**Review version**

Do not cite or circulate

**Towards personalized agronomy: Causal machine learning and data-driven personalized agronomic recommendations for wheat in Eastern India**

# Abstract

We introduce the concept of “personalized agronomy,” adapted from “personalized medicine,” as a computational modelling framework for policy decision making on where and to whom agronomic innovations should be targeted for maximum agronomic and economic returns across a given target population. We rely on a doubly robust machine learning model and Athey’s policy tree approach to generate individualized recommendations for farmers in a large survey sample, and then illustrate how such estimates can be used to generate tailored recommendations for new farmers in the target geography. We illustrate this computational modelling framework using a case study of wheat farmers in Bihar State, India. The recommended practices include broad-spectrum herbicides, early sowing, repeated irrigation, and using high-yielding long-duration varieties. This work demonstrates how field and farmer heterogeneity can be incorporated into a spatially explicit analytical framework for improved targeting of extension efforts and improved aggregate efficiency of related public and private investments.

# 1. Introduction

Evaluations of agricultural research and development (R&D) interventions have documented very high returns with global average benefit-cost ratio of 10:1, ranging from less than 2:1 to 42:1 (Alston et al 2020). However, the distribution of returns at the farm and community level is highly heterogenous, reflecting both variations in production conditions as well as management practices and other farm-level characteristics (Suri 2011; Suri and Udry 2021). P This heterogeneity in returns has prompted many researchers to recommend prioritization and targeting of interventions as key to unlocking the gains from agricultural R&D. However, the best methods for targeting remain uncertain. Many studies identify few factors they hypothesize matter and estimate the heterogeneous effects of these factors. In Malawi, Haile et al (2017) documented that on-farm agronomic trials were often biased towards better endowed farmers, drawing into question the broader relevance of experimental results. It is important to use observational data to check which agronomic practices are beneficial conditional on farmer characteristics and to match solution spaces to underlying agroecological heterogeneity (Snapp et al., 202x, McDonald et al, in review). In India for example, Krishna et al (2020) documents that zero tillage has high returns but mostly to larger farms therefore requires proper targeting. Other research in India suggests that returns from technology can be doubled if guided by remotely-sensed information (Jain et al., 2019). This heterogeneity can be well studied in a new framework of “personalized agronomy” borrowing on “personalized medicine”—a data-driven approach to making personalized recommendation on agronomic management.

Recent discussions of data-driven decision support in agriculture have suggested data-based geographical zonation using non-parametric approaches like random forest algorithms, classification and regressions trees (CART), and cluster analyses (e.g., Krupnik et al 2015, Urfels et al 2021). In suggested workflows, random forest algorithms are first used to identify variables of importance, the variation of which is then partitioned using regression trees (CART), followed by geographical targeting using k-means clustering. This approach however does not prescribe what the farmers in each cluster need to do to increase yields. In addition, this line of research still does not recommend who should receive the intervention and the welfare criteria for such decisions (i.e., individual level variation of recommendations within geographical recommendation zones). This means that even recommendations which are spatially well-targeted may be suboptimal for individual ricepients.

To address this challenge, we propose a targeting approach based on causal machine learning to define intervention assignment rules based on both farm and farmer characteristics. We use a casual random forest model, which provides a flexible framework for estimating heterogeneous treatment effects and which performs well at generating out-of-sample spatial predictions for clustered data (cite). We combine this approach with the policy tree algorithm (cite) to suggest who would likely benefit from targeted recommendations on agronomic practices. The combined causal random forest and policy tree optimization approach is different from the prediction oriented machine learning approach of using random forest or CART. Wager and Athey (2018) review the theoretical differences between causal random forest and random forest models.

Kakimoto et al (2022) discusses the differences between prediction-oriented machine learning methods (e.g., random forest, convolutional neural networks) and causal random forest in an agronomy related simulated data of N optimal management. They found that causal random forests are advantageous as compared to the prediction based machine learning approaches.

McCullough et al (2022) compared estimates from random forest and causal random forest on the nitrogen effect on maize yields in sub-Saharan Africa. They found that the predicted fertilizer response with a causal random forest model (1.49) was higher than that for a random forest model (1.26) but lower than that of a feasible generalized least squares estimator (2.07).

We depart from Kakimoto et al (2022) on the optimization framework in that instead of using individual comparison of the different categories, we instead using policy tree algorithm (Zhou et al 2022) which also specifies which variables would efficiently allocate farmers to the different categories.

Policy trees are different from CART in several ways. First, CART are for prediction while policy trees are for recommendation development.

The approach we proposed for personalized agronomy has been successful in personalized medicine (Inoue et al 2023) in which the causal machine learning approach to assign patients to treatment was almost five times more effective than the conventional approach used by doctors.

# 2. Materials and methods

## 2.1. Data sources and descriptive statistics

We use two main databases: the Landscape Diagnostic Survey (LDS) data for wheat, and geodata R package soil grids. The landscape diagnostic survey is a landscape level crop assessment survey covering about 8000 households administered in 2017-18 in Bihar and Eastern Uttar Pradesh by the Cereal Systems Initiative for South Asia (CSISA).

Figure 1 shows the count of farmers who use each of the agronomic practices[[1]](#footnote-1). Figure 1a shows the count of farmers using different weed management options. About ##% are not weeding at all.

