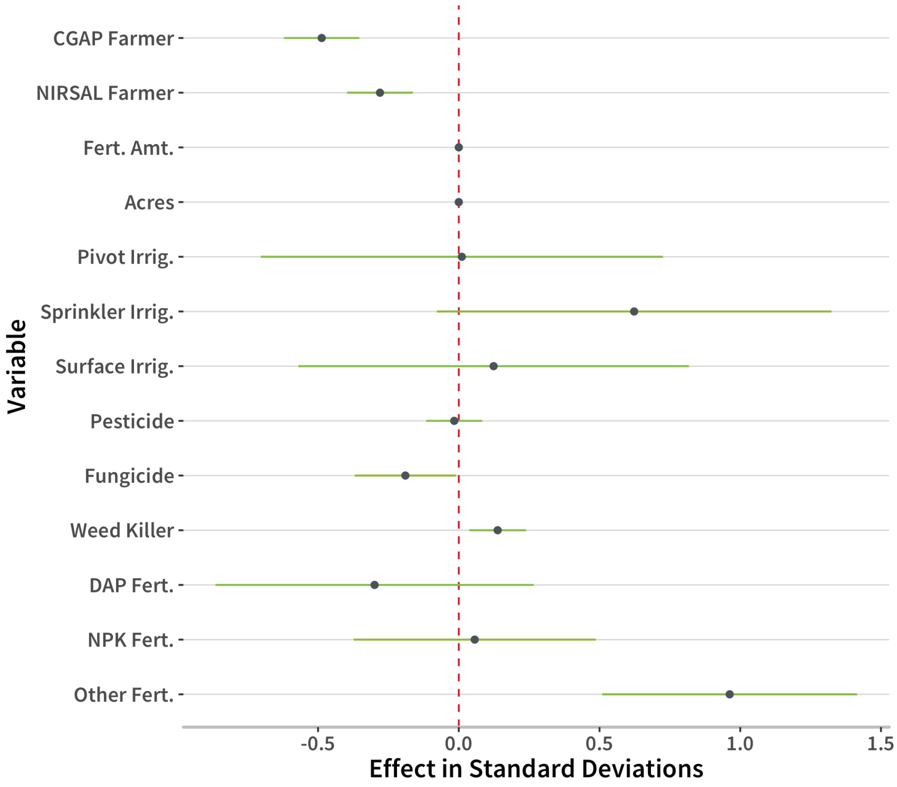
Regression Results

Fungicide is associated with lower yield due to the fact that farmers use it reactively rather than for prevention. Fungicide purchases could be used as an early warning indicator.

The effect of plot size is a "precise zero" once other variables are taken into account.

Farmer groups differ, before and as well as after controlling for input decisions.

Effects Of Inputs On Yield



State and Timing Variables Excluded.Reference categories for farmer type, irrigation type and fertilier type are AFEX, Other and None, respectively.

