Causal tree with instrumental variable: an extension of the causal tree framework to irregular assignment mechanisms

Article · May 2019

DOI: 10.1007/s41060-019-00187-z

CITATIONS
0

READS
217

2 authors:

Falco J. Bargagli Stoffi
KU Leuven
3 PUBLICATIONS **4** CITATIONS

SEE PROFILE

Giorgio Gnecco
IMT School for Advanced Studies Lucca
151 PUBLICATIONS **875** CITATIONS

SEE PROFILE

International Journal of Data Science and Analytics
https://doi.org/10.1007/s41060-019-00187-z

REGULAR PAPER

Causal tree with instrumental variable: an extension of the causal tree framework to irregular assignment mechanisms

Falco J. Bargagli Stoffi[1,2] · **Giorgio Gnecco**[1]

Received: 22 January 2019 / Accepted: 7 May 2019

Abstract
This paper provides a link between causal inference and machine learning techniques—specifically, Classification and Regression Trees—in observational studies where the receipt of the treatment is not randomized, but the assignment to the treatment can be assumed to be randomized (irregular assignment mechanism). The paper contributes to the growing applied machine learning literature on causal inference, by proposing a modified version of the Causal Tree (CT) algorithm to draw causal inference from an irregular assignment mechanism. The proposed method is developed by merging the CT approach with the instrumental variable framework to causal inference, hence the name Causal Tree with Instrumental Variable (CT-IV). An improved version, named Honest Causal Tree with Instrumental Variable (HCT-IV), able to estimate more reliably the heterogeneous causal effects, is also proposed. As compared to CT, the main strength of CT-IV and HCT-IV is that they can deal more efficiently with the heterogeneity of causal effects, as demonstrated by a series of numerical results obtained on synthetic data. Then, the proposed algorithms are used to evaluate a public policy implemented by the Tuscan Regional Administration (Italy), which aimed at easing the access to credit for small firms. In this context, HCT-IV breaks fresh ground for target-based policies, identifying interesting heterogeneous causal effects.

Keywords Machine learning · Causal inference · Causal trees · Instrumental variable · Application to social science · Policy evaluation

1 Introduction

Modern statistics is experiencing the growth in the usage of machine learning techniques, such as Classification and Regression Trees (CART) [14], and Random Forests [13], which can be applied to a wide range of statistical problems. In order to use these techniques to answer relevant statistical questions, it is appropriate to highlight some important features of many machine learning methods. These methods are largely about making good predictions and finding the model that fits the data best. Furthermore, their importance lies in the ability to deal with complex datasets, where the number of units is large, as well as the number of features connected with a single unit. In this framework, causality is often de-emphasized. However, in the last decades, the availability of increasingly larger datasets has brought to the attention a new important problem for causal inference, which machine learning techniques can "easily" solve. As a matter of fact, the necessity to deal with problems connected with the heterogeneity of the treatment effects is stronger than in the past: the availability of large datasets makes it possible to customize causal effect estimates for population's subsets and even for individuals. In the past, the analysis subsets for causal inference problems were specified in advance by trials' protocols, while with the new machine learning techniques presented in this paper, the subsets are selected by the algorithms themselves in a data-driven way. Classical approaches to the analysis of heterogeneous effects are non-parametric methods, such as nearest neighbour matching method, kernel method, and series estimation [38]. These techniques usually offer good results in terms of estimation

This paper is an extension version of the DSAA'2018 paper titled "Estimating Heterogeneous Causal Effects in the Presence of Irregular Assignment Mechanisms".

✉ Falco J. Bargagli Stoffi
falco.bargaglistoffi@imtlucca.it

Giorgio Gnecco
giorgio.gnecco@imtlucca.it

[1] IMT School for Advanced Studies, Piazza S. Francesco 19, 55100 Lucca, Italy

[2] KU Leuven, Naamsestraat 69, 3000 Leuven, Belgium

Published online: 21 May 2019

abilities. The drawback is that they perform well as far as the number of covariates is low. Machine learning techniques outperform other non-parametric methods when the number of covariates is relatively high. This can be seen as the reason that led recently to the application of machine learning techniques to causal discovery and inference. A good example of the use of machine learning techniques in these fields, and a very important inspiration for the present work, are [6–8,19,20,37,38], which consider the problem of causality in a supervised learning framework.

In this paper, we focus on causal inference in the presence of an irregular assignment mechanism, where the receipt of the treatment is not randomized, but the assignment to the treatment can be assumed to be randomized. A typical example is provided by an observational study in which a certain number of individuals are randomized to receive a drug, but not all the units that are assigned to receive it are actually treated (i.e. non-compliance in randomized designs). Non-compliance may arise in settings with individuals as units of the analysis, where the receipt of the treatment requires them to take, or subject themselves to, a particular action (e.g. taking a drug, entering a job training program, undergoing a surgery, entering an educational course) [25]. In these cases, one can rely on instrumental variables to estimate causal effects, which are heterogeneous when they depend on the values assumed by one or several covariates. In the work, an adaptation to causal inference of CART in its regression version, named Causal Tree (henceforth, CT) by Athey and Imbens [7,8], is developed to estimate heterogeneous causal effects with instrumental variables in the presence of an irregular assignment mechanism. While the goal of the method proposed in [7] is very similar to the one of the algorithm we develop in this paper (namely, to draw proper causal inference in the presence of irregular assignment mechanisms), the CT algorithm can identify the heterogeneity of causal effects with respect to a particular subset of selected covariates, where the selection needs to be done by the researcher herself. Conversely, our algorithm, named Causal Tree with Instrumental Variable (henceforth, CT-IV), provides a data-driven way to shed light on the heterogeneity of the treatment effects.[1] In the work, we also develop an improved version of CT-IV, named Honest Causal Tree with Instrumental Variable (HCT-IV), which is able to estimate more reliably the heterogeneous causal effects.

The paper is structured as follows. Section 2 provides a background on the causal inference framework, its link with machine learning as it is modelled via the CT algorithm, its "honest version", and basic concepts about instrumental variables. In Sect. 3, we describe our proposed CT-IV and HCT-IV algorithms. In Sect. 4, we provide a comparison on synthetic data between the CT, CT-IV and HCT-IV algorithms in the presence of an irregular assignment mechanism, showing numerically advantages of the proposed methods. Section 5 concludes the paper with a case study on firm level data where the proposed HCT-IV algorithm is used to assess the heterogeneity of the effects of an employment policy implemented by the Tuscan Regional Administration (Italy). The paper is a thoroughly extended version of [9], presented at the 5th IEEE International Conference on Data Science and Advanced Analytics (IEEE DSAA 2018). The completely new parts include Sects. 2C and 3, and two Appendices. Moreover, more extensive simulation results are provided in Sect. 4, whereas the application in Sect. 5B is completely reworked with the new HCT-IV algorithm. Furthermore, a comparison between the results of the CT-IV algorithm and the HCT-IV algorithm is provided.

2 Background

A. Rubin's causal model In order to set up the methods presented in this paper, it is important to remind some notions and notations of Rubin's potential outcome framework [34,35]. Rubin's framework is the milestone of causal inference. Together with the Pearl's causality approach [32], it is the most widely used model in the scientific literature about causal inference.

Given a set of N units, indexed by $i = 1, \ldots, N$, let W_i be the binary indicator of the receipt of the treatment:

$$W_i \in \{0, 1\}. \tag{1}$$

In order to develop a proper causal inference framework, one needs to assume that the potential outcomes for any unit do not vary with the treatments assigned to other units, and that, for each unit, there are no different forms or versions of each treatment level, which may lead to different potential outcomes [34,35]. This assumption is referred in the literature as the Stable Unit Treatment Value Assumption (SUTVA). Given SUTVA, one can postulate the existence of a pair of potential outcomes for each unit:

$$Y_i^{obs} = Y_i(W_i) = \begin{cases} Y_i(0) & \text{if} \quad W_i = 0, \\ Y_i(1) & \text{if} \quad W_i = 1. \end{cases} \tag{2}$$

[1] Algorithms such the ones proposed in [6,19] provide unit-wise estimation of the treatment effect. This feature is a very helpful tool in fields such as personalised medicine. While these techniques furnish unit level results, our proposed method is able to discriminate among bigger sub-populations. This can be useful when dealing with policy issues, since the targeted policies have to be as much general as possible. Indeed, when one needs to deliver a causal analysis to policy makers, one cannot provide unit level results since, in the spirit of the policies, the targeted intervention should be as universal as possible; in many countries, the possibility of targeted policy intervention at unit level (e.g. household, firm, organization) is ruled out to avoid personalized public spending [15,17].

Starting from the notion of potential outcomes, one can define a unit level causal effect as the difference between the potential outcome under treatment and the one under control:

$$\tau_i = Y_i(1) - Y_i(0). \tag{3}$$

The problem of this approach to causal inference is that one can observe just one potential outcome for every unit. It is impossible to observe both potential outcomes for the same unit at the same time. Therefore, from this perspective, causal inference is a *missing data problem* [25].

Is it then impossible to estimate any causal effect? No, it is not but, in order to draw proper causal inference, one needs to introduce the central concepts of the Rubin's Causal Model [25]. Let X_i be the vector of features (usually called also *covariates* or *pre-treatment* variables) associated with the i-th unit, and known not to be affected by the treatment. Let $\mathbf{X}$ be the $N \times K$ matrix of covariates values (where N is the number of units, and K the number of covariates per unit), W the N-dimensional vector of binary assignments to the treatment, and $Y(0)$ and $Y(1)$ the N-dimensional vectors of potential outcomes. Imbens and Rubin [25] define the *assignment mechanism* $P(W|\mathbf{X}, Y(0), Y(1))$, the *unit level assignment probability* $p_i(\mathbf{X}, Y(0), Y(1))$ and the *propensity score* $e(x) = P(W_i = 1|X_i = x)$, which is the probability for a unit to be treated, conditional on its covariates [33].

Following [25], one defines a *classical randomized experiment* as an assignment mechanism that has the following 4 properties:

1. it is *individualistic*, meaning that the treatment assignment for any unit is a function only of its own covariates and potential outcomes;
2. it is *probabilistic*, meaning that the unit level assignment probability belongs to the open interval (0, 1);
3. it is *unconfounded*, meaning that it does not depend on the potential outcomes;
4. it has a functional form that is *known* (and, to some extent, controlled) by the researcher.

Suppose that one is interested in the population average treatment effect:

$$\tau^p = \mathbb{E}[Y_i(1) - Y_i(0)] = \mu(1) - \mu(0), \tag{4}$$

where $\mu(1)$ is the expected value of $Y_i(1)$, and $\mu(0)$ is the expected value of $Y_i(0)$. In the case of a classical randomized experiment, an unbiased estimator of τ^p is:

$$\hat{\tau} = \hat{\mu}(1) - \hat{\mu}(0). \tag{5}$$

In the equation above, $\hat{\mu}(1) = \frac{1}{N_1} \sum_{i:W_i=1} Y_i^{obs}$, where N_1 is the number of units assigned to the treated group, and $\hat{\mu}(0) = \frac{1}{N_0} \sum_{i:W_i=0} Y_i^{obs}$, where N_0 is the number of units assigned to the control group. Finally, we set $\sum_{i\in\{0,1\}} N_i = N$.

By relaxing the fourth property of a known assignment mechanism, one ends up in a scenario that [25] define as a *Regular Assignment Mechanism*. Is it possible in such a scenario to still draw causal inference? The central property that needs to be invoked in order to do so is the unconfoundedness property 3) defined above. Unconfoundedness can be formalized as the conditional independence of the assignment variable W_i to the potential outcomes given (conditioning on) the covariates vector:

$$W_i \perp\!\!\!\perp (Y_i(0), Y_i(1)) \,|X_i. \tag{6}$$

The importance of this assumption is that, conditional on covariates, one can treat observations as they were coming from a randomized experiment. Let the Conditional Average Treatment Effect (CATE) be defined as:

$$\tau(x) = \mathbb{E}[Y_i(1) - Y_i(0)|X_i = x] = \mu(1, x) - \mu(0, x), \tag{7}$$

where $\mu(w, x)$ is the expected value of $Y_i(W_i = w)$ given $X_i = x$. Then it can be proven, by the law of iterated expectations, that:

$$\tau^p = \mathbb{E}[Y_i(1) - Y_i(0)] = \mathbb{E}[\tau(X_i)] = \mathbb{E}[\mu(1, x) - \mu(0, x)]. \tag{8}$$

It follows that τ^p is identified if $\mu(1, x)$ and $\mu(0, x)$ are identified over the support of $\mathbf{X}$. Under unconfoundedness, it can be proven that $\mu(1, x)$ and $\mu(0, x)$ are identified [7]. This gives the possibility to the researcher, if all the important confounding covariates are present in the data, to draw causal inference even when the assignment mechanism is not randomized but is regular. This is the typical case of observational studies, where the researcher does not know beforehand the assignment mechanism (i.e. property 4) above does not hold). Moreover, in observational studies, the assignment to the treatment may be different from the receipt of the treatment. In this scenario, where one allows for non-compliance between the treatment assigned and the treatment received, one can assume that the assignment is itself unconfounded, while the receipt is confounded. Following [25], this assignment mechanism is named *Irregular Assignment Mechanism*. How to draw inference in the presence of an irregular assignment mechanism will be the focus of Sect. 2C, and also the focus of our applied machine learning algorithms in Sect. 3.

Going back to the CATE, there is a variety of reasons for researchers to conduct estimation of $\tau(x)$ (see formula (7)). One is strictly related to the magnitude of the benefits of the treatment which can vary with the features of the individuals.

For instance, one can imagine the extreme case where the average treatment effect of a drug is positive on the overall population (in terms of curing a specific disease), but for a sub-population of patients, with certain characteristics, the average treatment effect is ineffective, or even negative. For these reasons, it is important to find a proper way to estimate causal effects not only on the entire population, but also on specific subsets of the population.

B. Regression trees for causal inference Machine learning offers new ways to investigate heterogeneous effects (i.e. ones that depend on the covariates vector X_i, see (7)), as suggested in [7,38]. Machine learning techniques developed so far in the literature can provide a useful tool to achieve this goal, in scenarios where the assignment mechanism is randomized or is regular.

