How Differently Do Farms Respond to Agri-environmental Policies? A Probabilistic Machine-Learning Approach

Silvia Coderoni *Assistant Professor, Department of Bioscience and Agro-Food and Environmental Technology, Università degli Studi di Teramo, Italy; scoderoni@unite.it*
Roberto Esposti *Professor, Department of Economics and Social Sciences, Università Politecnica delle Marche, Ancona, Italy; r.esposti@univpm.it*
Alessandro Varacca *Assistant Professor, Department of Economics and Social Sciences–DISES, Università Cattolica del Sacro Cuore, Piacenza, Italy;mailto: alessandro.varacca@unicatt.it*

ABSTRACT *This study evaluates the extent to which farmers respond heterogeneously to the agri-environmental policies implemented in the European Common Agricultural Policy (CAP). Our identification and estimation strategy combines a theory-driven research design formalizing all possible sources of heterogeneity with a Bayesian additive regression trees algorithm. Results from a 2015–2018 panel of Italian farms show that the responsiveness to these policies may differ substantially across farms and farm groups. This suggests room for improvement in implementing these policies. We also argue that the specific features of the CAP call for a careful implementation of these empirical techniques.* (JEL Q15, Q51)

1. Introduction

The Common Agricultural Policy (CAP) represents the primary ordinary policy instrument of the European Union, at least in terms of budget share. Starting with the 1992 MacSharry reform, environmental and ecological concerns have increasingly become one of the major justifications for maintaining the CAP expenditure. Indeed, environmental policy objectives are likely to be the most relevant for European agriculture in the coming decades (Coderoni et al. 2021). Given the growing concerns about environmental and ecological issues and the resulting policy orientations, researchers are left to wonder how much farmer behavior has changed in response to the new greener CAP and what those responses are (Brown et al. 2021). Answering these questions is rather challenging, mainly because there is no univocal answer for the very large heterogeneity typically encountered in agriculture.

Since EU farmers are known for their distinctive diversity (Esposti 2022a), we would typically expect equally diverse responses to these political shocks. Under this hypothesis, both academics and EU stakeholders have long advocated for a more targeted and tailored design of the EU policies (particularly CAP reforms; see Erjavec and Erjavec 2015; Ehlers, Huber, and Finger 2021). However, such a task is challenging without a deeper understanding of whether and to what extent the potential recipients of such measures respond differently. As most parametric/semiparametric (econometric) approaches to ex post policy evaluation can only produce aggregate (i.e., average) responses or represent limited and prespecified heterogeneity (e.g., Esposti 2017a, 2017b; Bertoni et al. 2020; Bartolini et al. 2021), the understanding of such heterogeneity has been rather limited so far.

Recent improvements in this field involve the use of specific causal inference (CI) methods (Imbens and Rubin 2015) for framing the evaluation of a policy as a treatment effect discovery problem, which exploits counterfactual thinking to define the estimands of interest (Uehleke, Petrick, and Hüttel 2022). In the rapidly evolving literature, causal machine learning (CML) has started to gain attention

Land Economics • May 2024 • 100 (2): 370–397
DOI:10.3368/le.100.2.060622-0043R1
ISSN 0023-7639; E-ISSN 1543-8325

as a useful extension to the more general CI framework, particularly when the objective of the evaluation regards highly complex and potentially heterogeneous responses to the treatment (Storm, Baylis, and Hekelei 2020; Stetter, Mennig, and Sauer 2022). Machine learning (ML) methods can be particularly beneficial when working with large, heterogeneous samples characterized by many interacting variables and nonlinear relationships but require suitable identifying assumptions and targeted technical adjustments (Chernozhukov et al. 2018; Hahn et al. 2018; Athey and Imbens 2019). This means that off-the-shelf ML algorithms (i.e., common ML methods designed for predictive purposes) may, at best, represent one of several components in the CML toolbox. Given these premises, CML represents a suitable instrument for understanding how and to what extent the impact of recent CAP environmental policies varies across diverse farms. Our work fits in the very recent and fast developing empirical literature that deals with this issue. In particular, we aim to disentangle the causal effect of two alternative treatment options expressing two different implementations of the agri-environmental policy (AEP) in the 2014–2020 CAP reform. On the one hand, we consider farms that not only fulfill the basic eligibility conditions to benefit from the whole Direct Payment (DP) but also apply for pillar 2 agri-environmental measures (AEMs).[1] On the other hand, we consider farms that choose not to comply with the conditional requirements (i.e., the so-called conditionality; see Section 2), thereby giving up the DP and not take up any pillar 2 AEM. We assume that the two treatments share the same control group, which consists of farms that only comply with the necessary environmental requirements to access the DP.

We begin by providing a theoretical background linking the determinants of AEP adoption by heterogeneous farmers to their production response and then linking it to the potential outcomes framework. We exploit these conceptual underpinnings to define the relevant confounding variables and treatments while providing a solid background for the necessary assumptions that characterize our identification strategy. The latter is grounded in the classical hypotheses that support most CI problems, including the stable unit treatment values assumption (SUTVA) that may be problematic given the multiple-treatment nature of the AEMs. These hypotheses are coupled with flexible surface estimation by a CML algorithm known as Bayesian causal forests (BCF) (Kapelner and Bleich 2016; Carnagie, Dorie, and Hill 2019; Hahn, Murray, and Carvalho 2020). Given their probabilistic nature, BCF can produce approximate posterior distributions for estimated heterogeneous treatment effects (HTEs), allowing the introduction of uncertainty into group comparisons or, more generally, when transforming individual-level estimands. This feature represents a further original contribution of this article, as it may provide a useful improvement over other comparable ML methods for which inference is less straightforward (Stetter, Mennig, and Sauer 2022).

Our research is closely related with the recent analysis presented by Stetter, Mennig, and Sauer (2022), as both studies share a common objective of assessing the heterogeneous response of farmers to AEPs through CML techniques. Nonetheless, as elaborated above and thoroughly discussed throughout, our approach diverges from and extends on their work in several fundamental aspects. These aspects encompass a more comprehensive delineation of the treatment set, a broader conceptualization of farmers' potentially heterogeneous response to AEPs, a distinct and relatively wider geographical coverage, and an investigation of the inherent limitations of conventional identification strategies used in cross-sectional observational studies.

2. Policy Relevance and Methodological Challenges

Over the past few decades, the EU CAP has undergone several structural reforms and has increasingly emphasized the primary sector's environmental dimension (Commission of European Communities 2000). Currently, the CAP includes objectives for protecting water,

[1] We consider AEMs as a subset of whole menu of AEPs.

soil, climate and air quality, landscape, and biodiversity (European Commission 2020). Following the 2014 CAP reform and the corresponding 2015–2020 CAP AEP design, these objectives are pursued by a diverse mix of policy instruments, three of which represent the subject herein.

The oldest of these three means of intervention (introduced in 1992) consists of the AEMs. These are voluntary measures belonging to CAP's pillar 2, which deliver compensatory payments to farmers to cover additional costs and forgone income from adopting more environmentally friendly practices. In our work with AEMs, we refer to two measures that, after the 2014 CAP reform, are named "measure 10" (agri-environment-climate commitments) and "measure 11" (organic farming). These measures provide monetary incentives for the voluntary adoption of eco-friendly farming techniques.[2]

Following the 2003 "Agenda 2000" reform, a second environmental measure was introduced: CAP's pillar 1 DPs became subject to the so-called cross-compliance (CC) requirements that made these monetary subsidies contingent on several environmental and ecological standards. Although these requirements are intended to be mandatory, *strictu sensu*, complying with part of them is like satisfying an eligibility condition for first-pillar payments, since noncompliance triggers administrative penalties up to the revocation of the DPs. Therefore, farmers may always give up applying for DP entirely, thus also ignoring part of the CC requirements.

The third policy instrument was introduced with the 2014 CAP reform through the so-called greening payment (GP). This measure represents the green component of the new modified DP scheme, in which the financial support now hinges on three mandatory practices intended to benefit both the environment and the climate. Since it builds on and reinforces CC, the GP is often regarded as a sort of additional (or super-) conditionality.[3] As in the previous case, noncompliance results in a loss of support directly delivered to farmers. Therefore, under the 2014–2020 CAP design, eligibility for the full DP related to environmentally friendly practices now depends on satisfying both CC and GP provisions.

It is worth noting that, in implementing such measures, there have been significant differences both across and within member states. For example, Italy has managed, implemented, and administered AEMs at the regional (NUTS-2) level through rural development plans (RDPs). Similarly, although CC requirements have been enforced following the EU conditionality principles, the list of commitments applicable at the local level has also been left to the regional authorities. These include commitments to prevent soil erosion, organic matter decline, and soil compaction; perform a minimum level of ecosystem maintenance; and prevent habitat and landscape deterioration (National Rural Network 2010). Finally, the GP is defined as a farm-specific, yearly, per hectare payment calculated as a proportion of a farm's DP total value. Once again, the actual implementation of the GP may be differentiated at the regional level.

Therefore, member states enforce and oversee these policy instruments acknowledging the existence of cross-country/cross-regional specificities, allowing for some degree of flexibility in their implementation (Guerrero 2021). Nevertheless, the content of these intervention tools (i.e., their monetary implications and associated requirements) remains rigid in comparison to the very diverse conditions to which they apply. In fact, the same policy menu is offered to very large farms and very small units, to extensive livestock farming in mountain areas and orchards in plain urban areas, and so on. This mismatch between highly heterogeneous farms and a

[2] Measure 10 supports (among other things) integrated production, manure management, increasing soil organic matter, sustainable management of extensive grassland, and management of buffer strips against nitrates. Measure 11 supports conversion to and maintenance of organic practices and methods. It is worth noticing that Stetter, Mennig, and Sauer (2022, 732) do not consider the organic farming measure "due to [the] distinctly different farming approach compared to conventional farms." As clarified in Section 4, we include this measure in the analysis to compare the results obtained on the whole sample.

[3] At the member state level, the total amount of GP must correspond to 30% of the total DPs. In several EU countries (including Italy), this condition is satisfied by automatically assigning to eligible farms 30% of total DP as the GP.

relatively homogeneous policy instrument is particularly delicate for Italy, whose primary sector mixes very different farming traditions and peculiar geographical characteristics (Coderoni and Esposti 2018). Such structural heterogeneity inevitably translates into behavioral heterogeneity in that the response of diverse farms to homogeneous policies may substantially diverge in terms of the size and nature of the response (i.e., the variables involved in the response). Moreover, even when farms exhibit analogous structural and behavioral characteristics, the uneven environmental effects that these policies may generate can result from very site-specific agronomic, ecological, and biophysical features, such as field slopes, soil types, hydrology, and crop rotation (e.g., Finn et al. 2009; Ó hUallacháin et al. 2016; OECD 2022).

These multiple and complex sources of heterogeneity suggest that AEPs should be more flexible in targeting diverse farms. Unsurprisingly, the need for a more tailored design of the CAP environmental policies has frequently been advocated over the past two decades (Erjavec and Erjavec 2015; Ehlers, Huber, and Finger 2021). In this respect, a policy rationalization through better targeting of specific farm characteristics might help achieve the declared environmental objectives, either through expenditure savings (for the same environmental performance) or through improved environmental performance (for the same level of expenditure) (Esposti 2022b). However, improving policy targeting and, ideally, tailoring also requires a better understanding of whether and how the potential beneficiaries of such measures respond differently. Borrowing from the CI jargon, one would wish to identify and estimate HTEs (or individual treatment effects) as the natural empirical counterpart of this knowledge gap.

Policy evaluation studies addressing the impact of agri-environmental policies have gained considerable attention in recent years. Chabé-Ferret and Subervie (2013), Arata and Sckokai (2016), Mennig and Sauer (2020), and Bertoni et al. (2020), to name a few recent examples, have applied difference-in-differences (DID) or matching techniques to assess the effects of different AEMs. Similarly, Bartolini et al. (2021) estimated the impact of AEMs in a multivariate treatment setting by adopting a generalized propensity score estimation. However, these studies typically have estimated average treatment effects (ATEs) without exploring treatment effect heterogeneity, if not by focusing on specific farm groups or considering quantile treatment effects (Esposti 2017a, 2017b). The main risk of working with such aggregate measures is that of hiding systematically different unit or group-level effects. In other words, what holds true on average might not hold true for specific clusters and vice versa. This may evidently lead to wrong policy conclusions.

In this respect, ML methods have recently proven a helpful toolbox for assessing AEPs. For example, Bertoni et al. (2021) used ML techniques to simulate the impact of GP in terms of land use change, although they did not touch on treatment effect heterogeneity. Among the latest contributions, Stetter, Mennig, and Sauer (2022) represent the only study explicitly addressing the heterogeneous response of (southeastern German) farms to AEMs in terms of environmental performances. We acknowledge that the proper identification of such HTEs can be problematic for at least two reasons: (1) using the participation to AEMs as a binary treatment variable can only proxy for a wide range of submeasures from which farmers can choose, and (2) measuring environmental performances is inherently hard because of the interconnected nature of many commonly adopted environmental indicators. Although HTEs can be particularly helpful for a better targeting of AEPs, thus improving their (cost) effectiveness, these two caveats may complicate their empirical tractability.

On the one hand, when policy measures are delivered via submeasures among which farmers can freely choose (i.e., a multivalued treatment), the standard identification strategies for HTEs may fail due to the presence of alternative versions of the treatment (VanderWeele and Hernán 2013; Lopez and Gutman 2017). Moreover, the interpretation of the resulting estimand could be misguided because the local differences in treatment effects could instead be driven by treatment heterogeneity (Heiler and Knaus 2022). On the other hand, had such disaggregation level been attainable, it would still be difficult to unambiguously link a specific scheme to a single environmental indicator. As previously mentioned, depending on the farm's specificity and the treatment, elementary environmental outcomes are always interdependent and hard to examine in isolation (Chabé-Ferret and Subervie 2013). In other words, for any treated unit, treatment effects can either differ across multiple indicators or, worse, trigger spillovers such that changes in one environmental outcome may impact others. Ignoring this output-dependent treatment effect heterogeneity (OTH) and focusing on elementary indicators may lead to misleading interpretations of the HTE.

While our interest lies in estimating the HTE of both DPs and AEMs in general, we also acknowledge and attempt to empirically address the two issues discussed above.