States in Nigeria are different

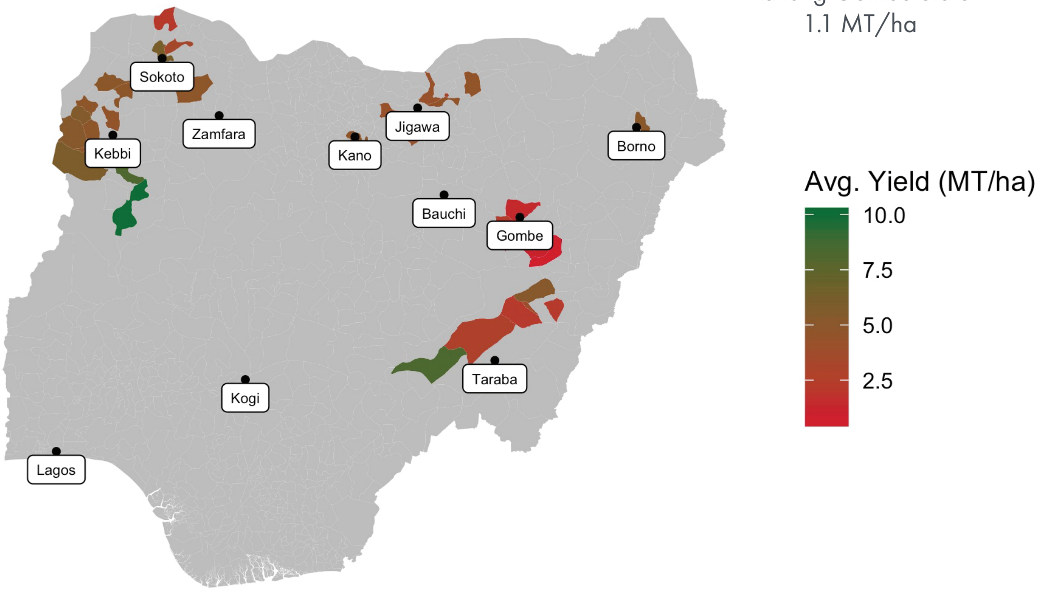
Yield is highly variable by lcoation which is why rice yields can differ by a factor of 2-3 across areas in similar climatic zones. The State alone explains 16% of variation.

In 2018, LGAs in Gombe State saw the lowest average yields of any surveyed area, despite normal rainfall. This is likely due to a plague of Quelea birds, which 44% of surveyed farmers in the State reported as a problem in 2018.

Average Yield By Surveyed District Rice Farmers, 2018

Ngaski Kebbi State
10.7 MT/ha

Kaltung Gombe State
1.1 MT/ha



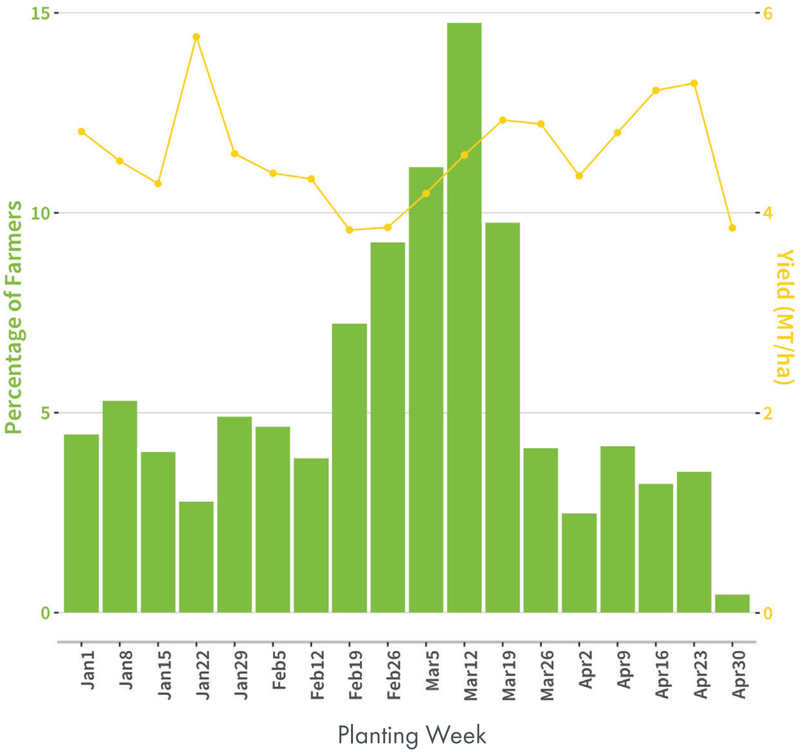
Gray indicates no data available. Pula yield survey conducted in 54 LGAs

When to plant?

Most rice planting is concentrated between mid-February and the end of March. However, a significant number of farmers plant as early as the first week of the year.

On average, farmers who keep rice in the ground longer have worse yields. For each additional 10 days of cultivating time, yield decreases by 0.4%.