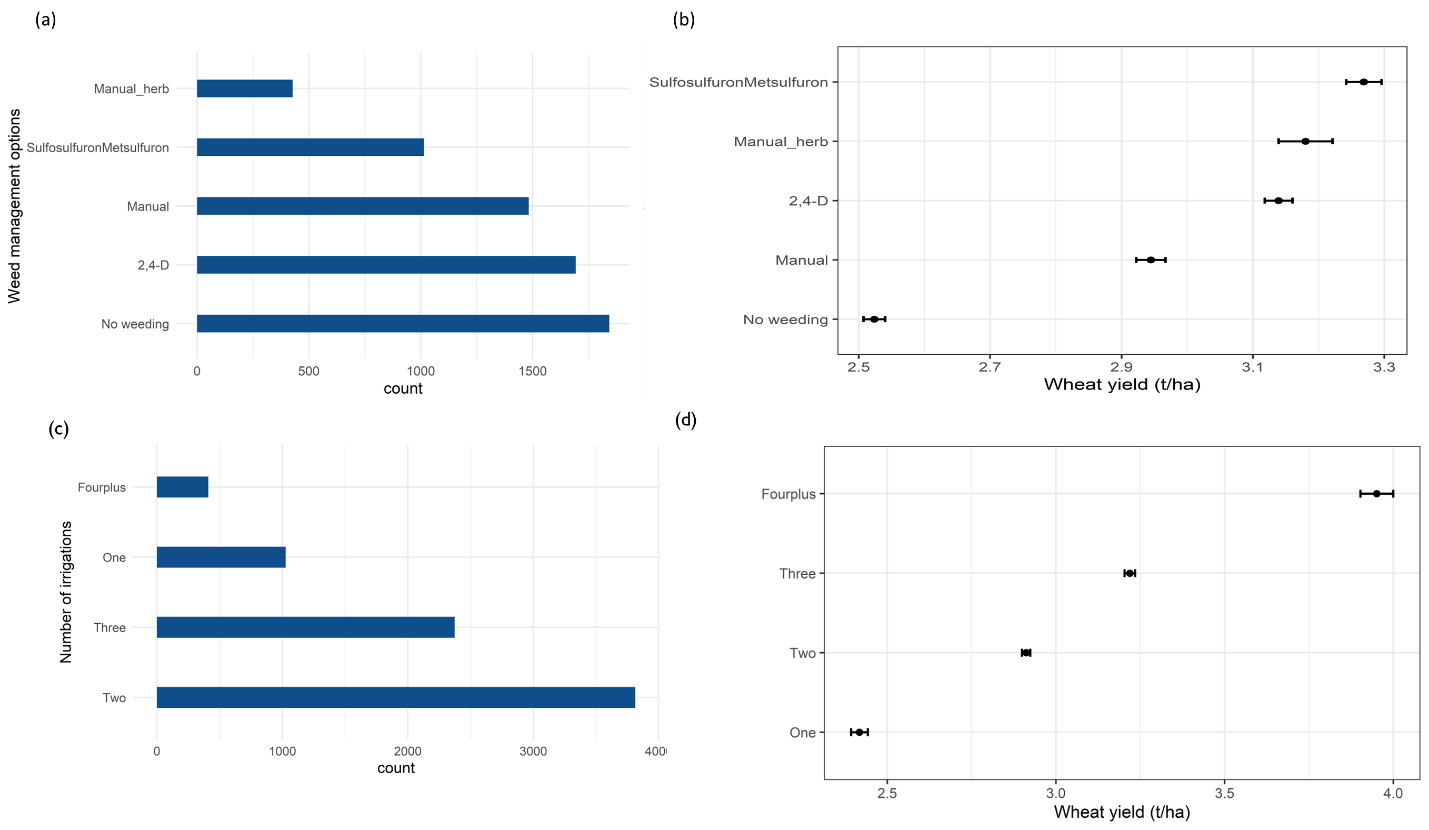


Figure 1: Frequency of farmers and yields by agronomic practices followed.

Note: Panel a and c show the number of farmers adopting each of the practices while panel b and d show the average and confidence interval of wheat yields by the agronomic practices.

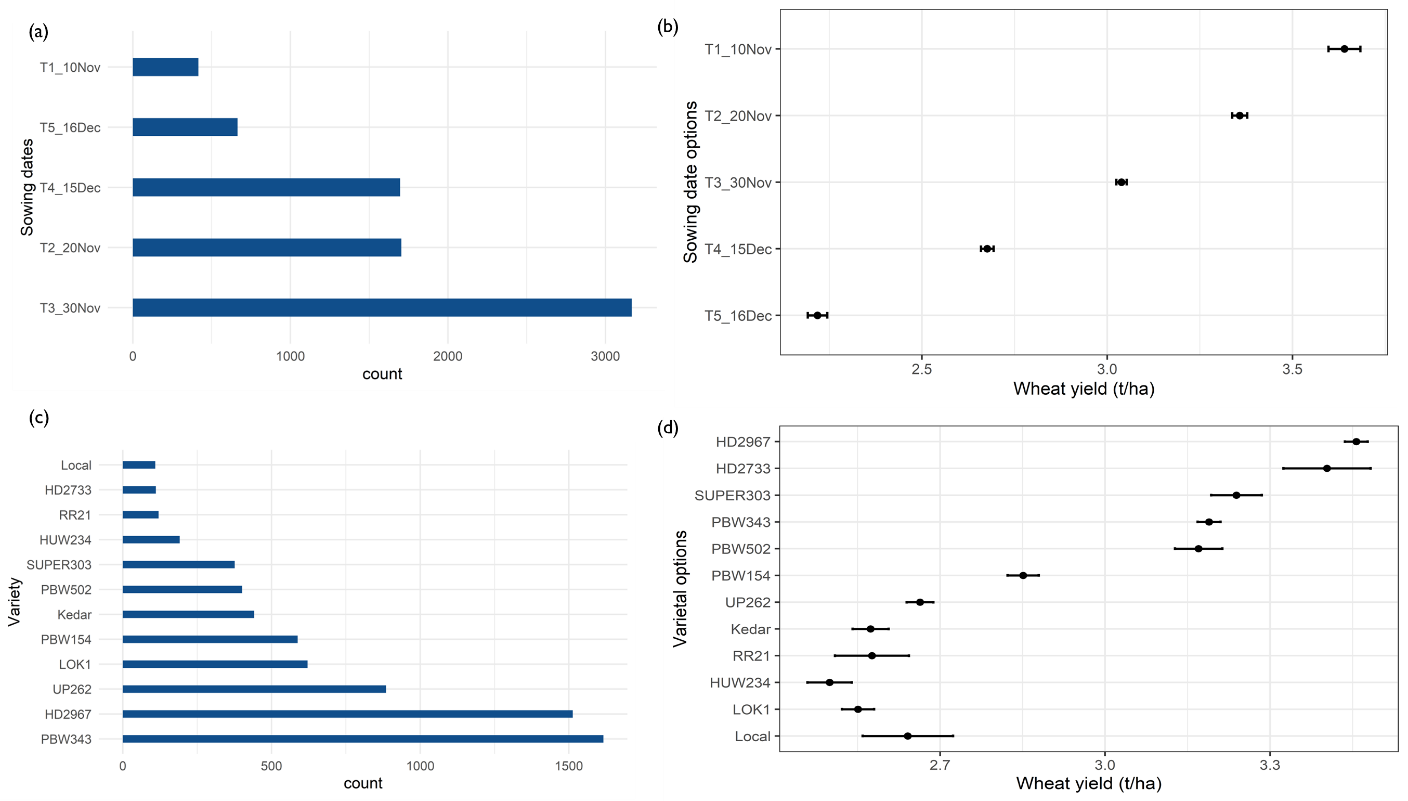


Figure 2: Frequency of farmers and mean yield comparisons across agronomic practices. Note: Panel a and c show the number of farmers adopting each of the practices while panel b and d show the average and confidence interval of wheat yields by the agronomic practices.

Figure 2 shows the mean yield distribution by agronomic management practices. In terms of weed management, Sulfosulfuron+Metsulfuron has the highest unconditional yield followed by 24D (figure 2a). The sowing date comparisons are clear with early sowing giving the highest yields (figure 2b). In terms of irrigation scheduling, adding four or more irrigations gives the highest yields (figure 2c). For varieties, the higher yielders include HD 2967, HD 2733 and SUPER 303 (figure 2d).

## 2.2. Empirical approach

We use a four step analytical process (figure 3) beginning with, (1) estimating a probability random forest model followed by a causal random forest model that uses the predicted probabilities from the first random forest as weights, (2) predicting individualized treatment effects, (3) estimating doubly robust scores, and (4) applying policy tree algorithm to optimally recommend a policy that can be targeted to each farmer. This procedure follows the policy tree algorithm by Zhou et al (2022) and Athey and Wager (2021). Following Athey and Wager (2021), a policy () is a rule or method by which we decide, who, on the basis of their characteristics will be targeted to receive intervention:

Where are characteristics of the farmers including their demographics, farm characteristics and plot characteristics. indicates the interventions to be targeted. It can be a dichotomous decision (e.g., plant early or late) or a discrete decision (e.g., weekly planting schedules). In most agronomic evaluation, the value of the agronomic practice is represented in form of yield gains which can be defined as a conditional average treatment effect (CATE)

Where is expectation term. is the yield (it can also be revenue or profits) with being yield with agronomic practice while is the control.