3. Theoretical Framework: Modeling Farmer Response to Agri-environmental Policies

We begin by discussing a simple theoretical framework conceptualizing farmer uptake of AEPs and providing a behavioral foundation for treatment effect heterogeneity. Unlike the model presented in Stetter, Mennig, and Sauer (2022), where HTEs only result from farm-specific production technologies, we postulate a stylized behavioral mechanism explaining how farms respond to different policy option and therefore how HTEs may emerge. Moreover, our framework formalizes how treatment heterogeneity and OTH can interfere with the identification of the HTEs of interest.

Consider a panel of N production units (i.e., farms) observed over T time periods. Each farm can choose among K alternative AEPs. Next assume that farmers are profit maximizers and, for simplicity, risk neutral. The latter greatly simplifies the following analytical treatment as it allows formulating farmer behavior in terms of actual profits ($\pi_{it,k}$) rather than expected profits.[4] In practice, we assume that none of the AEPs considered in this study imply a major change in the riskiness of farming activity.[5]

We postulate that each farm $i \in \{1,...,N\}$ is associated with an aggregated general multi-input multi-output farm-specific technology represented by the feasible production set $F_i \subset \mathbb{R}^M$. Given F_i, the $(M \times 1)$ vector of netputs $\boldsymbol{y}_i = (y_{1i},..., y_{Mi})'$ is feasible if $\boldsymbol{y}_i \in F_i$.[6] This netput vector contains both farm-specific outputs (with positive signs) and farm-specific inputs' use (with negative signs), possibly including nonmarket inputs and outputs. The adjective "farm-specific" implies that F_i contains all possible sources of heterogeneity in the farmer's production decisions that depend on both external and internal factors (Esposti 2022b).[7] We can express the ith farm's specific features with a Q-dimensional vector $\boldsymbol{Z}_{it}$.

To keep the notation consistent, we refer to the set $\{T_{it,1},...,T_{it,K}\}$ as the treatment set and to $T_{it,k}$ as treatment k. At period $t \in \{0,...,T\}$, any AEP chosen by farmer i, $T_{it,k}$, is expected to induce specific production choices, $\boldsymbol{y}_{it,k}$, via either output production or input use. Therefore, treatments can be univocally mapped to production choices ($T_{it,k} \leftrightarrow \mathbf{y}_{it,k}$). Notice that this argument holds for multiple treatments. For example, suppose that the kth treatment is delivered through V alternative versions

[4]Since production decisions must be taken ex ante, their consequences are evidently subject to some degree of uncertainty. Consequently, farmers actually maximize $E\{\Pi[g(T_{it,k},\boldsymbol{X}_i)]\}$ and, more importantly, the condition $E\{\Pi[g(T_{it,k},\boldsymbol{X}_i)]\} \geq E\{\Pi[g(T_{it,h},\boldsymbol{X}_i)]\}$, $\forall k, h \in K, k \neq h$ remains valid only if we are willing to assume farmer's risk neutrality. Otherwise, the variance of $\pi_{it,k}$ and $\pi_{it,h}$, and the possible impact of $T_{it,k}$ on them, would also matter.

[5]It can be argued that under risk aversion, farmers are expected to be more prudent and conservative; therefore, ceteris paribus, the participation in the treatment and the observed response, $\Delta\boldsymbol{y}$ should be smaller. At the same time, the monetary support granted to participant farmers may represent a guaranteed income, making participation in the measure a less risky situation. Also notice that under risk aversion, risk can be interpreted as an additional source of costs and/or forgone income that the AEP is expected to compensate. Therefore, as noted in previous studies (Esposti 2017a, 2017b), it is difficult to model and predict the differential impact of these support measures between risk-neutral and risk-averse farmers.

[6]Unlike the other vectors of model variables, the netput vector is here indicated with a small letter, $\boldsymbol{y}_{it}$, to avoid confusion with the conventional notation of potential outcomes, $Y_i(0)$ and $Y_i(1)$ (see Section 5).

[7]As will be clarified in Section 4, examples of internal factors are the farm size and the farmer's age and education. Examples of external factors are latitude and farm's location in a disadvantaged area.

($v = 1,...,V$) among which the farmers choosing the kth treatment can choose (VanderWeele and Hernán 2013). We can then indicate the treatment as $T_{it,kv}$. This does not affect the overarching structure of our theoretical model, as the new set of treatment option can be simply rewritten as $\{T_{it,1},...,T_{it,k1},...,T_{it,kV}, ...,T_{it,K}\}$, and it is always possible to express $(T_{it,kv} \leftrightarrow \mathbf{y}_{it,kv})$.

We can now express farmer production choices as functions of the policy treatments themselves, given a farm-specific technology F_i as expressed in $\boldsymbol{Z}_{it}$; that is, $\boldsymbol{y}_{it,k} = g(T_{it,k}, \boldsymbol{Z}_{it})$, where $g(.)$ is a vector-valued function. In addition, if farms are profit maximizers and can choose $T_{it,k}$, the policy support operates like market price changes in orienting production decisions (Esposti 2017a, 2017b). Consequently, we can generically express farms' individual profit functions as $\pi_{it,k} = \Pi[g(T_{it,k}, \boldsymbol{Z}_{it})]$, where $\Pi(.)$ is a single-valued function.[8]

This behavioral representation makes clear that farmer choice is not driven by $\boldsymbol{y}_{it,k}$, which is the main target of the policy, but by the associated profit $\pi_{it,k}$. Following this logic, each observed pair $(T_{it,k}, \boldsymbol{y}_{it,k})$ represents the profit-maximizing combination of each treatment and the resulting set of production choices. Without assuming any specific functional form for the underlying technology or profit function, an augmented version of the weak axiom of profit maximization can be formulated to identify the optimal netput vector $\boldsymbol{y}_{it,k}$ (Afriat 1972; Varian 1984; Chavas and Cox 1995; Esposti 2000). This implies that $\Pi[g(T_{it,k}, \boldsymbol{Z}_i)] \geq \Pi[g(T_{it,h}, \boldsymbol{Z}_{it})]$, $\forall k, h \in K, k \neq h$. Namely, the profit of the ith farmer choosing treatment k at time t ($\pi_{it,k}$) exceeds the profit that she would have achieved had the farmer chosen any other alternative T_h ($\pi_{it,h}$). For a given baseline treatment ($T_{it,0}$), farm i will choose treatment k at time t if $\Pi[\boldsymbol{y}_{it,k}(T_{it,k}, \boldsymbol{Z}_{it})] \geq \Pi[\boldsymbol{y}_{it,0}(T_{it,0}, \boldsymbol{Z}_{it})]$ or, alternatively, $\Pi[\Delta g(T_{it,k}, T_{it,0}, \boldsymbol{Z}_{it})] \geq 0$, where $\Delta g = \Delta \boldsymbol{y}_{it,k} = \boldsymbol{y}_{it,k} - \boldsymbol{y}_{it,0}$. Notice that in this conceptual framework, the full treatment set might not be feasible for all farms. In fact, $\boldsymbol{Z}_{it}$ might bind the choice of the netput vector $\boldsymbol{y}_{it,k}$, thereby limiting the choice of $T_{it,k}$ to a subgroup of $\{T_{it,1},...,T_{it,K}\}$. This may also apply when treatment is delivered through V alternative versions: given $\boldsymbol{Z}_{it}$, not all the subtreatments, $T_{it,k1},...,T_{it,kV}$, may be feasible for all farmers choosing the kth treatment.

The main goal of this article is to construct and identify an empirical counterpart of $\Delta \boldsymbol{y}_{it,k}$ and determine its distribution across heterogeneous farms.[9] Assuming that either $\boldsymbol{y}_{it,k}$ or $\boldsymbol{y}_{it,0}$ can be observed, this research question can be addressed using the CI analytical framework, where $\Delta \boldsymbol{y}_{it,k}$ indicates the TE of interest, and $\boldsymbol{y}_{it,0}$ represents the counterfactual state of $\boldsymbol{y}_{it,k}$, had the farm not chosen treatment k (Imbens and Rubin 2015). However, in the presence of multiple treatment versions ($T_{it,k1},...,T_{it,kv},...,T_{it,kV}$), $\Delta \boldsymbol{y}_{it,k}$ may differ from $\Delta \boldsymbol{y}_{it,kv}$, for some $v \in V$. Not only may these two quantities differ but, more importantly, we may also observe $(\Delta \boldsymbol{y}_{it,k} - \Delta \boldsymbol{y}_{jt,k}) \lesseqgtr (\Delta \boldsymbol{y}_{it,kv} - \Delta \boldsymbol{y}_{jt,kv}) \neq (\Delta \boldsymbol{y}_{it,k\mathcal{v}} - \Delta \boldsymbol{y}_{jt,k\mathcal{v}})$for any $i, j \in N$ and any two $\mathcal{v}$, $v \in V$. Heiler and Knaus (2022) show that the above inequality results from $(\Delta \boldsymbol{y}_{it,k} - \Delta \boldsymbol{y}_{jt,k})$ being a weighted average of all the treatment versions $\Delta \boldsymbol{y}_{it,kv}$, where the weights are proportional to the probability that farm i chooses $T_{it,kv}$. In other words, in presence of multiple treatment versions, we would erroneously mistake treatment effect heterogeneity for what is, in fact, a diverse treatment choice mechanism (i.e., treatment heterogeneity).

As introduced in Section 2, when it comes to evaluating the effect of a treatment, one could focus on one or multiple elements of the netput vector $\boldsymbol{y}_i = (y_i,..., y_{mi},..., y_{Mi})'$. However, since most entries in $\boldsymbol{y}_i$ can be highly

[8]Following the conventional terminology of production theory, this should be a direct profit function as opposed to the more frequently used indirect profit function, where profit is a function of only output and input prices. In fact, in addition to netput quantities, the direct profit function includes the respective prices expressed as $\Pi[\boldsymbol{v}'_{it} g(T_{it,k}, \boldsymbol{X}_i)]$, where $\boldsymbol{v}'_{it}$ is the $(M \times 1)$ vector of netput prices. For nonmarket netputs, there are no prices, but these elements in $\boldsymbol{v}'_{it}$ can still be interpreted as shadow prices. Nonetheless, prices have been excluded from the present notation under the assumption, maintained that the prices are constant or, more precisely, unaffected by the policy regime.

[9]The heterogeneity among farms is the core of this theoretical framework. With homogeneous farms, we would have $\pi_{it,k} = \pi_{jt,k} = \pi_t$, $\forall i \neq j$, $\forall k$ and $\forall t$, so all farmers would opt for the same policy, and we would observe only one treatment. A policy response would thus be only conjectural but not actually observable if not by comparing farms before and after the treatment.

interconnected (i.e., some y's can be positively or negatively correlated with one or more other y's), evaluating treatment effects though marginal evaluations of these elements could make results hard to interpret. For example, consider any two positively (or negatively) correlated items $y_{mi}, y_{li} \in \boldsymbol{y}_i$. Then, for any $i, j \in N$ and treatment $T_{it,k}$, we will have that $\Delta y_{mi,k}$ is also correlated with $\Delta y_{li,k}$. Therefore, comparing the marginal HTE for the two indicators—that is, comparing ($\Delta y_{mi,k} - \Delta y_{mj,k}$) against ($\Delta y_{li,k} - \Delta y_{lj,k}$)—can lead to misleading conclusions. We previously referred to this issue as OTH. In Section 4, we postulate that OTH can be addressed via dimension reduction, where we project a vector of correlated environmental indicators $\boldsymbol{y}_i^e \subset \boldsymbol{y}_i$ onto a lower-dimensional space through a synthetic environmental performance indicator. Nonetheless, it remains possible to empirical assess the potential interference of the OTH on HTE estimation by comparing the results obtained via the lower-dimensional index to those obtained on its individual components (see Section 6).[10]

If one can address treatment heterogeneity and OTH, then under suitable restrictions on the joint distribution of the potential outcomes ($\boldsymbol{y}_{it,k}, \boldsymbol{y}_{it,0}$) and given farm characteristics $\boldsymbol{Z}_{it}$, the identification of $\Delta \boldsymbol{y}_{it,k}$ can be achieved via unconfoundedness (see Section 5) if $\boldsymbol{Z}_{it}$ contains all the relevant variables that influence both the treatment choice, $T_{it,k}$, and the farmer's production choices (Angrist and Pischke 2008; Wooldridge 2010, ch. 21; Imbens and Rubin 2015, ch. 3).

Following Brown et al. (2021) and Stetter, Mennig, and Sauer (2022), we distinguish between four sets of farms attributes:[11] economic factors (i.e., factor endowment), sociodemographic characteristics (of the farm's holder and workforce), environmental (mostly geographical) factors, and idiosyncratic characteristics (of the farm's holder and workforce, such as ability, knowledge, motivations, beliefs, and values, as well as unobserved environmental features such as agronomic characteristics and fertility). To facilitate the illustration of our identification strategy, we assemble these characteristics into separate partitions of $\boldsymbol{Z}_{it}$, namely, $\boldsymbol{Z}_{it} = (\boldsymbol{X}_{it}, u_i)$, where $\boldsymbol{X}_{it}$ consists of a ($P \times 1$) array. Furthermore, we define $\boldsymbol{X}_{it} = (\mathbf{V}_{it}, \mathbf{S}_i)$, where $\mathbf{S}_i$ is a vector of observable time-invariant farm characteristics, $\mathbf{V}_{it}$ is a vector of observable time-variant farm attributes. u_i represents unobservable time-invariant farm features. According to this categorization, identifying HTEs requires two fundamental restrictions: first, $\mathbf{V}_{it}$ must be predetermined in that the treatment cannot affect $\boldsymbol{y}_{it}$ via $\mathbf{V}_{it}$; second, u_i must not be associated with both $T_{it,k}$ and $\boldsymbol{y}_{it}$, under penalty of introducing selection-on-unobservable bias (Imbens and Rubin 2015). Although the first condition can be satisfied using time-stable variables (i.e., $\mathbf{V}_{it} \approx \mathbf{V}_i$) or lagged values (see Section 4), the exogeneity of u_i is often assumed and tested via sensitivity analysis.

