De heterogene effecten van het GOK-beleid.

Falco Bargagli Stoffi, Kristof De Witte en Giorgio Gnecco

Promotor: Mike Smet; co-promotor: Kristof De Witte

Research paper SONO

Het Steunpunt Onderwijsonderzoek is een samenwerkingsverband van UGent, KU Leuven, VUB, UA en ArteveldeHogeschool.

Beleidssamenvatting

Inleiding

Sinds 2002 werd in Vlaanderen het Gelijke Onderwijskansen (GOK-)programma geïmplementeerd. Dit GOK-programma voorziet bijkomende lestijden (in het basisonderwijs) of uren-leraar (in het secundair onderwijs) voor scholen met een minimum aandeel leerlingen uit kansengroepen. Deze GOK-uren worden in principe toegekend in cycli van 3 jaar en dienen ingezet te worden op een vooraf bepaald thema.

In eerder onderzoek in de schoot van SONO (De Witte, Smet en Van Assche, 2017; De Witte, D'Inverno en Smet, 2018) toonden we aan dat de GOK-middelen geen effecten hadden op de geëvalueerde onderwijsuitkomsten noch op de efficiëntie van de scholen. Via innovatieve data-gedreven technieken die gebaseerd zijn op 'Machine Learning' gaan we nu dieper in op de heterogeniteit van de effecten, dit wil zeggen op het verschillend effect dat de GOK-middelen hebben voor verschillende subgroepen. We ontwikkelen hiervoor een innovatieve techniek die een oorzakelijke interpretatie van de resultaten toelaat, en die de effecten analyseert voor verschillende subpopulaties. Waar traditionele heterogeniteitsanalyses (zoals toegepast in de eerdere SONO rapporten) kijken naar vooraf bepaalde subpopulaties laten Machine Learning technieken toe om effecten te verkennen en detecteren in endogeen bepaalde subgroepen. Met andere woorden, deze data-gedreven techniek verkent alle mogelijke combinaties van subgroepen en gaat na waar er wel of geen effect is en kwantificeert dit effect. Dit is nuttig voor beleidsanalyses omdat de subgroepen waar een bepaalde interventie effect in sorteert niet noodzakelijk bij voorbaat bekend zijn, en omdat er heel veel combinaties van subgroepen mogelijk zijn waardoor het quasi onmogelijk is om alle potentiële subgroepen met traditionele technieken te verkennen.

De methode

In het bijzonder ontwikkelen we in deze paper een nieuw model dat een uitbereiding vormt op de traditionele Machine Learning technieken. Het model laat toe dat observaties (bijvoorbeeld leerlingen of scholen) niet perfect een vooraf bepaalde regel naleven. In een traditionele beleidsanalyse zijn we geïnteresseerd in het verschil tussen de uitkomsten in een controlegroep en een experimentele groep. In een quasi-experimenteel design kan het onderscheid tussen de observaties in de controlegroep en experimentele groep komen doordat er een extern vastgestelde regel is (in het jargon een 'exogeen criterium') is waardoor de observaties bijna willekeurig aan een controle- of experimentele groep worden toegewezen. In het geval dat deze toewijzing niet perfect is, moeten er bijkomende schattingen gemaakt worden. Traditioneel verloopt dit via een 'instrumentele variabelen' techniek. In deze paper bereiden we de traditionele Machine Learning techniek uit op twee manieren. Ten eerste geven we de resultaten een causale (of oorzakelijke) interpretatie doordat we focussen op de observaties die dicht rond de extern vastgestelde regel liggen. Het is met andere woorden bijna het toeval dat bepaalt of observaties wel of niet de extra middelen krijgen. Ten tweede innoveert de methode doordat het ook niet perfecte toewijzing toelaat. Hiervoor combineren we inzichten uit de literatuur van 'instrumentele variabelen' met 'machine learning'.

Aangezien de voorgestelde methode focust op het vinden van heterogene effecten, en vooral bruikbaar is op grote (administratieve) datasets, kan de methode toegepast worden voor diverse onderzoeksvragen. Om het gebruik te faciliteren werd een routine ontwikkeld in het open-source programma R. Deze routine kan gebruikt worden door onderzoekers of beleidsmedewerkers om de heterogene effecten van interventies te analyseren, en dus om de subpopulaties af te bakenen waar een interventie werkt. De routine is vrij beschikbaar.

Data

Om de heterogene effecten te meten maken we gebruik van bestaande administratieve databanken (Vlaams Ministerie van Onderwijs en Vorming) betreffende het secundair onderwijs. De gegevens zijn beschikbaar op leerlingniveau. Naast de data met betrekking tot de GOK-indicatoren en de uitkomstvariabelen (nl. zittenblijven en behalen van een A-attest), beschikken we ook over informatie over de onderwijscarrière en enkele kenmerken van leerlingen (zoals geslacht, thuistaal, zittenblijven in het lager onderwijs, en of een leerling schoolliep in het buitengewoon lager onderwijs). Op schoolniveau beschikken we over informatie over de leraren (leeftijd, ervaring, bekwaamheidsbewijs) en de schoolleiding (leeftijd en ervaring).

Resultaten eerste graad

In de empirische analyse focussen we allereerst op de eerste graad van het secundair onderwijs. Als scholen minimum 10% GOK-leerlingen in de eerste graad van het secundair onderwijs hebben, kunnen ze aanspraak maken op de GOK-financiering. Bovendien moeten scholen ook minstens 6 GOK-uren genereren. Scholen die een minder hoog aandeel GOK-leerlingen hebben of die minder dan 6 GOK-uren genereren, maken geen aanspraak op de middelen. Net als in vorige rapporten (De Witte et al., 2017 en 2018) focussen we op de observaties rond deze grens van 10% GOK-leerlingen. Op deze manier kunnen we voor de scholen rond de grens oorzakelijke uitspraken doen over het effect van de GOK-uren. Alle analyses worden op leerlingniveau uitgevoerd.

Bij het toepassen van de data-gedreven Machine Learning techniek observeren we, net als in voorgaande rapporten, geen effect op de gemiddelde leerling. Dit bevestigt het eerdere onderzoek met traditionele regressie discontinuïteitsmethode en efficiëntie-analyse. Machine Learning is echter een veel flexibelere techniek, waardoor we nu diepgaander zicht kunnen krijgen op mogelijke heterogene effecten.

We kijken allereerst naar het behalen van een A-attest als uitkomstvariabele. Het aandeel leerlingen in onze steekproef die een A-attest behaalt is 91,73%. Rond de exogene grens is er gemiddeld een positief, maar geen significant, effect van de GOK-middelen op het behalen van een A-attest (in lijn met de bevindingen van De Witte et al., 2017). De heterogeniteitsanalyses laten echter zien dat er differentiële effecten schuilen afhankelijk van het zittenblijven van de leerling in het lager onderwijs: rond de 10% grens zijn de effecten van de GOK-middelen groter voor leerlingen met een geschiedenis van zittenblijven. Hoewel dit effect niet statistisch significant is, toont het aan de GOK-middelen (in afwezigheid van betere uitkomstmaatstaven zoals testscores) vooral effect kunnen hebben bij leerlingen met lagere schoolprestaties. Als tweede drijfveer van de heterogeniteit observeren we verschillen in het effect volgens de gemiddelde leeftijd van het lerarenkorps: rond de 10% grens leiden de GOK-middelen voor leerlingen in scholen met jongere leraren tot het significant meer (7% meer) behalen van een A-attest als deze leerlingen geen verleden hebben van zittenblijven in het lager onderwijs. Ook voor leerlingen met zittenblijven in het basisonderwijs zien we nog steeds een positief (maar geen significant) effect van de GOK-middelen in scholen met jonge leerkrachten. Deze bevindingen raken aan de brede literatuur die het belang aantoont van ervaren leerkrachten (Goldhaber, 2019). In het bijzonder lijken onze resultaten te suggereren dat, rond de 10% grens, de GOK-financiering een positief significant effect heeft als deze wordt toegekend aan scholen met meer jonge leerkrachten. Als we het geobserveerde effect verder uitsplitsen naar de ervaring van de schoolleider, observeren we dat de GOK-middelen vooral effect hebben voor leerlingen zonder zittenblijven in het lager onderwijs, in scholen met jonge leerkrachten én minder ervaren schoolleiders.

