Everything under control : comparing machine learning and classical econometric impact assessment methods using FADN data

European Review of Agricultural Economics

Brignoli, P.L.; de Mey, Y.; Gardebroek, C.

https://doi.org/10.1093/erae/jbae034

European Review of Agricultural Economics Vol **00 (00)** (2024) pp. 1–32
doi:https://doi.org/10.1093/erae/jbae034

Everything under control: comparing machine learning and classical econometric impact assessment methods using FADN data

P. L. Brignoli[†,*], Y. de Mey[‡] and C. Gardebroek[†]

[†]Agricultural Economics and Rural Policy Group, Wageningen University & Research, Wageningen, The Netherlands; [‡]Business Economics Group, Wageningen University & Research, Wageningen, The Netherlands

Received 26 April 2023; Accepted 27 November 2024

Abstract

Machine learning (ML) methods have been proposed to improve the assessment of agricultural policies through enhanced causal inference. This study uses a simulation framework tailored to Farm Accountancy Data Network (FADN) data to scrutinize the performance of both ML and classical methods under diverse causal properties crucial for identification. Our findings reveal significant variations in performance across different treatment assignment rules, sample sizes and causal properties. Notably, the Causal Forest method consistently outperforms others in retrieving the causal effect and accurately characterizing its heterogeneity. However, the data-driven approach of ML methods proves ineffective in selecting the correct set of controls and addressing latent confounding.

Keywords: causal inference, machine learning, FADN, controlled simulation experiment

1. Introduction

As the objectives of the European Union's Common Agricultural Policy (CAP) continue to increase in number and ambition, and the need to justify them to the public remains unchanged, it is crucial for CAP interventions to be effective. In the period 2023–2027, the CAP can rely on its largest budget in history, with 387 billion euros earmarked for promoting a greener, fairer and more competitive agriculture (European Commission, 2022). However, in order to ensure progress towards these goals, it is essential to assess the effectiveness of CAP interventions.

*Corresponding author: E-mail: paolo.brignoli@wur.nl

The ex-post evaluation of CAP interventions has predominantly relied on conventional econometric methods, such as matching, as evidenced in the literature. For instance, the effects of agri-environmental schemes (AES) have been analysed using propensity score matching (PSM) (Pufahl and Weiss, 2009), or PSM combined with difference-in-differences (DiD) (Arata and Sckokai, 2016; Mennig and Sauer, 2020). A notable exception is Stetter, Mennig and Sauer (2022) who analysed AES participation in Southeastern Germany using a causal forest, a machine learning (ML) technique.

Despite their widespread use in the literature, classical matching methodologies have come under scrutiny due to several concerns that have been raised (Breiman, 2001; King and Zeng, 2007; King and Nielsen, 2019; Storm, Baylis and Heckelei, 2019). In general, these methods are (asymptotically) unbiased as long as the data generation process (DGP) is known, meaning that we can choose a method whose underlying assumptions for unbiasedness match the DGP structure. However, in scenarios where this is not the case, data-driven approaches based on more flexible assumptions may offer a more credible alternative than attempting to guess the correct specification (Hastie, Tibshirani and Friedman, 2009).

Since CAP evaluation involves sample selection bias and treatment heterogeneity, ML methods could greatly contribute to agricultural policy impact assessment (Storm, Baylis and Heckelei, 2019). These methods have emerged as a promising complement or even alternative to classical econometric methods for causal inference recently (Athey, 2018; Athey and Imbens, 2019; Storm, Baylis and Heckelei, 2019). Their key advantage lies in the ability to model both the treatment assignment and the outcome without restrictions on the functional form or number of variables, as well as the ability to explore heterogeneity across dimensions not specified previously (Athey, 2018). These strengths enable ML methods to overcome some of the limitations of classical methods (Athey, Tibshirani and Wager, 2019; Hahn, Murray and Carvalho, 2020).

However, due to the inherent challenge of causal inference, evaluating a new estimator entails a fundamental reliability issue. The fundamental problem of causal inference is that since the counterfactual is unobserved, one cannot know whether the true treatment effect was retrieved in any given observational scenario. To determine whether a method is trustworthy in a specific context, it must be validated in a controlled environment where the counterfactuals are known.

Therefore, the objective of this paper is to assess the reliability of classic and ML causal estimators in retrieving a treatment effect by comparing their performances in a simulation study tailored to EU agricultural policies. To tailor our simulation to the agricultural economics domain, we start from the European Farm Accountancy Data Network (FADN), the reference dataset for studying the CAP (Pufahl and Weiss, 2009; Arata and Sckokai, 2016; Mennig and Sauer, 2020; Stetter, Mennig and Sauer, 2022) and focus on participation in agri-environmental schemes. We restrict our analysis to causal inference with observational data as researchers in the agricultural policy domain are often

limited to using administrative data for their analyses and policy experiments are often not feasible.

Our simulation study is designed building on Lechner and Wunsch (2013), Huber, Lechner and Wunsch (2013), Wendling et al. (2018) and Knaus, Lechner and Strittmatter (2021). In addition to customizing such simulation to an agricultural economics setting, our research extends the existing simulation literature by investigating various treatment assignment procedures, and differentiating between confounding (selection into treatment) and sample selection bias. Incorporating treatment assignment under varying assumptions enables a more equitable comparison, reflecting the diverse ways farmers might choose to participate in the treatment. While the traditional econometric literature often uses confounding and sample selection bias interchangeably (Cinelli, Forney and Pearl, 2022), recent advancements in causal inference offer a framework to accurately model these distinct phenomena. This approach enhances our understanding of each method's performance across different contexts: whether farmers self-select into treatment (the outcome and the decision to participate are influenced by the same variables—confounding), or whether the sampling process or estimation procedure over- or under-represents a certain group (sample selection bias).

The paper makes three contributions to the literature on evaluating EU agricultural policies. First, it proposes an evaluation framework for causal methods that can be applied flexibly, enabling a nuanced understanding of the most suitable method for addressing diverse agricultural economic impact-assessment questions. Second, it examines the behaviour of the considered estimators across different scenarios and highlights factors that can cause a method to fail in retrieving the true effect. Third, it offers guidelines for the practical application of the considered methods, taking into account the scenarios' characteristics and the estimators' functioning. Overall, these contributions aim to support agricultural economists in making informed choices when selecting a causal estimator for a specific policy evaluation context.

2. Observational data: ML versus classical approaches

The primary strength of ML methods lies in their data-driven approach to selecting variables and functional forms. Researchers often aim to control for as many confounders as possible to enhance the credibility of the unconfoundedness assumption (Baiardi and Naghi, 2021). However, classical approaches typically focus on a restricted set of variables selected based on existing literature. In the absence of theoretical guidance for modelling choices, data-driven approaches may be preferable as they are not limited in the number of variables they can use to model the response function and treatment assignment. Moreover, data-driven variable selection enhances confidence that these choices were made to satisfy the unconfoundedness assumption, rather than through a trial-and-error process aimed at achieving specific results or engaging in questionable research practices. Therefore, data-driven approaches to variable selection are particularly relevant when the theoretical framework behind the

research question is not sufficiently developed to guide variable choice (Hastie, Tibshirani and Friedman, 2009).

Controlling for confounders is essential in causal inference, which requires both selecting the correct controls and functional form (King and Nielsen, 2019; Chernozhukov *et al*., 2018). However, most classical econometric methods rely on the assumption of linearity in parameters, which might not be credible in complex contexts like agriculture where biological, social and economic factors interact (Storm, Baylis and Heckelei, 2019). In such cases, linear functional forms might not be sufficiently flexible or powered to capture complex non-linearities and interactions, and at the same time not sufficiently rooted in theory for justifying their use (Storm, Baylis and Heckelei, 2019; Hastie, Tibshirani and Friedman, 2009). In contrast, flexible ML methods can account for non-linearities and high-order interactions intrinsically, allowing to obtain a better fit to the underlying DGP. Moreover, this flexibility is particularly relevant in the presence of a specific type of latent confounding, where the confounders are related to observable controls. Latent confounding occurs when a confounder is unobserved, either because it is missing from the dataset or because it is unmeasurable. However, if a complex combination of observable controls can sufficiently approximate the latent confounder, the estimator can still accurately identify the effect (Louizos *et al*., 2017; Kallus, Puli and Shalit, 2018; Bennett and Kallus, 2019; Wang and Blei, 2019).

The data-driven choice of variables and greater model flexibility could also help in dealing with selection into treatment afflicting CAP interventions, where factors involved and their relationships are unknown but observed. Theoretical arguments on CAP participation benefits might not always align in practice with participants' expectations. For instance, a researcher might believe that the take-up rate of an AES would be dictated by subsidies and the farm structure, while in practice it could be driven by a marketing campaign launched by an NGO operating only in certain areas. Allowing the algorithm to determine which variables to include and how to include them is expected to increase the robustness of the analysis in this situation.

Last, ML also allows to explore treatment heterogeneity without incurring the bias caused by multiple hypothesis testing. Understanding how the treatment effect varies by farmers' characteristics is highly relevant for policymaking (Koutchadé, Carpentier and Femenia, 2018), as it allows to design more cost-effective ways of achieving set objectives. Standard approaches to exploring treatment heterogeneity imply prespecifying a set of sub-populations over which to assess differences in the causal effect (which is prone to confirmation bias on the researcher's end). However, when proceeding with the evaluation, the researcher must account for the probabilistic nature of testing: out of 20 hypotheses on treatment effectiveness tested, we expect to reject incorrectly the null of at least one of them considering a 5 per cent significance level. Although several corrections are available for the issue (Bonferroni, 1936; Benjamini and Hochberg, 1995; Holm, 1979) they do not scale well in presence of many covariates considered, severely limiting the possibility

for testing effect heterogeneity. This is in stark contrast with tree-based methods, which naturally target the individual treatment effects (ITEs) (Athey, Tibshirani and Wager, 2019; Hahn, Murray and Carvalho, 2020).

