UNIVERSITÀ POLITECNICA DELLE MARCHE

DIPARTIMENTO DI SCIENZE ECONOMICHE E SOCIALI

IS ORGANIC AGRICULTURE SUSTAINABLE? QUASI-EXPERIMENTAL EVIDENCE WITH HETEROGENEOUS EFFECTS FROM ITALIAN FARMS.

ROBERTO ESPOSTI

QUADERNO DI RICERCA n. 508
ISSN: 2279-9575

July 2026

Abstract

This paper evaluates the sustainability of organic farming by testing whether it achieves a win–win–win outcome, improving farm-level environmental and economic performance without compromising aggregate agricultural output. We adopt a quasi-experimental design that accounts for voluntary adoption and allows for heterogeneous treatment effects. We combine Random Forest and LASSO methods, enabling flexible, fully interactive modeling with high-dimensional covariates. The analysis is based on a balanced panel of Italian farms from the FADN dataset covering the period 2008–2022. The results reveal substantial heterogeneity in the effects of organic farming adoption. A win–win–win outcome does not emerge on average but is observed only for a small subset of farms; for other groups, the full sustainability of organic adoption appears questionable. These findings call for more targeted and differentiated policy support to promote organic farming.

Keywords: Sustainable Farming, Organic Agriculture, Heterogeneous Treatment Effect, Double Machine Learning.

JEL Classification: C14, C23, Q01, Q12.

Address:

Roberto ESPOSTI
Department of Economics and Social Sciences - Università Politecnica delle Marche, 60121, Ancona (Italy). Email: r.esposti@staff.univpm.it

Is Organic Agriculture Sustainable?

Quasi-Experimental Evidence with Heterogeneous Effects from Italian Farms.

Roberto Esposti
Department of Economics and Social Sciences, Università Politecnica delle Marche, Ancona, Italy; r.esposti@staff.univpm.it

Abstract

This paper evaluates the sustainability of organic farming by testing whether it achieves a win–win–win outcome, improving farm-level environmental and economic performance without compromising aggregate agricultural output. We adopt a quasi-experimental design that accounts for voluntary adoption and allows for heterogeneous treatment effects. We combine Random Forest and LASSO methods, enabling flexible, fully interactive modeling with high-dimensional covariates. The analysis is based on a balanced panel of Italian farms from the FADN dataset covering the period 2008–2022. The results reveal substantial heterogeneity in the effects of organic farming adoption. A win–win–win outcome does not emerge on average but is observed only for a small subset of farms; for other groups, the full sustainability of organic adoption appears questionable. These findings call for more targeted and differentiated policy support to promote organic farming.

Keywords: Sustainable Farming, Organic Agriculture, Heterogeneous Treatment Effect, Double Machine Learning.

JEL Classification: C14, C23, Q01, Q12.

1. Introduction

The literature and public debate on the actual contribution of organic farming to the long-term sustainability of the agricultural sector is extensive and long-standing (Chotai and Augustine, 2024; Akhuli, 2025; Reddy, 2024). Yet, despite decades of discussion, no unequivocal and broadly accepted empirical evidence has emerged (Coderoni et al., 2026). Existing findings remain controversial, fragmented, and often inconsistent.

From an empirical perspective, settling this debate would ideally require properly designed experiments. Although randomized experiments ideally represent the platinum standard in socio-economic policy evaluation (Carattini et al., 2024; Dur et al., 2025), laboratory experiments are clearly unsuitable for capturing the complex agronomic, economic, and environmental processes involved in agricultural production. Field experiments, in turn, are extremely costly and difficult to implement, as the effects of adopting organic practices typically unfold over long time horizons, often several years, during which maintaining control over internal and external conditioning factors becomes highly challenging. A further source of inconclusiveness lies in the intrinsically multidimensional nature of sustainability, which calls for an empirical assessment capable of accounting simultaneously for multiple, and potentially conflicting, outcome dimensions.

These issues motivate the adoption of a quasi-experimental logic and a research design capable of approximating the conditions of an experiment using real-world observational data. In this perspective, a heterogeneous Treatment Effects (TE) framework represents a promising avenue for uncovering how the adoption of organic farming causally affects the sustainability dimensions in different ways. However, this approach also raises substantial identification and estimation challenges. They concern both upstream and downstream aspects of quasi-experimental design. Upstream issues relate to the conceptual framework required to model the behavioral responses of voluntarily adopting and non-adopting agents, farmers in the present study, and, consequently, to define the estimands of interest. Downstream issues arise once the estimands and the research design have been set, and pertain to the identification strategies and estimation methods required to recover causal effects from observational data.

The present paper focuses on three dimensions of sustainability associated with organic farming adoption. Environmental sustainability is captured by the use of potentially harmful chemical inputs. Private economic sustainability refers to the ability of farms to maintain or improve their net income. A third dimension concerns the capacity of the agricultural sector to maintain or expand overall production or supply levels. The central research question is how these three dimensions jointly combine and which specific configurations of sustainability organic farming is able to deliver. More specifically, the objective is to assess whether organic farming adoption can be characterized as a win-win-win strategy: improving environmental performance, enhancing the economic performance of individual farms, while not jeopardizing aggregate agricultural production.

The main contribution of this paper therefore lies in proposing and implementing an empirical strategy grounded in a fully quasi-experimental logic to assess the multidimensional sustainability of organic farming under heterogeneous responses. To this end, we develop a theoretical framework of farmers' behavior that links organic adoption decisions to the simultaneous emergence of these multiple sustainability outcomes. Despite a common underlying behavioral logic, there is no a priori reason to expect uniform effects across farms. Sustainability outcomes are likely to depend on farm-specific characteristics, implying substantial heterogeneity across observable internal and external features.

The key methodological challenge therefore lies in identifying and estimating heterogeneous treatment effects both at the group level (Group Average Treatment Effects, GATEs), and at the individual farm level (Individualized Average Treatment Effects, IATEs). Empirically, we address this challenge by combining Causal Random Forests and Least Absolute Shrinkage and Selection Operator (LASSO)-based estimation methods. These tools are used sequentially to estimate the GATEs and IATEs in a flexible and largely nonparametric manner, allowing for rich interactions across the covariate space.

The empirical analysis relies on a balanced panel of the Italian Farm Accountancy Data Network (FADN) farms observed over the period 2008–2022. Italian agriculture is characterized by pronounced heterogeneity in agro-climatic conditions, farm structures, and production specializations (Coderoni and Esposti, 2018; Baldoni et al., 2021; Coderoni et al., 2026), which translates into substantial variation in both the propensity to adopt organic practices and behavioral responses to policy incentives. Moreover, the detailed nature of the Italian FADN survey allows for a rich and high-dimensional set of covariates. To approximate an experimental setting, the analysis is restricted to always-treated (organic) and always-non-treated (conventional) farms. Time averages of the outcome variables are used to mitigate short-term volatility, while initial values of covariates are exploited to reduce the risk of endogeneity.

The remainder of the paper is organized as follows. Section 2 provides a structured review of the literature about the sustainability performance of organic farming. Section 3 presents the analytical framework linking farmers' decision-making to heterogeneous sustainability outcomes. Section 4 describes the dataset and the quasi-experimental research design, with particular attention to the definition of treatment and outcome variables. Section 5 details the estimation strategy based on the combined use of Machine Learning methods (Causal Random Forests and LASSO). Section 6 illustrates and discusses the estimation results. Section 7 presents a set of validation exercises while Section 8 concludes and outlines directions for future research.

2. Literature review: the multidimensional impact of organic farming

Sustainability is commonly conceptualized as a multidimensional construct encompassing economic viability, environmental integrity, and social equity. This view is often represented by the so-called sustainability triangle, which emphasizes the need to jointly balance these three interdependent dimensions (OECD, 2008). In the context of agriculture, such a framework reflects the sector's simultaneous role in generating economic value, exerting environmental pressures, as well as benefits, and sustaining rural communities and social outcomes (OECD, 2019). Accordingly, assessments of agricultural sustainability typically rely on multiple indicators capturing distinct, and potentially conflicting, responses across these dimensions, rather than on a single performance metric.

This general notion of sustainability, however, must be operationalized whenever a concrete and specific process is under scrutiny. In the context of the debate on organic farming and its value, two aspects have been most widely recognized and empirically investigated (Coderoni et al., 2024, 2026). The first concerns the environmental dimension, particularly the benefits associated with reducing the use of potentially harmful chemical inputs. The second relates more directly to the farmer's perspective, namely the potential contribution of organic farming to the maintenance, stabilization, or even improvement of farm profitability.

A third dimension also deserves attention, namely the role of agriculture in ensuring food security, defined as the capacity to provide sufficient and affordable food for the population (World Bank, 2025).[1] The need to reconcile this objective with improved environmental performance of farming has generated an extensive debate under the concept of so-called *sustainable* (or *smart*) *intensification* (Tilman et al., 2011). In the context of organic farming, this dimension relates to the ability to maintain, or potentially increase, overall food supply while preserving its affordability. However, the existing literature has generally failed to provide an integrated assessment of organic farming across all three dimensions of sustainability, typically focusing on one dimension, or at most two. As a consequence, the question of whether, and under what conditions, organic farming can deliver a "win–win–win" outcome across environmental, economic, and food security objectives remains largely unresolved.

A large scientific literature investigates the agronomic and environmental impacts of organic farming. Recent surveys reviewing the available evidence broadly agree on its positive contribution (Akhuli, 2025; Sanders et al., 2025). In fact, these studies also emphasize that such effects are neither generalizable to all environmental aspects nor unequivocal across different farming contexts (Coderoni et al., 2024). Besides the possibility of heterogeneous and sometimes ambiguous outcomes, however, *environmental sustainability* is commonly regarded as the first and most widely acknowledged "win" associated with organic farming.

By contrast, substantial doubts have been raised as to whether the adoption of organic farming generates an economic advantage for farmers, namely, the second "win" of the sustainability triangle, *economic sustainability*. This issue is closely related to the question of whether organic farming entails a net social cost in terms of public financial transfers required to sustain adoption (Coderoni

[1] It is also worth noting that a fourth dimension of sustainability, namely resilience, has recently gained prominence in the literature on agricultural systems (Akıf and Büyüksaatçı-Kiriş, 2026). This dimension relates to the capacity of farms and agri-food systems to respond and adapt to external shocks (energy price fluctuations, trade disruptions, water scarcity, geopolitical crises). This aspect is not explicitly considered in the present study for two main reasons. First, elements of resilience are, at least partially, reflected in the three dimensions already analyzed. Second, resilience is difficult to capture through a single, well-defined outcome variable, as it inherently involves dynamic and system-level responses. Nevertheless, incorporating this additional dimension represents a promising avenue for future research.

et al., 2026; Esposti, 2026). The literature addressing these aspects remains relatively limited and fragmented, and the available evidence is largely inconclusive (Baldoni and Esposti, 2025; Coderoni et al., 2024, 2026). Nonetheless, existing studies converge on one robust evidence: the economic viability of organic farming critically depends on public policy support, although the magnitude of this dependence varies substantially across farm types and regional contexts. In some settings, policy payments more than offset potential income losses associated with conversion, whereas in others support levels appear insufficient to generate effective adoption incentives (Esposti, 2024).

Recent contributions reinforce this conclusion. Kremmydas et al. (2024), for instance, assess the economic implications of increasing the EU share of organic utilized agricultural area to 25% and conclude that achieving this target would require substantial additional policy support. Earlier and influential evidence is provided by Zander et al. (2008), who combine FADN data with a survey of 550 organic farms across Western and Eastern Europe, documenting a strong reliance on organic payments, ranging from 4–6% of gross output in Western Europe to 4–19% in Eastern Europe, and reaching up to 75% of farm profits in some new Member States. Offermann et al. (2009) further show that when family farm income is used as the benchmark, dependence on organic payments becomes even more pronounced, particularly in countries such as Denmark and the UK, where income would deteriorate sharply in the absence of these transfers.

Overall, empirical research jointly assessing farm-level economic performance and the environmental outcomes of organic farming remains scarce. As a result, a genuine win–win outcome cannot be considered a general result, although it appears to hold in specific farming systems and contextual settings (Coderoni et al., 2024, 2026; Brutti and Freo, 2026). The most controversial aspect, however, and the one supported by the weakest empirical evidence, concerns the third "win" of the sustainability triangle: whether, beyond policy support, widespread adoption of organic farming is compatible with maintaining, let alone increasing, aggregate agricultural supply, and thus with ensuring food security. In this perspective, the issue is not the environmental or economic performance at the farm level, but the capacity of the agricultural sector as a whole to sustain aggregate output over time (*production-level sustainability*).

Several studies conclude that, in many circumstances, the positive environmental contribution of organic farming comes at the cost of lower overall agricultural production. If generalized, this trade-off would imply either a compensatory expansion of conventional agriculture or a contraction of aggregate supply, with potential upward pressure on food prices. The global meta-analysis by Smith et al. (2019), for instance, shows that organic systems tend to perform better on several environmental indicators, such as soil carbon and biodiversity, whereas conventional systems exhibit superior economic performance through higher and more stable yields.

The main difficulty in providing a conclusive answer about the three dimensions of organic farming sustainability lies in the methodological challenges it entails. Existing evidence is fragmented and unsystematic, ranging from anecdotal results to meta-analyses, largely because empirical assessments of sustainability outcomes are themselves heterogeneous, being based on different farm samples, observation periods, and definitions of outcome variables capturing sustainability performance.

Any rigorous ex-post evaluation of the impacts of organic farming adoption should be conducted at the farm level, both to avoid aggregation bias and to properly account for heterogeneity in responses. Moreover, such an assessment requires a causal framework grounded in counterfactual reasoning, in which organic farming adoption is treated explicitly as an intervention whose effects on environmental, economic, and production outcomes must be identified and estimated as treatment effects. Since field-level randomized experiments are generally infeasible, ex-post evaluation of organic farming policies must necessarily rely on appropriate quasi-experimental designs (Möhring et al., 2024).

Only a limited number of studies apply such a quasi-experimental logic to the issue considered here. Mennig and Sauer (2022), for example, analyze the effects of a policy encouraging conversion to organic farming in the German federal state of Bavaria using a Difference-in-Differences (DiD) approach, but focus exclusively on economic adoption incentives rather than on subsequent environmental outcomes. A similar limitation characterizes Esposti (2024), who applies a Causal Machine Learning (CML) framework to Italian farms over the 2008–2018 period to capture heterogeneous economic responses to policy support, without explicitly estimating environmental impacts as causal effects.

More recently, Biagini et al. (2026) attempt to bridge economic and environmental perspectives by exploiting a large panel of Italian FADN crop farms from 2015 to 2022 within a dynamic DiD framework. Their results indicate that conversion to organic farming entails significant trade-offs in resource use, particularly higher energy and water intensity per unit of output, and that environmental benefits are neither automatic nor uniform.

Relying on EU-wide FADN data over the 2014–2021 period and adopting a combined Propensity Score Matching – Difference-in-Differences (PSM-DiD) approach, a recent study by Brutti and Freo (2026) adopts and finds that organic farming adoption leads to a significant reduction in the use of potentially harmful inputs. At the same time, economic performance, measured as value added per unit of labor, does not decline on average and may even improve, although effects vary substantially across farm types. However, their analysis does not explicitly address treatment effect heterogeneity and its associated methodological and policy implications.

Although these contributions are closely related in spirit and objectives to the present study, none explicitly and jointly considers the three pillars of organic farming sustainability as simultaneous treatment outcomes while allowing for heterogeneity in treatment effects. To fill these gaps, this study proposes an ex-post causal analysis that simultaneously evaluates the environmental, economic, and production-level dimensions of organic farming sustainability. This requires a quasi-experimental design grounded in farmers' decision-making, from which multiple outcome responses naturally emerge, while accounting for heterogeneous effects across a high-dimensional set of farm characteristics.

Accordingly, the proposed research design exploits longitudinal data on the same group of organic and conventional farms over a sufficiently long time horizon to capture the full transition to organic practices. This allows identification of both average and heterogeneous treatment effects for all three outcome dimensions. Estimation is carried out using a combination of flexible parametric methods, based on LASSO regularization, and non-parametric Causal Machine Learning techniques, enabling the estimation of group-specific and individualized treatment effects within a unified empirical framework (Chernozhukov et al., 2018).

3. The conceptual framework

Consider a sample of N production units (farms). For any i-th farm ($i = 1, \dots, N$), an aggregated multi-input multi-output production technology can be represented by the feasible production set $F_i \subset \mathbb{R}^M$.[2] F_i is considered farm-specific and accounts for all sources of heterogeneity across units depending on external and internal factors that we generally indicate with the $(Q \times 1)$ vector, $\mathbf{X}_i$. Farmers' production choices are expressed by a $(M \times 1)$ vector of netputs $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{iM})'$, where $\mathbf{Y}_i \in F_i$. Netput vector $\mathbf{Y}_i$ can be partitioned as $\mathbf{Y}_i = (\mathbf{Y}_{iO}, \mathbf{Y}_{iI})$, where $\mathbf{Y}_{iO} \geq 0$ indicates the $(P \times 1)$ vector of outputs, while $\mathbf{Y}_{iI} \leq 0$ represents the $(R \times 1)$ vector of production inputs. In turn, inputs

[2] The notation presented in this section can be generalized to a panel data framework, where the N units are observed over Z time periods. In such a case, all model variables would be time-variant (e.g., $F_{it}, T_{it}, \mathbf{Y}_{it}, \mathbf{X}_{it}$). However, this generalization is not required here, as the empirical strategy adopted relies on cross-sectional samples (see Section 4).

can be partitioned into a group, thus a $(S \times 1)$ vector $\mathbf{Y}_{iS}$, that does not bring about environmental implications (e.g., labour) and a group with a negative environmental impact, here generically designated as *environmentally intensive* or *harmful inputs* (see Section 4), thus a $(U \times 1)$ vector $\mathbf{Y}_{iU}$. The netput set partitioning thus becomes $\mathbf{Y}_i = \left(\mathbf{Y}_{iO}, \mathbf{Y}_{iS,}\mathbf{Y}_{iU}\right)$ with *R*=*S*+*U* and *M*=*P*+*S*+*U*.