Time Trend In Planting And Yield
Rice Farmers, 2018



Irrigation informs when farmers plant

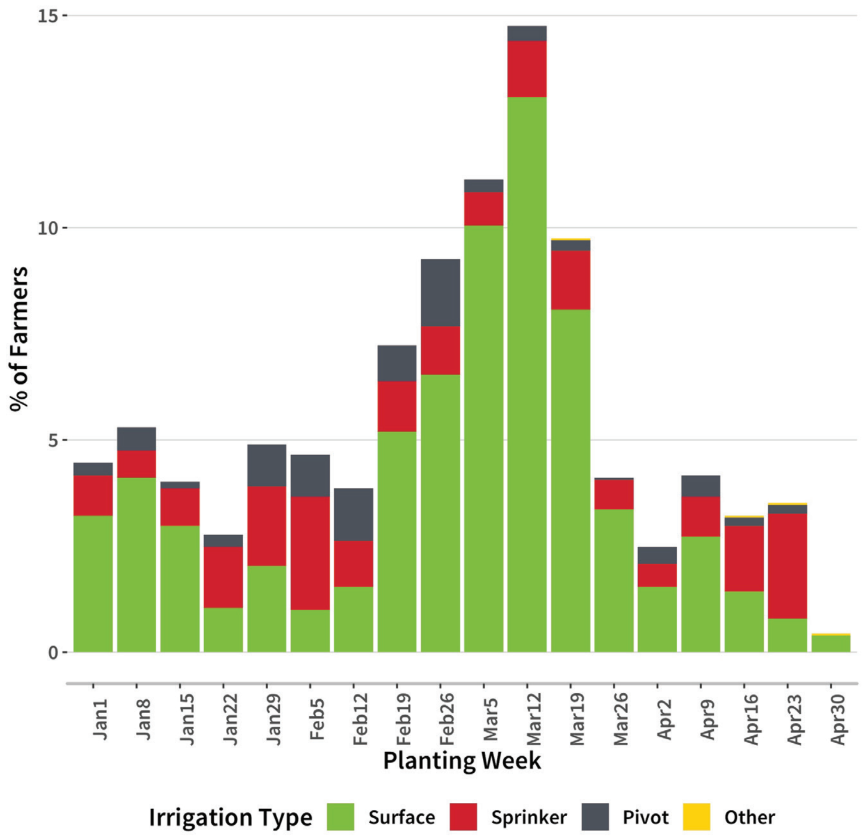
- 70% of farmers (1552) from this sample use surface irrigation and tend to plant closest to the rains.

- Farmers with sprinkler systems (466) plant evenly across the season.

- On average, 195 farmers who have pivot systems plant 10 days earlier than the
- rest**.

**The remaining 6 farmers in the analysis sample had an irrigation system described as "other"

Time Trend In Planting And Yield
Rice Farmers, 2018



Irrigation Access Affects Input Choices

The type of irrigation is related to many other decisions farmers make. For one, farmers with sprinkler irrigation systems are the most likely to be AFEX-affiliated (49% of sprinkler farmers are AFEX affiliates). Put another way: of CGAP farmers (who can be thought of as the general population), 10% have pivot systems, 9 % sprinkler and the remaining 81% surface. Of the AFEX farmers, 39% have sprinkler systems.

Input choices, unsurprisingly, also vary significantly across irrigation system. For example, 76% of pivot farmers use pesticide, compared to 53% of sprinkler irrigated plots and 33% of surface irrigation farmers. These relationships are reversed, however, when it comes to using weed killer.