A machine learning technique that was applied to this task is the CART method [14]. CART is suitable for this goal because, on one side, it is a fully supervised machine learning technique but, on the other side, it is a pretty flexible method that can be adapted to various learning tasks. Here, we provide a general overview of the basic ideas behind such method, referring the reader to [14] for other details about it. The primary goal of CART is to estimate the conditional expectation of an observed outcome on the basis of the information on features and outcomes for units in the training sample, and to compare the resulting estimates on a test sample. Practically, one can estimate these values by building a suitable tree (a classification or a regression tree, depending on the specific problem). The different admissible tree models one can construct entail alternative splits of the tree, based on the values of the features in the data. A possible way to choose the best among various admissible trees is provided by the following procedure, whose initial step consists in dividing the dataset into two different samples:
(a) a first sample, called training sample (or training set), which is used to construct a maximal depth tree, performing the splits using an in-sample goodness-of-fit measure Q^{is}. The size of this training sample is indicated by N^{tr}. Then, the maximal depth tree is pruned, with the aim of maximizing another criterion function Q^{crit}, for various choices of a suitable penalty parameter $\alpha > 0$ on which Q^{crit} depends;
(b) a second sample, called validation sample (or validation set), which is used, for each choice of α, to validate the associated pruned tree, through the use of an out-of-sample goodness of fit Q^{oos}. This second sample size is indicated by N^{va}.

Here, we consider the case in which a single training set and a single validation set are used. In the machine learning literature, this procedure is called the holdout method, and is particular form of cross-validation. In this case, α is chosen by maximizing Q^{oos} with respect to it, and the tree itself is re-trained using the full dataset, for the resulting value of α.

Finally, a different sample, called test sample (or test set), with cardinality N^{te}, is used to assess the performance of the resulting model.

In the following, we describe the Causal Tree (CT) method [7], which is a modification of the original CART method in its regression version, tailored to causal inference. The CT method differs from CART from the following features:

a. the CATE transformation of the outcome;
b. a rework of the in-sample goodness of fit;
c. a rework of the out-of-sample goodness of fit.

2.1 *The CATE transformation*

First of all, one needs to address the big issue of constructing an algorithm that leads to an accurate estimate $\hat{\tau}(x)$ of the conditional average treatment effect. In an ideal world, one would measure the quality of the estimator by looking at the value of the following goodness-of-fit measure, defined in terms of the mean squared error:

$$Q^{infeas}(\hat{\tau}) = -\mathbb{E}[(\tau_i - \hat{\tau}(X_i))^2], \tag{9}$$

However, it is infeasible to estimate the value of Q^{infeas}, because one does not know the values of both potential outcomes for each unit, as τ_i is unobservable. To address this issue, one can transform the observed outcome using the treatment indicator variable W_i and the propensity score $e(X_i)$, as proposed by Athey and Imbens [7]:

$$Y_i^* = Y_i^{obs} \cdot \frac{W_i - e(X_i)}{(1 - e(X_i)) \cdot e(X_i)}. \tag{10}$$

Since Y_i^{obs} is equivalent to $Y_i(W_i)$ then, using (1), one can express (10) as:

$$Y_i^* = Y_i(1) \cdot \frac{W_i}{e(X_i)} - Y_i(0) \cdot \frac{(1 - W_i)}{(1 - e(X_i))}. \tag{11}$$

What is the strength of this transformation? Athey and Imbens prove that, if the unconfoundedness assumption holds, then:

$$\mathbb{E}[Y_i^* | X_i = x] = \tau(x), \tag{12}$$

where Y_i^* in (11) is computed replacing the propensity score $e(X_i)$ with its suitable estimate $\hat{e}(X_i)$ (obtained, e.g. via logistic regression). However, there are some issues in building a tree using a straightforward transformation of the outcome like Y_i^*. In fact, Athey and Imbens argue that the within a leaf sample average of the transformed outcome Y_i^* is not the most efficient estimator of the treatment effect and, moreover, that the proportion of treated and control units within a leaf can be quite different from the overall sample

proportion. An easy way to solve this issue, proposed in [7], is to weigh the CATE transformation in a matter similar to the one developed in [21]. Every partition of the covariates space is identified by a set of leaves, and the treatment effect for the covariates vector X_i belonging to a generic leaf $\mathbb{X}_j$ is estimated as:[2]

$$\hat{\tau}^{CT}(X_i) = \frac{\sum_{l:X_l^{tr}\in\mathbb{X}_j} Y_l^{obs,tr} \cdot \frac{W_l^{tr}}{\hat{e}(X_l^{tr})}}{\sum_{l:X_l^{tr}\in\mathbb{X}_j} \frac{W_l^{tr}}{\hat{e}(X_l^{tr})}} - \frac{\sum_{l:X_l^{tr}\in\mathbb{X}_j} Y_l^{obs,tr} \cdot \frac{(1-W_l^{tr})}{(1-\hat{e}(X_l^{tr}))}}{\sum_{l:X_l^{tr}\in\mathbb{X}_j} \frac{(1-W_l^{tr})}{(1-\hat{e}(X_l^{tr}))}}. \quad (13)$$

2.2 *In-sample goodness of fit*

The second component of the algorithm, which also differs from the corresponding component in the original CART algorithm, is the in-sample goodness of fit. The big issue for defining a proper criterion function for the in-sample goodness of fit is that, in the causal inference framework, the criterion (9), and even its sample approximation $-\frac{1}{N^{tr}}\sum_{i=1}^{N^{tr}}(\tau_i - \hat{\tau}(X_i^{tr}))^2$, which is what would be implemented by a direct application of the original CART algorithm, are infeasible. Hence, [7] proposes to approximate (9) by:

$$Q^{is} = -\frac{1}{N^{tr}}\sum_{i=1}^{N^{tr}} \hat{\tau}^2(X_i^{tr}), \quad (14)$$

and to use the corresponding criterion function

$$Q^{crit} = Q^{is} - \alpha \cdot \kappa, \quad (15)$$

where $\alpha > 0$ is a penalty parameter, and κ is the number of leaves in the tree.

2.3 *Out-of-sample goodness of fit*

For cross-validation, there is no big need for any significant additional computational effort, given the fact that one has already obtained an estimate $\hat{\tau}^{CT}$ of the causal effect defined in terms of the training sample [see (13)], and one just needs to compare it with the causal effect drawn from the validation sample used for the cross-validation. One could easily rework the mean squared error with the transformed outcome Y_i^* to get the Transform-The-Outcome (TOT) loss function:

$$Q^{oos,tot} = -\frac{1}{N^{va}}\sum_{i=1}^{N^{va}}(Y_i^{va,*} - \hat{\tau}(X_i^{va}))^2. \quad (16)$$

The in-sample goodness of fit can be reworked in different ways, following [7]. It looks to us that the TOT-based out-of-sample goodness of fit in (16) fits in a better way in those frameworks in which the number of covariates would lead to very computationally demanding alternative estimators.

2.4 *Causal inference with causal tree*

Due to the specific construction of the Causal Tree, the learning problem reduces to that of estimating the treatment effect in each member of a partition of the covariate space. For the problem of estimating the treatment effect in each leaf of the partition, standard methods are valid. Once one has constructed the tree $\mathbb{T}$, one can consider the leaf that corresponds to the subset $\mathbb{X}_j$ (henceforth, identified with $\mathbb{X}_j$ itself). The tree is defined as a partition of the feature space $\mathbb{X}$, and one can write:

$$\mathbb{T} = \{\mathbb{X}_1, \ldots, \mathbb{X}_{\#(\mathbb{T})}\}, \text{ with } \bigcup_{j=1}^{\#(\mathbb{T})} \mathbb{X}_j = \mathbb{X}, \quad (17)$$

where $\#(\mathbb{T})$ indicates the number of leaves in the tree. Within the leaf $\mathbb{X}_j$, the average treatment effect is:

$$\tau_{\mathbb{T},\mathbb{X}_j} = \mathbb{E}[Y_i(1) - Y_i(0) | X_i \in \mathbb{X}_j], \quad (18)$$

which can be estimated as follows, by subtracting the average outcome

$$\overline{Y}_{\mathbb{X}_j}^{obs,te}(0) = \overline{Y}_{c,\mathbb{X}_j}^{obs,te} \quad (19)$$

on the control units in the leaf $\mathbb{X}_j$ from the average outcome

$$\overline{Y}_{\mathbb{X}_j}^{obs,te}(1) = \overline{Y}_{t,\mathbb{X}_j}^{obs,te} \quad (20)$$

on the treated units in the leaf $\mathbb{X}_j$, both evaluated over the test sample, which is different from the training and validation sample used for the cross-validation:

$$\hat{\tau}_{\mathbb{T},\mathbb{X}_j} = \overline{Y}_{\mathbb{X}_j}^{obs,te}(1) - \overline{Y}_{\mathbb{X}_j}^{obs,te}(0) = \overline{Y}_{t,\mathbb{X}_j}^{obs,te} - \overline{Y}_{c,\mathbb{X}_j}^{obs,te}. \quad (21)$$

One can also estimate, for each leaf $\mathbb{X}_j$, the variance of this estimator using the following Neyman estimator [31]:

$$\hat{\mathbb{V}}_{\mathbb{T},\mathbb{X}_j}^{Neyman} = \frac{s_{t,\mathbb{X}_j}^{te,2}}{N_{t,\mathbb{X}_j}^{te}} + \frac{s_{c,\mathbb{X}_j}^{te,2}}{N_{c,\mathbb{X}_j}^{te}}, \quad (22)$$

[2] Likewise next formulas (37) and (46), (13) can be applied, e.g. also to the validation sample and to the entire (training and validation) sample Ω, replacing the superscript "tr", respectively, with "va" and "Ω".

where $s^{te,2}_{t,\mathbb{X}_j}$ represents the sample variance of the treated group in the test sample, $N^{te}_{t,\mathbb{X}_j}$ its size, $s^{te,2}_{c,\mathbb{X}_j}$ the sample variance of the control group in the test sample, and $N^{te}_{c,\mathbb{X}_j}$ its size. These two sample variances are computed as follows:

$$s^{te,2}_{t,\mathbb{X}_j} = \frac{1}{N^{te}_{t,\mathbb{X}_j} - 1} \sum_{i:W_i=1, X^{te}_i \in \mathbb{X}_j} \left(Y^{te}_i(1) - \overline{Y}^{obs,te}_t\right)^2 , \quad (23)$$

$$s^{te,2}_{c,\mathbb{X}_j} = \frac{1}{N^{te}_{c,\mathbb{X}_j} - 1} \sum_{i:W_i=0, X^{te}_i \in \mathbb{X}_j} \left(Y^{te}_i(0) - \overline{Y}^{obs,te}_c\right)^2 . \quad (24)$$

These estimators of the two variances are unbiased, with respect to the finite-sample distribution of the test sample, if the treatment effect can be assumed to be *additive and constant* within a leaf [25]. However, it can be used to construct confidence intervals only under the normal approximation, which is typically reliable when the number of units inside a leaf is large enough.

C. Honest causal tree One of the most attractive features of Causal Trees is that, by using the CART method for causal inference, one can easily separate the construction of the tree from the treatment effect estimation. In order to do that, one needs to split the available dataset into a training dataset Ω^{tr}, a validation dataset Ω^{va}, an estimation dataset Ω^{est}, and a test dataset Ω^{te}, constructing at first a tree using the training dataset, then cross-validating the complexity of the tree on the validation dataset, estimating the causal effects using the estimation dataset on the selected tree, and eventually assessing the performance of the model on the test dataset. The split between training, validation, estimation, and test datasets is needed in order to draw what Athey and Imbens call *honest* causal inference [8].

The *honest algorithm* modifies the Causal Tree in 2 ways: by using independent samples to build the tree and to estimate the leaf means, and by modifying the splitting and cross-validation criteria to incorporate the unbiased estimates obtained by using the estimation sample.[3] The authors rework the target criterion to be minimized as the expected mean squared error, computed as the expected value of the adjusted-MSE of prediction over the estimation and the test samples:[4]

$$EMSE(\Omega^{te}, \Omega^{est}) = \mathbb{E}[MSE^{adj}(\Omega^{te}, \Omega^{est})]. \quad (25)$$

If one expands (25), following [8], with respect to the CATE in (7), one gets:

$$EMSE^{cate}(\Omega^{te}, \Omega^{est}) = -\mathbb{E}_{X^{te}_i}[\tau^2(X^{te}_i)] + \mathbb{E}_{X^{te}_i, \Omega^{est}}[Var(\hat{\tau}^{est}(X^{te}_i))] . \quad (26)$$

Approximating the expected mean squared error for a Honest Causal Tree, the authors of [8] get:

$$\widehat{EMSE}^{HCT} = -\frac{1}{N^{tr}} \sum_{i \in \Omega^{tr}} (\hat{\tau}^{tr}(X^{tr}_i))^2 + \left(\frac{1}{N^{est}} + \frac{1}{N^{tr}}\right) \sum_{\mathbb{X}_j \in \mathbb{T}} Var\left(\hat{\tau}^{tr}(\mathbb{X}_j)\right), \quad (27)$$

where $\hat{\tau}^{tr}(X^{tr}_i)$ is the CATE estimated on each element of the training set, and $Var(\hat{\tau}^{tr}(\mathbb{X}_j))$ is the within-leaf $(\mathbb{X}_j)$ sampling variance of the estimated CATE.

The training and estimation samples are built on independent data sources, and there is no effect on the training sample tree due to sampling variation in the estimation sample. The main features of the *honest* algorithm applied to the Causal Tree (HCT) is that it eliminates the bias connected to spurious correlations between the covariates and the outcome variable (what the authors of [8] call *dishonest selection*) and gives the possibility to construct valid confidence intervals that do not deteriorate if more covariates are added, or if the data-generation process gets more complex [8]. The price to pay for these benefits is in terms of less deep and less complex trees, which lead to less personalized prediction,[5] and in terms of a higher MSE for the treatment effects.

The central feature of both the CT and HCT algorithms is that conducting causal inference in each leaf of the tree helps one to deal with heterogeneous causal effects: the causal effects computed on different leaves are conditional on different sets of values for the covariates. For instance, the treatment effect $\hat{\tau}^{CT}(X_i)$ estimated over the leaf $\mathbb{X}_l$, where $X_i \in \mathbb{X}_l$, can be substantially different from the one $\hat{\tau}^{CT}(X_j)$ estimated over the leaf $\mathbb{X}_r$, where $X_j \in \mathbb{X}_r$.

