The heterogeneous effect of dropping out for higher education students : the French case

Gaspard Tissandier - Université Paris 1 Panthéon Sorbonne *

20/05/2022

Abstract:

For higher education system with general and vocational degrees, the question of how to allow resources for dropout policy is fundamental. In France, the university concentrates most of the focus and resources compared to the other vocational track. I use this setting to estimate the heterogeneous effect of dropping out on labor market outcomes (rate of employment and wages), conditional on the followed degree. I use a causal Random Forest methodology in order to account for the heterogeneous students composition of these degrees, with the distance to the closest higher education institution at 6th grade as an instrument for dropping out. I find that using 2SLS lead to underestimate the overall effect of dropping by 9 percentage points for the rate of employment, and by 4 percentage points for the average wage. Vocational degrees dropouts are actually more penalized than university dropouts on their average wage, but not on their time in employment. Finally, using a multidimensional categorization of students can be beneficial for creating targeted dropout policy.

JEL code : J01, J24, I2, I24

Keywords : dropout, higher education, generalized random forest, instrumental variable, labor market outcomes

*I am thankful to Pierre Kopp, Marc Arthur Diaye, and Carmen Aina for their helpful reading and remarks. I also thank the participants of the JMA 2022, IWAEE 2022 and Ifo Dresden WLE 2022 for their comments and useful remarks

1 Introduction

For a large part of the workforce, the higher education is a crucial phase for accumulating competencies, knowledge, and skills that will be later valued in the labor market. While it is decisive for students to acquire diplomas to testify to their abilities, dropping out is a recurrent event in higher education systems, whether voluntary or involuntary (Aina et al., 2018). Dropping out has a strong effect on labor market outcomes and can penalize new entrants in the long run, especially when the dropout happens at the beginning of the higher education period (Schnepf, 2014). Thus, the targeting of dropout policy is essential to ensure their effectiveness, especially if those policies are applied at a national or state scale.

Most education systems propose two type of degrees : vocational and general degrees, spanning from secondary education (Agarwal et al., 2021) to higher education (Powell & Solga, 2010). These tracks differ in their structure and objective, as the former's degrees are shorter and aim at specializing students for a precise type of job, while the latter's are longer and more general. They also differ in the characteristics of students attending them, either demographically, socially or geographically (Hanushek et al., 2017). Because the public and the dropout rate are heterogeneous from a degree to another, these multiple tracks institutional settings create a trade-off in the allocation of resources dedicated to higher education dropouts, where policy makers want to target individuals who benefit the most from the engaged resources. As reduce earnings and employment duration following a dropout can have long lasting impact on a former student's career, the variables of choice for measuring the dropout penalty and targeting the right populations are labor market outcomes such as wages and employment rate (Brodaty et al., 2008; Heigle & Pfeiffer, 2019; McNamara, 2020).

The heterogeneous responses to higher education policies are a growing concern in the economic literature (Belskaya et al., 2020), and the wide array of dropout determinants (Aina et al., 2018, 2022; Behr et al., 2020) as well as the heterogeneous composition of both degrees indicate that we need to account for high dimension interactions between students and degrees characteristics in order not only to measure the dropout propensity, but also to estimate the heterogeneous effect of dropping out on labor market outcomes.

To answer this dropout policy resources allocation trade-off, this paper proposes an evaluation of the dropout penalty conditional on a large vector of individuals characteristics in order to account for the varying composition of vocational and general higher education degrees. Using the french higher education setting, I estimate the heterogeneous effects of dropping out from the general or vocational degrees on labor market outcomes (wage and rate of employment) and then explore the structure of these effects conditional on the diploma and socio-economic status of the parents in order to propose a better targeting of dropout policies.