However, the benefits listed so far do not come without a price, with interpretability being a primary concern. Specifically, not being able to completely understand a model makes its debugging and utilization more difficult (Storm, Baylis and Heckelei, 2019). However, the loss of interpretability should not be deemed inherently connected to ML methods, but rather as part of a trade-off with model complexity. Therefore, in some cases, it may be preferable to have a more sophisticated model, able for instance to allow for treatment heterogeneity, than a simplified but more understandable approximation. In recent years, procedures have been developed to improve the interpretability of ML methods (Molnar, 2020). Nonetheless, interpretability is associated with issues such as transparency, fairness and manipulability, which further hinder the perceived reliability of ML methods (Athey, 2018; Storm, Baylis and Heckelei, 2019).

3. Simulation background and data

3.1. Underlying assumptions

Monte Carlo simulation using synthetic data are commonly used to provide insights into the functioning and properties of estimation methods, allowing the researcher to establish the ground truth by retaining control over both the treatment assignment and effect (Dorie *et al*., 2019; Wendling *et al*., 2018; Hahn, Dorie and Murray, 2019; Knaus, Lechner and Strittmatter, 2021). Notable examples of simulation studies using hypothetical DGPs to examine the properties of estimators include Frölich (2004), who explores different matching procedures; Zhao (2004), who compares matching based on propensity scores versus matching based on covariates; Busso, DiNardo and McCrary (2009, 2014), who compare matching versus reweighting procedures; and King et al. (2011), who delve into various classes of matching procedures and describes the PSM paradox.

Our simulation is based on FADN data and mimics participation in AES. AES are a series of interventions aimed at rewarding farmers for the provision of positive externalities beyond the mandatory requirements to obtain EU subsidies (Baylis *et al*., 2008). Over the years, AES have been one of the main subjects of the European agricultural policy-evaluation literature. This extensive research focus stems from a paradox: while AES receive the largest share of rural development program funding (Arata and Sckokai, 2016), they face significant criticism for their design (Massfeller *et al*., 2022). This chosen application does not affect the generalizability of the results, but it provides a framework for making decisions consistently, reducing researcher discretion. Therefore, it is possible to think of our treatment as an additional AES scheme that is guaranteed to have an effect on the designated outcome.

To obtain a representative and meaningful simulation, we have two targets: (i) adhering as closely as possible to reality, while (ii) being able to control

the relevant factors in observational causal inference in the domain of agricultural policy impact assessment. The first target is achieved by simulating an outcome and a treatment assignment tailored to the FADN dataset, ensuring their relationship with farm characteristics. Moreover, we compute the minimum detectable effect (MDE) based on the number of observations present in the FADN and stick to the percentage of treated found for the AES. The second target is achieved by including parameters in the outcome and treatment assignment equations that allow us to manipulate the causal properties of each DGP. Specifically, we analyse the effects of five supplementary causal properties that influence the structure of causality in empirical studies: (i) the degree of non-linearity, (ii) the share of common support, (iii) the strength of the treatment effect, (iv) the presence of sample selection bias and (v) the presence of latent confounding. We choose these properties, as well as their levels, so to include what we perceive as most relevant in agricultural policy impact assessment—more specifically, FADN-based analysis—while constraining the number of possible combinations.

To have an identifiable effect, the DGPs in this simulation are ensured to satisfy three assumptions: conditional independence (excluding DGPs with latent confounding), common support and the stable unit treatment value assumption (SUTVA). Conditional independence states that the treatment assignment is independent from the potential outcomes conditional on the controls; common support concerns the existence of overlap in the probability of being treated between the treatment and control group; and SUTVA requires the treatment effect of each individual to be independent from the treatment status of other individuals (Imbens and Rubin, 2015). Although our simulation is designed to guarantee conditional independence, it is worth noting that this is a strong assumption and the least favoured in economics. However, as applied researchers are often confined to working with (increasingly large) observational datasets, we see value in exploring how causal ML methodologies are able to identify causal effects in an observational context.

3.2. Data

Our starting point is the FADN dataset for all the EU-28 countries in the year 2020[1] (DG AGRI, 2023), consisting of 81,834 farms. We then conduct our simulation, comprising the DGPs combination and the treatment assignment procedures, on two different sample sizes. The larger dataset encompasses all available observations, while the smaller dataset is limited to 4000 observations. We focus on two different sample sizes for two reasons. First, CAP impact evaluations are usually conducted either aggregating data from various member states (Arata and Sckokai, 2016), or at a regional level (Stetter, Mennig and Sauer, 2022). Second, ML models are expected to manifest a slower convergence rate than classical models (Abadie and Imbens, 2011; Chernozhukov *et al.*, 2018; Hastie, Tibshirani and Friedman, 2009). Therefore, by having two different sample sizes, we can offer guidance to researchers who

1 The most recent year available to us for analysis.

may engage in either type of analysis, while analysing to what extent smaller or larger datasets can affect the performances of the estimators.

The nature of the FADN dataset, with a vast array of different variables, imposes us to start the simulation by limiting in a data-driven way which variables enter the DGPs. Not all variables can be expected to be equally relevant, due to the strong heterogeneity in European farming and the high specificity of each variable, so that many variables are excessively sparse.[2] Therefore, the first step in our simulation is the restriction of the dataset to the least sparse variables, described in this section as a matrix of covariates X plus a subscript identifying the set, viz. X_{TD} (treatment determinants) and X_{OD} (outcome determinants).

We run a LASSO regression on both a proxy outcome (farm revenues[3]) and a proxy treatment (participation in AES[4]), obtaining each variable ranking for how good it predicts its target (Hastie, Tibshirani and Friedman, 2009). To ensure the correct training of the LASSO model, we preprocess our data with two transformations. First, we apply the Yeo–Johnson transformation (Yeo and Johnson, 2000) to deal with both outliers and the skewness of the variables' distribution. Second, we divide each variable by its own maximum absolute value, to preserve the data sparsity[5] while constraining it between 0 and 1. Then, we select the first 75 variables,[6] in order of importance determined by the LASSO models, for each target (avoiding repetitions).

Afterwards, we randomly pick the variables used in the equations across the different DGPs and iterations. These variables are divided into observed confounders (X_{OC}), latent confounders predictors (X_{LC}—the observed controls correlating with the latent confounders) and other outcome predictors (X_{OP}) (Table 1). For each set, we start from either X_{TD} or X_{OD} and select the variables so that they contain both continuous and categorical variables. We obtain the following sets:

X_{OC}: 5 variables such that $X_{OC} \subset X_{TD}$
X_{LC}: 6 variables such that $X_{LC} \subset X_{TD}$
X_{OP}: 5 variables such that $X_{OP} \subset X_{OD}$

Following Kallus, Puli and Shalit (2018), we incorporate (a specific form) of latent confounding in our simulation. We do not directly employ the X_{LC} variable set in constructing our simulation, but rather we construct two latent confounders, $\widehat{X_{LC}}$, as a linear combination of three variables in X_{LC}, each plus

2 Variables designed to capture the particularities of a specific region or production system will necessarily report no values for most others. For instance, the variable indicating the number of hectares dedicated to strawberry cultivation will be zero for all farmers not growing strawberries.

3 Identified in the FADN with the code SE005.

4 Identified in the FADN with the code SAEAWSUB_2_V. This variable measures the amount of subsidies farmers received from AES—so in order to use it as a treatment variable, we encoded it as a categorical variable being 1 if the amount received was greater than zero and 0 otherwise.

5 Meaning that the relative distance between each point is maintained.

6 We chose this number arbitrarily, aiming to strike a balance between obtaining a dataset large enough to fully demonstrate ML methods capability and the implied running time. We chose the number of variables for each set with the same logic.

Table 1. Variable sets and related FADN identifying codes

Variable set	FADN identifier	Name
Observed confounders X_{OC}	NUTS2	Regional indicator
	SE025	Total utilized agricultural area
	SE010	Total labour input
	SE132	Total output/total input
	SE624	Total support for rural development
Latent confounders X_{LC}	SE042	Area in energy crops
	SE621	Environmental subsidies
	SE105	Number of poultry
	LEGAL	Farm legal status
	SE281	Total specific costs
	SE446	Land, permanent crops and quotas
Outcome predictors X_{OP}	AMCHQP_AD	Machinery accumulated depreciation value
	ACSHEQ_CV	Cash and equivalents closing value
	ALNDAGR_OV	Agricultural land opening value
	ARECV_CV	Receivables closing value
	TF8	Type of farming

a random term. This $\widehat{X_{LC}}$ set of variables becomes the one entering the DGPs, while each method will be provided with X_{LC} to evaluate its capability to correctly approximate and manage latent confounding. See Appendix A for more details on the variables employed.

4. Simulation design

We start our simulation from a linear DGP, wherein the treatment assignment is determined by a propensity score, ensuring the assumptions of unconfoundedness and common support are satisfied. This approach establishes a comprehensible reference for our study, serving as a benchmark for all subsequent comparisons. From this baseline, we incrementally introduce modifications to the DGP, the treatment assignment mechanism, or both, to evaluate how the performance of each method diverges from the outcomes observed under the base DGP.

We design 32 DGPs, obtained by manipulating 5 causally relevant properties (each taking on two different levels) of the base DGP. The causal properties considered are (i) degree of non-linearity, (ii) share of common support, (iii) strength of the treatment effect, (iv) presence of sample selection bias and (v) presence of latent confounding. Over the following sub-sections, we start by introducing the base DGP, and then proceed explaining how we extend that

DGP to modify each causal property. The full replication code is available is the online Supplementary Materials.

4.1. Base DGP

To establish our base DGP, we begin by simulating the outcome of interest Y following a linear model as specified in Equation 1:

$$Y = \alpha X_{OP} + \beta X_{OC} \tag{1}$$

Y is obtained as the linear combination of X_{OP}, the outcome predictors and X_{OC}, the observed confounders. The corresponding vectors of coefficients for these variables are denoted by α and β, respectively (Appendix B).