We maintain this assumption throughout, thus only focusing on $\mathbf{X}_{it}$ when discussing treatment effect identification. As discussed in Sections 4 and 6, however, we also resort to suitable robustness checks to test the validity of our identification strategy under endogenous u_i.

4. Data and Research Design

Observational Dataset

We use information from the Italian Farm Accountancy Data Network (FADN), which represents the only source of microeconomic agricultural data that is harmonized at EU level and collects physical, structural, economic, and financial data on farms in all EU member states (European Council 2009). The survey is representative of the farms that can be considered professional and market oriented, due to their economic size (that is equal or more than €8,000 of standard output). In Italy these correspond to 95% of utilized agricultural area, 97% of the value of standard production, 92% of labor units, and 91% of livestock units. The representativeness of the dataset is ensured on three dimensions, namely region, economic

[10]Notice that this assessment applies to both single treatment and multiple-treatments versions.

[11]See Zimmerman and Britz (2016), Dessart, Barreiro-Hurlé, and van Bavel (2019), Brown et al. (2021) for recent and extensive reviews of structural and behavioral factors underlying farmer's decisions.

size, and farm typology. For these reasons, the FADN is the most (and only) widely used farm-level dataset for, among others, CAP evaluations and specifically for the assessments of the AEP impacts (among others, Arata and Sckokai 2016; Bartolini et al. 2021; Stetter, Mennig, and Sauer 2022).

Our research focuses on the 2014–2020 programming period of the CAP.[12] However, unlike Stetter, Mennig, and Sauer (2022), we exclude the initial year (2014) for two reasons: first, payments of one of the policies under consideration (the GP) only started in 2015; second, many of the farms observed in 2014 may still benefit from measures of the previous programming period. We thus focus on the 2015–2020 period, although we only have detailed and validated information until 2018. Therefore, our initial sample consists of a representative collection of Italian commercial farms that produces an unbalanced panel consisting of 9,580, 10,135, 10,792, and 10,386 observations in 2015, 2016, 2017, and 2018, respectively. Because our analysis does not address regime-switching dynamics, we only consider farms for which the treatment status did not change over the period analyzed; that is, $T_{it,k} = T_{i,k}$ for all $i \in (1,...,N)$. For this reason, we first extract a balanced panel consisting of 5,836 units observed over 2015–2018 and then drop all entries satisfying $T_{it,k} \neq T_{is,k}$ for any s, $t \in \{2015,..., 2018\}$ and $s \neq t$.[13] The resulting dataset consists of 4,001 farms repeated over four years, for a total of 16,004 observations. Compared with other related works (Bertoni et al. 2020; Stetter, Mennig, and Sauer 2022), our study provides wide coverage of the agricultural sector by focusing on the entire national area instead of a single region. Furthermore, since the treatments presented in Section 4 are likely to affect the agri-environment over several years, our outcome variable uses information from the last two years in the series to account for potential accumulation effects (see Section 4 for details).

[12] The programming period has been subsequently extended to 2022, also because of the COVID-19 pandemic. Validated data from 2021 and 2022 have still to be released.

[13] It is worth noticing that extracting the balanced sample from the unbalanced one does not imply a relevant loss in terms of representativeness of the sample; see Baldoni, Coderoni, and Esposti (2021) for a detailed explanation.

Definition of Treatments

As mentioned in Section 2, the 2015–2020 CAP AEP design is based primarily on two main policy instruments that belong to CAP's pillar 1, pillar 2, or both. On the one hand, we observe pillar 1 subsidies that are conditional on a set of compulsory requirements (i.e., CC and the GP) with which farmers must comply to preserve the DP. On the other hand, we have voluntary measures aimed at compensating farmers for income losses or increased costs resulting from the voluntary adoption of more sustainable farming practices (i.e., the AEM of pillar 2). Consequently, farms are subscribed to—in fact, they voluntarily choose—one of three possible policy alternatives, which effectively reflect the interplay between the two pillars of the CAP: (1) farms failing to meet all the CC and GP requirements, that is, farms receiving neither pillar 1 nor pillar 2 payments; (2) farmers receiving both pillar 1 (DP and GP) and pillar 2 (AEM) payments; and (3) farms complying with the CC and GP requirements but not adopting any AEM.

Table 1 indicates how the farms in our sample are distributed across the three policy categories. The third cohort is the largest group, which includes approximately 71% of the observed farms (2,841 units). Using the terminology introduced in Section 3, we consider the corresponding policy option as the baseline treatment, $T_{i,0}$, associated with the netput vector $\boldsymbol{y}_{it,0}$. Next, all farms choosing not to benefit from pillar 1 and pillar 2 payments (i.e., the first cohort, corresponding to approximately 13% of the sample) take up the first treatment, $T_{i,k=1}$, which implies giving up both pillar 1 and pillar 2 resources. We assume that this decision follows the behavioral model stylized in Section 3, according to which, conditional on $\boldsymbol{X}_{it}$, $T_{i,1}$ produces higher profits than $T_{i,0}$. Similarly, farms applying for pillar 2 AEM supports (i.e., the second cohort, corresponding to approximately 16% of the sample) choose treatment $T_{i,k=2}$ through the same profit-maximizing mechanism. In this respect, our work extends the analysis in Stetter, Mennig, and Sauer (2022) by distinguishing between the two different AEPs (i.e., the AEMs and the pillar 1 environmental requirements).

We postulate that treatments $T_{i,1}$ and $T_{i,2}$ belong to two nonoverlapping choice sets; in other words, we rule out a multiple treatment setup by positing treatment $T_{i,1}$ as infeasible for farms choosing $T_{i,2}$ and vice versa. Although this assumption is quite strong, it is necessary to identify the treatment effects of interest. Given that $T_{i,1}$ and $T_{i,2}$ represent two ends of a rather wide spectrum of policy options, it is plausible that both treatments may appeal to (i.e., are feasible for) farms with very distinctive characteristics. Conversely, our setup implies that both $T_{i,1}$ and $T_{i,2}$ are feasible alternatives to the baseline treatment $T_{i,0}$. This presupposes that farms in the control group are characterized by features $\boldsymbol{X}_{it}$ that overlap with the characteristic of the units in $T_{i,1}$ or $T_{i,2}$. That is, we can always find comparable farms in either of the two groups in different strata of $\boldsymbol{X}_{it}$, that is, $0 < Pr(T_{i,k} = 1 \mid \boldsymbol{X}_{it} = \boldsymbol{x}_{it}) < 1$. This restriction is also commonly known as common support (or positivity), and as we discuss in Section 5 and Appendix E, it limits extrapolation issues, thus preventing unreliable treatment effects.

One caveat in our setup is that unlike $T_{i,1}$ farms choosing $T_{i,2}$ may in fact opt for one among four treatment versions. As outlined in Section 2, $T_{i,2}$ aggregates measure 10 and 11 which in turn can be decomposed in two submeasures: agri-environment-climate commitments (10.1); conservation and sustainable use and development of genetic resources in agriculture (10.2); payment to convert to organic farming practices and methods (11.1); and payment to maintain organic farming practices and methods (11.2). While measure 10.2 only concerns a small share of farms (roughly 3% of our sample) and can be thus excluded or safely merged into measure 10.1 (our current choice), submeasures 11.1 and 11.2 are substantially equivalent in terms of farmer behavior, the only difference being the amount of support granted. For this reason, we de facto consider submeasure 11.1 and 11.2 as a unique measure (i.e., measure 11). As put forward in Sections 2 and 3, disregarding such distinctions may greatly affect the interpretation of the HTEs via treatment heterogeneity.

It is also worth mentioning that in principle, the submeasures could be further disaggregated into specific actions (using the RDP jargon).

Table 1
Policy Treatments Set

Treatment	Agri-environmental Policy	Total Farms (2015–2018 FADN Balanced Sample)
T_1	None	512
Control group	Only those implied by the pillar 1 direct payments	2,841
T_2	Those implied by both pillar 1 direct payments and pillar 2 agri-environmental measures	648

Note: FADN = Italian Farm Accountancy Data Network.

Unfortunately, the Italian FADN data do not provide enough information on AEM actions. In fact, to our knowledge, there are no high-quality representative datasets that can provide more detail on AEMs (e.g., Stetter, Mennig, and Sauer 2022, who use the German version of our dataset). Had this level of disaggregation been observable, it would imply a very large number of actions (i.e., treatment versions), as evidenced by the 21 RDPs implemented in Italy.[14] Clearly, expanding the treatment options well beyond the four submeasures would greatly affect the sample size of each subgroup and challenge the estimation of any HTE under the standard conditions discussed in Section 5 (Heiler and Knaus 2022). Finally, focusing on more specific measures does not necessarily imply a more refined outcome variable (see Section 4 for further discussion).[15]

Since organic farming (measure 11) is homogeneous across the RDPs and involves a reasonable number of farms (271), we repeat our analysis by redefining treatment T_2 as a two-versions treatment $T_2 = (T_{2o}, T_{2n})$, where

[14] More specifically, from a survey carried out at national level, it emerged that there are 65 different versions of measure 10 that can be applied at regional programming level, corresponding to a total of 100 commitment categories for the whole 21 RDPs; see https://www.reterurale.it/flex/cm/pages/ServeBLOB.php/L/IT/IDPagina/23816.

[15] The support for organic farming is exemplary in this respect. The nature of the response may vary largely across different farming types, even under such a very specific measure. The same argument applies to CC requirements, where each element and constraint becomes applicable to the farm depending on the characteristics of the farmland or the agricultural activities carried out.

o = organic and n = nonorganic. Given our initial definition of the treatments, T_{2n} coincides with measure 10 which, unlike measure 11, is not entirely homogeneous across RDPs and could be exposed to further treatment heterogeneity. We therefore estimate the HTE of T_2 under two different setups: (1) we analyze the HTE of participating to AEMs as in Stetter, Mennig, and Sauer (2022); (2) we break down the treatment in setup 1 into T_{2o} and T_{2n} and obtain the corresponding HTE; and (3) we compare the results from setups 1 and 2 and discuss their implications for the interpretation of the HTE of interest (see Section 6).

Outcome Variable

The theoretical framework presented in Section 3 expresses the farm response to the treatment as $\Delta \boldsymbol{y}_{it,k}$, that is, a vector whose nonzero elements represent all of the farmer's production choices associated with the treatment in terms of both input and output.[16] These elements may consist of a long list of the farmer's specific production decisions, ranging from crop and livestock management practices to water and nutrient use (Burton and Schwarz 2013; Guerrero 2021, 11). One way to reduce the dimensionality of $\Delta \boldsymbol{y}_{it,k}$ consists of identifying and extracting the elementary indicators expressing the change in farming practices toward extensification or environmentally friendly practices. However, as discussed in Sections 2 and 3, focusing on elementary indicators might cause ambiguity when interpreting treatment effect heterogeneity because of the OTH problem. Given the potential correlation among the components of $\Delta \boldsymbol{y}_{it,k}$, one way to retain all the information in the netput vector while avoiding multiple marginal evaluations is to perform dimension reduction (Chipman and Gu 2005) to obtain composite dimensional indices (Bartolini et al. 2021). This strategy not only provides an insulation against OTH but also resonates the need for a comprehensive evaluation of complex policy instruments such as the AEM discussed in Sections 2 and 4. As also argued by Stetter, Mennig, and Sauer (2022, 727), despite the articulation of AEMs in specific submeasures, the goal of the AEPs remains more general, aiming to improve the overall environmental performance of the agricultural sector. Although many studies have tried to evaluate the effectiveness of distinct AEPs with respect to specific policy targets (e.g., the impact on biodiversity), the integrated assessment of multifaceted goals involving, for example, soil and water protection and the curbing of greenhouse gas (GHG) emissions have received relatively little attention until recently (Hudec et al. 2007; Zhen et al. 2022). However, the literature has long suggested that the intricate and ecosystemic nature of the agri-environment requires that any assessment should be based on a comprehensive integration of indicators across many environmental dimensions (Wascher 2003; Purvis et al. 2009).

In this respect, Purvis et al. (2009) propose an interesting, harmonized approach to evaluating AEMs: the so-called agri-environmental footprint index (AFI). The AFI expresses a multidimensional assessment as a univariate index that can be flexibly adapted to diverse contexts. We use the AFI framework as adapted by Westbury et al. (2011) with the FADN data. We refer to this methodology as FADN-AFI, as the resulting index uses elementary information included in the FADN dataset. We extend the FADN-AFI to evaluate whether and to what extent the implementation of the CC requirements, GPs, and AEMs meet the CAP 2015–2020 environmental objectives.[17]

Table 2 presents the elementary components of our FADN-AFI (see Appendix Table B2). The land use diversity indicator (the Shannon index) is detailed in Appendix

[16] For elements of $\boldsymbol{y}_{it,k}$ that are only marginally (or not at all) affected by the policy treatment under consideration, we have $\Delta \boldsymbol{y}_{it,k} \approx 0$. Therefore, we may restrict the analysis only to input and output decisions that are related to the environmental measures, all the rest being orthogonal by assumption.

[17] These goals are related to (1) the mandatory practices devised to benefit the environment (soil and biodiversity in particular) and climate (with the GP of pillar 1), and (2) the new RDP priority areas specifically addressing the environment and climate change (pillar 2). The latter are aimed at restoring, preserving and enhancing ecosystems dependent on agriculture and forestry (priority 4) and promoting resource efficiency and supporting the shift toward a low-carbon and climate-resilient economy in the agriculture, food and forestry sectors (priority 5).

Table 2
Elementary Indicators Used to Assemble the Outcome Variable (FADN-AFI)

Environmental Issue	Assessment Criterion	Indicator	Measurement Unit	Weight
Natural resources protection	Intensity crop husbandry / livestock production	Fertilizer cost of UAA	€/ha	−1
		Crop protection costs UAA	€/ha	−1
		Average number of livestock units	LU/ha	−1
	Energy consumption	Energy costs UAA	€/ha	−1
		% UAA irrigated	Share	−1
Biodiversity and land use	Land use diversity	Crop diversity (Shannon diversity) index (BI_{it})	Index (0–1)	+1
	Provision woodland habitats	% total farm area that is woodland	Share	+1
Climate change mitigation	GHG emissions	GHG at farm level	kg CO_{2eq}	−1

Note: AFI = agri-environmental footprint index; FADN = Italian Farm Accountancy Data Network; UAA = utilized agricultural area.