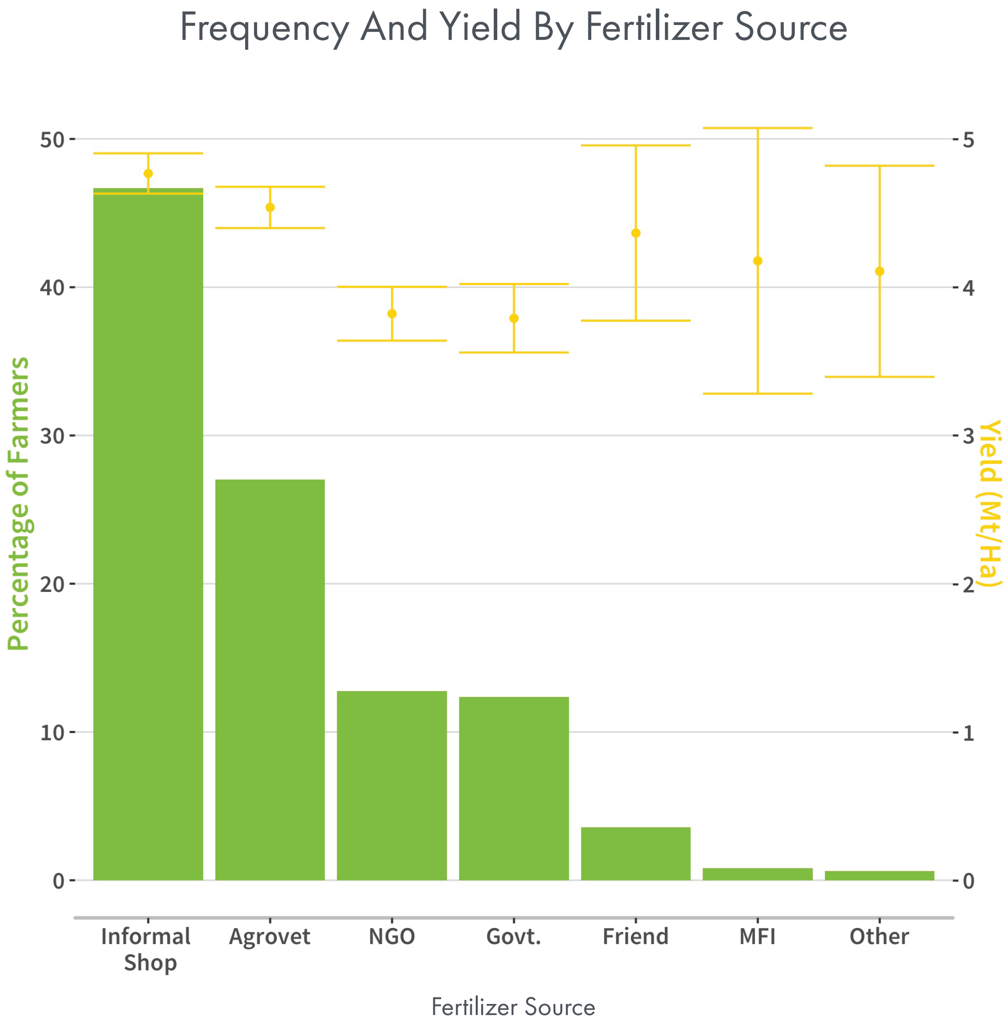
Sprinkler irrigation farms have the highest yield (5.5MT/ha vs approximately 4.3 MT/ha for both other types), but these farmers also use much more fertilizer (311kg per acre on average, vs 216kg for those with pivot systems and 182kg for farmers relying on surface irrigation). Plot size is similar for all types.

Overall, these findings paint a picture of sprinkler and pivot farmers as more sophisticated in their input decisions and market affiliations than the general population.

Not all fertilizer sources are equal

Farmers who rely on NGO or government programs, such as ABP, have significantly lower average yields. This may be due to custom preferences/broad options that are better accessed with self-purchases in informal shops. Further exploration on costing models could reveal price sensitivity of farmers and possible opportunities to invest in farming.

Farmers who received NGO fertilizer planted the earliest on average, followed by government programs. On average, these two groups planted 10 days earlier than farmers who got fertilizer from other sources. These relationships are statistically significant and remain so when controlling for location.