In the second step, we compute each observation's probability of receiving the treatment (the propensity score) using the logit function defined in Equation 2:

$$\begin{aligned} \pi(x) &= \Pr(T = 1) = \frac{1}{1 + e^{-C}} \\ C &= \kappa X_{OC} + \varepsilon \end{aligned} \tag{2}$$

The propensity scores are calculated as a linear combination of X_{OC}, the observed confounders, with an added noise term ε. The vector of coefficients associated with the observed confounders is denoted by κ. Consequently, confounding is introduced by including X_{OC} in both the outcome regression and the treatment assignment process.

The treatment assignment rule is then constructed using the computed propensity scores in a Bernoulli distribution (Millimet and Tchernis, 2009; Huber, Lechner and Wunsch, 2013; Knaus, Lechner and Strittmatter, 2021), enforcing a 90 per cent overlap[7] and a fixed percentage of treated units. This percentage is set as the actual percentage of farmers that participate in the AES reported in the FADN in 2020 (26.1 per cent). While the first constraint will later on be used to test the robustness of the methods to variation in the overlap share, the second is a feature we added to the simulation to strengthen the link with actual FADN-based impact assessment studies.

The third step involves estimating θ_{ATT}, the average treatment effect on the treated (ATT). To add the possibility to evaluate the ability of ML methods to address treatment heterogeneity, we first draw the individual treatment effects for each farm based on its farm type (τ_{i,TF_i} in Equation 3) from a uniform distribution $U \sim (-1, 1)$. Then, to fix the ATT at a defined value, we compute a correction term ($\theta_{correction}$), which is the difference between the average of the individual farm type effects and the target ATT (Equation 3). The target ATT is determined as double the minimum detectable effect (MDE) calculated through power analysis based on the size of the FADN dataset (Huntington-Klein, 2021) using linear regression. This approach yields two

7 Meaning that 90 per cent of the treated units fall into the propensity score interval of the control units.

distinct estimands of interest: the ATT, which allows comparison across all considered methods, and the conditional ATT (CATT), which exclusively assesses the performance of ML techniques.

$$\theta_{ATT} = \left(\frac{1}{n}\sum_{i}^{n}\tau_{i,TF_i}\right) - \theta_{correction} \tag{3}$$

By allowing the treatment effect to vary in this specific manner, we are able to assess the effectiveness of ML models in detecting treatment heterogeneity across sub-groups. In scenarios like agricultural policy evaluation, it is reasonable to assume that the treatment effect varies among subgroups given the vast heterogeneity in farmers' characteristics (Stetter, Mennig and Sauer, 2022). In this context, understanding how the treatment effects vary conditionally on a specific farm characteristic (in our simulation, the farm type) becomes crucial for better comprehending the policy instrument in question, improving its design and enhancing targeting strategies.

Finally, we introduce clustered errors into our base DGP to prevent the model from being deterministic. Specifically, we consider a scenario where errors are clustered at the country level (Equation 4), as assuming identically and independently distributed errors would be unrealistic given the structure and data collection methods of the FADN (DG AGRI, 2023). In Equation 4, ε_s is a normally distributed error term, while $\varepsilon_{country}$ is a country-specific error term. This specification implies that 10 per cent of the residual variation is country-specific. Lastly, σ_y denotes the standard deviation of the outcome (Hahn, Dorie and Murray, 2019).

$$\begin{aligned} Y &= \alpha X_{OP} + \beta X_{OC} + \theta T + \sigma_y\left(0.9\varepsilon_s + 0.1\varepsilon_{country}\right) \\ \sigma_y &= \sqrt{Var(Y + \pi\tau(x_i))} \end{aligned} \tag{4}$$

Note that the errors are defined additively in this way. While non-additive errors could arise in high-dimensional and non-linear settings, considering them here would add unnecessary complexity to our simulation setup (for an example, see Hahn, Dorie and Murray (2019)).

4.2. Introducing non-linearity

The base DGP is modified to violate the linearity assumption by changing both the outcome regression and treatment assignment functional form. Specifically, Equation 5 describes the three functions $r(x)$, $s(x)$ and $t(x)$ that we use to introduce non-linearity.

$$r(x) = (x+1)^3$$

$$s(x) = \begin{cases} \frac{1}{r(x)} & if \quad x \le -\rho \\ round(r(x)) + \eta & if \quad -\rho \le x \le \rho \\ \frac{1}{r(x)} & if \quad x > \rho \end{cases} \tag{5}$$

$$t(x) = \frac{\log(|s(x)|)}{e^{\frac{\cos(s(x))}{s(x)}}}$$

The first transformation involves cubing the linear combination of variables sets in consideration ($X_{OP}$and X_{OC} for the outcome regression, X_{OC} for the treatment assignment), and an added constant, thereby incorporating[8] up to third-order polynomial and interaction terms. This polynomial is then passed through a step function that includes hyperbolic terms and rounding to disrupt the continuity of the function. The values of η and ρ are chosen to prevent spikes in the outcome distribution. Finally, a combination of logarithmic, exponential and cosine functions is used to further increase the complexity of the outcome. A more conventional choice for a non-linear functional form could have been to utilize well-known forms from the economic literature. However, we opted for this arbitrary form, inspired by the simulation literature (Dorie *et al*., 2019), to allow a broader applicability of our analysis across various evaluation contexts. To introduce non-linearity in the treatment assignment, we apply the same procedure to C (Equation 2) before passing it through the logit function. Consequently, in DGPs where non-linearity is introduced, Equations 1 and 2 can be rewritten as Equations 6 and 7, respectively:

$$Y = t(\alpha X_{OP} + \beta X_{OC}) \tag{6}$$

$$\begin{aligned} \pi(x) &= \Pr(T = 1) = \frac{1}{1 + e^{-C}} \\ C &= t(\kappa_1 X_{OC} + \varepsilon) \end{aligned} \tag{7}$$

The reason why we give this much importance to the degree of non-linearity is because treatments taking place at the border of different domains (economic, social and environmental) are expected to be highly complex (Börner *et al*., 2017; Schlüter *et al*., 2023). Moreover, the importance of accounting for higher-order variables and their interactions stems from the necessity of correctly incorporating agricultural system characteristics. These higher-order variables are assumed to represent elements of a farmer's decision-making process that display varying marginal effects, either increasing or decreasing. For example, when assessing the impact of subsidies aimed at modernizing machinery and farm equipment, the probability of a particular farmer enrolling in the programme is expected to increase progressively with each additional unit of land they possess. In order to capture the diverse range of farm structures stemming from various combinations of factors, we incorporate interaction terms into our modelling approach. To illustrate, we examine the production costs of a tomato farmer, where factors such as being located in a region characterized by high temperatures and proximity to a body of

8 By expanding the cubic polynomial (i.e. multiplying each term by itself and by the other terms in the polynomial), the higher-order and interaction terms become apparent.

water play a significant role. Independently, both these factors might reasonably be assumed to decrease the production costs, as high temperatures and ample water supply favour tomato production, reducing the need for external inputs. However, the combination of these two factors could potentially increase the production costs due to a higher presence of insects, necessitating the use of more pesticides. This nuanced consideration underscores the complexity inherent in real-world agricultural systems and highlights the necessity of incorporating interaction terms in our models.

4.3. Changing common support

To change the overlap in common support from the base DGP, we enforce the treatment assignment rule so that only 50 per cent of the treated units fall into the propensity score interval of the control units (Appendix B). The assumption of common support is critical for all methods considered in this analysis, and this modification will help us understand the extent to which these methods depend on that assumption.

In agricultural economics, it is often observed that the control group is not ideal for comparison with the treated group. This discrepancy can arise, for instance, when examining the switch to organic farming or when treatments are administered based on geographical location or production orientation, resulting in non-treated observations that are structurally different from treated observations.

4.4. Changing treatment effect size

To vary the treatment effect size, we adjust $\theta_{correction}$ in Equation 3 accordingly. Specifically, while the base data DGP sets the treatment effect at double the MDE from the power analysis, we also explore scenarios where the treatment effect equals the MDE itself. Including the size of the causal effect in simulations is a standard practice in the literature, but examining method performance near the detection threshold is particularly relevant in agricultural economics. This field often experiences gradual rather than abrupt changes in response to interventions. Agricultural practices evolve slowly due to factors such as the steep learning curve involved in altering farm practices or the time required for soil and crop modifications to yield noticeable effects. Therefore, assessing method performance under conditions where treatment effects are small yet detectable is crucial for accurately evaluating intervention impacts in agricultural settings.

4.5. Introducing sample selection bias

Sample selection bias is typically described as a preferential selection to the pool, arising during the sampling phase or from controlling for a 'bad control' or collider[9] (Bareinboim and Pearl, 2012; Cinelli, Forney and Pearl, 2022).

9 A collider is defined as a common effect shared by the two variables on a given causal path.

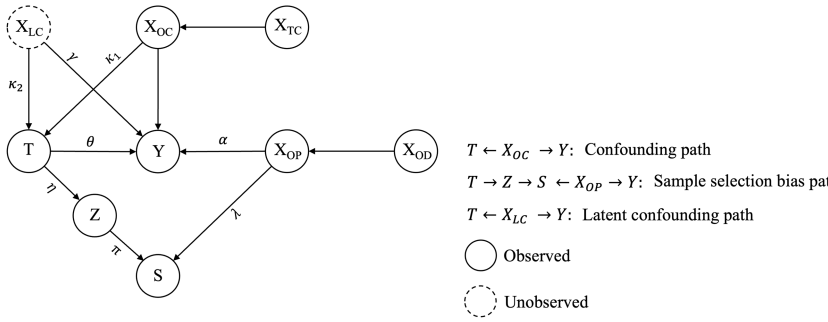

Fig. 1. Directed acyclic representation of the simulation causality structure.

This preferential selection leads to the under- or over-representation of specific population subgroups with particular characteristics in the treated sample, resulting in bias in the estimation of treatment effects. When the sampling phase introduces selection bias, the outcome and treatment distributions are always conditioned on the set of characteristics driving this preferential selection (Bareinboim and Pearl, 2012). Alternatively, conditioning on a collider introduces a spurious correlation between the outcome and treatment.