In Sect. 3, we describe two algorithms we propose to extend to an instrumental variable scenario first the Causal Tree (CT), then the Honest Causal Tree (HCT).

D. General instrumental variable framework In observational studies, the assignment mechanism may be irregular. For example, dependence on the assignment of the potential outcomes may be not ruled out even after conditioning on a rich set of covariates. These are the cases where the unconfoundedness assumption is violated. In these settings, instrumental variable methods [25] can still help to estimate causal effects. To briefly make the context clear, one can

[3] For further details on the *Honest Causal Tree* algorithm, the reader is referred to [8].

[4] For the definition of the adjusted-MSE and a detailed mathematical derivation of (25), (26) and (27), we refer to "Appendix A".

[5] This problem is not really an issue in the policy-related framework studied in this paper, as explained in detail in the introduction of this paper (see Sect. 1).

consider the following example of an irregular assignment mechanism, for which, in a study population of N units, a certain number of individuals are randomized to receive a treatment (read a drug), but not all the units that are assigned to receive it are actually treated.

Let us denote by W_i the receipt of the treatment, and by Z_i the assignment to the treatment (instrumental variable). Throughout this paper, we will assume both the W_i and Z_i to be binary, even if one could get similar results by relaxing this assumption. In the following, $W_i(Z_i)$ represents the treatment received as a function of the treatment assigned. This leads one to distinguish four different sub-populations G_i of units: those that always comply with the assignment (compliers), those who never comply with the assignment (defiers), those that even if not assigned to the treatment take it (always-takers), and those who do not take the treatment even if assigned to it (never-takers). Formally, one can highlight these sub-populations as follows:

1. Compliers ($G_i = C$): $W_i(Z_i = 0) = 0$ and $W_i(Z_i = 1) = 1$;
2. Defiers ($G_i = D$): $W_i(Z_i = 0) = 1$ and $W_i(Z_i = 1) = 0$;
3. Always-takers ($G_i = AT$): $W_i(Z_i = 0) = 1$, $W_i(Z_i = 1) = 1$;
4. Never-takers ($G_i = NT$): $W_i(Z_i = 0) = 0$, $W_i(Z_i = 1) = 0$.

How can one conduct causal inference in such a setting, if one decides to use the CART method? The first thing to do is to assume the classical Instrumental Variable (IV) assumptions to hold [25]: monotonicity, existence of compliers, unconfoundedness of the instrument, and exclusion restriction. These four assumptions can be written in detail as follows:

1. monotonicity: $W_i(1) \geq W_i(0)$;
2. existence of compliers: $P(W_i(0) < W_i(1)) > 0$;
3. unconfoundedness of the instrument (expressed in terms of conditional independence notation): $Z_i \perp\!\!\!\perp (Y_i(0,0), Y_i(0,1), Y_i(1,0), Y_i(1,1), W_i(0), W_i(1))|X_i$;
4. exclusion restriction: $Y_i(0) = Y_i(1)$ for each element of the sub-population $G_i \in \{AT, NT\}$ where, for each sub-population and $z \in \{0,1\}$, the shortened notation $Y_i(z)$ is used to denote $Y_i(z, W_i(z))$.

The monotonicity assumption leads us to exclude the existence of units that do exactly the opposite of what they are assigned to (read defiers). In the case of one-sided noncompliance, when units that are not assigned to take the drug cannot take it, this assumption is automatically satisfied as $W_i(0) = 0$ for each unit, excluding the presence of defiers and always-takers. In the case of two-sided noncompliance, when treated and control units can access the opposite treatment status, the monotonicity assumption is very plausible but not directly verifiable. The second assumption is the so-called existence of compliers assumption. This assumption states that the sub-population of compliers exists with positive probability. The third assumption states that the instrument is unconfounded. As we saw in Sect. 2, the importance of unconfoundedness is that, conditional on covariates, the assignment to the treatment is as good as if the assignment mechanism was randomized. The last but not least assumption is the exclusion restriction, which rules out any direct effect of Z_i on Y_i. According to this assumption, there is no effect of the assignment on the outcome, in the absence of an effect of the assignment of the treatment on the treatment received, being the treatment of primary interest. Further details on two crucial assumptions for our analysis (exclusion restriction and monotonicity) are provided later, respectively, in footnotes 6 and 7.

Complier Average Causal Effect In the setting considered above, what "one can get from the data" (without invoking any of the previous assumptions) is the Intention To Treat ITT_Y, which is defined as the effect of the intention to treat a unit on the outcome of the same unit (effect of the assignment):

$$ITT_Y = \mathbb{E}[Y_i|Z_i = 1] - \mathbb{E}[Y_i|Z_i = 0]. \tag{28}$$

If one does not assume any of the classical IV assumptions above to hold, then the global ITT_Y may be written as the weighted average of the ITT_Y effects across the four sub-populations of compliers, defiers, always-takers and never-takers:

$$\begin{aligned} ITT_Y = {} & \pi_C ITT_{Y,C} + \pi_D ITT_{Y,D} \\ & + \pi_{NT} ITT_{Y,NT} + \pi_{AT} ITT_{Y,AT}, \end{aligned} \tag{29}$$

where $ITT_{Y,G}$ $(G = C, D, NT, AT)$ is the effect of the treatment assignment on units of type G and π_G is the proportion of units of type G.

We can then proceed by adding step by step the four assumptions. The first assumption that we impose is the exclusion restriction.[6] If it holds, then we get

[6] See Sect. 23.6 in [25] for a discussion about the validity of the exclusion restriction, which is a common assumption in causal inference with instrumental variables. The exclusion restriction states that there is no effect of the assignment to the treatment on the outcome, in the absence of an effect of the assignment to the treatment on the treatment received [25]. This assumption is not directly testable, but there are plenty of studies in economics and social sciences where this assumption is assumed to hold. Examples can be found in fundamental econometric works such as Angrist and Imbens [1], Angrist, Imbens and Rubin [2], Angrist and Pischke [4] and Wooldridge [40]. A famous example is the study by Angrist and Kruger [3] on the effects of different lengths of schooling

$$ITT_{Y,AT} = ITT_{Y,NT} = 0, \tag{30}$$

since, for both always-takers and never-takers, one has

$$Y_i(1) - Y_i(0) = 0. \tag{31}$$

If for an individual the assignment has no effect on the treatment received, then it has also no effect on the outcome. This is a substantial assumption, and is not implied by the design. It is generally stated as the assignment not affecting the outcome other than through the treatment received, as we saw above. Such an assumption can be used to attribute the effect of assignment to the treatment received as follows, taking into account only compliers and defiers:

$$ITT_Y = \pi_C ITT_{Y,C} + \pi_D ITT_{Y,D}. \tag{32}$$

Under monotonicity, we rule out the existence of defiers:[7] $\pi_D = 0$. If we add the unconfoundedness assumption, we can estimate the distribution of compliance types as follows:

a. $\pi_{AT} = P(W_i(0) = W_i(1) = 1) = \mathbb{E}[W_i|Z_i = 0] = \mathbb{E}[W_i(0)]$, estimated as $\hat{\pi}_{AT} = \frac{1}{N_0}\sum_{i=1}^{N}(1 - Z_i)W_i$;
b. $\pi_{NT} = P(W_i(0) = W_i(1) = 0) = 1 - \mathbb{E}[W_i|Z_i = 1] = 1 - \mathbb{E}[W_i(1)]$, estimated as $\hat{\pi}_{NT} = \frac{1}{N_1}\sum_{i=1}^{N} Z_i(1 - W_i)$;
c. $\pi_C = P(W_i(0) = 0, W_i(1) = 1) = \mathbb{E}[W_i|Z_i = 1] - \mathbb{E}[W_i|Z_i = 0]$, estimated as $\hat{\pi}_C = \frac{1}{N_1}\sum_{i=1}^{N} Z_i W_i - \frac{1}{N_0}\sum_{i=1}^{N}(1 - Z_i)W_i$,

where N_1 is the number of units assigned to the treatment and N_0 is the number of units assigned to the control. Once one has estimated the distribution of compliers, when one adds also the "existence of compliers" assumption, one finally gets:

$$ITT_Y = \pi_C ITT_{Y,C}. \tag{33}$$

From this formula, as being $\pi_C \neq 0$, it comes out that $ITT_{Y,C}$, the so-called Complier Average Causal Effect (CACE), is [26]:

$$ITT_{Y,C} = ITT_Y/\pi_C = \frac{\mathbb{E}[Y_i|Z_i = 1] - \mathbb{E}[Y_i|Z_i = 0]}{\mathbb{E}[W_i|Z_i = 1] - \mathbb{E}[W_i|Z_i = 0]}. \tag{34}$$

In general, the global ITT_Y may be viewed as a lower bound on the treatment effect on the compliers: with the assumptions $\pi_D = 0, \pi_C > 0$, and that both $ITT_{Y,NT}$ and $ITT_{Y,AT}$ are strictly less than $ITT_{Y,C}$, one gets $ITT_Y < ITT_{Y,C}$. The complier average treatment effect, $ITT_{Y,C}$, is a local effect, since it makes reference just to the population of compliers, hence it can also be referred as a Local Average Treatment Effect (LATE). It can be estimated as the coefficient associated with the instrumental variable regression [24] as we will see in detail in Sect. 3B. Invoking unconfoundedness, exclusion restriction and monotonicity, one can also infer the outcome distribution for compliers, $\mathbb{E}[Y_i(0)|G_i = C]$, and $\mathbb{E}[Y_i(1)|G_i = C]$. Under the same assumptions, one can estimate the entire marginal distribution of $Y_i(0)$ and $Y_i(1)$ for compliers.

3 Causal tree with instrumental variable

A. Causal tree with randomized instrumental variable In the following, we extend the CT algorithm to the case of an irregular assignment mechanism where the assignment-to-the-treatment variable is itself randomized, but its receipt is not. If we assume the instrumental variable to be randomized, we can draw causal inference from a Causal Tree by making some changes in the structure of the tree. The first difference is that we need to rework the outcome variable, substituting in (10) the indicator variable W_i with the instrumental variable Z_i, as follows:

$$Y_i^{*,IV} = Y_i^{obs} \cdot \frac{Z_i - e(X_i)}{(1 - e(X_i)) \cdot e(X_i)}, \tag{35}$$

where the propensity score is now reworked as $e(x) = P(Z_i = 1|X_i = x)$. In this case, when the assignment mechanism corresponds to a classical randomized experiment, the propensity score is a constant (i.e. $e(x) = p$ for all x), and the transformation above simplifies to:

Footnote 6 continued
time on earnings later in life. The authors used as an instrumental variable, Z_i, the quarter of birth of the students. Indeed, they observed that most States required pupils to enter school in the calendar year in which they turned 6, and that students were required to stay in school until the 16h birthdays. Hence, the length of time in school, which is the treatment variable W_i, was a function of date of birth [4]. By exploiting the fact that there is no direct effect of date of birth on earnings, hence the exclusion restriction holds, they were able to consistently estimate the effect of schooling on earnings later in life. Moreover, another example of scenarios in which the exclusion restriction can be assumed to hold is the case of double-blind assignments [25]. In such settings, since the individuals in the study do not know whether they were assigned to the treatment group or to the control group, there is no effect of the assignment on the outcome, and all the effects on the outcome are mediated by the treatment received.

[7] In many situations, the monotonicity assumption is reasonable [25], because the behaviour of a defier would be in contradiction to its own interest. We refer to [12] for a discussion about this issue. In particular, it is important to highlight that, in many scenarios, defiers are ruled out by not allowing individuals in the control group to have access to the treatment (and vice versa). For instance, this is the case of settings where people that are not assigned to the treatment (i.e. taking a drug, entering a job training program, undergoing a surgery) are excluded by design from the treatment (namely, they cannot possibly get the drug, enter the job program, etc.). In these scenarios of so-called one-sided non-compliance, defiers are ruled out by design. This is also the case of the application that we propose in Sect. 5.

$$Y_i^{*,IV} = \frac{Y_i(1) \cdot Z_i}{p} - \frac{Y_i(0) \cdot (1 - Z_i)}{(1 - p)}. \tag{36}$$

Likewise in (13), one can also use a weighted version of the transformation of the outcome to provide an estimate $\widehat{ITT}_Y(X_i)$ of the intention to treat ITT_Y, for X_i belonging to a generic leaf $\mathbb{X}_j$, as follows:

$$\widehat{ITT}_Y(X_i) = \frac{\sum_{l:X_l^{tr} \in \mathbb{X}_j} Y_l^{obs,tr} \cdot \frac{Z_l^{tr}}{\hat{p}}}{\sum_{l:X_l^{tr} \in \mathbb{X}_j} \frac{Z_l^{tr}}{\hat{p}}} - \frac{\sum_{l:X_l^{tr} \in \mathbb{X}_j} Y_l^{obs,tr} \cdot \frac{(1-Z_l^{tr})}{(1-\hat{p})}}{\sum_{l:X_l^{tr} \in \mathbb{X}_j} \frac{(1-Z_l^{tr})}{(1-\hat{p})}}, \tag{37}$$

where $\hat{p}$ is the estimated value of p. Again, following [21], (37) is an unbiased and efficient estimator of (28) within every leaf.

We also need to rework the in-sample and out-of-sample goodness-of-fit measures:

a. In-sample goodness of fit:

$$Q^{is,IV} = -\frac{1}{N^{tr}} \sum_{i=1}^{N^{tr}} \widehat{ITT}_Y^2(X_i^{tr}); \tag{38}$$

b. Out-of-sample goodness of fit:

$$Q^{oos,IV} = -MSE = -\frac{1}{N^{va}} \sum_{i=1}^{N^{va}} (Y_i^{va,*} - \widehat{ITT}_Y(X_i^{va}))^2. \tag{39}$$

For the sake of clarity, here (and in the following subsections), to fit the instrumental variable framework, we have reworked the out-of-sample goodness-of-fit based on TOT (see (16)). This rework could easily be adapted to other out-of-sample goodness-of-fit measures.

