Heterogeneous causal effects with imperfect compliance: a novel Bayesian machine learning approach[*]

Falco J. Bargagli-Stoffi[†] Kristof De Witte[‡] Giorgio Gnecco[§]

Abstract

This paper introduces an innovative Bayesian machine learning algorithm to draw inference on heterogeneous causal effects in the presence of imperfect compliance (e.g., under an irregular assignment mechanism). We show, through Monte Carlo simulations, that the proposed Bayesian Causal Forest with Instrumental Variable (BCF-IV) algorithm outperforms other machine learning techniques tailored for causal inference (namely, Generalized Random Forest and Causal Trees with Instrumental Variable) in estimating the causal effects. Moreover, we show that it converges to an optimal asymptotic performance in discovering the drivers of heterogeneity in a simulated scenario. BCF-IV sheds a light on the heterogeneity of causal effects in instrumental variable scenarios and, in turn, provides the policy-makers with a relevant tool for targeted policies. Its empirical application evaluates the effects of additional funding on students' performances. The results indicate that BCF-IV could be used to greatly enhance the effectiveness of school funding on students performance.

Keywords: Machine Learning; Bayesian Causal Forest; Honest Causal Trees; School Funding; Students' Performance

JEL Codes: H52; I21; I28

[*] We are grateful to participants at Harvard University, KU Leuven, IMT School for Advanced Studies and Flemish Ministry of Education seminars and at the Atlantic Causal Inference Conference (ACIC 2019). We want to thank Bart Baesens, Giovanna D'Inverno, Francesca Dominici, Avi Feller, Steven Groenez, Michael William Johnson, Alessandra Mattei, Fabrizia Mealli, Mike Smet, Melissa Tuytens, and Paolo Zacchia for the useful comments provided. Kristof De Witte acknowledges funding from Steunpunt SONO.

[†] Corresponding author. IMT School for Advanced Studies, Lucca, Italy and KU Leuven, Leuven, Belgium. Mail to: falco.bargaglistoffi@imtlucca.it. Laboratory for the Analysis of Complex Economic Systems, IMT School for Advanced Studies, piazza San Francesco 19 - 55100 Lucca, Italy. LEER - Leuven Economics of Education Research, Faculty of Economics and Business, KU Leuven, Naamsestraat 69 - 3000 Leuven, Belgium.

[‡] KU Leuven, Leuven, Belgium and Maastricht University, Maastricht, The Netherlands. Mail to: kristof.dewitte@kuleuven.be. LEER - Leuven Economics of Education Research, Faculty of Economics and Business, KU Leuven, Naamsestraat 69 - 3000 Leuven, Belgium. UNU-Merit, Maastricht University, Minderbroedersberg 4 - 6211 LK Maastricht, The Netherlands.

[§] IMT School for Advanced Studies, Lucca, Italy. Mail to: giorgio.gnecco@imtlucca.it. Laboratory for the Analysis of Complex Economic Systems, IMT School for Advanced Studies, piazza San Francesco 19 - 55100 Lucca, Italy.

1 Introduction

In recent years the ability of machines to solve increasingly more intricate tasks has grown exponentially. At the core of this *revolution* (Sejnowski, 2018), there is the staggering predictive power of machine learning algorithms. However, prediction does not imply causation (Lechner, 2019). In social and health sciences the largest part of scientific research questions deals with inferring a causal relationship (e.g., evaluating the impact of a policy, the effects of drug, the returns from a marketing or business strategy and so on). Moreover, following the growing availability of large datasets, the necessity to deal with problems connected with potentially heterogeneous treatment effects is stronger than what was observed in the past. The availability of large datasets makes it possible to investigate and, in turn, customize the causal effect estimates for population subsets as well as for individuals (Athey, 2018). In this scenario, machine learning techniques are increasingly used to address causal inference tasks and, in particular, to estimate heterogeneous causal effects (Foster et al., 2011; Hill, 2011; Su et al., 2012; Green and Kern, 2012; Athey and Imbens, 2016; Hahn et al., 2017; Wager and Athey, 2018; Lee et al., 2018; Lechner, 2019). A growing literature seeks to apply supervised machine learning techniques to the problem of estimating heterogeneous treatment effects. In their seminal contributions, Hill (2011) and Foster et al. (2011) propose to directly apply machine learning algorithms to estimate the unit level causal effect as a function of the units' attributes. In other papers, machine learning algorithms are adapted to estimate the heterogeneous causal effects (Athey and Imbens, 2016; Athey et al., 2016; Hahn et al., 2017; Wager and Athey, 2018; Lechner, 2019). However, most of these techniques are tailored for causal inference in settings where the treatment is randomly assigned to the units and do not address imperfect compliance issues. Nevertheless, in the real world, the implementation of policies or interventions often results in imperfect compliance, which makes the policy evaluation complicated. Imperfect compliance may arise in observational studies where the assignment to the treatment can be different from the receipt of the treatment (e.g., individuals are randomly assigned to a treatment, but not all the units that are assigned to it actually receive it). Recently, some machine learning algorithms have been proposed to deal with imperfect compliance (Athey et al., 2016; Hartford et al., 2016; Wang et al., 2018; Bargagli Stoffi and Gnecco, 2019; Johnson et al., 2019). However, these methods exhibit three principal limitations: (i) random forest-based algorithms for causal inference require large samples to converge to a good asymptotic behaviour for the estimation of causal effects, as shown in Hahn et al. (2018b) and Wendling et al. (2018); (ii) deep learning-based algorithms lack interpretability of the machine learning black-box which can expose them to critiques, especially in the context of social sciences; (iii) the algorithms proposed by Wang et al. (2018) and Bargagli Stoffi and Gnecco (2018, 2019) are based on single learning algorithms that perform worse as compared to multiple learning algorithms (i.e., ensemble methods)[1]. Moreover, in machine learning applications, inference and uncertainty quantification are of secondary importance after predictive performance. However, in policy decision settings it is crucial to know the credibility and variance of the counterfactual predictions (Athey et al., 2016).

To address and accommodate these shortcomings this paper innovates the literature in both a methodological and an empirical perspective. First, we develop a machine learning algorithm tailored to draw causal inference in situations where the assignment mechanism is irregular, namely the assignment depends on the observed and unobserved potential outcomes (Imbens and Rubin, 2015). This methodology contributes to the increasing use of machine learning techniques to draw causal inference (Athey and Imbens, 2017). In particular, we propose to modify a machine learning technique, namely, Bayesian Causal Forests, developed for causal inference goals (Hahn et al., 2017) to fit an instrumental variable setting (Angrist et al., 1996). The proposed method, Bayesian Instrumental Variable Causal Forest (BCF-IV), is an ensemble semi-parametric Bayesian regression model that directly builds on the Bayesian Additive Regression Trees (BART) algorithm (Chipman et al., 2010). BART is an ensemble-of-trees approach to nonparametric regression (Starling et al., 2018a), which is in turn a *refined* version of the random forest algorithm (Breiman, 2001): BART obtains more precise estimates both in non-causal inference scenarios (Chipman et al., 2010; Murray, 2017; Linero and Yang, 2018; Linero, 2018; Hernández et al., 2018; Starling et al., 2018a) and in causal inference settings (Hill, 2011; Hahn et al., 2017, 2018b; Logan et al., 2019) by employing a full set of prior distributions on the depth of the trees, on the noise, and on the outcome in their nodes.

Second, we evaluate the fit of the proposed algorithm by comparing it with two alterna-

[1] Ensemble methods have extensively been shown to outperform single learning algorithms in prediction tasks (Van der Laan et al., 2007).

tive machine learning methods explicitly developed to draw causal inference in the presence of irregular assignment mechanisms: namely, the Generalized Random Forests (GRF) algorithm (Athey et al., 2016) and the Honest Causal Trees with Instrumental Variables (HCT-IV) algorithm (Bargagli Stoffi and Gnecco, 2019). Using Monte Carlo simulations, we evaluate each algorithm with respect to three dimensions: (i) the choice of the correct source of heterogeneity (i.e., the choice of the right splitting variable); (ii) the choice of the correct cutoff, in the case of a continuous splitting variable, and (iii) the estimation of the heterogeneous causal effects. These dimensions are consistent with recent evaluations of various machine learning methods for causal inference that highlight the excellence of Bayesian algorithms for causal inference (Hahn et al., 2018b; Wendling et al., 2018). We show that for each dimension, BCF-IV outperforms both GRF and HCT-IV in small samples and converges to an optimal asymptotic behaviour.

Third, we show how the proposed algorithm can be used for targeted policies, which are increasingly relevant as the call for personalized interventions has unfurled in all social sciences, especially in economics and management sciences (Athey and Imbens, 2017). The main objective of targeted policies studies is to inform policy-makers about the best allocation of treatments to individuals or sub-populations (Kitagawa and Tetenov, 2018). The idea behind these policies is to target those observations that benefit the most from a certain intervention in order to get either of two possible welfare gains: (i) reducing the costs of an intervention with constant effect sizes, or (ii) increasing the intervention effects for given costs (Kleinberg et al., 2017).

Fourth, in an empirical application, BCF-IV is used for the evaluation of an educational policy. The evaluation of educational policies is a promising field for the discovery of heterogeneous causal effects and, in turn, targeted policies. This is due to at least two factors: (i) in the education context, there is a clear source of heterogeneity given by the disparate profiles of schools and students; and (ii) it is possible to gather large (administrative) datasets. In a similar framework, machine learning provides a tailored, data-driven tool for the evaluation of the heterogeneity in the causal effects, and, consequently, the implementation of targeted policies. In particular, we evaluate the effects of additional resources for disadvantaged students on students' performance in a fuzzy Regression Discontinuity Design (RDD) scenario (Hahn et al., 2001). The fuzzy RDD that we implement in this paper is described in detail in Section 4.

By using a unique administrative dataset, we employ BCF-IV to evaluate the heterogeneity in the effects of the 'Equal Educational Opportunities Program' promoted by the Flemish Ministry of Education starting from 2002. The program is aimed at providing additional funding for secondary schools with high share of disadvantaged students (De Witte et al., 2018). We focus on the effects of additional funding on two outcomes: (i) students' performance, namely if a student gets the most favorable outcome (*A-certificate*); and (ii) students' progresses to the following year without retention in their grades. The Flemish Ministry of Education provided us with data on the universe of pupils in the first stage of secondary education in the school year 2010/2011 with a total of 135,682 students. We obtained data on student level and school level characteristics. Moreover, this setting provides us with a quasi-experimental identification strategy since the additional funding is provided to schools based on being above or below an exogenously set threshold regarding the proportion of disadvantaged students. There is also a second, exogenously set, eligibility criterion stating that schools have to generate a minimum number of teaching hours. This provides us with an imperfect compliance setting, as not all the schools fulfill both the criteria, in which we are able to exploit a fuzzy regression discontinuity design to draw causal effects.

The results of our empirical application suggest that, although the effects of additional funding on the overall population of students are found to be not statistically significant[2], there is appreciable heterogeneity in the causal effects: the effects on students' progress are positive and significant if we focus on the sub-population of students in schools with less senior principals (namely, principals with less than 25 years of experience) and younger principals (namely, principals younger than 55 years). These results can advise policy-makers in multiple ways: the heterogeneous drivers could, on the one hand, help them enhancing the policy effectiveness by targeting just the schools with the highest shares of pupils that benefit the most from additional funding. On the other hand, policy-makers could investigate more in depth the reason why some schools do not benefit from the policy and ultimately provide additional tools to these schools to enhance the policy outcomes.

The methodology proposed in this paper can be more widely applied to evaluations of the

[2] This is in line with further research of additional funding on school level outcomes (De Witte et al., 2018).

heterogeneous impact of an intervention in the presence of an irregular assignment mechanism in social and biomedical sciences.

The remainder of this paper is organized as follows: in Section 2 we provide a general overview on causal inference and the applied machine learning frameworks and we introduce our algorithm. In Section 3 we compare the performance of our algorithm with the performance of other methods already established in the literature. In Section 4 we depict the usage of our algorithm in an educational scenario to evaluate the heterogeneous causal effects of additional funding to schools. Section 5 discusses the results and highlights the further applications of heterogeneous causal effects discovery and targeted policies in education as well as in social and biomedical sciences.

2 Bayesian Instrumental Variable Causal Forest

2.1 Notation

This paper contributes to the literature by establishing a novel machine learning approach for the estimation of conditional causal effects in the presence of an irregular assignment mechanism.

We follow the standard notation of the Rubin's causal model (Rubin, 1974, 1978; Imbens and Rubin, 2015). Given a set of N units, indexed by $i = 1, ..., N$, we denote with Y_i a generic outcome variable, with W_i a binary treatment indicator and, with $\mathbf{X}$ a $N \times P$ matrix of P control variables. Given the Stable Unit Treatment Value Assumption (SUTVA), that excludes interference between the treatment assigned to one unit and the potential outcomes of another (Imbens and Rubin, 2015), we can postulate the existence of a pair of potential outcomes: $Y_i(W_i)$. Specifically, the potential outcome for a unit i if assigned to the treatment is $Y_i(W_i = 1) = Y_i(1)$, and the potential outcome if assigned to the control is $Y_i(W_i = 0) = Y_i(0)$. We cannot observe for the same unit both the potential outcomes at the same time. However, we observe the potential outcome that corresponds to the assigned treatment: $Y_i^{obs} = Y_i(1)W_i + Y_i(0)(1 - W_i)$.

In order to draw proper causal inference in observational studies researchers need to assume *strong ignorability* to hold. This assumption states that:

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

and

$$0 < Pr(W_i = 1 | X_i = x) < 1 \ \forall \, x \in \mathbb{X}, \tag{2}$$

where $\mathbb{X}$ is the features space. The first assumption (unconfoundedness) rules out the presence of unmeasured confounders while the second condition needs to be invoked to be able to estimate the unbiased treatment effect on all the support of the covariates space. If these two conditions hold, we are in the presence of the so-called *regular assignment mechanism.* In such a scenario the Average Treatment Effect (ATE) can be expressed as:

$$\tau = \mathbb{E}[Y_i^{obs} | W_i = 1] - \mathbb{E}[Y_i^{obs} | W_i = 0], \tag{3}$$

and one can define, following Athey and Imbens (2016), the Conditional Average Treatment Effect (CATE) simply as:

$$\tau(x) = \mathbb{E}[Y_i^{obs} | W_i = 1, X_i = x] - \mathbb{E}[Y_i^{obs} | W_i = 0, X_i = x]. \tag{4}$$

CATE is central for targeted policies as it enables the researcher to investigate the heterogeneity in causal effects. For instance, we may be interested in assessing how the effects of an intervention vary within different sub-populations.