Each farm can voluntarily participate in a policy (a treatment), T_i. When a farm participates ($T_i = 1$), this choice is expected to induce specific production choices, $\mathbf{Y}_i$, given F_i thus $\mathbf{X}_i$. Therefore, it is possible to express these production choices as a function of both the treatment and the aforementioned farm characteristics, $\mathbf{Y}_i^T = (\mathbf{Y}_{iO}^T, \mathbf{Y}_{iS}^T, \mathbf{Y}_{iU}^T) = f^T(\mathbf{X}_i)$, where $f^T(\cdot)$ is a vector-valued function that depends on the treatment status T.

Since adoption is voluntary and following standard adoption theory (Coderoni et al., 2026), the *i*-th farm opts into the treatment ($T_i = 1$) only if its expected utility from adoption exceeds its expected utility from non-adoption. In this respect, the variable of interest to the farmers is their profit (or net farm income; see Section 4), π_i^T, which can be itself expressed by the function $\pi_i^T = g(\mathbf{Y}_i) = g[f^T(\mathbf{X}_i)]$, where $g(\cdot)$ is a single-valued function.[3] π_i^T expresses the farm-level economic dimension of organic farming adoption. It is important to clarify that π_i^T here refers to farm profits including the policy monetary support S_i (see Section 4.1) since this is the value that eventually affects the farmers' adoption decision. However, from the societal perspective the economic performance, and sustainability, of organic farming should be assessed in terms of $(\Delta\pi_i - S_i)$.[4]

The central implication of this conceptual framework is that the multidimensional impacts of organic farming adoption originate from a unified decision-making process. This allows the same analytical structure to be extended to the other dimensions considered in the analysis, ensuring internal consistency across outcome definitions. As anticipated, a second dimension concerns the environmental implications expressed by vector $\mathbf{Y}_{iU}^T$ than can be collapsed into a scalar by simply aggregating all these unsustainable inputs in terms of value using the respective prices, that is, converting $\mathbf{Y}_{iU}^T$ into the respective expenditure $\mathbf{Y}_{iU}^T \rightarrow_i P_i^T = h_P[f^T(\mathbf{X}_i)]$ where $h_P(\cdot)$ is a single-valued function. P_i^T expresses the farm-level environmental dimension of organic farming adoption. The third dimension concerns the overall supply level expressed by vector $\mathbf{Y}_{iO}^T$ than can itself be collapsed into a scalar by aggregating all outputs in terms of value using the respective prices, that is, converting $\mathbf{Y}_{iO}^T$ into the respective revenue, or Gross Production Value (see Section 4): $\mathbf{Y}_{iO}^T \rightarrow GPV_i = h_{GPV}[f^T(\mathbf{X}_i)]$ where $h_{GPV}(\cdot)$ is a single-valued function.

Since function $f^T(\cdot)$ univocally maps the treatment status into the set of production choices conditional on farm-specific characteristics, $\mathbf{Y}_i^T = f^T(\mathbf{X}_i)$, it is natural to derive a TE interpretation of organic farming adoption. Following the conventional Potential Outcome (PO) logic (Rubin, 1974; Imbens and Rubin, 2015), the TE can be identified as $\Delta\mathbf{Y}_i = (\mathbf{Y}_i^{T=1}|\mathbf{X}_i) - \left(\mathbf{Y}_i^{T=0}|\mathbf{X}_i\right) = f^{T=1}(\mathbf{X}_i) - f^{T=0}(\boldsymbol{X}_i)$, where $\mathbf{Y}_i^{T=1}$and $\mathbf{Y}_i^{T=0}$ indicate the production choices the i-th farm would make with and without the treatment (organic farming), respectively, and $\Delta\mathbf{Y}_i$ is the consequent $(M \times 1)$ vector of changes in outputs and inputs following adoption. Therefore, $\Delta\mathbf{Y}_i$ already contains the three dimensions of sustainability. Consequently, the TE on economic sustainability is $\Delta\pi_i = (\pi_i^{T=1}|\mathbf{X}_i) - \left(\pi_i^{T=0}|\mathbf{X}_i\right) = g[f^{T=1}(\mathbf{X}_i)] - g[f^{T=0}(\mathbf{X}_i)]$; the TE on environmental sustainability is $\Delta\mathbf{Y}_{iU} \rightarrow$

[3] Standard causal inference frameworks are formulated in terms of scalar potential outcomes, which implies that multi-dimensional outcome vectors are typically handled by projecting them into scalar components and estimating treatment effects separately for each dimension (Imbens and Rubin, 2015).

[4] This distinction highlights that the present study does not aim to evaluate the effectiveness of policy instruments supporting organic adoption, as in previous contributions (Baldoni and Esposti, 2023; Esposti, 2024), but rather to identify the causal impact of organic farming adoption *per se*, whether induced by subsidies or not, on farm performance. It is worth stressing that not all Italian organic farms receive the correspondent support. In the FADN dataset, both the dummy 'organic' and the extent of policy support are provided. Here, we consider as the treatment variable only the former while we exclude the latter from the calculation of π_i.

$\Delta P_i = (P_i^{T=1}|\mathbf{X}_i) - (P_i^{T=0}|\mathbf{X}_i) = h_P[f^{T=1}(\mathbf{X}_i)] - h_P[f^{T=0}(\mathbf{X}_i)]$; the TE on production-level sustainability is $\Delta\mathbf{Y}_{iO} \rightarrow \Delta GPV_i = (GPV_i^{T=1}|\mathbf{X}_i) - (GPV_i^{T=0}|\mathbf{X}_i) = h_{GPV}[f^{T=1}(\mathbf{X}_i)] - h_{GPV}[f^{T=0}(\mathbf{X}_i)]$.

The empirical identification of these treatment effects is inherently challenging. In observational data, each farmer is observed under either $T_i = 1$ or $T_i = 0$, but never under both. As a consequence, only one of the two potential outcomes is observed, while the other, by construction, remains unobserved and is referred to as the counterfactual outcome. This fundamental problem of causal inference implies that a proper empirical strategy is required to recover causal effects. Such a strategy typically combines a quasi-experimental design with appropriate estimation methods. The specific approach adopted in this study is described in Sections 4 and 5.

Once identified, the combination of the three treatment effects provides a coherent assessment of the organic farming sustainability triangle. In particular, $\Delta\pi > 0$ reflects private economic sustainability; $\Delta P < 0$ expresses environmental sustainability; $\Delta GPV \geq 0$ indicates production-level sustainability. The combination of these three treatment effects allows us to define eight possible outcomes:

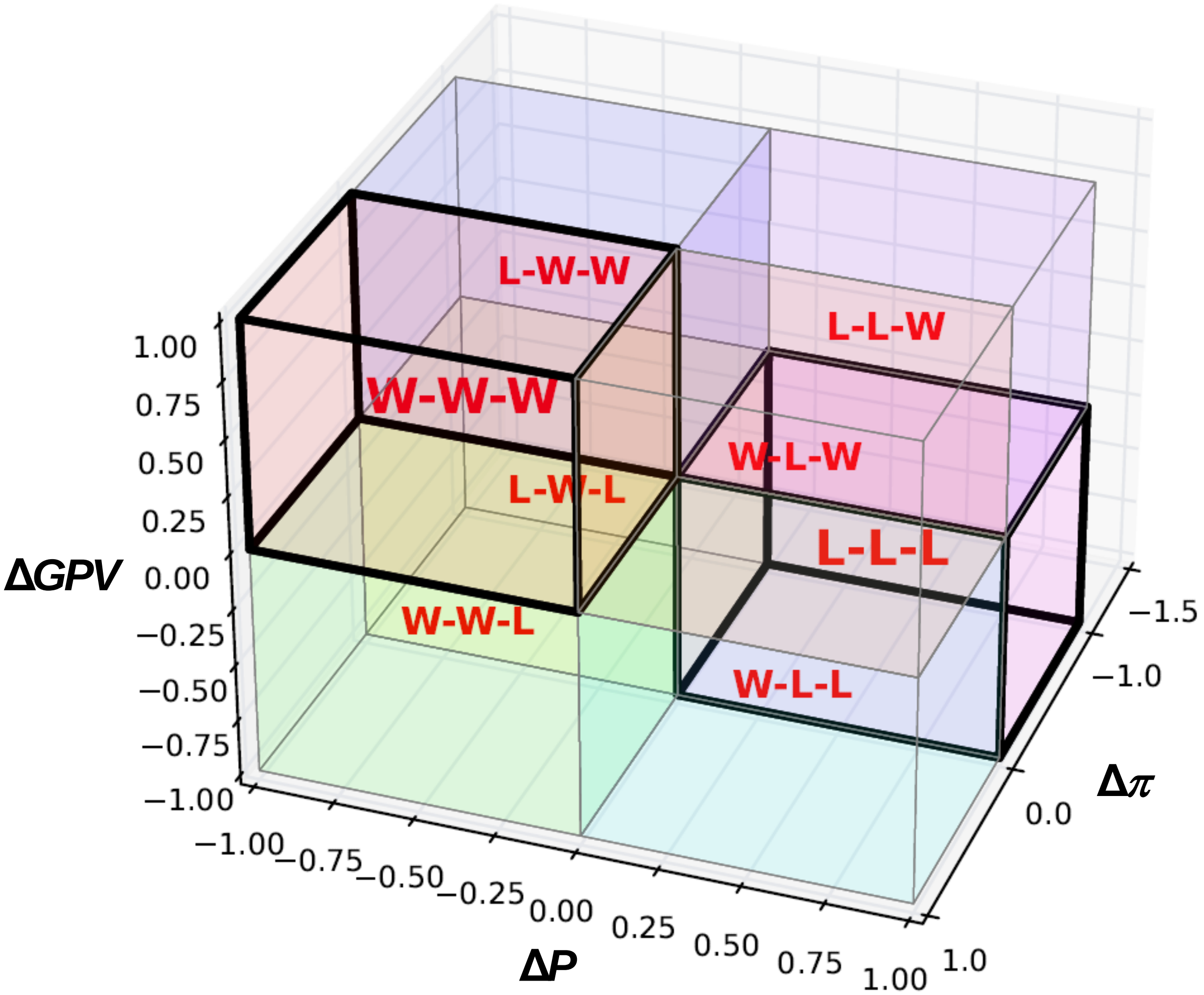
- $\Delta\pi>0$, $\Delta P<0$, $\Delta GPV>0$ (WIN-WIN-WIN): fully sustainable organic farming.
- $\Delta\pi>0$, $\Delta P>0$, $\Delta GPV>0$ (WIN-LOSS-WIN): environmentally unsustainable organic farming.
- $\Delta\pi>0$, $\Delta P<0$, $\Delta GPV<0$ (WIN-WIN-LOSS): production-level unsustainable organic farming.
- $\Delta\pi>0$, $\Delta P>0$, $\Delta GPV<0$ (WIN-LOSS-LOSS): environmentally and production-level unsustainable organic farming.
- $\Delta\pi<0$, $\Delta P<0$, $\Delta GPV>0$ (LOSS-WIN-WIN): economically unsustainable organic farming.
- $\Delta\pi<0$, $\Delta P>0$, $\Delta GPV>0$ (LOSS-LOSS-WIN): economically and environmentally unsustainable organic farming.
- $\Delta\pi<0$, $\Delta P<0$, $\Delta GPV<0$ (LOSS-WIN-LOSS): economically and production-level unsustainable organic farming.
- $\Delta\pi<0$, $\Delta P>0$, $\Delta GPV<0$ (LOSS-LOSS-LOSS): fully unsustainable organic farming.

With $\Delta\pi$, ΔP, ΔGPV as continuous variables, the conventional sustainability triangle associated with organic farming can be represented as a three-dimensional continuous space which, following the taxonomy above, defines a "sustainability octant" framework, as illustrated in Figure 1. Accordingly, each observation (farm) can be assigned to one of the eight subspaces. Apart from the two extreme cases (win–win–win and loss–loss–loss), all other combinations reflect mixed sustainability outcomes, providing no conclusive answer to whether organic farming is more sustainable than conventional farming.

In fact, positioning within the octant framework is more informative than the classification itself, as it does not capture the magnitude of the effects. While the framework provides a useful heuristic for interpreting multidimensional causal effects, it can be further refined by incorporating information on the relative magnitude of the three components. In particular, even when the three dimensions are not aligned and yield mixed outcomes, one effect may dominate the others, thereby shaping the overall sustainability assessment.

This relative dominance is reflected in the position within a given octant, but it can also be captured by considering ratios between the three indicators. Three ratio-based indicators are of particular interest. First, $\Delta(P/\pi)$: even when $\Delta\pi < 0$ or $\Delta P > 0$, organic farming may improve the joint economic–environmental balance if the ratio of environmentally harmful inputs to net farm income declines, i.e. $\Delta(P/\pi) < 0$. Second, $\Delta(P/GPV)$: a negative value indicates that, even when $\Delta GPV < 0$ or $\Delta P > 0$, organic farming reduces environmental impact per unit of agricultural output. Third, $\Delta(\pi/GPV)$: a positive value suggests that, even when $\Delta\pi < 0$ or $\Delta GPV < 0$, organic farming improves profitability per unit of production.

Figure 1 – The sustainability space and its octants (axis values are purely illustrative).



4. **The research design and the dataset**

4.1. Sample construction

As anticipated, the fundamental challenge in identifying and estimating the treatment effects of interest is that, for any given farm, the counterfactual outcome is unobserved. Each unit is observed under either $T_i = 1$ or $T_i = 0$. In an ideal experimental setting, this problem is addressed through random assignment. A sufficiently large sample of units is drawn from the target population, and the treatment is randomly allocated to a subset of units, leaving the remaining ones as controls. Randomization ensures that treated and control units are statistically comparable, so that the observed outcomes of the control group provide a valid proxy for the counterfactual outcomes of treated units.

After a sufficient period for treatment effects to materialize, outcomes are observed for all units. In contexts where outcomes are inherently volatile, such as farm-level economic or environmental indicators, it is often appropriate to compute time-averaged measures over a sufficiently long window in order to mitigate short-term fluctuations and obtain more stable estimates of performance. The comparison between the average outcomes of treated and control groups yields an estimate of the Average Treatment Effect (ATE). At the same time, individual-level outcomes allow for the analysis of treatment effect heterogeneity.

When observational rather than experimental data are used, this ideal empirical strategy is no longer feasible. Consequently, conditional on the available information, a suitable quasi-experimental design must be implemented in order to approximate as closely as possible the identifying conditions of randomized experiments.

The present empirical analysis is based on the Italian FADN (RICA) dataset over the 2008–2022 period. This dataset provides a representative sample of commercial farms at the national level, observed as an unbalanced panel. The number of observations ranges from 11,391 in 2008 to 11,084 in 2022, with a maximum of 11,398 in 2013 and a minimum of 9,580 in 2015. It contains detailed information on farms' structural, economic, and geographical characteristics, and ensures representativeness across multiple dimensions, including the NUTS 2 regional level.

The 2008–2022 time span represents the longest period currently available that allows the extraction of sufficiently large and consistent farm samples.[5] This fifteen-year horizon satisfies two key requirements for approximating an experimental setting. First, it allows for the completion of the conversion process from conventional to organic farming. According to EU policy rules, this transition typically involves a two- to three-year conversion phase followed by a five-year maintenance period, after which support is phased out.[6] A meaningful comparison between organic and conventional farms should therefore be conducted only after this adjustment period has been completed.[7] Second, a sufficiently long observation window makes it possible to compute time-averaged measures of farm performance, thereby mitigating the substantial year-to-year volatility that characterizes economic and environmental outcomes in agriculture. In observational settings, this volatility is particularly problematic because it varies substantially across units (e.g. across farm specializations) and may be correlated with both observed and unobserved characteristics. As a result, it can induce significant heteroskedasticity in the estimation stage, potentially affecting the precision and reliability of the estimated treatment effects.

In principle, there are well-known advantages to working with panel data in observational settings. The time dimension allows controlling for time-invariant unobserved heterogeneity across farms and enables identification strategies based on within-unit variation, such as difference-in-differences (DiD) approaches (Esposti, 2026). However, in the present context, these advantages are limited and are more than offset by the corresponding disadvantages associated with the time dimension, which hinder the attempt to approximate an experimental (i.e. controlled) setting. Two main issues emerge in this regard.