Dropouts are numerous in the French higher education system: in 2018, 23.9% of students enrolled in their first year of higher education dropped out. This phenomenon is persistent through time as 4.1% of students who began their study in 2014 dropped out at the end of

the second year, and 10.4% at the end of the third year[1].The French higher education system proposes two main tracks after the high school diploma (the *Baccalauréat*): the vocational (or technical) path, from which students obtain a BTS or a DUT degree, is labeled STS/IUT, and the general path, from which students obtain a Licence degree, mostly done at the University[2]. The first one is dedicated to training students for technical jobs, and offers mostly two years degrees, while the University path proposes three years general degrees, and can bring students to the Master's level. As the 2007 national policy against dropouts shows, most of the efforts and funds are concentrated on the University (or general track) dropouts, as a large part of the Licence students exit this degree before graduating. (Morlaix & Perret, 2013). Moreover, the STS/IUT (or vocational) degrees are historically known to have less difficulty to integrates the labor market in case of dropout, mainly because of the technical nature of the jobs the students are trained for. However, recent studies show that since 2007, this statistical fact does not seem to hold anymore, and so policies regarding dropout must be re-thinked to account for the actual structure of the effect of dropping out conditional on the degree of the students and other socio-demographic characteristics (Ménard, 2018) [3].

According to both fundamental economic models of education (Becker, 1993 or Spence, 1973), acquiring more year of education or degrees bring return on education, as detailed in the review of Oreopoulos and Petronijevic, 2013 for the American context, or Psacharopoulos and Patrinos (2018) for a more global scope. The structure of the return on education for the french context is detailed in the work of Courtioux et al., 2014, where the authors develop a micro-simulation methodology to model the distribution of return on education for various degrees and individual characteristics.

The literature about potential impact of dropping out on academic or labor market outcomes for higher education students is organized split in two parts, depending on the adopted comparison. The first group compares higher education dropouts with students who didn't engage in a higher education degree, and aims at measuring the premium of acquiring human capital through post-secondary education, even without graduating (Hällsten, 2017; Heigle & Pfeiffer, 2019; McNamara, 2020; Schnepf, 2014). The second type of literature identify the effect of not completing a degree, by comparing individuals engaged in the same diploma, but without the same completion status (Brodaty et al., 2008; Matkovic and Kogan, 2012; McNamara, 2020; Scholten and Tieben, 2017, and Bjerk, 2012 on high school dropouts). This paper, by studying the dropout penalty across degrees and comparing them in order to find students who could benefit more from dropout policy, fits in the second group.

The evidence concerning the effect of dropping out on labor market outcomes are mostly

[1]Repères et références Statistiques 2019 - Direction de l'évaluation de la prospective et de la performance

[2]BTS stands for Brevet de technicien supérieur, DUT for Diplôme universitaire de technologie and the Licence is the equivalent of a Bachelor

[3]Merlin Fanette, Le « décrochage » en STS : l'autre échec dans l'enseignement supérieur, Céreq Bref, n° 366, 2018, 4 p.

positive for works comparing higher education dropouts' outcomes with high school graduate, and negative for papers comparing individuals who dropped out of a degree with graduates. Schnepf (2014) find that most European higher education students benefits from following their degrees compared to high school graduates, measured on the rate of employment. McNamara (2020) finds the same positive effect for higher education dropouts on employment level and wage on the United Kingdom case, as well as Hällsten (2017) for Swedish higher education dropouts and Heigle and Pfeiffer (2019) in the German context. Regarding the second part of the literature, most higher education dropouts do worse than same degree graduates. For amercian high school dropouts, Bjerk (2012) finds a negative effect of dropping out on the employment rate and the criminal activity propensity. Brodaty et al. (2008) find the same results on the wage and the employment rate for the french higher education dropouts, by using an instrumental variable setting. Finally, Matkovic and Kogan (2012) find mixed results on Croatia and Serbia, mostly due to the varying economics context of these countries.

The literature dedicated to dropout proposes many methodology to account for potential heterogeneous effect of dropping out, either in the propensity or the effect on academic and labor market outcomes. It has been shown that the social origin and individual characteristics of students are highly determinant in the dropout process (Aina et al., 2018, Vignoles and Powdthavee, 2009), and that specific variables such as the academic path or gender have an impact on the structure of the effect of dropout. Gury (2011) emphasises the heterogeneous probability of dropping out in the french higher education, especially conditional on socio-economic background of the students. This heterogeneity is more pronounced at the beginning of the university or vocational track path. Regarding the heterogeneous effect of dropping out on labor market outcomes, Brodaty et al. (2008) propose a linear estimation of the dropout effect conditional on the followed degrees. McNamara (2020) use a double machine learning approach to estimate the dropout premium (compared on high school graduate) conditional on the gender, and find heterogeneity in the occupational status but an overall negative effect of dropping out for both genders.