In observational studies, the assignment to the treatment may be different from the reception of the treatment. In these scenarios, where one allows for non-compliance between the treatment assigned and the treatment received, one can assume that the assignment is unconfounded, wherein the receipt is confounded (Angrist et al., 1996). In such cases, one can rely on an instrumental variable (IV), Z_i, to draw proper causal inference[3]. Z_i can be thought as

[3] Throughout this Section and throughout the paper we assume the instrumental variable to be binary but, one could relax this assumption. However, as there are currently only a few studies that develop machine learning algorithms for the estimation of heterogeneous causal effects with a continuous treatment variable, we leave the application of such algorithms to further research.

a randomized assignment to the treatment, that affects the receipt of the treatment W_i, without directly affecting the outcome Y_i (*exclusion restriction*). Thus, one can then express the treatment received as a function of the treatment assigned: $W_i(Z_i)$.

If the classical four IV assumptions[4] (Angrist et al., 1996) hold, one can get the causal effect of the treatment on the sub-population of compliers, the so-called Complier Average Causal Effect (CACE), that is:

$$\tau^{cace} = ITT_{Y,C} = \frac{\mathbb{E}[Y_i|Z_i=1] - \mathbb{E}[Y_i|Z_i=0]}{\mathbb{E}[W_i|Z_i=1] - \mathbb{E}[W_i|Z_i=0]} = \frac{ITT_Y}{\pi_C}. \tag{5}$$

CACE is also sometimes referred as LATE (Local Average Treatment Effects) and represents the estimate of causal effect of the assignment to treatment on the principal outcome, Y_i, for the subpopulation of compliers (Imbens and Rubin, 2015). In this paper we consider the following conditional version of CATE. The conditional CACE, $\tau^{cace}(x)$, can be thought as the CACE for a sub-population of observations defined by a vector of characteristics x:

$$\tau^{cace}(x) = ITT_{Y,C}(x) = \frac{\mathbb{E}[Y_i|Z_i=1, X_i=x] - \mathbb{E}[Y_i|Z_i=0, X_i=x]}{\mathbb{E}[W_i|Z_i=1, X_i=x] - \mathbb{E}[W_i|Z_i=0, X_i=x]} = \frac{ITT_Y(x)}{\pi_C(x)}. \tag{6}$$

2.2 Estimating Conditional Causal Effects with Machine Learning

In recent years, various algorithms have been proposed to estimate conditional causal effects (i.e, CATE and $\tau^{cace}(x)$). Most algorithms focus on the estimation of CATE (Hill, 2011; Su et al., 2012; Green and Kern, 2012; Athey and Imbens, 2016; Hahn et al., 2017; Wager and Athey, 2018; Lee et al., 2018; Lechner, 2019) while just a few (Athey et al., 2016; Hartford et al., 2016; Wang et al., 2018; Bargagli Stoffi and Gnecco, 2019) focus on the estimation of $\tau^{cace}(x)$. In this paper, we propose an algorithm for the estimation of CATE in an irregular assignment mechanism scenario. In particular, we adapt the Bayesian Causal Forest (BCF) algorithm (Hahn et al., 2017) for such a task. BCF was originally proposed for regular assignment mechanisms. This algorithm builds on the Bayesian Additive Regression Trees (BART) algorithm (Chipman et al., 2010) which in turn is a Bayesian version of an ensemble of Classification and Regression Trees (CART) (Breiman et al., 1984)[5].

[4] See Appendix A for a detailed discussion of the four assumptions and how they are assumed to hold in our application reported in Section 4.

[5] Chipman et al. (2010) highlight how their algorithm is different from other ensemble methods such as the Random Forest algorithm (Breiman, 2001).

CART is a widely used algorithm for the construction of binary trees (namely, trees where each node is splitted into only two branches). A binary tree is constructed by splitting a node into two child nodes repeatedly, beginning with the root node that contains the whole learning sample and proceeding with the splits to the final nodes (leaves). Figure 1 illustrates how the binary partitioning works in practice in a simple case with just two regressors $x_1 \in [0, 1]$ and $x_2 \in [0, 1]$.

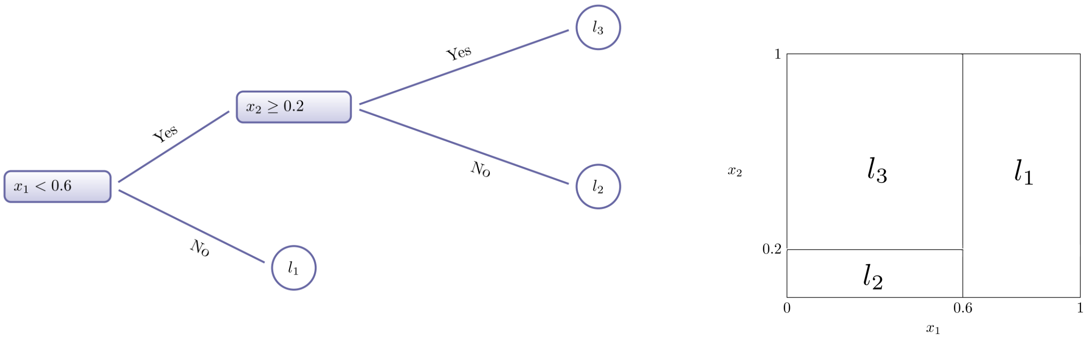


Figure 1: (Left) An example binary tree. The internal nodes are labelled by their splitting rules and the terminal nodes labelled with the corresponding parameters l_i.
(Right) The corresponding partition of the sample space.

Binary trees are named *classification trees* when the outcome variable can take a discrete set of values, and *regression trees* when the outcome variable takes continuous values. The CART algorithm associates, for every individual belonging to a partition of the feature space, a conditional prediction for the outcome variable. The task of a binary tree is to estimate the conditional expectation of the 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 to tune the complexity of the tree, in order to minimize the "error"[6] between the true and estimated values of $Y_i(x)$ within each partition.

The accuracy of the predictions of binary trees, $\hat{Y}_i(x)$, can be dramatically improved by iteratively constructing the trees. A Random Forest (RF) consists in an ensemble of trees, where each tree is constructed by randomly sampling the observations and randomly drawing the covariates (predictors, in the machine learning literature) that are used to build each tree

[6] There are various "error" measures used to optimize binary trees. The most widely used are the mean-squared-error for regression trees and the entropy or the Gini index for classification trees.

(Breiman, 2001). One of the main problems of RFs is that they tend to "overfit" the data on which they are trained. "Overfitting" leads to a scarce generalizability of the predictions on samples different from the training set. In order to avoid this, Bayesian Additive Regression Trees were proposed by Chipman et al. (2010).

BART, as well as BCF, are "refined versions" of the RF algorithm. BART is, as the RF, a sum-of-trees ensemble algorithm, but its estimation approach used to obtain the values of $Y_i(x)$ relies on a fully Bayesian probability model (Kapelner and Bleich, 2013). In particular, the BART model can be expressed as:

$$Y_i = f(X_i) + \epsilon_i \approx \mathbb{T}_1(X_i) + ... + \mathbb{T}_q(X_i) + \epsilon_i, \qquad \epsilon_i \sim \mathcal{N}(0, \sigma^2), \tag{7}$$

where the q distinct binary trees are denoted by $\mathbb{T}$[7].

The Bayesian component of the algorithm is incorporated in a set of three different priors on: (i) the structure of the trees (this prior is aimed at limiting the complexity of any single tree $\mathbb{T}$ and works as a regularization device); (ii) the probability distribution of data in the nodes (this prior is aimed at shrinking the node predictions towards the center of the distribution of the response variable Y_i); (iii) the error variance σ^2 (which bounds away σ^2 from very small values that would lead the algorithm to overfit the training data)[8]. The aim of these priors is to "regularize" the algorithm, preventing single trees to dominate the overall fit of the model (Kapelner and Bleich, 2013). Moreover, BART allows the researcher to tune the variables' importance by departing from the original formulation of the Random Forest algorithm where each variable is equally likely to be chosen from a discrete uniform distribution (i.e., with probability $\frac{1}{p}$) to build a single tree learner. These Bayesian tools give researchers the possibility to mitigate the "overfitting" problem of RFs and to tune the algorithm with prior knowledge. Give these characteristics BART has shown particular flexibility and an excellent performance in both prediction tasks

[7] $\mathbb{T}$ represents the entire tree: its structure, its nodes and its leaves (terminal nodes).

[8] The choice of the priors, and the derivation of the posterior distributions, is discussed in depth by Chipman et al. (2010) and Kapelner and Bleich (2013). Namely, (i) the prior on the probability that a node will split at depth k is $\beta(1+k)^{-\eta}$ where $\beta \in (0,1), \eta \in [0,\infty)$ (these hyper-parameters are generally chosen to be $\eta = 2$ and $\beta = 0.95$); (ii) the prior on the probability distribution in the nodes is a normal distribution with zero mean: $\mathcal{N}(0, \sigma_q^2)$ where $\sigma_q = \sigma_0/\sqrt{q}$ and σ_0 can be used to calibrate the plausible range of the regression function; (iii) the prior on the error variance is $\sigma^2 \sim InvGamma(v/2, v\lambda/2)$ where λ is determined from the data in a way that the BART will improve 90% of the times the RMSE of an OLS model.

(Murray, 2017; Linero and Yang, 2018; Linero, 2018; Hernández et al., 2018; Starling et al., 2018a) and in causal inference tasks (Hill, 2011; Hahn et al., 2017; Logan et al., 2019).

Thus far, the algorithms that we discussed are tailored to find heterogeneity in the response variable $Y_i(x)$, but are not developed to estimate the heterogeneity in the causal effects. The BCF algorithm proposed by Hahn et al. (2017) is a semi-parametric Bayesian regression model that directly builds on BART. It, however, introduces some significant changes in order to estimate heterogeneous treatment effects in regular assignment mechanisms (even in the presence of strong confounding). The principal novelties of this model are the expression of the conditional mean of the response variable as a sum of two functions and the introduction, in the BART model specification for causal inference, of an estimate of the propensity score, $E[W_i = 1|X_i = x] = \pi(x)$, in order to improve the estimation of heterogeneous treatment effects[9]. As depicted from the results of the Atlantic Causal Inference Conference (ACIC) competition in 2016 and 2017, reported by Hahn et al. (2018b), it was observed that BCF performs dramatically better than other machine learning algorithms for causal inference in the presence of randomized and regular assignment mechanisms.

2.3 Extending BCF to an IV Scenario: Bayesian Instrumental Variable Causal Forest

Bargagli Stoffi and Gnecco (2019) show that a naïve application of methods developed for the estimation of heterogeneous causal effects in randomized or regular assignment mechanisms would introduce a large bias in the estimation of the heterogeneous causal effects in imperfect compliance settings. Moreover, the authors show that this would lead to very imprecise heterogeneous causal effects estimators. This reason drives the need for a new algorithm, the Bayesian Instrumental Variable Causal Forest (BCF-IV), tailored for causal inference on heterogeneous effects in the presence of irregular mechanisms.

[9] It is important to highlight that the propensity score is not used to estimate the causal effects but to moderate the distortive effects in treatment heterogeneity discovery due to strong confounding. Moreover, since BCF includes the entire predictors' vector, $\mathbf{X}$, even if the propensity score is mis-specified or poorly estimated, the model allows for the possibility that the response remains correctly specified (Hahn et al., 2017). In Appendix B, we show that even if the estimate $\hat{\pi}(x)$ of the propensity score is incorrectly specified the results are still widely robust.

The BCF-IV algorithm is constructed in two steps:

1. Discovering heterogeneity for the conditional intention-to-treat ($ITT_Y(x)$);
2. Estimation of the conditional CACE ($\tau^{cace}(x)$) within the sub-populations defined in the first step.

We will discuss these two steps in detail in the next Sections.

2.3.1 Heterogeneity in the Conditional ITT

The BCF-IV algorithm starts from modifying (7) to adapt it for the estimation of the intention-to-treat, by including the instrumental variable Z_i:

$$Y_i = f(X_i, Z_i) + \epsilon_i \approx \mathbb{T}_1(X_i, Z_i) + ... + \mathbb{T}_q(X_i, Z_i) + \epsilon_i, \qquad \epsilon_i \sim \mathcal{N}(0, \sigma^2), \tag{8}$$

where, for simplicity, we assume the error to be a mean zero additive noise as in Hill (2011); Hahn et al. (2017); Logan et al. (2019). The conditional expected value can be expressed as:

$$\mathbb{E}[Y_i|Z_i = z, X_i = x] = \mu(z, x), \tag{9}$$

and in turn the conditional intention-to-treat, $ITT_Y(x)$, is:

$$ITT_Y(x) = \mathbb{E}[Y_i|Z_i = 1, X_i = x] - \mathbb{E}[Y_i|Z_i = 0, X_i = x] = \mu(1, x) - \mu(0, x). \tag{10}$$

Then, adapting to an irregular assignment mechanism the model proposed by Hahn et al. (2017), we adopt the following functional form for (9):

$$\mathbb{E}[Y_i|Z_i = z, X_i = x] = \mu(x, \hat{\pi}(x)) + ITT_Y(x) \cdot z \tag{11}$$

where $\hat{\pi}(x)$ is the estimated propensity score for the instrumental variable:

$$\pi(x) = E[Z_i = 1|X_i = x]. \tag{12}$$

The expression of $\mathbb{E}[Y_i|Z_i = z, X_i = x]$ as a sum of two functions is central: the first component of the sum, $\mu(x, \hat{\pi}(x))$, directly models the impact of the control variables on the conditional mean of the response (the component that is independent from the treatment effects) while the second component $ITT_Y(x)z$ models directly the intention-to-treat effect as a nonlinear function of the observed characteristics (this second components captures the heterogeneity in the intention-to-treat). Both the functions μ and ITT_Y are given independent priors. These priors are chosen in line with Hahn et al. (2017) to be for the first component the same priors of Chipman et al. (2010) (see Section 2.2). However, for the second component the priors are changed in a way that allows for less deep, hence simpler trees[10].