[5] Years prior to 2008 are, in principle, available in the FADN dataset but the change in the data collection protocols made serval information not comparable with the following years and, in any, would imply a substantial reduction of the balanced sample size and, in particular, of treated units. At the moment this paper is written the period of coverage could be extended also to 2023 and 2024. Also in this case, however, the unbalanced panel consists of 10988 and 11006 units, respectively, but the balanced sample would reduce to 1112 and 1028 units (961 and 891 units if staggered entrants are excluded) and the number of always-treated farms falls to 33 and 29, respectively. Moreover, years 2023 and 2024 are excluded to preserve consistency with a stable institutional environment. In particular, 2023 marks the beginning of the new Common Agricultural Policy (CAP) programming period, which introduces substantial changes in policy architecture, including the shift towards National Strategic Plans, the enhanced role of eco-schemes, and a stronger integration of environmental and climate objectives into farm-level decision-making. Although the present study does not explicitly model CAP support for organic farming, such changes are likely to affect both the adoption decision and the distribution of outcomes through multiple channels, including altered incentive structures, revised conditionality requirements, and increased heterogeneity in national implementation. These factors may induce a structural break in the underlying data-generating process. Restricting the analysis to the pre-2023 period therefore allows us to work within a more homogeneous institutional framework, avoiding potential confounding effects associated with regime changes and strengthening the internal validity of the quasi-experimental identification strategy.

[6] In the 2015-2022 EU Rural Development Policy (RDP), in particular, there is a clear distinction between support for the adoption of organic farming (Measure 11.1) and support for the maintenance of organic farming (Measure 11.2) (Stolze et al., 2016). The former support can be delivered only for the aforementioned period of conversion. The latter support can last up to five years after the end of the conversion period, though this duration can be extended by EU member states in their Rural Development Policy programming.

[7] A similar discussion on the appropriate length of the analysis window, accounting for the conversion period in organic farming adoption, albeit with different empirical strategies, can be found in Brutti and Freao (2026, p. 10).

First, the FADN sample is characterized by significant turnover (attrition), implying that some farms are observed only for limited time spans. To ensure consistency with an experimental logic requiring stable units, we restrict the analysis to a balanced panel covering the entire 2008–2022 period, yielding 1,219 farms. Second, within this balanced panel, several farms switch into organic production during the observation period. The heterogeneous treatment timing implied by such staggered and progressive entry might be informative for investigating the history dependence of the TE (Sun and Abraham, 2021; Baldoni and Esposti, 2025). In fact, it complicates the experimental logic adopted here, which requires a clear distinction between controls and treated units, namely farms that have already gone through the whole conversion phase and are to all intents and purposes organic. We therefore exclude farms that change treatment status over time. This procedure results in a balanced panel of 1,055 farms that are either always treated (36 farms corresponding to 3.4% of the sample) or never treated (1,019 farms). These latter provide the control group used to construct counterfactual outcomes.

Finally, to further align the empirical design with an experimental framework, this balanced panel is collapsed into a cross-sectional sample of 1,055 units. For each farm, outcome variables are defined as multi-year averages computed over the last five years of observation (2018–2022). As discussed below, the earlier period (2008-2010) is then used to construct the set of exogenous covariates. This choice reduces the impact of short-term volatility in farm performance and abstracts from the short-term adjustment dynamics associated with the conversion phase. Thus it is intended to approximate long-term or persistent production conditions and multidimensional performance.

4.2. Outcome and treatment variables

Another key difference between experimental and observational settings is that, in controlled experiments, researchers can pre-define the outcome of interest, the treatment assignment, and the set of confounding variables to be measured and controlled for. In observational data, by contrast, the available information cannot be designed *ex ante* but must be exploited *ex post*. In this context, the richness of the FADN dataset provides an opportunity to leverage a large set of observed characteristics in order to mitigate the lack of experimental control. In other words, the availability of detailed information may partially compensate for the absence of a controlled environment. However, this strategy raises the issue of high dimensionality, as the inclusion of a large number of covariates may lead to overfitting and complicate the identification of causal effects (see Section 5).

The treatment variable is here defined as the adoption of organic farming. Farms are classified as treated ($T_i = 1$) if they are reported as organic in the FADN (RICA) dataset. This status relies on the official certification released upon complying with the European Union's regulatory framework for organic production.[8] Accordingly, treated units can be interpreted as farms that have formally adopted organic practices in compliance with EU regulatory requirements, regardless of whether they receive the respective CAP monetary support.

Following Section 3, three level-based outcome variables are considered, each corresponding to one dimension of sustainability. Economic sustainability at the farm level ($\Delta\pi$) is measured as the variation in net farm income reported in the FADN dataset. As anticipated, this income measure includes the monetary CAP support for organic farming. The last column of Table 1 shows that this support, on average, amounts to about €8,000 per supported organic farm, corresponding to about 15% of the respective net farm income, but it is also highly skewed (the median is about 40% lower than the mean) and exhibits substantial variability (the coefficient of variation exceeds 1.3).

[8] The current framework is defined by Regulation (EU) 2018/848 on organic production and labelling, which replaced earlier legislation and builds on a regulatory tradition originally established with Regulation (EEC) No 2092/91. These regulations specify production standards, input restrictions, and certification procedures required for organic status.

Environmental sustainability (ΔP) is proxied by total expenditure on potentially harmful inputs, as reported in the FADN data. This aggregate includes expenditures on fertilizers, pesticides and herbicides, veterinary products, externally sourced animal feed, water, energy and fuels, and is intended to capture input-based pressure on natural resources. Finally, production-level sustainability (ΔGPV) is measured by gross production value, as reported in the FADN dataset. From these level-based outcome variables, three relative (or ratio-based) outcome variables can then be computed: $\Delta(P/\pi)$, $\Delta(P/GPV)$ and $\Delta(\pi/GPV)$.

Table 1 reports the sample mean and median values together with the standard deviation over the whole sample and the treatment groups for these six outcome variables. On average, organic farms appear to be less profitable in absolute terms (π) but this gap is largely compensated by the respective average CAP support (S). Moreover, they are more profitable in relative terms (π/GPV). Organic farms also show a lower overall level of production (GPV) but also a lower use of potentially harmful inputs both in level-based (P) and ratio-based terms, (P/π) and (P/GPV). Within the sample, however, the empirical distribution of the three level-based variables is highly asymmetric with the median value always lower than the mean value and with a large variability (standard deviation). These simple descriptive statistics make the fundamental research question of the present study explicit. What holds on average may not hold at the farm level, where heterogeneous patterns across indicators and their combinations may emerge.

Table 1 – Annual average values of the outcome variables (2008–2022): descriptive statistics.

	π (€)	P (€)	GPV (€)	P/π	P/GPV	π/GPV	S
All units (n. 1055):							
Mean	56848	25259	138499	0.378	0.172	0.477	-
Median	30099	9498	66315	0.354	0.154	0.439	-
Standard deviation	108879	55158	322366	3.773	0.124	0.519	-
Control units (n. 1029):							
Mean	56965	25802	139704	0.385	0.174	0.471	0
Median	110169	55973	326441	3.835	0.124	0.518	0
Standard deviation	29726	9594	66603	0.364	0.158	0.436	0
Treated units (n. 36):							
Mean	53534	9875	104396	0.198	0.108	0.657	8075
Median	62966	15545	169571	0.161	0.087	0.518	4885
Standard deviation	33326	4389	59453	0.153	0.092	0.576	10814
Mean difference: control vs. treatment	3430	15927	35308	0.187	0.066	-0.186	-8075

4.3. Confounding variables

In an experimental setting, the set of confounding variables is defined prior to the implementation of the experiment and is explicitly measured as part of its design. This choice is driven by two main principles. First, confounders are selected to control for factors that may be correlated with both treatment assignment and the outcome, thereby potentially biasing the identification of treatment effects (see Section 5). Second, confounders are identified and measured so as to ensure their exogeneity, i.e. that they are not themselves affected by the treatment or by the outcome. Proper experimental design and data collection are therefore intended to minimize both selection bias and endogeneity concerns.

In the present observational setting, such ex-ante design is not feasible. The only viable strategy is to exploit, as effectively as possible, the information available in the FADN dataset. As discussed in Section 3, the set of confounders is intended to capture the main determinants of farm-specific production technologies that shape the response to the treatment. However, the set of observed characteristics $\mathbf{X}_i$ may not be exhaustive. The use of a high-dimensional set of covariates can mitigate the potential bias from selection on unobservables, which is likely to arise when dealing with highly heterogeneous units. The strategy adopted here thus consists of leveraging the richness of the available covariate information and employing estimation methods specifically designed for high-dimensional settings, which allow for the selection of relevant covariates while shrinking the influence of less informative ones. This approach is commonly referred to as regularization in the Machine Learning (ML) literature (see Section 5).

Following Brown et al. (2021) and Stetter et al. (2022), farm technology is characterized using 47 observed covariates, grouped into the following four sets: size and factor endowment; production specialization; socio-demographic characteristics of farmers; environmental and geographical factors. These variables jointly form the $(Q \times 1)$ vector $\mathbf{X}_i$ of Section 3. Annex 1 provides the complete list of covariates included in the analysis. Some of these covariates are time-invariant (e.g. geographical location) and can therefore be reasonably considered exogenous with respect to both treatment assignment and outcome variables. Several other elements of $\mathbf{X}_i$, however, are time-varying (and should therefore be denoted as $\mathbf{X}_{it}$). Even when they reflect structural characteristics of the farm (e.g. physical size), it cannot be excluded that these variables are endogenously influenced over time by treatment status and the associated production choices.

To address these endogeneity concerns, a common empirical strategy, often referred to as *lag identification*, consists in using only pre-treatment values of time-varying covariates. This approach helps prevent unintended causal pathways from treatment to covariates, thereby mitigating reverse causality and post-treatment bias (Stetter et al., 2022; Coderoni et al., 2024). Under the assumption that anticipation effects are negligible, lagged covariates are often considered sufficient to approximate exogeneity. However, as emphasized by Bellemare et al. (2017), this strategy does not necessarily eliminate endogeneity concerns, since pre-treatment variables may still be correlated with unobserved factors, such as spatial or institutional forces, that jointly influence treatment assignment and outcomes.

In the present case, however, identification of lagged effects is not feasible. The analysis is restricted to always-treated and never-treated units over the 2008–2022 period and changes in FADN sampling procedures and variable definitions prevent the consistent use of pre-2008 data across units. Nevertheless, given that outcome variables are defined as time-averages over the last five years of observation, a similar logic can be implemented by constructing time-averaged covariates using only the initial period (2008–2010). This approach allows for a temporal separation between covariates and outcomes, while also reducing the impact of short-term volatility in the covariates themselves. Considering the relatively long period separating these initial covariates from the observed outcomes, this strategy can help approximate the exogeneity of time-varying confounders with respect to both treatment status and outcome variables, and mitigate remaining sources of endogeneity associated with initial years.

4.4. Limitations of the observational design

The quasi-experimental design presented above not only approximates an ideal experimental setting, but also makes explicit the identification challenges that arise from the observational nature of the data. A first fundamental difference between an ideal randomized experiment and the present setting concerns sample size and its allocation. In randomized controlled trials, researchers can perform ex-ante power analyses and design sample sizes and treatment allocation to ensure adequate precision and statistical power. In observational studies, by contrast, both the total sample size and the distribution of treated and control units are not determined by the researcher. Instead, they are inherited from the underlying data collection process. As a result, standard ex-ante power analysis is not directly applicable. In our case, while the overall sample size is relatively large (1,055 observations), the effective sample for identification is driven by the number of treated units, which is small (36 farms) relative to the pool of controls (1019 units). This imbalance limits both statistical precision and the scope of identification.

In this context, a key challenge is the limited overlap between treated and control units. While some farm characteristics display comparable distributions across groups, others are clearly unbalanced. This partial lack of common support becomes particularly critical when moving beyond the estimation of the average treatment effects (ATEs) to conditional and individual treatment effects (CATEs and IATEs). The estimation procedure discussed in Section 5 is designed to address these imbalances using high-dimensional parametric and nonparametric methods. However, identification remains inherently local, confined to the support of treated units, and extrapolation beyond this region is not warranted.

A second key discrepancy with respect to the experimental benchmark concerns the already mentioned endogenous treatment assignment. While this risk is minimized by construction in experimental settings, here it remains somehow implicit since, in any case, the treatment is not randomly assigned but it is always a voluntary choice of the farms and may reflect pre-existing differences between treated and control farms. These differences may both reduce overlap and introduce selection bias, particularly when relevant determinants are unobserved or imperfectly measured. As discussed above, strict lag identification is not fully feasible in the present setting, and therefore residual endogeneity concerns cannot be entirely ruled out.

Taken together, these challenges motivate a twofold empirical strategy: the use of identification and estimation methods capable of mitigating these sources of bias (Section 5), and the implementation of complementary validation exercises to assess the sensitivity and robustness of the results (Section 7).

5. The estimation approach

Given the premises above, we seek an estimation approach able to simultaneously account for potential treatment endogeneity, highly heterogeneous (possibly unit-level) causal effects, and a high-dimensional covariate space capturing observed heterogeneity, with the resulting risk of limited overlap. A doubly robust, machine learning–based estimation framework (Chernozhukov et al., 2018; Athey et al., 2019) provides a suitable approach to address these challenges.

Let y_i denote the observed outcome for unit i, conditional on the binary treatment indicator T_i and the vector of observable characteristics $\mathbf{X}_i$. Following the PO framework discussed in Section 3, the expected value of the difference between the two potential outcomes identifies the TE of interest:

(1) $$\tau(\mathbf{X}_i) = E[y_i^{T=1} - y_i^{T=0}|\mathbf{X}_i]$$

When (1) is extended to all units, assuming they are all random observations of the same underlying data generating process thus neglecting TE heterogeneity, we obtain the ATE: $\tau = E[y_i^{T=1} - y_i^{T=0}]$. The ATE characterizes the whole distribution of treatment effects when the treatment effect is constant across the population. However, when the treatment effect is heterogeneous since units react differently to the same treatment, estimating only the mean of treatment effects may lead to misleading conclusions (Esposti, 2026).

But equation (1) allows for heterogeneous treatment effects, as it defines the average treatment effect conditional on covariates $\mathbf{X}_i$, i.e., the Conditional Average Treatment Effect (CATE). The granularity of the conditional set $\mathbf{X}_i$ thus determines the subpopulation to which the ATE refers. At the finest level, when $\mathbf{X}_i$ exclusively identifies a specific observation, the CATE identifies the Individualized Average Treatment Effect (IATE). Alternatively, the conditional set $\mathbf{X}_i$ may identify a prespecified grouping (for instance, farms classified by physical size), denoted by g_i. In this case, the CATE is also known as Group Average Treatment Effect (GATE): $\tau(g_i) = E[y_i^{t=1} - y_i^{t=0}|g_i]$. The group variable, g_i, has to be specified ex-ante to avoid p-value hacking, as discussed in Head et al. (2015). However, if we do not have an ex-ante grouping, we can elicit the groups in a data-driven way by generating the groups using the quantiles of the IATE estimates.

While (1) provides the fundamental logic behind the identification of the CATE and its variants (ATE, IATE, GATE), it leaves open the issue of estimating $y_i^{T=1}$ and $y_i^{T=0}$ for all units since only one of these two values is actually observed for any given unit. This requires modelling the stochastic process generating the outcome variable conditional on $\mathbf{X}_i$ and given the treatment status. The most general way to express this process, also known as *outcome model*, is the following:

(2) $$y_i = T_i(\mathbf{X}_i) + \varphi(\mathbf{X}_i) + \varepsilon_i$$

where $\tau(\mathbf{X}_i)$ is a function of $\mathbf{X}_i$ that interacts with the binary treatment status t_i while $\varphi(\mathbf{X}_i)$ is a function expressing the dependence of the outcome from the covariates independently from the treatment status. ε_i is a generic additive error term with $E(\varepsilon_i|\mathbf{X}_i) = 0$.

In observational studies the treatment is not randomly assigned and in the present case it depends on the i-th unit voluntary choice. Consequently, T_i can be itself dependent on a set of unit-specific characteristics and modelled as follows (*treatment model*):

(3) $$T_i = \theta(\mathbf{X}_i) + u_i$$

where $\theta(\mathbf{X}_i)$ is a function expressing the dependence of the treatment status from the covariates and u_i is another additive error term $E(u_i|\mathbf{X}_i) = 0$. In fact, as the dependent variable in equation (3) is a binary variable, (3) takes the form of a Propensity Score (PS) equation*:*

(4) $$P(T_i = 1|\mathbf{X}_i) = F[\theta(\mathbf{X}_i)]$$

where $P(\cdot)$ denotes the propensity score, i.e., the probability of receiving the treatment conditional on observed covariates $\mathbf{X}_i$ and $F(\cdot)$ denotes the cumulative distribution function of the error term u_i. In parametric settings, this probability is typically modelled using binary response models such as logit or probit; however, in high-dimensional contexts it can be estimated more flexibly using ML methods.

Upon the estimation of (2) and (4), it becomes possible to compute the two potential outcomes as follows:

(5) $$\begin{aligned} y_i^{T=1} &= \tau(\mathbf{X}_i) + \varphi(\mathbf{X}_i) + \varepsilon_{i1} \\ y_i^{T=0} &= \varphi(\mathbf{X}_i) + \varepsilon_{i0} \end{aligned}$$

with $E(\varepsilon_{i1}|\mathbf{X}_i) = E(\varepsilon_{i0}|\mathbf{X}_i) = 0$. Consequently, the CATE is simply estimated as: $E[y_i^{T=1} - y_i^{T=0}|\mathbf{X}_i] = \tau(\mathbf{X}_i)$.