Als tweede uitkomstvariabele kijken we naar de effecten op doorstroom zonder zittenblijven. Rond de 10% grens gaat ongeveer 98% van de leerlingen in de twee jaren van de eerste graad over zonder zittenblijven. Ook hier observeren we in lijn met het eerder onderzoek een positief maar niet significant effect op de gemiddelde leerlingen. Wel observeren we rond de grens heterogene (maar niet significante) effecten voor bepaalde leerlingen. In het bijzonder leidt de GOK-financiering tot positieve effecten voor jongens, en negatieve effecten voor meisjes. Dit is deels te wijten aan de grotere kans voor jongens (63%) op zittenblijven. Ten tweede observeren we dat leerlingen in scholen met jonge schoolleiders een positief (maar niet significant) effect van de GOK-financiering ondervinden.

Resultaten tweede en derde graad

Vervolgens focussen we op de tweede en derde graad van het secundair onderwijs. Als scholen minimum 25% GOK-leerlingen in de tweede en derde graad van het secundair onderwijs hebben, kunnen ze aanspraak maken op de GOK-financiering. Bovendien moeten scholen ook minstens 6 GOK-uren genereren. Doordat de 25% grens een stuk hoger ligt dan de grens in de eerste graad kan verwacht worden dat het tweede criterium (6 GOK-uren) minder bindend zal zijn. Hierdoor zal de voorgestelde innovatie in de machine learning techniek (nl. we houden rekening met de niet perfecte toewijzing) minder relevant worden, en komt het model dichter bij een traditionele machine learning benadering. Desalniettemin is het voorgestelde algoritme voldoende flexibel om hier rekening mee te houden.

De resultaten voor de tweede en derde graad suggereren rond de 25% grens geen heterogene effecten voor het behalen van een A-attest als uitkomstvariabele. Het algoritme kan met andere woorden geen subgroepen vinden waar het effect significant sterker of minder sterk wordt. Voor de tweede en derde graad secondair onderwijs vinden we rond de 25% grens wel heterogene effecten voor doorstroom zonder zittenblijven. Waar meisjes in scholen met financiering het minder goed lijken te doen, vinden we vooral een sterk positief heterogeen effect voor jongens zonder zittenblijven.

Belang van professionele ontwikkeling

Hoewel de resultaten inzicht kunnen bieden in potentiële beleidsacties, is de stap van resultaten naar beleid niet eenduidig. Immers, de geobserveerde resultaten zijn waarschijnlijk indicaties van onderliggende factoren en processen, zodat ze vooral richting geven voor verder onderzoek. De heterogene effecten kunnen beleidmakers wel helpen om te verkennen voor welke type scholen de observeerde effecten het hoogst zijn. Zo kan er nagegaan worden waarom rond de 10% grens vooral scholen met minder ervaren leerkrachten en directies baat lijken te hebben bij de bijkomende GOK-uren. Gezien het geobserveerde verband met ervaring kunnen de resultaten ook suggereren om meer middelen te voorzien voor professionele ontwikkeling. Zo gaf eerder SONO onderzoek (Vanblaere, Tuytens en Devos, 2017) reeds aan dat professionele ontwikkeling van leraren, als onderdeel van een breder personeelsbeleid, kan bijdragen tot kwaliteitsvol onderwijs. De resultaten versterken ook het belang aan professionele ontwikkeling van beginnende leraren in de vorm van mentoring, coaching of aanvangsbegeleiding (Van Hoof en Van Peteghem, 2013; Compen, De Witte en Schelfhout, 2019).

Referenties

Compen, B., De Witte, K., & Schelfhout, W. (2018). The role of teacher professional development in financial literacy education: A systematic literature review. *Educational Research Review* 26, 16-31.

De Witte, K., D'Inverno, G., en Smet, M. (2018). The effect of additional resources for schools with disadvantaged students: Evidence from a conditional efficiency model. *Steunpunt Onderwijs Onderzoek*. SONO/2018.OL3.1/1, pp. 160.

De Witte, K., Smet, M. en Van Assche (2017). The impact of additional funds for schools with disadvantaged pupils - A regression discontinuity design. *Steunpunt Onderwijs Onderzoek*. SONO/2017.OL3.1/3, pp. 60.

Goldhaber, D. (2019). SONO Lezing door Dan Goldhaber. *Steunpunt Onderwijs Onderzoek*, April 3, 2019 Leuven.

Van Hoof, J. en Van Petegem, P. (2013). Professionele ontwikkeling en samenwerking van leraren en schoolleiders in Vlaanderen. *Edubron, Universiteit Antwerpen*. Pp. 42.

Vanblaere, B., Tuytens, M. en Devos, G. (2017). Personeelsbeleid in onderwijs: een review van veelvoorkomende HRM-praktijken in scholen. *Steunpunt Onderwijs Onderzoek*. SONO/2017.OL2.3/3, pp. 25.

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

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

September 2019.

Abstract

This paper introduces an innovative Bayesian machine learning algorithm to draw heterogeneous causal effects in the presence of imperfect compliance (e.g., 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 behaviour 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 policy-makers 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 enhance the effectiveness of school funding on students' performance by 3.2 to 3.5 times.

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

JEL Codes: H52; I21; I28

*We are grateful to participants to KU Leuven, Flemish Ministry of Education, Harvard University and IMT Lucca seminars and at the Atlantic Causal Inference Conference (ACIC 2019). We want to thank Giovanna D'Inverno, Steven Groenez, Francesca Dominici, 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 complex 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, etc.). Moreover, following the growing availability of large datasets, the necessity to deal with problems connected with potentially heterogeneous treatment effects is stronger than in the past. The availability of large datasets makes it possible to investigate and, in turn, customize causal effect estimates for populations subsets and even 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 (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. Foster et al. (2011) estimate $\mu(1,x) = \mathbb{E}[Y_i(1)|X_i = x]$ and $\mu(0,x) = \mathbb{E}[Y_i(0)|X_i = x]$ (both using random forests), then calculate $\hat{\tau}_i = \hat{\mu}(1,x) - \hat{\mu}(0,x)$. The machine learning algorithms to estimate $\hat{\tau}_i$ as a function of the units' attributes, X_i, can then be used. 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 the imperfect compliance issues. Nevertheless, in the real world, the implementation of policies or interventions often result 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 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). However, these methods exhibit three principle 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].

To address and accommodate these shortcomings, this paper proposes a novel ensemble Bayesian machine learning algorithm to draw inference on the heterogeneity of causal effects in scenarios with imperfect compliance to an intervention, or so-called irregular assignment mechanisms. This novel algorithm consistently outperforms all the alternative techniques in precisely estimating the causal effects, as shown through Monte Carlo simulations, especially in small samples. Monte Carlo simulations are used to model the probability of different outcomes in a process that cannot easily be predicted due to the intervention of random variables. It is a technique used to understand the impact of risk and uncertainty in prediction and forecasting models. Moreover, the method suggested in this paper performs in a manner similar to other non-Bayesian machine learning techniques such as Generalized Random Forests in discovering the heterogeneity driving variables, eventually converging to an optimal asymptotic behaviour in a simulated scenario. The method we propose contributes the literature with furnishing interpretable results and an easy-to-modify algorithm that allows researchers to incorporate prior knowledge on the distribution of the outcome into the model and to tune the complexity of the algorithm. Using this approach, we can evaluate the heterogeneous impact of an intervention with imperfect compliance, and consequently, target only those observations which benefit most from the intervention (or not target those observations with negative effects). Targeted policies are relevant as the call for personalized interventions has unfurled in all social sciences and in economics in particular (Athey and Imbens, 2017). The main objective of targeted policies studies is to inform policy-makers about the best alloca-

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

tion 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 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). The evaluation of education policies is a promising field for the application of heterogeneous causal effects discovery and, in turn, targeted policies. This is due to, at least, two factors: (i) in this 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.

The 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 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). This technique is 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) and in causal inference settings (Hahn et al., 2018b) by employing a full set of prior distributions on the depth of the trees, on the noise and on the outcome in the leaves. Second, we evaluate the fit of the proposed algorithm by comparing it with two alternative 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 given a discrete or 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, in an empirical application, BCF-IV is used for the evaluation of the effects of additional resources for disadvantaged students on students' performance in a fuzzy regression discontinuity design scenario (Hahn et al., 2001). In particular, 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 certain student can progress to the next year without restrictions; (ii) students' progresses to the following year without grade retention. 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 (135,682 students). We obtained data on student level characteristics 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 significant[2], there is significant heterogeneity in the causal effects: the effects on students' performance are positive and significant if we focus on the sub-population of students in schools with younger teachers (namely, teachers with lower age than the average) and less senior principals (namely, principals with a level of seniority below the average)[3]. These results can inform 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 the additional funding. On the other hand, policy-makers could investigate more in depth why some schools do not benefit from the policy and ultimately provide additional tools to these schools in order to enhance the policy outcomes.