A. Appendix B discusses the definition of a farm-level GHG emissions indicator using farm-level information. This measure should provide a reliable proxy of the contribution of a farm's practices to climate change mitigation (Dabkiene, Balezentis, and Streimikiene 2021). The FADN-AFI's elementary components are then standardized to obtain dimensionless z-scores that we eventually aggregate using the weights indicated in the last column of Table 2 (i.e., giving a positive or negative sign for positive or negative environmental externalities, respectively).[18] The resulting FADN-AFI is monotonic in farms' environmental performance in that higher FADN-AFI scores correspond to "better" environmental performance. Since the range of the FADN-AFI is not bounded, the index might be difficult to interpret per se. However, since HTEs are defined through pairwise differences, these can easily be understood comparatively. Finally, we average the FADN-AFI in 2017–2018 to provide more stable values for the outcome variable.[19]

Confounding Variables

As discussed in Section 3, the choice of covariates entering the $\mathbf{X}_{it}$ vector becomes crucial for identifying the HTEs of interest. These should encompass farm heterogeneity as extensively as possible, thereby allowing fair comparisons between treated and untreated units. Selecting all the relevant confounders such that the assumptions outlined in Section 5 are satisfied may follow multiple routes. On the one hand, one may construct a very large collection of internal farm characteristics and external socioeconomic indicators that might explain the individual decision of adopting one of the treatments. In this case, we would let ML algorithm choose which feature contributes the most to predict farmer behavior through a regularization mechanism. However, as recently outlined by Hünermund, Louw, and Caspi (2023), this strategy may lead to severely biased treatment effects if the covariate set includes potentially endogenous confounding variables. Ultimately, the authors advocate that when the goal is conducting CI, researchers need to justify the controls they want to include and, more importantly, make sure that these are exogenous (i.e., pretreatment).

For these reasons, we begin by defining the confounders in $\mathbf{S}_i$ and $\mathbf{V}_{it}$ through an extensive literature review covering several empirical studies addressing farmer participation in AEPs and the impact of AEPs on farms' economic and environmental performance. The results of this survey are displayed in Table 3, where the list of covariates resulting from this desk research is classified using the taxonomy elaborated by Brown et al. (2021) and discussed in Section 3. We invite the reader to refer to the individual studies for a throughout

[18] Following Purvis et al. (2009), all the indicators and assessment criteria in the FADN-AFI receive a subjectively equal weighting.

[19] Averaging only over the last two years reduces the risk of integrating out potential accumulation effects by smoothing over a longer period (i.e., the cumulative benefit of environmentally friendly practices).

Table 3
Covariates Used in Analysis

	Unit of Measure	Component	Reference
Economic characteristic			
Total arable land	ha	V_{it}	(1), (2), (3), (4), (6), (7), (8)
Share of rented land	%	V_{it}	(2), (3), (4), (7), (8)
Farm revenue	€ per ha	V_{it}	(4), (7), (8)
Farm fixed costs	€ per ha	V_{it}	New
Farm variable costs	€ per ha	V_{it}	(4)
Fertilizer expenditure	€ per ha	V_{it}	(2), (4), (7)
Pesticides expenditure	€ per ha	V_{it}	(2), (4), (7)
Livestock density	Units per ha	V_{it}	(2), (5), (6), (8)
Family labor	Count	$S_i(*)$	(2), (3), (4), (8), (9)
Nonfamily labor	Count	V_{it}	(2), (3), (4), (9)
Share of most important crop	%	V_{it}	(5), (6)
Share of second most important crop	%	V_{it}	(5), (6)
Share of grassland	%	V_{it}	(4), (7), (8)
Machinery horsepower	Kw per ha	V_{it}	(3)
Machinery value	€ per ha	V_{it}	(3)
Machinery endowment	Units per ha	V_{it}	(3)
Farm specialization	Categorical	S_i	(3), (6), (7), (8)
Sociodemographic characteristic			
Farmer's age	Years	$S_i(*)$	(1), (3), (5), (6), (7), (8), (9)
Farmer's gender	Categorical	S_i	New
Farmer's education	Categorical	$S_i(*)$	(1), (3), (8)
Experience with previous AEPs	Categorical	S_i	(6), (7)
Environmental/geographical characteristic			
Disadvantaged area	Categorical	S_i	(5), (7), (8)
Latitude and longitude	Degrees	S_i	(6), (7)
Average altitude	Meters	S_i	(9)

Note: These characteristics are not strictly time invariant, but we assume they are approximately so ($V_{it} \approx S_i$) for the period 2015–2018. The references are (1) Vanslembrouck, Van Huylenbroeck, and Verbeke (2002); (2) Pufhal and Weiss (2009); (3) Pascucci et al. (2013); (4) Arata and Sckokai (2016); (5) Zimmerman and Britz (2016); (6) Bertoni et al. (2020); (7) Uehleke, Petrick, and Hüttel (2022); (8) Wąs et al. 2021; (9) Varacca et al. (2023).

explanation of how these regressors are relevant for the research questions. The abundance of controls compiled in this long list might suggest some form of preliminary selection to avoid redundancy and achieve a more parsimonious set of variables. Nevertheless, unlike most parametric econometric tools, forest-based ML algorithms can easily accommodate multiple overlapping information sources and use them to either create intermediate features or discard redundant ones through regularization. Therefore, our empirical analysis makes use of all the covariates in Table 3.[20]

To satisfy the identifying conditions anticipated in Section 3, the time-varying controls, $\mathbf{V}_{it}$, must be exogenous with respect to the treatment (i.e., predetermined). In theory, this would preclude the use of certain direct measures of farm physical and economic size, such as utilized arable land, profit, revenue, costs, and total workforce. To circumvent this issue, some authors suggest using covariates measured before the introduction of the treatment (for studies assessing AEMs, see, e.g., Bertoni et al. 2020; Uehleke, Petrick, and Hüttel 2022; Stetter, Mennig, and Sauer 2022). However, this strategy is sometimes infeasible, as such measurements may not be available if the policies under investigation were introduced several years before the outcome

[20]This explains the presence of insurance expenditure among covariates. This variable might seem contradictory to the risk neutrality assumed in deriving the theoretical framework (Section 3). However, it is worth remembering that in most cases, farms incur these costs not because of their risk aversion but because taking out an insurance contract is mandatory to receive public or private investment support. For this reason, this variable was considered in previous studies and thus in the present study.

is measured. When this happens, going back in time may imply a major loss of observations. This concern is particularly relevant for our application, as the rotating structure of the Italian FADN panel shows that 582 farms (approximately 15% of the sample) included in the 2015–2018 dataset are not present in the 2014 data. Therefore, our choice is to follow the strategy of Arata and Sckokai (2016) and Pufhal and Weiss (2009), which consists of using the first year since the introduction of the policy as the pretreatment period (2015, in this case).[21] Notice that since our outcome variable is calculated using the years 2017 and 2018, $\mathbf{V}_{it}$ contains lagged (by two years) elementary components of the FADN-AFI. Moreover, since farms usually sign up for participating in certain AEMs over several years (Bertoni et al. 2020; Uehleke, Petrick, and Hüttel 2022), we also include information on previous participation to such programs in $\mathbf{V}_{it}$ (Chabé-Ferret and Subervie 2013). Appendix Tables C1 and C2 report descriptive statistics for the outcome variable and all the control variables discussed above.

Unobservable Characteristics

The theoretical derivation in Section 3 provides the behavioral foundation of the farmer's treatment choice and response to the treatment. This behavior depends on some observable characteristics but also on unobservable farm characteristics, u_i. The conditional independence between any of the treatments and the corresponding potential outcomes also hinges on the last component of the conditioning vector $\mathbf{Z}_{it}$, namely, the unobservable farm characteristics, u_i. If these latent features influence the choice between $T_{i,1}$ and $T_{i,2}$ and the corresponding potential outcomes, the identification of the HTE becomes challenging because of the violation of unconfoundedness. Even though $\boldsymbol{X}_{it}$ can be extended to collect as many observable farm characteristics as possible, this strategy may be insufficient to insulate against selection-on-unobservable. Policy conclusions drawn from the HTE estimation could be problematic and even erroneous if the relevance of these unobservables and their possible association with the observable characteristics are not properly investigated and understood.

In these situations, ML methods (including BCFs) can help in identifying automatically creating nonlinearities and complex interactions among the variables in $\boldsymbol{X}_{it}$, generating artificial strata that allow more precise comparisons between treated/untreated units and their counterfactuals. These "synthetic traits" not only greatly expand the initial set of confounders but also correlate with the unobservable characteristics, thereby making the unconfoundedness assumption more credible. This argument is also put forward by Stetter, Mennig, and Sauer (2022, 738–39, 744), who provide a nice example of how this property of ML techniques may help to control for farmer attitudes toward environmental issues.[22] Since this is not directly testable, we check the robustness of the above propositions through several sensitivity analysis tests. As illustrated in Appendix H, we probe the stability of our results in the presence of omitted variable bias from unobserved endogenous heterogeneity by introducing synthetically generated u_i into the covariate set. See Section 6 for more details and caveats of this approach.

5. Methodology

Research on the estimation of HTE has flourished recently, stimulated by an increasing interest in the development of ML methods able to provide theoretically sound inferences in such research settings (Athey and Imbens 2019; Athey, Tibshirani, and Wager 2019; Hahn, Murray, and Carvalho 2020; Knaus, Lechner, and Strittmatter 2021, 2022). Recent studies have proposed two ways ML can be used to estimate HTE. First, off-the-shelf ML

[21] This requires assuming no anticipation and no instantaneous impact of either T_1 or T_2 on $\mathbf{V}_{it}$. With no anticipation, we refer to the assumption that farmers have not changed their characteristics $\mathbf{V}_{it-1}$ in response to the foreseen implementation of the policy at time t.

[22] In short, the authors discuss how a construct resulting from the interaction between farm type, farm size, farmer's age, farm capital intensity, and proxies for risk behavior is conceivably strongly correlated with the unobservable trait, thereby contributing to deconfounding the treatment effect.

algorithms can be tweaked to address some of the relevant identification issues of CI directly (Imai and Ratkovic 2013; Athey and Imbens 2016;; Wager and Athey 2018; Hahn, Murray, and Carvalho 2020).[23] Second, direct modifications of the loss functions and data-splitting techniques can also help address one challenging problem of traditional ML techniques in causal settings: regularization-induced confounding (RIC) (Chernozhukov et al. 2018, and references therein; Hahn et al. 2018; Hahn, Murray, and Carvalho 2020; Nie and Wager 2021). We broadly refer to all these methods as CML.

Among the diverse approaches proposed in the literature, BART-based algorithms (Chipman, George, and McCulloch 2010; Hill 2011; Hill, Linero, and Murray 2020) stand out as promising additions to the CML toolbox. These methods not only exhibit encouraging performance in terms of unbiasedness and coverage rates (Carvalho et al. 2019; Dorie et al. 2019; Hahn, Murray, and Carvalho 2020; Lee, Bargagli-Stoffi, and Dominici 2020) but also take advantage of a fully probabilistic (i.e., Bayesian) inferential approach, which enables the introduction of uncertainty measures when comparing groups of individuals (an aspect that currently limits the extent of other comparable ML methods; Stetter, Mennig, and Sauer 2022) and facilitates investigating the extent of overlap between treated and untreated groups (see Appendix E for details; Hill and Su 2013; Li, Ding, and Mealli 2022). The latter is particularly important when it comes to treatment T_1, as the farms associated with this group are likely to exhibit very specific characteristics (see Appendix E; Esposti 2017a, 2017b). Both traits hinge on the full posterior distributions of, on the one hand, the estimated HTE and, on the other hand, the fitted individual-level conditional expectations.

As with many other tree-based methods, BART can flexibly fit complex response surfaces by creating regularized ensembles of shallow Bayesian regression trees (Chipman, George, and McCulloch 1998), making it possible to perform predictive inference using the resulting posterior distributions (Chipman, George, and McCulloch 2010). This flexibility is achieved via recursive partitioning of the covariate space at the tree level, a procedure that is adept at defining nonlinearities and interactions between the observed covariates without the need to prespecify them (Hill 2011). However, since the original BART was not purposely designed for CI, a naive application of such methods for the estimation of HTE might potentially introduce RIC. For this reason, Hahn, Murray, and Carvalho (2020) recently proposed an extension of the original algorithm, which they refer to as BCFs.[24] In addition to exploiting the estimated propensity score (PS) to deal with potential distortions attributable to RIC (see Appendix D), the BCF algorithm also provides for a more flexible structure that separates the prognostic component from the heterogeneous treatment effect, thereby enabling direct control over the latter to avoid overfitting.

Estimating Treatment Effects via BCF

The estimation of HTEs using the BCF algorithm requires the usual assumptions of unconfoundedness and SUTVA, which can be expressed as follows:

$$Y_i(0),\ Y_i(1) \perp T_{i,k} \mid \boldsymbol{X}_i, \qquad [1]$$

where Y_i represents the FADN-AFI defined in Section 4, $\boldsymbol{X}_i$ indicates the vector of confounders defined in Section 4, while $Y_i(1)$ and $Y_i(0)$ indicate potential outcomes for individuals in a treatment group ($T_{i,k} = 1$) or control group ($T_{i,k} = 0$), respectively (Imbens and Rubin 2015, ch. 1). Notice that SUTVA implies no hidden variations of the treatment. As discussed in Sections 2 and 3, binarized multiple-versions treatments can lead to violations of this assumption unless one imposes stringent restrictions on the treatment assignment mechanism. For example, in case any individual i with characteristics $\boldsymbol{X}_i$ can only choose one of the hidden treatments, SUTVA is still a credible assumption (VanderWeele

[23] For an inventory of these methods, see Nie and Wager (2021).