As detailed below, the estimand of interest, $\tau(\mathbf{X}_i)$, can then be estimated nonparametrically via the Generalized Random Forest proposed in Athey et al. (2019) while the nuisance functions $\varphi(\mathbf{X}_i)$ and $\theta(\mathbf{X}_i)$ can be estimated using the high-dimensional techniques LASSO or Random Forest estimation (Chernozhukov et al., 2018). Therefore, as far as we do not impose any parametric assumption on $\tau(\mathbf{X}_i)$ and on the nuisance functions $\varphi(\mathbf{X}_i)$ and $\theta(\mathbf{X}_i)$, model (5) is sufficiently flexible and general to accommodate potentially complex and unknown forms of dependence of the treatment assignment, the outcome variable, and the treatment effect itself on the covariates $\mathbf{X}_i$.

The outcome model (2) is additive (or partial linear) since it assumes that the observed outcome is generated by the sum of the treatment effect, the function $\varphi(\mathbf{X}_i)$ and an error term. However, model flexibility, can be improved by specifying (2) as a fully interactive model:

(6) $$y_i = (1 - T_i)\mu_0(\mathbf{X}_i) + t_i\mu_1(\mathbf{X}_i) + \varepsilon_i$$

The CATE implied by (6) thus becomes : $E[y_i^{T=1} - y_i^{T=0} | \mathbf{X}_i] = \mu_1(\mathbf{X}_i) - \mu_0(\mathbf{X}_i) = \tau(\mathbf{X}_i)$.[9]

The full interactive model to be estimated thus becomes:

$$(7) \quad \begin{aligned} y_i^{T=1} &= \mu_1(\mathbf{X}_i) + \varepsilon_{i1} \\ y_i^{T=0} &= \mu_0(\mathbf{X}_i) + \varepsilon_{i0} \\ P(T_i = 1 | \mathbf{X}_i) &= F[\theta(\mathbf{X}_i)] \end{aligned}$$

In observational studies, where treatment is not randomly assigned, identifying causal effects as in (7) generally requires additional identifying assumptions (Imbens and Rubin, 2015). The first is the standard unconfoundedness assumption (or Conditional Independence Assumption, CIA), whereby potential outcomes are independent of treatment assignment conditional on observed covariates, $(y_i^{T=0}, y_i^{T=1}) \perp T_i \mid \mathbf{X}_i$. The second is the overlap condition, requiring that the propensity score is bounded away from zero and one, $0 < P(\cdot) < 1$. The third is the Stable Unit Treatment Value Assumption (SUTVA), which ensures that each unit's potential outcomes depend only on its own treatment status and that the treatment is well defined. As discussed in section 4.4., CIA and overlap condition may be critical in the present study. Their validity of these assumptions cannot be easily verified and should be supported by appropriate testing procedures and robustness checks.

Under the validity of these identifying assumptions, the CATE and its variants (IATE, GATE and ATE) are here estimated adopting the fully interactive specification in (7). In practice, functions $\mu_1(\mathbf{X}_i)$, $\mu_0(\mathbf{X}_i)$ and $F[\theta(\mathbf{X}_i)]$ are approximated using flexible learners, such as LASSO or Random Forest, which enable the modeling of complex nonlinear patterns in high-dimensional covariate spaces. Using these methods, recent literature in the field has proposed doubly robust ML–based (DML) estimators of the CATE. In particular, following Knaus (2022) and Kennedy (2023), this approach relies on the estimation of Augmented Inverse Probability Weighting (AIPW) scores, which provide doubly robust pseudo-outcomes for the individual treatment effects (IATEs). Based on these, group average treatment effects (GATEs) and the average treatment effect (ATE) can be readily obtained (see below).

A key property of these scores is Neyman orthogonality, which implies that the resulting estimates are locally insensitive to first-order errors in the estimation of the nuisance functions, including both the outcome and the treatment (propensity score) models in (7) (Chernozhukov et al., 2018; Knaus, 2022). This property is particularly relevant when machine learning (ML) methods are used to estimate these components, as such methods may be subject to regularization bias or estimation errors.

The AIPW pseudo-outcomes are computed as follows:

$$(8) \quad \begin{aligned} y_{i,AIPW}^{T=1} &= \mu_1(\mathbf{X}_i) + \tfrac{T_i}{F[\theta(\mathbf{X}_i)]}[y_i - \mu_1(\mathbf{X}_i)] \\ y_{i,AIPW}^{T=0} &= \mu_0(\mathbf{X}_i) + \tfrac{1-T_i}{1-F[\theta(\mathbf{X}_i)]}[y_i - \mu_0(\mathbf{X}_i)] \end{aligned}$$

Thus, the CATE $\tau(\mathbf{X}_i)$ can also be written as:

$$(9) \quad \tau(\mathbf{X}_i) = E\left[y_{i,AIPW}^{T=1} - y_{i,AIPW}^{T=0}\right] = E\left\{[\mu_1(\mathbf{X}_i) - \mu_0(\mathbf{X}_i)] + \tfrac{T_i}{F[\theta(\mathbf{X}_i)]}[y_i - \mu_1(\mathbf{X}_i)] - \tfrac{1-T_i}{1-F[\theta(\mathbf{X}_i)]}[y_i - \mu_0(\mathbf{X}_i)]\right\}.$$

A consistent estimation of $\tau(\mathbf{X}_i)$ thus requires a consistent estimation of nuisance functions $\mu_0(\mathbf{X}_i)$, $\mu_1(\mathbf{X}_i)$ and the propensity score $F[\theta(\mathbf{X}_i)]$. To this end, we employ ML–based regularization methods, which are particularly well suited for high-dimensional observational data, where nuisance functions must be estimated flexibly to accommodate nonlinearities, complex interactions, and rich covariate structures. We estimate the nuisance functions using LASSO regression. Specifically, Logit LASSO

[9] Compared to (6), (2) thus imposes the following restriction: $\mu_0(\mathbf{X}_i) = \varphi(\mathbf{X}_i)$ and $\mu_1(\mathbf{X}_i) = \varphi(\mathbf{X}_i) + \tau(\mathbf{X}_i)$.

is employed to estimate $F[\theta(\mathbf{X}_i)]$, while square-root LASSO is used for $\mu_0(\mathbf{X}_i)$ and $\mu_1(\mathbf{X}_i)$ ensuring scale invariance and robustness to unknown error variance.[10]

Given the estimates $\hat{\mu}_0(\mathbf{X}_i)$, $\hat{\mu}_1(\mathbf{X}_i)$, and the propensity score $F\left[\hat{\theta}(\mathbf{X}_i)\right]$, the AIPW pseudo-outcomes $\hat{y}_{i,AIPW}^{T=1}$ and $\hat{y}_{i,AIPW}^{T=0}$ can be computed according to (8). These are then used to construct the AIPW scores $\hat{\Gamma}_i^{(-i)} = \hat{y}_{i,AIPW}^{T=1} - \hat{y}_{i,AIPW}^{T=0}$, where $\hat{\Gamma}_i^{(-i)}$ denotes the cross-fitted AIPW score for the i-th observation obtained using nuisance function estimates based on samples that exclude i-th observation. The IATE for each unit, $\hat{\tau}(\mathbf{X}_i)$, is ultimately estimated by solving the following local moment condition:

(10) $\sum_{i=1}^{N} \alpha(\mathbf{X}_i)\left[\hat{\Gamma}_i^{(-i)} - \hat{\tau}(\mathbf{X}_i)\right] = 0$

where $\alpha(\mathbf{X}_i)$ defines local weights that assign greater importance to observations with covariates close to $\mathbf{X}_i$. In the present study, the moment condition is solved by fitting an honest regression forest of the AIPW scores $\hat{\Gamma}_i^{(-i)}$ on $\mathbf{X}_i$, following Wager and Athey (2018) and implemented via the Generalized Random Forest approach of Athey et al. (2019).[11]

For the sake of notation simplicity, we are assuming for the moment that the three functions that identify the CATE according to (5), namely $\tau(\mathbf{X}_i)$, $\varphi(\mathbf{X}_i)$ and $\theta(\mathbf{X}_i)$, and estimation via (10), depend on the same set of covariates $\mathbf{X}_i$. In fact, as will be clarified further in Section 6 (see also Table A1 in Annex 1), in order to limit dimensionality and enhance the interpretability of heterogeneity patterns, the set entering the last estimation stage expressed by (10) can be just a more parsimonious subset of the whole high-dimensional set entering $\mu_1(\mathbf{X}_i)$, $\mu_0(\mathbf{X}_i)$ and $F[\theta(\mathbf{X}_i)]$.

Based on these estimated IATEs, GATE and ATE estimates can be readily obtained. The ATE is computed as the average of the estimated AIPW scores over the full sample. For pre-determined groups, following Semenova and Chernozhukov (2021), the GATE is obtained as the OLS regression of the AIPW scores on group indicator variables, g_i. In the case of data-driven groups, the estimated AIPW scores are first ranked and partitioned into quantiles. The GATE is then obtained by regressing the AIPW scores on the corresponding quantile indicator dummies.

Under the aforementioned identification assumptions (CIA, overlap condition and SUTVA), this estimation procedure yields consistent and asymptotically normal causal effect estimates with valid standard errors, allowing for interpretable confidence intervals.[12] This allows us to handle the two major limitations brought about by this kind of datasets discussed in Section 4.4.

In this respect, the proposed approach is particularly well suited to our empirical setting, which relies on the observational, unbalanced, and high-dimensional structure of the Italian FADN dataset. One key advantage of this approach is that it helps address limitations arising from poor overlap, as

[10] LASSO can be viewed as a linear approximation method within the broader class of machine learning estimators, combining variable selection and regularization in high-dimensional settings. In the ML terminology, “regularization” refers to the introduction of a penalty term in the estimation problem to control model complexity and prevent overfitting in high-dimensional settings. Relative to more flexible nonlinear learners, such as Random Forest, LASSO may provide more stable estimates in settings characterized by high dimensionality and a limited number of treated units, as in the present application.

[11] To perform this estimation procedure, sample-splitting and cross-fitting is used to obtain out-of-sample predictions of the nuisance functions, which are then employed in the construction of the AIPW pseudo-outcomes. This ensures that, for each observation, the estimated nuisance functions are obtained from models trained on independent data, thereby preventing overfitting and preserving the orthogonality properties of the estimator. See Chernozhukov et al. (2018) for details.

[12] It is worth stressing that Athey et al. (2019) highlight additional regularity conditions required to ensure these desirable properties. In particular, while the double machine learning framework accommodates high-dimensional covariates in the estimation of the nuisance functions, the nonparametric estimation of the IATE requires that the dimension of the conditioning variables remains relatively low in order to avoid the curse of dimensionality. For this reason, as anticipated, the set of covariates used in the final stage of the estimation procedure as expressed by (10) is restricted to a subset of those employed in the first-stage estimation of the PS and nuisance functions.

imbalances between treated and control units may lead to instability in classical parametric models. Rather than imposing common support globally, the estimation procedure adapts to the empirical distribution of the data, placing less weight on observations with extreme propensity scores and focusing on regions where overlap is stronger. As a result, the proposed approach mitigates overfitting, reduces sensitivity to extreme propensity scores, and supports more reliable inference. At the same time, it allows heterogeneous treatment effects to be identified only in regions of the covariate space where sufficient overlap exists, i.e., where treated and control units are comparable.

A second advantage concerns potential endogeneity arising from imperfect observability of relevant covariates. While the proposed approach does not fully eliminate this issue, it mitigates its impact by reducing reliance on restrictive functional form assumptions and by orthogonalizing the estimation with respect to the nuisance functions. As a result, first-order bias due to estimation errors in the nuisance functions is attenuated and does not directly carry over to the estimated treatment effects.

Nonetheless, limitations related to imperfect overlap and potential residual endogeneity remain in our empirical setting. We therefore perform a range of robustness checks and sensitivity analyses, which are discussed in Section 7.

6. Results

The estimation procedure described in the previous section is here applied to the six outcome variables: π, P, GPV, P/π, P/GPV, π/GPV.[13] As the interest here concerns the sustainability of organic farming adoption only IATE, GATE, ATE estimates are reported and commented.[14] The set of covariates used in the different estimation stages are reported in Table A1 (Annex 1). When not otherwise specified, estimates are obtained using the reference AIPW estimator combining square-root LASSO for the outcome model, Logit LASSO for the treatment model, and Generalized Random Forest for CATE estimation.

6.1. The average treatment effect (ATE)

The upper part of Table 2 reports the ATE estimates for all outcome variables. The direction of the effects is consistent with the descriptive evidence reported in Table 1 for P and GPV. In particular, the adoption of organic farming is associated with a reduction in the overall value of agricultural production, alongside a decrease in the use of environmentally harmful inputs. In contrast, the effect on π displays the opposite sign, as organic adoption is associated with a slight increase in net farm income. Although only indicative, this comparison suggests that economic sustainability is achieved from the farmer's perspective, while from a societal perspective the income gains alone appear smaller than the public resources devoted to supporting organic farming.

Moreover, statistical significance varies across outcomes. The estimated ATE is significantly different from zero for the reduction in gross production value (GPV) and in the use of unsustainable inputs (P), whereas the effect on net farm income is not statistically significant. The relatively large standard error associated with income suggests substantial variability in treatment effects across farms in this respect.

The estimates for the ratio-based indicators largely confirm these findings. Organic farming adoption leads to lower expenditure on environmentally harmful inputs relative to both net farm income, $\Delta(P/\pi)$, and gross production value, $\Delta(P/GPV)$, although statistical significance is only observed for the former. At the same time, when profitability is measured in relative terms, namely per unit of

[13] All estimates, tests and calculations presented in this study have been performed with software STATA 19.
[14] Other ancillary results (like coefficients estimation of the treatment and outcome model) are available upon request.

gross production value $\Delta(\pi/GPV)$, it appears to be positively and significantly affected by organic farming adoption.

Overall, the ATE estimates point to mixed evidence regarding the sustainability implications of organic farming. Not only do different outcome dimensions respond differently to treatment, but the uncertainty surrounding some estimates is also consistent with substantial heterogeneity in farm-level responses. This variability indicates that average effects may mask important distributional patterns, thus motivating a more detailed investigation of heterogeneous treatment effects.

Table 2 – ATE and quintile-based GATE estimates (estimated standard errors in parenthesis).

Outcome variable:	$\Delta\pi$ (€)	ΔP (€)	ΔGPV (€)	$\Delta(P/\pi)$	$\Delta(P/GPV)$	$\Delta(\pi/GPV)$
ATE	1697 (2973)	-10912*** (3101)	-11598*** (4294)	-0.863*** (0.128)	-0.091 (0.072)	0.502** (0.235)
GATE:						
Q1	-20855 (11568)	-39262*** (12792)	-30109* (17706)	-2.45*** (0.486)	-0.093*** (0.023)	0.043 (0.293)
Q2	-8947 (6488)	-5567*** (1899)	-18401** (9024)	-0.857*** (0.244)	-0.089** (0.041)	0.253 (0.182)
Q3	-1003 (8905)	-7688 (5231)	-2118 (2695)	-0.494** (0.214)	-0.309 (0.213)	0.330* (0.197
Q4	11507 (21729)	-2032** (1019)	-114.2 (3279)	-0.225*** (0.068)	-0.066 (0.049)	0.417*** (0.114)
Q5	26363* (15416)	-17.52 (2668)	769.8 (2669)	-0.196 (0.142)	0.143 (0.273)	1.29 (1.08)
GATE heterogeneity test (H0: GATEs are homogeneous): $\chi^2(4)$	5.52	12.78**	9.56**	27.36***	2.03	3.28

Levels of statistical significance: *** = 1%, ** = 5%, * = 10%.

LEGEND: Q1: Lowest IATEs quintile (0-20%); Q2: Low–medium IATEs quintile (20–40%); Q3: Medium IATEs quintile (40–60%); Q4: Medium–high IATEs quintile (60-80%); Q5: Highest IATEs quintile (80-100%).

6.2. Heterogeneous treatment effect: the GATEs

6.2.1. Data-driven GATEs: quantiles

ATE estimates largely confirm the evidence reported in previous studies, showing that economic and environmental sustainability do not necessarily coincide in organic farming adoption (e.g. Esposti, 2024; Brutti and Freo, 2026). Such findings may partly reflect the substantial heterogeneity in responses to organic adoption. To explore this heterogeneity, we first consider data-driven GATEs, namely ATEs computed over observations grouped into quantiles of the estimated IATE distribution. Here, quantiles correspond to quintiles (Q1–Q5). In principle, one might expect $\mathrm{GATE}_{\mathrm{Q1}} \leq \mathrm{GATE}_{\mathrm{Q2}} \leq \cdots \leq \mathrm{GATE}_{\mathrm{Q5}}$, with Q1 and Q5 representing the lowest and highest predicted effects, respectively. Because the ranking is based on estimated rather than true treatment effects, however, monotonicity of the GATEs is not guaranteed. In practice, deviations from monotonicity may arise due to estimation noise and finite-sample variability. Consequently, GATE estimates should be interpreted as exploratory rather than structural.