The expression of $\mathbb{E}[Y_i|Z_i = z, X_i = x]$ as a sum of two functions has a double effect: (i) on the one hand, it allows the algorithm to learn which component in the heterogeneity of the conditional mean of the outcome is driven by a direct effect of the control variables and which component is the true heterogeneity in the effects of the assignment to the treatment Z_i on Y_i; (ii) on the other hand, it allows the predictions of the treatment effect driven by the BART to be modelled directly and separately with respect to the impact of the control variables (Hahn et al., 2017).

The estimated propensity score, in the BCF model, is not used for the estimation of the effects but is included, as an additional covariate, in the first component of (11) to mitigate possible problems connected to *regularization induced confounding* (RIC)[11] and *targeted selection*[12]. Moreover, in scenarios where the instrumental variable is not randomized ex-ante, the inclusion of $\hat{\pi}(x)$ leads to an improvement in the discovery of the heterogeneity in the causal effect (Hahn et al., 2018b). Furthermore, it is important to highlight that choosing a mis-specified definition of $\hat{\pi}(x)$ does not impact in a significant way the quality of the results as shown in Appendix B. This is due to the fact that this first step of our algorithm is not about directly estimating the conditional CACE but is tailored to discover the heterogeneity in $ITT_Y(x)$.

[10] The depth penalty parameters are set to be $\eta = 3$ and $\beta = 0.25$ (instead of $\eta = 2$ and $\beta = 0.95$).

[11] RIC is analyzed in depth in Hahn et al. (2018a). RIC issues rise when the ML algorithm used for regularizing the coefficient does not shrink to zero some coefficients due to a nonzero correlation between Z_i and X_i resulting in an additional degree of bias that is not under the researcher's control.

[12] Targeted selection refers to settings where the treatment (or in an IV scenario the assignment to the treatment) is assigned based on an ex-ante prediction of the outcome conditional on some characteristics X_i. We refer to Hahn et al. (2017) for a discussion of targeted selection problems.

Once one estimated with the BCF-IV the unit-level intention-to-treat, one can build a simple binary tree, using a CART model (Breiman, 1984), on the fitted values ($\widehat{ITT}_Y(X_i)$) to discover the drivers of the heterogeneity.

2.3.2 Heterogeneous Conditional ITT with a Binary Outcome Variable

As pointed out by Starling et al. (2018b) the original BCF algorithm by Hahn et al. (2017) is not directly tailored to handle binary outcomes. Following Starling et al. (2018b), we propose an extension of our BCF-IV algorithm to binary outcomes. This novel version of the algorithm is described in detail in the following.

To deal with the binary outcome case, in line with the BART model of Chipman et al. (2010), we modify the BCF-IV algorithm by introducing a random variable c_i, modelled as:

$$c_i = \mu(x_i, \hat{\pi}(x_i)) + ITT_Y(x_i) \cdot z_i + \epsilon_i, \tag{13}$$

where $\epsilon_i \sim \mathcal{N}(0,1)$, and is independent from z_i and x_i. Hence, c_i is a Gaussian latent variable given z_i and x_i. We model the outcome $y_i \in \{0,1\}$ as:

$$y_i = \begin{cases} 1 \text{ if } \phi\big(\mu(x_i, \hat{\pi}(x_i)) + ITT_Y(x_i) \cdot z_i\big) \geq 0.5, \\ 0 \text{ if } \phi\big(\mu(x_i, \hat{\pi}(x_i)) + ITT_Y(x_i) \cdot z_i\big) < 0.5, \end{cases} \tag{14}$$

where ϕ is the standard normal cumulative distribution function (CDF). The counterfactual probabilities are:

$$\begin{aligned} \omega_i(0) &= \phi\big(\mu(x_i, \hat{\pi}(x_i))\big), & (15) \\ \omega_i(1) &= \phi\big(\mu(x_i, \hat{\pi}(x_i)) + ITT_Y(x_i) \cdot z_i\big). & (16) \end{aligned}$$

2.3.3 Estimation of Conditional CACE

Once the heterogeneous patterns in the intention-to-treat (ITT) are learned from the algorithm, one can estimate the conditional CACE, $\tau^{cace}(x)$. To do so, one can simply use the method of moments estimator in Equation (6) within all the different sub-populations that were detected in the previous step.

The conditional CACE can be estimated in a generic sub-sample (i.e., for each $X_i \in \mathbb{X}_j$, where $\mathbb{X}_j$ is a generic node of the tree, like a non-terminal node or a leaf) as:

$$\hat{\tau}^{cace}(X_i) = \frac{\widehat{ITT}_Y(X_i)}{\hat{\pi}_C(X_i)}, \tag{17}$$

where $\hat{\pi}_C(X_i)$ is estimated as:

$$\hat{\pi}_C(X_i) = \frac{1}{N_{1,l}} \sum_{l:X_l \in \mathbb{X}_j} W_l \cdot Z_l - \frac{1}{N_{0,l}} \sum_{l:X_l \in \mathbb{X}_j} W_l \cdot (1 - Z_l), \tag{18}$$

and $\widehat{ITT}_Y(X_i)$ as:

$$\widehat{ITT}_Y(X_i) = \frac{1}{N_{1,l}} \sum_{l:X_l \in \mathbb{X}_j} Y_l^{obs} \cdot Z_l - \frac{1}{N_{0,l}} \sum_{l:X_l \in \mathbb{X}_j} Y_l^{obs} \cdot (1 - Z_l), \tag{19}$$

where $N_{k,l}$ (where $k \in \{0,1\}$) is the number of observations with $Z_l \in \{0,1\}$ in the sub-sample of observations with $X_l \in \mathbb{X}_j$[13].

To show in detail how this second step works let us use a toy example. Let's imagine a simple heterogeneity structure for $ITT_Y(x)$ where $ITT_Y(X_{i,p} > 0) \gg ITT_Y(X_{i,p} \leq 0)$ and $X_{i,p} \in (-1,1)$ is a single regressor. This is namely, the case where the average intention-to-treat for those individuals with positive values of $X_{i,p}$ is greater than for individuals with non-positive values. Then, the conditional CACE can be estimated in the two different sub-populations defined with respect to X_p as[14]:

$$\hat{\tau}^{cace}(X_{i,p} > 0) = \frac{\widehat{ITT}_Y(X_{i,p} > 0)}{\hat{\pi}_C(X_{i,p} > 0)} \text{ and } \hat{\tau}^{cace}(X_{i,p} \leq 0) = \frac{\widehat{ITT}_Y(X_{i,p} \leq 0)}{\hat{\pi}_C(X_{i,p} \leq 0)}. \tag{20}$$

[13] It is worth highlighting that, since the supervised machine learning technique is used in the discovery phase and not in the estimation phase, the estimators that are proposed here could be used in a more "traditional way", in settings where the subgroups are defined ex-ante by the researcher.

[14] Alternatively, one can perform a Two Stage Least Squares (2SLS) regression within the different sub-populations. This is our preferred estimation strategy and is the one used both for the simulations and the real application.

2.4 Properties of the Conditional CACE Estimator

In the case of a binary instrument ($Z_i \in \{0,1\}$) and a binary treatment variable ($W_i \in \{0,1\}$), Angrist et al. (1996) and Imbens and Rubin (2015) revealed that the population versions of (17)-(19) correspond to a Two Stage Least Squares (henceforth referred as 2SLS) estimator of τ^{cace}, in the cases where the four IV assumptions can be assumed to hold. Hence, since this case is analogous to our setting, one can apply the 2SLS method in every node $\mathbb{X}_j$ of the tree $\mathbb{T}$ for the estimation of the effect on the compliers population, as it is presented by Imbens and Rubin (1997).

The two simultaneous equations of the 2SLS estimator are, in the population,

$$Y_i^{obs} = \alpha + \tau^{cace} \cdot W_i + \epsilon_i, \tag{21}$$

$$W_i = \pi_0 + \pi_C \cdot Z_i + \eta_i, \tag{22}$$

where $\mathbb{E}(\epsilon_i) = \mathbb{E}(\eta_i) = 0$, and $\mathbb{E}(Z_i\eta_i) = 0$[15]. In the econometric terminology, the explanatory variable W_i is *endogenous*, while the IV variable Z_i is *exogenous*.

We can express the 2SLS equations, conditional on a subpopulation of a node $\mathbb{X}_j$, as

$$Y_{i,\mathbb{X}_j}^{obs} = \alpha_{\mathbb{X}_j} + \tau_{\mathbb{X}_j}^{cace} \cdot W_{i,\mathbb{X}_j} + \epsilon_{i,\mathbb{X}_j}, \tag{23}$$

$$W_{i,\mathbb{X}_j} = \pi_{0,\mathbb{X}_j} + \pi_{C,\mathbb{X}_j} \cdot Z_{i,\mathbb{X}_j} + \eta_{i,\mathbb{X}_j}, \tag{24}$$

where $\mathbb{E}(\epsilon_{i,\mathbb{X}_j}) = \mathbb{E}(\eta_{i,\mathbb{X}_j}) = 0$, and $\mathbb{E}(Z_{i,\mathbb{X}_j}\eta_{i,\mathbb{X}_j}) = 0$.

Moreover, the following reduced equation (obtained plugging (24) into (23)) holds:

$$\begin{aligned} Y_{i,\mathbb{X}_j}^{obs} &= \left(\alpha_{\mathbb{X}_j} + \tau_{\mathbb{X}_j}^{cace} \cdot \pi_{0,\mathbb{X}_j}\right) + \left(\tau_{\mathbb{X}_j}^{cace} \cdot \pi_{C,\mathbb{X}_j}\right) \cdot Z_{i,\mathbb{X}_j} + \left(\epsilon_{i,\mathbb{X}_j} + \tau_{\mathbb{X}_j}^{cace} \cdot \eta_{i,\mathbb{X}_j}\right) \\ &= \bar{\alpha}_{\mathbb{X}_j} + \gamma_{\mathbb{X}_j} \cdot Z_{i,\mathbb{X}_j} + \psi_{i,\mathbb{X}_j}. \end{aligned} \tag{25}$$

In the case of a single instrument, the logic of IV regression is that one can estimate the respective parameters $\pi_{C,\mathbb{X}_j}$ and $\gamma_{\mathbb{X}_j} = \tau_{\mathbb{X}_j}^{cace} \cdot \pi_{C,\mathbb{X}_j}$ of the regressions (24) and (25) above by least squares, when the observations in each node are independent and identically distributed, then obtaining

[15] The latter comes from the fact that (22) is assumed to represent the linear projection of W_i onto Z_i.

an estimate of the parameter $\tau_{\mathbb{X}_j}^{cace}$ in (23). In particular, for every element X_i of a node $\mathbb{X}_j$, one can estimate $\tau^{CACE}(X_i) = \tau_{\mathbb{X}_j}^{cace}$ through 2SLS, as the following ratio (Imbens and Rubin, 2015):

$$\hat{\tau}^{CACE}(X_i) \equiv \hat{\tau}_{\mathbb{X}_j}^{2SLS} = \frac{\hat{\gamma}_{\mathbb{X}_j}}{\hat{\pi}_{C,\mathbb{X}_j}}. \tag{26}$$

The 2SLS estimator associated with (23)-(25) satisfies the next properties. They can be proved likewise in the application of 2SLS to the population case (see, e.g., Imbens and Rubin (2015)).

Theorem 1: Consistency of the Conditional 2SLS Estimator.
Let $\mathbb{E}(Z_{i,\mathbb{X}_j}^2) \neq 0$ (Assumption 1), $\mathbb{E}(Z_{i,\mathbb{X}_j}\epsilon_{i,\mathbb{X}_j}) = 0$ (Assumption 2) and $\pi_{C,\mathbb{X}_j} \neq 0$ (Assumption 3) hold. Then

$$\hat{\tau}_{\mathbb{X}_j}^{2SLS} - \tau_{\mathbb{X}_j} \xrightarrow{p} 0 \text{ as } N_{\mathbb{X}_j} \to \infty, \tag{27}$$

where $\xrightarrow{p}$ denotes convergence in probability, and $N_{\mathbb{X}_j}$ is the number of observations within the node $\mathbb{X}_j$.

It should be noted that Assumption 3 is not necessarily guaranteed even if the overall instrument is strong. However, this assumption is standard in treatment effects variation papers such as the contribution of Ding et al. (2019)[16].

Theorem 2: Asymptotic Normality of the Conditional 2SLS Estimator.
Let Assumptions 1, 2, and 3 hold. Let also $\mathbb{E}(Z_{i,\mathbb{X}_j}^2\epsilon_{i,\mathbb{X}_j}^2)$ be finite (Assumption 4). Then

$$\sqrt{N_{\mathbb{X}_j}}\big(\hat{\tau}_{\mathbb{X}_j}^{2SLS} - \tau_{\mathbb{X}_j}\big) \xrightarrow{d} \mathcal{N}\big(0, N_{\mathbb{X}_j} \cdot avar(\hat{\tau}_{\mathbb{X}_j}^{2SLS})\big) \text{ as } N_{\mathbb{X}_j} \to \infty, \tag{28}$$

where $\xrightarrow{d}$ denotes convergence in distribution, $\mathcal{N}$ stands for normal distribution, and $avar(\hat{\tau}_{\mathbb{X}_j}^{2SLS})$ is the asymptotic variance of the 2SLS estimator.