The methodology proposed in this paper can be more widely applied to evaluations of the heterogeneous impact of an intervention in the presence of an irregular assignment mechanism[4]. The first method developed in this field was the Generalized Random Forest algorithm by Athey et al. (2016). Following this contribution Hartford et al. (2016) adapted a deep-learning based algorithm for counterfactual predictions in an IV scenario. The latest contributions by Wang et al. (2017, 2018) and Bargagli Stoffi and Gnecco (2019) built methods tailored for imperfect compliance based on decision tree algorithms. This was done in order to provide more interpretable algorithms and simpler and easier to implement policy rules. The need to provide to policy-makers simple, interpretable and general rules was highlighted also in a recent the contribution by Lee et al. (2018).

The remainder of this paper is organized as follows: in Section 2 we provide a general overview on the causal inference and the applied machine learning frameworks and we introduce our algorithm; in Section 3 we compare the performance of our algorithm with

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

[3] As we show in Section 4.3, these results are in line with the literature of aging on teachers' performance.

[4] An R function, for BCF-IV is available upon request to the corresponding author. The function is built on the R package *BCF* by Hahn et al. (2017).

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 further applications of heterogeneous causal effects discovery and targeted policies in education.

2 Bayesian Instrumental Variable Causal Forest

2.1 Notation

This paper contributes to the literature by developing 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 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: $Y_i(W_i)$. Namely, 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 feature 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 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 because 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, while 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[5]. Z_i can be thought as 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[6] (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

[5] Throughout this Section and in all the paper we assume the instrumental variable to be binary but, one could, in principle, relax this assumption. However, at the moment there are no algorithms available for the estimation of causal effects with continuous treatment variable: we leave this to further research.

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

Causal Effect (CACE), that is:

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

In this paper we introduce a novel 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 focus on the estimation of $\tau^{cace}(x)$ (Athey et al., 2016; Hartford et al., 2016; Wang et al., 2018; Bargagli Stoffi and Gnecco, 2019). In this paper, we rework an algorithm used for the estimation of CATE to fit an irregular assignment mechanism scenario. In particular, the algorithm that we modify is the Bayesian Causal Forest (BCF) algorithm (Hahn et al., 2017). BCF 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) (Friedman et al., 1984)[7].

CART is a widely used algorithm for the construction of binary trees (namely, trees where each node is splitted into only two branches). Figure 1 illustrates how the binary partitioning works in practice in a simple case with just two regressors $x_1 \in [0, 1]$ and $x_2 \in [0, 1]$.

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

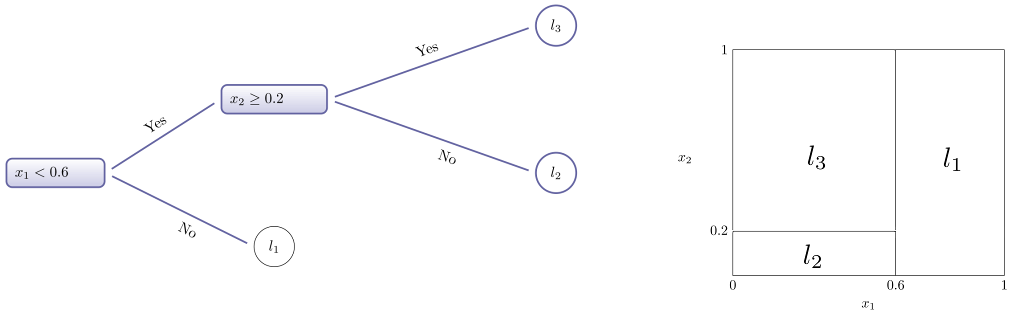


Figure 1: (Left) An example binary tree. The internal nodes are labelled by their splitting rules and the terminal nodes labelled with the corresponding parameters l_i.
(Right) The corresponding partition of the sample space.

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"[8] between the true and estimated values of $Y_i(x)$ within each partition.

The accuracy of binary trees predictions, $\hat{Y}_i(x)$, can be dramatically improved by constructing trees iteratively. 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) used to build each tree (Breiman, 2001). One of the main problems of RF 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 issue, 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

[8] 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.

values of $Y_i(x)$ relies on a fully Bayesian probability model (Kapelner and Bleich, 2013). In particular, the BART algorithm 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 $\mathbb{T}$[9].

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 it works as a regularization device); (ii) the probability distribution of data in the leaves (this prior is aimed at shrinking the leaf 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)[10]. 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 (with probability $\frac{1}{p}$) to build a single tree learner. These Bayesian tools give the researcher the possibility to mitigate the "overfitting" problem of RFs and to tune the algorithm with prior knowledge.

Thus far, the algorithms that we discussed were tailored to find the heterogeneity in the response variable $Y_i(x)$, but are not developed to estimate the heterogeneity in the causal effects. The BCF algorithm (Hahn et al., 2017) is a semi-parametric Bayesian regression model that directly builds on BART, but introduces some significant changes in order to estimate heterogeneous treatment effects in regular assignment mechanisms (even in the presence of strong confounding). The main novelties of this model are the

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

[10] 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 $\beta \in (0,1), \eta \in [0,\infty)$ (the hyper-parameters are generally chosen to be $\eta = 2$ and $\beta = 0.95$); (ii) the prior on the probability distribution in the leaves 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.

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 heterogeneous treatment effects estimation[11] (in the next Subsection we will see in detail these changes). BCF performs dramatically better than other machine learning algorithms for causal inference in the presence of randomized and regular assignment mechanisms, as shown by the results of the Atlantic Causal Inference Conference (ACIC) competition in 2016 and 2017 (Hahn et al., 2018b).

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

What would happen if the assignment mechanism was neither randomized nor regular but irregular? As shown in Bargagli Stoffi and Gnecco (2019), in imperfect compliance settings a naïve application of methods developed for the estimation of heterogeneous causal effects in randomized or regular assignment mechanisms introduces a large bias in the estimation of the heterogeneous causal effects. 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.

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 discuss the two steps in detail in the next Section.

[11] 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 $\hat{\pi}(x)$ is incorrectly specified the results are still widely robust.

2.3.1 Heterogeneity in the Conditional ITT

The BCF-IV algorithm starts from modifying (7) to make it tailored 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 Hahn et al. (2017). 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, we model (9), in a regression scenario, reworking the objective function proposed by Hahn et al. (2017), as:

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

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. The 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), while for the second component the priors are changed in a way that allows for less deep, hence simpler, trees[12].

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 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)[13] and *targeted selection*[14]. In our application, none of these problems is present since the specification of the propensity score for Z_i is known by the researcher (namely, $\hat{\pi}(x) = \pi(x)$). However, 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 misspecified definition of $\hat{\pi}(x)$ does not impact in a significant way the quality of the results as shown in B. This is due to the fact that this first step of our algorithm is not directly about estimating the conditional CACE but is tailored to discover the heterogeneity in $ITT_Y(x)$.

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

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

[13] 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 and X resulting in an additional degree of bias that is not under the researcher's control.

[14] 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.

2.3.2 Estimation of Conditional CACE

Once heterogeneous patterns in the intention-to-treat 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 (6) within all the different sub-populations detected in the previous step.

The conditional CACE can be estimated in a generic sub-sample (i.e., for each $X_i \in \mathbb{X}_j$) as:

$$\hat{\tau}^{cace}(X_i) = \frac{\widehat{ITT}_Y(X_i)}{\hat{\pi}_C(X_i)} \tag{13}$$

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

$$\hat{\pi}_C(X_i) = \frac{1}{N_{1,l}} \sum_{l:X_l \in \mathbb{X}_j} Z_l W_l - \frac{1}{N_{0,l}} \sum_{l:X_l \in \mathbb{X}_j} (1 - Z_l) W_l, \tag{14}$$

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

$$\widehat{ITT}_Y(X_i) = \frac{1}{N_{1,l}} \sum_{l:X_l \in \mathbb{X}_j} Y_l^{obs} \cdot Z_l - \frac{1}{N_{0,l}} \sum_{l:X_l \in \mathbb{X}_j} Y_l^{obs} \cdot (1 - Z_l) \tag{15}$$

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 $X_l \in \mathbb{X}_j$.