The last part of our algorithm based on the instrumental variable focuses on the estimation of the complier average treatment effects. As we highlighted before, by using the instrumental variable Z_i, we are substantially assuming four different types in our population: compliers, always-takers, never-takers, and defiers. As before, our interest lies on the effect on the compliers. Within every leaf, the complier average causal effect is:

$$\tau_{\mathbb{X}_j}^{cace} = \frac{\mathbb{E}[Y_i | Z_i = 1, X_i \in \mathbb{X}_j] - \mathbb{E}[Y_i | Z_i = 0, X_i \in \mathbb{X}_j]}{P(G_i = C | X_i \in \mathbb{X}_j)} = \frac{ITT_{Y,\mathbb{X}_j}}{\pi_{C,\mathbb{X}_j}}. \tag{40}$$

This formula is analogous to (34) and can be estimated in every leaf assuming the existence of compliers. Then, $\tau_{\mathbb{X}_j}^{cace}$ can be estimated as:

$$\hat{\tau}_{\mathbb{X}_j}^{cace} = \frac{\widehat{ITT}_{Y,\mathbb{X}_j}}{\hat{\pi}_{C,\mathbb{X}_j}}, \tag{41}$$

where $ITT_{Y,\mathbb{X}_j}$ is estimated following (37), and $\pi_{C,\mathbb{X}_j}$ can be estimated as:

$$\hat{\pi}_{C,\mathbb{X}_j} = \frac{1}{N_{1,\mathbb{X}_j}} \sum_{i=1}^{N_{\mathbb{X}_j}} Z_i W_i - \frac{1}{N_{0,\mathbb{X}_j}} \sum_{i=1}^{N_{\mathbb{X}_j}} (1 - Z_i) W_i, \tag{42}$$

where $N_{1,\mathbb{X}_j}$ and $N_{0,\mathbb{X}_j}$ are the numbers of units assigned, respectively, to the treated and control group within a certain leaf $\mathbb{X}_j$, and $N_{\mathbb{X}_j}$ is the number of units within the leaf.

B. Causal tree with unconfounded instrumental variable

Now, we extend the analysis above to the case of an irregular assignment mechanism, where both the assignment and receipt of the treatment are not randomized, but the assignment can be assumed to be unconfounded when conditioning on important covariates. When the instrumental variable is not randomized a priori, the property of unconfoundedness of the instrument does not necessarily hold. If we think of Z_i as our assignment mechanism, then the unconfoundedness of the instrument holds when:

$$Z_i \perp\!\!\!\perp (Y_i(0), Y_i(1)) | X_i. \tag{43}$$

Due to the propensity score properties, this assumption holds even conditioning on the propensity score:

$$Z_i \perp\!\!\!\perp (Y_i(0), Y_i(1)) | e(X_i). \tag{44}$$

In fact, when one conditions on the propensity score, one can assume the instrumental variable to be unconfounded. While in most of the cases, propensity score and instrumental variables have been used alternatively to deal with possible confoundings [11,25], in this paper the propensity score is redefined with regard to the instrumental variable, then the treatment assignment is replaced by the instrumental variable for further estimation. To the extent of our knowledge, this is a novelty of the work.[8] When the assumption (44) holds, one can rework the transformed outcome variable in a similar way as in the previous subsection, obtaining

$$Y_i^* = Y_i(1) \cdot \frac{Z_i}{e(X_i)} - Y_i(0) \cdot \frac{(1 - Z_i)}{(1 - e(X_i))}. \tag{45}$$

[8] Moreover, since the aim of the tree is to create nodes that include units with the highest index of similarity, the proposed procedure can be related to a matching procedure (see [22]).

Assuming that the exclusion restriction and the monotonicity assumptions hold, it is possible to provide an estimate $\widehat{ITT}_Y(X_i)$ of the intention to treat ITT_Y for X_i belonging to a generic leaf $\mathbb{X}_j$, as follows:

$$\widehat{ITT}_Y(X_i) = \frac{\sum_{l:X_l^{tr}\in\mathbb{X}_j} Y_l^{obs,tr}\cdot\frac{Z_l^{tr}}{\hat{e}(X_l^{tr})}}{\sum_{l:X_l^{tr}\in\mathbb{X}_j}\frac{Z_l^{tr}}{\hat{e}(X_l^{tr})}} - \frac{\sum_{l:X_l^{tr}\in\mathbb{X}_j} Y_l^{obs,tr}\cdot\frac{(1-Z_l^{tr})}{(1-\hat{e}(X_l^{tr}))}}{\sum_{l:X_l^{tr}\in\mathbb{X}_j}\frac{(1-Z_l^{tr})}{(1-\hat{e}(X_l^{tr}))}}, \tag{46}$$

where $\hat{e}(X_l^{tr})$ is the estimated value of $e(X_l^{tr})$. The difference between (37) and (46) is that, given the complete randomization of the instrument, in (37) the probability $\hat{e}(X_l^{tr})$ was fixed to $\hat{p}$ for any given unit, while in (46) the assignment-to-the-treatment probability is modelled by the estimated propensity score $\hat{e}(X_l^{tr})$. Finally, the complier average treatment effect in each leaf is still estimated using (41), replacing (37) with (46) to determine the estimate $\widehat{ITT}_{Y,\mathbb{X}_j}$.

3.1 *Overall CACE*

Starting from all the leaves, one can reconstruct the overall effect over all of them as a weighted average of the estimates $\hat{\tau}^{cace}_{\mathbb{X}_j}$ over every leaf $\mathbb{X}_j$. One can represent this weighted average as

$$\hat{\tau}^{cace}_{overall} = \sum_{j=1}^{N^l}\frac{\hat{\tau}^{cace}_{\mathbb{X}_j}\cdot N^{co}_{\mathbb{X}_j}}{N^{co}}, \tag{47}$$

where N^l represents the number of leaves, $N^{co}_{\mathbb{X}_j}$ the number of compliers for every leaf $\mathbb{X}_j$, and N^{co} the overall number of compliers in all the leaves. One can also compute the proportion of compliers in every leaf $\mathbb{X}_j$ simply as:

$$\pi^{co}_{\mathbb{X}_j} = \frac{N^{co}_{\mathbb{X}_j}}{N^{co}}. \tag{48}$$

3.2 *Estimating CACE in every leaf with two-stage least squares regressions*

A suitable possibility to estimate the treatment effect in every leaf is to use, within every leaf $\mathbb{X}_j$ of the tree $\mathbb{T}$, the Two-Stage Least Squares (henceforth, TSLS) method for the estimation of the effect on the complier population, as it is presented in [26]. If one assumes the receipt of the treatment variable W_i and the instrumental variable Z_i to be binary variables, our problem can be expressed in terms of 2 simultaneous regressions:

$$Y_i^{obs} = \alpha + \tau^{CACE}\cdot W_i + \epsilon_i, \tag{49}$$

$$W_i = \pi_0 + \pi_1\cdot Z_i + \eta_i. \tag{50}$$

In the econometric terminology, the explanatory variable W_i is *endogenous*, while the IV variable Z_i is *exogenous*.

The logic of IV regression is that one can estimate the above two reduced form regressions in the case of a single instrument by least squares. In particular, one can estimate τ^{CACE} through TSLS, as the following ratio [25,26]:

$$\hat{\tau}^{CACE} = \hat{\tau}^{IV} = \frac{\hat{\gamma}}{\hat{\pi}_1} = \frac{\widehat{ITT}_Y}{\hat{\pi}_C}, \tag{51}$$

where $\hat{\tau}^{IV}$ is an unbiased estimator of the average causal effect on the population of compliers, and $\hat{\gamma}$ is the estimate of ITT_Y obtained from the reduced form regression

$$Y_i^{obs} = \gamma_0 + \gamma_1\cdot Z_i + \nu_i\,. \tag{52}$$

If one runs a TSLS regression within every leaf $\mathbb{X}_j$ of the tree $\mathbb{T}$, then one is able to obtain an estimate $\hat{\tau}^{cace}_{\mathbb{X}_j}$ for every such leaf.

A possible extension of (50) would be to include in the first-stage regression all the possible confounding variables available in the dataset (in this case, $\pi_2\cdot X_i$ denotes a scalar product):

$$W_i = \pi_0 + \pi_1\cdot Z_i + \pi_2\cdot X_i + \eta_i. \tag{53}$$

The idea is that, if the instrument is unconfounded only conditional on confounding variables, then one could include these covariates in the estimation of the treatment effect on the complier population within each leaf.

In every leaf, using the TSLS method, we can also obtain an estimate of the variance of our $\hat{\tau}^{cace}_{\mathbb{X}_j}$ estimator, which corresponds to the Neyman estimated variance $\hat{\mathbb{V}}^{Neyman}_{\mathbb{T},\mathbb{X}_j}$ for the leaf $\mathbb{X}_j$ of the tree $\mathbb{T}$ (see (22)).

C. The CT-IV algorithm Our proposed CT-IV algorithm is summarized as follows. Compared with the CT algorithm, its main novelties are the replacement of the treatment variable with the instrumental variable, its "redefinition" of the propensity score, and its dropping mechanism for units with either too small or too large estimated propensity scores. These variations are particularly relevant. On the one side, not including the instrumental variable could lead to a biased estimation of the heterogeneous causal effects and, in turn, to potentially harmful targeted policies. This is discussed in Sect. 5 where we compare, in a case study, the application of CT and HCT algorithms and of our novel CT-IV and HCT-IV methodologies. On the other side, not dropping the units with either too small or too large estimated propensity score would provide too much (or too low) weight to units with

extreme values of the propensity score, leading to a scarce precision in the estimation of the causal effects for these units [23].

Causal tree with instrumental variable (CT-IV)

Inputs: N units i $(X_i, Z_i, W_i, Y_i^{obs})$, where X_i is the feature vector, Z_i is treatment assignment (instrumental variable), W_i is the treatment receipt, and Y_i^{obs} is the observed response.

Outputs: (1) a Causal Tree (determined by the use of the instrumental variable), and (2) estimates of the Complier Average Causal Effects within its leaves.

1. First Step of the Algorithm (Building the Tree)
   - Draw a random subsample Ω without replacement and divide it into two disjoint sets: a training set (Ω^{tr}) and a validation set (Ω^{va}) with sizes N^k $(k \in \{tr, va\})$ such that $\sum_k N^k = \#(\Omega)$.
   - Grow a Causal Tree, following the next procedure to take into account the presence of the instrumental variable Z_i:
     (i) estimate the propensity $e(x) = P(Z_i = 1|X_i = x)$ of getting assigned to the treatment;
     (ii) drop units with an estimated propensity score below 0.1 or above 0.9 (in order not to weigh too much units with extreme values of the estimated propensity score);[9]
     (iii) grow a tree by maximizing the following in-sample goodness-of-fit criterion, for several values of $\alpha > 0$:

$$Q^{crit,IV} = -\frac{1}{N^{tr}} \sum_{i=1}^{N^{tr}} \widehat{ITT}_Y^2(X_i^{tr}) - \alpha \cdot \kappa,$$

     where $\widehat{ITT}_Y(X_i^{tr})$ is estimated on the training sample as in (37) in the case of randomization of the instrument or as in (46) if the instrument is not randomized, α is the penalty parameter, and κ is the number of leaves, which measures the complexity of the model;
     (iv) cross-validate the tree, using the following out-of-sample goodness of fit:

$$Q^{oos,IV} = -MSE = -\frac{1}{N^{va}} \sum_{i=1}^{N^{va}} (Y_i^{va,*} - \widehat{ITT}_Y(X_i^{va}))^2$$

     where $\widehat{ITT}_Y(X_i^{va})$ is estimated on the validation sample as in (37) in the case of randomization of the instrument or as in (46) if the instrument is not randomized.

2. Second Step of the Algorithm (Estimating the Complier Average Causal Effects)
   - The complier average causal effect within a leaf $\mathbb{X}_j$ can be estimated on the entire sample Ω in two alternative ways:
     (a) if Z_i is randomized (Sect. 3A) then one can directly estimate the complier average causal effect within every leaf as:

$$\hat{\tau}^{cace}_{\mathbb{X}_j} = \frac{\widehat{ITT}_{Y,\mathbb{X}_j}}{\hat{\pi}_{C,\mathbb{X}_j}}$$

     where $\widehat{ITT}_{Y,\mathbb{X}_j}$ is estimated as in (37) and $\hat{\pi}_{C,\mathbb{X}_j}$ is estimated following (42);
     (b) if Z_i is not randomized but can be assumed to be unconfounded (Sect. 3B) then run a TSLS conditioning on the confounding covariates in the first-stage regression.

D. The HCT-IV algorithm A "honest" extension of the CT-IV algorithm can be obtained, following the HCT algorithm introduced in Sect. 2, by using different data for training the model and for estimating the causal effects. As we discussed previously, this approach allows one to rule out the possibility that the causal effect would be driven by spurious correlations between regressors and outcomes, and allows to construct credible confidence intervals. In this case, which we call Honest Causal Tree with Instrumental Variable (HCT-IV), we divide our data in a training sample Ω^{tr}, a validation sample Ω^{va}, an estimation sample Ω^{est}, and a test sample Ω^{te}. The first sample Ω^{tr} is used to build the tree, the second sample Ω^{va} to cross-validate the tree in order to tune its complexity, the third sample Ω^{est} is used to compute $\widehat{ITT}_Y$ once the tree is built, and the last sample Ω^{te} to assess the performance of the resulting model. As we described in Sect. 2, the target criterion to be minimized, in a honest tree scenario, is the expected adjusted mean squared error $EMSE(\Omega^{te}, \Omega^{est})$ reported in (25). Expanding $EMSE(\Omega^{te}, \Omega^{est})$ with respect to the Intention To Treat, one gets the following:

$$EMSE(\Omega^{te}, \Omega^{est}) = -\mathbb{E}_{X_i^{te}}[ITT_Y^2(X_i^{te})] + \mathbb{E}_{X_i^{te},\Omega^{est}}[Var(\widehat{ITT}_Y^{est}(X_i^{te}))]. \quad (54)$$

Reworking the expected mean squared error for a Honest Causal Tree with Instrumental Variables we get:[10]

[9] We introduce this trimming step following [23]: including values of the covariate such that there is a small overlap between treated and control units leads to a large variance estimator for the average causal effect for these units. The cutoffs are chosen following [16].