[24] Notice that although the terminology "causal forests" resembles that used in Wager and Athey (2018), BCF differs substantially from the frequentist counterpart in their definition, functioning, and in how inference is performed.

and Hernán 2013; Lopez and Gutman 2017). As previously discussed, we make this assumption for the treatments defined in Section 4, except for the distinction between organic and nonorganic farming. We therefore set k to $k \in \{1,2\}$ such that $T_{i,k} = 1$ indicates either $T_{i,1} = 1$ or $T_{i,2} = 1$, while $T_{i,k} = 0$ always refers to farms in the control group. We discuss the implication for disaggregating $T_{i,2}$ into $T_{i,2o}$ and $T_{i,2n}$ in Section 6. For notational convenience, we drop the subscript k. Of these elements, we only observe the potential outcome that corresponds to the realized T_i, namely, $Y_i = T_iY_i(1) + (1 - T_i)Y_i(0)$. Equation [1] postulates independence between the potential outcomes and the treatment, conditional on the set of exogenous variables, $\boldsymbol{X}_i$.

Combining unconfoundedness, SUTVA, and overlap (as discussed in Section 4) allows the estimation of causal effects via strong ignorability; that is, $\mathbb{E}[Y_i(t)\,|\,\boldsymbol{X}_i = \boldsymbol{x}_i] = \mathbb{E}[Y_i\,|\,T_i = t_i, \boldsymbol{X}_i = \boldsymbol{x}_i]$, with $t_i \in \{0,1\}$. The latter implies that the estimand of interest is simply the difference between two conditional expectation functions:

$$\begin{aligned}\tau(\boldsymbol{x}_i) &= \mathbb{E}[Y_i\,|\,T_i = 1, \boldsymbol{X}_i = \boldsymbol{x}_i] - \mathbb{E}[Y_i\,|\,T_i = 0, \boldsymbol{X}_i = \boldsymbol{x}_i] \\ &= \mu_1(\boldsymbol{x}_i) - \mu_0(\boldsymbol{x}_i),\end{aligned} \tag{2}$$

where $\tau(\boldsymbol{x}_i)$ is typically referred to as a conditional average treatment effect (CATE). Since one can use $\mu_T(\boldsymbol{x}_i)$ to impute conditional treatment effects at the individual level, equation [2] is sometimes referred to as individualized average treatment effect (IATE) (Lechner 2018; Knaus, Lechner, and Strittmatter 2021, 2022). This estimand represents the most disaggregated form of HTE.

Often researchers may be interested in subgroups or intermediate aggregation levels of the exogenous covariates, leading to the definition of group average treatment effects (GATEs):

$$\tau(\boldsymbol{g}_i) = \int d\boldsymbol{x}_i \tau(\boldsymbol{x}_i)\phi_{\boldsymbol{X}_i|\boldsymbol{G}_i=\boldsymbol{g}_i}(\boldsymbol{x}_i), \tag{3}$$

where $\phi(.)$ represents a generic probability density of mass function, $\boldsymbol{G}_i$ denotes the collection of possible groups, and $\boldsymbol{g}_i$ denotes one such group. GATEs have recently gained considerable attention in the applied literature as treatment effect heterogeneity is often better understood for subsets of the population (Lechner 2018; Lee, Bargagli-Stoffi, and Dominici, 2020). ATEs can also be obtained by averaging the IATEs over the full distribution of $\boldsymbol{X}_i$:

$$\tau = \int d\boldsymbol{x}_i \tau(\boldsymbol{x}_i)\phi_{\boldsymbol{X}_i}(\boldsymbol{x}_i) \tag{4}$$

To estimate the IATEs (and then the GATEs and ATEs), we assume that the data-generating process for follows a stochastic process defined as follows:

$$Y_i = f(\boldsymbol{X}_i, T_i) + \varepsilon_i, \tag{5}$$

where f indicates an arbitrarily complex function[25] and ε_i represents an additive idiosyncratic error term $\varepsilon_i \sim \mathcal{N}(0,\sigma^2)$, independently distributed.

In this context, $\mathbb{E}[Y_i\,|\,T_i = t_i, \boldsymbol{X}_i = \boldsymbol{x}_i] = f(\boldsymbol{x}_i, t_i)$ therefore, at least in principle, $\tau(\boldsymbol{x}_i)$ can be estimated by the simple difference $f(\boldsymbol{x}_i, t_i = 1) - f(\boldsymbol{x}_i, t_i = 0) = \mu_1(\boldsymbol{x}_i) - \mu_0(\boldsymbol{x}_i)$, as illustrated above. However, as discussed by Künzel et al. (2019) and Nie and Wager (2021), training two separate conditional mean functions and taking their difference may produce highly unstable estimates. For this reason, Hahn, Murray, and Carvalho (2020) proposed a slightly different approach, wherein the expected value of the outcome of interest has two components: a prognostic function, $\mathcal{m}(\boldsymbol{x}_i)$ plus an additive heterogeneous treatment effect, $\tau(\boldsymbol{x}_i)$:

$$\begin{aligned}\mathbb{E}[Y_i\,|\,T_i = t_i, \boldsymbol{X}_i = \boldsymbol{x}_i] &= f(\boldsymbol{x}_i, t_i) \\ &= \mathcal{m}(\boldsymbol{x}_i) + \tau(\boldsymbol{x}_i)t_i\end{aligned} \tag{6}$$

where both $\mathcal{m}(.)$ and $\tau(.)$ represent stochastic functions with BART priors, namely, $\mathcal{m} \sim \mathrm{BART}(\boldsymbol{\theta}\,|\,\widehat{\mathrm{PS}}(\boldsymbol{x}_i), \boldsymbol{x}_i)$ and $\tau \sim \mathrm{BART}(\boldsymbol{\vartheta}\,|\,\boldsymbol{x}_i)$, and $\widehat{\mathrm{PS}}(\boldsymbol{x}_i)$ indicates the estimated PS. The two vectors $\boldsymbol{\theta}$ and $\boldsymbol{\vartheta}$ collect the hyperparameters regulating the number of trees in the BART ensembles, their depth, and the splitting rule associated with each single tree (see

[25] $f(\mathbf{X}_i, T_i)$ could be specified. as a fully parametric function, although this would inevitably constraint the cross-farm technological and behavioral heterogeneity. Admitting an arbitrarily complex function is thus more consistent with the assumption of a farm-specific production set F_i.

Appendix F for details). As previously mentioned, the specification in equation [6] allows regularizing $\tau(\boldsymbol{x}_i)$ directly and independently, thereby reducing the noisiness of the IATEs with respect to the same estimates obtained from simple differences in conditional mean functions. Furthermore, the additive nature of equation [6] ensures that the prior on $f(\boldsymbol{x}_i, t_i)$ is also a BART (Chipman, George, and McCulloch 2010; Hill, Linero, and Murphy 2020). Finally, notice that the model presented in equation [6] also appears in Nie and Wager (2021), who propose a frequentist approach to estimating $\tau(\boldsymbol{x}_i)$. In contrast to the setup discussed above, however, the authors propose a residuals-on-residuals reparameterization of equation [6] which is then used to obtain (regularized) consistent estimates of $\tau(\boldsymbol{x}_i)$ via a two-stage optimization procedure.

The full Bayesian model requires the definition of a likelihood function for the outcome variable (Gelman et al. 2013; McElreath 2020). Consistent with equation [5] and Chipman, George, and McCullogh (2010), Hill (2011), and Hahn, Murray, and Carvalho (2020), we employ a normal model for Y_i, along with a semiconjugate inverse chi square prior for its variance:

$$
\begin{aligned}
&Y_i \sim \text{Normal}(\mathcal{m}(\boldsymbol{x}_i) + \tau(\boldsymbol{x}_i)t_i, \sigma^2)\\
&\mathcal{m} \sim \text{BART}(\boldsymbol{\theta} \mid \widehat{\text{PS}}(\boldsymbol{x}_i), \boldsymbol{x}_i\\
&\tau \sim \text{BART}(\boldsymbol{\vartheta} \mid \boldsymbol{x}_i)\\
&\sigma^2 \sim \text{Inv}\chi^2(\omega) \qquad [7]
\end{aligned}
$$

where ω is set following Chipman, George, and McCullogh (2010) (see Appendix F for further details). Samples from the posterior distribution of $\tau(\boldsymbol{x}_i)$ are obtained via Markov chain Monte Carlo sampling, as implemented in the `R` package `bcf`. We indicate posterior draws from $\phi(\tau(\boldsymbol{x}_i) \,|\, \boldsymbol{x}_i, t_i, y_i, \ldots, y_N)$ as $\{\tau^s(\boldsymbol{x}_i)\}_{s=1}^{S}$, where S indicates the number of Markov chain Monte Carlo simulations.

Subgroup Search via Shallow Regression Trees

The approximated posterior $\{\tau^s(\boldsymbol{x}_i)\}_{s=1}^{S}$ is a multivariate probability distribution over a complex P-dimensional function, and as such, it might be difficult to interpret directly. One way to compress such information consists of obtaining marginal distributions of the IATEs for one covariate of interest and plotting them against the full range of that variable. A similar approach was adopted by Stetter, Mennig, and Sauer (2022), who used Shapley values (Shapley 1953) to identify the marginal contributions of several treatment effect drivers and used these indicators to construct partial dependence plots. Another sensible approach to investigating IATE heterogeneity consists of comparing farm subgroups obtained by projecting the full posterior distribution onto a lower-dimensional covariate space. In this respect, we follow the work of Yeager et al. (2019), Hahn, Murray, and Carvalho (2020), Woody, Carvalho, and Murray (2021) and (and partially Lee, Bargagli-Stoffi, and Dominici 2020), who suggest eliciting the relvant subgroups by partitioning the IATE maximum a posteriori (MAP) estimates, $\check{\tau}_i = S^{-1}\sum_{s=1}^{S} \tau_q^s(\boldsymbol{x}_i)$, using shallow regression trees (CART) (Breiman 1984). Specifically, the authors propose to split $\check{\tau}_i$ along $\boldsymbol{w}_i$, where $\boldsymbol{w}_i \subseteq \boldsymbol{x}_i$ indicates a vector of policy-relevant variables and setting $\boldsymbol{w}_i \subset \boldsymbol{x}_i$ implies using domain knowledge to enforce an initial regularization of the resulting tree. We restrict our attention to a subset of simple and understandable characteristics that policy makers might find helpful to improve the targeting of AEMs (see Section 6). Once farm subgroups have been identified, GATEs can be obtained as weighted averages of the IATEs that fall into each cluster. This approach to calculating GATEs is also consistent with Lechner (2018) in that group-level effects are obtained as convex combinations of the IATEs. In our application, however, weighting is automatically performed when fitting a tree to $\check{\tau}_i$.

Finally, for some potential effect moderator $x_p \in \boldsymbol{x}$, the comparison between pointwise estimates (or intervals) computed at different levels of x_p ignores any potential correlation between IATEs along other variables x_l, for all p, $l \in \{1,\ldots,P\}$. In other words, the marginal distribution of $\tau(x_{p,i})$ disregards the information encoded in the correlation between $\tau(x_{l,i})$ and $\tau(x_{l,-i})$ when $x_{l,i}$ and $x_{l,-i}$ are close. This might lead to misleading comparisons along

x_p and, consequently, unreliable policy implications. Therefore, once the relevant subgroups have been identified, one can obtain the full posterior distribution of each pairwise difference as: $\phi_{g_1,g_2} = \phi(\tau_{i|i\in g_1} - \tau_{i|i\in g_2})$, where g_1 and g_2 indicate any two subsets of $\check{\tau}_i$.

6. Results

IATEs

The two graphs numbered "1" in Figure 1 display the MAP; that is, the average over the S samples from the posterior distribution of $\tau(\boldsymbol{x}_i)$ estimates and corresponding 95% confidence intervals (CrI) of the IATEs over the two treatment comparisons. These are ordered across the respective samples from the lowest to the highest individual value. We start our discussion by presenting the results for T_2, the treatment that is more frequently addressed by the literature. First, it is worth noting that overall, the modal direction of the responses to the treatment (T_2) is fully consistent with theoretical expectations: adding the AEM to the environmental standards implied by the CC and the GP (Figure 1a [1]) induces an improvement in the FADN-AFI, that is, in the farm-level environmental performance. The opposite response is observed when the environmental standards implied by the CC and GP are dropped (i.e., treatment T_1) (Figure 1b [1]). Whereas in the first case, most estimated IATEs exhibit CrI not including zero (black dots), the converse applies to the second comparison group, for which a large proportion of farms have inconclusive individual-level TEs (light gray dots). The first graph in Figure 1b also indicates that some farms might even exhibit opposite responses, although the corresponding IATEs appear quite noisy. This evidence is presented in greater detail in Table 4, which provides descriptive summaries of our main results.

The two graphs numbered "2" in Figure 1 show the IATE's MAP frequency distribution for the two cases. These plots highlight the variability of the responses, with few cases showing a treatment effect direction that conflicts with the expected direction (despite exhibiting CrI including positive and negative values). Apart from these rare extreme cases, however, our MAP estimates range between roughly 0.1 and 1.0 for treatment T_2 and between approximately −3 and 1.5 for treatment T_1. The nature and determinants of these different patterns can be further investigated by estimating GATEs, as addressed in the next section.

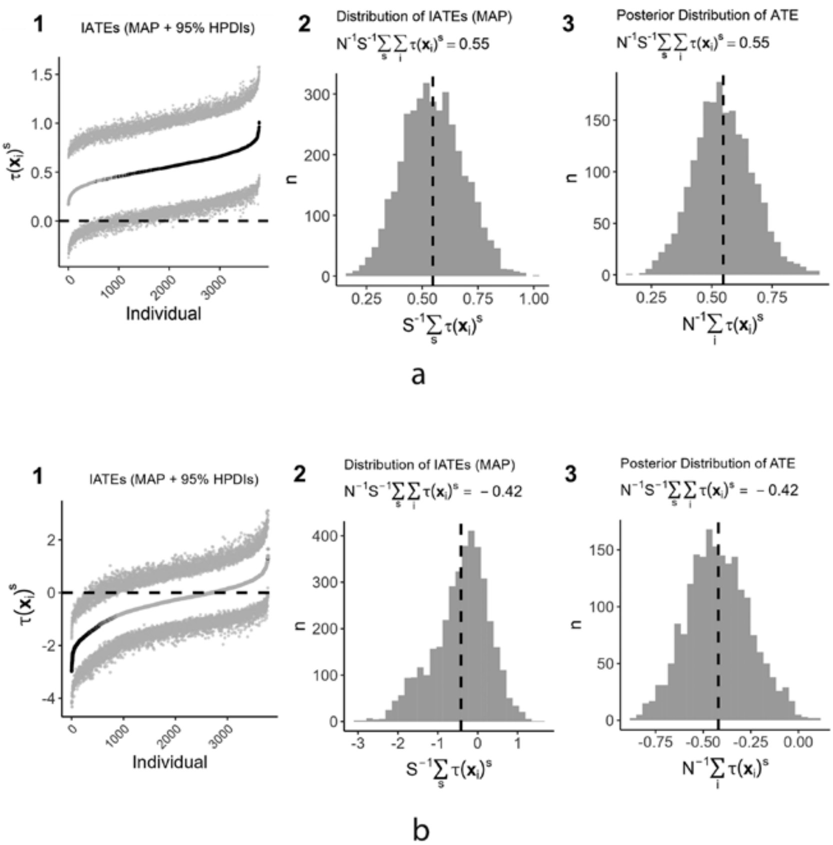
The irregularity of farms' responses to the treatments is a clear sign of heterogeneity, one that would be lost by the mere inspection of ATEs (see the two graphs numbered "3" in Figure 1). Whereas these latter aggregated estimands provide clear indications of policy effectiveness (as both show an effect in the expected direction), the inspection of the IATEs tells a different and more subtle story. This is especially true for the treatment T_1, whereas the responses seem more homogeneous when studying the treatment effect of implementing CC and GP requirements together with AEMs (treatment T_2).