**Step 1 & 2: [Causal random forest model]—**Estimate a yield crop response model using causal random forest model with augmented inverse weighted propensity score (AIWP).

Unlike random forest models, causal random forests use honest splitting in which splitting is done on training sample and uses independent sample to estimate leaf means. Causal random forests also allow orthogonalization for observational (survey) data using Inverse-Probability Weights predicted from probit model. Augmented Inverse-Probability Weighted Conditional Average Treatment Effect (AIPW-CATE) is given by

Where are treatment assignment probabilities. In all the specifications, we control for all confounders including variety type, date of sowing, fertilizer applied, soil quality variables including soil nitrogen, precipitation, temperature, and farm size.

**Step 3 & 4: Estimate doubly robust scores and policy recommendations**

If our goal is to get the highest yields or profits for the individual farmer, then we can define a utilitarian value of the policy

The type of loss function to use when deciding who should get the treatment on the basis of this value can be through threshold target (e.g., if , where is some threshold which can represent cost of delivery), inverse propensity score weighting loss function (Kitagawa and Tetenov 2018) or a doubly robust value estimator (Athey and Wager 2021).

In observational data settings, doubly robust value estimator (Athey and Wager (2021) has been documented to be effective for generating personalized recommendations. We present next the derivation of policy function based on this method. The goal is to choose a policy that maximizes the value function

Using k-fold cross validation, we can test the value of the targeting procedure by comparing to the current levels or random assignment.

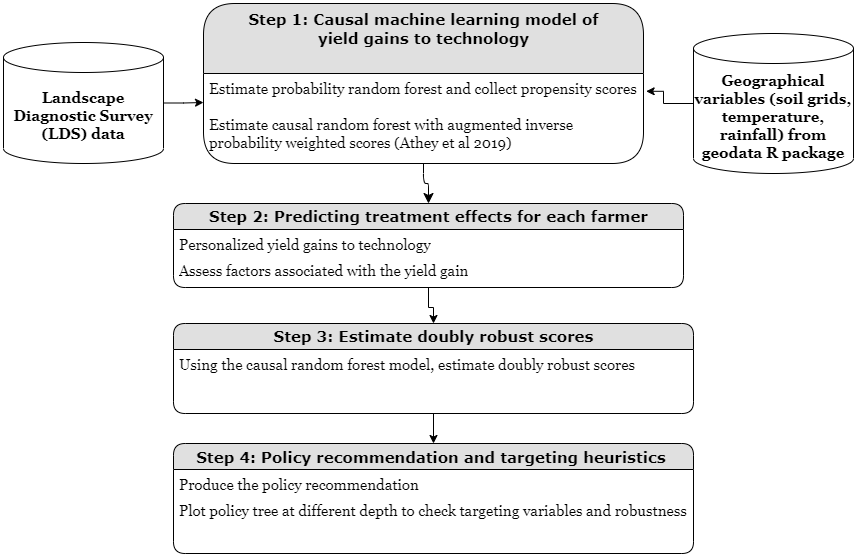


Figure 3: Model workflow

# 3. Results

## 3.1. Heterogeneous treatment effects

Table 1 shows the conditional average treatment effects of various agronomic practices on wheat yields.

Panel (a) shows the yield gains to weed management with Sulfosulfuron+Metsulfuron giving the highest yield gains of about 0.3t/ha over no weeding followed by 24D and manual weeding. The combination of manual weeding and herbicides did not have any effect probably due to lack of power and small sample size.

Panel (b) shows the effects of early sowing with with sowing earlier (1-10 Nov) getting yield gains as high as 0.7t/ha over those sowing after 16th Dec. Panel (c) reports yield gains to additional irrigation as compared to 1 irrigation. We find that an additional irrigation results in about 0.4t yield gains. Panel (d) shows the yield losses to selecting a variety other than HD 2967. Except for Super 303 and HD 2733, which has no significant yield differences to HD 2967, the rest of the varieties examined result in yield losses as compared to HD 2967.

Table 1: Conditional average treatment effects (yield gain as compared to base) using multi-armed bandit causal random forest models

|  |  |  |  |
| --- | --- | --- | --- |
| Practice | Contrasts | Estimate  (t/ha) | Standard error |
| *Panel (a)* |  |  |  |
| Weed management | Manual - No weeding | 0.27 | 0.02 |
| Weed management | Manual+ herbicides –  No weeding | 0.31 | 0.03 |
| Weed management | 24D - No weeding | 0.37 | 0.02 |
| Weed management | Sulfosulfuron+ Metsulfuron –  No weeding | 0.43 | 0.03 |
| *Panel (b)* |  |  |  |
| Sowing dates | 15Dec - 16Dec | 0.24 | 0.03 |
| Sowing dates | 30Nov - 16Dec | 0.43 | 0.02 |
| Sowing dates | 20Nov - 16Dec | 0.58 | 0.03 |
| Sowing dates | 10Nov - 16Dec | 0.71 | 0.05 |
| *Panel (c)* |  |  |  |
| Number of irrigations | Two - one | 0.38 | 0.02 |
| Number of irrigations | Three - one | 0.75 | 0.03 |
| Number of irrigations | Four - one | 1.34 | 0.04 |
| *Panel (d)* | |  |  |  |
| Variety choice | Local - HD2967 | -0.48 | 0.03 |
| Variety choice | PBW343 - HD2967 | -0.21 | 0.03 |
| Variety choice | UP262 - HD2967 | -0.51 | 0.03 |
| Variety choice | LOK1 - HD2967 | -0.56 | 0.03 |
| Variety choice | PBW154 - HD2967 | -0.45 | 0.03 |
| Variety choice | Kedar - HD2967 | -0.49 | 0.04 |
| Variety choice | SUPER303 - HD2967 | -0.05 | 0.03 |
| Variety choice | PBW502 - HD2967 | -0.19 | 0.03 |
| Variety choice | HD2733 - HD2967 | 0.03 | 0.18 |
| Variety choice | HUW234 - HD2967 | -0.59 | 0.04 |
| Variety choice | RR21 - HD2967 | -0.41 | 0.05 |

These average returns are also consistent for distributional comparisons as shown by figure 4. These distributions help to demonstrate the heterogeneity in the returns. For varieties, the only varieties that are match the yield levels of HD 2967 for all farmers are long duration varieties (timely sown varieties) like SUPER 303, HD 2733, PBW 502 and PBW 343.