[10] See "Appendix A" for further details on the derivation of $EMSE(\Omega^{te}, \Omega^{est})$ and $\widehat{EMSE}^{HCT-IV}$.

$$\widehat{EMSE}^{HCT-IV} = -\frac{1}{N^{tr}} \sum_{i \in \Omega^{tr}} \left(\widehat{ITT}_Y^{tr}(X_i^{tr})\right)^2 + \left(\frac{1}{N^{est}} + \frac{1}{N^{tr}}\right) \sum_{j=1,\ldots,\#(\mathbb{T})} Var\left(\widehat{ITT}_Y^{tr}(\mathbb{X}_j)\right), \quad (55)$$

where $\widehat{ITT}_Y^{tr}(X_i^{tr})$ is the ITT estimated on each element of the training set and $Var(\widehat{ITT}_Y^{tr}(\mathbb{X}_j)$ is the within-leaf $(\mathbb{X}_j)$ variance of the estimated ITT.

Honest causal tree with instrumental variable

Inputs: N units i $(X_i, Z_i, W_i, Y_i^{obs})$, where X_i is the feature vector, Z_i is treatment assignment, W_i is the treatment receipt, and Y_i^{obs} is the observed response.

Outputs: (1) a Honest Causal Tree (determined by the use of the instrumental variable and of the estimation sample), and (2) honest estimates of the Complier Average Causal Effects within its leaves.

1. First Step of the Algorithm (Building the tree)
   - Draw a random sample Ω without replacement and divide it into three disjoint sets: a training sample (Ω^{tr}), a validation sample (Ω^{va}) and an estimation sample (Ω^{est}) of size N^{tr}=N^{est}, with sizes N^k $(k \in \{tr, va, est\})$ such that $\sum_k N^k = \#(\Omega)$.
   - Grow a Honest Causal Tree, following the next procedure to take into account the presence of the instrumental variable Z_i and of the estimation sample:
     (i) estimate the propensity $e(x) = P(Z_i = 1|X_i = x)$ of getting assigned to the treatment;
     (ii) drop units with an estimated propensity score below 0.1 or above 0.9;
     (iii) grow a tree by maximizing the following in-sample goodness-of-fit criterion, for several values of $\alpha > 0$:

$$\widehat{EMSE}^{HCT-IV} = -\frac{1}{N^{tr}} \sum_{i \in \Omega^{tr}} \left(\widehat{ITT}_Y^{tr}(X_i)\right)^2 + \left(\frac{1}{N^{est}} + \frac{1}{N^{tr}}\right) \cdot \sum_{j=1,\ldots,\#(\mathbb{T})} Var\left(\widehat{ITT}_Y^{tr}(\mathbb{X}_j)\right) - \alpha \cdot K,$$

where $\widehat{ITT}_Y(X_i^{tr})$ is estimated on the training sample as in (37) in the case of randomization of the instrument or as in (46) if the instrument is not randomized, α is the penalty parameter, and κ is the number of leaves, which measures the complexity of the model;

(iv) cross-validate the tree, using the following out-of-sample goodness of fit:

$$Q^{oos,IV} = -MSE = -\frac{1}{N^{va}} \sum_{i=1}^{N^{va}} (Y_i^{va,*} - \widehat{ITT}_Y(X_i^{va}))^2$$

where $\widehat{ITT}_Y(X_i^{va})$ is estimated on the validation sample as in (37) in the case of randomization of the instrument or as in (46) if the instrument is not randomized.

2. Second Step of the Algorithm (Estimating the Complier Average Causal Effects using Ω^{est})
   - The complier average causal effect within a leaf $\mathbb{X}_j$ can be estimated, using observations just from the estimation sample, in two alternative ways:
     (a) if Z_i is randomized (Sect. 3), then one can directly estimate the complier average causal effect within every leaf as:

$$\hat{\tau}_{\mathbb{X}_j}^{cace,est} = \frac{\widehat{ITT}_{Y,\mathbb{X}_j}^{est}}{\hat{\pi}_{C,\mathbb{X}_j}^{est}},$$

where $\widehat{ITT}_{Y,\mathbb{X}_j}^{est}$ is estimated as in (37), and $\hat{\pi}_{C,\mathbb{X}_j}^{est}$ is estimated following (42);
     (b) if Z_i is not randomized but can be assumed to be unconfounded (Sect. 3), then run a TSLS conditioning on the confounding covariates in the first-stage regression.

4 Comparison of the CT and HCT-IV algorithms on synthetic data

In this section, we conduct simulations on several synthetic data, to compare the performance of the proposed CT-IV and HCT-IV algorithms with that of CT. As goodness-of-fit measure, we use the opposite of the Mean Squared Error of prediction (MSE) on the test sample, and to assess the relative performance of the two algorithms, we consider the following relative gap measure based on such MSE [39]:

$$Relative\ Gap = \frac{MSE_{CT} - MSE_{(H)CT\text{-}IV}}{MSE_{CT}} \times 100. \quad (56)$$

Moreover, we run some robustness checks. In this section, our focus will be also on what happens in the presence of a weak instrument, namely when the instrument Z_i is weakly correlated with the treatment variable, and when the instrument directly affects the outcome. While the presence of weak instruments is directly testable (typically, with an F-test on the first-stage regression), what is not testable and could be

Table 1 Simulation models

Design	Form of the model	Error (ϵ_i)
1	$Y_i^{obs} = 1 + X_{i1} + W_i \cdot X_{i1} + \eta_i + \epsilon_i$	$\mathcal{N}(0, 1)$
2	$Y_i^{obs} = 1 + \sum_{k=1}^{10} X_{ik} + W_i \cdot X_{i,10} + \eta_i + \epsilon_i$	$\mathcal{N}(0, 1)$
3	$Y_i^{obs} = 1 + \sum_{k=1}^{10} X_{ik} + W_i \cdot X_{i,10} + \eta_i + \epsilon_i$	$Exp(10)$
4	$Y_i^{obs} = 1 + \sum_{k=1}^{10} X_{ik} + W_i \cdot X_{i,10} + \eta_i + \epsilon_i$	$\mathcal{U}(0, 1)$
5	$Y_i^{obs} = 1 + \sum_{k=1}^{10} X_{ik} + W_i \cdot \sum_{k=9}^{10} X_{ik} + \eta_i + \epsilon_i$	$\mathcal{N}(0, 1)$

potentially harmful is a violation of the exclusion restriction at the leaf level. Alternative algorithms, such as the one in [39], take into account the exclusion restriction just at a general level while, in this paper, we take into account that assumption at the leaf level. In a non-synthetic scenario, this assumption is not directly testable, but our algorithms seem to be more "transparent" than other algorithms by taking into account possible violations of this assumption.

A. Synthetic data construction To compare our (H)CT-IV algorithm with the CT one, we first consider some scenarios where the assignment mechanism is irregular. As we saw in Subsection 2.C, this means that the assignment to the treatment is randomized, but the receipt of the treatment is not. The general model that we use for our data simulation is built by considering the following variation of the typical IV setting reported in (49) and (50). The major differences are that we introduce in the main equation (57) a nuisance term η_i and an interaction term between regressors and the treatment indicator, in order to *heterogenise* the treatment effects. The nuisance term η_i can be thought as a not-observable feature that affects both the treatment assignment and the outcome. The general setting looks as follows:

$$Y_i^{obs} = 1 + f(X_i^{out}) + W_i \cdot g(X_i^{tre}) + \eta_i + \epsilon_i, \tag{57}$$

$$W_i = H(Z_i + f(X_i^{out}) - \theta), \tag{58}$$

where $X_i = (x_{i1}, \ldots, x_{iK})$ is a K-dimensional vector of covariates, X_i^{out} highlights those covariates that have an effect on the outcome, and X_i^{tre} (with $\{X_i^{tre}\} \subseteq \{X_i^{out}\}$) those covariates that affect the treatment effect, $H(\cdot)$ is the Heaviside unit step function

$$H(z) = \begin{cases} 0\,, & \text{if } z < 0\,, \\ 1\,, & \text{if } z \geq 0\,, \end{cases} \tag{59}$$

and $\theta \in \mathbb{R}$ is a constant threshold. We consider various functional forms for f and g and for the error distribution in the main equation (57), as well as for f in the first-stage equation (58). The designs investigated (with $K = 1$ for design 1, and $K = 10$ for the other cases) are reported in Table 1.

We train all the five models using incrementally bigger samples, with cardinality ranging from 500 to 50000 (i.e. 500, 1000, 5000, 50000). We implement a hold-one-out cross-validation, bootstrapping the observations assigned to the training set (and validation set) and to the test sample. We let $X_{ik} \sim \mathcal{N}(0, 0.1)$ (considering independent features), $Z_i \sim Bern(0.5)$, and the nuisance parameter η_i be a Gaussian white noise with zero mean and unit variance. Moreover, we set the threshold θ in such a way to get $W_i \sim Bern(0.5)$, and the correlations between W_i and Z_i and W_i and η_i to be, respectively, $Cor(W_i, Z_i) \simeq 0.65$ and $Cor(W_i, \eta_i) \simeq 0.50$. To make the trees comparable, we set the maximal depth of the tree to be 2, and the minimal leaf size to be one tenth of the sample size.

B. Simulation results The results of the simulations are evaluated in Table 2, in terms of the mean squared error of prediction on the test sample. The results are obtained by aggregating the outcomes of 30 different bootstrapped samples. As one can see from the relative gaps reported in Table 2, both the CT-IV and the HCT-IV algorithms outperform the CT one in all the different designs. Comparing the various models by column, one observes that with respect to the baseline case (design 1), the relative gap between the (H)IV-CT and CT algorithms widens as we add covariates (design 2), or change the errors distribution (designs 3 and 4), or change the functional form (design 5). Moreover, it is important to notice that, as the sample size increases, the relative gap widens as well. From the values of the MSE it seems that, while the CT performance is quite stable, (H)CT-IV performance increases as the sample size grows larger. This is especially true in designs 1 and 5. It is important to notice that eventually both CT-IV and HCT-IV reach a coverage rate of 100%. The convergence rate of the two algorithms seems to be fairly the same. However, the relative strength of the HCT-IV algorithm as compared to the CT-IV algorithm is that HCT-IV rules out the possibility that the causal effect would

Table 2 Simulation results

Design	Approach	Sample size			
		500	1000	5000	50,000
1	MSE (CT-IV)	0.128	0.086	0.006	0.001
	MSE (HCT-IV)	0.163	0.098	0.009	0.001
	MSE (Causal Tree)	0.861	0.727	0.650	0.544
	Relative Gap (CT-IV)	85%	88%	99%	100%
	Relative Gap (HCT-IV)	81%	87%	99%	100%
2	MSE (CT-IV)	0.052	0.075	0.031	0.002
	MSE (HCT-IV)	0.042	0.065	0.034	0.002
	MSE (Causal Tree)	0.517	0.722	0.503	0.522
	Relative Gap (CT-IV)	90%	90%	94%	100%
	Relative Gap (HCT-IV)	92%	91%	93%	100%
3	MSE (CT-IV)	0.063	0.072	0.005	0.002
	MSE (HCT-IV)	0.071	0.091	0.002	0.002
	MSE (Causal Tree)	0.824	0.661	0.499	0.518
	Relative Gap (CT-IV)	92%	89%	99%	100%
	Relative Gap (HCT-IV)	91%	86%	100%	100%
4	MSE (CT-IV)	0.019	0.023	0.005	0.001
	MSE (HCT-IV)	0.023	0.027	0.005	0.001
	MSE (Causal Tree)	0.620	0.623	0.501	0.517
	Relative Gap (CT-IV)	97%	96%	99%	100%
	Relative Gap (HCT-IV)	96%	96%	99%	100%
5	MSE (CT-IV)	0.238	0.208	0.054	0.004
	MSE (HCT-IV)	0.250	0.216	0.068	0.005
	MSE (Causal Tree)	5.677	6.655	5.820	8.010
	Relative Gap (CT-IV)	96%	97%	99%	100%
	Relative Gap (HCT-IV)	96%	97%	99%	100%

be driven by spurious correlations between regressors and outcomes.[11]

C. Robustness checks Once we have checked that the (H)CT-IV algorithm outperforms the CT algorithm on synthetic data, it is worth asking what happens when some of the assumptions on which the consistency of the (H)CT-IV is built are partially violated. The main problem that can arise when applying the (H)CT-IV method is a well-known issue in the econometric literature, known as the *weak-instrument* problem [36]. This problem, in our framework, deals with the fact that the number of compliers within every leaf can be particularly small. In an econometric framework, the goal that one would like to achieve is to ensure that $Cor(W_i, Z_i)$ is bounded away from zero. In the following, we test what happens when the instrument is weak on the overall population, and what happens when the exclusion restriction is violated in a specific sub-population. We test these violations on the second model design in Table 1.

In particular, we assume two different scenarios. In the first scenario, we let the instrument be weak on the overall population, by setting $Cor(W_i, Z_i) \simeq 0.5$. In the second scenario, we impose a partial violation of the exclusion restriction, by letting the instrumental variable Z_i directly affect the outcome Y_i when the feature X_{i10} satisfies the condition $X_{i10} < 0$.

In this case, the results from the simulations, reported in Table 3, show that both the CT-IV and HCT-IV algorithms outperform the CT even in the presence of weak instruments. It is important to notice that, within every leaf, the weak-instrument test leads to the rejection of the null hypothesis of weak instrument: our algorithm is able to identify those leaves where there is no weak-instrument problem. Moreover, our algorithm is robust even when the exclusion restriction is partially violated (second scenario). In this case, while the CT algorithm shows a better performance compared with the other scenario, by partially reducing the relative gap, the CT-IV and HCT-IV still perform better in terms of the mean squared error of prediction. Since the estimation of the causal effects is performed in a second stage

[11] The comparative advantage of using the HCT-IV algorithm will be made more clear in the application part where we will compare the results from both algorithms in a case study.

with respect to the building of the tree, our algorithms seem to handle in a good way possible problems due to the violation of exclusion restriction within every leaf. This could not hold true if the exclusion restriction is taken into account just at a general level, as in [39].