Finally, for each individualized treatment effect, we calculate the posterior probability that the corresponding IATE is either greater than zero or lower than zero for the T_2 and T_1, respectively. Our results show that when comparing farms implementing CC and GP requirements plus AEMs with the control group, most of the IATEs' posterior distributions lie above zero. For example, the proportions of IATEs with at least 60%, 75%, and 90% positive posterior are 100%, 88.5%, and 5%, respectively. Conversely, when comparing the control group to farms with no adherence to CC or GP, the posterior distributions of their IATEs are largely negative. In this case, the proportions of IATEs with at least 60%, 75%, and 90% negative posterior are 83%, 15%, and 0%, respectively.

Notice that all the results discussed thus far are based on observations satisfying the common support as defined by rule I in Appendix E. Under such a restriction to the range of $\boldsymbol{X}_i$, however, our dataset does not suffer drops. The sensitivity of these figures to different exclusion rules is discussed in Section 6 (robustness check), in which the selection method we used based on the estimated PS is also discussed.

Figure 1

Estimated Individualized Average Treatment Effects: (a) Treatment Group T_2 (Farmers Implementing Agri-environmental Measures); (b) Treatment Group T_1 (Farmers Not Fulfilling Conditionality Restraints or Implementing Agri-environmental Measures)



Note: 1 = point estimates of individualized average treatment effect (median line) ordered from the smallest to the largest, with the upper and lower dots representing the posterior 95% CrI endpoints associated with each individual MAP point estimate; 2 = distribution of the MAP point estimate displayed in 1; 3 = Monte Carlo approximation of the full posterior distribution of average treatment effect defined in equation [4].
CrI = credible interval; MAP = maximum a posteriori.

Table 4

Individualized Average Treatment Effects Estimates for Model [6]

Treatment	$0 \notin$ CrI	MAP > 0	MAP < 0	ATE > 0	ATE < 0	ATE	CrI ATE Lower	CrI ATE Upper
All Observations								
	(%)	(%)	(%)	(%)	(%)	(AFI)	(AFI)	(AFI)
T_2	72.3	100	0	100	0	0.55	0.31	0.79
T_1	15.3	28.5	71.5	0.5	99.5	–0.42	–0.73	–0.08
Observations with $0 \notin$ CrI								
T_2	1	100	0	100	0	0.61	0.364	0.86
T_1	1	0	100	0	100	–0.85	–1.02	–0.46

Note: $0 \notin$ CrI = proportion of IATE CrI that do not include zero; MAP > (<) 0 = proportion of IATE with MAP > (<) than zero; ATE > (<) 0 = posterior probability that ATE (as defined in equation [4]) is > (<) than zero; CrI ATE = 95% CrI for ATE. AFI = agri-environmental footprint index; ATE = average treatment effect; CrI = credible interval; MAP = maximum a posteriori.

GATEs

We partition the posterior distribution of $\tau(\boldsymbol{x}_i)$ using a set of policy-relevant measures $\boldsymbol{w}_i$ covering the most relevant dimensions of heterogeneity, as evidenced by the measures of feature importance produced by the BCF. Our characterization of $\boldsymbol{w}_i$ involves (1) examining the variable importance metrics generated as a by-product of the fitting model [7],[26] and (2) choosing the 10 most predictive dimensions that policy makers might target to improve the effectiveness of AEMs. We fit a CART algorithm to $\check{\tau}_i$ using the attributes selected using the procedure illustrated above: latitude, longitude, altitude (geographical location); total arable land, share of rented land, revenue (physical or economic size); farm specialization (relative importance of the first and second crop, farms specialized in livestock, crop and livestock farms, farms specialized in annual crops, and farm specialized in perennial crops). The results for the two treatments are shown in Figures 4 and 5, wherein, for the sake of interpretability, we do not allow the trees to split more than three times.

When we consider the adoption an AEM in addition to CC and GP requirements (treatment T_2) (Figure 2b [1]), we find that TE heterogeneity is mostly associated with five variables: latitude, physical farm size, altitude, crop specialization (share of the second crop in the crop mix), and livestock intensity. These covariates trace out eight subgroups with different levels of treatment effects. For example, subgroup g_8 exhibits the lowest treatment effect and consists of farms in southern Italy with less than 85 ha arable land. On the opposite end of the spectrum, we find subgroup g_{15}, which comprises crop-specialized farms in northern Italy with low livestock intensity. One can then obtain the full posterior distribution of $g_{15} - g_8$ with 95% CrI between –0.27 and 0.49 (Figure 2b [2]), which indicates that the difference between the two subgroups is in fact small, if not zero. Interestingly, if we repeat this exercise across all the leaves defined by the tree in Figure 2b [1], no group differences emerge (see Appendix G). These results are consistent with our discussion in Section 6 (i.e., our preliminary findings suggested limited treatment effect heterogeneity for treatment T_2).

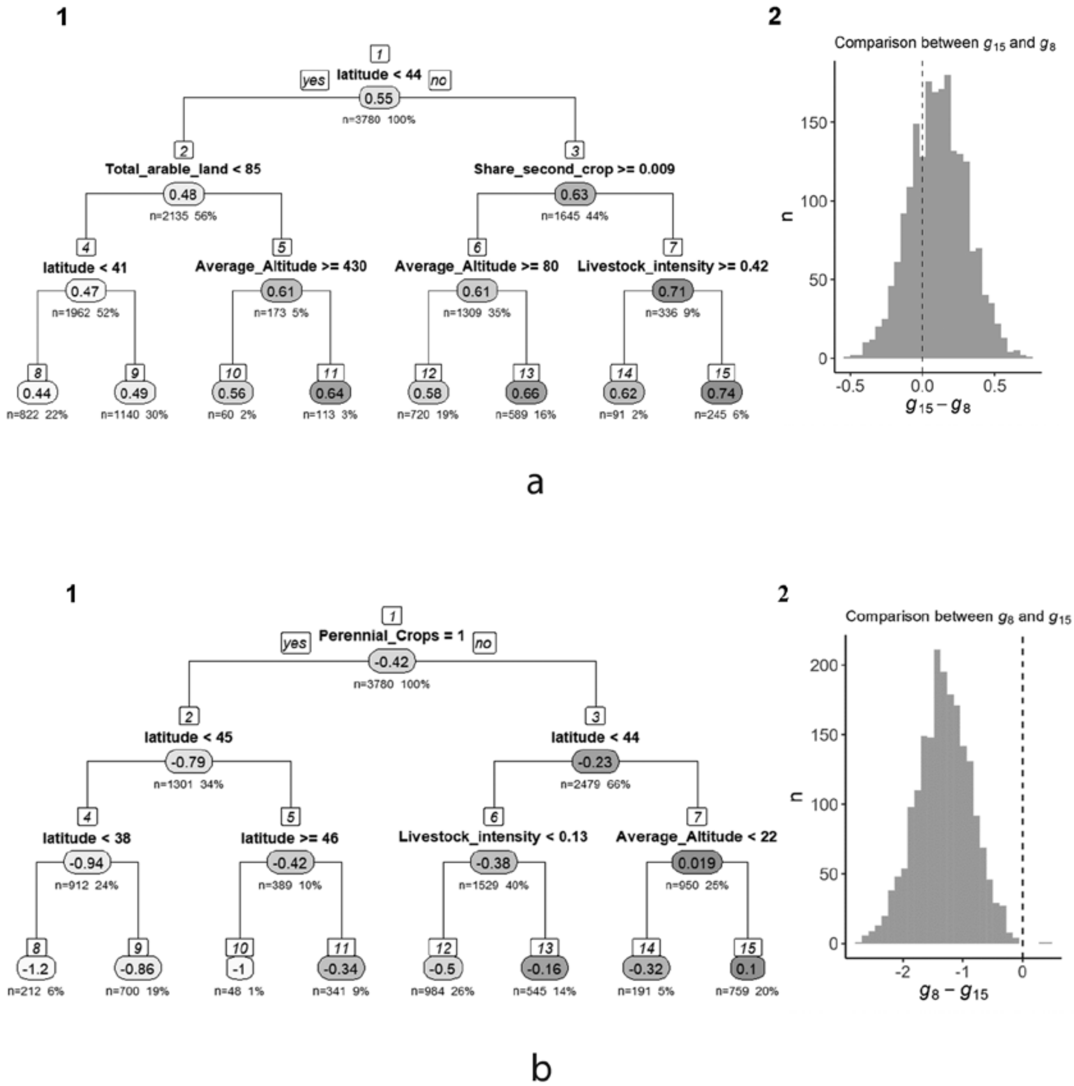
In the case of treatment T_1 (Figure 2a [1]), we see that the shallow tree picks up four moderating variables: specialization in perennial crops, latitude, altitude, and livestock intensity. In this case, the subgroup with the strongest TE is g_8, which consists of farms specialized in perennial crops in Italy's southmost regions. Subgroup g_{15} includes observations from farms in the Po Valley that are not specialized in perennial crops. The difference in TE between these subgroups lies approximately between –2.2 and –0.41 (95% CrI; Figure 2a [2]), indicating the presence of treatment effect heterogeneity. Repeating this exercise across all the terminal nodes, we find that unlike treatment T_2, when the treatment consists of dropping both CC and GP requirements, many groups exhibit diversified responses. These further details are provided in Appendix G, where we also provide a deeper tree to gain further insights into these HTEs and a graphic representation of the geographical distribution of the IATEs.

It is finally worth stressing that although our main goal is to explore which observable farm characteristics exhibit a greater heterogeneity of response, some of these features might not be easily addressed by AEPs due to cost constraints or infeasibility or because they could potentially lead to discriminatory outcomes. From a policy perspective, it would be more useful to evaluate the level of heterogeneity associated with covariates that can be targeted more easily and effectively through policy measures. Most of the geographical features considered in our study, along with variables indicating long-term farm production specialization, appear particularly suitable for this purpose. In this respect, our results confirm that most of these geographical features significantly contribute to the observe heterogeneity of response.

[26] The importance metric is obtained from a BART that includes the PS (PS-BART). Unlike the algorithm in equation [7], the PS-BART does not distinguish the prognostic from the treatment effect component. However, in terms of variable importance, the difference between the two techniques is negligible.

Figure 2

Shallow Regression Tree Fitted to the Maximum a Posteriori (MAP) Individualized Average Treatment Effects: (a) Treatment Group T_2; (b) Treatment Group T_1



Note: 1 = structure of the penalized regression tree; 2 = posterior distribution for the comparison between subgroups 8 and 15.

Similarly, the presence of perennial crops, crop specialization, and livestock density, all of which relate to distinct and consistent farming practices, pinpoint to patterns of strong heterogeneity. This suggests that AEPs could significantly enhance their effectiveness by specifically targeting these features. For a more detailed discussion on this matter, please refer to Appendix G.

Robustness Checks

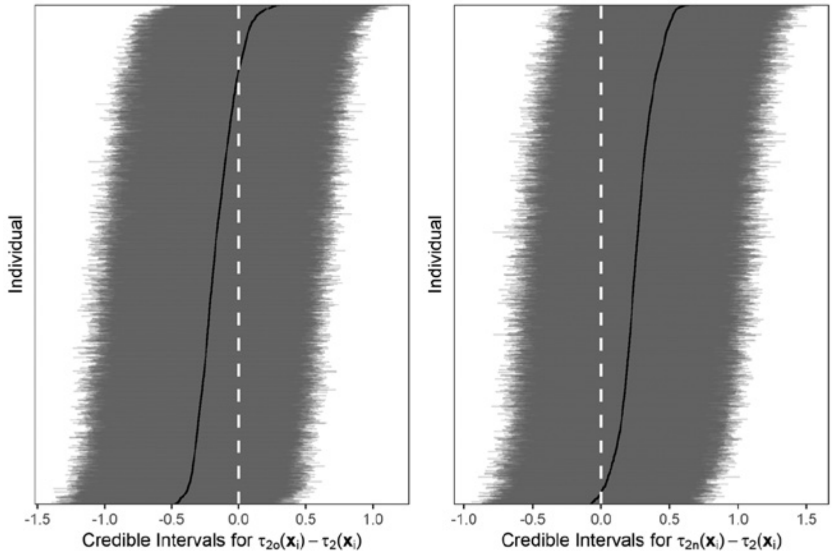
We check the consistency of our results to the assumptions formulated in Sections 4 and 5. Our first robustness check concerns the common support condition. As anticipated in Section 5 and further detailed in Appendix E, we use both the posterior distribution of the BART algorithm and a PS-based algorithm to investigate common support. Our tests show that the results presented in Section 6 are robust to these different methods to achieve overlap (see Appendix Tables H1 and H2).