The lower panel of Table 2 reports the estimated GATEs. Overall, monotonicity is largely respected, with only a minor deviation between Q3 and Q4 in the case of $\Delta(P/\pi)$. More generally, estimated treatment effects tend to increase when moving from the lower to the upper quintiles. In addition, sign changes across quintiles are observed for $\Delta\pi$, ΔGPV, and ratio $\Delta(P/GPV)$, further confirming the presence of substantial heterogeneity in treatment effects.

In fact, differences across quintiles are not always statistically robust. For $\Delta\pi$, GATE estimates largely mirror the ATE results, indicating no statistically significant impact for most quintiles. The only exception is Q5, where the effect is positive and weakly significant, exceeding €25,000 in terms of net farm income. Nevertheless, the formal test for GATE heterogeneity reported at the bottom of Table 2 does not provide strong support for significant differences across quintiles.

A different pattern emerges for ΔP. In this case, no change in sign is observed across quintiles, with all groups displaying a reduction in the use of environmentally harmful inputs under organic farming. This reduction is statistically significant for three groups (Q1, Q2, and Q3), which explains the rejection of the null hypothesis of homogeneous effects across quintiles.

For the outcome variable ΔGPV, the heterogeneity test also rejects homogeneity. However, the pattern is less clear-cut: only the lowest two quintiles (Q1 and Q2) exhibit statistically significant and negative effects, while estimates for the remaining groups are weaker and a change in sign is observed for the highest quintile (Q5).

When ratio-based indicators are considered, the results reveal a more pronounced divergence across cases. For $\Delta(P/\pi)$, the estimated effect is consistently negative and statistically significant for all groups except Q5, and the heterogeneity test strongly rejects the null of homogeneous effects. By contrast, for both $\Delta(P/GPV)$ and $\Delta(\pi/GPV)$, the test fails to reject homogeneity. In the former case, a change in sign (from negative to positive) is observed in the highest quintile (Q5), although statistical significance is limited to Q1 and Q2. In the latter case, the estimated effects remain positive across all quintiles, but statistical significance is observed only for Q3 and Q4.

6.2.2. Gender and age

As discussed in Section 5, an alternative way to assess treatment effect heterogeneity concerns pre-defined groups. We therefore examine whether the impact of organic farming varies with key structural characteristics of the farmer and the farm. Table 3 reports GATE estimates for the six outcome variables, distinguishing by gender and age of the farm holder and by farm economic size.

Focusing first on gender, the null hypothesis of homogeneous treatment effects cannot be rejected for five of the six outcomes, and is rejected only for $\Delta(P/\pi)$. However, for both $\Delta\pi$ and $\Delta(\pi/GPV)$, treatment effects are not statistically different from zero for either gender. For the remaining outcomes, ΔP, ΔGPV, $\Delta(P/\pi)$, and $\Delta(P/GPV)$, at least one gender group exhibits statistically significant effects. In these cases, farms managed by female holders tend to perform slightly better in terms of sustainability than those managed by male holders, most notably for $\Delta(P/\pi)$.

Turning to age, we compare younger (below 40 years) and older (above 40 years) farmers. The results broadly mirror those for gender, as the heterogeneity tests consistently fail to reject homogeneity across all outcomes. In addition, for $\Delta\pi$ the treatment effect is not statistically different from zero for either age group. However, some differences emerge in the magnitude and sign of the effects. In particular, older farmers exhibit statistically significant and negative impacts on ΔP and ΔGPV, whereas no significant effects are observed for younger farmers. This pattern reverses for $\Delta(P/\pi)$, while no significant differences are found for the other ratio-based indicators, $\Delta(\pi/GPV)$ and $\Delta(P/GPV)$. Overall, unlike the gender comparison, age-related differences hint at possible structural heterogeneity, potentially reflecting different strategic motivations in the adoption of organic farming.

Table 3 – GATE estimates by selected groups: gender, age and size (estimated standard errors in parenthesis).

Outcome variable:	$\Delta\pi$ (€)	ΔP (€)	ΔGPV (€)	$\Delta(P/\pi)$	$\Delta(P/GPV)$	$\Delta(\pi/GPV)$
Gender:						
Female	-3665	-13186***	-9678	-1.48***	-0.151**	0.180
	(5792)	(4112)	(10286)	(0.408)	(0.060)	(0.481)
Male	8223	-10279***	-13634***	-0.718***	-0.067	0.373
	(7858)	(3572)	(3783)	(0.125)	(0.085)	(0.480)
GATE heterogeneity test (H0: GATE are homogeneous): $\chi^2(1)$	2.18	0.28	0.46	3.28*	0.66	0.08
Age:						
Young farm holder[a]	8190	-8989	-6380	-1.37***	0.147	1.379
	(7089)	(4821)	(4615)	(0.414)	(0.269)	(1.479)
Older farm holder	1914	-12751***	-14761***	-0.741***	-0.104	0.232
	(2250)	(2359)	(5023)	(0.126)	(0.071)	(0.413)
GATE heterogeneity test (H0: GATE are homogeneous): $\chi^2(1)$	1.34	0.37	2.30	2.13	0.81	0.56
Size:[b]						
Small	8346**	-908*	-30726	-2.179**	0.983*	1.061***
	(3807)	(535)	(34332)	(0.880)	(0.558)	(0.408)
Medium	2363	-1333	-2081	-0.907***	0.028	0.718
	(7025)	(1186)	(1777)	(0.179)	(0.095)	(0.538)
Large	-1042	-21169***	-19293*	-0.677***	-0.218**	-0.324
	(3394)	(6476)	(11177)	(0.173)	(0.104)	(0.491)
GATE heterogeneity test (H0: GATE are homogeneous): $\chi^2(2)$	7.12**	9.76***	3.18	3.31	6.65**	8.52**

Levels of statistical significance: *** = 1%, ** = 5%, * = 10%.

[a] ≤ 40 years old.

[b] Small farms: €8.000-<€25,000 Standard Output; Medium farms: €25,000–€100,000 Standard Output; Large farms: > €100,000 Standard Output.

6.2.3. Farm size and production specialization

Heterogeneity across farm size classes (small, medium, and large) is more pronounced and informative. The homogeneity hypothesis is rejected for four outcomes: $\Delta\pi$, ΔP, $\Delta(P/GPV)$, and $\Delta(\pi/GPV)$. For medium-sized farms, treatment effects are generally not statistically significant, with the exception of $\Delta(P/\pi)$. A clear contrast instead emerges between small and large farms. Small farms display statistically significant effects for all outcomes except ΔGPV: $\Delta\pi$, $\Delta(P/GPV)$ and $\Delta(\pi/GPV)$ are positive, while ΔP and $\Delta(P/\pi)$ are slightly negative. Conversely, large farms show significant and strongly negative effects for ΔP, ΔGPV, $\Delta(P/\pi)$, and $\Delta(P/GPV)$, whereas the effects on $\Delta\pi$ and $\Delta(\pi/GPV)$ are not statistically significant. These patterns suggest that, even more clearly than in the age comparison (to which they may be partly related), farm size is associated with divergent strategies in the adoption of organic farming. Large farms appear to prioritize cost reduction, as reflected in a markedly lower use of unsustainable inputs, while small farms seem to focus more on profitability.

The structural characteristic exhibiting the most pronounced and policy-relevant TE heterogeneity is farm production specialization. Figure 2 reports the GATE estimates by eight Type of Farming (TF) categories. In fact, it should be recalled that these groups are not internally homogeneous, and their degree of within-group heterogeneity varies across categories. For example, the "Horticulture" category is relatively homogeneous, whereas categories such as "Mixed crops & livestock" and "Other field crops" are, by construction, more residual and internally diverse. As a result, Figure 2 shows that the standard errors associated with the estimated GATE vary considerably across categories. Importantly, for none of the outcome variables are the estimated GATEs statistically different from zero across all TF categories. Nevertheless, whenever statistically significant effects are detected, they display a consistent sign across categories and are aligned with the corresponding ATE estimates (see Table 2).

For $\Delta\pi$ we obtain a positive and significant effect for categories “Cereals” and “Granivores”. For ΔP we observe a negative and significant impact on all TFs implying livestock activities and “Cereals”. For ΔGPV a negative and significant impact is found for “Mixed crops&livestock”, “Horticulture” and “Cereals”. For $\Delta(P/\pi)$ a negative and significant estimate is obtained for all categories but for “Mixed crops&livestock” while category “Granivores” shows the lowest value. For $\Delta(P/GPV)$ the only statistically significant GATE is negative and concerns categories “Cereals” and “Granivores” while for $\Delta(\pi/GPV)$ the positive significant effect is found for “Horticulture”, “Cereals” and “Mixed crops&livestock”. A broad pattern across TFs is that the categories most affected by the adoption of organic farming are livestock systems and extensive crops (notably cereals). However, even in these cases, the combination of effects does not point to a fully sustainable outcome.

6.2.4. Treatment status

A final assessment on the heterogeneity of the TE concerns the comparison between treated units (always organic farms) and untreated units (always conventional farms). This comparison is reported in Table 4. These estimates should not be interpreted as GATEs in a strict sense since the grouping variable (treatment status) is directly involved in the definition of the potential outcomes framework, rather than representing an exogenous partition of the covariate space. Accordingly, Table 4 simply reports averages of the estimated IATEs within the treated and untreated subpopulations. Therefore, the reported subgroup averages of the estimated individual treatment effects are presented for descriptive purposes only. Because these quantities are themselves estimated objects obtained from the same first-stage models, standard tests for differences in means are not valid. Consequently, no formal statistical test is conducted for the difference between subgroup averages, and the comparison should be interpreted as illustrative rather than inferential.

Although the sign of the estimated impact is consistent between treated and untreated farms for all outcome variables, the magnitude of the effect is systematically more favorable for treated units. This points to a higher level of sustainability among adopters: higher farm net income and gross production value, as well as of $\Delta(\pi/GPV)$ coupled with a lower relative use of harmful inputs, $\Delta(P/\pi)$ and $\Delta(P/GPV)$. This evidence points to a positive selection into treatment based on expected gains thus suggesting that the voluntary adoption of organic farming, i.e. farmers’ self-selection into treatment, is already oriented towards more sustainable outcomes. Adopters are primarily driven by private economic gains (higher π), yet their choices also generate positive societal effects, as reflected in reductions in P and, relative to the counterfactual, improvements in ΔGPV. Overall, self-selection thus appears to be an efficient mechanism for the diffusion of organic farming. However, this also implies that extending adoption to non-adopters through mandatory policies may yield smaller gains than those observed among voluntary adopters.

Table 4 – Mean estimated ATE by treatment status for the six outcome variables (within-group standard deviation in parenthesis).

Outcome variable:	Treated units	Control units	Difference
$\Delta\pi$ (€)	4652 (361)	3585 (2483)	977
ΔP (€)	-10069 (795)	-6612 (2051)	-3457
ΔGPV (€)	-6929 (1724)	-11786 (633)	4947
$\Delta(P/\pi)$	-0.894 (0.085)	-0.864 (0.194)	-0.030
$\Delta(P/GPV)$	-0.198 (0.033)	-0.087 (0.005)	-0111
$\Delta(\pi/GPV)$	0.414 (0.025)	0.203 (0.150)	0.211

6.3. The Individualized treatment Effects (IATEs) and the sustainability octants

All GATE estimates are derived from the underlying IATEs, which therefore constitute the most informative object of analysis. Figure 3 reports the distributions of IATEs for the six outcome variables, along with the corresponding tests for treatment effect heterogeneity. As expected, the distributions of level-based outcome indicators display wider tails than those of ratio-based measures. Nonetheless, statistical evidence of heterogeneity is more pronounced for the latter. In particular, for $\Delta\pi$ the null hypothesis of homogeneous treatment effects cannot be rejected. In this case, the distribution of IATEs is strongly concentrated around zero, as might be expected for a key indicator of economic performance in a highly competitive private-sector context. However, the distribution also exhibits relatively long tails on both sides, with a small number of extreme values, more frequent on the right tail, likely reflecting idiosyncratic and possibly temporary conditions.

For ΔP, the null hypothesis of homogeneous treatment effects is rejected. The distribution is concentrated to the left of zero, with a long left tail and relatively few positive values clustered around zero. This pattern provides strong support for one of the main findings of the analysis: despite the presence of heterogeneity, organic farming adoption is predominantly associated with a reduction in the use of environmentally harmful inputs. A similar pattern emerges for ΔGPV, for which the homogeneity hypothesis is also rejected. However, in this case the distribution displays a sharper concentration of negative values around zero.

For the ratio-based indicators, the null of homogeneous treatment effects is rejected for all outcomes at the 10% significance level. For $\Delta(P/\pi)$, the distribution exhibits substantial dispersion and asymmetry, with all values being negative, no mass around zero, and a long and dense left tail. For $\Delta(P/GPV)$, heterogeneity appears less pronounced. Although negative values still prevail, the distribution shows a marked concentration around zero and a relatively longer right tail. This pattern reflects the fact that, in several cases, GPV declines more than P, leading to a sharper reduction in the ratio P/GPV. Finally, for $\Delta(\pi/GPV)$, the distribution is approximately symmetric and close to normal, with a strong concentration around zero and a slight shift to the right, accompanied by a longer right tail. This indicates that, for a subset of farms, either π increases while GPV decreases, or π changes more favorably than GPV (increasing more or declining less).

Another major advantage in estimating IATE is a deeper investigation of the linkage between the observed TE heterogeneity and some major farms characteristics. In particular, as stressed in Section 6.2.3, the economic size seems to play a major role. This relevance can be investigated further by jointly considering the physical size, as expressed by the farm's utilized agricultural area (UAA), and the economic size as expresses by revenue instead of the size classes considered above. Figure 4 reports the heatmaps for outcome variables $\Delta\pi$, ΔP and ΔGPV across the two size dimensions.[15] The three heatmaps provide a clear and informative representation of the substantial heterogeneity in the effects of organic farming adoption across these two key structural dimensions. They offer a useful visual complement to the GATE and IATE results, highlighting patterns that would otherwise remain hidden in average effects.

For $\Delta\pi$, the heatmap reveals an uneven distribution of impacts across the two dimensions. While positive effects prevail in most of the support, their magnitude varies considerably. In particular, the stronger gains, namely highest treatment effects (yellow/red area), are concentrated in the upper-left region of the heatmap corresponding to farms' small land endowments (lowest UAA quantiles) and higher revenue levels (highest revenue quantiles). This pattern suggests the presence of scale–intensity interactions, whereby land-constrained but revenue-intensive farms experience particularly large gains from organic adoption. Conversely, farms with medium to large UAA and intermediate

[15] Heatmaps of the ratio-based outcome variables are available upon request.

revenue levels (central-right part of the heatmap) exhibit much weaker impacts, namely the lowest IATE values (dark blue area).

This pattern suggests that the private economic benefits of organic adoption are not uniformly distributed, but rather concentrated among specific types of farms, possibly those better able to exploit price premiums or market niches. Importantly, the variation in treatment effects is not monotonic along either dimension when considered separately. Instead, the heatmap reveals a joint, nonlinear interaction between land size and revenue performance: treatment effects do not systematically increase or decrease with farm size alone, nor do they follow a clear pattern across revenue quantiles.

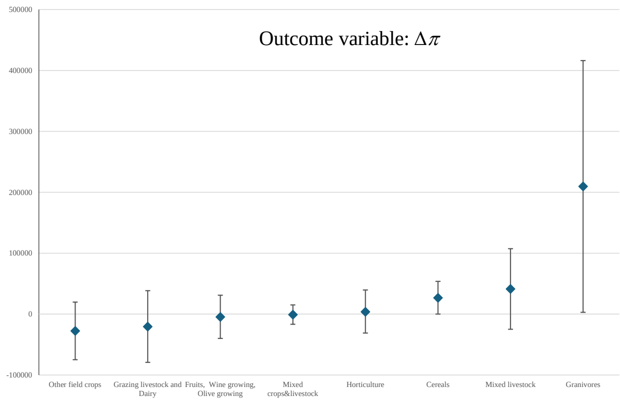
The heatmap for ΔP displays a more regular and monotonic pattern. A reduction in the use of environmentally harmful inputs is observed across all combinations, confirming the environmental effectiveness of organic farming. However, the intensity of this reduction is clearly heterogeneous and increases markedly with farm size. Larger farms exhibit substantially stronger reductions in unsustainable input use (dark blue area). Not surprisingly, this result points to a scale-related environmental benefit, whereby larger farms have greater potential for reducing environmental pressure.

The heatmap for ΔGPV is quite similar to that for ΔP in showing a consistently negative effect that becomes more pronounced as farm size increases. Smaller farms display relatively limited reductions in gross production value, whereas larger farms experience substantially larger decreases (dark blue area). This pattern further highlights the potential trade-off between environmental and production performance: the stronger the environmental gains, the larger the reduction in production levels, particularly for larger farms.