Farmer groups

All surveyed farmers are designated as affiliated with one of Pula's Nigeria partner organizations : AFEX, NIRSAL or CGAP.

Farmers who participate in NIRSAL (the Nigerian Incentive-based Risk-sharing System for Agricultural Lending) receive inputs from the government, which is known to deliver the inputs late. However, we find that in 2018 NIRSAL farmers actually planted significantly earlier (average planting date was February 28, relative to March 9 for AFEX farmers and March 17 in the CGAP pool) and use more fertilizer per acre than either of the other groups.

Affiliates of AFEX, an agricultural commodities exchange and exporting company, can rely on improved inputs on time, which is the most probable reason for their better yield averages despite similar average plot size and other characteristics.

CGAP farmers are actually randomly selected neighbors of farmers in the other two programs, who were surveyed for research purposes. Therefore, the CGAP group can be interpreted as representative of the general population in the sample areas.

Farming group effectiveness

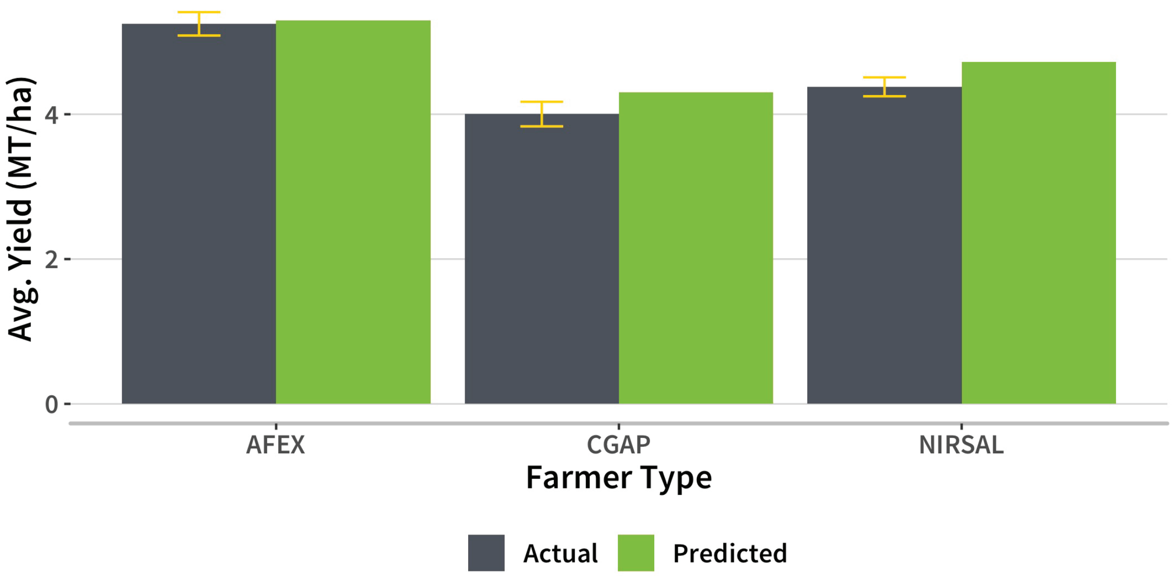
All surveyed farmers are customers of one of Pula's agribusiness partners: AFEX, NIRSAL or CGAP.

Differences in productivity between these groups persist after, size, location and input choices, are taken into account.

Once these factors are taken into account, NIRSAL and CGAP farmers underperform their predicted yields, suggesting that those groups are using suboptimal practices not captured by the harvest survey.

These results also suggest predictive modeling performs better for AFEX farmers than for other groups.

Performance By Farmer Group
Before And After Controlling For Inputs And Practices



Adjusted means reflect controlling for state, farm size, crop timing, fertilizer, irrigation and chemicals.

LEARNING SUMMARY

RECOMMENDATIONS FOR SURVEY METHODOLOGY

To improve the speed of learning from data, implement better checks for data validation.