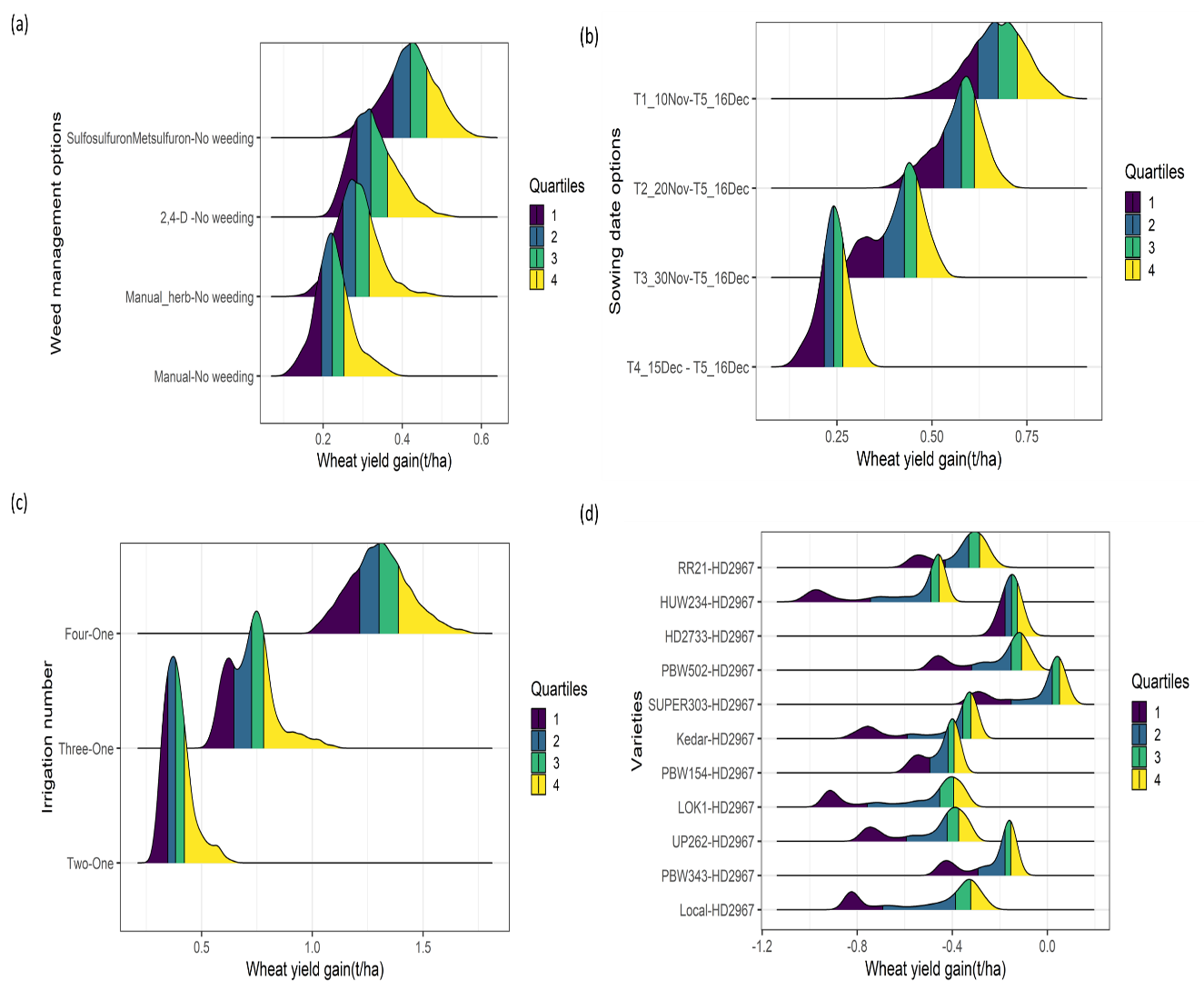


Figure 4: Treatment effect heterogeneity

## 3.2. How many farmers will optimally benefit from the different agronomic management options?

The density functions in figure 4 are estimates of how many farmers would find it beneficial to use each of the candidate agronomic recommendations. Figure 5 shows the proportion of farmers who can be assigned to each of the interventions using policy tree optimization. For weed management, figure 5 panel (a) shows that the optimal transitions for yield gains are to apply broad spectrum herbicides especially Sulfosulfuron+Metsulfuron (55%) followed by 24-D (32%) and combined manual and herbicides (13%).

In the case of sowing dates, farmers are advised to sow early, i.e., between 1-10 Nov for 73% of the farmers and between 11-20th Nov for 27% of the farmers.

Perhaps the most radical shift is recommended for irrigation management with farmers having to shift from 3 or less irrigations to four irrigations (98%).

We find that on the basis of yield maximization, three varieties can be recommended, viz. HD 2967 (52%), Shri Ram SUPER 303 (43%) AND HD 2733 (5%).