The proofs of the two Theorems above directly follow from their unconditional versions[17]. In

[16] In cases where Assumption 3 is not violated but the proportion of compliers within the nodes approaches zero, there could be potential problems related to heterogeneity driven by these small values of $\pi_{C,\mathbb{X}_j}$ and weak instruments issues within the nodes. In our scenario, since the heterogeneity in the treatment effect is detected with respect to the conditional ITT (which does not take directly into account $\pi_{C,\mathbb{X}_j}$ for its computation), our model is robust to possible treatment effects heterogeneity variations driven by smaller values of $\pi_{C,\mathbb{X}_j}$. Moreover, we do run weak-instrument tests within every node and we discard those nodes where a weak-instrument issue is detected.

[17] For further details on these proofs we refer to (Wooldridge, 2002, Section 5.2).

this case, for the convergence of our estimator to $\tau_{\mathbb{X}_j}$ and its normality to hold approximately we need to have a sufficient number of observations within every node. Hence, we suggest to perform our algorithm on sufficiently large datasets and to trim those nodes where the number of observations is not large enough[18].

3 Monte Carlo Simulations

To evaluate the performance of the BCF-IV algorithm we compare it, using Monte Carlo Simulations, with two methods that are directly tailored for drawing causal inference in irregular assignment mechanism scenarios: the Honest Causal Trees with Instrumental Variable (HCT-IV) algorithm (Bargagli Stoffi and Gnecco, 2019) and the Generalized Random Forests (GRF) algorithm (Athey et al., 2016). Both the latter algorithms outperform other machine learning methodologies which are not tailored for irregular assignment mechanisms (Bargagli Stoffi and Gnecco, 2019), so we focus, in this context, just on a comparison within these three algorithms.

Since the foremost focus of this paper is on discovery of the heterogeneity in the causal effects, we compare the algorithms on three dimensions: (i) the correct choice of the variable that drives the heterogeneity (*heterogeneity driving variable* [HDV]), (ii) the correct choice of the threshold value used to perform the binary split of the data given the right identification of HDV, and (iii) the mean-squared error for the heterogeneous causal effects given the correct choice of HDV.

For Monte Carlo Simulations we build two different designs. The functional forms of the designs are built following the simulation designs in Wang et al. (2018). The first design takes the form of $Y_i = \sum_{p=1}^{k} X_{i,p} + W_i \cdot X_{i,1} + \xi_i + \epsilon_i$ where $X_{i,p} \sim \mathcal{N}(0,1)$, $W_i \sim Bern(0.5), \xi_i \sim \mathcal{N}(0, 0.01)$ and $\epsilon_i \sim \mathcal{N}(0,1)$. The interaction term between the regressor $X_{i,1}$ and the treatment indicator W_i is functional to heterogenise the treatment effects, while the nuisance parameter ξ_i is an unobserved variable that affects both W_i and the response variable Y_i. The second design has the same functional form but $x_{i,p} \sim Bern(0.5)$. In both the designs we set the correlations between W_i, the instrument Z_i and the nuisance parameter ξ_i to be: $Cor(W_i, Z_i) \in (0.55, 0.65)$ and $Cor(W_i, \xi_i) \in (0.45, 0.55)$, while k assumes values 5 and 10 and the sample sizes are 500,

[18] An `R` function for BCF-IV, and the code used for both the simulations and the application study in the paper are available at the GitHub page of the corresponding author: `https://github.com/barstoff`.

1,000, 5,000. For both designs the results are aggregated over 30 rounds of simulations.

The results from the simulations are shown in Table 1. As shown in Panel A, the correct identification of the HDV is very similar for BCF-IV and GRF in the designs with 500 and 5,000 units. GRF is asymptotically faster in identifying the right HDV, as it outperforms both BCF-IV and HCT-IV when the sample size is 1,000. Panel B depicts the results in terms of mean squared error between the true and the predicted threshold used to perform the binary split of the data. The threshold is not available in Design 2 where the regressors are binary variables. BCF-IV outperforms both GRF and HCT-IV with all the sample sizes and with both 5 and 10 features (with the exception of Design 1 in the sample of 5,000 units with 5 features). Panel C depicts the mean squared error of prediction for the causal effects given the correct identification of the HDV. Another clear advantage of using BCF-IV is given by the correct identification of the treatment effects. Indeed, BCF-IV outperforms, in terms of lower mean-squared-error of prediction for the treatment effects, the other algorithms in both the designs with 5 and 10 features (with the exception of Design 1 in the samples of 5,000 units with 5 features and 500 units with 10 features). Hence, in a scenario with binary regressors, BCF-IV is preferable irrespective of the sample size. However, in a scenario with continuous regressors[19], there is a tradeoff between the capacity of getting the right HDV and the capacity of correctly estimating the causal effect. In designs with samples sizes of 1,000, GRF outperforms BCF-IV in correctly identifying the HDV but fails to precisely estimate the causal effect. As the sample size increases, particularly in the scenario with 5 features, both the algorithms get to the same asymptotic results in terms of correct HDV identification and the mean-squared-error of prediction.

We argue that, in small samples, BCF-IV would be preferable to GRF because, while the proportion of correctly identified HDVs is very similar, the gains obtained both in terms of mean-squared-error between the true and predicted threshold and the true and predicted causal effects are much larger. In fact, the relative gap[20] between the true and predicted causal effects ranges between 15% and 81% in favour of BCF-IV, while the relative gap in the proportion of correctly

[19] For instance, when the regressors are distributed according to a standardized normal distribution.

[20] The formula for the relative gap is, for the MSE of prediction, the following (Wang et al., 2018):

$$\text{Relative Gap} = \frac{MSE_{GRF} - MSE_{BCF\text{-}IV}}{MSE_{GRF}} \times 100.$$

The relative gap is positive when BCF-IV outperforms GRF and negative viceversa.

identified HDV ranges from -10% (in favour of GRF) to 30% (in favour of BCF-IV). Hence, we claim that the gain in the mean-squared-error of prediction for the causal effect outweighs the slower identification of HDVs. This holds true as BCF-IV and GRF converge to a very similar fit, as the sample size increases, with respect to the three dimensions that are the object of our analysis. Moreover, the asymptotic behaviour of BCF-IV is slightly better than the one of all the other techniques.

Table 1: **Monte Carlo Comparison of BCF-IV, GRF and HCT-IV**[21]

Panel A: Proportion of Correctly Identified Heterogeneity Driving Variables (HDV)

		Sample Size					
#Features	Approach	500	1,000	5,000	500	1,000	5,000
		Design 1			Design 2		
		HDV					
5	BCF-IV	0.57	0.57	**1.00**	0.90	0.96	**1.00**
	GRF	**0.63**	**0.83**	**1.00**	**0.93**	**1.00**	**1.00**
	HCT-IV	0.43	0.40	0.70	0.53	0.56	0.86
10	BCF-IV	**0.30**	0.33	0.73	**0.53**	0.93	**1.00**
	GRF	0.23	**0.43**	**1.00**	0.50	**0.96**	**1.00**
	HCT-IV	0.17	0.30	0.77	0.20	0.63	0.83

[21] Note: In this Table we show the results from Monte Carlo simulations. Panel A depicts the proportion of correctly identified heterogeneity driving variables (HDVs). Panel B shows the mean squared error between true and predicted threshold (which is available just for the model with normally distributed variables). Panel C depicts the mean squared error between true and predicted causal effects. We highlighted in bold the best results for every round of simulations. In case of the same mean-squared-error of prediction (that occurs with BCF-IV and HCT-IV since both techniques are based on the same estimator) we assigned the best performance to the technique that performs better in identifying the correct HDVs. The results are obtained by aggregating 30 bootstrap samples.

Panel B: Mean Squared Error between True and Predicted Threshold

		Sample Size					
#Features	Approach	500	1,000	5,000	500	1,000	5,000
		Design 1			Design 2		
		Threshold					
5	BCF-IV	**0.062**	**0.037**	0.006	-	-	-
	GRF	0.063	0.069	**0.002**	-	-	-
	HCT-IV	0.185	0.188	0.045	-	-	-
10	BCF-IV	**0.046**	**0.014**	**0.017**	-	-	-
	GRF	0.190	0.125	0.040	-	-	-
	HCT-IV	0.096	0.023	0.167	-	-	-

Panel C: Mean Squared Error between True and Predicted Causal Effects

		Sample Size					
#Features	Approach	500	1,000	5,000	500	1,000	5,000
		Design 1			Design 2		
		Causal Effects					
5	BCF-IV	**0.047**	**0.026**	0.002	**0.048**	**0.005**	**0.010**
	GRF	0.330	0.030	**0.001**	0.055	0.006	0.011
	HCT-IV	0.067	0.051	0.012	0.048	0.005	0.010
10	BCF-IV	0.017	**0.021**	**0.020**	**0.036**	**0.005**	**0.002**
	GRF	0.230	0.197	0.128	0.190	0.105	0.012
	HCT-IV	**0.013**	0.039	0.046	0.036	0.005	0.002

In Appendix B, we provide a number of robustness checks of the Monte Carlo simulation. In particular, we focus on what happens to the fit of the three algorithms when one: (i) changes the correlation between Z_i and W_i (possible weak-instrument problems)[22]; (ii) introduces a violation in the exclusion restriction; (iii) changes the specification of the propensity score for the BCF-IV; (iv) introduces multiple heterogeneity variables; (v) changes the error distribution. The results that we highlighted before hold true also in the robustness checks: BCF-IV converges slowly to an optimal identification of the HDVs but largely outperforms GRF with respect to the mean-squared-error of prediction for the causal effects[23]. Moreover, the performance of BCF-IV does not seem to widely deteriorate, as compared to the baseline models in Table 1, in any of the robustness designs.

4 Heterogeneous Causal Effects of Education Funding

There is a wide consensus that education positively influences labor market outcomes (see the review by Psacharopoulos and Patrinos (2018)). Students' performance can be driven by multiple factors connected with students' characteristics and environmental characteristics. However, to the best of our knowledge, this is the first paper to study the impact of additional school funding on students' performance using machine learning techniques tailored for causal inference. In this Section we apply the BCF-IV algorithm to evaluate the impact and estimate the heterogeneity in the effects of additional funding to schools with disadvantaged students on students' performance. First, we describe the data used for this application. Next, we depict the identification strategy. Finally, we describe the results obtained and their relevance in the economics of education literature.

4.1 Data

Starting from the year 2002, the Flemish Ministry of Education promoted the "Equal Educational Opportunities" program (henceforth referred as EEO) to ensure equal educational opportunities to all the students (OECD, 2017). The EEO program provides additional funding for secondary

[22] It is important to highlight that in order to avoid weak-instrument problems within a node our algorithm performs a weak-instrument test in every sub-sample (namely, an F-test on the first stage regression) and discards the nodes where the null hypothesis of weak instrument is not rejected.

[23] However, when we introduce a partial violation of the exclusion restriction assumption (design 2) we see exactly the opposite: BCF-IV outperforms GRF with respect to the identification of the correct HDV while GRF outperforms BCF-IV in precisely estimating the causal effects.

schools with a significant share of disadvantaged students. Owing to the funding schools can hire additional teachers and increase the number of teaching hours. Pupils are considered to be disadvantaged on the basis of five different indicators: (i) the pupil lives outside the family; (ii) the pupil does not speak Dutch as a native language; (iii) the mother of the pupil does not have a secondary education degree; (iv) the pupil receives educational grant guaranteed for low income families; and (v) one of the parents is part of the travelling population. In order for a school to be eligible for the EEO funding, it needs to satisfy two conditions: the first condition is that the share of students with at least one of the five characteristics has to exceed an exogenously set threshold; to avoid fragmentation of resources, the second condition requires that the additional resources should be at least larger than six teaching hours a week. The exogenous cutoff is, for students in the first two years of secondary education (first stage students), a minimum share of 10% disadvantaged students.

The Flemish Ministry of Education provided us with data on the universe of pupils in the first stage of education in the school year 2010/2011 (135,682 students). In particular, we have data on student level characteristics and school level characteristics. The student level characteristics cover the gender of the pupil (*gender*), the grade retention in primary school (*retention*) and the inclusion of the pupil in the special needs student population in primary school (which serves as a proxy of student's low cognitive skills). The school level characteristics include both the teacher characteristics, such as the teachers' age, seniority and education, in addition to principal characteristics, such as the principals' age and seniority. Teacher and principal seniority measures the level of experience of the teachers and principals, respectively. These variables assume values in the range of 1 to 7, where the teachers (and principals) with a seniority level of 1 are the least experienced (0-5 years of experience) and teachers (and principals) with a seniority level of 7 are the most experienced (more than 30 years of experience)[24]. Similarly, the ages of teacher and principal are reported as categorical variables that range from 1 to 8, where teachers/principals in the first category are the youngest (less than 30 years old) and teachers/principals in the last category are the oldest (more than 60 years old)[25]. Teachers' education records whether

[24] Teachers and principals' seniority classes are the following: class 1: between 0 and 5 years of experience; class 2: between 6 and 10; class 3: between 11 and 15; class 4: between 16 and 20; class 5: between 20 and 25; class 6: between 26 and 30; class 7: more than 30.

[25] Teachers and principals' age classes are the following: class 1: less than 30 years old; class 2: between 30

or not the teacher holds a pedagogical training (in the following we will refer to it as "teacher training"). All these variables are aggregated at school level in the form of averages (for age and seniority) and shares (for teachers' education) and assigned to each student with respect to the school where he/she is enrolled.

The outcome variables are two dummy variables defined as follows: the variable *progress school* assumes value 1 if the student progresses to the following year without any grade retention and 0 if not (this variable is a complement of school retention); the variable *A-certificate* assumes value 1 if the student gets an "A-certificate" at the end of the school year (which is the most favorable outcome) and 0 if not. Since we do not have data on standardized test scores for Flemish students, *A-certificate* is a good, available proxy of student performance. Every year, each student performs a final test and gets a ranking from "A" to "C". Students that get an "A" can progress school without any restriction, while the students that get either "B" or "C" can progress school but only in specific programs or have some grade retention. Both these outcome variables are proxies for different levels of students' performance: a positive *A-certificate* proxies for a higher level of performance than a positive *progress school.* In principle, the target of a policy-maker could be to have the highest possible share of students getting "A-certificates" and the lowest share of students not progressing through school.