We perform a battery of tests that largely encompass those discussed by Stetter, Mennig, and Sauer (2022) in that we reestimate our BCF multiple times, each time manipulating different model features. We begin by probing unconfoundedness through a recursive procedure in which we fit model [7] after dropping: (1) the most important feature in terms of relative frequency in the forest, (2) the three most important features, and (3) the five most important features. As detailed in Appendix Figures H1–H3), this exercise yields the first indication that the BCF in equation [7] is fundamentally resilient against unobserved heterogeneity as long as this is associated with the set of observed confounders. Put differently, the complex interactions and nonlinearities generated by the tree ensemble seem to work as additional synthetic controls associated with the left-out covariates, thus compensating for their absence in the model. However, this line of reasoning hinges on the (strong) assumption that the most predictive features are also associated with both Y and T_k. In case this assumption fails, the procedure discussed above cannot be interpreted as a robustness check for unconfoundedness. For this reason, we build on these preliminary results and devise an additional test targeting endogenous unobserved heterogeneity directly. Our strategy consists of generating a random variable correlated with both Y and T_k, forming the vector $\boldsymbol{Z}_{it}$ as described in Section 3, and rerunning the model. As shown in Appendix Figure H4, our results do not change substantially, even under a strong imposed association between the unobserved variable and (Y, T_k). This stability could result from the properties of the BART ensemble in that when the forest is dense, the marginal contribution of each covariate becomes increasingly small (Chipman, George, and McCulloch 2010). Alternatively, it could be that the correlation between the nonlinear interactions generated by the BCF and the new confounder is strong enough to prevent distortions in the IATEs. In either case, it is worth warning that treatment effect estimates might deteriorate quickly when unobserved heterogeneity is more abundant and complex. This test is in fact only restricted to a single unobserved factor, which we model as linearly associated with Y and T_k (i.e., through correlations, which do not necessarily imply a direct effect of the synthetic u_i on the outcome or the treatment). We thus expect that in presence of multiple endogenous latent confounders, possibly related to the treatment and the outcome (or other elements of $\boldsymbol{X}_i$ in a nonlinear fashion, our estimates might tun out sensibly different. Although the literature offers other methods to perform sensitivity analysis with respect to omitted confounders (Dorie et al. 2016; VanderWeele and Ding 2018), we believe that they either do not overcome the limitations discussed above or they remain difficult to implement in HTE estimation. Therefore, despite the promising results presented so far, we stress that these only hold if several important restrictions are met.

The following robustness check consists of creating both a placebo treatment and a placebo outcome, replacing their observed counterparts in equation [7], and fitting the model two more times. If the model is correctly specified, the IATEs resulting from these "fake" variables should be uncorrelated with $\tau(\boldsymbol{x}_i)$. As Appendix Figures H5 and H6 show, the new results obtained through placebo treatments and outcomes not only have no correlation with our estimated IATEs but also produce zero ATE with minimal treatment effect heterogeneity.

Finally, we assess the robustness of the estimated IATEs with respect to the OTH problem discussed in previous sections. We proceed by replacing the FADN-AFI with its elementary components and reestimating model [7] as discussed. Appendix Figure H7 suggests that focusing on marginal indicators produces TEs whose individual directions are essentially in line with those presented in Section 6. For example, implementing AEMs seems to yield lower GHGs, higher crop diversity, lower fertilizer expenditure, and more woodland areas. Nonetheless, a noteworthy difference emerges in terms of treatment effect heterogeneity. Whereas adopting the FADN-AFI points to a limited diversity across farms, using marginal measurements would suggest that treatment T_2 is environmentally beneficial only when

Figure 3
Individualized Average Treatment Effects Differentials for the Two Versions of T_2



Note: The left graph shows 95% credible intervals for the distribution of $\tau_{2o}(\mathbf{x}_i) - \tau_2(\mathbf{x}_i)$ (gray lines) and corresponding posterior means (black line); the right graph shows 95% credible intervals for the distribution of $\tau_{2n}(\mathbf{x}_i) - \tau_2(\mathbf{x}_i)$ (gray lines) and corresponding posterior means (black line).

the treatment effect is large. For this reason, our results invite to caution when it comes to choosing the dependent variable of model [7]. Although addressing individual indicators may appear more attractive and interpretable, it is worth stressing that missing out on the potential correlation or interdependence among them can affect the TE estimates in a nontrivial way.

Role of Heterogeneous Treatments

As discussed in Sections 3 and 4, one potential limitation of our results (as well as other works investigating HTE of aggregated treatments) is that part of the estimated treatment effect heterogeneity in T_2 might be a statistical artifact. This would result from the fact that T_2 is a multiple-versions treatment as it aggregates two distinct measures which admit, in turn, several submeasures (see Section 4). As introduced in Section 5, the presence of treatment heterogeneity may affect our results by violating SUTVA (Heiler and Knaus 2022). Since in this case, the resulting interpretation of $\tau(\boldsymbol{x})$ would be misleading, we reestimate model [7] replacing T_2 with the two respective measures (measure 11 and measure 10) and approach the problem from a multiple-versions treatment perspective as discussed in Lopez and Gutman (2017).

To assess the possible bias in HTE estimation due to treatment heterogeneity, we compare the posterior distribution of the IATEs presented in Section 6 with the posterior density of the IATEs estimated using either T_{2o}, $\tau_{2o}(\boldsymbol{x}_i)$, or T_{2n}, $\tau_{2n}(\boldsymbol{x}_i)$. Figure 3 shows the 95% CrI for the differences $\tau_{2o}(\boldsymbol{x}_i) - \tau_2(\boldsymbol{x}_i)$ and $\tau_{2n}(\boldsymbol{x}_i) - \tau_2(\boldsymbol{x}_i)$, respectively, where $\tau_2(\boldsymbol{x}_i)$ indicates the IATE for individual i under treatment T_2. As we can see from these plots, the difference between our initial estimates and those obtained by substituting T_2 with T_{2o} are minimal. Indeed, although $\tau_{2o}(\boldsymbol{x}_i)$ is on average (black line in the left graph in Figure 3) slightly smaller than $\tau_2(\boldsymbol{x}_i)$ for all $i \in N_o$, where N_o indicates the number of units choosing T_{2o}, all the CrI include both positive and negative values. At the same time, when focusing on T_{2n}, we see that $\tau_{2n}(\boldsymbol{x}_i) - \tau_2(\boldsymbol{x}_i)$ are on average higher than zero for all $i \in N_n$, where N_n indicates the units choosing T_{2n}. However, the CrI once again includes zero for all such comparisons, although they are all moderately skewed toward positive values. Moreover, as mentioned in Section 4, T_{2n} could still entail some degree of treatment heterogeneity, which recommends caution when interpreting the corresponding estimates. Overall, examining the two measures separately highlights that the posterior distribution of the IATEs does not seem to change markedly when the aggregated (T_2) or the disaggregated (T_{2o}, T_{2n}) treatment is considered. This would suggest a limited impact of treatment heterogeneity on our interpretation of the HTEs discussed above. Nonetheless, further research effort remains desirable to better clarify the possible role of multiple versions in the correct identification and estimation of the HTE.

7. Concluding Remarks

Giving the CAP a more explicit environmental orientation and justification has been at the core of all its recent reforms. This necessarily means shifting the support from undifferentiated and unconditional payments to more tailored and target measures. The efficiency and effectiveness of AEPs in this respect critically depend on how farmers respond to these measures. This response, in turn, largely depends on the individual characteristics of supported farms. This makes the response itself highly heterogeneous and, consequently, suggests that there is still room for substantial improvement through better policy targeting.

In this article, we present a CML approach to assessing the heterogeneous response of farmers to different AEPs implemented through the 2015–2020 CAP reform. Building on the existing literature, this study's main contribution is twofold. First, we explicitly conceptualize and investigate the different sources of heterogeneity that we expect influence farms' environmental performances under such policies. Second, we take advantage of the most recent developments in Bayesian nonparametrics and conduct the analysis using a relatively unexplored algorithm called Bayesian causal forest. This method allows using the posterior distribution of the individualized treatment effect (the IATEs) to draw inferences about arbitrary transformations of these highly disaggregated estimands. We leverage this property, particularly when discussing group-level treatment effects and testing the robustness of our results against identification assumptions.

More generally, estimating IATEs can prove insightful in that some beneficiaries of an AEP may exhibit limited or unsatisfactory responses, thereby calling for an intensification of the support, while others may show responses that are well beyond the policy target, suggesting a reduction of support. Our results illustrate how informative the approach can be in detecting the extent, nature, and source of this heterogeneous response. For instance, we demonstrate that contrasting different farm subgroups can provide additional information on the nature of the heterogeneous response. Specifically, we highlighted that the treatment effect from implementing pillar 2 agri-environmental measures and fulfilling pillar 1 conditionality requirements seems more homogeneous than the response to adopting none of the above.

The primary policy implication of our results concerns the need for a better targeting of AEPs. In this respect, caution is necessary, as not all farm characteristics considered can be easily targeted due to practical or political constraints. Nonetheless, our analysis suggests that significant heterogeneity in treatment effects is concentrated in farm subgroups that can be feasibly targeted. These subgroups often involve geographical features and specific production specializations. Therefore, delivering some CAP measures at a local scale and tailoring them to specific production orientations, along with broader adoption of results-based payment schemes, may represent a sensible initial step toward better targeting. The new CAP acknowledges greater flexibility for member states through the new delivery model, allowing them to address the environmental aspects of pillar 1 (the reinforced CC and the eco-schemes replacing the GP) and the AEMs in pillar 2 more effectively. In principle, this flexibility seems to go along with the goal of improved targeting for these AEPs.

Although our empirical results provide valuable insights, our work also contributes to the constructive discussion on the potential and limitations of these relatively new policy assessment methods. How useful is CML and the analysis of heterogeneous treatment effects in informing policy improvements related to the CAP? Our conceptual framework and empirical investigation suggest that they can be useful. However, as with all emerging econometric approaches, several issues require careful consideration.

Because standard causal ML methods cannot be used for policy analysis without additional identifying restrictions and assumptions, selecting appropriate confounders and ensuring overlapping/treatment-stable units necessitates a solid theoretical understanding of treatment selection mechanisms. Developing these conceptual foundations also facilitates result interpretation, as the complex output of these estimation methods can be challenging to put into perspective. Among the standard assumptions presented herein, unconfoundedness and the stable unit treatment value are often regarded as restrictive. Although the former can be corroborated via robustness checks and the use of ML algorithms, the latter finds little practical help from flexible estimation techniques and thus remains debatable. In this respect, specifying the correct treatment variable(s) is quintessential for an unbiased interpretation of the resulting treatment effect, an aspect that is still relatively underdiscussed in the literature.

More generally, investigating the effectiveness of CAP's agri-environmental policies in a binary-treatment logic may prove limiting when the analysis targets heterogeneous causal effects. The risk is that the elicited estimates do not entirely reflect farms' heterogeneous responses to a treatment but encapsulate the heterogeneity of the treatment itself. Besides the prototypical case of multiple-versions treatments (whether hidden or observable), problems can also arise when a policy measure is not only adopted (i.e., a discrete choice) but also exhibits different intensity levels in different cohorts of farms. In such cases, binary treatments should be extended to incorporate dosage information. How to define the treatment intensity (i.e., the "dose") of different agri-environmental policies is an ambitious empirical question that we leave to future research.

References

Afriat, S. N. 1972. "Efficiency Estimation of Production Function." *International Economics Review* 13 (3): 568–98.

Angrist, J. D., and J. S. Pischke. 2008. *Mostly Harmless Econometrics*. Princeton, NJ: Princeton University Press.

Arata, L., and P. Sckokai. 2016. "The Impact of Agri-environmental Schemes on Farm Performance in Five EU Member States: A DID-Matching Approach." *Land Economics* 92 (1): 167–86. https://doi.org/10.3368/le.92.1.167.

Athey, S., and G. W. Imbens. 2016. "Recursive Partitioning for Heterogeneous Causal Effects." *Proceedings of the National Academy of Sciences* 113 (27): 7353–60.

———. 2019. "Machine Learning Methods that Economists Should Know About." *Annual Review of Economics* 11: 685–725.

Athey, S., J. Tibshirani, and S. Wager. 2019. "Generalised Random Forests." *Annals of Statistics* 47 (2): 1148–78.

Baldoni, E., S. Coderoni, and R. Esposti. 2021. "Immigrant Workforce and Agriculture Productivity: Evidence from Italian Farm-Level Data," *European Review of Agricultural Economics* 48 (4): 805–34.

Bertoni, D., G. Aletti, D. Cavicchioli, A. Micheletti, and R. Pretolani. 2021. "Estimating the CAP Greening Effect by Machine Learning Techniques: A Big Data Ex-Post Analysis." *Environmental Science and Policy* 119: 44–53.

Bertoni, D., D. Curzi, G. Aletti, and A. Olper. 2020. "Estimating the Effects of Agri-environmental Measures Using Difference-in-Difference Coarsened Exact Matching." *Food Policy* 90: 101790. https://doi.org/10.1016/j.foodpol.2019.101790.

Breiman, L. 1984. *Classification and Regression Trees*. New York: Routledge.

Brown, C., E. Kovács, I. Herzon, S. Villamayor-Tomas, A. Albizua, A. Galanaki, I. Grammatikopoulou, *et al*. 2021. "Simplistic Understandings of Farmer Motivations Could Undermine the Environmental Potential of the Common Agricultural Policy." *Land Use Policy* 101: 105136. https://doi.org/10.1016/j.landusepol.2020.105136.

Burton, R., and G. Schwarz. 2013. "Result-Oriented Agri-environmental Schemes in Europe and Their Potential for Promoting Behavioural Change." *Land Use Policy* 30 (1): 628–41.

Carnagie, N., V. Dorie, and J. J. Hill. 2019. "Examining Treatment Effect Heterogeneity Using BART." *Observational Studies* 5 (2): 52–70.

Carvalho, C., A. Feller, J. Murrary, S. Woody, and D. Yeager. 2019. "Assessing Treatment Effect Variation in Observational Studies: Results from a Data Challenge." *Observational Studies* 5 (2): 21–35.

Chabé-Ferret, S., and J. Subervie. 2013. "How Much Green for the Buck? Estimating Additional and Windfall Effects of French Agri-environmental Schemes by DID-Matching." *Journal of Environmental Economics and Management* 65 (1): 12–27.