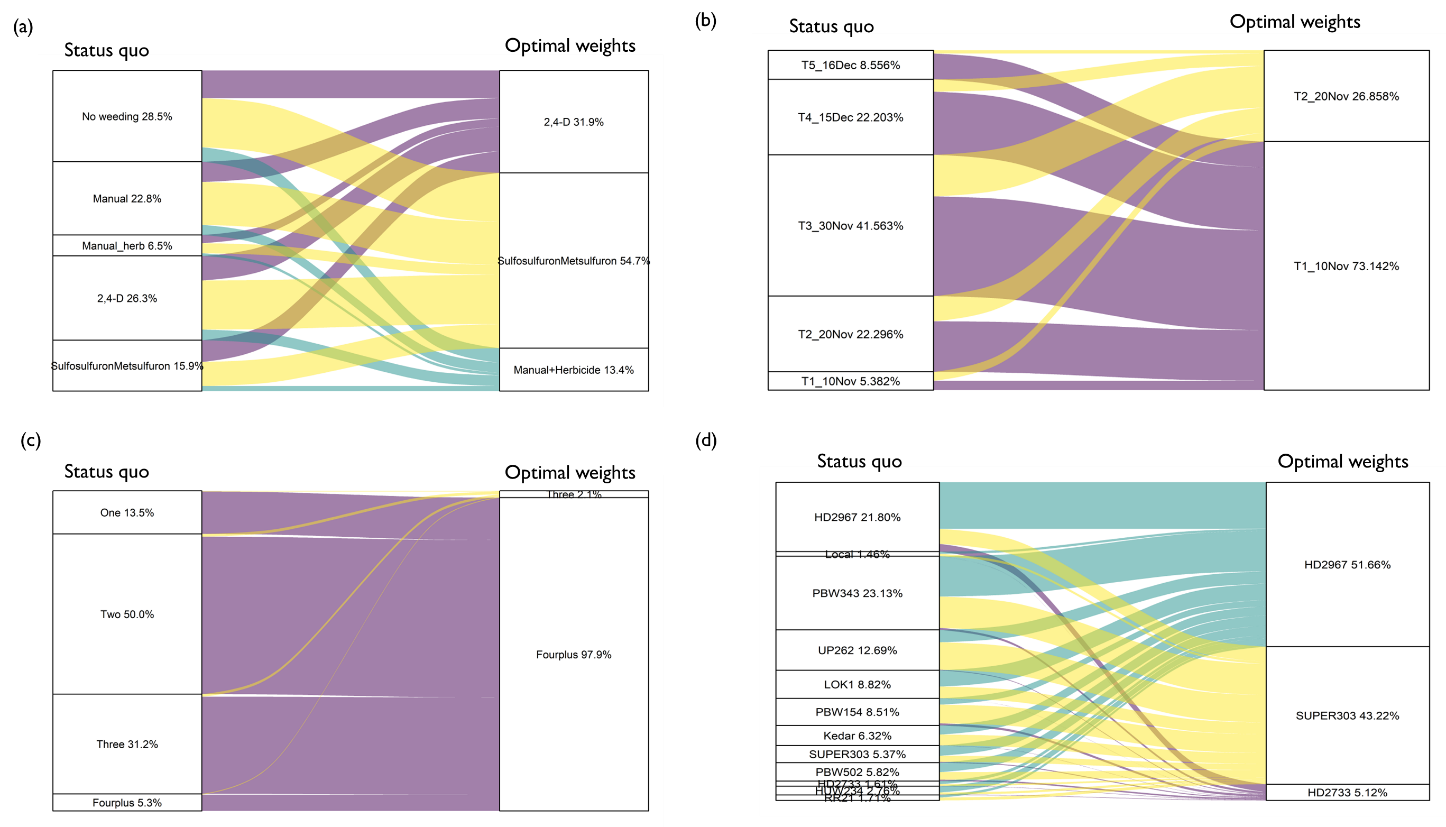


Figure 5: Optimal transition from status quo practices to optimal practices

## 3.3. How and where to target farmers to receive a recommendation?

Figure 6 panel (a) shows the policy tree for weed management options. The policy tree shows that weed management options can be targeted to farmers on the basis of their plot size, market distance, and phosphorus application.

Consistent with prior studies, all farmers are recommended to plant before 20th November with about 73% recommended for 1-10th Nov and 27% for 11th to 20th Nov (Figure 5). This recommendation can be made on the basis of latitude (and longitude) as can also be seen in Figure 6b.

Conventionally, farmers apply at least two irrigations to the wheat crop in Bihar. Growing evidence suggests that three or four irrigations would provide higher yield returns. The question remains as to where and to who should these additional irrigations be targeted. Using the policy tree based algorithm, we find that 93% of the farmers would benefit more from applying four irrigations. Figure 6c shows the policy tree for this recommendation. The key variables for delineating the groups of farmers include longitude, sand, and elevation.

The supposedly simple task of choosing a wheat variety to plant can be complicated due to variety choice overload with dozens of varieties on the market. To aid in the choosing of a variety that gives the maximal yield gains, we apply the policy tree algorithm on 12 wheat varieties that are planted by over 90% of the farmers. These varieties can be targeted on the basis longitude, and latitude.

Figure 6 shows policy trees (depth 2) for the different agronomic management practices.

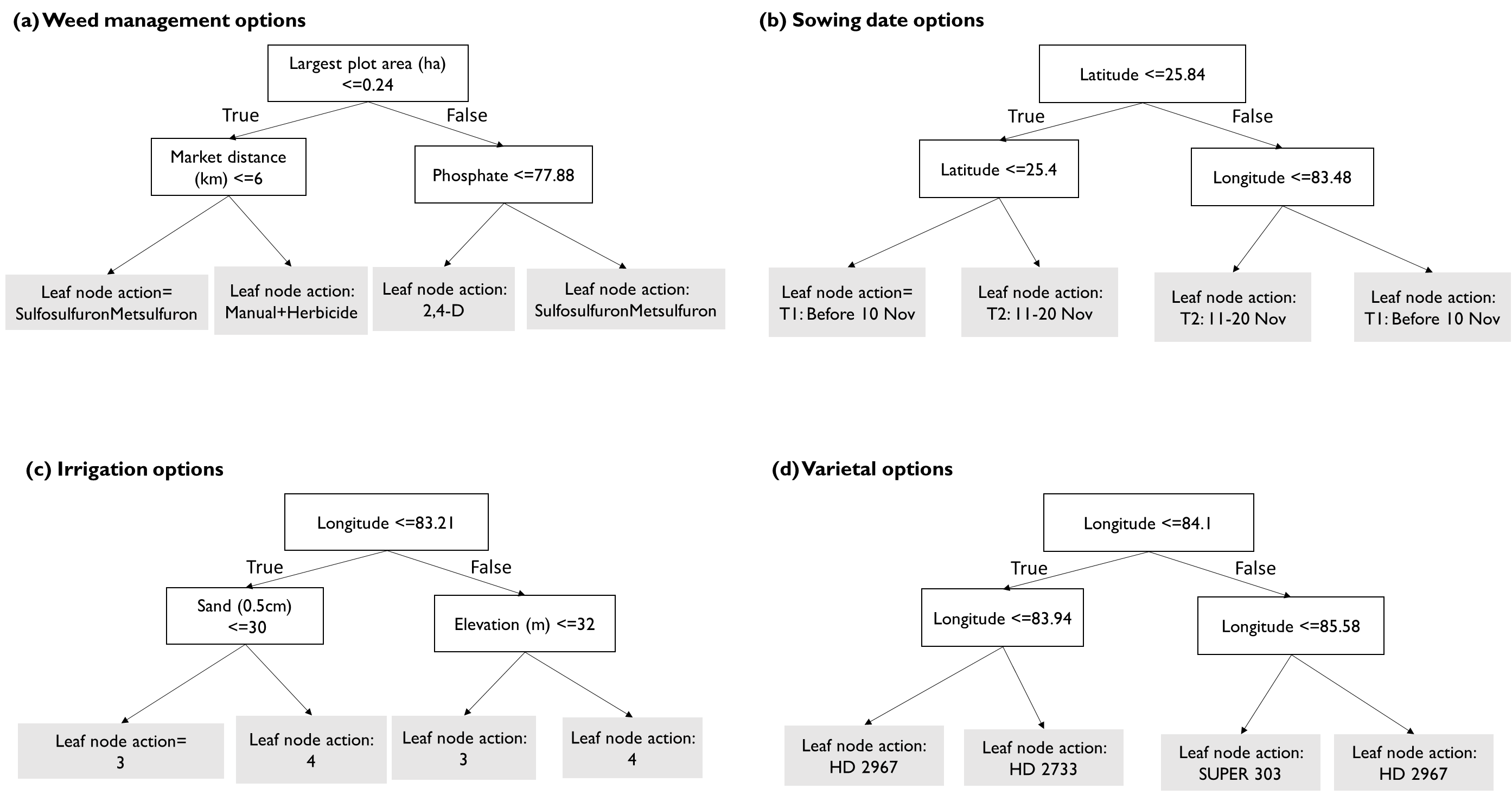


Figure 6: Depth 2 policy trees for decision support on agronomic practices

With this algorithm, we are able to assign each of the farmers to particular agronomic practices. We can therefore produce a spatial representation of these farmers. Figure 7 shows maps of recommended practices for each of the farmers. We find that weed management has more spatially dispersed recommended practices while the other practices exhibit spatial clustering.