4.2 Identification Strategy

To evaluate the impact of the policy on students' performance we apply the BCF-IV within a regression discontinuity design (Hahn et al., 2001). Regression Discontinuity Design (RDD) is a method that aims at evaluating the causal effects of interventions in settings where the assignment to the treatment is determined (at least partly) by the values of an observed covariate lying on either sides of a threshold point. The idea is that subjects just above and below this threshold are very similar and one can assume a quasi-randomization around the threshold (Mealli and Rampichini, 2012). RDDs are categorized in sharp RDDs and fuzzy RDDs. In sharp RDDs, the central assumption is that, around the threshold, there is a sharp discontinuity (from 0 to 1) in the probability of being treated. This is due to the fact that in sharp RDDs there

and 34; class 3: between 35 and 39; class 4: between 40 and 44; class 5: between 45 and 49; class 6: between 50 and 54; class 7: between 55 and 60; class 8: more than 60.

is no room for imperfect compliance. In many real world scenarios, however, thresholds are not strictly implemented, as in the case of our application. To deal with these situations, one can use fuzzy RDDs, which are applicable when around the threshold the probability of being actually treated changes discontinuously, but not sharply from 0 to 1 (i.e., the jump in the probability of being treated is less than 1). In our application of the fuzzy RDD technique, we exploit two cutoffs around the 10% share of disadvantaged students in the first stage of secondary education. The students in schools just below the threshold are assigned to the control group ($Z_i = 0$), while the students in schools just above the threshold are assigned to the treated group ($Z_i = 1$). The bandwidth around the threshold (from which one obtains the two cutoffs) is determined using the "*rdrobust package*" in `R` (Calonico et al., 2015). The optimal, bias-corrected bandwidths around the threshold are 3.5% and 3.7%, respectively for the outcome variables *A-certificate* and *progress school.* Accordingly to these two bandwidths, we obtain two refined samples where the sample with the 3.5% bandwidth is the smallest and the sample with the 3.7% bandwidth the largest. We run a series of robustness checks for the selection of the bandwidths around the cutoff. In order to validate the bandwidths selected using the method of Calonico et al. (2015), we run additional analyses implementing the Bayesian methods proposed by Li et al. (2015) and by Mattei and Mealli (2016). In these papers, the authors implement a hierarchical Bayesian model for assessing the balance of the covariates between the groups of observations assigned to the treatment and the ones assigned to the control. For both the thresholds selected following Calonico et al. (2015) (i.e., 3.5% and 3.7%), the probability of the pre-assignment variables being well-balanced is high for the subpopulations defined by values of the cutoff strictly lower than 3.5% and 3.7%. Indeed, these probabilities are larger than or close to 0.8, indicating that the covariates are balanced in the two groups.

Moreover, to guarantee an equal representation to all the schools, and to avoid biases related to the over-representation of biggest schools' students, we sample 50 pupils from each school. In turn, this leads to a higher balance among the averages between the observations assigned to the treatment and the observations assigned to the control, as shown in panel (a) of Figure 4. In Appendix C, we run a series of tests to show that the RDD (Regression Discontinuity Design) is valid for this application. Moreover, as a robustness check we sample a higher number of students according to the size of the smallest school (62 pupils) from every school. In Appendix D, we show the balance in the samples of units assigned to the treatment and to the control in the second scenario.

4.3 Results

This Section assesses the effects of the additional funding on students' performances and highlights the main drivers of the heterogeneity in causal effects. These analyses are made for both the outcome variables: *A-certificate* and *progress school*[26].

4.3.1 A-Certificate

Proceeding from the seminal contributions of Coleman (1966) and Hanushek (2003) to recent contributions by Jackson et al. (2015) and Jackson (2018), the question on whether or not school spending affects students' performances has been central in the economic literature.

In our study, the variable *A-certificate* serves as a proxy for positive performance. In our sample, the students that got an "A-certificate" are the 91.73% of the total population. In Figure 2, the heterogeneous Complier Average Causal Effects (CACE) estimated using the proposed model are depicted[27] [28]. The darker the shade of blue in the node the higher the causal effect.

Although positive, the overall effect of the additional funding is not significant. This finding is in line with the recent literature on school spending and students' performance in a cross-country scenario (Hanushek et al., 2016; Hanushek and Woessmann, 2017) and in the Flanders, in particular (De Witte et al., 2018). Nevertheless, rather than focusing on the overall average effect it is more interesting to explore the heterogeneous effects.

The first driver in the heterogeneity of the effects is the variable *teacher seniority*: for students in schools with more senior teachers, the effects of funding are larger. These results, even if not significant, show that the treatment effects are higher for students that are in schools

[26] It is important to highlight that the results for both the outcomes, considered separately, in terms of effects and heterogeneity drivers, remain roughly the same when we widen the sample of units included in the analysis (results are reported in the Appendix D).

[27] The nodes for whom (i) it was not possible to compute the CACE or, (ii) the weak-IV test was not rejected were excluded from the plot.

[28] In Figures 2, 3, 7, 8, the so-called summarizing trees (Hahn et al., 2017) are depicted. A summarizing tree is a classification or regression tree that is built using the fitted values estimated from the BCF-IV. These summarizing trees are used to provide a visualization of the heterogeneity in the causal effects.

with teachers that have more than 16 years of experience. The second driver of heterogeneity is the age of the teacher: students in schools with younger teachers (namely, when *teaching age* assumes values lower or equal to 3.5 on a scale from 1 to 8, referring to teachers younger than 40 years old) have an increase in their performance even if they are in schools with less experienced teachers. Both these heterogeneity drivers, namely, the seniority and age of the teacher, are particularly appreciable, as there are evidences in the education literature that connect teachers' seniority (Rice, 2010; Harris and Sass, 2011) and teachers' age (Holmlund and Sund, 2008) to their teaching performance, and in turn teaching performance to students' positive achievements (Goldhaber and Hansen, 2010).

Further heterogeneous effects come from the interaction between teacher's seniority and principal's age. The effect for students in schools with more experienced teachers and principals younger than 60 is higher than the effect on students in schools with teachers with similar experience but older principals. This evidence can be interpreted in the following way: the additional funding has a positive, but not statistically significant, effect in boosting the performance of students in the overall population, but it increases its effect in a notable way for those students in schools with more senior teachers and younger teachers and principals. These results are in line with the evidence that additional school funding does not boost the performance of the overall population of students (Hanushek et al., 2016; Hanushek and Woessmann, 2017; De Witte et al., 2018) and with the literature that connects students' achievements with teaching performance (Holmlund and Sund, 2008; Rice, 2010; Harris and Sass, 2011) and, in turn, teaching performance with students' performance (Goldhaber and Hansen, 2010). It is important to highlight that even if we find some evidence of treatment effects variation connected to teachers' seniority and age and principal age, the conditional causal effects are not significant.

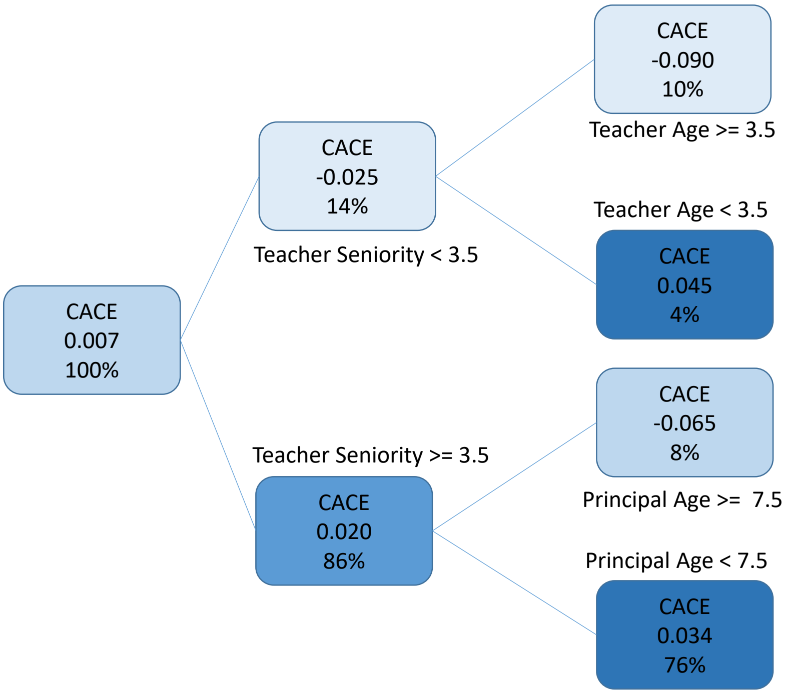


Figure 2: Visualization of the heterogeneous Complier Average Causal Effects (CACE) of additional funding on *A-certificate* estimated using the proposed model. The tree is a summarizing classification tree fit to posterior point estimates of individual treatment effects as in (Hahn et al., 2017). The significance level is * for a significance level of 0.1, ** for a significance level of 0.05 and *** for a significance level of 0.01.

4.3.2 Progress in School

The second outcome variable, *progress school*, assumes value 1 if the student progresses to the following year without any grade retention and 0 if not: roughly 98% of the students in the sample manage to progress in school in the first two years of secondary education. For the students unable to progress in school, this variable is used as a proxy of negative achievements. Therefore, it is relevant to understand if additional funding was effective in driving students away from negative performance. Figure 3 depicts the heterogeneous conditional CACEs: the darker the shade of green in the node, the higher the causal effect.

The additional funding has a slightly negative, but statistically insignificant, impact on the chance of progress in school for the overall students in the sample (again, this is in line with what was found by De Witte et al. (2018) at the school level). However, it is compelling to observe the main drivers of the heterogeneity in the causal effect. The first driver is the seniority level of principals: the treatment effect is positive and statistically significant (the effect in this case is 0.044**, meaning that being treated leads to an increase of 4.4% in the probability of progressing through school) for students in schools with less senior principals (less than 25 years of experience), and it is negative and statistically significant for students in schools with more experienced principals (the conditional effect is -0.029**, meaning that being treated leads to a decrease of 2.9% in the probability of progressing through school). The second driver of the heterogeneity in CACE is the principal age. As in the case of the previous outcome, students in schools with younger principals (namely, principals younger than 55 years) show higher causal effects (the conditional effect is 0.056**). This holds true even when we do not condition on the principal's seniority (in this case the conditional effect is 0.062*). The sub-populations of students with positive and statistically significant effect account for the 36% and 32% of the overall sample (respectively, when conditioning, or not, on the *principal seniority* variable) and show effects that are in their absolute values between 49 and 62 times larger than the overall effect[29]. This evidence of higher and statistically significant effects of the funding in schools for those schools with younger and less senior principals (with respect to the average seniority and age, respectively) is a novel finding of this research. There is a compelling evidence in the literature regarding the role of principals in driving higher students' achievements (Eberts and Stone, 1988; Gentilucci and Muto, 2007), however this is, to the extent of our knowledge, the first research that highlights the role of principal's age and seniority as drivers of treatment effect variations. Clearly, these characteristics could possibly correlate with unobservables, such as the effectiveness of principals (which may decrease as principals grow older). In any case, this finding opens up new fields for further investigation, in line with the newly established role of machine learning in the economic literature as a "theory-driving/theory-testing" tool (Mullainathan and Spiess, 2017).

[29] 49 is the ratio between the conditional treatment effect for the sub-population of students in schools with principals with less than 25 years of experience (0.0440) and the absolute value of the average treatment effect for the overall population (0.0009). 62 is the ratio between the conditional treatment effect for the sub-population of students that are in schools with less senior and younger principals (0.0557), corresponding to the darker shade of green leaf in Figure 3, and the absolute value of the average treatment effect on the overall population (0.0009).

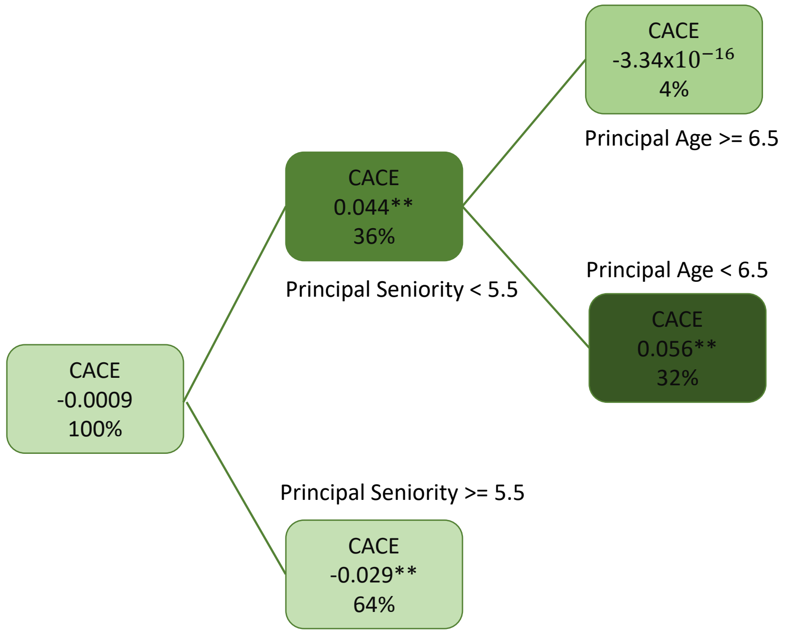


Figure 3: Visualization of the heterogeneous Complier Average Causal Effects (CACE) of additional funding on *Progress School* estimated using the proposed model. The tree is a summarizing classification tree fit to posterior point estimates of individual treatment effects as in (Hahn et al., 2017). The significance level is * for a significance level of 0.1, ** for a significance level of 0.05 and *** for a significance level of 0.01.

5 Conclusion and Discussion

This paper developed a novel Bayesian machine learning technique, BCF-IV, to draw causal inference in scenarios with imperfect compliance. By investigating the heterogeneity in the causal effects, the technique expedites targeted policies. We manifested that the BCF-IV technique outperforms other machine learning techniques tailored for causal inference in precisely estimating the causal effects and converges to an optimal asymptotic performance in identifying the heterogeneity driving variables (HDVs). Moreover, using Monte-Carlo simulations, we show that the competitive advantages of using BCF-IV, as compared to GRF or HCT-IV, are substantial. Peculiarly, the performance of BCF-IV in precisely estimating the heterogeneous causal effects shadows its slower convergence to an optimal identification of HDVs as compared to GRF. This is especially true if we look at the relative gaps between the BCF-IV and the other techniques.