For questions regarding crop failure and other negative outcomes, split the question into two: one multiple choice giving category of response, and another for more detailed feedback.

Build checks for inconsistency into survey programming, to prevent data points which are difficult to interpret (e.g. crop failure question is left blank in a variety of circumstances and it is unclear why).

To improve precision of the model, measure variables known to better predict yield: gender, age, years of farming experience and last season's sale price.

Combine the questionnaires (for the box placement, wet and dry harvest) to better consolidate repeated information otherwise filled by field offices.

Engage Pula partners to gather historical yield data on farmers in their network, to allow for better predictive modeling.

RECOMMENDATIONS FOR FARMER BEST PRACTICES

Tracking of fungicide purchases as an early warning indicator for fungus infections.

Explore the impact of different sources of fertilizers on farmer investment through quantitative research methods.

Explore the different partner farmer engagement models through quantitative research to establish replicable and unique insights that drive yield.

Creating awareness and reminder systems to discourage lengthy grounding of crops among farmers (this could be channeled through partner groups).

Develop systems for using information from crop cut surveys to connect farmers to extension services, such as input providers or lenders.

Consider partnering with local or national governments to target resources to areas predicted to have high incidence of crop failure.

mastercard foundation

MERCY CORPS | AGRIFIN

THANK YOU

Daniel Mellow
Data Specialist, Busara Center
daniel.mellow@busaracenter.org

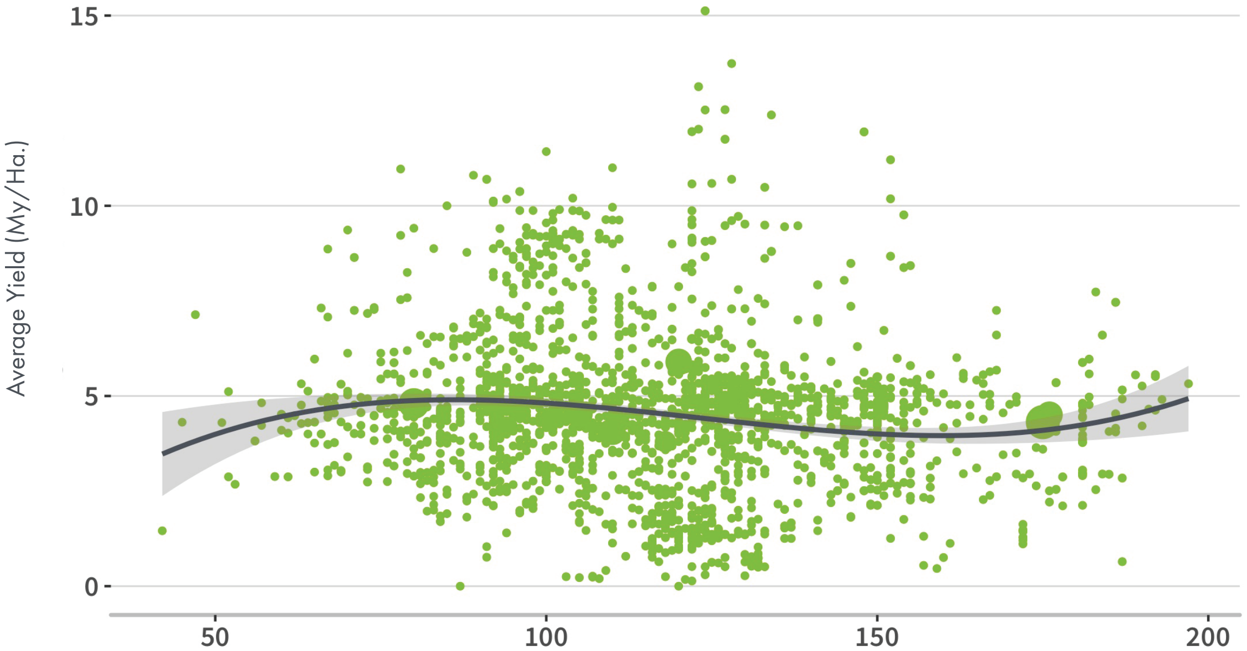
ANNEX

Yield is related to maturity period

Yield increases in maturity up to 75 days, then declines between 75 and 160 days, with a slight upward bump in the outliers at 160-200 days.