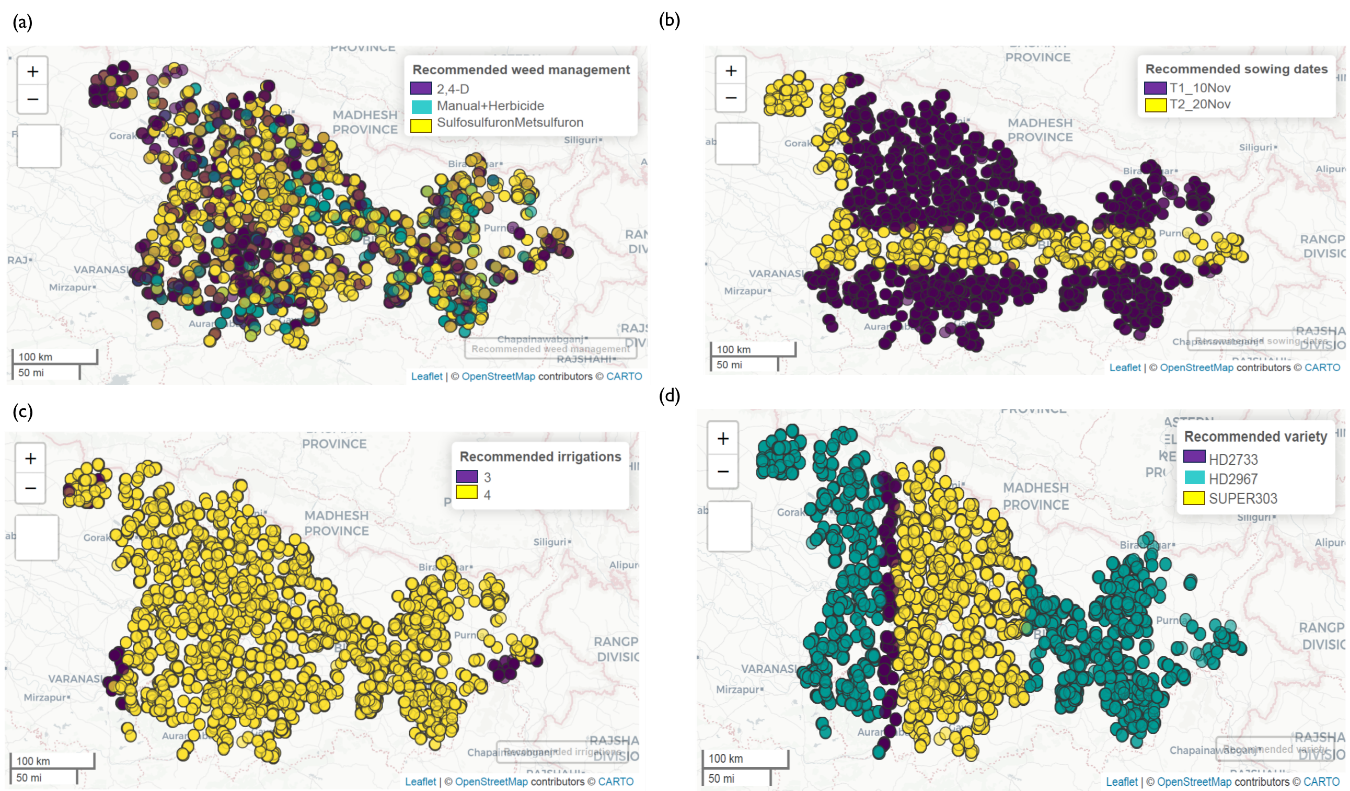


Figure 7: Spatial distribution of recommended agronomic practices

# 4. Discussion

## 3.1. Economic assessment example: Irrigation cost

The results so far are based on yield benefits of the technology options. The policy tree algorithm however allows one to embed a cost function or any other complex functions of the yield gains. For cash based inputs with differential costs across technology options.

Even after doubling the cost of the fourth irrigation, almost 49% are recommended to apply four irrigations and 51% to apply three irrigations (Figure 7). The evidence seems to be clear that farmers would gain from either three or four irrigations and that the cost of the irrigation seems to matter less in this decision because the revenue gains are so enormous.

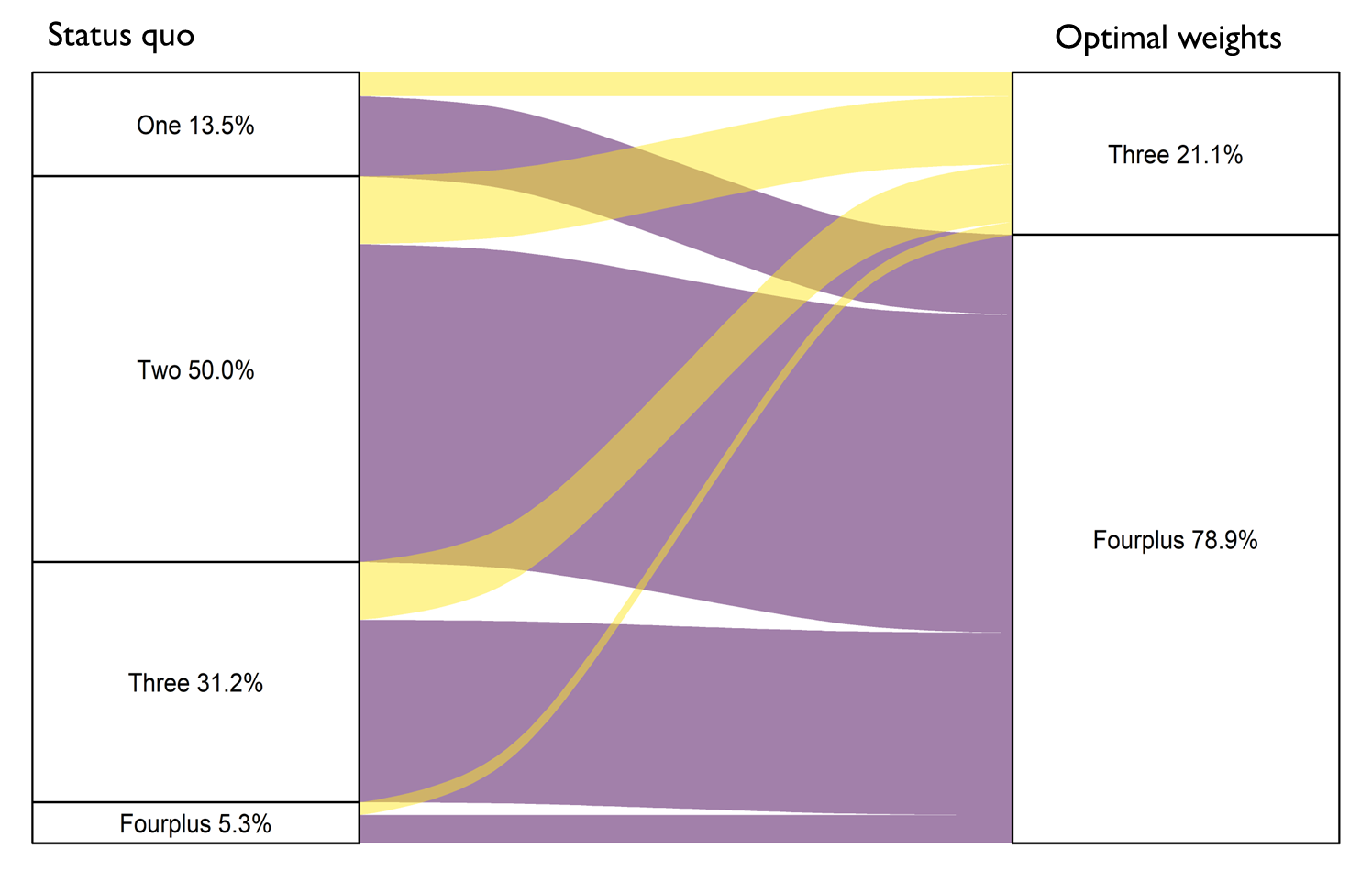


Figure 8: Targeting irrigation assuming costs (note: Revenue gain: Rs. 15/kg x Yield gain, irrigation cost: Rs. 120/hr x 10 hrs=1200/ha for 1 irrigation (80kg), Rs. 2400 for 2 irrigations (160kg), Rs. 3600 for 3 irrigations (240kg wheat equiv.), and Rs. 9600 for 4 irrigations (640kg wheat equiv.).