To see 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 (namely, this is 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[15]:

$$\hat{\tau}^{cace}(X_{i,p} > 0) = \frac{\widehat{ITT}_Y(X_{i,p} > 0)}{\hat{\pi}_C(X_{i,p} > 0)} \text{ and } \hat{\tau}^{cace}(X_{i,p} \leq 0) = \frac{\widehat{ITT}_Y(X_{i,p} \leq 0)}{\hat{\pi}_C(X_{i,p} \leq 0)}. \tag{16}$$

[15] Alternatively, one can perform a Two Stage Least Squares (TSLS) regression within the different sub-populations. This is our favourite estimation strategy and is the one used both for the simulations and the 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) show that the population versions of (13)-(15) correspond to a Two Stage Least Squares (henceforth, TSLS) 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 TSLS method in every leaf $\mathbb{X}_j$ of the tree $\mathbb{T}$ for the estimation of the effect on the complier population, as it is presented in Imbens and Rubin (1997).

The two simultaneous equations of the TSLS estimator are, in the population,

$$Y_i^{obs} = \alpha + \tau^{cace} \cdot W_i + \epsilon_i, \tag{17}$$

$$W_i = \pi_0 + \pi_C \cdot Z_i + \eta_i, \tag{18}$$

where $\mathbb{E}(\epsilon_i) = \mathbb{E}(\eta_i) = 0$, and $\mathbb{E}(Z_i \eta_i) = 0$[16]. In the econometric terminology, the explanatory variable W_i is *endogenous*, while the IV variable Z_i is *exogenous*.

We can express the TSLS equations, conditional on a sub-population of a leaf $\mathbb{X}_j$, as

$$Y_{i,\mathbb{X}_j}^{obs} = \alpha_{\mathbb{X}_j} + \tau_{\mathbb{X}_j}^{cace} \cdot W_{i,\mathbb{X}_j} + \epsilon_{i,\mathbb{X}_j}, \tag{19}$$

$$W_{i,\mathbb{X}_j} = \pi_{0,\mathbb{X}_j} + \pi_{C,\mathbb{X}_j} \cdot Z_{i,\mathbb{X}_j} + \eta_{i,\mathbb{X}_j}, \tag{20}$$

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 (20) into (19)) holds:

$$\begin{aligned} Y_{i,\mathbb{X}_j}^{obs} &= \Big(\alpha_{\mathbb{X}_j} + \tau_{\mathbb{X}_j}^{cace} \cdot \pi_{0,\mathbb{X}_j}\Big) + \Big(\tau_{\mathbb{X}_j}^{cace} \cdot \pi_{C,\mathbb{X}_j}\Big) \cdot Z_{i,\mathbb{X}_j} + \Big(\epsilon_{i,\mathbb{X}_j} + \tau_{\mathbb{X}_j}^{cace} \cdot \eta_{i,\mathbb{X}_j}\Big) \\ &= \bar{\alpha}_{\mathbb{X}_j} + \gamma_{\mathbb{X}_j} \cdot Z_{i,\mathbb{X}_j} + \psi_{i,\mathbb{X}_j}. \end{aligned} \tag{21}$$

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 (20) and (21)

[16] The latter comes from the fact that (18) is assumed to represent the linear projection of W_i onto Z_i.

above by least squares, when the observations in each leaf are independent and identically distributed, then obtaining an estimate of the parameter $\tau^{cace}_{\mathbb{X}_j}$ in (19). In particular, for every element X_i of a leaf $\mathbb{X}_j$, one can estimate $\tau^{CACE}(X_i) = \tau^{cace}_{\mathbb{X}_j}$ through TSLS, as the following ratio (Imbens and Rubin, 2015):

$$\hat{\tau}^{CACE}(X_i) \equiv \hat{\tau}^{TSLS}_{\mathbb{X}_j} = \frac{\hat{\gamma}_{\mathbb{X}_j}}{\hat{\pi}_{C,\mathbb{X}_j}}. \tag{22}$$

The TSLS estimator associated with (19)-(21) satisfies the next properties. They can be proved likewise in the application of TSLS to the population case (see, e.g., Imbens and Rubin (2015)).

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

$$\hat{\tau}^{TSLS}_{\mathbb{X}_j} - \tau_{\mathbb{X}_j} \ \xrightarrow{p}\ 0 \ \text{ as } \ N_{\mathbb{X}_j} \to \infty, \tag{23}$$

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

Theorem 2: Asymptotic Normality of the TSLS Estimator.
Let both Assumptions 1 and 2 hold. Then

$$\sqrt{N_{\mathbb{X}_j}}\big(\hat{\tau}^{TSLS}_{\mathbb{X}_j} - \tau_{\mathbb{X}_j}\big) \ \xrightarrow{d}\ \mathcal{N}\big(0, N_{\mathbb{X}_j} \cdot avar(\hat{\tau}^{TSLS}_{\mathbb{X}_j})\big) \ \text{ as } \ N_{\mathbb{X}_j} \to \infty, \tag{24}$$

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

The proofs of the two Theorems above directly follow from their unconditional versions[17]. Here we want to stress that in order 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 leaf. Hence, we suggest to perform our algorithm on sufficiently large datasets and to trim those leaves where the number of observation is not large enough.

[17] For further details on these proofs we refer to (Greene, 2003, Chapter 12).

3 Monte Carlo Simulations

To evaluate the performance of the BCF-IV algorithm we compare it, through 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 Generalized Random Forests (GRF) algorithm (Athey et al., 2016). Both the latter algorithms outperform other machine learning algorithms not tailored for irregular assignment mechanisms, so we focus, in this context, just on a comparison within these three algorithms.

Since the main focus of this paper is on the 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 given the right identification of the HDV, and (iii) the mean-squared error for the heterogeneous causal effects given the correct choice of the 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, 1000, 5000. For both the designs the results are aggregated over 30 rounds of simulations.

The results from the simulations are shown in Table 1. As we can see from 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 predicted threshold (clearly, 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. The 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 standardized regressors[18], there is a trade off 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 estimate precisely the causal effect. As the sample size increases, in particular in the scenario with 5 features, both the algorithms get to the same asymptotic results in terms of correct HDV identification and 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. Indeed, the relative gap[19] 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 identified HDV ranges from -10% (in favour of GRF) to 30% (in favour of BCF-IV). Moreover, Figure 2 depicts the distribution of the estimated causal effects over 100 repeated bootstrapped samples for both BCF-IV and

[18] Namely, when the regressors are distributed as a standardized normal distribution.

[19] The formula for the relative gap is, for the MSE of prediction, is 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.

GRF. In both cases the distribution of the estimated effect for BCF-IV is centered closer to the true values (one and two, respectively) than its distribution for GRF. Hence, we claim that the gain in the mean-squared-error of prediction for the causal effect outweights the slower identification of HDVs. This holds true as BCF-IV and GRF converge to a very similar fit, as the sample size increase, with respect to the three dimensions that are 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** [20]

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

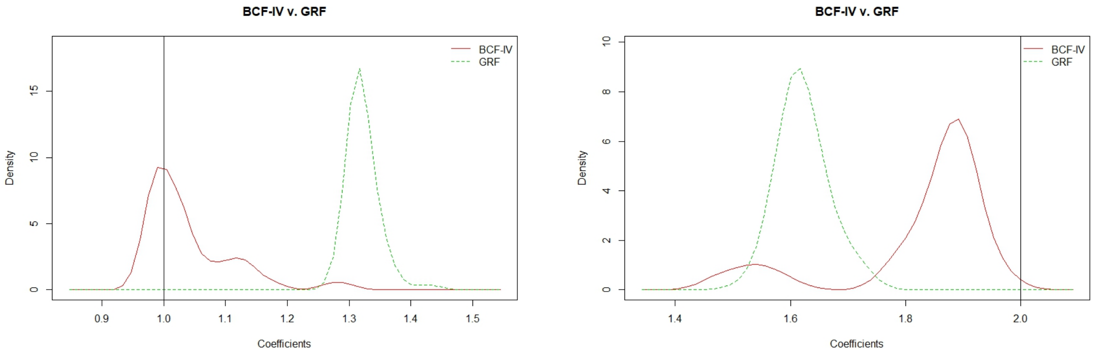
		Sample Size					
#Features	Approach	500	1,000	5,000	500	1,000	5,000
		Design 1			Design 2		
		HDV					
5	BCF-IV	0.57	0.57	**1.00**	0.90	0.96	**1.00**
	GRF	**0.63**	**0.83**	**1.00**	**0.93**	**1.00**	**1.00**
	HCT-IV	0.43	0.40	0.70	0.53	0.56	0.86
10	BCF-IV	**0.30**	0.33	0.73	**0.53**	0.93	**1.00**
	GRF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

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



((a)) The true heterogeneous effect is one

((b)) The true heterogeneous effect is two.