Chavas, J. P., and T. L. Cox. 1995. "On Nonparametric Supply Response Analysis." *American Journal of Agricultural Economics* 77 (1): 80–92.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. 2018. "Double/Debiased Machine Learning for Treatment and Structural Parameters." *Econometrics Journal* 21 (1): C1–68. https://doi.org/10.1111/ectj.12097.

Chipman, H. A., E. I. George, and R. E. McCulloch. 1998. "Bayesian CART Model Search." *Journal of the American Statistical Association* 93 (443): 935–48.

———. 2010. "Bart: Bayesian Additive Regression Trees." *Annals of Applied Statistics* 4 (1): 266–98.

Chipman, H. A., and H. Gu. 2005. "Interpretable Dimension Reduction." *Journal of Applied Statistics* 32 (9): 969–87.

Coderoni, S., and R. Esposti. 2018. "CAP Payments and Agricultural GHG Emissions in Italy. A Farm-Level Assessment." *Science of the Total Environment* 627: 427–37.

Coderoni, S., J. Helming, M. Pérez-Soba, P. Sckokai, and A. Varacca. 2021. "Key Policy Questions for Ex-Ante Impact Assessment of European Agricultural and Rural Policies." *Environmental Research Letters* 16 (9): 094044. https://doi.org/0.1088/1748-9326/ac1f45.

Commission of European Communities. 2000. "Communication from the Commission to the Council and the European Parliament. Indicators for the Integration of Environmental Concerns into the Common Agricultural Policy." Brussels.

Dabkiene, V., T. Balezentis, and D. Streimikiene. 2021. "Development of Agri-environmental Footprint Indicator Using the FADN Data: Tracking Development of Sustainable Agricultural Development in Eastern Europe." *Sustainable Production and Consumption* 27: 2121–33.

Dessart, F. J., J. Barreiro-Hurlé, and R. van Bavel. 2019. "Behavioural Factors Affecting the Adoption of Sustainable Farming Practices: A Policy-Oriented Review." *European Review of Agricultural Economics* 46 (3): 417–71.

Dorie, V., M. Harada, N. B. Carnegie, and J. Hill. 2016. "A Flexible, Interpretable Framework for Assessing Sensitivity to Unmeasured Confounding." *Statistics in Medicine* 35 (20): 3453–70.

Dorie, V., J. Hill, U. Shalit, M. Scott, and D. Cervone. 2019. "Automated versus Do-It-Yourself Methods for Causal Inference: Lessons Learned from a Data Analysis Competition." *Statistical Science* 34 (1): 43–68.

Ehlers, M. H., R. Huber, and R. Finger. 2021. "Agricultural Policy in the Era of Digitalisation." *Food Policy* 100: 102019. https://doi.org/10.1016/j.foodpol.2020.102019.

Erjavec, K., and E. Erjavec. 2015. "Greening the CAP: Just a Fashionable Justification? A Discourse Analysis of the 2014–2020 CAP Reform Documents." *Food Policy* 51: 53–62.

Esposti, R. 2000. "The Impact of Public R&D and Extension Expenditure on Italian Agriculture. An Application of a Mixed Parametric/Nonparametric Approach." *European Review of Agricultural Economics* 27 (3): 365–84.

———. 2017a. "The Empirics of Decoupling: Alternative Estimation Approaches of the Farm-Level Production Response." *European Review of Agricultural Economics* 44 (3): 499–537.

———. 2017b. "The Heterogeneous Farm-Level Impact of the 2005 CAP-First Pillar Reform: A Multivalued Treatment Effect Estimation." *Agricultural Economics* 48 (3): 373–86.

———. 2022a. "The Coevolution of Policy Support and Farmers Behaviour and Performance. An Investigation on Italian Agriculture over the 2008–2019 Period." *Bio-Based and Applied Economics* 11 (3): 231–64.

———. 2022b. "Non-monetary Motivations of Agri-environmental Policies Adoption. A Causal Forest Approach." Quaderno di Ricerca no. 459. Ancona, Italy: Dipartimento di Scienze Economiche e Sociali, Università Politecnica delle Marche.

European Council. 2009. "Council Regulation (EC) No. 1217/2009 of 30 November 2009: Setting Up a Network for the Collection of

Accountancy Data on the Incomes and Business Operation of Agricultural Holdings in the European Community." OJ L 328. Brussels.

Finn, J. A., F. Bartolini, D. Bourke, I. Kurz, and D. Viaggi. 2009. "Ex-post Environmental Evaluation of Agri-environment Schemes Using Experts' Judgements and Multicriteria Analysis." *Journal of Environmental Planning and Management* 52: 717–37.

Gelman, A., H. S. Stern, J. B. Carlin, D. B. Dunson, A. Vehtari, and D. B. Rubin. 2013. *Bayesian Data Analysis*. Boca Raton, FL: Chapman and Hall/CRC.

Guerrero, S. 2021. "Characterising Agri-environmental Policies: Towards Measuring Their Progress." OECD Food, Agriculture and Fisheries Paper 155. Paris: OECD.

Hahn, P. R., C. M. Carvalho, D. Puelz, and J. He. 2018. "Regularisation and Confounding in Linear Regression for Treatment Effect Estimation." *Bayesian Analysis* 13 (1): 163–82.

Hahn, P. R., J. S. Murray, and C. M. Carvalho. 2020. "Bayesian Regression Tree Models for Causal Inference: Regularisation, Confounding, and Heterogeneous Effects (With Discussion)." *Bayesian Analysis* 15 (3): 965–1056.

Heiler, P., and M. Knaus. 2022. "Effect or Treatment Heterogeneity? Policy Evaluation with Aggregated and Disaggregated Treatments." IZA Discussion Paper 15580. Bonn, Germany: IZA.

Hill, J. L. 2011. "Bayesian Nonparametric Modelling for Causal Inference." *Journal of Computational and Graphical Statistics* 20 (1): 217–40.

Hill, J., A. Linero, and J. Murray. 2020. "Bayesian Additive Regression Trees: A Review and Look Forward." *Annual Review of Statistics and Its Application* 7: 251–78.

Hill, J., and Y. S. Su. 2013. "Assessing Lack of Common Support in Causal Inference Using Bayesian Nonparametrics: Implications for Evaluating the Effect of Breastfeeding on Children's Cognitive Outcomes." *Annals of Applied Statistics* 7 (3): 1386–1420.

Hudec, B., C. Kaufmann, R. Landgrebe-Trinkunaite, and S. Naumann. 2007. "Evaluation of Soil Protection Aspects in Certain Programmes of Measures Adopted by Member States." Final Report. Brussels: European Commission.

Hünermund, P., B. Louw, and I. Caspi. 2023. "Double Machine Learning and Automated Confounder Selection: A Cautionary Tale." *Journal of Causal Inference* 11 (1): 20220078. https://doi.org/10.1515/jci-2022-0078.

Imai, K., and M. Ratkovic. 2013. "Estimating Treatment Effect Heterogeneity in Randomised Program Evaluation." *Annals of Applied Statistics 7*(1): 443–70.

Imbens, G. W., and D. B. Rubin. 2015. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge, UK: Cambridge University Press.

Kapelner, A., and J. Bleich. 2016. "bartMachine: Machine Learning with Bayesian Additive Regression Trees." *Journal of Statistical Software* 70 (4): 1–40.

Knaus, M. C., M. Lechner, and A. Strittmatter. 2021. "Machine Learning Estimation of Heterogeneous Causal Effects: Empirical Monte Carlo Evidence." *Econometrics Journal* 24 (1): 134–61.

———. 2022. "Heterogeneous Employment Effects of Job Search Programmes: A Machine Learning Approach." *Journal of Human Resources* 57 (2): 597–636.

Künzel, S. R., J. S. Sekhon, P. J. Bickel, and B. Yu. 2019. "Metalearners for Estimating Heterogeneous Treatment Effects Using Machine Learning." *Proceedings of the National Academy of Sciences* 116 (10): 4156–65. https://doi.org/10.1073/pnas.1804597116.

Lechner, M. 2018. "Modified Causal Forests for Estimating Heterogeneous Causal Effects." arXiv preprint arXiv:1812.09487. https://doi.org/10.48550/arXiv.1812.09487.

Lee, K., F. J. Bargagli-Stoffi, and F. Dominici. 2020. "Causal Rule Ensemble: Interpretable Inference of Heterogeneous Treatment Effects." arXiv preprint arXiv:2009.09036. https://doi.org/10.48550/arXiv.2009.09036.

Li, F., P. Ding, and F. Mealli. 2022. "Bayesian Causal Inference: A Critical Review." arXiv preprint arXiv:2206.15460. https://doi.org/10.48550/arXiv.2206.15460.

Lopez, M. J., and R. Gutman. 2017. "Estimation of Causal Effects with Multiple Treatments: A Review and New Ideas." *Statistical Science* 32 (3): 432–54.

McElreath, R. 2020. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. Boca Raton, FL: CRC Press.

Mennig, P., and J. Sauer. 2020. "The Impact of Agri-environment Schemes on Farm Productivity: A DID-Matching Approach." *European*

Review of Agricultural Economics 47 (3): 1045–93.

National Rural Network. 2010. "Report on Cross Compliance Implementation in Italy." Available at http://www.reterurale.it.

Nie, X., and S. Wager. 2021. "Quasi-oracle Estimation of Heterogeneous Treatment Effects." *Biometrika* 108 (2): 299–319.

ÓhUallacháin, D., J. A. Finn, B. Keogh, R. Fritch, and H. Sheridan. 2016. "A Comparison of Grassland Vegetation from Three Agri-environment Conservation Measures." *Irish Journal of Agricultural and Food Research* 55: 176–91.

OECD (Organisation for Economic Co-operation and Development) 2003. "Agriculture and Biodiversity: Developing Indicators for Policy Analysis." Proceedings from an OECD Expert Meeting, Zurich, November 2001. Paris.

Pascucci, S., T. de-Magistris, L. Dries, F. Adinolfi, and F. Capitanio. 2013. "Participation of Italian Farmers in Rural Development Policy." *European Review of Agricultural Economics* 40 (4): 605–31.

Pufahl, A., and C. R. Weiss. 2009. "Evaluating the Effects of Farm Programmes: Results from Propensity Score Matching." *European Review of Agricultural Economics* 36 (1): 79–101.

Purvis, G., G. Louwagie, G. Northey, S. Mortimer, J. Park, A. Mauchline, J. Finn, *et al.* 2009. "Conceptual Development of a Harmonised Method for Tracking Change and Evaluating Policy in the Agri-environment: The Agri-environmental Footprint Index." *Environmental Science and Policy* 12: 321–37.

Shapley, L. S. 1953. "A Value for n-Person Games." *Contributions to the Theory of Games* 2 (28): 307–17.

Stetter, C., P. Mennig, and J. Sauer. 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): 723–39.

Storm, H., K. Baylis, and T. Heckelei. 2020. "Machine Learning in Agricultural and Applied Economics." *European Review of Agricultural Economics* 47 (3): 849–92.

Uehleke, R., M. Petrick, and S. Hüttel. 2022. "Evaluations of Agri-environmental Schemes Based on Observational Farm Data: The Importance of Covariate Selection." *Land Use Policy* 114: 105950.

VanderWeele, T. J., and P. Ding. 2017. "Sensitivity Analysis in Observational Research: Introducing the E-Value." *Annals of Internal Medicine* 167 (4): 268–74.

VanderWeele, T. J., and M. A. Hernán. 2013. "Causal Inference under Multiple Versions of Treatment." *Journal of Causal Inference* 1 (1): 1–20.

Vanslembrouck, I., G. Van Huylenbroeck, and W. Verbeke. 2002. "Determinants of the Willingness of Belgian Farmers to Participate in Agri-environmental Measures." *Journal of Agricultural Economics* 53 (3): 489–511.

Varacca, A., L. Arata, E. Castellari, and P. Sckokai. 2023. "Does CAP Greening Affect Farms' Economic and Environmental Performances? A Regression Discontinuity Design Analysis." *European Review of Agricultural Economics* 50 (2): 272–303.

Varian, H. 1984. "The Nonparametric Approach to Production Analysis." *Econometrica* 52 (3): 579–97.

Wager, S., and S. Athey. 2018. "Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests." *Journal of the American Statistical Association* 113 (523): 1228–42. https://doi.org/10.1080/01621459.2017.1319839.

Wąs, A., A. Malak-Rawlikowska, M. Zavalloni, D. Viaggi, P. Kobus, and P. Sulewski. 2021. "In Search of Factors Determining the Participation of Farmers in Agri-environmental Schemes: Does Only Money Matter in Poland?" *Land Use Policy* 101: 105190. https://doi.org/10.1016/j.landusepol.2020.105190.

Wascher, D. M. 2003. "Overview of Agricultural Landscape Indicators across OECD Countries." Proceedings of the NIJOS/OECD Expert Meeting on Agricultural Landscape, 7–9 October 2002, Oslo, Norway. Available at https://research.wur.nl/en/publications/overview-on-agricultural-landscape-indicators-across-oecd-countri.

Westbury, D. B., J. R. Park, A. L. Mauchline, R. T. Crane, and S. R. Mortimer. 2011. "Assessing the Environmental Performance of English Arable and Livestock Holdings Using Data from the Farm Accountancy Data Network (FADN)." *Journal of Environmental Management* 92 (3): 902–9.

Woody, S., C. M. Carvalho, and J. S. Murray. 2021. "Model Interpretation through Lower-Dimensional Posterior Summarisation." *Journal* *of Computational and Graphical Statistics* 30 (1): 144–61.

Wooldridge, J. M. 2010. *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.

Yeager, D. S., P. Hanselman, G. M. Walton, J. S. Murray, R. Crosnoe, C. Muller, and C. S. Dweck. 2019. "A National Experiment Reveals Where a Growth Mindset Improves Achievement." *Nature* 573 (7774): 364–69.

Zhen, H., Y. Qiao, H. Zhao, X. Ju, R. Zanoli, M. A. Waqas, F. Lun, and M. T. Knudsen. 2022. "Developing a Conceptual Model to Quantify Eco-compensation Based on Environmental and Economic Cost-Benefit Analysis for Promoting Ecologically Intensified Agriculture." *Ecosystem Services* 56: 101442.

Zimmermann, A., and W. Britz. 2016. "European Farms' Participation in Agri-environmental Measures." *Land Use Policy* 50: 214–28.