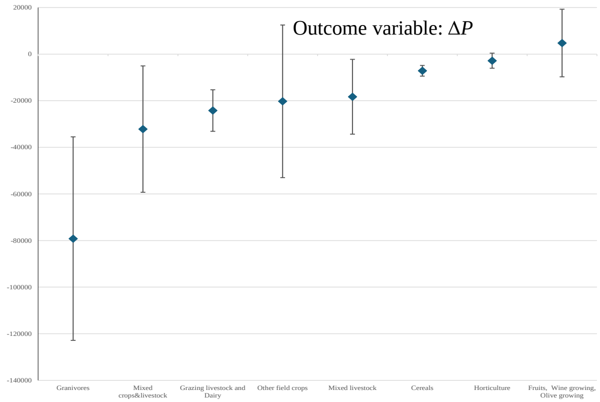
Overall, the three heatmaps confirm that the effects of organic farming adoption are strongly heterogeneous and vary systematically across farm characteristics. In particular, in areas where environmental improvements are most pronounced, production tends to decline, while economic benefits appear confined to specific segments. This evidence reinforces not only the idea that ATEs provide an incomplete picture, and that a disaggregated analysis is essential to capture the multidimensional and context-dependent nature of outcomes associated with organic farming. More importantly, this evidence indicates that this heterogeneity is inherently multidimensional and cannot be adequately captured by stratifying along a single covariate. Rather, it emerges from the interaction of multiple structural characteristics.

This multidimensionality of organic farming adoption is what eventually motivates its multidimensional sustainability performance. In the present analysis, the primary interest of the IATE estimates lies in their use for assigning farms to the sustainability octant framework introduced in Section 3 (Figure 1). Table 5 classifies all units according to the combination of their estimated IATEs across the different outcome variables. The upper panel of the table reports combinations based on the three level-based outcome variables ($\Delta\pi$, ΔP, ΔGPV), whereas the lower panel presents the corresponding combinations for the ratio-based measures, $\Delta(P/\pi)$, $\Delta(P/GPV)$, $\Delta(\pi/GPV)$. Fully sustainable outcomes are very rare: only two farms display a win–win–win configuration when absolute changes are considered, and none when ratio-based measures are used, as these capture not only the direction but also the relative magnitude of changes. By contrast, the opposite octant framework, representing a simultaneous loss across all dimensions, although still a minority, is substantially more populated, with 52 and 81 farms, respectively.

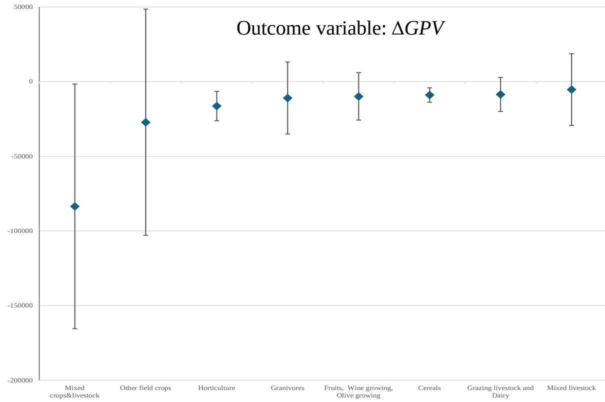
Figure 2 – GATE estimates for 8 Type of Farming (TF) categories in ascending order: point estimates with 95% confidence interval error bars and treatment-effects heterogeneity test.



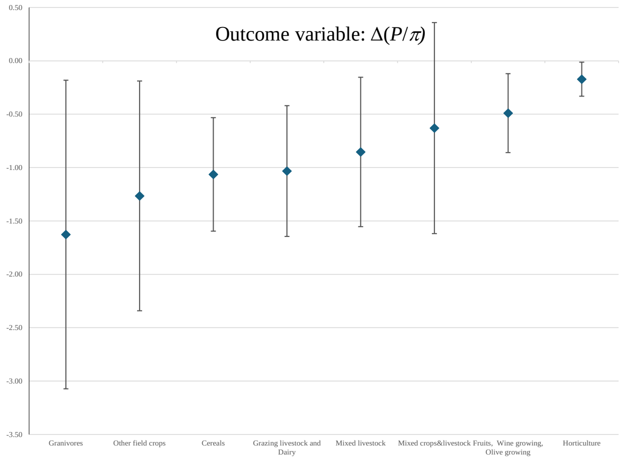
Treatment-effects heterogeneity test: $\chi^2(7) = 10.76$



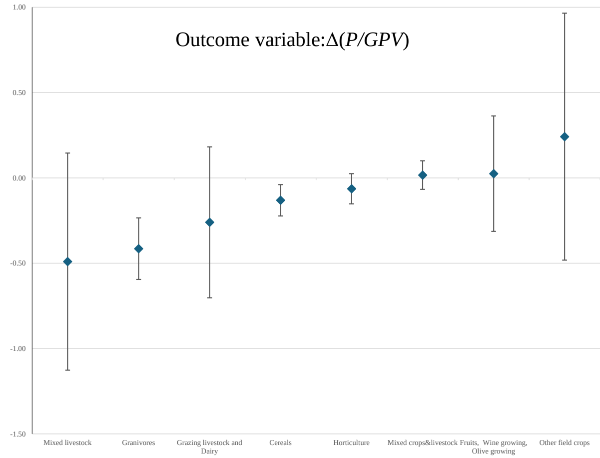
Treatment-effects heterogeneity test: $\chi^2(7) = 39.52$***



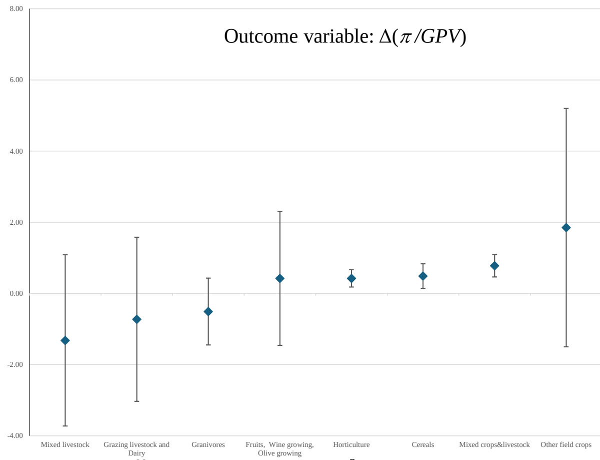
Treatment-effects heterogeneity test: $\chi^2(7) = 5.26$



Treatment-effects heterogeneity test: $\chi^2(7) = 25.21$***



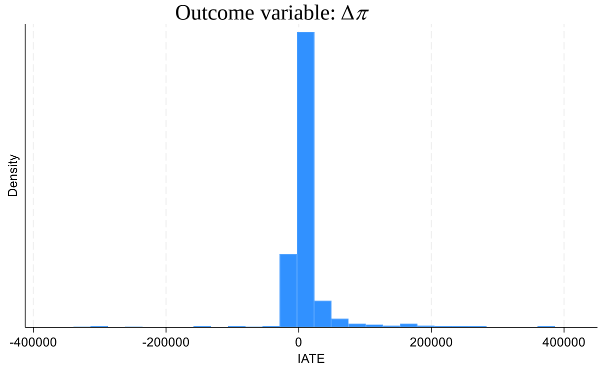
Treatment-effects heterogeneity test: $\chi^2(7) = 22.01$***



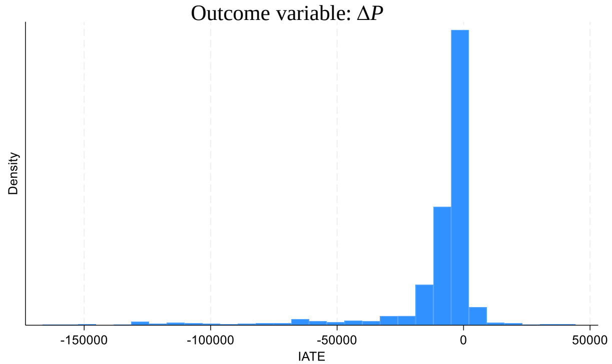
Treatment-effects heterogeneity test: $\chi^2(7) = 12.17$*

Levels of statistical significance: *** = 1%, ** = 5%, * = 10%

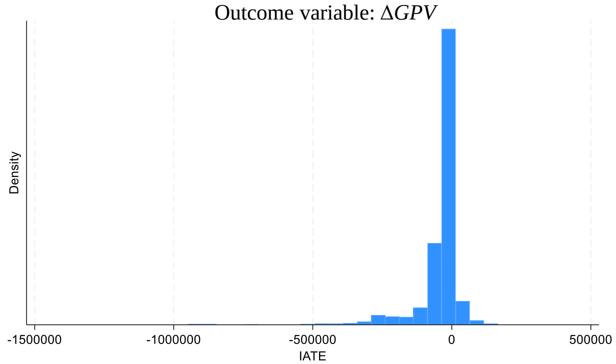
Figure 3 – IATE histograms and treatment-effects heterogeneity test (H0: treatment effects are homogeneous; p-values in parenthesis).



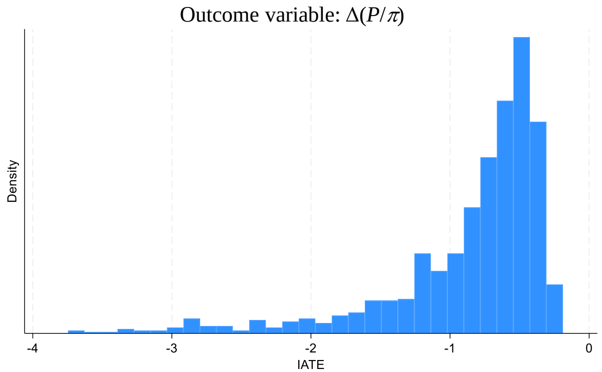
Treatment-effects heterogeneity test: $\chi^2(1)$ = 1.48 (0.22)



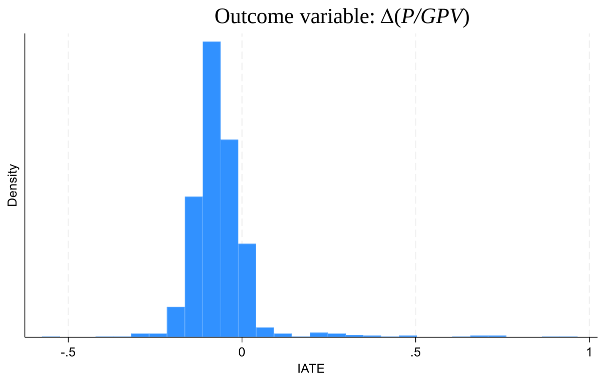
Treatment-effects heterogeneity test: $\chi^2(1)$ = 30.66 (0.00)***



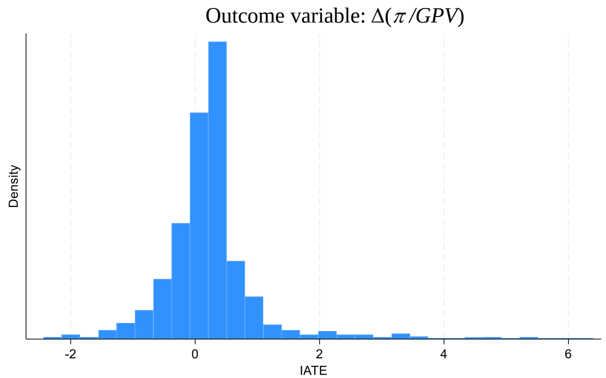
Treatment-effects heterogeneity test: $\chi^2(1)$ = 7.96 (0.01)***



Treatment-effects heterogeneity test: $\chi^2(1)$ = 16.42 (0.00)***



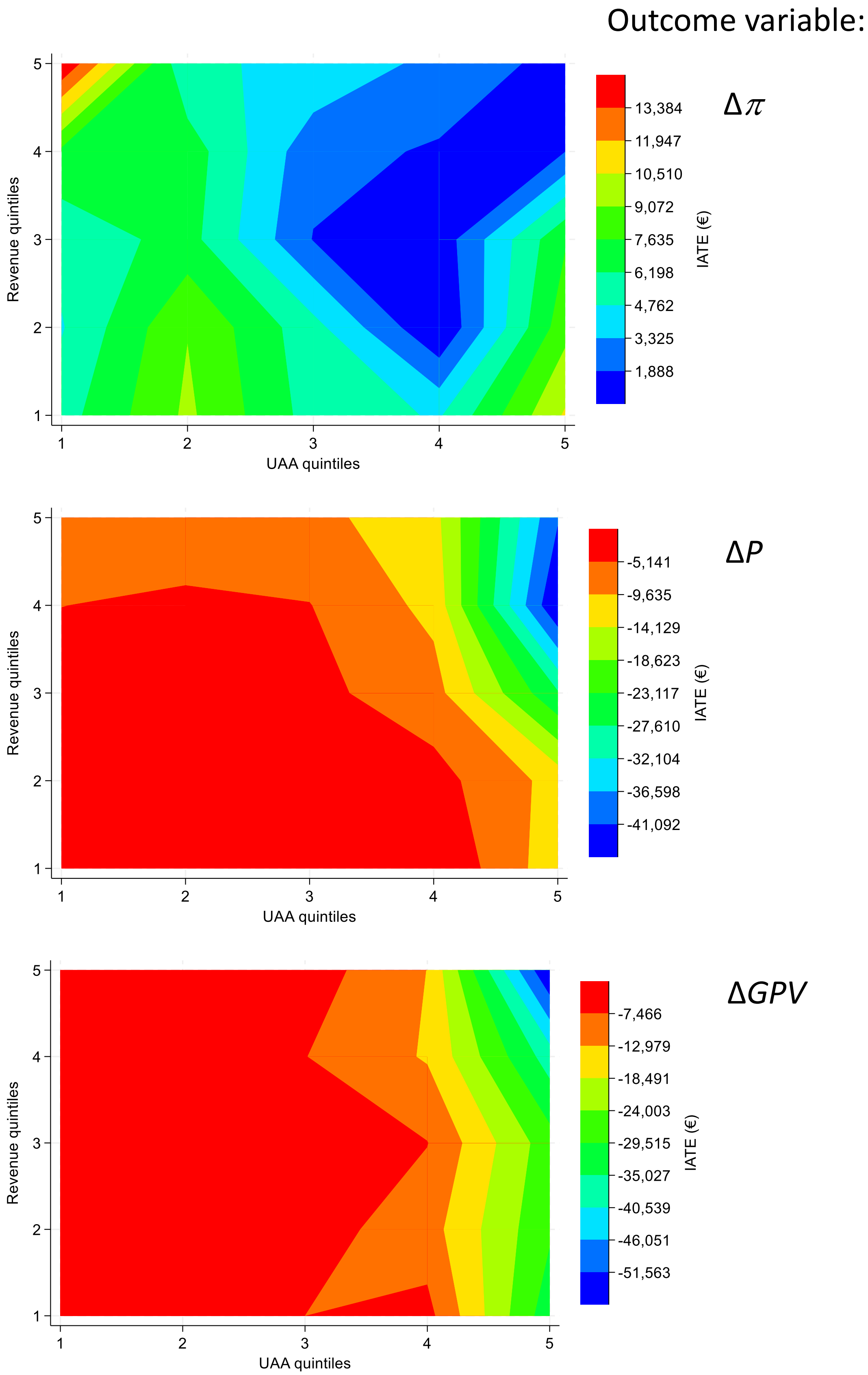
Treatment-effects heterogeneity test: $\chi^2(1)$ = 2.98 (0.09)*



Treatment-effects heterogeneity test: $\chi^2(1)$ = 4.40 (0.04)**

Levels of statistical significance: *** = 1%, ** = 5%, * = 10%.

Figure 4 – Heatmap of estimated IATEs by quintiles of utilized agricultural area (UAA) and farm revenue.[a]



[a] Each cell represents the estimated IATE within a subgroup defined by the joint distribution of farm's UAA and revenues. Colors transition from dark blue (low IATE) to yellow/red (high IATE), with the color scale reported on the right-hand side.

Table 5 – Sustainability octants (bold entries identify the extreme profiles, W-W-W and L-L-L, and the most frequent group).