BCF-IV can assist the researchers to shed light on the heterogeneity of causal effects in IV scenarios in order to provide to policy-makers a relevant knowledge for targeted policies.

In our application, we evaluated the effects of additional funding on students' performances. While the overall effects are positive but not significant, there are significant differences among different sub-populations of students. Indeed, for students in schools with less senior and younger principals (with respect to the average seniority and age, respectively) the effects of the policy are between 49 and 62 times bigger than the effects on the overall population (in the most conservative scenario), and significant for the *Progress School* output.

On one hand, as an underlying mechanism, the need for additional funds can be higher in schools with younger teachers and principals, who are more often observed in the most disadvantaged schools. This phenomenon arises as senior teachers and principals select themselves out of the most disadvantaged schools and more into advantaged schools, thereby creating relatively more vacancies in disadvantaged schools. Therefore, on average, younger teachers and principal lack a real choice but to start working in the most disadvantaged schools. Moreover, owing to the additional funds, schools could use the funds to reduce class sizes, which might be more effective for younger (and less senior) teachers. On the other hand, we can think of the motivation for both teachers and principals to decrease as they grow older and this, in turn, have an impact on their performance. Favouring this hypothesis Ololube (2006) finds that motivation enhances productivity and has an impact on teachers' performances. However, teachers' motivation might positively and significantly affect teacher's job satisfaction, but it might not affect their performance. To the best of our knowledge, this is the first study that investigates the effects of age and seniority of principals on enhancing the effectiveness of school funding on students' performance. The investigation of the true causal channel is beyond the goals of this paper and is left to further investigation where more granular teachers' and principals' characteristics are available.

These results are relevant to the policy as they furnish the instruments to policy-makers to enhance the effects of additional funding on students' performance. Indeed, on one side policy-makers could target just students in school with positive, statistically significant effects reducing the overall costs of the policy and using the savings to experiment more effective policies in the other schools. On the other side, policy-makers could analyze the reason of lack of the effectiveness of funding in schools with certain characteristics and implement policies to boost the effects of future funding.

Furthermore, the added value of our algorithm is that it could enable policy-makers to target just those units that benefit the most from the treatment and it provides an insight on possible inefficiencies in the allocation and/or usage of funding. From our analysis it seems that there is room for policies that support less senior principals since students in their schools show higher returns in terms of performance from additional funding.

The extension of these methods to other fields of economic investigation and the development of novel machine learning algorithms for targeted policies and welfare maximization can form the future scope of further research. In particular, the development of an algorithm that could deal with welfare maximization in the context of multiple outcomes is of interest. The "usual" Bayesian way for estimating CACE is via a data augmentation scheme (e.g., imputing compliance status and estimating impacts among estimated compliers [Imbens and Rubin, 2015]). In our algorithm we do not implement such a methodology, however our algorithm could be extended including a data augmentation scheme. Moreover, further research should focus on connecting BCF-IV and GRF into a single ensemble algorithm, following Van der Laan et al. (2007), to obtain a novel algorithm that combines the small and large sample properties of both BCF-IV and GRF to obtain possible gains in imperfect compliance scenarios.

References

Angrist, J.D., Imbens, G.W., Rubin, D.B., 1996. Identification of causal effects using instrumental variables. Journal of the American Statistical Association 91, 444–455.

Athey, S., 2018. The impact of machine learning on economics, in: The Economics of Artificial Intelligence: An Agenda. University of Chicago Press.

Athey, S., Imbens, G., 2016. Recursive partitioning for heterogeneous causal effects. Proceedings of the National Academy of Sciences 113, 7353–7360.

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

Athey, S., Tibshirani, J., Wager, S., 2016. Generalized random forests. arXiv preprint, arXiv:1610.01271 .

Bargagli Stoffi, F.J., Gnecco, G., 2018. Estimating heterogeneous causal effects in the presence of irregular assignment mechanisms. In Proceedings of the 5th IEEE Conference in Data Science and Advanced Analytics , 1–10.

Bargagli Stoffi, F.J., Gnecco, G., 2019. Causal tree with instrumental variable: An extension of the causal tree framework to irregular assignment mechanisms. International Journal of Data Science and Analytics DOI: https://doi.org/10.1007/s4106.

Breiman, L., 1984. Classification and regression trees. Routledge, New York, New York.

Breiman, L., 2001. Random forests. Machine Learning 45, 5–32.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and regression trees. Belmont, CA: Wadsworth & Brooks .

Calonico, S., Cattaneo, M.D., Titiunik, R., 2015. rdrobust: An r package for robust nonparametric inference in regression-discontinuity designs. R Journal 7, 38–51.

Chipman, H.A., George, E.I., McCulloch, R.E., et al., 2010. Bart: Bayesian additive regression trees. The Annals of Applied Statistics 4, 266–298.

Coleman, J.S., 1966. Equality of educational opportunity. Washington DC: US Government Printing Office , 1–32.

De Witte, K., Smet, M., D'Inverno, G., 2018. The effect of additional resources for schools with disadvantaged students: Evidence from a conditional efficiency model. Steunpunt Sono Research Report .

Ding, P., Feller, A., Miratrix, L., 2019. Decomposing treatment effect variation. Journal of the American Statistical Association 114, 304–317.

Eberts, R.W., Stone, J.A., 1988. Student achievement in public schools: Do principals make a difference? Economics of Education Review 7, 291–299.

Foster, J.C., Taylor, J.M., Ruberg, S.J., 2011. Subgroup identification from randomized clinical trial data. Statistics in Medicine 30, 2867–2880.

Gentilucci, J.L., Muto, C.C., 2007. Principals' influence on academic achievement: The student perspective. NASSP bulletin 91, 219–236.

Goldhaber, D., Hansen, M., 2010. Using performance on the job to inform teacher tenure decisions. American Economic Review 100, 250–55.

Green, D.P., Kern, H.L., 2012. Modeling heterogeneous treatment effects in survey experiments with bayesian additive regression trees. Public Opinion Quarterly 76, 491–511.

Hahn, J., Todd, P., Van der Klaauw, W., 2001. Identification and estimation of treatment effects with a regression-discontinuity design. Econometrica 69, 201–209.

Hahn, P.R., Carvalho, C.M., Puelz, D., He, J., et al., 2018a. Regularization and confounding in linear regression for treatment effect estimation. Bayesian Analysis 13, 163–182.

Hahn, P.R., Dorie, V., Murray, J.S., 2018b. Atlantic causal inference conference (acic) data analysis challenge 2017.

Hahn, P.R., Murray, J.S., Carvalho, C.M., 2017. Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects. arXiv preprint, arXiv:1706.09523 .

Hanushek, E.A., 2003. The failure of input-based schooling policies. The Economic Journal 113, F64–F98.

Hanushek, E.A., Machin, S.J., Woessmann, L., 2016. Handbook of the Economics of Education. Elsevier.

Hanushek, E.A., Woessmann, L., 2017. School resources and student achievement: A review of cross-country economic research, in: Cognitive abilities and educational outcomes. Springer, pp. 149–171.

Harris, D.N., Sass, T.R., 2011. Teacher training, teacher quality and student achievement. Journal of public economics 95, 798–812.

Hartford, J., Lewis, G., Leyton-Brown, K., Taddy, M., 2016. Counterfactual prediction with deep instrumental variables networks. arXiv preprint, arXiv:1612.09596 .

Hernández, B., Raftery, A.E., Pennington, S.R., Parnell, A.C., 2018. Bayesian additive regression trees using bayesian model averaging. Statistics and Computing 28, 869–890.

Hill, J.L., 2011. Bayesian nonparametric modeling for causal inference. Journal of Computational and Graphical Statistics 20, 217–240.

Holmlund, H., Sund, K., 2008. Is the gender gap in school performance affected by the sex of the teacher? Labour Economics 15, 37–53.

Imbens, G.W., Rubin, D.B., 1997. Estimating outcome distributions for compliers in instrumental variables models. The Review of Economic Studies 64, 555–574.

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

Jackson, C.K., 2018. Does School Spending Matter? The New Literature on an Old Question. Technical Report. National Bureau of Economic Research.

Jackson, C.K., Johnson, R.C., Persico, C., 2015. The effects of school spending on educational and economic outcomes: Evidence from school finance reforms. The Quarterly Journal of Economics 131, 157–218.

Johnson, M., Cao, J., Kang, H., 2019. Detecting heterogeneous treatment effect with instrumental variables. arXiv preprint arXiv:1908.03652 .

Kapelner, A., Bleich, J., 2013. Bartmachine: Machine learning with bayesian additive regression trees. arXiv preprint, arXiv:1312.2171 .

Kitagawa, T., Tetenov, A., 2018. Who should be treated? Empirical welfare maximization methods for treatment choice. Econometrica 86, 591–616.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., Mullainathan, S., 2017. Human decisions and machine predictions. The Quarterly Journal of Economics 133, 237–293.

Van der Laan, M.J., Polley, E.C., Hubbard, A.E., 2007. Super learner. Statistical Applications in Genetics and Molecular Biology 6.

Lechner, M., 2019. Modified causal forests for estimating heterogeneous causal effects. CEPR Discussion Paper No. DP13430 .

Lee, D.S., Lemieux, T., 2010. Regression discontinuity designs in economics. Journal of Economic Literature 48, 281–355.

Lee, K., Small, D.S., Dominici, F., 2018. Discovering effect modification and randomization inference in air pollution studies. arXiv preprint, arXiv:1802.06710 .

Li, F., Mattei, A., Mealli, F., 2015. Evaluating the causal effect of university grants on student dropout: evidence from a regression discontinuity design using principal stratification. The Annals of Applied Statistics , 1906–1931.

Linero, A.R., 2018. Bayesian regression trees for high-dimensional prediction and variable selection. Journal of the American Statistical Association 113, 626–636.

Linero, A.R., Yang, Y., 2018. Bayesian regression tree ensembles that adapt to smoothness and sparsity. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 80, 1087–1110.

Logan, B.R., Sparapani, R., McCulloch, R.E., Laud, P.W., 2019. Decision making and uncertainty quantification for individualized treatments using bayesian additive regression trees. Statistical Methods in Medical Research 28, 1079–1093.

Mattei, A., Mealli, F., 2016. Regression discontinuity designs as local randomized experiments. Observational Studies 66, 156–173.

McCrary, J., 2008. Manipulation of the running variable in the regression discontinuity design: A density test. Journal of Econometrics 142, 698–714.

Mealli, F., Rampichini, C., 2012. Evaluating the effects of university grants by using regression discontinuity designs. Journal of the Royal Statistical Society: Series A (Statistics in Society) 175, 775–798.

Mullainathan, S., Spiess, J., 2017. Machine learning: an applied econometric approach. Journal of Economic Perspectives 31, 87–106.

Murray, J.S., 2017. Log-linear bayesian additive regression trees for categorical and count responses. arXiv preprint, arXiv:1701.01503 .

OECD, 2017. Educational opportunity for all: Overcoming inequality throughout the life course.

Ololube, N.P., 2006. Teachers job satisfaction and motivation for school effectiveness: An assessment. Essays in Education 18, 1–10.

Psacharopoulos, G., Patrinos, H.A., 2018. Returns to investment in education: a decennial review of the global literature. Education Economics 26, 445–458.

Rice, J.K., 2010. The impact of teacher experience: Examining the evidence and policy implications. brief no. 11. National center for analysis of longitudinal data in education research .

Rubin, D.B., 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology 66, 688.

Rubin, D.B., 1978. Bayesian inference for causal effects: The role of randomization. The Annals of Statistics , 34–58.

Sejnowski, T.J., 2018. The deep learning revolution. MIT Press.

Starling, J.E., Murray, J.S., Carvalho, C.M., Bukowski, R., Scott, J.G., 2018a. Functional response regression with funbart: an analysis of patient-specific stillbirth risk. arXiv preprint, arXiv:1805.07656 .

Starling, J.E., Murray, J.S., Carvalho, C.M., Bukowski, R.K., Scott, J.G., 2018b. Bart with targeted smoothing: An analysis of patient-specific stillbirth risk. arXiv preprint arXiv:1805.07656 .

Su, X., Kang, J., Fan, J., Levine, R.A., Yan, X., 2012. Facilitating score and causal inference trees for large observational studies. Journal of Machine Learning Research 13, 2955–2994.

Wager, S., Athey, S., 2018. Estimation and inference of heterogeneous treatment effects using random forests. Journal of the American Statistical Association 113, 1228–1242.

Wang, G., Li, J., Hopp, W.J., 2018. An instrumental variable tree approach for detecting heterogeneous treatment effects in observational studies. Available at SSRN 3045327 .

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

Wooldridge, J.M., 2002. Econometric analysis of cross section and panel data. MIT Press, Cambridge, Massachusetts.

A Discussion on the Instrumental Variables Assumptions

In a typical IV scenario one can express the treatment received as a function of the treatment assigned: $W_i(Z_i)$. This leads to distinguish four sub-populations of units (G_i) (Angrist et al., 1996; Imbens and Rubin, 2015): (i) those that comply with the assignment (*compliers*: $G_i = C : W_i(Z_i = 0) = 0$ and $W_i(Z_i = 1) = 1$); (ii) those that never comply with the assignment (*defiers*: $G_i = D : W_i(Z_i = 0) = 1$ and $W_i(Z_i = 1) = 0$); (iii) those that even if not assigned to the treatment always take it (*always-takers*: $G_i = AT : W_i(Z_i = 0) = 1, W_i(Z_i = 1) = 1$); (iv) those that even if assigned to the treatment never take it (*never-takers*: $G_i = NT : W_i(Z_i = 0) = 0, W_i(Z_i = 1) = 0$). In such a scenario what "one directly gets from the data" is the so-called Intention-To-Treat (ITT_Y):

$$ITT_Y = \mathbb{E}[Y_i|Z_i = 1] - \mathbb{E}[Y_i|Z_i = 0], \tag{29}$$

which is defined as the effect of the intention to treat a unit on the outcome of the same unit. (29) can be written as the weighted average of the intention-to-treat effects across the four sub-populations of compliers, defiers, always-takers and never-takers:

$$ITT_Y = \pi_C ITT_{Y,C} + \pi_D ITT_{Y,D} + \pi_{NT} ITT_{Y,NT} + \pi_{AT} ITT_{Y,AT}, \tag{30}$$

where $ITT_{Y,G}$ is the effect of the treatment assignment on units of type G and π_G is the proportion of units of type G.