## 3.2. The value of data-driven personalized recommendations

On sowing date optimization, prior analytics on using machine learning for understanding divers of timely planting of wheat has focused on how to advance rice planting to allow early harvesting of rice (e.g., Urfels et al 2021 and McDonald et al 2022). In this paper, we build on that evidence to understand where a recommendation to plant early would be prudent. Recent evidence has shown that sowing before 21st November is ideal for achieving good yield returns. We considered five sowing date schedules to recommend the optimal sowing window. These sowing date options include: (1) 1-10th Nov, (2) 11-20th Nov, (3) 21st -30th Nov, (4) 1-15th Dec, and (5) After 16th Dec.

## 3.3 Generation of recommendations in real world settings…

# 5. Conclusion

The paper has introduced the concept of “personalized agronomy” made possible by causal machine learning models for estimating individual treatment effects and policy tree models for generating optimal recommendations. These models have been effective in “personalized medicine” and we conjecture agricultural scientists would also use these models in making personalized crop management recommendations. In agricultural context, a recent study by Athey et al (2023) has piloted the use of causal machine learning based recommender systems and have found them to be better than conventional approaches. One challenge of this approach is that each agronomic practice is analysed separately. What combinations of these are likely to be the most profitable for each farmer? For our application, this seems not to be an issue as the recommended practices can be implemented together but this may not be the case in other applications.

other

# Acknowledgements

This methods report was prepared with support from the Bill and Melinda Gates Foundation through the Cereal Systems Initiative for South Asia (CSISA) and the Excellence in Agronomy Initiative. We thank all the CSISA field teams who supported the collection of the Landscape Diagnostic Survey (LDS). We also thank the farmers who provided their precious time to respond to survey questions. The content of this report solely reflects the opinions and findings of the authors.

# References

Alston, J.M., Pardey, P.G., and Rao, X. 2022. “Payoffs to a half century of CGIAR research.” *American Journal of Agricultural Economics* 104 (2): 502-529. Doi: <https://doi.org/10.1111/ajae.12255>.

Athey, S., and Wager, S. 2021. “Policy learning with observational data.” *Econometrica* 89(1): 133-161. Doi: <https://doi.org/10.3982/ECTA15732>.

Haile, B., Azzarri, C., Roberts, C., Spielman, D.J. 2017. “Targeting, bias, and expected impact of complex innovations on developing-country agriculture: evidence from Malawi.” *Agricultural Economics* 48 (3): 317-326. Doi:  <https://doi.org/10.1111/agec.12336>.

Inoue, K., Athey, S., and Tsugawa, Y. 2023. “Machine-learning-based high-benefit approach versus conventional high-risk approach in blood pressure management.” *International Journal of Epidemiology* 1-14. Doi: <https://doi.org/10.1093/ije/dyad037>.

Kakimoto, S., Mieno, T., Tanaka, T.S.T., and Bullock, D.S. 2022. “Causal forest approach for site-specific input management via on-farm precision experimentation”. *Computers and Electronics in Agriculture* 199: 107164. Doi: <https://doi.org/10.1016/j.compag.2022.107164>.

Kitagawa, T., and Tetenov, A. 2018. “Who should be treated? Empirical welfare maximization methods for treatment choice.” *Econometrica* 86 (2): 591-616. Doi: <https://doi.org/10.3982/ECTA13288>.

Krishna, V.V., Keil, A., Jain, M., Zhou, W., Jose, M., Serendran-Padmaja, S., Barba-Escoto, L., Balwinder-Singh, Jat, M.L., Erenstein, O. 2022. “Conservation Agriculture Benefits Indian Farmers, but Technology Targeting Needed for Greater Impacts.” *Frontiers in Agronomy*. Doi:10.3389/fagro.2022.772732.

Krupnik, T.J., Ahmed, Z.U., Timsina, J., Yasmin, S., Hossain, F., Al Mamun, A., Mridha, A.I., and McDonald, A.J. 2015. “Untangling crop management and environmental influences on wheat yield variability in Bangladesh: An application of non-parametric approaches.” *Agricultural Systems* 139: 166-179. Doi: <https://doi.org/10.1016/j.agsy.2015.05.007>.

McCullough, E.B., Quinn, J.D., Simons, A.M. 2022. “Profitability of climate-smart soil fertility investment varies widely across sub-Saharan Africa”. Nature Food

Suri, T. and Udry, C., 2022. Agricultural technology in Africa. Journal of Economic Perspectives, 36(1), pp.33-56.

Suri, T., 2011. Selection and comparative advantage in technology adoption. Econometrica, 79(1), pp.159-209.

Stokes, M.E., and McCourt, P. 2014. “Towards personalized agriculture: what chemical genomics can bring to plant biotechnology.” *Frontiers in Plant Science* 5. Doi: <https://doi.org/10.3389/fpls.2014.00344>.

Urfels, A., McDonald, A.J., Halsema, G., Struik, P.C., Kumar, P., Malik, R.K., Poonia, S.P., Balwinder-Singh, Singh, D.K., Singh, M., and Krupnik, T.J. 2021. “Social-ecological analysis of timely rice planting in Eastern India.” *Agronomy for Sustainable Development* 41: 14. Doi: <https://doi.org/10.1007/s13593-021-00668-1>.

Wager, S., and Athey, S. 2018. “Estimation and Inference of Heterogeneous Treatment Effects using Random Forests”. *Journal of the American Statistical Association* 113(523): 1228-1242. Doi: <https://doi.org/10.1080/01621459.2017.1319839>.

Zhou, Z., Athey, S., Wager, S. 2022. “Offline multi-action policy learning: Generalization and optimization.” *Operations Research* 71 (1): 148-183. Doi: <https://doi.org/10.1287/opre.2022.2271>.