To introduce sample selection bias into our base DGP, we stratify the sample based on S, a continuous variable reflecting the probability of entering the data pool (Bareinboim and Pearl, 2012). We then include a control variable Z to mitigate[10] the bias resulting from sample selection (Cinelli, Forney and Pearl, 2022) (Figure 1). By stratifying the sample on S, we can better simulate real-world scenarios, where researchers are often provided with datasets potentially affected by sample selection.

In order for S to be a collider on the causal path between the treatment and the outcome, causal paths must exist from both the treatment and the outcome to S. Therefore, we start by computing the intermediate $\dot{S}$, the outcome predictors (X_{OP}) contribution to S, as in Equation 8.

$$\dot{S} = \omega_1 X_{OP_1} + \omega_2 X_{OP_2} + \omega_3 X_{OP_3} + \omega_4 X_{OP_4} + \omega_5 X_{OP_5} \tag{8}$$

In Equation 8, each X_{OP_p} represents a variable in X_{OP}, while ω_p denotes the coefficient associated with each variable. We select the coefficients ω_p ensuring that the signs correspond with vector α in Equation 5 for each X_{OP}. Then, we construct Z as a mediator on the path between the treatment T and the collider S (Equation 9), meaning that treatment T has an impact on S only through Z (Cinelli, Forney and Pearl, 2022). U is a random uniform variable having the same minimum and maximum values as $\dot{S}$, representing the part of the mediator independent from T:

10 It is important to note that our analysis focuses on a specific instance of sample selection bias that can be corrected using Z, although this may not always be the case (Bareinboim and Pearl, 2012).

$$Z = TU + (1 - T)(-1)U \tag{9}$$

Last, we combine the two components to obtain S (Equation 10), ensuring on one end that S is caused by the treatment and that by controlling for Z we are able to identify the treatment effect.

$$S = 0.5\dot{S} + 0.5Z \tag{10}$$

To introduce sample selection bias in our simulation, we stratify our dataset dropping all the observations below the first 33 per cent quantile of S. This approach ensures that observations with higher values of S, and consequently X_{OP}, and Y are overrepresented in our treated sample. Since our procedure to introduce sample selection bias removes a third of the observations, we also randomly discard a third of the observations in those DGPs where sample selection bias is absent to maintain comparability. As a result, the final sample sizes are 57,620 observations (from 81,834) for the larger dataset and 2814 from (4000) observations for the smaller dataset. We selected the first 33 per cent quantile as it strikes a balance between introducing sufficient sample selection bias (Appendix B) while avoiding dropping too many observations.

Sample selection might be serious issues for impact studies using EU FADN data. While the EU FADN dataset represents around 90 per cent of the total EU production, it only represents around 42 per cent of the holdings—with small part-time farmers being under-represented (Bradley and Hill, 2016). When evaluating AES impact, sample selection bias might play an important role as smaller and less intensive farmers are often the ones more likely to opt into such subsidies. In contrast, larger, more economically driven farms may be less inclined to participate (Zimmermann and Britz, 2016). In this context, the variable Z can be seen as a mediator, such as farm production orientation, distinguishing between economically or environmentally oriented farms. Failing to account for sample selection bias could lead to over-estimated policy effects and hence the policy instrument failing to produce an additional effect in practice by encouraging more intensive farmers to adopt greener practices. Instead, it may also result in windfall effects, where funds are transferred for the adoption of practices that would have been adopted even without the treatment (Chabé-Ferret and Subervie, 2013).

4.6. Introducing latent confounding

The base DGP is modified to violate the unconfoundness assumption by adding the variables set $\widehat{X_{LC}}$ in both the outcome equation and treatment assignment. Therefore, in the DGPs where latent confounding is present, it is possible to rewrite Equations 6 and 7 as Equations 11 and 12, respectively:

$$Y = t\left(\alpha X_{OP} + \beta X_{OC} + \gamma \widehat{X_{LC}}\right) \tag{11}$$

$$\pi(x) = \Pr(Z = 1) = \frac{1}{1 + e^{-C}}$$
$$C = t\left(\kappa_1 X_{OC} + \kappa_2 \widehat{X_{LC}} + \varepsilon\right) \tag{12}$$

Where κ_1 and κ_2 are the sets of coefficients associated with X_{OC} and $\widehat{X_{LC}}$, respectively. By setting t equal to the identity function, it is still possible to retrieve the original linear specification.

The presence of latent confounding poses a significant challenge in all empirical applications, making selection on observables probably one of the least favoured assumptions in economics. As discussed in Section 4.2, impact-assessment studies in agricultural economics intersect multiple disciplines, suggesting that numerous variables influencing a process may remain unobservable. However, the interdisciplinary nature of these linkages might offer potential for finding observed variables sufficiently connected to those unobservables to approximate them. This prospect is supported by the expanding availability of large datasets and advancements in causal ML methods capable of leveraging them. Consequently, it is important to ascertain the capability of these ML techniques in addressing latent confounding and its implications.

4.7. Tree-based treatment assignment

To further evaluate the methods' dependence on the presence of an underlying true propensity score, we replicate the previously described DGPs[11] with treatment assignment using a tree-based model. The tree-based treatment assignment starts with an unsupervised clustering procedure of all the observations in the FADN based on X_{OC}. The clustering rules are then learned with a random forest algorithm (Breiman, 2001) and are consequently used to discriminate among observations that participate or not in the schemes. Rather than opting for a single decision tree, we chose to utilize a random forest method to avoid arbitrary decisions regarding treatment distribution and to eliminate any concerns that the choice of a specific tree could bias the performance evaluation of the methods.

The two treatment assignment procedures reflect two distinct perspectives on how farmers make decisions regarding their enrolment in a particular policy intervention. On one hand, the logit assignment represents a scenario in which farmers are aware of how their farm structure (elements in X_{OP}, X_{OC} and X_{LC}) interacts with the treatment. Consequently, they derive a probability of enrolling in the intervention based on these considerations. However, the final decision may be influenced by other factors, such as ethical or political beliefs (ε in Equation 7). This means that even farmers with lower probabilities of enrolling might still decide to do so, and vice versa. On the other hand, the tree-based assignment represents a scenario in which farmers construct a mental map based on their farm structure, and this map deterministically guides their

11 Except for the DGPs obtained by changing the level of overlap, this is feasible and only makes sense when there exists an underlying true propensity score.

Table 2. Causal properties varied in the simulation and respective settings

Causal property	Levels	
	Easy	Complex
Degree of non-linearity	Linear	Complex
Common support	90% of treated	50% of treated
Strength of effect	0.3	0.15
Sample selection bias	Absent	Present
Latent confounding	Absent	Present

decision on whether to participate or not. Both assumptions can be considered realistic and are therefore put to the test. This approach allows researchers to make reasoned arguments about which sorting mechanism is likely to be more relevant depending on the specific policy tool under consideration, and consequently, they can choose the appropriate method accordingly.

4.8. Running the simulation

To summarize, Figure 1 represents the causality structure we establish in our simulation, while Table 2 provides a brief recap of the causal properties and related settings explained so far, including their possible levels.

In Figure 1, X_{OD}, X_{TD}, X_{OP}, X_{OC}, and X_{LC}, are the variables sets introduced in Section 3.2. T represents the treatment (Equation12), Y represents the outcome (Equation 11), S represents a covariate necessary to introduce sample selection bias (Equation 10), and Z is a control placed to address sample selection bias (Equation 9).

These combinations are compared under two treatment assignment procedures based on two different assumptions (and consequently models). The first procedure mirrors an assignment based on a propensity score, implying that farmers have a certain probability of participating in a treatment based on their characteristics. The second procedure mirrors a scenario in which the farmers decide on whether to apply for the intervention based on a tree-based scheme. Due to the impossibility of controlling the share of common support under the tree-based schemes, this causal property is disregarded under this scheme.

To compare the performances of the considered methods, we run each of them on the 32 DGPs for the propensity-based and the 16 DGPs for the tree-based assignment obtained from all the combinations of causal properties levels.[12] Each combination of method and DGPs for the large sample size is run 50 times to ensure an accurate representation of bias (Huber, Lechner and Wunsch, 2013; Knaus, Lechner and Strittmatter, 2021; Wendling *et al.*, 2018; Advani, Kitagawa and Słoczyński, 2019; Künzel *et al.*, 2019; Parikh *et al.*, 2022). Each run involves drawing a new sample data from the FADN through

12 We have 32 for the propensity-based assignment, while 16 for the tree-based assignment since common support cannot be controlled in the latter case.

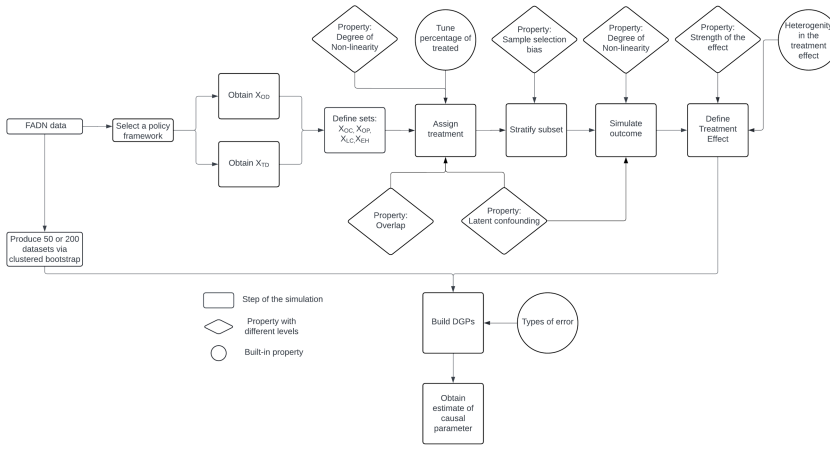

Fig. 2. Simulation steps.

a clustered bootstrap at the NUTS2 region level, while maintaining the original sample size. For the smaller sample size, we adhere to the same procedure, but each combination of method and DGP is executed 200 times to account for the reduced sample size. Figure 2 provides a summary of the steps involved in this simulation, with an indication of where the causal properties of the DGPs can be manipulated.

