Causal Effect Estimation				Refutations						
				Placebo		RCC		UCC	RSR	
Method	**ATE**	**CI**	**p-value**	**Effect***	**p-value**	**Effect***	**p-value**	**Effect***	**Effect***	**p-value**
Linear Regression	546	(211, 880)	0.0015	-25.74	0.39	546	0.49	85	543	0.45
Matching	448	(186, 760)	0.0060	50.82	0.39	432	0.40	116	438	0.48
IPS weighting	471	(138, 816)	0.0010	38.82	0.40	470	0.40	113	462	0.45
T-Learner (RF)	372	(215, 528)	0.0240	9.26	0.49	373	0.46	-	353	0.42
X-Learner (RF)	437	(300, 574)	0.0050	5.10	0.50	430	0.37	-	409	0.36

Table 3: Results of Average Treatment Effect estimation. Includes point estimates, 95% confidence intervals, and four refutation tests. For the Placebo, RCC and RSR refutations, the new ATE estimate is reported (denoted as Effect*), alongside the respective p-value (< 0.05 indicates a failed test). The UCC column reports the mean ATE estimate of the corresponding heatmap (for full heatmaps and details see Sec. 2 of Appendix). Numbers are in cotton kg/ha, rounded to the nearest integer.

from each variable and dividing by its standard deviation. The treatment T is binarized, with 1 indicating that a farmer sowed on a favorable day, and 0 indicating the opposite.

Propensity modeling is a prerequisite of IPS weighting. We thus begin by discussing the propensity model that is fit. Given the relatively small dataset size, logistic regression is used on the scaled back-door adjustment set Z for classifying each field into the treatment/control group. We subsequently trim the dataset by removing all rows with extreme propensity scores (< 0.2 or > 0.8) to aid the overlap assumption (**?**). The resulting distribution of propensity scores can be seen at Figure 3. The model scores 0.81 in accuracy, 0.64 in F1-score, and 0.88 in ROC-AUC. After trimming extreme propensity scores, a subset of 48 treated and 37 control units remains. There is decent overlap between the propensity score distributions of the treatment and control group, indicating that they are comparable and enabling reliable propensity-based ATE estimation.

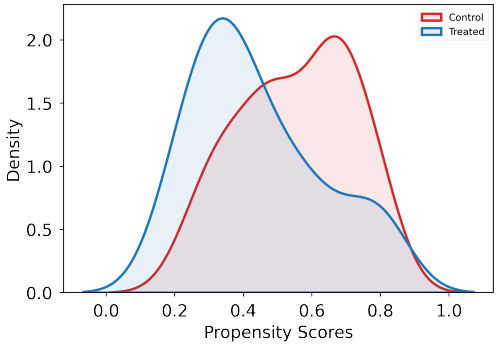

Figure 3: Distribution of propensity scores for the control and treatment group after trimming extreme scores.

Table 3 and Figure 4 show the results of the ATE estimation per method, alongside the corresponding 95% confidence intervals and p-values. Besides Linear Regression, other methods do not provide confidence intervals by default. For matching, IPS, and meta-learners confidence intervals and the resulting p-values are hence bootstrapped. Both the T-learner and X-learner use a Random Forest for modeling the outcome Y.

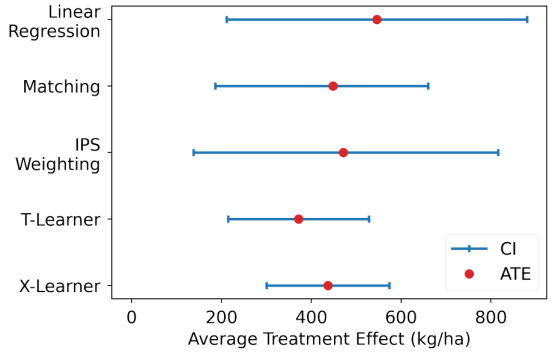

Figure 4: ATE point estimates and 95% confidence intervals for all estimation methods.

All methods detect a significant ATE at 95% confidence level, with point estimates ranging from 372 to 546 kilograms of cotton per hectare. For context, the average observed yield is 3,145 kg/ha. We thus infer that the causal effect of following the sowing recommendation on yield is significantly positive, driving a yield increase ranging from 12% to 17%, depending on the estimation method used.

Of central importance are the refutation tests we run after having estimated the recommendation impact. Table 3 features analytic results for all method / refutation test combinations. All estimation methods are robust against performing the following data manipulations and re-estimating the ATE: randomly permuting the treatment (Placebo test), adding a confounder (RCC test), sampling a subset of data (RSR test) and creating unobserved confounding (UCC test). Specifically, Placebo ATE estimates do not differ significantly from 0, while RCC and RSR estimates do not differ significantly from the already obtained ATE. For the UCC test, the mean ATE estimates are reduced yet remain positive, despite unobserved confounding of significant magnitude. Confidence intervals and p-values are bootstrapped (1000 iterations).

The results indicate that the recommendation system's advice drove a net increase in yield that was deemed both significant and robust from a statistical perspective. By utilizing

the theory of graphical causal models, the analysis transparently puts forward its assumptions and explicitly incorporates domain knowledge in it. Combined with accurate and performant systems, such analyses can benefit the reliability and adoption of digital agriculture as well as farmers' trust. The provision of information on the actual impact expected from a recommendation system may also enable a cost-benefit analysis on behalf of the farmer, by simply comparing the digital tool cost to the expected yield gain. Even though the analysis is transparent, it is as good as the causal assumptions it makes and the DAG it develops. Our graph is consistent with agri-environmental knowledge on cotton, however there is always a possibility that bias exists, either due to a missed confounder or due to a missed interaction between observed variables. The robustness checks we performed were all successful; noting that when we add strong unobserved confounding, the UCC test estimates become volatile - an expected behavior to a certain degree.