5 Case study

A. Programs for the development of crafts in Tuscany (Italy) In the period 2003-2005, the Tuscan Regional Administration (Italy) introduced the "Programs for the Development of Crafts" (PDC). These programs were aimed at Tuscan small-sized handicraft firms, with the goal of promoting innovation and regional development [5,30]. The firms could access PDC by a voluntary application and eligibility criteria. The objective of PDC was to ease access to credit for small-sized firms to boost investments, sales and employment levels. The PCD call guaranteed soft-loans to the firms that were considered eligible for the grant. The eligibility was evaluated on the basis of an investment project. The minimal admissible investment cost was 25,000 Euros, and the grant covered 60% of the financed investment [29]. Among firms participating in the PDC, the large majority of the projects were funded, and the percentage of insolvencies was lower than 3%. Data are available for firms that received the PDC, firms that applied for the funding but were not eligible, and firms that did not apply for the PDC. For our analysis, we use an integrated dataset including information collected by the "Artigian Credito Toscano" and information coming from the archives of the Chamber of Commerce ($2001 - 2004$). The data are available for 266 assisted firms (participating in 2003/05 PDC) and 721 non-assisted firms. The firms in the dataset are operating in 4 economic sectors that comprise the majority of the Tuscan artisan firms: construction; manufacturing activities; wholesale and retail trade; real estate business, rental services, computer, research, business services. The covariates X_{ik} are time-varying covariates, such as sales and employees, and time-invarying covariates, such as location of the firm, year of start-up, legal status, and main distribution channel. The location of each firm is recorded at the provincial level. A central variable for our analysis is the amount of firm's sales in 2002 (pre-treatment year). We created six different sales' groups (up to 50,000; 50,000–100,000; 100,000–250,000; 250,000–500,000; 500,000–1,000,000; greater than 1,000,000). The outcome variable Y_i is a categorical variable that takes the values $1, -1, 0$, respectively, if the number of employees in the firm in the year immediately subsequent to the treatment increased, decreased, or remained the same. The covariate that catches the receipt of the treatment is a dummy variable W_i, which is recorded as 1 if the firm received the financial aid during the two years 2003/2005, and 0 otherwise. However, the treatment variable is not randomized. To draw proper causal inference in this scenario (irregular assignment mechanism), one needs to use an instrumental variable. Luckily, we have data on firms' applications for the funding, which is also represented by a dummy variable, recorded as 1 if the firm applied for the funding, and 0 otherwise. This is a good instrument Z_i, since those firms that applied for the funding were very likely to get it, and it seems that the application itself should not have affected the outcome (exclusion restriction). Moreover, we know that the population of compliers exists (existence of compliers) and, since this is a case of one-sided non-compliance, defiers are ruled out by the design of the policy (monotonicity). However, the instrument is not randomized and, in order to draw proper causal inference in this scenario, one needs to build a propensity score for the instrument itself, in order for the unconfoundedness assumption to hold. This is a very central point and a main novelty of the approach proposed in this paper. Finally, we discard firms with missing values on relevant variables and we exclude from the control sample those observations with extreme values of the propensity score. The selection procedure leads to a sample of 98 assisted firms and 662 non-assisted firms.

B. Application of the (H)CT-IV algorithm to assess heterogeneous causal effects In Fig. 1, results are shown for the HCT-IV algorithm applied on our data. Within every node, the estimated value of the intention to treat and the percentage of observations associated with each node are shown. The name of the variable used for splitting the tree is shown just below each node. The nodes are numbered according to their level in the tree (with the same numbering that one would use for a full binary tree with the same number of levels). The different effects are recorded with different colours: the pinker the node, the closer the effect to zero, while the greener the node, the stronger the effect. No weak-instrument problems occur in the construction of the tree. Table 4 reports the estimated values of the ITT and of the CACE, for each leaf of the constructed tree. For the latter, we report also its standard error, estimated using the TSLS method within every leaf. As one can see, there is a large variation in the CACE among the different leaves. For instance, the observations in the *node 2* (those firms with sales lower than 100,000 Euros) provide a slightly negative estimated causal effect. On the opposite side, those firms within the *node 15* (firms with more than 11 employees, and

Table 3 Robustness checks for violations of IV assumptions

Scenario	Approach	Sample size			
		500	1000	5000	50,000
1	MSE (CT-IV)	0.071	0.043	0.034	0.002
	MSE (HCT-IV)	0.039	0.066	0.041	0.002
	MSE (Causal Tree)	0.771	0.803	0.675	0.662
	Relative Gap (CT-IV)	91%	95%	95%	100%
	Relative Gap (HCT-IV)	95%	92%	94%	100%
2	MSE (CT-IV)	0.163	0.066	0.115	0.074
	MSE (HCT-IV)	0.148	0.055	0.116	0.076
	MSE (Causal Tree)	0.457	0.893	0.707	0.713
	Relative Gap (CT-IV)	64%	93%	84%	90%
	Relative Gap (HCT-IV)	68%	94%	84%	89%

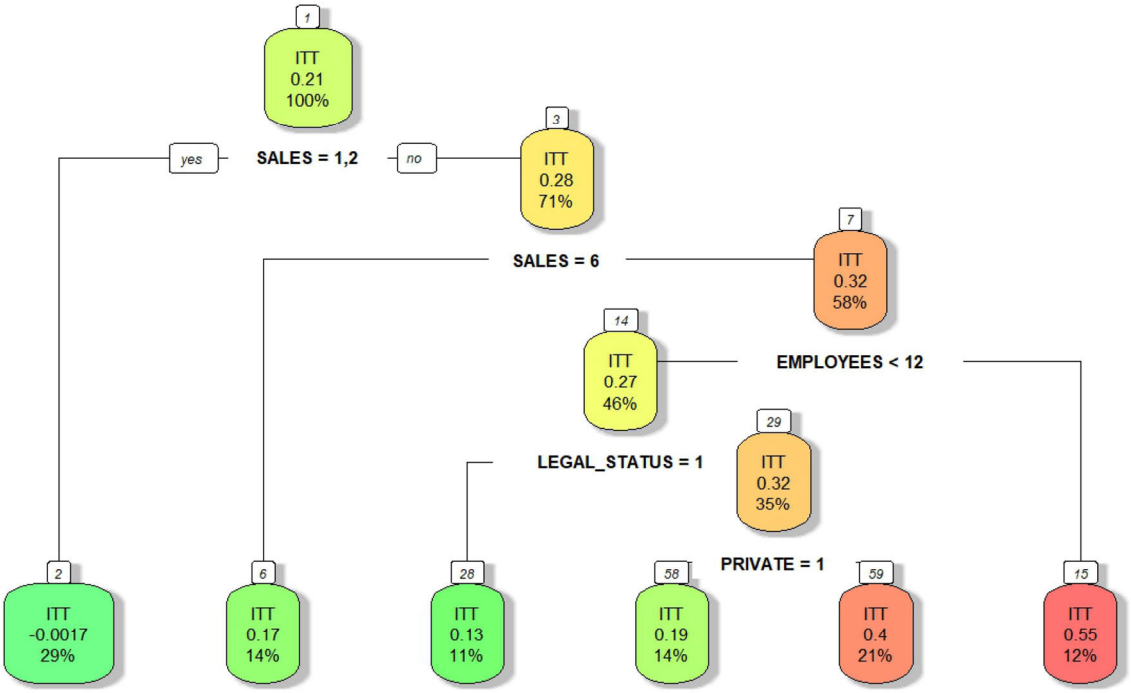

Fig. 1 HCT-IV built on the PDC data (the colors in the figure are reported in the online version of the article)

Table 4 Estimated intention to treat and complier average causal effect on the final nodes (leaves)

Node # j	#1	#2	#3	#6	#7	#14	#15	#28	#29	#58	#59
$\widehat{ITT}_{Y,\mathbb{X}_j}$	0.213	−0.002	0.285	0.166	0.324	0.267	0.551	0.133	0.317	0.192	0.402
$\hat{\tau}^{cace}_{\mathbb{X}_j}$	0.222	−0.002	0.297	0.176	0.337	0.279	0.568	0.139	0.331	0.200	0.420
S.E. $\hat{\tau}^{cace}_{\mathbb{X}_j}$	0.040	0.062	0.049	0.125	0.055	0.063	0.127	0.138	0.076	0.141	0.087

with sales between 100,000 and 1,000,000 Euros) show a 0.55 estimated causal effect in the chance of increasing the number of employees when they get the funding. In our case study, the difference between $\widehat{ITT}^{est}_{Y,\mathbb{X}_j}$ and $\hat{\tau}^{cace,est}_{\mathbb{X}_j}$ is quite small, because the percentage of compliers within every leaf is around 95%, and the overall percentage of compliers is exactly 96%. The overall estimated CACE is positive (0.21) and significant.

In "Appendix B", we depict the tree grown using the CT-IV algorithm (Fig. 2). It is central here to highlight that the variables selected from the algorithm are roughly the same. The CT-IV algorithm however includes also the province

where the firm is active as a variable used to split the population. However, this is a spurious effect due to the correlation between the location of the firm and its outcome. By using the honest version of the CT-IV algorithm, we rule out this spurious causal effect gaining a more meaningful insight in terms of policy evaluation. Moreover, we depict, in "Appendix C", the results that one would obtain by naïvely implementing the CT and HCT algorithms. The differences between CT and HCT and their instrumental variables counterparts are of two types: (i) differences in the estimated causal effects; (ii) differences in the ranking of the variables selected for the split. The results in the root and in leaf 3 are very similar to the results obtained with the CT-IV and HCT-IV algorithms (compare Figs. 1, 2, 3, and 4). However, if we look at the heterogeneous effect estimated in leaf 2 (which includes roughly 29% of firms in the case study) in both CT-IV and HCT-IV as compared to CT and HCT, we find a large difference in the estimated causal effects. Indeed, in the case of CT and HCT the estimated causal effects are positive, while in the case of CT-IV and HCT-IV they are negative. This is due to the fact that the naïve estimator of CT and HCT is largely biased due to the fact that it neglects the irregular assignment mechanism. This would lead to the implementation of potentially harmful targeted policies. Indeed, if policy makers would base their decisions on CT and HCT, they could wrongly target the incentives to firms with sales lower than 100,000 Euros, then they would obtain negative effects in terms of employment. Moreover, these differences in the estimated causal effects lead to differences in the ranking of the splitting variables. The first splitting variable is consistent across all the algorithms, and it is the sales level of the firms. However, both CT and HCT include as a very important splitting variable (i.e. a variable that is used for splitting the first leaves) the legal status of the firm (i.e. sole proprietor, partnership, or capital company). On the contrary, CT-IV does not include this variable, whereas HCT-IV includes it, but in a very low ranking. This highlights how the sources of heterogeneity selected are ranked in different ways according to the different algorithms. Concluding, in this case study, not considering the irregular nature of the assignment mechanism by naïvely implementing CT and HCT would lead to large biases in the estimation of the heterogeneous causal effects and, in turn, to potentially harmful targeted policies.

6 Conclusions

The main aim of this paper is to strengthen the link between machine learning techniques and causal inference, as well as to provide, in this regard, an innovative approach. From this point of view, the CT-IV algorithm developed in this paper has shown to fit in a good way our causal inference goals in the presence of an irregular assignment mechanism. The results obtained from the simulations show that, on one side, CT-IV provides a robust estimation of the overall causal effect on the population under study. On the other side, it outperforms the Causal Tree, providing a very good insight into the heterogeneity of the effects. The CT-IV algorithm has been also combined with the Honest Causal Tree framework [8], to improve the quality of the estimates. Moreover, it could be interesting to extend to the case of an irregular assignment mechanism the ensemble learner proposed in [18], including in that framework our CT-IV and HCT-IV algorithms to fit an IV scenario.

Studying the heterogeneity of causal effects is growing in importance as the size of the dataset, and thus of the population under study, grows; indeed, as shown in the case study, taking into account heterogeneous effects on different subgroups of the study population can help optimizing public interventions and making them more cost-effective. Other possible applications are in fields such as management, health sciences, economics, sociology, and political science.

Concluding, CT-IV and HCT-IV have the following peculiar strengths:

1. they can be applied directly, even when the instrument is not randomized (confounded instrumental variable);
2. they do not directly need a theoretical derivation of the consistency of their estimators because they are based on robust estimators (ITT & TSLS estimators);
3. they provide robust causal effect estimators within every leaf, even when the exclusion restriction is partially violated;
4. they are also robust in settings where the instrument is weak.

On top of these characteristics, HCT-IV provides an additional tool to rule out possible spurious causal effects due to correlations between the regressors and the outcome variable. From this perspective, we suggest to use the HCT-IV algorithm in any case where it is not possible to rule out correlations between the variables available and the output.

One weakness of the proposed algorithms is that they require, of course, the presence of an instrumental variable. However, statistical tests can be still used to check whether the candidate instrumental variable really satisfies (or not) some of the assumptions contained in the definition of an instrumental variable, allowing, e.g. to detect the presence of weak instruments in certain leaves of the tree. However, even when an instrumental variable is not present, previous papers showed that it is possible to draw proper causal inference through two-stages least squares [28] or a synthetic instrument [27]. These techniques could be adopted before running our algorithms to draw inference on heterogeneous causal

effects, even when an instrumental variable is not available from the beginning. Moreover, our proposed algorithms may be extended to the case of multiple candidate instrumental variables, by performing instrumental variable selection as in [10].

In this paper, we focused on the extension of a non-parametric machine learning technique to an irregular assignment mechanism scenario. A further direction of investigation in this field could be the extension to the same framework of ensemble methods for causal inference (e.g. Bayesian Causal Forests [19]).

Acknowledgements Both the authors are members of GNAMPA-INDAM (Gruppo Nazionale per l'Analisi Matematica, la Probabilità e le loro Applicazioni - Istituto Nazionale di Alta Matematica).