ITT_Y does not represent the effect of the treatment itself but just the effect of the assignment to the treatment. If we want to draw proper causal inference in such a scenario we need to invoke the four classical IV assumptions (Angrist et al., 1996):

1. *exclusion restriction*: $Y_i(0) = Y_i(1), \quad \text{for } G_i \in \{AT, NT\}$ where, for each sub-population and $z \in \{0, 1\}$, the shortened notation $Y_i(z)$ is used to denote $Y_i(z, W_i(z))$

2. *monotonicity*: $W_i(1) \geq W_i(0) \rightarrow \pi_D = 0$;

3. *existence of compliers*: $P(W_i(0) < W_i(1)) > 0 \rightarrow \pi_C \neq 0$;

4. *unconfoundedness of the instrument*:

   $Z_i \perp\!\!\!\perp (Y_i(0,0), Y_i(0,1), Y_i(1,0), Y_i(1,1), W_i(0), W_i(1))$.

In our application, these four assumptions are assumed to hold. Let us look at them in detail. The exclusion restriction is assumed to hold since we can reasonably rule out a direct effect of being eligible (around the threshold) on the performance of students. The effect, in this case, can be reasonably assumed to go through the actual reception of additional funding. Monotonicity holds by design: since we are in a one-sided non-compliance scenario there is no possibility for those who are not assigned to the treatment to defy and get the treatment. The same can be said about the existence of compliers. Since the sub-populations of always-takers and defiers can be ruled out by design, this leads to the fact that units receiving the treatment are compliers. Unconfoundedness of the instrument can also reasonably be assumed to hold since observations around the exogenous threshold are as good as if they were randomized to the assigned-to-the-treatment group and the assigned-to-the control group. This holds true especially since we do not observe any manipulation around the threshold and sorting of the units into the treated group.

B Robustness Checks in Monte Carlo Simulations

We introduce some changes in the synthetic models used to test the fit of the BCF-IV (as compared to GRF and HCT-IV). The model from which we start is the simplest model introduced in Section 3: $Y_i = \sum_{p=1}^{k} X_{i,p} + W_i \cdot X_{i,1} + \xi_i + \epsilon_i$ where $X_{i,p} \sim \mathcal{N}(0,1), W_i \sim Bern(0.5), \xi_i \sim \mathcal{N}(0, 0.01)$, $\epsilon_i \sim \mathcal{N}(0,1)$ and $k = 5$. We introduce 5 different variations in this model (each one corresponds to a different design in Table 6):

1. we change the correlation between Z_i and W_i in order to introduce possible weak-instrument problems: we decrease the correlation to $Cor(W_i, Z_i) \in (0.45, 0.55)$ and we do so by introducing in half of the population a very weak instrument $Cor(W_i, Z_i | X_{i,5} < 0) \in (0.35, 0.45)$;

2. we introduce a partial violation in the exclusion restriction;

3. we introduce multiple heterogeneity driving variables (HDVs):

$$Y_i = \sum_{p=1}^{k} X_{i,p} + \sum_{p=1}^{2} (W_i \cdot X_{i,p}) + \xi_i + \epsilon_i, \tag{31}$$

   where this variation is introduced to test if the HDVs are correctly selected even when they are multiple;

4. we change the error distribution, $\epsilon_i \sim \mathcal{U}(0, 1)$, to test if the algorithm is robust to changes in the noise parameter;

5. we manipulate the propensity score function for the BCF-IV, to test if this model is robust to a mis-specification of $\hat{\pi}(x)$.

The results from these five different designs are reported in Table 2. In the presence of a weak-instrument (design 1), the fit of all the three algorithms deteriorates. As we saw in the Monte Carlo simulations in Section 3, GRF is better in identifying the correct HDV but BCF-IV outperforms both GRF and HCT-IV in picking the right threshold and in precisely estimating the causal effect. As we introduce a partial violation of the exclusion restriction (design 2), BCF-IV outperforms the other algorithms with respect to all the dimensions both in the cases with small sample and large sample sizes. When we introduce multiple heterogeneity driving variables, the capacity of correctly estimating the causal effects for GRF deteriorates while BCF-IV outperforms the other algorithms. In the last two designs (design 4 and 5), we again see a trade-off, for the designs with 500 and 1,000 units, between the capacity of correctly identifying the HDV (GRF outperforms the other techniques) and precisely estimating the causal effects (BCF-IV outperforms the other algorithms). In both the latter designs, BCF-IV and GRF get to fairly similar asymptotic results.

Table 2: **Robustness Checks**

		Sample Size								
#Design	Approach	500	1000	5000	500	1000	5000	500	1000	5000
		HDV			Threshold			MSE given HDV		
1	BCF-IV	0.37	0.63	0.83	**0.065**	**0.031**	**0.007**	**0.078**	**0.117**	0.017
	GRF	**0.56**	**0.73**	**1.00**	0.157	0.107	**0.007**	0.230	0.122	**0.011**
	HTC-IV	0.23	0.36	0.73	0.289	0.090	0.065	0.110	0.140	0.022
2	BCF-IV	**0.63**	0.40	**0.77**	**0.022**	**0.061**	**0.002**	**0.005**	0.107	0.043
	GRF	0.43	**0.57**	0.23	0.052	0.067	0.003	0.171	**0.082**	**0.035**
	HTC-IV	0.43	0.37	0.53	0.189	0.170	0.064	0.018	0.102	0.056
3	BCF-IV	0.60	0.76	**1.00**	0.352	0.242	0.021	**0.183**	**0.077**	**0.004**
	GRF	**0.77**	**1.00**	**1.00**	**0.169**	**0.230**	**0.004**	0.776	0.365	0.275
	HTC-IV	0.53	0.63	0.73	0.323	0.289	0.180	0.312	0.116	0.031
4	BCF-IV	0.63	0.80	**1.00**	**0.043**	0.065	0.002	**0.006**	**0.009**	0.002
	GRF	**0.80**	**0.97**	**1.00**	0.068	**0.014**	**0.001**	0.211	0.031	**0.001**
	HTC-IV	0.47	0.40	0.70	0.207	0.103	0.087	0.029	0.014	0.017
5	BCF-IV	0.53	0.63	**1.00**	0.112	**0.020**	0.016	**0.018**	**0.011**	0.007
	GRF	**0.63**	**0.83**	**1.00**	**0.062**	0.069	**0.002**	0.330	0.030	**0.001**
	HTC-IV	0.43	0.40	0.70	0.185	0.188	0.045	0.067	0.051	0.012

Table 2: Results from the robustness checks. HDV refers to the proportion of correctly identified Heterogeneity Driving Variables (HDV); Threshold refers to the mean-squared-error between the true threshold and the predicted one; MSE given HDV refers to the mean-squared-error of prediction for the true causal effects. We highlighted in bold the best results for every round of simulations.

C RDD Checks

In order to check whether or not the RDD (Regression Discontinuity Design) setting is valid, we implement the following checks (Lee and Lemieux, 2010)[30]: (i) we check the balance in the sample

[30] The checks depicted in this Subsection are made on the sample of 50 students introduced in Subsection 4.2.

of units assigned to the treatment just above and below the threshold (this is done to check if the randomization holds); (ii) we examine if there are manipulations in the distribution of schools with respect to the share of disadvantaged students around the threshold, (iii) we employ a formal manipulation test, the McCrary test (McCrary, 2008), to discover potential sorting around the threshold; (iv) we check if there is a discontinuity in the probability of being assigned to the treatment around the threshold. Table 3 shows that the averages of the control variables are not statistically different for the group of units assigned to the treatment and assigned to the control around the cutoff, with the exception of *teacher seniority.* Thus, there is evidence that more senior teachers self-select in schools with lower disadvantaged students. However, as we will show in Section 4.3, this variable does not surface in any model as a heterogeneity driver. We argue that including this variable in our model results in more robust findings. This is due to the fact that our model is robust to *spurious* heterogeneity coming from unbalances in the samples, as shown by Hahn et al. (2017) in randomized and regular assignment mechanisms' scenarios. Moreover, panel (b) of Figure 4 shows the standardized difference in the means for these two groups with the relative standardized confidence intervals. The McCrary manipulation test implemented in Calonico et al. (2015) through a Local-Polynomial Density Estimation leads to the rejection of the null hypothesis of the threshold manipulation[31]. Both these results and the plot of the distribution of schools with respect to the share of disadvantaged students around the threshold in Figure 5 indicate that there is no evidence of manipulation. Finally, Figure 6 shows a clear discontinuity in the probability of being assigned to the treatment around the threshold.

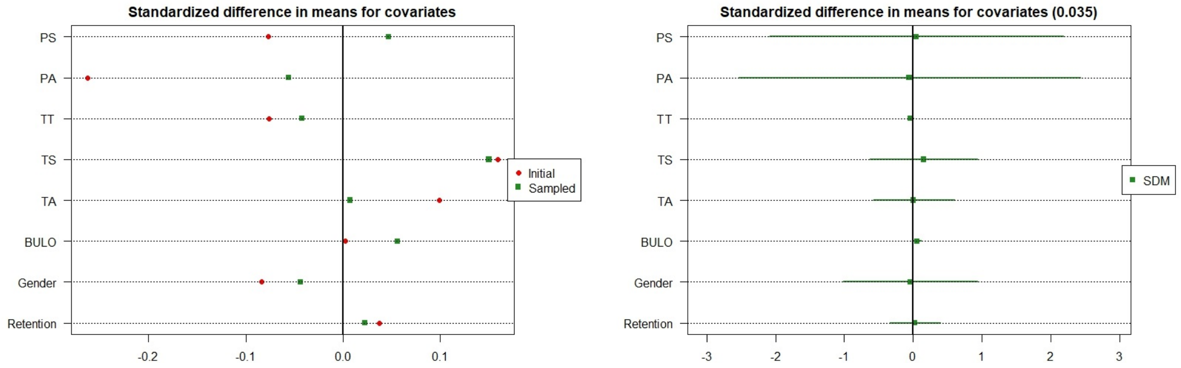
However, as we pointed out in Section 4, schools that are assigned to the treatment actually *receive* the treatment if they satisfy an additional condition of a minimum of six teaching hours. This leads to a fuzzy-regression discontinuity design where the jump in the probability of being assigned to the treatment around the cutoff is not sharp. This scenario is characterized by imperfect compliance.

Students can be sorted, with respect to their compliance status, into two types: (i) students in schools above the cutoff with more than six teaching hours or students in schools below the

[31] The McCrary test leads to a T-value of -0.7497 corresponding to a p-value of 0.4534. The test is performed aggregating the student data at school level.

cutoff (*compliers*: $W_i(Z_i = 1) = 1$ or $W_i(Z_i = 0) = 0$); (ii) students in schools above the cutoff but with less than six teaching hours (*never-takers*: $W_i(Z_i = 1) = 0)^{32}$.

The assignment to the treatment variable (i.e., studying in a school just below or above the cutoff) is a relevant instrumental variable in our scenario (namely, the correlation between Z_i and W_i is roughly 0.62). Moreover, we can reasonably assume both the exogeneity condition and the exclusion restriction to hold in this situation. On one side, since the randomization of the instrument holds there is no reason not to assume conditional independence between the instrument and the unobservables. On the other side, the exclusion restriction seems to hold as well since we can believe that being just below or above the threshold does not affect the performance of students in any way other than through the additional funding. In this imperfect compliance setting, the causal effect of the additional funding on the students' performance can be assessed through the Complier Average Causal Effect in (5). Moreover, using our novel BCF-IV algorithm we can estimate the Conditional Complier Average Causal Effect, (6), to assess the heterogeneity in the causal effects.



a: Balance improvement obtained with sampling. "Initial" refers to the initial sample, while "Sampled" refers to the bootstrapped sample.

b: Standardized difference in means (SDM) and 95% confidence interval around the cutoff with a bandwidth of 3.5%.

Figure 4: The label "PS" refers to Principal Seniority, the label "PA" to Principal Age, the label "TS" to Teacher Seniority, the label "TA" to Teacher Age, the label "TT" to Teacher Training and the label "BULO" refers to students with special needs in primary education.

[32] This a so-called case of one-sided-non-compliance, in which we do not observe any *always-takers* since for those that are sorted out of the assignment to the treatment ($Z_i = 0$) there is no possibility to access the treatment.

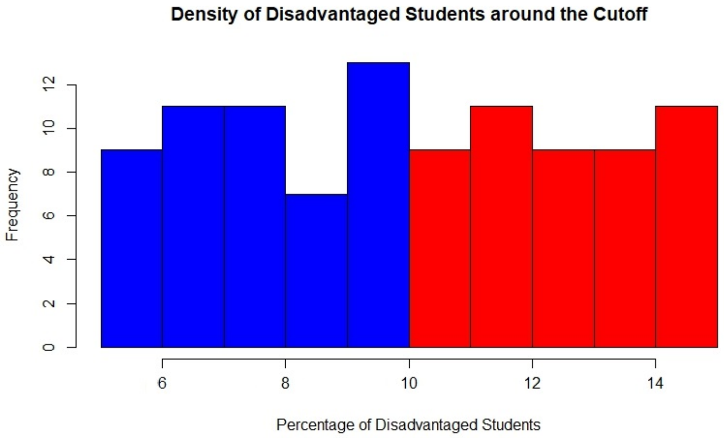


Figure 5: Frequency distribution of disadvantaged students around the threshold (10%). In red the density of the disadvantaged students in the units assigned to the treatment and in blue the density for the units assigned to the control. The densities are aggregated at school level.