Combination	Description	N. of cases
$\Delta\pi>0$, $\Delta P<0$, $\Delta GPV>0$: WIN-WIN-WIN	**Private economic gain, environmental conservation, supply maintenance**	**2**
$\Delta\pi>0$, $\Delta P>0$, $\Delta GPV>0$: WIN-LOSS-WIN	Private economic gain, environmental depletion, supply maintenance	1
$\Delta\pi>0$, $\Delta P<0$, $\Delta GPV<0$: WIN-WIN-LOSS	**Private economic gain, environmental conservation, supply shortage**	**735**
$\Delta\pi>0$, $\Delta P>0$, $\Delta GPV<0$: WIN-LOSS-LOSS	Private economic gain, environmental depletion, supply shortage	61
$\Delta\pi<0$, $\Delta P<0$, $\Delta GPV>0$: LOSS-WIN-WIN	Private economic loss, environmental conservation, supply maintenance	31
$\Delta\pi<0$, $\Delta P>0$, $\Delta GPV>0$: LOSS-LOSS-WIN	Private economic loss, environmental depletion, supply maintenance	3
$\Delta\pi<0$, $\Delta P<0$, $\Delta GPV<0$: LOSS-WIN-LOSS	Private economic loss, environmental conservation, supply shortage	170
$\Delta\pi<0$, ΔP>0, $\Delta GPV<0$: LOSS-LOSS-LOSS	**Private economic loss, environmental depletion, supply shortage**	**52**
$\Delta(\pi/GPV)>0$, $\Delta(P/\pi)<0$, $\Delta(P/GPV)<0$: WIN-WIN-WIN	**Productive efficiency gain, environmental efficiency, eco-economic synergy**	**0**
$\Delta(\pi/GPV)>0$, $\Delta(P/\pi)>0$, $\Delta(P/GPV)<0$: WIN-LOSS-WIN	Productive efficiency gain, environmental inefficiency, eco-economic synergy	0
$\Delta(\pi/GPV)>0$, $\Delta(P/\pi)<0$, $\Delta(P/GPV)>0$: WIN-WIN-LOSS	Productive efficiency gain, environmental efficiency, eco-economic deterioration	2
$\Delta(\pi/GPV)>0$, $\Delta(P/\pi)>0$, $\Delta(P/GPV)>0$: WIN-LOSS-LOSS	Productive efficiency gain, environmental inefficiency, eco-economic deterioration	1
$\Delta(\pi/GPV)<0$, $\Delta(P/\pi)<0$, $\Delta(P/GPV)<0$: LOSS-WIN-WIN	**Productive inefficiency, environmental efficiency, eco-economic synergy**	**663**
$\Delta(\pi/GPV)<0$, $\Delta(P/\pi)>0$, $\Delta(P/GPV)<0$: LOSS-LOSS-WIN	Productive inefficiency, environmental inefficiency, eco-economic synergy	53
$\Delta(\pi/GPV)<0$, $\Delta(P/\pi)<0$, $\Delta(P/GPV)>0$: LOSS-WIN-LOSS	Productive inefficiency, environmental efficiency, eco-economic deterioration	255
$\Delta(\pi/GPV)<0$, $\Delta(P/\pi)>0$, $\Delta(P/GPV)>0$: LOSS-LOSS-LOSS	**Productive inefficiency, environmental inefficiency, eco-economic deterioration**	**81**

While only a limited number of farms are scattered across the remaining octants, the most prevalent configurations are the win–win–loss combination (approximately 70% of farms) when considering level-based indicators, and the loss–win–win combination (about 63%) when using ratio-based measures. Overall, these patterns suggest that the adoption of organic farming is often associated with both private economic gains and environmental improvements. However, these benefits are not accompanied by the preservation of the overall value of agricultural production. In particular, the predominance of $\Delta(\pi/GPV)$<0 points less to a conflict between private and societal interests than to tensions among competing societal objectives. While environmental improvements appear robust, as indicated by $\Delta(P/\pi)$<0 and $\Delta(P/GPV)$<0, the results suggest that organic farming adoption may not fully preserve productive efficiency from a broader societal perspective.

7. Validation

The ATE, GATE and IATE estimates presented above require careful validation for two main reasons. First, these estimates rely on data-intensive machine learning methods. While such approaches offer considerable flexibility in capturing complex, high-dimensional relationships, they may also exhibit sensitivity to extreme observations and sample-specific patterns, potentially leading to instability and limited robustness (Belloni et al., 2014; Chernozhukov et al., 2018; Athey and Imbens, 2019). Second, the quasi-experimental design adopted in this study is subject to inherent limitations in approximating the ideal conditions of a randomized experiment. In particular, identification may be challenged by violations of key assumptions such as overlap and unconfoundedness (CIA), which are difficult to fully verify in observational settings. These limitations may raise concerns about the validity of the estimated treatment effects.

This section therefore presents a set of sensitivity analyses and robustness checks aimed at assessing the stability of the results with respect to both the statistical properties of the machine learning estimators and the identification challenges arising from the observational nature of the data.

7.1. Alternative estimations

To assess the sensitivity of the results to alternative specifications of the nuisance functions, we consider a set of estimation variants relative to the baseline approach described in Section 5. First, the fully interactive formulation of the outcome model in (7) is replaced by a restricted specification with an additive structure, in which the outcome equation is assumed to be identical across treatment groups and the treatment effect is separable (equation (2)). In this case, the AIPW-based construction of the pseudo-outcome is replaced by a conventional estimation of the nuisance functions $\tau(\mathbf{x}_i)$ and $\varphi(\mathbf{x}_i)$, followed by the same second-stage procedure. This approach corresponds to the so-called partialling-out (PO) estimation (Nie and Wager, 2021).

Second, within both the AIPW and PO frameworks, alternative machine learning estimators are employed for the nuisance functions. We consider replacing the square-root and Logit LASSO used in the baseline specification with three alternative estimation strategies. Alternative 1 adopts the PO approach, using linear LASSO for the outcome model and Logit LASSO for the treatment equation. Alternative 2 relies on the AIPW framework, with Random Forest estimators applied to both outcome and treatment functions. Finally, Alternative 3 combines the PO estimation with Random Forest methods for both nuisance functions. In all cases, the moment condition (10) is solved via the Generalized Random Forest approach.

Table 6 provides the results of these alternative estimations. For space limitations, only ATE estimates for the six outcome variables are reported.[16] Relative to the baseline estimates reported in Table 2, the alternative estimators generally exhibit lower statistical precision. However, the main qualitative conclusions remain remarkably stable across estimation approaches, despite some variability in the magnitude and statistical significance of the estimated effects. In particular, the baseline result of a non-significant effect of organic adoption on farm profitability ($\Delta\pi$) is largely confirmed across all alternative specifications. While point estimates are always positive, their magnitude varies considerably and they remain statistically insignificant in all cases.

[16] Respective GATE and IATE estimates are available upon request.

For environmental sustainability (ΔP), all alternative estimations confirm a reduction in the use of environmentally harmful inputs, although both magnitude and significance exhibit some sensitivity. The strongest and most robust results are obtained under Alternative 2, where the effect remains statistically significant, while other specifications produce smaller and less precise estimates. This suggests that, although the environmental effect is stable in sign, its magnitude is partly driven by modelling choices. A similar pattern is observed for production outcomes (ΔGPV). The negative effect associated with organic adoption is confirmed across all specifications, but statistical significance is only achieved under Alternative 2.

When considering ratio-based indicators, the most robust result concerns $\Delta(P/\pi)$, which remains negative and statistically significant across all estimators, indicating a consistent improvement in the balance between environmental and economic performance. By contrast, the remaining ratios show greater variability, with statistical significance depending on the estimation method, although their qualitative interpretation remains broadly unchanged.

Taken together, these results suggest that the core findings of the analysis, namely, the environmental effectiveness of organic farming and the absence of systematic economic gains, are robust to alternative modelling choices. At the same time, the variability observed in the magnitude and statistical significance of several effects highlights the sensitivity of the findings to both model specification and the choice of machine learning techniques. This also reinforces the need for complementary robustness and sensitivity analyses whenever highly flexible, nonlinear methods are employed.

Table 6 – ATE estimates from alternative estimation approaches (estimated standard errors in parenthesis).

Outcome variable:	$\Delta\pi$ (€)	ΔP (€)	ΔGPV (€)	$\Delta(P/\pi)$	$\Delta(P/GPV)$	$\Delta(\pi/GPV)$
Alternative 1:	2142 (8749)	-3106 (2633)	-9310 (10351)	-0.882*** (0.124)	-0.046 (0.061)	0.379 (0.343)
Alternative 2:	3344 (3237)	-6588*** (1687)	-19419* (11012)	-0.855*** (0.123)	-0.053* (0.027)	0.503** (0.234)
Alternative 3:	2884 (5093)	-8494 (7340)	-16698 (14353)	-0.799*** (0.137)	-0.044 (0.037)	0.307* (0.182)

Levels of statistical significance: *** = 1%, ** = 5%, * = 10%.
LEGEND:
Alternative 1: PO estimation; Outcome: Linear LASSO; Treatment: Logit LASSO; IATE: Generalized Random Forest.
Alternative 2: AIPW estimation; Outcome: Random Forest; Treatment: Random Forest; IATE: Generalized Random Forest.
Alternative 3: PO estimation; Outcome: Random Forest; Treatment: Random Forest; IATE: Generalized Random Forest.

7.2. Overlap issues

Although our quasi-experimental design cannot replicate the full control of a randomized experiment, the estimation framework adopted here restricts identification to regions of common support, thereby avoiding extrapolation across fundamentally non-comparable units. The estimated IATEs are thus effectively defined over the subset of the covariate space where treated and control units coexist. As a result, the estimated TEs should be interpreted as local causal effects defined over the empirically supported overlap region. Despite this, overlap may still be weak or incomplete in parts of the covariate space, especially due to the limited number of treated units.

In the present case, although the total sample size is relatively large (1,055 observations), the effective sample for identification is determined by the number of treated units (36 farms) and by the extent of their overlap with the control group. Consequently, statistical power is primarily driven by the availability of comparable treated and control observations rather than by total sample size. Figure A1 (Annex 2) illustrates the distribution of estimated propensity scores across the two treatment groups. While a region of common support exists, overlap is imperfect, implying that identification is inherently local. Although the AIPW estimator improves efficiency through its double-robust structure, it cannot overcome the fundamental information constraints imposed by the limited number of treated units. As a result, the precision of the estimates largely reflects the structure of the data rather than the efficiency of the estimator. This aspect requires an ex-post assessment of how treatment effect estimates and their precision vary across regions with different degrees of overlap. This is implemented through the re-estimation of the model on subsamples defined by increasingly restrictive common support conditions.[17]

Two complementary approaches are considered. The first is sensitivity to overlap thresholds, whereby the estimation is replicated on subsamples defined by progressively tighter propensity score bounds (e.g. [0.01, 0.99], [0.05, 0.95], and [0.10, 0.90]). This approach identifies treatment effects for increasingly comparable subpopulations, trading external validity for stronger identification. The second approach is trimming sensitivity, which evaluates the robustness of the estimates to the exclusion of observations with extreme propensity scores. Unlike the previous approach, trimming is primarily a robustness check on the influence of extreme weights and does not redefine the target population. While both methods reduce the effective sample size, they address distinct aspects of the overlap problem: the former focuses on identification, the latter on estimator stability. Taken together, these analyses provide a comprehensive assessment of how overlap limitations affect both identification and statistical precision, and therefore the overall credibility of the estimated treatment effects.

Table 7 reports the results of both approaches. For reasons of space, only ATE estimates are presented.[18] Overall, the results indicate that the main qualitative conclusions are largely robust to alternative overlap definitions, although both magnitude and precision exhibit some sensitivity as the effective sample size decreases. Starting with the sensitivity to overlap thresholds, the core patterns observed in the baseline estimation broadly persist. In particular, the negative and statistically significant effects for ΔP and ΔGPV are confirmed across specifications, especially when a relatively large common support is retained (e.g. [0.01, 0.99]). As the overlap region becomes more restrictive and the sample size decreases, the magnitude of these effects tends to attenuate and statistical significance weakens. A similar pattern is observed for the ratio-based indicator $\Delta(P/\pi)$, which remains negative and statistically significant across all overlap specifications, although its magnitude decreases under more restrictive support conditions. By contrast, the remaining indicators, $\Delta\pi$, $\Delta(P/GPV)$, and $\Delta(\pi/GPV)$, display substantial variability across specifications and generally lack statistical significance.

The trimming sensitivity analysis leads to consistent conclusions. While trimming reduces the influence of extreme propensity scores and improves comparability, it also reduces the effective sample size, resulting in increased standard errors. Nonetheless, the qualitative patterns remain stable, particularly for ΔP, ΔGPV, and $\Delta(P/\pi)$, confirming that the main findings are not driven by observations with extreme weights or by poor overlap.

Taken together, these results suggest that the environmental effectiveness of organic farming and the associated production trade-offs are robust to alternative overlap definitions and trimming procedures. At the same time, they confirm that identification remains local and that statistical precision is constrained by the limited number of treated observations. This underscores the importance of interpreting the results within the region of reliable common support and of complementing point estimates with uncertainty and sensitivity analyses.

[17] Although the estimation approach here adopted internally relies on a cross-fitted treatment-assignment model, standard overlap diagnostics based on a single propensity score are not directly applicable. We therefore assess overlap using a separately estimated propensity score based on the same covariate set, following standard practice in double machine learning applications.

[18] Respective GATE and IATE estimates are available upon request.

Table 7 – Overlap assessment through trimming sensitivity[a] and sensitivity to overlap thresholds[b]: ATE estimates (estimated standard errors in parentheses).

Outcome variable:	$\Delta\pi$ (€)	ΔP (€)	ΔGPV (€)	$\Delta(P/\pi)$	$\Delta(P/GPV)$	$\Delta(\pi/GPV)$
			Sensitivity to overlap thresholds			
Overlap [0.01, 0.99]	5979	-8941***	-13959***	-0.613***	-0.045	0.328***
	(6326)	(3002)	(3239)	(0.151)	(0.076)	(0.140)
Overlap [0.01, 0.99]	4899	-12947***	-9047*	-0.398**	-0.349	0.223
	(5160)	(3154)	(4791)	(0.197)	(0.345)	(0.327)
Overlap [0.01, 0.99]	2763	-8325	-5973	-0.273***	-0.388	0.315
	(3725)	(1084)	(4276)	(0.097)	(0.505)	(0.458)
			Trimming sensitivity			
Trimming [0.01, 0.99]	1057	-4185***	-14956***	-0.599***	-0.093	0.555**
	(2558)	(1605)	(3651)	(0.150)	(0.179)	(0.262)
Trimming [0.01, 0.99]	1838	-12895***	-9432**	-0.732**	-0.161	0.628**
	(2160)	(2843)	(4552)	(0.299)	(0.318)	(0.310)
Trimming [0.01, 0.99]	4575	-7698*	-5459	-0.895***	-0.146	0.351
	(2889)	(4214)	(4941)	(0.298)	(0.355)	(0.292)

Levels of statistical significance: *** = 1%, ** = 5%, * = 10%.

[a] Common support is imposed by trimming observations with estimated propensity scores outside the intervals [0.01, 0.99], [0.05, 0.95], and [0.10, 0.90].

[b] Common support is imposed by restricting the sample to observations whose estimated propensity scores fall within increasingly narrow intervals. Population becomes narrower and overlap stronger moving from case .01 to .10.

7.3. Endogeneity issues

A further key challenge in approximating a randomized experimental setting within an observational study lies in adequately controlling for all relevant confounders. These can be numerous and often difficult to observe, and in some cases may themselves be affected by treatment, thereby violating the unconfoundedness assumption. In randomized experiments, such risks are mitigated through design, by ensuring random assignment and by measuring confounders prior to treatment exposure. In observational settings, however, these strategies are inherently limited.

As discussed in Section 4, we adopt two complementary approaches to mitigate these concerns. First, we construct a high-dimensional set of covariates by exploiting the richness of the FADN dataset and implement a fully interactive modelling strategy. The use of DML estimation further helps to reduce bias arising from model misspecification and improves inference in high-dimensional settings. Second, we impose a temporal separation between covariates and outcomes by measuring the former using only the early years of the observation period and the latter using the later years. This strategy aims to mitigate simultaneity and reverse causality.

Despite these efforts, endogeneity concerns cannot be entirely ruled out. In particular, since the analysis focuses on always-treated units, all observed characteristics for treated farms are measured after treatment adoption. As a result, the true pre-treatment information set remains unobserved. Moreover, treatment assignment is not random but reflects farmers' voluntary decisions, often influenced by policy incentives. This implies that treated and control units may differ systematically due to self-selection, generating potential selection-on-unobservables bias that cannot be fully eliminated. In our case, there are factors (production specialization and economic size, in particular) that emerge as important sources of heterogeneity. Although these characteristics evolve slowly over time, their potential endogeneity with respect to treatment cannot be excluded.

To further assess the presence of residual endogeneity, we implement two complementary validation exercises. First, we replicate the baseline estimation after excluding from the covariate set the aforementioned variables that can themselves be affected by treatment: production specialization and economic size. This provides a test of the sensitivity of the estimates to potentially endogenous controls. Second, we perform an alternative estimation based on a control function approach (Baldoni and Esposti, 2023).[19] Although this framework does not allow for heterogeneous treatment effects, it offers a direct test for endogeneity by estimating the correlation between the unobserved determinants of treatment assignment and the outcome. The corresponding parameter, ρ, captures residual endogeneity and a statistically significant estimate provides evidence against the validity of the unconfoundedness assumption.

Table 8 reports the results of these two endogeneity checks, comparing ATE estimates obtained with a restricted covariate set to the baseline estimates,[20] and presenting the significance tests of the residual correlation parameter (ρ) from the control function approach. Overall, the results provide mixed but informative evidence regarding the presence of residual endogeneity.

In terms of ATE estimates, the main qualitative patterns remain broadly consistent with the baseline results. In particular, the negative and statistically significant effects for ΔP and ΔGPV are confirmed, with magnitudes that are comparable, if not slightly stronger, under the restricted specification. Similarly, the key ratio-based indicator $\Delta(P/\pi)$ remains negative and statistically significant, suggesting that the environmental–economic trade-off identified in the baseline analysis is robust to the exclusion of potentially endogenous covariates. By contrast, the effect on profitability ($\Delta\pi$) remains statistically insignificant, while the other ratio indicators display some variability in magnitude and significance, consistent with previous sensitivity analyses.