According to this information, it appears the optimal maturity is approximately 75 days, though this might be more useful as a recommendation to farmers if broken down by State and irrigation type.

Interestingly, the highest-yielding individual farmers all have maturity dates of approximately 125 days, though the average at that time frame is lower.

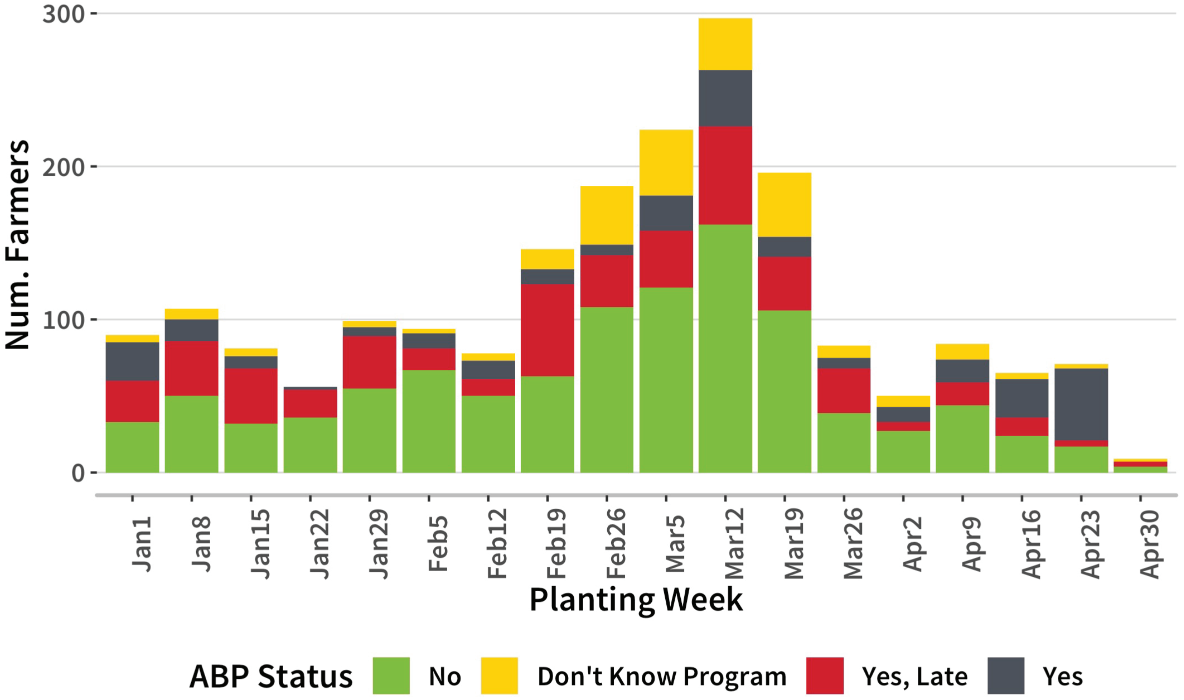


Maturity Period
(Size of point corresponds to size of plot)

ABP Fertilizer and Planting Time

Farmers who received ABP subsidized fertilizer on time actually planted about 15 days later on average than those who received the fertilizer, but "not on time." This relationship holds even after controlling for location and type of irrigation.

Does the ABP program influence planting time?


Num. Farmers
Jan1
Jan8
Jan15
Jan22
Jan29
Feb5
Feb12
Feb19
Feb26
Mar5
Mar12
Mar19
Mar26
Apr2
Apr9
Apr16
Apr23
Apr30
Planting Week
ABP Status
No
Don't Know Program
Yes, Late
Yes