Compliance with ethical standards

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

Appendix A: Estimation of the expected mean squared error

Let $ITT_Y(x)$ be the true intention to treat, conditional on a certain set of covariates' values $X_i = x$:[12]

$$ITT_Y(x) = \mathbb{E}\big[Y_i(Z_i = 1) - Y_i(Z_i = 0) | X_i = x\big].$$

The (adjusted) Expected Mean Squared Error (henceforth, $EMSE$[13]) is the expectation over the test sample Ω^{te} and the estimation sample Ω^{est} of the following adjusted Mean Squared Error (MSE^{adj}), whose precise expression is given later in equation (60):

$$EMSE(\Omega^{te}, \Omega^{est}) = \mathbb{E}_{\Omega^{te},\Omega^{est}}\big[MSE^{adj}(\Omega^{te}, \Omega^{est})\big].$$

First, the MSE can be defined as the average over the test sample of the squared error of prediction associated with the conditional estimator obtained on the estimation sample. It is expressed as:

$$MSE(\Omega^{te}, \Omega^{est}) = \frac{1}{\#(\Omega^{te})} \sum_{i \in \Omega^{te}} \big(ITT^{te}_{Y,i} - \widehat{ITT}^{est}_Y(X^{te}_i)\big)^2,$$

where $\#(\Omega^{te})$ is the number of observations in the test sample, $ITT_{Y,i}$, the unit level intention to treat, is

$$ITT_{Y,i} = Y_i(Z_i = 1) - Y_i(Z_i = 0),$$

and $ITT^{te}_{Y,i}$ denotes its value on an element of the test sample. Following [8], we can adjust the MSE by the empirical mean (on the test sample) of $(ITT^{te}_{Y,i})^2$. Since this term does not depend on the choice of the estimator, subtracting it does not affect the way the criterion ranks different estimators [8]. The adjusted version of the MSE is the following:

$$MSE^{adj}(\Omega^{te}, \Omega^{est}) = \frac{1}{\#(\Omega^{te})} \sum_{i \in \Omega^{te}} \Big\{\big(ITT^{te}_{Y,i} - \widehat{ITT}^{est}_Y(X^{te}_i)\big)^2 - (ITT^{te}_{Y,i})^2\Big\}. \tag{60}$$

Nevertheless, the unit level intention to treat $ITT^{te}_{Y,i}$ is infeasible, since one cannot observe for the same unit i, and at the same time, the effects under its assignment to the treatment and under its assignment to the control. However, if one puts aside this problem of infeasibility for a moment, one can expand the $EMSE$, on a partition of a given tree $\mathbb{T}$, as follows:[14]

$$\begin{aligned}
&EMSE(\Omega^{te}, \Omega^{est}, \mathbb{T}) \\
&= \mathbb{E}_{i \in \Omega^{te}, \Omega^{est}}\left[\left(ITT^{te}_{Y,i} - \widehat{ITT}^{est}_Y(X^{te}_i)\right)^2 - (ITT^{te}_{Y,i})^2\right] \\
&= \mathbb{E}_{i \in \Omega^{te}, \Omega^{est}}\Bigg[\Big(ITT^{te}_{Y,i} - ITT_Y(X^{te}_i) \\
&\quad + ITT_Y(X^{te}_i) - \widehat{ITT}^{est}_Y(X^{te}_i)\Big)^2 - (ITT^{te}_{Y,i})^2\Bigg] \\
&= \mathbb{E}_{i \in \Omega^{te}}\left[\left(ITT^{te}_{Y,i} - ITT_Y(X^{te}_i)\right)^2 - (ITT^{te}_{Y,i})^2\right] \\
&\quad + \mathbb{E}_{i \in \Omega^{te}, \Omega^{est}}\left[\left(ITT_Y(X^{te}_i) - \widehat{ITT}^{est}_Y(X^{te}_i)\right)^2\right] \\
&\quad + \mathbb{E}_{i \in \Omega^{te}\Omega^{est}}\Bigg[2\Big(ITT^{te}_{Y,i} - ITT_Y(X^{te}_i)\Big) \\
&\qquad \cdot \Big(ITT_Y(X^{te}_i) - \widehat{ITT}^{est}_Y(X^{te}_i)\Big)\Bigg].
\end{aligned}$$

Since $\mathbb{E}_{i \in \Omega^{te}}\big[\big(ITT^{te}_{Y,i} - ITT_Y(X^{te}_i)\big)\big]$ is zero and the covariance between the two terms $\big(ITT^{te}_{Y,i} - ITT_Y(X^{te}_i)\big)$ and $\big(ITT_Y(X^{te}_i) - \widehat{ITT}^{est}_Y(X^{te}_i)\big)$ is zero,[15] then the term

[12] The following mathematical derivation can be easily extended to the case in which $ITT_Y(x)$ is replaced by $\tau(x)$.

[13] For the seek of brevity of notation, we do not include the superscript "adj" in $EMSE$.

[14] The following expected value depends for its estimation on the tree $\mathbb{T}$. Again, to avoid burdening the terminology, we omit this dependence from the formulas.

[15] This is due to the fact that $ITT^{te}_{Y,i}$ comes from a sample independent of Ω^{est}.

$\mathbb{E}_{i\in\Omega^{te},\Omega^{est}}\left[2\left(ITT^{te}_{Y,i}-ITT_Y(X^{te}_i)\right)\left(ITT_Y(X^{te}_i)-\widehat{ITT}^{est}_Y(X^{te}_i)\right)\right]$ cancels out:[16]

$$
\begin{aligned}
&EMSE(\Omega^{te},\Omega^{est},\mathbb{T})\\
&=\mathbb{E}_{(X^{te}_i,Y^{te}_i)}\left[\left(ITT^{te}_{Y,i}-ITT_Y(X^{te}_i)\right)^2-(ITT^{te}_{Y,i})^2\right]\\
&\quad+\mathbb{E}_{X^{te}_i,\Omega^{est}}\left[\left(ITT_Y(X^{te}_i)-\widehat{ITT}^{est}_Y(X^{te}_i)\right)^2\right]\\
&=\mathbb{E}_{(X^{te}_i,Y^{te}_i)}\Big[(ITT^{te}_{Y,i})^2+ITT^2_Y(X^{te}_i)\\
&\qquad\qquad-2ITT^{te}_{Y,i}ITT_Y(X^{te}_i)-(ITT^{te}_{Y,i})^2\Big]\\
&\quad+\mathbb{E}_{X^{te}_i,\Omega^{est}}\left[\left(ITT_Y(X^{te}_i)-\widehat{ITT}^{est}_Y(X^{te}_i)\right)^2\right]\\
&=\mathbb{E}_{X^{te}_i}\left[ITT_Y(X^{te}_i)^2-2ITT_Y(X^{te}_i)ITT_Y(X^{te}_i)\right]\\
&\quad+\mathbb{E}_{X^{te}_i,\Omega^{est}}\left[\left(ITT_Y(X^{te}_i)-\widehat{ITT}^{est}_Y(X^{te}_i)\right)^2\right],
\end{aligned}
$$

leading to the following:

$$
\begin{aligned}
&EMSE(\Omega^{te},\Omega^{est},\mathbb{T})\\
&=-\mathbb{E}_{X^{te}_i}\left[ITT^2_Y(X^{te}_i)\right]+\mathbb{E}_{X^{te}_i}\left[Var_{\Omega^{est}}\left(\widehat{ITT}^{est}_Y(X^{te}_i)\right)\right],
\end{aligned}
$$

where $Var_{\Omega^{est}}\left(\widehat{ITT}^{est}_Y(X^{te}_i)\right)$ denotes the conditional variance of $\widehat{ITT}^{est}_Y(X^{te}_i)$ given Ω^{est}.

Now it is possible to proceed with the estimation of $EMSE^{HCT-IV}$ for the Honest Causal Tree with Instrumental Variable. For $X^{te}_i\in\mathbb{X}_j$, the conditional variance in the second term of (54) can be approximated by the within-leaf conditional variance estimated on the training sample divided by the number of observations in the leaf (in the estimation sample):

$$
\widehat{Var}_{\Omega^{est}}(\widehat{ITT}^{est}_Y(X^{te}_i))\approx\frac{Var\left(\widehat{ITT}^{tr}_Y(\mathbb{X}_j)\right)}{N^{est}_{\mathbb{X}_j}}.
$$

The expected value can be estimated as:

$$
\begin{aligned}
&\hat{E}_{X^{te}_i}\left[\widehat{Var}_{\Omega^{est}}(\widehat{ITT}^{est}_Y(X^{te}_i))\right]\\
&=\sum_{j\in1,\ldots,\#(\mathbb{T})}\mathcal{P}^{est}_{\mathbb{X}_j}\frac{Var\left(\widehat{ITT}^{tr}_Y(\mathbb{X}_j)\right)}{N^{est}_{\mathbb{X}_j}},
\end{aligned}
$$

where the $\mathcal{P}^{est}_{\mathbb{X}_j}$'s are the leaf shares on the estimation sample. Assuming approximately equal leaf size, we get:

$$
\begin{aligned}
&\hat{E}_{X^{te}_i}\left[\widehat{Var}_{\Omega^{est}}(\widehat{ITT}^{est}_Y(X^{te}_i))\right]\\
&\approx\frac{1}{N^{est}}\sum_{j=1,\ldots,\#(\mathbb{T})}Var\left(\widehat{ITT}^{tr}_Y(\mathbb{X}_j)\right).
\end{aligned}
$$

With respect to the first term in (54), $ITT^2_Y(X^{te}_i)$ can be now approximated using the square of the estimated $\widehat{ITT}^{tr}_Y(X^{te}_i)$ in the training sample minus an estimate of the within-leaf variance of $\widehat{ITT}^{tr}_Y(X^{te}_i)$, obtained by taking into account the number of observations (in the training sample) in the leaf $\mathbb{X}_j$ associated with X^{te}_i:[17]

$$
ITT^2_Y(X^{te}_i)\approx\left(\widehat{ITT}^{tr}_Y(X^{te}_i)\right)^2-\frac{Var\left(\widehat{ITT}^{tr}_Y(\mathbb{X}_j)\right)}{N^{tr}_{\mathbb{X}_j}}.
$$

Assuming again that the leaves are of equal size, the expected value of $ITT^2_Y(X^{te}_i)$ in (54) can be approximated as follows:

$$
\begin{aligned}
&\mathbb{E}_{X^{te}_i}\left[ITT^2_Y(X^{te}_i)\right]\\
&\approx\frac{1}{N^{tr}}\sum_{i\in\Omega^{tr}}\left(\widehat{ITT}^{tr}_Y(X^{tr}_i)\right)^2\\
&-\frac{1}{N^{tr}}\sum_{j=1,\ldots,\#(\mathbb{T})}Var\left(\widehat{ITT}^{tr}_Y(\mathbb{X}_j)\right).
\end{aligned}
$$

Merging the formulas above, we get an estimator of $EMSE^{HCT-IV}(\Omega^{te},\Omega^{est})$ for every partition:

$$
\begin{aligned}
&\widehat{EMSE}^{HCT-IV}\\
&=-\frac{1}{N^{tr}}\sum_{i\in\Omega^{tr}}\left(\widehat{ITT}^{tr}_Y(X^{tr}_i)\right)^2
\end{aligned}
$$

[16] This comes from the fact that one can decompose the covariance between two random variables A and B as $Cov(AB)=\mathbb{E}(AB)-\mathbb{E}(A)\mathbb{E}(B)$. Then, $\mathbb{E}(AB)=Cov(AB)+\mathbb{E}(A)\mathbb{E}(B)$ and, since $Cov(AB)$ and $\mathbb{E}(A)$ are zero, $\mathbb{E}(AB)$ is zero as well.

[17] This is derived from the fact that:

$$
Var\left(\widehat{ITT}_Y(X^{te}_i)\right)=\mathbb{E}\left[\widehat{ITT}^2_Y(X^{te}_i)\right]-\left[\mathbb{E}\left[\widehat{ITT}_Y(X^{te}_i)\right]\right]^2,
$$

whose two members can be approximated as follows:

$$
\frac{Var\left(\widehat{ITT}^{tr}_Y(\mathbb{X}_j)\right)}{N^{tr}_{\mathbb{X}_j}}\approx\left(\widehat{ITT}^{tr}_Y(X^{te}_i)\right)^2-ITT^2_Y(X^{te}_i).
$$

$$+\left(\frac{1}{N^{est}}+\frac{1}{N^{tr}}\right)\sum_{j=1,\ldots,\#(\mathbb{T})} Var\left(\widehat{ITT}_Y^{tr}(\mathbb{X}_j)\right).$$

The first component of (55) is the conventional causal tree criterion, which rewards the partitions with a stronger heterogeneity in the causal effect, while the second component penalizes those partitions that create variance in the leaf causal estimates. This algorithm tends to balance the causal tree tendency to reward heterogeneity in the causal estimates by penalizing imprecise causal estimates within the leaves.

Moreover, one can estimate the terms $\left(ITT_Y^{tr}(X_i^{tr})\right)^2$ and $Var\left(\widehat{ITT}_Y^{tr}(\mathbb{X}_j)\right)$, respectively, as follows:

$$\left(\widehat{ITT}_Y^{tr}(X_i^{tr})\right)^2 \approx \left(\frac{1}{N_{1,\mathbb{X}_i}^{tr}}\sum_{X_j^{tr}\in\mathbb{X}_i, Z_j^{tr}=1} Y_j^{tr} - \frac{1}{N_{0,\mathbb{X}_i}^{tr}}\sum_{X_j^{tr}\in\mathbb{X}_i, Z_j^{tr}=0} Y_j^{tr}\right)^2,$$

$$\widehat{Var}\left(\widehat{ITT}_Y^{tr}(\mathbb{X}_j)\right) \approx \left(\frac{s_{1,\mathbb{X}_j}^2}{N_{1,\mathbb{X}_j}^{tr}}+\frac{s_{0,\mathbb{X}_j}^2}{N_{0,\mathbb{X}_j}^{tr}}\right),$$

where $\mathbb{X}_i$ is the leaf to which X_i^{tr} is assigned by the tree $\mathbb{T}$, $N_{1,\mathbb{X}_i}^{tr}$ is the number of units assigned to treatment within the leaf $\mathbb{X}_i$, $N_{0,\mathbb{X}_i}^{tr}$ is the number of units assigned to control within the leaf $\mathbb{X}_i$, and, for a generic leaf $\mathbb{X}_j$ of the tree $\mathbb{T}$, $s_{1,\mathbb{X}_j}^2$ is the within-leaf variance of ITT_Y for the units assigned to the treatment, and $s_{0,\mathbb{X}_j}^2$ is the within-leaf variance of ITT_Y for the units assigned to the control.[18] Concluding, one can estimate the overall $EMSE^{HCT-IV}$ as follows:

$$\widehat{EMSE}^{HCT-IV} = -\frac{1}{N^{tr}}\sum_{i\in\Omega^{tr}}\left(\frac{1}{N_{1,\mathbb{X}_i}^{tr}}\sum_{X_j^{tr}\in\mathbb{X}_i, Z_j^{tr}=1} Y_j^{tr} - \frac{1}{N_{0,\mathbb{X}_i}^{tr}}\sum_{X_j^{tr}\in\mathbb{X}_i, Z_j^{tr}=0} Y_j^{tr}\right)^2 + \left(\frac{1}{N^{est}}+\frac{1}{N^{tr}}\right)\sum_{j=1,\ldots,\#(\mathbb{T})}\left(\frac{s_{1,\mathbb{X}_j}^2}{N_{1,\mathbb{X}_j}^{tr}}+\frac{s_{0,\mathbb{X}_j}^2}{N_{0,\mathbb{X}_j}^{tr}}\right).$$

In practice, one can use the same sample size for both Ω^{tr} and Ω^{est}, so the estimator above becomes:

$$\widehat{EMSE}^{HCT-IV} = -\frac{1}{N^{tr}}\sum_{i\in\Omega^{tr}}\left(\frac{1}{N_{1,\mathbb{X}_i}^{tr}}\sum_{X_j^{tr}\in\mathbb{X}_i, Z_j^{tr}=1} Y_j^{tr} - \frac{1}{N_{0,\mathbb{X}_i}^{tr}}\sum_{X_j^{tr}\in\mathbb{X}_i, Z_j^{tr}=0} Y_j^{tr}\right)^2 + \frac{2}{N^{tr}}\sum_{j=1,\ldots,\#(\mathbb{T})}\left(\frac{s_{1,\mathbb{X}_j}^2}{N_{1,\mathbb{X}_j}^{tr}}+\frac{s_{0,\mathbb{X}_j}^2}{N_{0,\mathbb{X}_j}^{tr}}\right).$$

Appendix B: Case study with causal tree with IV

Figure 2 depicts the CT-IV built using the data from the case study in Sect. 5.