	Above Cutoff		Below Cutoff		Full Sample		p-value
Retention	0.036	(0.187)	0.037	(0.189)	0.037	(0.188)	0.913
Gender	0.492	(0.500)	0.471	(0.499)	0.482	(0.500)	0.155
Special Needs	0.000	(0.000)	0.002	(0.044)	0.001	(0.030)	0.045
Teacher Age	4.022	(0.333)	4.024	(0.269)	4.023	(0.304)	0.814
Teacher Seniority	3.867	(0.452)	3.927	(0.342)	3.895	(0.404)	0.000
Teacher Training	0.982	(0.025)	0.981	(0.026)	0.982	(0.026)	0.169
Principal Age	6.022	(1.308)	5.951	(1.229)	5.988	(1.271)	0.067
Principal Seniority	5.778	(1.228)	5.829	(0.935)	5.802	(1.098)	0.120
Observations	2250		2050		4300		

Table 3: Results for 3.5% discontinuity sample with bootstrapped samples of size 50. Standard deviations are in parentheses and the p-value corresponds to a t-test for the difference between the means in the group above and below the threshold.

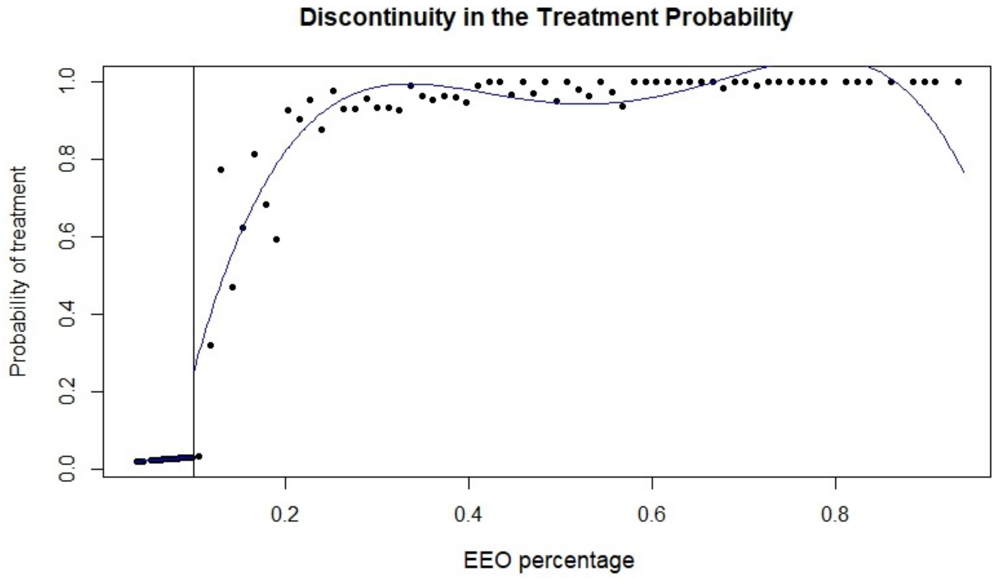


Figure 6: Probability of treatment given the share of disadvantaged students (EEO percentage) in the first stage of secondary education (cutoff 10%).

D Robustness Checks for Policy Evaluation

This Appendix tests the robustness of our models to sampling variations. The sampling variations introduced come from the following two sources: (i) a wider bandwidth around the threshold (changing from 3.5% to 3.7%); (ii) an expansion in the number of sampled units (from 50 up to the lowest number of students per school, which is 62). Moreover, an algorithm which detects the heterogeneity in the ITT effects is applied. To understand if the balance and the results are robust, we manifest the balance in the averages in the samples of units assigned to the treatment and assigned to the control (Tables 4, 5, 6) and the results of the causal effects when we increase the number of units sampled (Figures 7).

In all the different samples the school level characteristics remain widely balanced (with the exception of teacher seniority[33]). *Primary retention* and *Gender* seem to be slightly unbalanced when we widen the bandwidth, this however holds true just in the case where we sample through bootstrap 50 units (*Gender* in this case gets back to a good balance).

[33] This could be due to the fact that less senior principals select themselves in schools with a lower percentage of disadvantaged students.

	Above Cutoff		Below Cutoff		Full Sample		p-value
Gender	0.471	(0.499)	0.493	(0.500)	0.482	(0.499)	0.110
Retention	0.039	(0.194)	0.035	(0.184)	0.037	(0.189)	0.418
Special Needs	0.001	(0.039)	0.000	(0.000)	0.001	(0.027)	0.045
Teacher Age	4.024	(0.269)	4.002	(0.333)	4.023	(0.304)	0.793
Teacher Seniority	3.926	(0.341)	3.867	(0.452)	3.895	(0.404)	0.000
Teacher Training	0.982	(0.025)	0.981	(0.026)	0.982	(0.026)	0.126
Principal Age	5.951	(1.228)	6.002	(1.308)	5.988	(1.271)	0.041
Principal Seniority	5.829	(0.934)	5.777	(1.227)	5.802	(1.097)	0.083
Observations	2790		2542		5332		

Table 4: Results for 3.5% discontinuity sample with bootstrapped samples of size 62. Standard deviations are in parentheses and the p-value corresponds to a t-test for the difference between the means in the group above and below the threshold.

	Above Cutoff		Below Cutoff		Full Sample		p-value
Retention	0.030	(0.170)	0.042	(0.201)	0.036	(0.186)	0.025
Gender	0.497	(0.500)	0.461	(0.499)	0.479	(0.500)	0.015
Special Needs	0.000	(0.021)	0.001	(0.037)	0.001	(0.030)	0.309
Teacher Age	4.022	(0.333)	4.023	(0.260)	4.022	(0.299)	0.955
Teacher Seniority	3.867	(0.452)	3.932	(0.330)	3.899	(0.398)	0.000
Teacher Training	0.982	(0.025)	0.983	(0.026)	0.983	(0.026)	0.805
Principal Age	6.022	(1.308)	6.000	(1.206)	6.011	(1.259)	0.556
Principal Seniority	5.778	(1.228)	5.818	(0.912)	5.798	(1.083)	0.212
Observations	2250		2200		4450		

Table 5: Results for 3.7% discontinuity sample with bootstrapped samples of size 50. Standard deviations are in parentheses and the p-value corresponds to a t-test for the difference between the means in the group above and below the threshold.

	Above Cutoff		Below Cutoff		Full Sample		p-value
Retention	0.029	(0.168)	0.040	(0.196)	0.034	(0.182)	0.026
Gender	0.490	(0.500)	0.464	(0.499)	0.477	(0.500)	0.058
Special Needs	0.000	(0.019)	0.001	(0.038)	0.001	(0.030)	0.174
Teacher Age	4.022	(0.333)	4.023	(0.260)	4.022	(0.299)	0.950
Teacher Seniority	3.867	(0.452)	3.932	(0.330)	3.899	(0.398)	0.000
Teacher Training	0.982	(0.025)	0.983	(0.026)	0.983	(0.026)	0.784
Principal Age	6.022	(1.308)	6.000	(1.206)	6.011	(1.259)	0.512
Principal Seniority	5.778	(1.227)	5.818	(0.912)	5.798	(1.083)	0.165
Observations	2790		2728		5518		

Table 6: Results for 3.7% discontinuity sample with bootstrapped samples of size 62 (the smallest school in the sample). Standard deviations are in parentheses and the p-value corresponds to a t-test for the difference between the means in the group above and below the threshold.

With respect to the results of the BCF-IV algorithm, when we increase the number of sampled units we report the results just for the outcome variable *Progress School*, for which we find evidence of significant treatment effect variation. The main differences between the results for the sample of 50 students (Figure 3) and the ones for the sample of 62 students (Figure 7) are the following: (i) the overall effect is positive (but, again, not statistically significant); (ii) the conditional effects estimated are larger and with a higher statistical significance (which is not surprising given the larger sample).

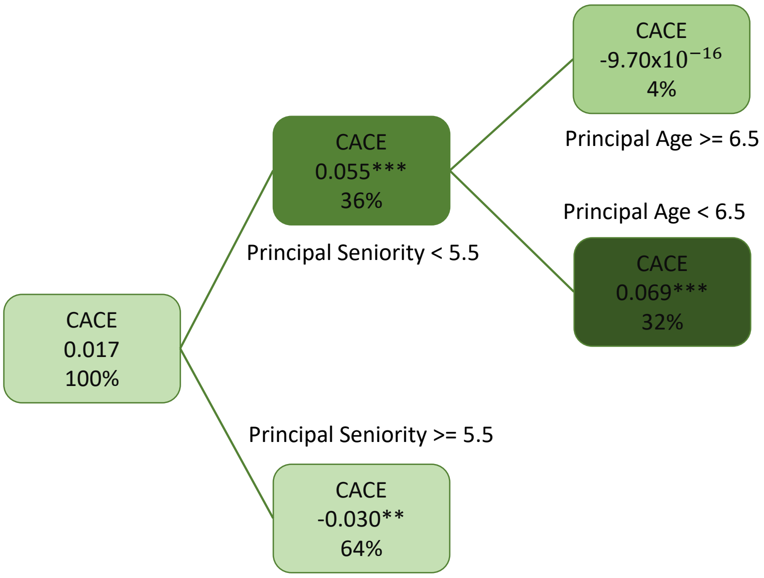


Figure 7: Visualization of the heterogeneous Complier Average Causal Effects (CACE) of additional funding on *Progress School* estimated using the proposed model. The tree is a summarizing classification tree fit to posterior point estimates of individual treatment effects as in Hahn et al. (2017). The significance level is * for a significance level of 0.1, ** for a significance level of 0.05 and *** for a significance level of 0.01.

Figures 8 and 9 depict the results obtained for the estimation of the Intention-To-Treat (ITT) effect using the same set-up used for the estimation of CACE in Figures 2 and 3. Figure 8 depicts the results for the ITT for the *A-certificate* outcome and can be directly compared with Figure 2, while Figure 9 depicts the results for the ITT for the *Progress School* outcome and can be directly compared with Figure 3. The results for CACE and ITT are fairly similar as the complier sub-populations vary between a maximum of 83% of compliers to a minimum of 58% of them. Hence, we can argue that our algorithm catches treatment variations in CACE that are directly driven by the effect variations in the estimated ITT and not by variations in the proportion of compliers within the different subpopulations.

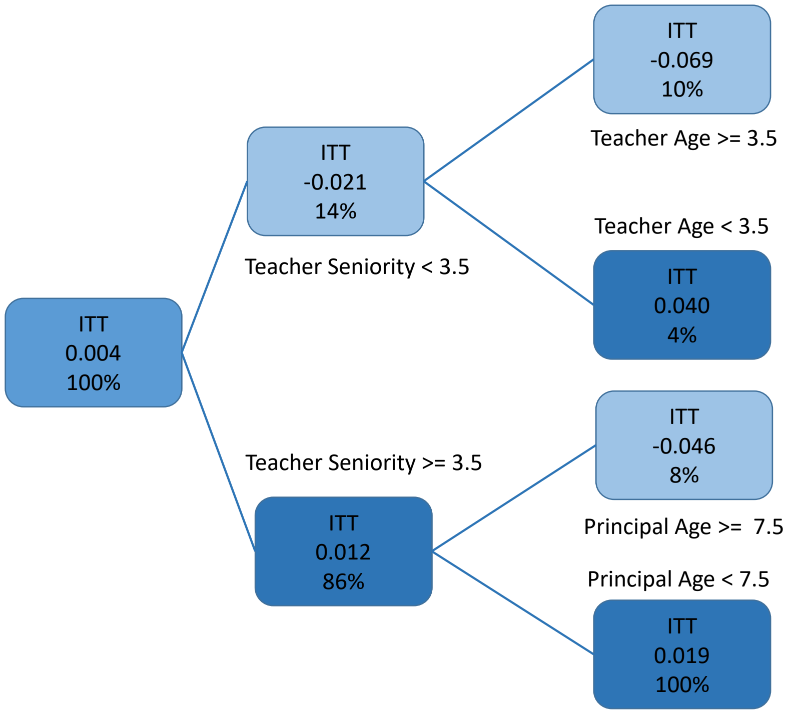


Figure 8: Visualization of the heterogeneous Intention-To-Treat (ITT) effect of additional funding on *A-certificate*. The tree is a summarizing classification tree fit to posterior point estimates of individual treatment effects as in Hahn et al. (2017).

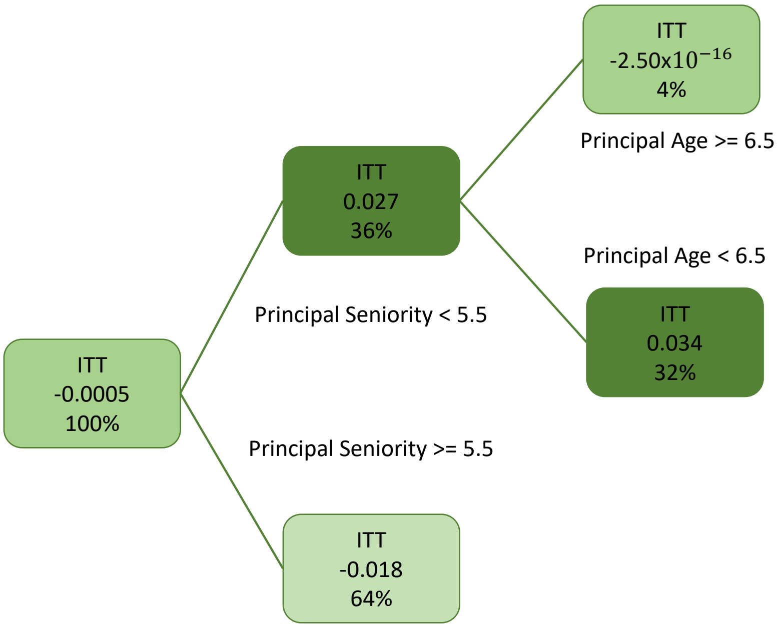


Figure 9: Visualization of the heterogeneous Intention-To-Treat (ITT) effect of additional funding on *Progress School* estimated using the proposed model. The tree is a summarizing classification tree fit to posterior point estimates of individual treatment effects as in Hahn et al. (2017).