5. Methods

In this section, we outline the various evaluation methods used. For each method, we briefly explain the underlying mechanism, highlighting their strengths and drawbacks, which will later be used in assessing their performances. Despite their differences, all the methods require three basic assumptions to make their estimates causal: unconfoundedness, common support and SUTVA.

5.1. Classical matching methods

5.1.1. Propensity Score Matching

PSM (Rosenbaum and Rubin, 1983) is commonly used as a subset selection procedure[13] to mitigate confounding in observational settings in economics (Athey and Imbens, 2017; King and Nielsen, 2019). It matches treated and control observations based on their propensity score, being the probability of an observation to receive the treatment conditional on its covariates. The key in PSM is Rosenbaum and Rubin's Theorem 3 (T3) stating that unconfoundedness based on raw covariates implies unconfoundedness on the propensity score (assuming common support and SUTVA). Following T3, the biggest

13 See Appendix C for a performance comparison between PSM subsetting and re-weighting procedures.

advantage of this method is that it avoids the curse of dimensionality. When matching over multiple dimensions, it becomes increasingly difficult to find observations which are similar across all of them. The researcher is then left either with bad matches or a drastically smaller dataset.

However, the way this method has been traditionally applied presents several drawbacks (Ho *et al*., 2007; Hill, 2008; Austin, 2009) and in recent years also its underlying mechanism has been criticized (King and Nielsen, 2019). There is a fundamental contrast between the reason a researcher would apply matching and the underlying theoretical properties of PSM. Matching is a non-parametric procedure used to reduce the model dependence and discretion bias, two consequences of making functional form assumptions during estimation (Hill, 2008). However, T3 holds true only for the true propensity score—which would require us to know the true treatment assignment process. Since this is rarely the case (Ho *et al*., 2007; King and Nielsen, 2019), the researcher is left with making functional form assumptions—which as a consequence introduces bias.

Despite its weaknesses, we decided to focus on two different versions of PSM due to its huge popularity. The two specifications differ in their underlying mechanism as well as the function used in computing the propensity scores. The first is the bias-corrected missing value imputation procedure proposed by Abadie and Imbens (2011), employing a logistic function. The second alternative is a matching procedure employing Bayesian additive regression trees (BART). We focus on the first version as it is used in most applied work (Austin, 2009; King and Nielsen, 2019), while we selected the BART version as a possible way to mitigate model dependence, being non-parametric and highly flexible (Lee, Lessler and Stuart, 2010; Westreich, Lessler and Funk, 2010). Both approaches are based on nearest-neighbor (NN) matching, as it is the most widely applied procedure (Austin, 2009) and because asymptotically all matching algorithms should yield the same result (Caliendo and Kopeinig, 2008). Trimming rules commonly studied in other simulation studies are not considered in our simulation because they are known to be problematic with heterogenous treatment effects (Busso, DiNardo and McCrary, 2009). To code the models, we relied on the MatchIt R package (Ho *et al*., 2011). After matching, following Ho et al. (2007), we regress the outcome on the treatment indicator and observed confounders, identifying the ATT.

In addition to the two specifications above, we added a humble PSM model, where at random (across the 50 iterations) two controls are removed from the specification. While this estimator is expected to always fail to retrieve the true ATT, it remains interesting to assess both the size and sign of the bias occurring in an FADN-like setting.

The PSM models are expected to perform well in limited common support scenarios (Busso, DiNardo and McCrary, 2009, 2014), and the version exploiting a BART model for the estimation of the scores should be able to tackle non-linear DGPs (Hill, Weiss and Zhai, 2011).

5.1.2. Coarsened exact matching

In response to the shortcomings of PSM, Iacus, King and Porro (2011; 2012) introduced a method to address these issues and surpass PSM's performance. This method, known as coarsened exact matching (CEM), has gained significant traction in recent years and has been recognized for its effectiveness (King *et al*., 2011). CEM consists of two steps: in the first one, variables are coarsened in bins based on domain knowledge or algorithmically; in the second step, the bins are exactly matched. More specifically, variables are coarsened so that indistinguishable values (from an analytical perspective) are grouped together. This step facilitates finding exact matches—as the definition of exact gets broader the more the data are coarsened—even on continuous variables, while avoiding dropping too many observations.

CEM features two major advantages over PSM, allowing it to deal better with imbalances in covariates. First, CEM guarantees covariates imbalance reduction at the expense of sample size—the opposite of what PSM entails (Iacus, King and Porro, 2011). The maximum imbalance for a given variable depends on how fine it is coarsened, where higher coarsening leads to larger imbalance. Second, CEM allows the researcher to set the maximum level of unbalance for each covariate independently. In PSM, different specifications will lead to different balancing in an unpredictable way.

The major drawback in applying CEM is that despite the coarsening it discards all the observations not exactly matched. This leads to a decrease in sample size, which implies reductions in power and precision. Iacus, King and Porro (2012) defend their method following Rubin (2006), stating that unbiasedness comes before efficiency, and that in observational studies functional form assumptions are a bigger problem than sample size. However, as noted by Black, Lalkiya and Lerner (2020), CEM drops observations in a non-obvious way, misidentifying average and heterogeneous effects.

Following Iacus, King and Porro (2012), we used domain knowledge (meaning consistency with the framework we simulate) to coarsen factor variables and applied Sturges' (1926) algorithm for continuous variables. Also in this case, we coded the model relying on the MatchIt R package (Ho *et al*., 2011).

5.2. Tree-based matching methods

5.2.1. Causal forests

A causal forest (CF) is an algorithm built on top of the classic random forests (RFs) (Breiman, 2001) and belonging to the family of generalized random forests (GRFs) (Athey, Tibshirani and Wager, 2019). The basic building block of a RF is the decision tree, which is an algorithm that partitions the variable space to obtain clusters as homogenous as possible. From that, RFs are obtained as ensembles of several decision trees—in order to reduce the variance of the estimate.

While RFs are limited to expected outcomes, thanks to an adaptation of the optimization criterion GRFs allow for the estimation of other quantities (for

instance, a causal estimand for CFs—assuming all identifying assumptions are met). The algorithm looks for the partitions of the variable space, which maximizes the heterogeneity of the target estimand between different regions.

In our analysis, we consider two different CF specifications: one in which the input variables are selected with a data-driven approach by the CF (Data—CF), and another in which we feed the CF the same subset of variables we feed to classical models (Theory—CF). Thus, it is possible to test the extent to which it is possible to rely on the CF in scenarios where the choice of the controls to be employed cannot be guided completely by theory.

In our analysis, we implement the CFs through the grf package in R (Tibshirani *et al.*, 2022), which relies on three augmentations compared to standard approaches. First, the estimation is conducted on the residualized outcome and treatment, in order to increase the efficiency and reduce bias. The orthogonalization follows Robinson (1988), and relies on marginal outcomes and propensity scores computed via separated RFs. Second, the trees are grown honestly (to avoid overfitting): observations used to build the structure of the tree cannot be used to make predictions. Last, average treatment effects are not obtained as averages of the ITEs but are rather plugged in a doubly robust estimator following Chernozhukov et al. (2018). The values for the tunable hyper-parameters for each CF are obtained via cross-validation. Last, our target estimand, to maintain comparability with classical methods, is the ATT.

A limitation of the CF method is its inability to deal with categorical variables if not encoded suitably for a sparsity-seeking algorithm (Johannemann *et al.*, 2019). Since this assumption is violated in our framework, we preprocess the data through a cross-validated target-encoding procedure (Appendix D).

5.2.2. Bayesian additive regression trees

BART is a tree-based method relying on Bayesian inference (Chipman, George and McCulloch, 2010). Conceptually, its frequentist analogue would be a boosted RF, as subsequent trees are fit on the residual of the previous tree. However, this procedure is only effective as long as we are able to avoid overfitting—i.e. picking up contingent[14] noise. Therefore, it is important to limit the contribution that each individual tree has in explaining the overall variability of the response. To do so, tree-based methods (including CF) rely on the hard choice of two hyper-parameters values via cross-validation: one governing the number of splits in a given tree (its depth), and the other being a penalty shrinking the fit of the tree. BART, thanks to its Bayesian framework, employs regularization prior only, encouraging the trees to be small. Overall, the Bayesian framework allows a principled approach to regularization, leading to flexible trees that can fit complex functions. However, while also frequentist tree-based methods have this flexibility, the Bayesian framework

14 Since the training of the ML models takes place in batches, contingent here refers to the specificities of each batch as opposed to the properties of the overall DGP.

also allows the quantification of uncertainty, to compute coherent confidence intervals.

Similarly to CF, BART's main advantages are that it naturally identifies heterogeneous effects and its flexibility in estimation—which is however combined with an easier usage (Carnegie and Wu, 2019). The default regularization prior has already been developed by Chipman, George and McCulloch (2010), who found it to be robust in a wide range of settings. Therefore, it allows the researcher to avoid the computational step of cross-validating the hyperparameters (Hill, 2011). Moreover, it is able to handle a larger number of controls than classical methods.

We deploy our BART model without cross-validating it (off-the-shelf), since one of its most praised abilities is the versatility of its regularization prior (Hill, 2011; Hill, Weiss and Zhai, 2011; Hahn, Murray and Carvalho, 2020). Similarly to the approach we use with CF, we specify two different BART models, trained on either a subset of variables selected by the algorithm (Data—BART) or the same variables used in classical methods (Theory—BART). Therefore, we specify our models following Hahn, Dorie and Murray (2019) who adapt BART to causal settings and exploit the bartCause R package (Hill, 2011).

5.3. Computation

To conduct our analyses, we utilized a high-performance computing cluster and partitioned the computations into smaller batches based on the required computational resources. Specifically, for each of the 9 models, we divided the 1600 runs (32 DGPs with 50 cluster-bootstrapped samples each) across 50 distinct nodes. This procedure was repeated for each treatment assignment scheme and each sample size. Each node provided computational resources as follows: 8GB of RAM for the three PSM-based and CEM models, 16GB of RAM for the CFs models and 64GB of RAM for the BART models. The entire procedure required a total runtime of 120 h.