Given the homogeneous management practices among farmers in our data, we remark that external validity of estimates is low, as results cannot be expected to generalize to other farms that might follow different routines. Nevertheless, it is not uncommon for farmers to follow similar practices in other regions or even entire countries. While the transfer of effect estimates warrants caution, the same does not hold for the proposed framework itself. Given relevant data and knowledge, a graph-based empirical evaluation of an agricultural recommendation system can normally proceed. If consensus among estimation methods in terms of ATE significance is reached, the tool is deemed beneficial; otherwise, more work is required. All in all, this system equips farmers with a provably valuable tool based on cotton knowledge and weather forecasts. It contributes to a successful growing season and lowers the likelihood of farmers resorting to expensive actions, e.g., replanting a field. For the growing season of 2022, the recommendation system was deployed at national scale and extended to two other crops (maize, sunflower). These new pilot applications will allow us to practically test the external validity of our results across different seasons, crops and locations. Moreover, given the developed causal graph G, the crop growth (CG) variable that is sufficiently captured through its NDVI proxy, mediates the effect of T on Y. The front-door criterion (**?**) might thus provide an alternative identification method for the ATE, and we plan on exploring it in collaboration with domain experts. Finally, the more growing seasons the recommendation system has seen, the more data are obtained. Going beyond ATE estimation by learning Conditional ATEs and using causal machine learning methods for providing personalized effect estimates is another next step. Due to the rich and well-established domain knowledge, we finally believe that the potential of causal reasoning in agriculture extends far beyond effect identification. Fitting Structural Causal Models and performing counterfactual inference can enable a greater understanding of the farm system and supercharge decision support tools.

Most generally, the essential condition that allowed us to utilize causal inference for empirically evaluating an agricultural recommendation system is the ample, long-established domain knowledge that exists. Decision support systems are being used on multiple fields (**?**) such as medical decision making (**?**) or forest and fire management (**??**). The aforementioned fields possess accumulated domain knowledge on the interactions a good system exploits; the same way we possess information on environmental conditions related to cotton planting. We thus expect graphical approaches to be valuable for the empirical evaluation of decision support systems of diverse domains.

Conclusion

In this study, we design, implement, and test a digital agriculture recommendation system for the optimal sowing of cotton. Using the collected data and leveraging domain knowledge, we evaluate the impact of system recommendations on yield. To do so, we utilize and propose causal inference as an ideal tool for empirically evaluating decision support systems. This idea can be upscaled to other digital agriculture tools as well as to different fields with well-established domain knowledge. This paradigm is in principle different to decision support systems that frequently use black-box algorithms to predict variables of interest, but are oblivious to the evaluation of their own impact. In that sense, this work comes to the defence of the farmer, by introducing an AI framework for elaborating on the assumptions, reliability, and impact of a system before discussing service fees.

Acknowledgements

We thank the Agriculture Cooperative of Orchomenos for the collaboration and data provision, and Corteva Hellas for their support. This work was supported by the EU H2020 Research and Innovation program through the eshape project (grant agreement No. 820852). It was also supported by the MICROSERVICES project (2019-2020 BiodivERsA joint call / BiodivClim ERA-Net COFUND programme, and with GSRI, Greece - No. T12ERA5-00075).

Animi fuga eveniet illum ipsum, ipsum iure minus magnam ducimus harum, molestias corrupti ab aut, temporibus magnam tempora nostrum quidem quis autem ratione, officia iste iure assumenda asperiores atque eligendi iusto illo animi nemo?Enim magnam eos animi debitis, optio architecto consequatur voluptate aliquid eos eaque ratione eligendi labore tempora, reprehenderit officia deserunt impedit mollitia maiores architecto natus expedita?Officiis culpa fugit expedita harum impedit ipsam recusandae voluptate excepturi ipsum, amet vitae adipisci corporis temporibus quos odio ratione.Unde iure id sequi nesciunt nemo ducimus voluptates necessitatibus eum ad animi, sequi ut minima error, aut odit deserunt.Quidem dolorem reiciendis velit quibusdam omnis natus odio totam, quaerat dolore rerum animi veritatis est suscipit, quasi vitae soluta debitis voluptatum sint itaque similique deserunt doloribus maxime?Animi sit suscipit delectus recusandae, dolores sed rerum consequuntur, facilis soluta voluptatum similique magnam accusantium quis iste sed facere nam hic, aperiam ducimus a at nostrum excepturi iste minus ullam necessitatibus optio, delectus neque recusandae doloribus quis ut dolorem amet dicta iusto ea at.Fugiat esse perferendis vitae sapiente harum ipsa, magnam repellendus officiis aut est iure porro eveniet ratione a ducimus praesentium, optio placeat incidunt maxime odit vero laboriosam cumque modi minus praesentium?Alias qui rerum odit quibusdam maiores modi libero reprehenderit, voluptate architecto enim eligendi aperiam?Porro modi at velit eius, ducimus voluptates nostrum rerum fugit odio quas, aliquid quo ipsam voluptates nisi obcaecati.