[18] $s_{0,\mathbb{X}_j}^2$ and $s_{1,\mathbb{X}_j}^2$ can be estimated following (23) and (24) and replacing $Y_i(W_i)$ with $Y_i(Z_i)$.

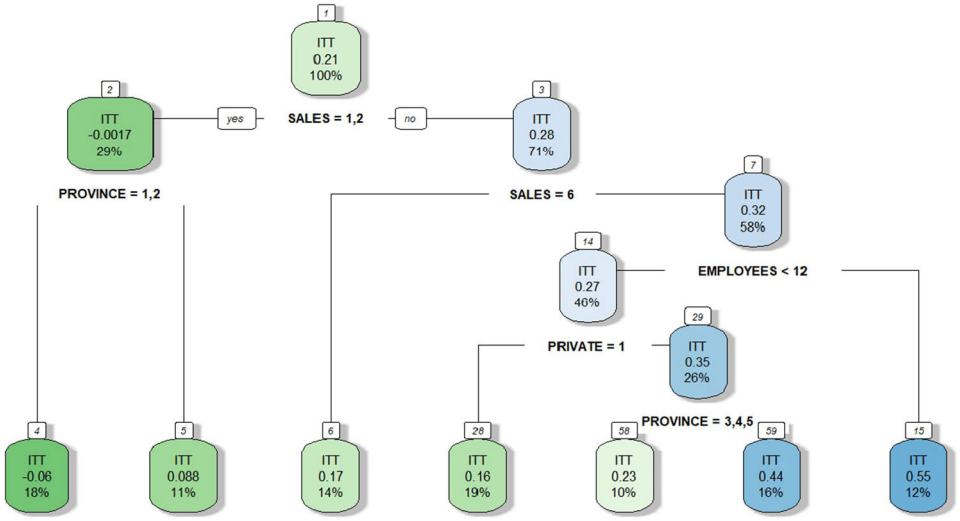

Fig. 2 CT-IV built on the PDC data (the colors in the figure are reported in the online version of the article)

Appendix C: Case study with causal tree and honest causal tree

Figures 3 and 4 depict the CT and the HCT built using the data from the case study in Sect. 5, respectively.

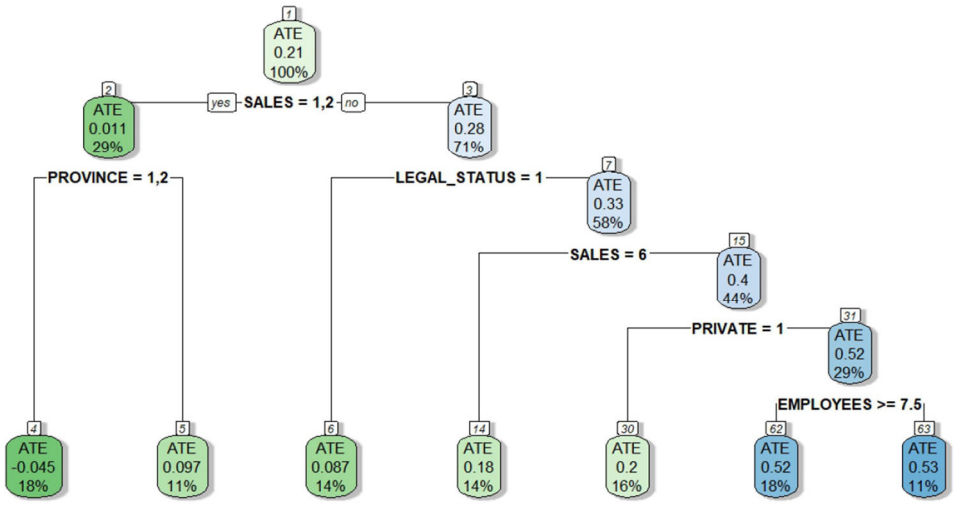

Fig. 3 CT built on the PDC data (the colors in the figure are reported in the online version of the article)

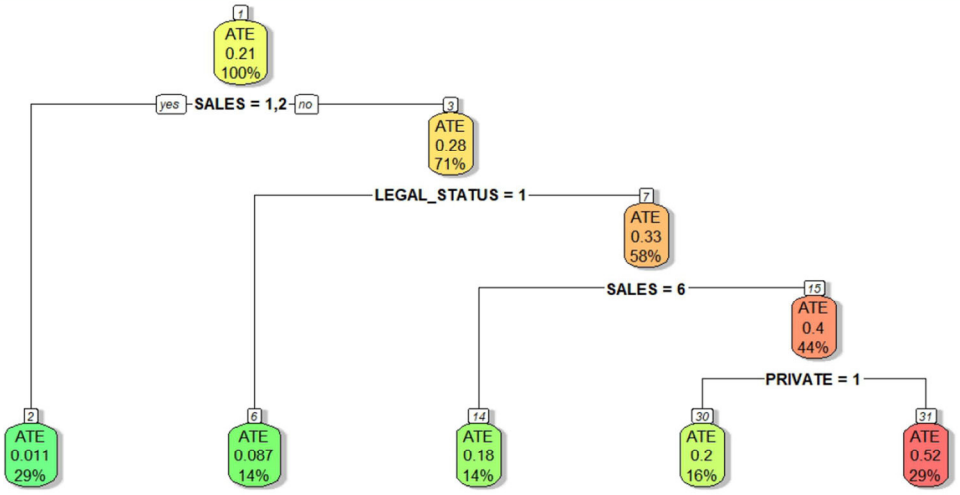

Fig. 4 HCT built on the PDC data (the colors in the figure are reported in the online version of the article)

References

1. Angrist, J.D., Imbens, G.W.: Two stage least squares estimates of average causal response in models with variable treatment intensity. J. Am. Stat. Assoc. **90**(430), 431–442 (1995)
2. Angrist, J.D., Imbens, G.W., Rubin, D.B.: Identification of causal effects using instrumental variables (with discussion). J. Am. Stat. Assoc. **91**(434), 444–472 (1996)
3. Angrist, J.D., Keueger, A.B.: Does compulsory school attendance affect schooling and earnings? Q. J. Econ. **106**(4), 979–1014 (1991)
4. Angrist, J.D., Pischke, J.S.: Mostly Harmless Econometrics: An Empiricist's Companion. Princeton University Press, Princeton (2008)
5. Arpino, B., Mattei, A.: Assessing the causal effects of financial aids to firms in Tuscany allowing for interference. Ann. Appl. Stat. **10**(3), 1170–1194 (2016)
6. Athey, S., Tibshirani, J., Wager, S.: Solving Heterogeneous Estimating Equations with Gradient Forests. arXiv preprint arXiv:1610.01271 (2016)
7. Athey, S., Imbens, G.W.: Machine learning methods for estimating heterogeneous causal effects. Stat **1050**(5), 1–26 (2015)
8. Athey, S., Imbens, G.W.: Recursive partitioning for heterogeneous causal effects. Proc. Natl. Acad. Sci. **113**(27), 7353–7360 (2016)
9. Bargagli Stoffi, F., Gnecco, G.: Estimating heterogeneous causal effects in the presence of irregular assignment mechanisms. In: Proceedings of the 5th IEEE International Conference on Data Science and Advanced Analytics (IEEE DSAA 2018), p. 10, Turin, Italy, October 1st–4th (2018)
10. Belloni, A., Chernozhukov, V., Hansen, C.: High-dimensional methods and inference on structural and treatment effects. J. Econ. Perspect. **28**(2), 29–50 (2014)
11. Bhattacharya, J., Vogt, W.B.: Do instrumental variables belong in propensity scores? NBER Technical Working Paper No. 343 (2009)
12. Bielby, R.M., House, E., Flaster, A., DesJardins, S.L.: Instrumental variables: conceptual issues and an application considering high school course taking. In: Paulsen, M.B. (ed.) Higher Education: Handbook of Theory and Research, vol. 28, pp. 263–321. Springer, Berlin (2013)
13. Breiman, L.: Random Forests. Mach. Learn. **45**(1), 5–32 (2001)
14. Breiman, L., Olshen, J.H., Stone, C.J.: Classification and Regression Trees. CRC Press, Boca Raton (1984)
15. Brown, R., Mawson, S.: Targeted support for high growth firms: theoretical constraints, unintended consequences and future policy challenges. Environ. Plan. C Gov. Policy **34**(5), 816–836 (2016)
16. Crump, R.K., Hotz, V.J., Imbens, G.W., Mitnik, O.A.: Nonparametric tests for treatment effect heterogeneity. Rev. Econ. Stat. **90**(3), 389–405 (2008)
17. Devereux, S.: Is targeting ethical? Glob. Soc. Policy **16**(2), 166–181 (2016)
18. Grimmer, J., Messing, S., Westwood, S.J.: Estimating heterogeneous treatment effects and the effects of heterogeneous treatments with ensemble methods. Polit. Anal. **25**(4), 413–434 (2017)
19. Hahn, P.R., Murray, J.S., Carvalho, C.M.: Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects. arXiv preprint arXiv:1706.09523 (2017)
20. Hill, J.L.: Bayesian nonparametric modeling for causal inference. J. Comput. Gr. Stat. **20**(1), 217–240 (2011)
21. Hirano, K., Imbens, G.W., Ridder, G.: Efficient estimation of average treatment effects using the estimated propensity score. Econometrica **71**(4), 1161–1189 (2003)
22. Ho, D.E., Imai, K., King, G., Stuart, E.A.: Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. Polit. Anal. **15**(3), 199–236 (2007)
23. Imbens, G.W.: Matching methods in practice: three examples. J. Hum. Resour. **50**(2), 373–419 (2015)
24. Imbens, G.W., Angrist, J.D.: Identification and estimation of local average treatment effects. Econometrica **62**(2), 467–475 (1994)
25. Imbens, G.W., Rubin, D.B.: Causal Inference for Statistics, Social, and Biomedical Sciences. An Introduction. Cambridge University Press, Cambridge (2015)
26. Imbens, G.W., Rubin, D.B.: Estimating outcome distributions for compliers in instrumental variables models. Rev. Econ. Stud. **64**(4), 555–574 (1997)
27. Le Gallo, J., Páez, A.: Using synthetic variables in instrumental variable estimation of spatial series models. Environ. Plan. A **45**(9), 2227–2242 (2013)
28. Lewbel, A.: Using heteroscedasticity to identify and estimate mismeasured and endogenous regressor models. J. Bus. Econ. Stat. **30**(1), 67–80 (2012)
29. Mattei, A., Mauro, V.: Valutazione di Politiche per le Imprese Artigiane. Research Report, IRPET - Istituto Regionale Programmazione Economica della Toscana (2007)
30. Mariani, M., Mealli, F.: The effects of R&D subsidies to small and medium-sized enterprises. Evidence from a regional program. Ital. Econ. J. **4**(2), 249–281(2018)

31. Neyman, J.: On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. J. R. Stat. Soc. **97**(4), 558–625 (1934). https://doi.org/10.2307/2342192
32. Pearl, J.: Causality. Cambridge University Press, Cambridge (2009)
33. Rosenbaum, P., Rubin, D.B.: Assessing the sensitivity to an unobserved binary covariate in an observational study with binary outcome. J. R. Stat. Soc. Ser. B **45**(2), 212–218 (1983)
34. Rubin, D.B.: Estimating causal effects of treatments in randomized and nonrandomized studies. J. Educ. Psychol. **66**(5), 688–701 (1974)
35. Rubin, D.B.: Randomization analysis of experimental data: The Fisher randomization test comment. J. Am. Stat. Assoc. **75**(371), 591–593 (1980)
36. Stock, J.H., Yogo, M.: Testing for weak instruments in linear IV regression. In: Andrews, D.W.K. (ed.) Identification and Inference for Econometric Models, pp. 80–108. Cambridge University Press, New York (2002)
37. Su, X., Kang, J., Fan, J., Levine, R.A., Yan, X.: Facilitating score and causal inference trees for large observational studies. J. Mach. Learn. Res. **13**(Oct), 2955–2994 (2012)
38. Wager, S., Athey, S.: Estimation and inference of heterogeneous treatment effects using random forests. J. Am. Stat. Assoc. **113**(523), 1228–1242 (2017)
39. Wang, G., Li, J., Hopp, W.J.: An Instrumental Variable Tree Approach for Detecting Heterogeneous Treatment Effects in Observational Studies. Technical report. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3045327 (2017)
40. Wooldridge, J.M.: Introductory Econometrics: A Modern Approach. Nelson Education, Scarborough (2015)