5.4. Evaluation metrics

To compare our estimators, we rely[15] on the absolute bias in the ATT (Equation 13), being the mean of the absolute value of the percentage deviation of the causal estimate $\widehat{\theta_{ATT_j}}$ from its correct value across m replications.

$$absolute\,bias_{ATT} = \frac{1}{m}\sum_{j=1}^{m}\left|\frac{\widehat{\theta_{ATT}} - \theta_{ATT}}{\theta_{ATT}}\right| \times 100 \tag{13}$$

To assess the effectiveness of ML methods in estimating CATT, we present a direct comparison of true effects versus estimated ones. Utilizing a metric that aggregates the effects might obscure the nuances of how well the methods approximate the distribution of the treatment effect.

15 For additional metrics describing the comparison please refer to Appendix E.

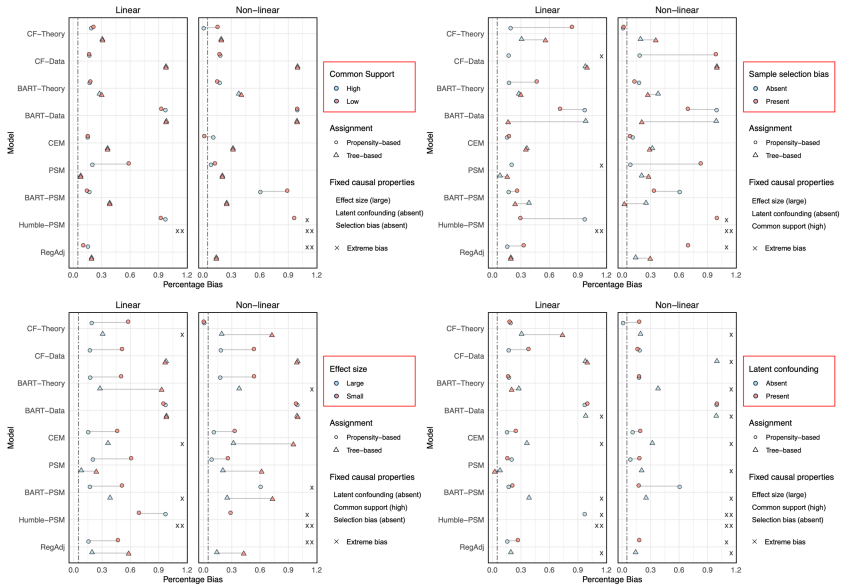

Fig. 3. Percentage bias across different causal properties.

6. Results and discussion

6.1. Treatment effect across causal properties

Figure 3 presents the average percentage bias scored by each model across various causal property levels, both for propensity-based and tree-based treatment assignments. Within each subplot, three causal properties are held constant at the easier level, illustrating the variation in percentage bias when transitioning from the easier to the more challenging level of the remaining causal property in both linear (left box) and non-linear (right box) scenarios. In each box, the vertical line represents the 5 per cent bias threshold that we use to determine whether a method has sufficiently satisfying performances. Then, to keep the illustration compact, when a model percentage bias surpasses 1, we replace it with a cross placed close to its box right border.

To illustrate interpretation, the top left subplot shows the difference in average percentage bias as the overlap changes from low to high across linear and non-linear scenarios—limiting the analysis in DGPs where the effect size is large, and both latent confounding and sample selection bias are absent. The rationale behind presenting results in this manner is to offer a concise representation of the relevance of specific causal properties and when they come into play, while minimizing the aggregation of bias stemming from different causal properties as much as possible. As a reference point, the light blue dots in the linear box represent the base DGP and therefore are in the same positions across the four subplots. Similarly, the light blue dots in the non-linear box represent the non-linear version of the base DGP. What changes within each

subplot is the position of the more challenging level dots. Therefore, comparing the position between the light blue dots in the linear and non-linear box within a single subplot allows us to understand the effect of non-linearity. Comparing the relative distance between the dots in the linear and non-linear boxes within a single subplot illustrates the interaction effect between the considered causal property and the degree of non-linearity. Finally, comparing the absolute positions of the orange dots in respect to the light blue dots gives the impact of moving from a scenario with no particular hindrances to the estimation to one where we expect serious bias from the considered causal property. For a more complete presentation of the results, Appendix E contains tables reporting all biases from the various DGPs for each model. Appendix F includes supplementary results from the simulation analysis carried out on a smaller sample size. Although the results for the smaller sample are not graphically presented in the main body to maintain brevity, they are briefly discussed in the following section, and the corresponding images are provided in Appendix F. In Appendix H, we compare results at different levels of common support. Last, Appendix G reports bias-variance plots for each method over all scenarios.

Examining the impact of common support (Figure 3—top left subplot), it appears that it only has an impact on the propensity score-based methods, affecting some exclusively in the context of linear DGPs and others solely in nonlinear scenarios. This pattern persists when considering a smaller sample size, as indicated in the supplementary materials (Appendix F).

Examining the treatment effect magnitude (Figure 3—bottom left subplot), we observe that, for each method (excluding the Data-BART), distinct levels of treatment effects correspond to varying bias levels. Notably, two scenarios highlight pronounced differences: in the case of ML models, particularly with propensity-based treatment and linear DGPs, and for classical models when applied on tree-based assignment with non-linear DGPs. Comparing these findings with those obtained from a smaller sample size reveals a generally similar pattern, albeit with some nuanced differences among models. Specifically, CFs encounter challenges in tree-based assignments and non-linear DGPs, whereas BARTs demonstrate overall improved performance. Conversely, propensity-based methods appear more adept at handling non-linear scenarios.

Exploring sample selection bias (Figure 3—top right subplot), we observe distinct responses among various methods. Notably, the disparities in performance are more pronounced in the propensity-based treatment assignment compared to the tree-based assignment. In the former, theory-driven ML methods exhibit proficiency in handling sample selection bias within non-linear DGPs, whereas data-driven ML models struggle with this challenge. However, it is worth noting that data-BART outperforms theory-BART in this specific instance. A possible explanation is that BART, being a sparsity-seeking algorithm, tends to discard Z (the control needed to address sample selection bias) when fed only the relevant variables (theory-BART). Among X_{OP} and X_{OC}, Z

has the lower predictive power for both the treatment and the outcome. However, when BART is provided with the entire set of variables, Z is more likely to be retained because it has greater predictive power than the irrelevant controls (data-BART). Classical models, in general, prove susceptible to sample selection bias, with many estimations displaying significant bias. A noteworthy exception is CEM, which consistently manages sample selection bias across all four scenarios. For the smaller sample size, we observe an overall reduction in bias attributed to sample selection bias. This reduction is particularly evident for classical models, which, while not yet achieving a fully satisfactory level of handling sample selection bias, demonstrates improvement in their performance. We believe this to be a consequence of how sample selection bias is introduced: while we drop the same percentage of units in both the large and small samples, the relative number of treated units discarded changes. This happens because of the stochastic component of S: in larger samples, there is a higher possibility that control units happen to have higher S, leading to more treated units being discarded and thus exacerbating the overrepresentation of high S—high Y observations.

A notable contrast emerges in method performances when latent confounding is introduced (Figure 3—bottom right subplot). Concerning the tree-based assignment, no method demonstrates efficacy in handling latent confounding. This is particularly pronounced when coupled with non-linear DGPs. For propensity score-based assignment, while the overall bias levels are lower, most models still exhibit disappointing performance. Remarkably, two models display notable robustness to sample selection bias: BART, consistent with findings by Hahn, Dorie and Murray (2019), and the bias-corrected PSM. Upon conducting an analysis with a smaller sample size, the inability of any model to address latent confounding persists when the assignment is tree-based. However, it becomes evident that the overall bias induced by latent confounding has a smaller impact on propensity-based treatment assignment.

Finally, it is worth noting a few points when discussing the importance of the underlying treatment assignment assumption. First, PSM tends to perform relatively well with the tree-based assignment in respect to the propensity-based assignment, but mostly if the forest does not include interaction and higher order terms. A possible reason for this lies in random forests functioning, clustering similar observations into the same group and assigning all of them the same probabilities. Consequently, overlap is enhanced while the scores variability is decreased, resulting in a simpler setting for PSM. Conversely, ML methods tend to perform better under the tree-based assignment when the setting is more complex (for instance with complex non-linear variable space).

In general, the performance of the models raises concerns, as the majority of methods consistently fail to stay below the 5 per cent bias threshold. The standout performer in this evaluation is the Theory-CF, demonstrating consistent accuracy in retrieving the correct treatment effect in non-linear DGPs, except when faced with low common support or latent confounding. A parallel scenario unfolds in the comparison involving a smaller sample size, with the

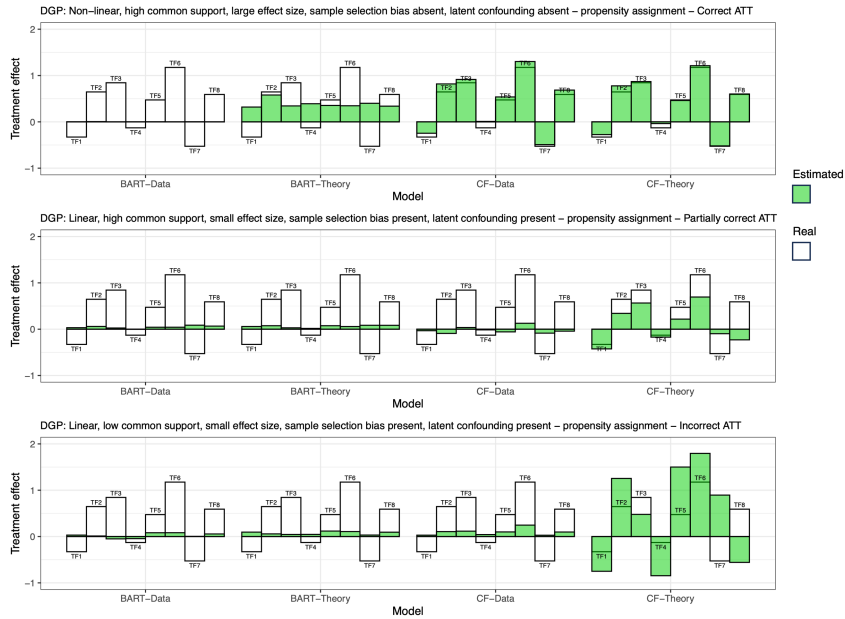

Fig. 4. CATT estimation.

notable difference that the best performance is by Data-BART. Despite grappling with challenges related to sample selection bias, latent confounding and treatment effect size (excluding common-support), Data-BART consistently retrieves the correct effect under these conditions.