Figure 2: Distribution of the estimated causal effects for both BCF-IV and GRF in 100 bootstrapped samples in the first design.

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)[21]; (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[22]. 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.

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

[22] 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.

4 Heterogeneous Causal Effects of Education Funding

Students' performance has been object of many studies among the last decades, starting from the seminal paper by Beckman (1970). This is due to the fact that there is a wide consensus around the impact of students' performance (and in turn their cultural capital) on job achievements later in life (Pascarella and Terenzini, 1991; Wise, 1975). 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 2002, the Flemish Ministry of Education promoted the "Equal Educational Opportunities" program (henceforth EEO) to ensure equal educational opportunities to all the students (OECD, 2017).

The EEO program provides additional funding for secondary schools with a significant share of disadvantaged students. Thanks 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; the second condition is that the school has to generate at least six teaching hours. 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 need's student population in primary school in primary school (which serves as a proxy of student's low cognitive skills). The school level characteristics include both teacher characteristics, such as the teachers' age, seniority and education, and 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 a range from 1 to 7, where 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)[23]. Similarly, teacher and principal age 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)[24]. Teachers' education records whether 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

[23] 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.

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

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 highest grade) 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 grade ranking from "A" to "C". Students that get an "A" can progress school without any restriction, while 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 exploit the cutoff around the 10% share of disadvantaged students in the first stage of secondary education. In order to draw proper causal inference, we focus on the observations just around the threshold. 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 ($Z_i = 1$). We determine the optimal bandwidth using the "*rdrobust package*" in R (Calonico et al., 2015). The optimal, bias-corrected bandwidths around the cutoff 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. Moreover, to guarantee an equal representation to all the schools, and 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 in the averages between the observations assigned to the treatment and the observations assigned to the control, as shown in panel (a) of Figure 5. 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 this second scenario.

4.3 Results

This Section assesses the effects of additional funding on students' performances and highlights which are the main drivers of the heterogeneity in causal effects. These analyses are made for both the outcome variables: *A-certificate* and *progress school*[25].

4.3.1 A-Certificate

Starting from the seminal contributions of Coleman (1966) and Hanushek (2003) to more 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 population. In Figure 3, are depicted the heterogeneous Complier Average Causal Effects (CACE) estimated using the proposed model: the darker the shade of blue in the node the higher the causal effect[26][27].

The overall effect of the additional funding is not significant, although positive. This finding is in line with 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., 2017, 2018). Nevertheless, rather than

[25] 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).

[26] 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.

[27] In Figures 3, 4, 8, 9, are depicted the so-called summarizing trees (Hahn et al., 2017). 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.

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 dummy variable *primary retention*: the effects of additional funding are larger for students that experienced grade retention during primary school. These effects, even if not significant, show that the effect is slightly higher for students that had a lower performance in the past. 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 4 on a scale from 1 to 8, referring to teachers with less than 40 years old) have a significant increase in their performance if they did not experience any retention in primary school. This effect is positive and significant (although slightly lower: 0.07, meaning that being treated leads to an increase of 7% in the probability of getting the best grade) even if we rule out the conditioning on the students with no primary retention. This heterogeneity driver (namely, the age of the teacher) is particularly significant because there are evidences in the education literature that connect teachers' age to their teaching performance (Young and Place, 1988; Kinney and Smith, 1992) and in turn teaching performance to students' positive achievements (Kosgei et al., 2013). In our scenario, it seems that the additional funding has a positive significant effect when the additional teaching hours are granted to schools with more younger teachers.

Further heterogeneous effects come from the interaction between teacher age and principal seniority. The effect for students without primary retention in schools with younger teachers and with principals with lower levels of seniority (namely, lower or equal than 5 on a 1 to 7 scale, referring to principals with less than 26 years of experience) are significantly higher than the effects on the overall population. Again, this holds true even if we rule out the conditioning on the *primary retention* variable (the effect in this case is 0.17**, meaning that being treated leads to an increase of 17% in the probability of getting the best grade). This sub-population of students that accounts for the 32% and 34% of the overall sample (respectively, when conditioning, or not, on the *primary retention*) shows effects that are between 3.2 and 3.5 times bigger than the overall effect[28]. This evidence can interpreted in the following way: the additional funding has a positive, but not significant, effect in boosting the performance of students in the overall population, but it increases its effect in a significant way for those students in schools with younger teachers and less senior 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., 2017, 2018) and with the literature that connects students' achievements with teaching performance (Young and Place, 1988; Kinney and Smith, 1992) and teaching performance with students' performance (Kosgei et al., 2013). The novel evidence from this research is that also principals play a role: the more senior is the principal, the lower are the causal effects. 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; Kleinberg et al., 2017; Peysakhovich and Naecker, 2017).

These results are policy relevant because they furnish the instruments to policy-makers to enhance the effects of additional funding on students performance. Indeed, on the one side policy-makers could target just students in school with positive, 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 dig in the reason of the lack of effectiveness of the funding in schools with certain characteristics and implement policies to boost the effects of future funding.

[28] 3.2 is the ratio between the conditional treatment effect for the sub-population of students that did not experience primary retention, and that are in schools with younger teachers and principals (0.1651), corresponding to the bluer leaf in Figure 3, and the average treatment effect for the overall population (0.0511). 3.5 is the ratio between the conditional treatment effect for the sub-population of students that are in schools with younger teachers and principals (0.1769) and the average treatment effect on the overall population (0.0511).

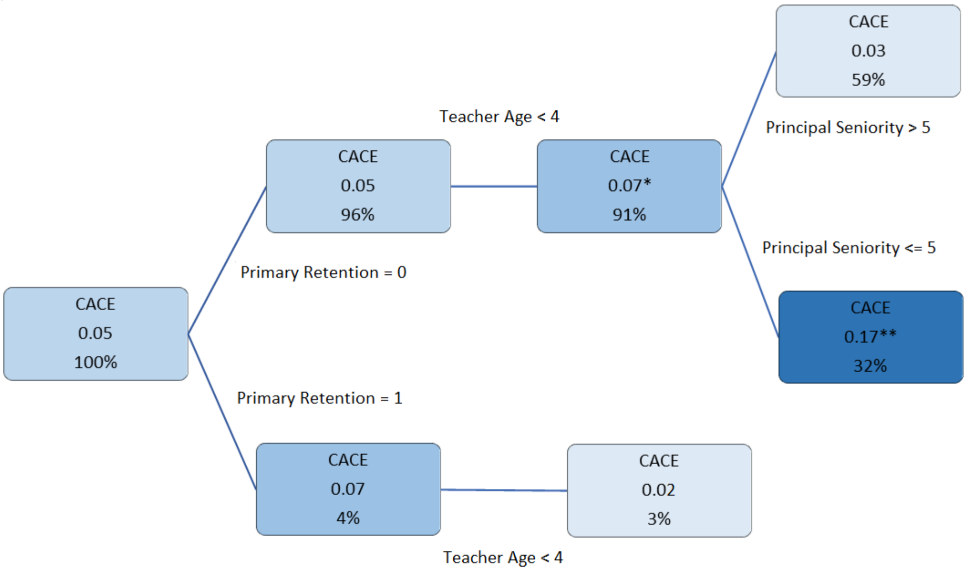


Figure 3: Visualization of the heterogeneous Complier Average Causal Effects (CACE) of additional funding on *A-certificate* estimated using the proposed model. The tree is a summarizing classification tree fit to posterior point estimates of individual treatment effects as in (Hahn et al., 2017). It is important to highlight that the leaves for *Teacher Age* greater than four were trimmed from the tree (both when *Primary Retention* is equal to 0 and to 1) because of small sample size issues. The significance level is * for a significance level of 0.05, ** for a significance level of 0.01 and *** for a significance level of 0.001.