The control function results provide a more direct assessment of residual endogeneity. The null hypothesis that treatment and outcome unobservables are uncorrelated ($\rho = 0$) is strongly rejected for two outcomes, $\Delta\pi$ and $\Delta(\pi/GPV)$, indicating the presence of residual endogeneity in these dimensions. In contrast, for all other outcomes, ΔP, ΔGPV, $\Delta(P/\pi)$ and $\Delta(P/GPV)$, the null cannot be rejected, suggesting that these indicators may be less affected by unobserved confounding.

Taken together, these findings highlight two key points. On the one hand, the stability of the main effects under the restricted specification supports the robustness of the baseline conclusions. On the other hand, endogenous selection remains a relevant concern, especially for economic and production outcomes. This suggests that, while the adopted quasi-experimental design and estimation strategy substantially mitigate endogeneity, they may not fully eliminate it. Consequently, the results should be interpreted with caution, particularly when drawing conclusions on economic performance, whereas the evidence on environmental impacts appears more robust.

Table 8 – Tests of residual endogeneity: ATE estimates using a restricted covariate set (standard errors in parentheses); [a] *significance test of the residual correlation parameter (ρ) from the control function approach.* [b]

Outcome variable:	ATE (restricted covariate set)	Test results $\chi^2(1)$
$\Delta\pi$ (€)	2106 (3507)	49.5***
ΔP (€)	-11227*** (3509)	2.58
ΔGPV (€)	-14842*** (4643)	2.66
$\Delta(P/\pi)$	-0.843*** (0.129)	0.08
$\Delta(P/GPV)$	-0.073 (0.059)	1.89
$\Delta(\pi/GPV)$	0.429* (0.248)	28.7***

Levels of statistical significance: *** = 1%, ** = 5%, * = 10%.

[a] The covariate set is restricted by removing the TF and economic size class variables (see Annex 1)

[b] H0: Treatment and outcome unobservables are uncorrelated, i.e., ρ=0.

[19] The control function approach relies on functional form and distributional assumptions that are not explicitly imposed in our baseline specification. For this reason, it is used here as a sensitivity check rather than as a benchmark estimator.

[20] Respective GATE and IATE estimates are available upon request.

8. Concluding remarks

This paper investigates the multidimensional nature of organic farming sustainability by jointly considering three key dimensions: private economic returns, reductions in environmentally harmful input use, and the maintenance of production capacity. Building on recent evidence based on EU-wide FADN data, which shows that organic adoption leads to significant environmental improvements while economic performance does not decline on average (Brutti and Freo, 2026), this study extends the analysis by incorporating an additional sustainability dimension and explicitly admitting heterogeneity.

The analysis adopts a causal inference framework based on farm-level data, accounting for treatment effect heterogeneity through the estimation of individualized treatment effects (IATE). This approach requires the construction of a quasi-experimental design capable of approximating the conditions of a randomized experiment. The results indicate that "win–win–win" outcomes, simultaneous gains in all dimensions, are rare. While environmental benefits are nearly universal and positive economic outcomes are often observed, these economic gains appear to depend critically on policy support. This suggests that the economic viability of organic farming is, at least in part, sustained by public transfers, raising concerns about its long-term efficiency from a societal perspective. Moreover, organic adoption often entails reductions in production levels, pointing to a potential trade-off between environmental performance and food supply capacity.

From a policy standpoint, the main implication of the study is that, although organic farming represents an important component of a broader sustainability strategy, it cannot be considered a standalone or universally applicable solution. Considering the relevance of this policy conclusion, these findings call for further empirical validation, ideally using more comprehensive datasets (e.g. the full EU FADN sample) and enriched covariate spaces capturing spatial and interaction effects (Baldoni and Esposti, 2023, 2025). However, the most pressing research need is methodological. Despite the use of advanced machine learning methods, some fundamental limitations associated with observational data remain unresolved.

Two issues appear particularly relevant. First, the use of a high-dimensional covariate space may induce weak overlap in parts of the covariate distribution, especially when the number of treated units is limited. Although doubly robust estimation methods mitigate this problem, they cannot eliminate it entirely. As a result, estimated treatment effects are valid only locally and should be interpreted accordingly. This also underscores the importance of systematically assessing robustness across alternative overlap definitions.

Second, endogenous treatment assignment remains a central concern in settings characterized by voluntary adoption. Lag-based identification strategies may not fully address this issue and may be infeasible when the design relies on always-treated and never-treated units. Future research should therefore develop methods that explicitly account for endogeneity, such as instrumental variable approaches, while preserving the flexibility of high-dimensional and nonparametric frameworks.

REFERENCES

Akhuli, S. (2025). Organic Farming: A Pathway to Sustainable Development. *International Journal of Research and Review* 12(1), 208-211.

Akıf, B., Büyüksaatçı-Kiriş, S. (2026). A sustainability and resilience-oriented multi-criteria assessment of agri-food systems across representative European countries *Frontiers in Sustainable Food Systems* 10:1801663. doi: 10.3389/fsufs.2026.1801663.

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

Athey, S., Tibshirani, J., Wager, S. (2019). Generalized random forests. *Annals of Statistics* 47: 1148‑1178. https://doi.org/10.1214/18-AOS1709.

Baldoni, E., Coderoni, S., Esposti, R. (2021). Immigrant workforce and agriculture productivity: Evidence from Italian farm-level data. *European Review of Agricultural Economics* 48(4): 805–834. DOI:10.1093/erae/jbaa033.

Baldoni, E., Esposti, R. (2023). Estimating the impact of policies under spatial interference. The case of CAP support to organic farming. Quaderno di Ricerca n. 475, Dipartimento di Scienze Economiche e Sociali, Università Politecnica delle Marche.

Baldoni, E. Esposti, R. (2025). The impact of environmental policies on adopters under general interference. The case of EU support to organic farming. Quaderno di Ricerca n. 497, Dipartimento di Scienze Economiche e Sociali, Università Politecnica delle Marche.

Bellemare, M.F., Masaki, T., Pepinsky, T.B. (2017). Lagged explanatory variables and the estimation of causal effect. *The Journal of Politics* 79(3): 949-963.

Belloni, A., Chernozhukov, V., Hansen, C. (2014) Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies* 81(2): 608–650.

Biagini, L., Severini, S., & Antonioli, F. (2026). Rethinking organic efficiency: A dynamic analysis of resource trade-offs in Italian agriculture through the FWEE nexus. *Resources, Conservation & Recycling*, 108853. https://doi.org/10.1016/j.resconrec.2026.108853

Brown, C. Kovács, E., Herzon, I., Villamayor-Tomas, S., Albizua, A., Galanaki, A., Grammatikopoulou, I., McCracken, D., Alkan Olsson, J., Zinngrebe, Y. (2021). Simplistic understandings of farmer motivations could undermine the environmental potential of the common agricultural policy. *Land Use Policy* 101, 105136.

Brutti, Z., Freo, M. (2026), Organic Transition in EU Agriculture: An Analysis of Environmental and Economic Outcomes. European Commission, Brussels, JRC145454.

Carattini, S., Dur, R., List, J. (2024). Policy evaluation and the causal analysis of public support. *Science* 386 (6721): 490-492.

Cerulli, G. 2025. Optimal policy learning using Stata. *Stata Journal* 25: 309‑343.

Chernozhukov, V., Hansen, C.B. (2004). The effects of 401(k) participation on the wealth distribution: An instrumental quantile regression analysis. *Review of Economics and Statistics* 86: 735‑751. https://doi.org/10.1162/0034653041811734.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. B. Hansen, W. K. Newey, Robins, J.M. (2018). Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal* 21: C1‑C68. https://doi.org/10.1111/ectj.12097.

Chernozhukov, V., M. Demirer, E. Duflo, Fernández-Val, I. (2006). Generic machine learning inference on heterogeneous treatment effects in randomized experiments, with an application to immunization in India. NBER Working Paper 24678, National Bureau of Economic Research. https://doi.org/10.3386/w24678.

Chotai, N.A., Agastine A. (2024). A Comprehensive Review of Organic Farming's Role in Advancing Sustainable Agriculture: Environmental, Socio-Economic, and Policy Dimensions. *ISBR Management Journal* 9(2): 106–120.

Coderoni, S., Esposti, R., Varacca, A. (2024). How differently do farms respond to agri-environmental policies? A probabilistic machine-learning approach. *Land Economics* 100(2): 370-397. DOI: 10.3368/le.100.2.060622-0043R1.

Coderoni, S., Esposti, R., Varacca, A. (2026). The Dual Impact of Organic Farming Support: Evaluating Greenhouse Gas Emissions Abatement and Private Payoffs. *Business Strategy and the Environment* (in press). DOI: 10.1002/bse.71010.

Dur, R., Non, A., Prottung, P., Ricci, B. (2025). Who's afraid of policy experiments? *The Economic Journal* 135 (666): 538–555.

Esposti, R., 2024. Non-Monetary Motivations of Agri-Environmental Policy Adoption. A Causal Forest Approach. *Journal of Environmental Management* 352, 119992. DOI: 10.1016/j.jenvman.2023.119992.

Esposti, R., 2026. Evaluating Policy Impact under Sparse and Staggered Adoption. A Synthetic Difference-in-Differences application to EU Rural Development Measures. *Evaluation and Program Planning* 116, 102751. https://doi.org/10.1016/j.evalprogplan.2026.102751.

Golub Capital Social Impact Lab. 2023. Machine Learning-based Causal Inference Tutorial. Stanford, CA, https://bookdown.org/stanfordgsbsilab/ml-ci-tutorial/hte-i-binary-treatment.html.

Head, M. L., L. Holman, R. Lanfear, A. T. Kahn, and M. D. Jennions. 2015. The extent and consequences of p-hacking in science. *PLOS Biology* 13: e1002106. https://doi.org/10.1371/journal.pbio.1002106.

Imbens, G.W., Rubin, D.B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Kennedy, E. H. 2023. Towards optimal doubly robust estimation of heterogeneous causal effects. *Electronic Journal of Statistics* 17: 3008‐3049. https://doi.org/10.1214/23-EJS2157.

Knaus, M. C. 2022. Double machine learning‐based programme evaluation under unconfoundedness. *Econometrics Journal* 25: 602‐627. https://doi.org/10.1093/ectj/utac015.

Koch, B., D. M. Vock, and J. Wolfson. 2018. Covariate selection with group lasso and doubly robust estimation of causal effects. *Biometrics* 74: 8‐17. https://doi.org/10.1111/biom.12736.

Kremmydas, D., Beber, C., Baldoni, E., Ciaian, P., Fellmann, T., Gocht, A., Hristov J., Pignotti D., Rey Vicario D., Stepanyan D., Tillie, P. (2024). The EU target for organic farming: Potential economic and environmental impacts of two alternative pathways. *Applied Economic Perspectives and Policy* 47(2): 602-623. DOI: 10.1002/aepp.13470

Mennig, P., Sauer, J. (2022). Promoting organic food production through flagship regions. *Q Open* 2(1), qoac010. https://doi.org/10.1093/qopen/qoac010

Möhring, N., Muller, A., Schaub, S. (2024). Farmers' Adoption of Organic Agriculture—A Systematic Global Literature Review. *European Review of Agricultural Economics* 51(4): 1012–1044.

Nie, X., Wager, S. 2021. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika* 108: 299‐319. https://doi.org/10.1093/biomet/asaa076.

OECD (2008). *Sustainable Development: linking economy, society, environment*. OECD, Paris.

OECD (2019). *Agriculture and Sustainability*. OECD policy framework, OECD, Paris.

Offermann, F., Nieberg, H., Zander, K. (2009). Dependency of organic farms on direct payments in selected EU member states: Today and tomorrow. *Food Policy* 34(3): 273–279. https://doi.org/10.1016/j.foodpol.2009.03.002

Reddy, R.N. (2023). Assessing the Economic and Environmental Impacts of Organic Farming. *International Journal of Agro Studies and Life Sciences* 2(1): 15–21.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 66: 688–701.

Sanders, J., Brinkmann, J., Chmelikova, L., Ebertseder, F., Freibauer, A., Gottwald, F., Haub, A., Hauschild, M., Hoppe, J., Hülsbergen, K., Jung, R., Kusche, D., Levin, K., March, S., Schmidtke, K., Stein-Bachinger, K., Treu, H., Weckenbrock, P., Wiesinger, K., Gattinger, A., Hess, J. (2025). Benefits of organic agriculture for environment and animal welfare in temperate climates. *Organic Agriculture* 15: 213–231.

Semenova, V., Chernozhukov, V. (2021). Debiased machine learning of conditional average treatment effects and other causal functions. *Econometrics Journal* 24: 264‑289. https://doi.org/10.1093/ectj/utaa027.

Smith, O. M., Cohen, A. L., Rieser, C. J., Davis, A. G., Taylor, J. M., Adesanya, A. W., Jones, M. S., Meier, A. R., Reganold, J. P., Orpet, R. J., Northfield, T. D., Crowder, D. W. (2019). Organic farming provides reliable environmental benefits but increases variability in crop yields: A global meta-analysis. *Frontiers in Sustainable Food Systems* 3, 82. https://doi.org/10.3389/fsufs.2019.00082

Stetter, C., Mennig, P., Sauer, J. (2022). Using Machine Learning to Identify Heterogeneous Impacts of Agri-Environment Schemes in the EU: A Case Study. *European Review of Agricultural Economics* 49(4): 723-739.

Stolze, M., Sanders, J., Kasperczyk, N., Madsen, G., Meredith, S., (2016). CAP 2014-2020: Organic farming and the prospects for stimulating public goods. IFOAM EU, Brussels.

Sun, L., Abraham, S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics* 225(2): 175-199.

Tilman, D., Balzer, C., Hill, J., Befort, B.L. (2011). Global food demand and the sustainable intensification of agriculture. *Proceedings of the National Academy of Sciences* (PNAS) 108 (50), 20260-20264. https://doi.org/10.1073/pnas.1116437108.

Twala, B. E. T. H., M. C. Jones, Hand, D.J. 2008. Good methods for coping with missing data in decision trees. *Pattern Recognition Letters* 29: 950‑956. https://doi.org/10.1016/j.patrec.2008.01.010.

Wager, S., Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association* 113: 1228‑1242. https://doi.org/10.1080/01621459.2017.1319839.

World Bank (2025). What is food security? There are four dimensions. In Food Security Update. World Bank. Retrieved August 26, 2025. https://www.worldbank.org/en/topic/agriculture/brief/food-security-update/what-is food-security.

Zander, K., Nieberg, H., & Offermann, F. (2008). Financial relevance of organic farming payments for Western and Eastern European organic farms. *Renewable Agriculture and Food Systems* 23(1): 53–61. https://doi.org/10.1017/S1742170507002050.

ANNEX 1

Table A1 - List of covariates considered in the study grouped in four categories. All the covariates are used in the estimation of the PS and nuisance functions (equation (7) of the main text); only variables highlighted in grey are used in the final estimation of $\hat{\tau}(\boldsymbol{X}_i)$ (equation (10) of the main text).

Variable	*Type*	*Unit of Measure*	*Variable*	*Type*	*Unit of Measure*
Size and economic endowment			**Production Specialization (continues)**		
Variable costs	Numeric	€	Type of farm (TF) (dummy: yes; no):		
Fixed costs	Numeric	€	Cereals	Categorical	-
Gross Production Value (GPV)[a]	Numeric	€	Fruits, Wine growing, Olive growing	Categorical	-
Net value added	Numeric	€	Grazing livestock and Dairy	Categorical	-
Non-organic CAP support	Numeric	€	Granivores	Categorical	-
Utilized Agricultural Area (UAA)	Numeric	Ha	Horticulture	Categorical	-
UAA – Rented	Numeric	%	Mixed crops&livestock	Categorical	-
UAA – Irrigated	Numeric	%	Mixed livestock	Categorical	-
Machinery	Numeric	Kw	Other field crops	Categorical	-
Livestock Units (LSU)	Numeric	Units	**Socio-Demographic Characteristics**		
Worked hours	Numeric	Hours	Age (dummy: young; older)	Categorical	-
Annual Working Units (AWU)	Numeric	Units	Gender (dummy: female; male)	Categorical	-
Number of farm centers	Numeric	Units	Farm labour organization	Categorical	-
Number of land blocks	Numeric	Units	Legal form	Categorical	-
Size class (dummy: yes; no):			Farm strategic profile	Categorical	-
Small	Categorical	-	Farm acquisition mode	Categorical	-
Medium	Categorical	-	**Environmental and geographical factors**		
Large	Categorical	-	Latitude	Numeric	Degrees
Production Specialization			Longitude	Numeric	Degrees
GPV crops	Numeric	%	Disadvantaged Area (dummy: yes; no)	Categorical	-
GPV animal products	Numeric	%	Altitude class	Categorical	-
GPV quality products	Numeric	%	NUTS 1 region (dummy: yes; no)		
GPV transformed products	Numeric	%	North-West Italy	Categorical	-
Other Gainful Activities (OGA)	Numeric	€	North-East Italy	Categorical	-
Forest area	Numeric	Ha	Central Italy	Categorical	-
Diversification (dummy: yes; no)	Categorical	-	Southern Italy	Categorical	-
			Insular Italy	Categorical	-

[a] This variable is excluded when outcome variables ΔGPV, $\Delta(P/GPV)$ and $\Delta(\pi/GPV)$ are considered.

ANNEX 2

Figure A1 – Overlap of propensity score distributions between treated and control units: (a) Histogram; (b) Kernel density estimate.