# Appendices

Table A1: Descriptive statistics

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Variables | Mean | Stdev | Minimum | Lower quartile | Median | Upper quartile | Maximum |
| Wheat yield (ton/ha) | 2.99 | 0.85 | 0.20 | 2.40 | 3.00 | 3.43 | 6.50 |
| Number of irrigations | 2.29 | 0.77 | 1.00 | 2.00 | 2.00 | 3.00 | 5.00 |
| One irrigation (share) | 0.13 | 0.34 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 |
| Two irrigations (share) | 0.50 | 0.50 | 0.00 | 0.00 | 1.00 | 1.00 | 1.00 |
| Three irrigations (share) | 0.31 | 0.46 | 0.00 | 0.00 | 0.00 | 1.00 | 1.00 |
| Four or more irrigations (share) | 0.05 | 0.23 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 |
| Sowing date: after 16th Dec | 0.09 | 0.28 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 |
| Sowing date: 1st Dec-15th Dec | 0.22 | 0.42 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 |
| Sowing date: 21st Nov-30th Nov | 0.41 | 0.49 | 0.00 | 0.00 | 0.00 | 1.00 | 1.00 |
| Sowing date: 11th Nov-20th Nov | 0.22 | 0.42 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 |
| Sowing date: 1-10th Nov | 0.05 | 0.23 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 |
| No schooling | 0.27 | 0.45 | 0.00 | 0.00 | 0.00 | 1.00 | 1.00 |
| Primary | 0.30 | 0.46 | 0.00 | 0.00 | 0.00 | 1.00 | 1.00 |
| Matriculation | 0.21 | 0.41 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 |
| Senior secondary | 0.11 | 0.31 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 |
| Bachelors | 0.09 | 0.28 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 |
| Post grad | 0.02 | 0.14 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 |
| Nitrogen per ha | 130.22 | 37.04 | 0.00 | 105.19 | 132.54 | 156.00 | 298.47 |
| Phosphate per ha | 59.04 | 19.63 | 0.00 | 45.43 | 59.78 | 72.69 | 212.96 |
| Weeded | 0.76 | 0.43 | 0.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| Timely sown variety | 0.53 | 0.50 | 0.00 | 0.00 | 1.00 | 1.00 | 1.00 |
| Weed severity (1-4) | 2.72 | 0.76 | 1.00 | 2.00 | 3.00 | 3.00 | 4.00 |
| Disease severity (1-4) | 1.47 | 0.70 | 1.00 | 1.00 | 1.00 | 2.00 | 4.00 |
| Insect severity (1-4) | 1.60 | 0.76 | 1.00 | 1.00 | 1.00 | 2.00 | 4.00 |
| Drought severity (1-4) | 1.94 | 0.88 | 1.00 | 1.00 | 2.00 | 3.00 | 4.00 |
| Temperature | 26.06 | 0.31 | 24.98 | 25.92 | 26.05 | 26.23 | 26.61 |
| Total annual precipitation (mm) | 953.77 | 254.25 | 599.80 | 741.20 | 890.90 | 1191.60 | 1874.40 |
| Elevation (m) | 68.33 | 21.49 | 27.00 | 54.00 | 67.00 | 79.00 | 327.00 |
| Distance to market (km) | 4.41 | 4.11 | 0.00 | 2.00 | 3.00 | 6.00 | 55.00 |
| Largest plot area (ha) | 0.25 | 0.27 | 0.01 | 0.11 | 0.18 | 0.28 | 8.10 |
| Marginalized caste | 0.76 | 0.43 | 0.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| Plot owned | 0.81 | 0.39 | 0.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| Gender (female) | 0.03 | 0.17 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 |
| Soil nitrogen | 1.60 | 0.21 | 1.10 | 1.50 | 1.60 | 1.70 | 2.70 |
| Soil sand | 29.88 | 3.02 | 19.00 | 28.00 | 30.00 | 32.00 | 47.00 |
| Soil organic carbon | 12.70 | 2.32 | 7.30 | 11.10 | 12.60 | 14.10 | 24.30 |

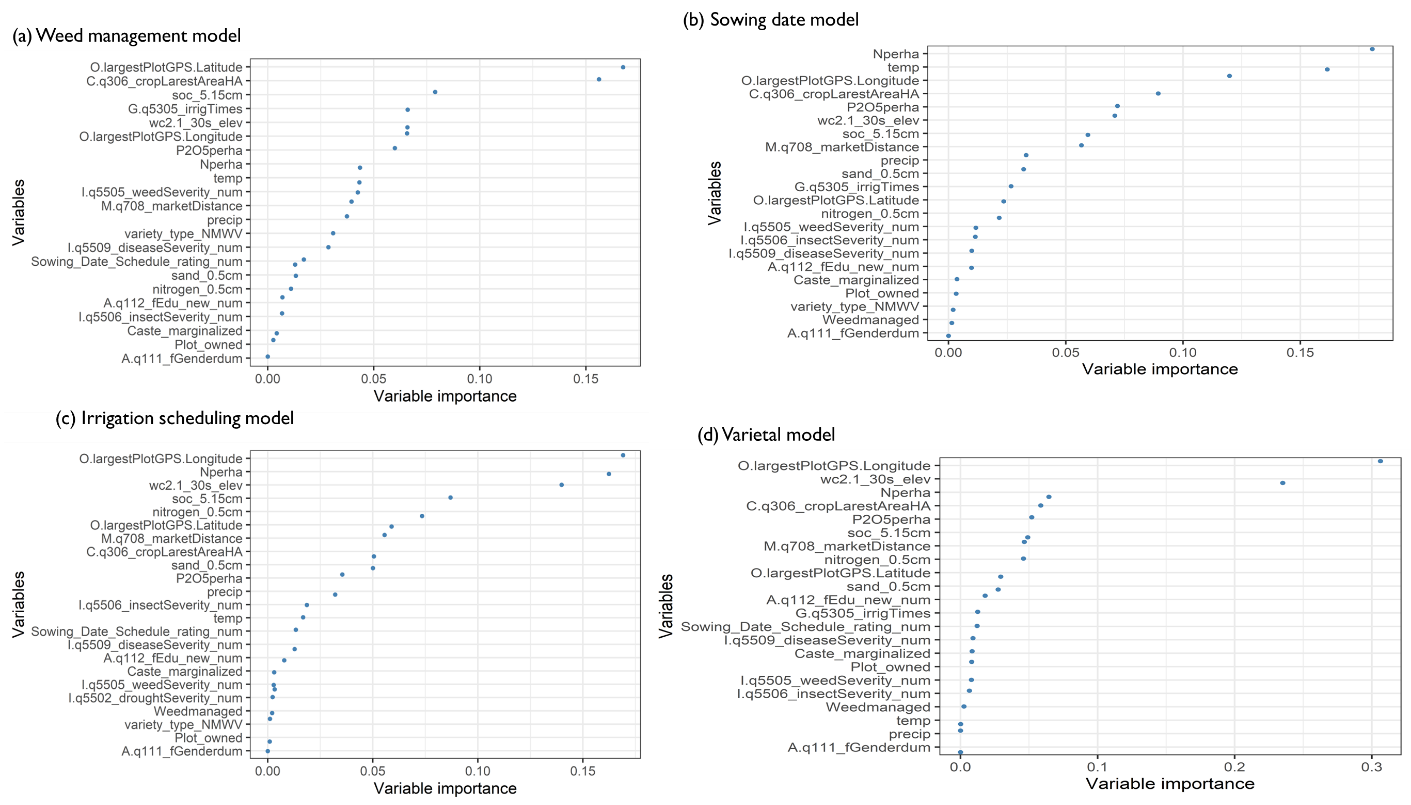


Figure C1: Variable importance plot

1. Table A1in the appendices shows the descriptive statistics of the rest of the variables. [↑](#footnote-ref-1)