4.3.2 Progress 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 school in the first two years of secondary education. This variable is a proxy for negative achievements for those students that were not able to progress school. Therefore, it is quite interesting to understand if additional funding were effective in driving students away from negative performance. Figure 4 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 positive impact on the chance of progress school for the overall students in the sample[29]. This effect, as well as the heterogeneous causal effects are not significant (again, this is in line with what was found by De Witte et al. (2018) at school level). However, it is interesting to see which are the main drivers of the

[29] However, in the robustness checks this effect is slightly negative. In any case, the overall effect is not significant.

heterogeneity in the causal effect. The first driver is the gender of the student: the effect seems to be positive for male students and negative for female students. This is particularly interesting given the fact that the 63% of the students that do not progress school are males. The second driver of the heterogeneity in CACE is the principal seniority. As in the case of the previous outcome, students in schools with more senior principals (more than 25 years of experience) show lower causal effects. This holds true even when we do not condition on the gender of the students.

Again, 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 gives 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.

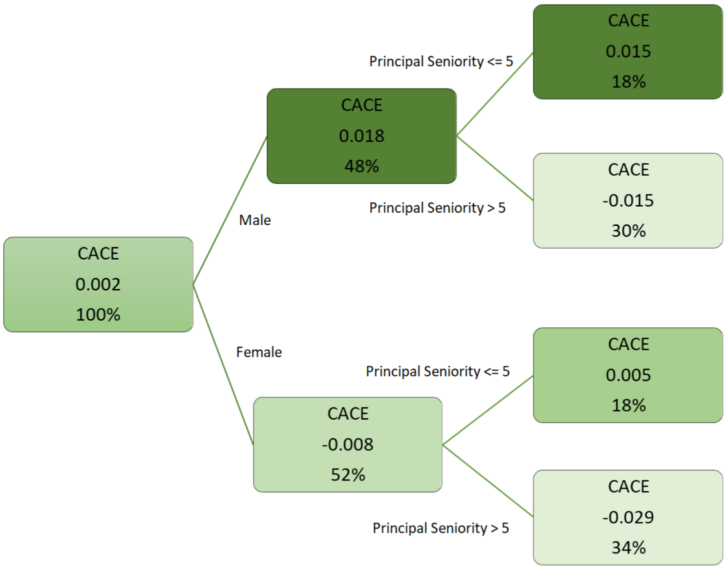


Figure 4: Visualization of the heterogeneous Complier Average Causal Effects (CACE) of additional funding on *Progress School* estimated using the proposed model. The tree is a summarizing classification tree fit to posterior point estimates of individual treatment effects as in (Hahn et al., 2017). The significance level is * for a significance level of 0.05, ** for a significance level of 0.01 and *** for a significance level of 0.001.

5 Conclusion and Discussion

In this paper, we developed a novel Bayesian machine learning technique, BCF-IV, to draw causal inference in scenarios with imperfect compliance. By exploring the heterogeneity in the causal effects, the technique facilitates targeted policies. We show 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 behaviour in identifying the heterogeneity driving variables (HDVs). Moreover, we show that the competitive advantages of using BCF-IV, as compared to GRF or HCT-IV, are large. In particular, 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 help researchers to shed a 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 younger teachers and younger principals (with respect to the average age and seniority, respectively) the effects of the policy are between 3.2 and 3.5 times bigger than the effects on the overall population (in the most conservative scenario) and significant for the *A-certificate* output.

As an underlying mechanism, on the one hand, 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, thus creating relatively more vacancies in disadvantaged schools. Therefore, on average, younger teachers lack a real choice but to start teaching in the most disadvantaged schools. Moreover, thanks to the additional funds, schools can use the funds to reduce class sizes, which might be more effective for younger (and less senior) teachers. 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.

What is worth highlighting in this context is that policy-makers could use the results from this study to target more intensively those students and schools that really benefit from the additional funding in order to increase the welfare effects of the policy and to enhance targeted policies to boost the effects of additional funding for those students that seem to benefit the less.

We leave to further research the extension of these methods to other fields of economic investigation and the development of novel machine learning algorithms for targeted policies and welfare's maximization. In particular, the development of an algorithm that could deal with welfare's maximization in the context of multiple outcomes of interest. 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 (434), 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 (27), 7353–7360.

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

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.

Beckman, L., 1970. Effects of students' performance on teachers' and observers' attributions of causality. Journal of Educational Psychology 61 (1), 76.

Breiman, L., 2001. Random forests. Machine learning 45 (1), 5–32.

Calonico, S., Cattaneo, M. D., Titiunik, R., 2015. rdrobust: An r package for robust nonparametric inference in regression-discontinuity designs. R Journal 7 (1), 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 (1), 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.

De Witte, K., Smet, M., Van Assche, R., 2017. The impact of additional funds for schools with disadvantaged students: a regression discontinuity design. Steunpunt Sono Research Report.

Foster, J. C., Taylor, J. M., Ruberg, S. J., 2011. Subgroup identification from randomized clinical trial data. Statistics in medicine 30 (24), 2867–2880.

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

Green, D. P., Kern, H. L., 2012. Modeling heterogeneous treatment effects in survey experiments with bayesian additive regression trees. Public opinion quarterly 76 (3), 491–511.

Greene, W. H., 2003. Econometric analysis. Pearson Education India.

Hahn, J., Todd, P., Van der Klaauw, W., 2001. Identification and estimation of treatment effects with a regression-discontinuity design. Econometrica 69 (1), 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 (1), 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.

Hanushek, E. A., 2003. The failure of input-based schooling policies. The economic journal 113 (485), 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.

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

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

Imbens, G. W., Rubin, D. B., 1997. Estimating outcome distributions for compliers in instrumental variables models. The Review of Economic Studies 64 (4), 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. Tech. rep., 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 (1), 157–218.

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

Kinney, D. P., Smith, S. P., 1992. Age and teaching performance. The Journal of Higher Education 63 (3), 282–302.

Kitagawa, T., Tetenov, A., 2018. Who should be treated? empirical welfare maximization methods for treatment choice. Econometrica 86 (2), 591–616.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., Mullainathan, S., 2017. Human decisions and machine predictions. The Quarterly Journal of Economics 133 (1), 237–293.

Kosgei, A., Mise, J. K., Odera, O., Ayugi, M. E., 2013. Influence of teacher characteristics on students' academic achievement among secondary schools. Journal of Education and Practice 4 (3), 76–82.

Lechner, M., 2019. Modified causal forests for estimating heterogeneous causal effects.

Lee, D. S., Lemieux, T., 2010. Regression discontinuity designs in economics. Journal of economic literature 48 (2), 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.

McCrary, J., 2008. Manipulation of the running variable in the regression discontinuity design: A density test. Journal of econometrics 142 (2), 698–714.

Mullainathan, S., Spiess, J., 2017. Machine learning: an applied econometric approach. Journal of Economic Perspectives 31 (2), 87–106.

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

Pascarella, E. T., Terenzini, P. T., 1991. How college affects students. Vol. 1991. Jossey-Bass San Francisco.

Peysakhovich, A., Naecker, J., 2017. Using methods from machine learning to evaluate behavioral models of choice under risk and ambiguity. Journal of Economic Behavior & Organization 133, 373–384.

Rubin, D. B., 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of educational Psychology 66 (5), 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.

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 (Oct), 2955–2994.

Van der Laan, M. J., Polley, E. C., Hubbard, A. E., 2007. Super learner. Statistical applications in genetics and molecular biology 6 (1).

Wager, S., Athey, S., 2018. Estimation and inference of heterogeneous treatment effects using random forests. Journal of the American Statistical Association 113 (523), 1228–1242.

Wang, G., Li, J., Hopp, W. J., 2017. Personalized health care outcome analysis of cardiovascular surgical procedures. Available at SSRN 2891517.

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.

Wise, D. A., 1975. Academic achievement and job performance. The American Economic Review, 350–366.

Young, I. P., Place, A. W., 1988. The relationship between age and teaching performance. Journal of Personnel Evaluation in Education 2 (1), 43–52.

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):

$$ITT_Y = \mathbb{E}[Y_i|Z_i = 1] - \mathbb{E}[Y_i|Z_i = 0], \tag{25}$$

which is defined as the effect of the intention to treat a unit on the outcome of the same unit. (25) 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:

$$ITT_Y = \pi_C ITT_{Y,C} + \pi_D ITT_{Y,D} + \pi_{NT} ITT_{Y,NT} + \pi_{AT} ITT_{Y,AT}, \tag{26}$$

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* (expressed in terms of conditional independence notation):
$Z_i \perp\!\!\!\perp (Y_i(0,0), Y_i(0,1), Y_i(1,0), Y_i(1,1), W_i(0), W_i(1))$.