6.2. Treatment effect heterogeneity

Building upon the preceding results, there are three distinct scenarios warranting investigation concerning the accuracy of CATT: when the method successfully retrieves the ATT, when the ATT is partially biased, and when the ATT is entirely biased. The main question revolves around whether the model accurately discerns the CATT and aggregates them to retrieve the correct ATT, or alternatively, it explores the possibility of a biased ATT resulting from an incorrect aggregation of the CATT, even if the latter are reasonably approximated. In Figure 4, we present the estimated CATT versus the actual CATT for three DGPs, each representing one of the aforementioned scenarios. Our focus remains on DGPs where the theory-CF can successfully retrieve the ATT, given its unique capability in this regard.

Our findings emphasize that when the CF is precisely specified and successfully retrieves the correct CATTs, it concurrently captures the correct ATT. Even small deviations from the correct CATTs result in significant biases in the estimation of the ATT. A noteworthy example is the data-driven CF, which remarkably performs well in predicting the CATTs despite being moderately

distant from the ATT. On the other hand, neither the data-driven nor the theory-driven BART can successfully retrieve the CATT in any scenarios. As the bias in ATT increases, a corresponding bias is consistently observed in CATTs across all DGPs, as detailed in Appendix E. Examining the results for the smaller sample size, we observe a similar phenomenon, albeit with the notable difference that the data-driven CF performs significantly better than in previous scenarios, as outlined in Appendix F. It is important to note, however, that the hyperparameters of the BART models were not optimized, leaving for further research whether such optimization could have improved model performance.

7. General discussion

This simulation study compares the performance of ML methods with classical econometrics for causal inference using observational data. The first observation from the results is that classical methods, whether parametric or non-parametric, consistently fail to retrieve the true causal parameter in non-linear DGPs. This underscores the potential risks associated with assuming linearity, leading to unrealistic dynamics in intervention outcomes. For instance, assuming constant marginal values for AES subsidies may be inappropriate, given the non-constant marginal costs of farms of varying sizes. Therefore, the assumption of linearity should be approached cautiously, and its implications should be contrasted with empirical knowledge. Our results indicate that in cases where linearity assumptions do not hold, ML methods are preferable.

Second, comparing performances under high or low common support initially suggested that methods were not significantly affected by this property. However, recognizing the common support assumption required by each model, we further investigated the issue by comparing results at different levels of support, reported in Appendix H. What we find with the new simulation design is that ML methods maintain nearly a constant level of bias across different shares of common support, indicating their robustness to low overlap as long as it exists. In contrast, classical methods exhibit a significantly higher bias below a certain threshold. Specifically, the bias-corrected PSM registers an almost fourfold increase in bias when the share of common support falls below 50 per cent. Similar trends are observed for BART-PSM and CEM when the share of common support drops below 25 per cent. These results, however, should be viewed as preliminary, and a dedicated simulation is required to thoroughly examine how different shares of common support influence the identification of a causal effect.

Third, comparing results under large or small treatment effects revealed that each method exhibits different power characteristics. Estimates of smaller treatment effects were more biased for every method compared to larger treatment effects. Consequently, a power analysis using the chosen methods is imperative to ensure the capability of detecting effects based on anticipated intervention results or targets, even if it entails running computationally intensive models (Ioannidis, Stanley and Doucouliagos, 2017).

Fourth, comparing estimates in the presence or absence of sample selection bias reveals the susceptibility of both ML and classical methods when incorrectly specified. Sample selection bias exacerbates existing biases when other factors hinder the correct treatment assignment retrieval. Conversely, when provided with the correct controls, the model can address sample selection bias if capable of retrieving the treatment effect. While direct policy implications are absent, the results emphasize the importance of selecting an appropriate model for the analysis, underscoring the significance of simulation studies.

Fifth, considering latent confounding, neither ML nor classical methods prove suitable for addressing it. Despite the potential mitigation of bias in simple DGPs by providing a correct proxy for latent confounders, latent confounding remains a persistent issue without a clear solution. This serves as a reminder that any analysis based on the FADN is constrained by the control variables present within it.

Finally, our analysis shows substantial differences in performance when the underlying treatment assignment is assumed to follow either a probability-based or a decision tree procedure. The performance of each method undergoes drastic changes when transitioning between these two treatment assignment procedures. In linear DGPs, a bias-corrected PSM is sufficient to retrieve the true treatment effect. However, in non-linear DGPs, theory-based counterfactuals are shown to be the best performers.

8. Conclusions

To evaluate agricultural policies, economists used a variety of econometric impact evaluation techniques the last few decades. More recently, ML techniques have been proposed as a panacea for all their ills. In this simulation study, we compared the performances of classical econometric methods versus ML methods in retrieving a causal estimand across different DGPs tuned to the European FADN. We tested four main advantages ML methods are supposed to have derived from their data-driven approach: functional form selection, relevant controls selection, treatment effect heterogeneity exploration and latent confounding control. Accordingly, we first assessed the reliability of each method over a wide range of scenarios, and second, investigated how ML methods can enhance causal analysis for agricultural policy evaluation.

Our findings reveal that, on the whole, ML methods exhibit superior performance compared to classical methods. Specifically, the theory-driven CF is the best performing model in large sample sizes, while the theory-driven BART excels in small sample sizes. The success of the CFs might be due to the doubly robust correction they employ, differently from the other methods. The success of the data-driven approach to approximate the functional form is particularly noteworthy, as ML methods perform significantly better on non-linear DGPs. Surprisingly, this advantage does not extend to situations where the underlying treatment assignment is assumed to follow a tree-based procedure. However, the data-driven variable selection does not consistently enhance performance and can even detrimentally impact both CF and BART. While these

forest methods identify relevant controls, the relative importance of these controls is obscured by other variables that exhibit higher correlation with the magnitude of the treatment effect rather than the assignment. A notable exception is observed with data-driven BART, which, in small sample sizes and only in certain scenarios, performs similarly or better than its theory-driven counterpart, underscoring the specific advantages of a data-driven approach in certain contexts. The most substantial contribution of ML methods, setting them significantly apart from classical methods, lies in their capability to correctly capture complex and heterogeneous treatment effects when appropriately specified. In this regard, CF appears better equipped than BART, making a considerable contribution to the toolbox of agricultural economists. Finally, our results suggest that the ability of tree-based methods to handle latent confounding does not solely hinge on providing the correct variables to generate control combinations.

There are three limitations to this study. First, the results of our models are only valid over the causality structures specified. In particular, we focus on cross-sectional methods relying on the conditional independence assumption, while we leave for future research the possibility of extending this framework to panel data (to leverage methodologies based on the DiD framework for instance) or designing scenarios resembling natural experiments. Second, we chose to focus on a subset of relevant properties that could pose challenges to identification in a causal study. Third, despite our efforts to make the simulation framework as general as possible, there are still some arbitrary choices, such as the number of variables included in each DFP or the specification of the non-linear functional form. While these limitations may negatively affect the generalizability of our results, we are confident that our findings can be extended across different scenarios since the causality structure we have specified encompasses a wide range of scenarios, and the entire framework allows for both implementations and modifications.

Further research on this topic could investigate the impact of framework modifications, such as the consideration of different causal properties. Another direction could be implementing additional preprocessing procedures to make the ML estimators more effective. Considering BART, a question that naturally follows our analysis is assessing the impact cross-validating its hyperparameters would have on its performances. Moving to CFs, another open question remains the design of preprocessing procedures to ensure their correct usage by the tree-based methods. The objective of these preprocessing procedures is to ensure that controls known from theory to be relevant have priority over remaining variables. Finally, future research could explore exploring the capabilities of ML models to deal with sample selection bias due to unobservables.

Acknowledgements

We would like to thank DG AGRI for providing access to the EU-FADN data (agreement IFD 2022_08).

Supplementary data

Supplementary data are available at *ERAE* online.

References

Abadie, A. and Imbens, G. W. (2011). Bias-corrected matching estimators for average treatment effects. *Journal of Business & Economic Statistics* 29(1): 1–11.

Advani, A., Kitagawa, T. and Słoczyński, T. (2019). Mostly harmless simulations? Using Monte Carlo studies for estimator selection. *Journal of Applied Economics* 34(6): 893–910.

Arata, L. and Sckokai, P. (2016). The impact of agri-environmental schemes on farm performance in five E.U. member states: a DID-matching approach. *Land Economics* 92(1): 167–186.

Athey, S. (2018). The impact of machine learning on economics. In: Agrawal A., Gans J. and Goldfarb A (eds), *The Economics of Artificial Intelligence: An Agenda*. Chicago: University of Chicago Press, 507–547.

Athey, S. and Imbens, G. W. (2017). The state of applied econometrics: causality and policy evaluation. *Journal of Economic Perspectives* 31(2): 3–32.

Athey, S. and Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics* 11: 685–725.

Athey, S., Tibshirani, J. and Wager, S. (2019). Generalized random forests. *The Annals of Statistics* 47(2): 1148–1178.

Austin, P. C. (2009). Some methods of propensity-score matching had superior performance to others: results of an empirical investigation and Monte Carlo simulations. *Biometrical Journal* 51(1): 171–184.

Baiardi, A. and Naghi, A. A. (2021). The value added of machine learning to causal inference: evidence from revisited studies. *arXiv preprint arXiv:2101.00878*.

Bareinboim, E. and Pearl, J. (2012). Controlling selection bias in causal inference. In: Lawrence N. D. and Girolami M (eds), *Artificial Intelligence and Statistics*. La Palma, Canary Islands: PMLR 22: 100–108.