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-population of always-takers 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 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 5):

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):

$$Y_i = \sum_{p=1}^{k} X_{i,p} + \sum_{p=1}^{2} (W_i \cdot X_{i,p}) + \xi_i + \epsilon_i; \tag{27}$$

   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 1. 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 designs, BCF-IV and GRF get to fairly similar asymptotic results.

Table 1: **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 1: 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 bal-

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

ance in the sample of units assigned to the treatment just above and below the threshold (this is done to check if the randomization holds); (ii) we check 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 find out 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 2 shows 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 show up in any model as an 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 5 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 6 indicate that there is no evidence of manipulation. Finally, Figure 7 shows a clear discontinuity in the probability of being assigned to the treatment around the threshold.

However, as we pointed out in the 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.

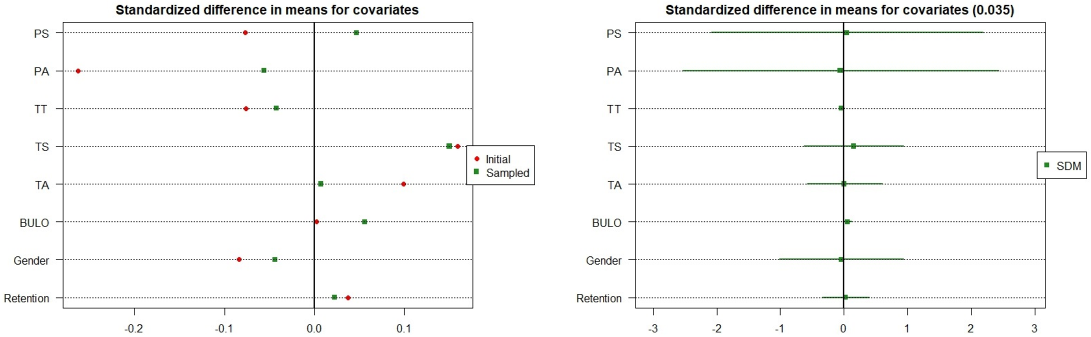
Students can be sorted, with respect to their compliance status, into two types: (i)

[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.

students in schools above the cutoff with more than six teaching hours or students in schools below the 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 scenario. On the 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.

[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.



(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 5: 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.

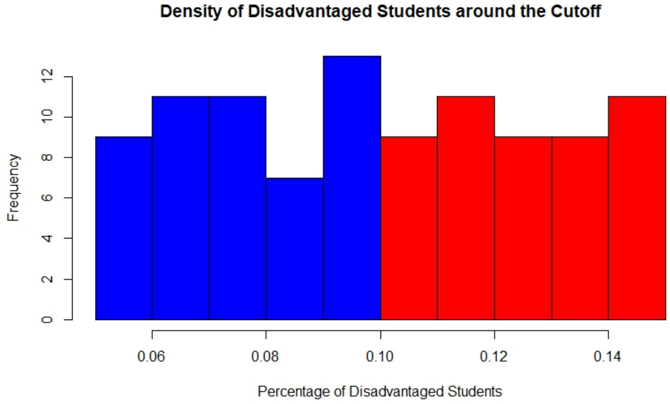


Figure 6: Frequency distribution of disadvantaged students around the cutoff (0.10) with a 0.5 bandwidth. In red the density of the disadvantaged students in the units assigned to the treatment and in blue the density for the units assigned to the control. The densities are aggregated at school level.

	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 2: 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.

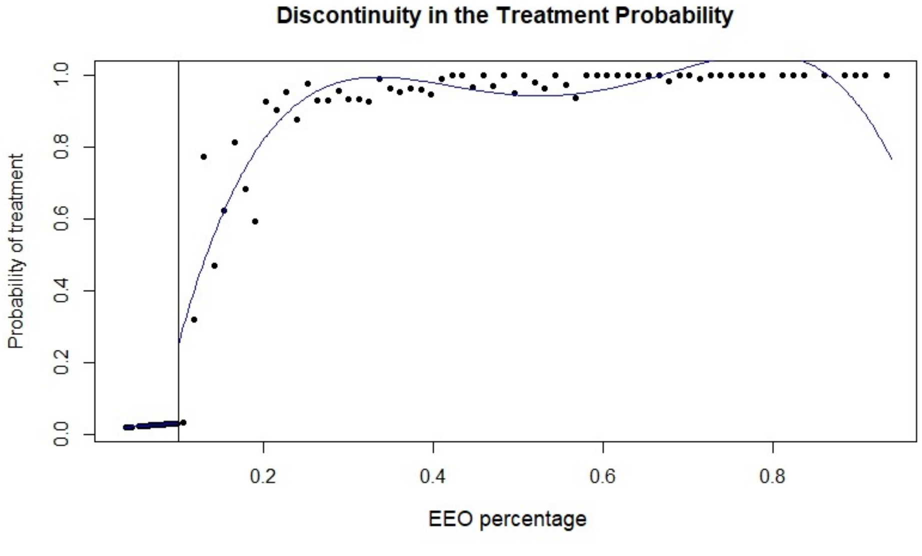


Figure 7: Probability of treatment given the share of disadvantaged students (EEO percentage) in the first stage of secondary education (cutoff 10%).

D Robustness Checks for Policy Evaluation

In this Section of the Appendix, we test the robustness of our models to sampling variations. The sampling variations introduced are of two sources: (i) a wider bandwidth around the threshold (from 3.5% to 3.7%) and (ii) an increase in the number of sampled units (from 50 up to the lowest number of students per school which is 62). To understand if the balance and the results are robust we show the balance in the averages in the samples of units assigned to the treatment and assigned to the control (Tables 3, 4, 5) and the results of the causal effects when we increase the number of units sampled (Figures 8, 9).

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 but this holds true just in the case where we sample through bootstrap 50 units (*Gender* in this case gets back to a good balance).

	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 3: 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.

[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
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 4: 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 5: 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 they remain widely the same when we increase the number of sampled units. There are two slight differences: (i) in the case of the summarizing tree for *A-certificate*, in Figure 8, for units with *Primary retention* equal to one the effect is now negative but still not significant (it is important to notice that these observations account just for the 4% of the overall units); (ii) in the case of the summarizing tree for *Progress School* the overall effect is now negative but still not significant. What is more important is that there are no differences in the heterogeneity drivers and in the direction of the effects (i.e., for more senior principals the effect for both the outcomes are still lower).

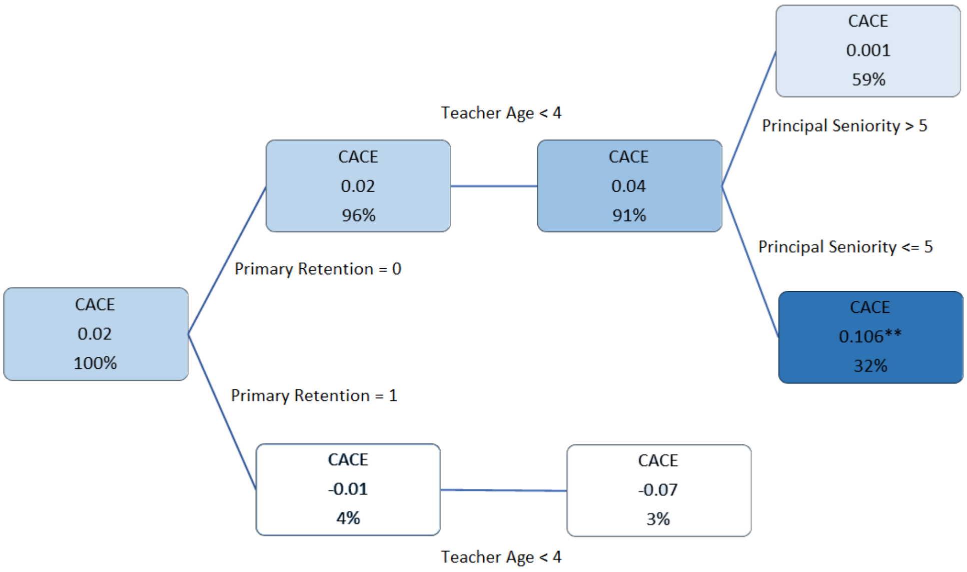


Figure 8: Visualization of the heterogeneous Complier Average Causal Effects (CACE) of additional funding on *A-certificate* estimated using the proposed model. The tree is a summarizing classification tree fit to posterior point estimates of individual treatment effects as in Hahn et al. (2017). The significance level is * for a significance level of 0.05, ** for a significance level of 0.01 and *** for a significance level of 0.001.