Baylis, K., Peplow, S., Rausser, G. and Simon, L. (2008). Agri-environmental policies in the EU and United States: A comparison. *Ecological Economics* 65(4): 753–764.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple hypothesis testing. *Journal of the Royal Statistical Society* 57: 289–300.

Bennett, A. and Kallus, N. (2019). Policy evaluation with latent confounders via optimal balance. *Advances in Neural Information Processing Systems* 32: 4826–4836.

Black, B. S., Lalkiya, P. and Lerner, J. Y. (2020). The trouble with coarsened exact matching. *Forthcoming in Northwestern Law & Econ Research Paper*.

Bonferroni, C. (1936). Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del Regio Istituto Superiore di Scienze Economiche e Commerciali di Firenze* 8: 3–62.

Börner, J., Baylis, K., Corbera, E., Ezzine-de-blas, D., Honey-Rosés, J., Persson, U. M. and Wunder, S. (2017). The effectiveness of payments for environmental services. *World Development* 96: 359–374.

Bradley, D. and Hill, B. (2016). Diversity and Innovation in the FADN Data Collection Systems in the EU-28. *EuroChoices* 15: 5–10. 10.1111/1746-692X.12137

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

Busso, M., DiNardo, J. and McCrary, J. (2009). Finite sample properties of semiparametric estimators of average treatment effects.

Busso, M., DiNardo, J. and McCrary, J. (2014). New evidence on the finite sample properties of propensity score reweighting and matching estimators. *Review of Economics and Statistics* 96(5): 885–897.

Caliendo, M. and Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys* 22(1): 31–72.

Carnegie, N. B. and Wu, J. (2019). Variable selection and parameter tuning for BART modelling in the fragile families challenge. *Socius* 5: 2378023119825886.

Chabé-Ferret, S. and Subervie, J. (2013). How much green for the buck? Estimating additional and windfall effects of French agro-environmental schemes by DID-matching. *Journal of Environmental Economics and Management* 65(1): 12–27.

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

Chipman, H. A., George, E. I. and McCulloch, R. E. (2010). BART: Bayesian additive regression trees. *Annuals of Applied Statistics* 4(1): 266–298.

Cinelli, C., Forney, A. and Pearl, J. (2022). A crash course in good and bad controls. *Sociological Methods & Research* 53(3): 00491241221099552.

DG AGRI. (2023). Farm Accounting Data Network. https://agridata.ec.europa.eu/extensions/FarmEconomyFocus/FADNDatabase.html. Accessed 20 December 2022.

Dorie, V., Hill, J., Shalit, U., Scott, M. and Cervone, D. (2019). Automated versus do-it-yourself methods for causal inference: lessons learned from a data analysis competition. *Statistical Science* 34(1): 43–46.

European Commission. (2022). Strategic Plans and Commissions Observations. *Directorate-General for Agriculture and Rural Development. Publications Office*.

Frölich, M. (2004). Finite-sample properties of propensity-score matching and weighting estimators. *Review of Economics and Statistics* 86(1): 77–90.

Hahn, P. R., Dorie, V. and Murray, J. S. (2019). Atlantic causal inference conference data analysis challenge 2017. *arXiv preprint arXiv:1905.09515*.

Hahn, P. R., Murray, J. and Carvalho, C. (2020). Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects. *Bayesian Analysis* 15(3): 965–1056.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning*. New York, NY, USA: Springer New York Inc.

Hill, J. (2008). Discussion of research using propensity-score matching: comments on "a critical appraisal of propensity-score matching in the medical literature between 1996 and 2003" by Peter Austin, Statistics in medicine. *Statistics in Medicine* 27(12): 2055–2061.

Hill, J. (2011). Bayesian nonparametric modelling for causal inference. *Journal of Computational and Graphical Statistics* 20(1): 217–240.

Hill, J., Weiss, C. and Zhai, F. (2011). Challenges with propensity score strategies in a high-dimensional setting and a potential alternative. *Multivariate Behavioral Research* 46: 477–513.

Ho, D. E., Imai, K., King, G. and Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis* 15: 199–236.

Ho, D. E., Imai, K., King, G. and Stuart, E. A. (2011). MatchIt: nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software* 42(8): 1–28.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 6(2): 65–70.

Huber, M., Lechner, M. and Wunsch, C. (2013). The performance of estimators based on the propensity score. *Journal of Econometrics* 175(1): 1–21.

Huntington-Klein, N. (2021). *The Effect: An Introduction to Research Design and Causality*, 1st edn. New York: Chapman and Hall/CRC.

Iacus, S. M., King, G. and Porro, G. (2011). Multivariate matching method that are monotonic imbalance bounding. *Journal of the American Statistical Association* 106(493): 345–361.

Iacus, S. M., King, G. and Porro, G. (2012). Causal inference without balance checking: coarsened exact matching. *Political Analysis* 20(1): 1–24.

Imbens, G. and Rubin, D. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge: Cambridge University Press.

Ioannidis, J. P. A., Stanley, T. D. and Doucouliagos, H. (2017). The power of bias in economics research. *Economic Journal* 127: F236–F265.

Johannemann, J., Hadad, V., Athey, S. and Wager, S. (2019). Sufficient representations for categorical variables. *arXiv preprint arXiv:1908.09874*.

Kallus, N., Puli, A. M. and Shalit, U. (2018). Removing hidden confounding by experimental grounding. *Advances in Neural Information Processing Systems* 31: 10911–10920.

King, G. and Nielsen, R. (2019). Why propensity scores should not be used for matching. *Political Analysis* 27(4): 435–454.

King, G., Nielsen, R., Coberley, C., Pope, J. E. and Wells, A. (2011). Comparative effectiveness of matching methods for causal inference. *Unpublished manuscript, Institute for Quantitative Social Science, Harvard University, Cambridge, MA*.

King, G. and Zeng, L. (2007). When can history be our guide? The pitfalls of counterfactual inference. *International Studies Quarterly* 51(1): 183–210.

Knaus, M. C., Lechner, M. and Strittmatter, A. (2021). Machine learning estimation of heterogeneous causal effects: Empirical Monte Carlo evidence. *The Econometrics Journal* 24(1): 134–161.

Koutchadé, O. P., Carpentier, A. and Femenia, F. (2018). Modelling heterogeneous farm responses to European Union biofuel support with a random parameter multicrop model. *American Journal of Agricultural Economics* 100(2): 434–455.

Künzel, S. R., Sekhon, J. S., Bickel, P. J. and Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences* 116(10): 4156–4165.

Lechner, M. and Wunsch, C. (2013). Sensitivity of matching-based program evaluations to the availability of control variables. *Labour Economics* 21: 111–121.

Lee, B. K., Lessler, J. and Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Statistics in Medicine* 29(3): 337–346.

Louizos, C., Shalit, U., Mooij, J. M., Sontag, D., Zemel, R. and Welling, M. (2017). Causal effect inference with deep latent-variable models. *Advances in Neural Information Processing Systems* 30: 6449–6459.

Massfeller, A., Meraner, M., Hüttel, S. and Uehleke, R. (2022). Farmers' acceptance of results-based agri-environmental schemes: a German perspective. *Land Use Policy* 120: 106281.

Mennig, P. and Sauer, J. (2020). The impact of agri-environment schemes on farm productivity: a DID-matching approach. *European Review of Agricultural Economics* 47(3): 1045–1093.

Millimet, D. L. and Tchernis, R. (2009). On the specification of propensity scores, with applications to the analysis of trade policies. *Journal of Business & Economic Statistics* 27(3): 397–415.

Molnar, C. (2020). Interpretable machine learning. Lulu.com.

Parikh, H., Varjao, C., Xu, L. and Tchetgen, E. T. (2022). Evaluating causal inference methods. *arXiv preprint arXiv:2202.04208*.

Pufahl, A. and Weiss, C. (2009). Evaluating the effects of farm programmes: results from propensity score matching. *European Review of Agricultural Economics* 36: 79–101.

Robinson, P. M. (1988). Root-N-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society* 56: 931–954.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* 70(1): 41–55.

Rubin, D. B. (2006). *Matched Sampling for Causal Effects*. Cambridge: Cambridge University Press.

Schlüter, M., Brelsford, C., Ferraro, P. J., Orach, K., Qiu, M. and Smith, M. D. (2023). Unraveling complex causal processes that affect sustainability requires more integration between empirical and modeling approaches. *Proceedings of the National Academy of Sciences* 120(41): e2215676120.

Stetter, C., Mennig, P. and Sauer, J. (2022). Using machine learning to identify heterogeneous impacts of agri-environment schemes in the EU: a case study. *European Review of Agricultural Economics* 49(4): 723–759.

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

Sturges, H. A. (1926). The choice of a class interval. *Journal of the American Statistical Association* 21: 65–66.

Tibshirani, J., Athey, S., Sverdrup, E. and Wager, S. (2022). grf: Generalized Random Forests. R package version 2.2.0. https://CRAN.R-project.org/package=grf. Accessed 20 December 2022.

Wang, Y. and Blei, D. M. (2019). The blessings of multiple causes. *Journal of the American Statistical Association* 114(528): 1574–1596.

Wendling, T., Jung, K., Callahan, A., Schuler, A., Shah, N. H. and Gallego, B. (2018). Comparing methods for estimation of heterogeneous treatment effects using observational data from health care databases. *Statistics in Medicine* 37(23): 3309–3324.

Westreich, D., Lessler, J. and Funk, M. J. (2010). Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. *Journal of Clinical Epidemiology* 63(8): 826–833.

Yeo, I. K. and Johnson, R. A. (2000). A new family of power transformations to improve normality or symmetry. *Biometrika* 87(4): 954–959.

Zhao, Z. (2004). Using matching to estimate treatment effects: data requirements, matching metrics, and Monte Carlo evidence. *Review of Economics and Statistics* 86(1): 91–107.

Zimmermann, A. and Britz, W. (2016). European farms' participation in agri-environmental measures. *Land Use Policy* 50: 214–228.

European Review of Agricultural Economics Vol 00 (00) (2024) pp. 1–32, doi:https://doi.org/10.1093/erae/jbae034,