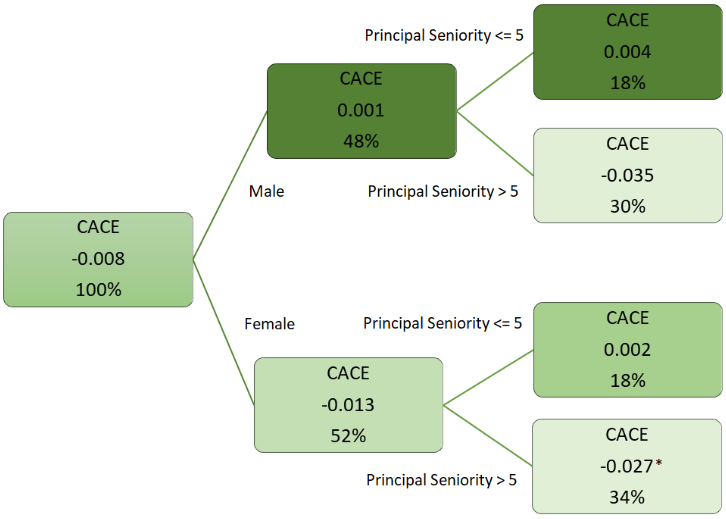


Figure 9: Visualization of the heterogeneous Complier Average Causal Effects (CACE) of additional funding on *Progress School* estimated using the proposed model. The tree is a summarizing classification tree fit to posterior point estimates of individual treatment effects as in Hahn et al. (2017). The significance level is * for a significance level of 0.05, ** for a significance level of 0.01 and *** for a significance level of 0.001.

E Robustness Checks on the full sample

This section explores the robustness of the results on the full sample for the first cycle of secondary education. As such, we do not focus anymore on the students in schools close to the exogenous cutoff point, but include all schools in the analysis. Extending the analysis to a wider set of schools comes at the cost of the interpretation: we cannot interpret the findings as causal anymore as schools far away from the exogenous cutoff point have different observed and unobserved characteristics. It can be shown that the schools which are far away from the threshold, have other observed characteristics than schools close to the threshold. These observed characteristics might drive the results. Therefore, the findings should be interpreted with significant caution and can only be considered as exploratory evidence for future research.

For the outcome variable “A-certificate” in the first cycle of secondary education, the underlying algorithm does not find any heterogeneity for the outcome “A-certificate”. This signals that the average effect size of -0.081 does not correlate with any other variable. Given the absence of heterogeneity, a visual representation is not included in this Appendix.

For the outcomes variable "Progress School" in the first cycle, we observe a more interesting pattern. The results are presented in Figure 10. The results indicate an average, non-significant, influence of the policy on "Progress school". However, there is some heterogeneity by the principal seniority. In particular, more senior principals seem to have lower school progress of their students. It should be noted that this finding should be interpreted with sufficient caution, as we now consider the full sample. It is very likely that the more senior principals are selecting themselves in the schools with more advantaged students, such that this finding is likely to be an endogenous estimate hiding unobserved heterogeneity.

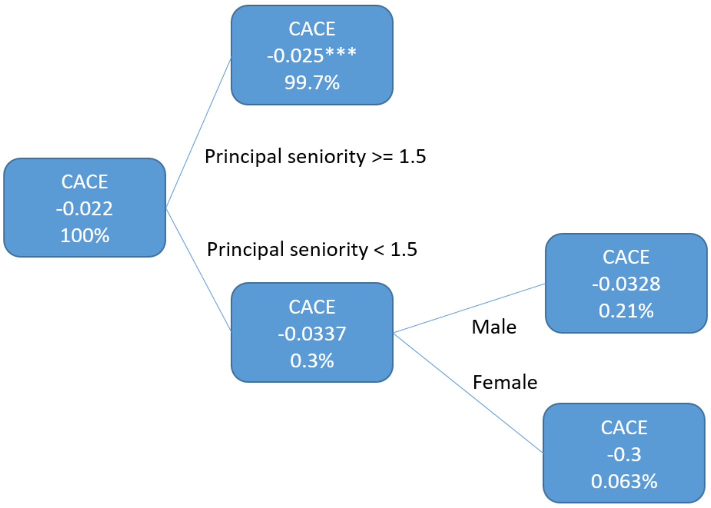


Figure 10: Visualization of the heterogeneous Complier Average Causal Effects (CACE) of additional funding on *Progress School* estimated using the proposed model on the full sample of students in the first cycle of secondary education. The tree is a summarizing classification tree fit to posterior point estimates of individual treatment effects as in Hahn et al. (2017). The significance level is * for a significance level of 0.05, ** for a significance level of 0.01 and *** for a significance level of 0.001.

F Robustness Checks on the second and third cycle

This section presents the robustness test on the second and third cycle of secondary education. In the first cycle (i.e., the first two years) of secondary education funding is provided to schools with more than 10 percent of disadvantaged students. This contrasts to the second and third cycle of secondary education, where the exogenous threshold is set at 25 percent. As the methodology suggested in this paper introduces a procedure to include imperfect compliance, it can be expected that the instrumental variables part of the suggested routine becomes less relevant as the second eligibility criterion (i.e., generate at least 6 teaching hours) becomes less an issue for the schools. As such, the estimation resembles closer to the traditional BCF-procedure, rather than to the newly proposed BCF-IV procedure. Nevertheless, the underlying R-code is sufficiently flexible such that it is accurate also in the case of full compliance.

First, consider the "A certificate" as an outcome variable. It is clear that schools far away from the 25 percent threshold are different in their observed and unobserved characteristics such that the analysis becomes highly endogenous, and that the results can be misleading. Therefore, to allow for a causal interpretation of the results, we focus on the schools close to the 25 percent threshold and estimated the optimal bandwidth using the procedure of Calonico et al. (2015). As in the full sample analysis of the first stage, we only observe an average, non-significant treatment effect of -0.052. The underlying algorithm does not find any heterogeneity for the outcome "A-certificate" in the second and third stage. This signals that the average effect size does not correlate with any other variable. Given the absence of heterogeneity, a visual representation is not included in this Appendix.

Second, consider the progress through school in the second and third cycle of secondary education. The results are presented in Figure 11. Although they suggest a not statistically significant average treatment effect, we observe interesting heterogenous average causal effects (CACE) in the nodes. In a first node, the suggested alogoritm splits the sample between the males and the females. While we do not observed a significant CACE for the males, a small statistically significant negative effect is observed for females suggesting that in schools with more funding females progress slower through school. This negative effect is offset by the positive effect for males with no grade retention in primary education. For this group, we observe a significant positive impact of the funding. Future research could use these heterogeneous results to analyse the underlying mechanisms causing the effects.

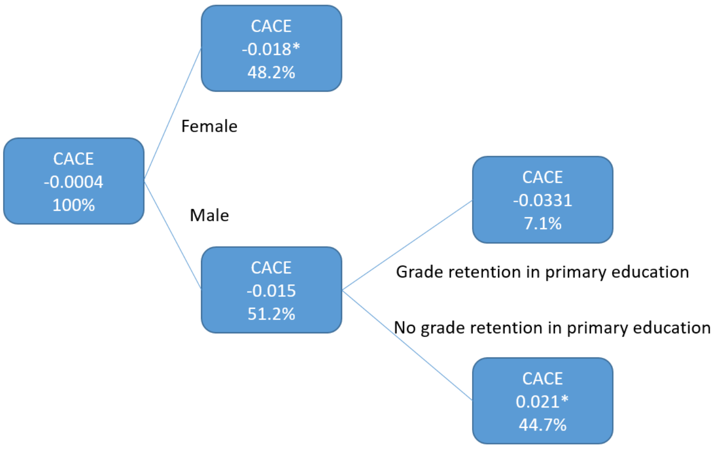


Figure 11: Visualization of the heterogeneous Complier Average Causal Effects (CACE) of additional funding on *Progress School* estimated using the proposed model on the sample of students in the second and third cycle of secondary education close to the 25 percent threshold. The tree is a summarizing 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 0.05, ** for a significance level of 0.01 and *** for a significance level of 0.001.