The main issue in estimating the effect of dropping out or having delay in graduation is the endogeneity of the event with the underlying ability of the student: following Spence, 1973, dropping out sends a negative signal to the labor market about the ability of the individual, where ability is defined as an underlying variable reflecting the capacity of a worker to perform well in a task, or a set of tasks. Propensity score matching can be used to solve this issue, as in Schnepf, 2014. On the other hand, recent papers like Mahjoub, 2017 use the period of birth as an instrument, inspired by Angrist and Krueger, 1991. An alternative instrument is the distance to the closest higher education institution, as proposed by Card, 1993. In Brodaty et al., 2008, the authors use a dense system of geographical IV with the distance to the closest university in 6th grade, and the number of openings of higher education institutions in the geographical area during secondary education.

My paper, by using a causal machine learning approach, takes a similar stance as McNa-

mara (2020) in considering the propensity and the effect of dropping out as high-dimensional phenomenons, but differs on many points. First, the objective is to provide a evaluation framework dedicated to policy maker in dual-track higher education system, and not an estimation of the effect of dropping out from higher education. Second, this paper uses causal random forest, which allow for non-linear high-dimension interaction to model the relationship between individuals characteristics and the dropout propensity, and between the dropout and the effects on labor market outcomes. This methodology is particularly adapted when considering the dropout as an event concerning very diverse individuals, either in the socio-demographic characteristics or their chosen degrees. Finally, I use an instrumental variable setting in order to provide unbiased estimation of the dropout penalty for various degrees.

To allow the estimation of heterogeneous treatment effect, I apply the Generalized Random Forest (GRF) method, developed by Athey et al., 2019, on a French database of 12000 young workers who finished their education in 2010. Their work records are surveyed from 2010 to 2013, which helps to construct two indicators of the average wages and the time in employment for every individual. The database gathers individuals who left the education system at the same time (in 2010), regardless of their educational advancements, allowing to compare individuals with different degrees and human capital, but exposed to the same labor market conditions. This setting helps to eliminate the effect of the labor market cyclical structure on outcomes, to focus on the comparison of individuals who dropped out or not, without conjectural effect in the estimation. To explore the highly dimensional structure of the effect of dropping out on labor market outcomes, I use the GRF algorithm, based on the Random Forest structure (Breiman, 2001), to estimate individual Conditional Average Treatment Effect (CATE). The CATE are however not in themselves indicative of the average treatment effect of dropping out for a group of individuals, as they are an individualized treatment effect. Thus, the analysis relies on the Average (Conditional) Local Average Treatment Effect, estimated on multiple sub-samples (students from general and vocational degrees, or coming from different socio-economic background) to understand the heterogeneity of dropping out on labor market outcomes conditional on individuals' characteristics. The A(C)LATE is computed per quartiles of CATE effects, per diplomas, per socio-economics status (SES) and parents' diploma, and the interaction of diploma and SES in order to identify sub-populations to target more intensively by dropout policies.

The endogeneity of the dropout process is tackled with an instrumental variable setting adapted to the Random Forest structure of the GRF. I use a second degree polynomial of the distance to the closest Higher Education institution in 6th grade as the instrument. Paired with a vector of controls to estimate the predicted probabilities of dropout and using the three steps methods proposed in (Adams et al., 2009), I obtain an efficient instrumental variable setting to identify heterogeneous causal effects of dropout on labor market outcomes. The distance is measured at 6th grade to avoid the endogeneity of the high school location choice, highly conditional on the student's performance and social origin (Brodaty et al., 2008).

I find that the effect of droppint out is statistically heterogeneous on the whole distribution and conditional on degrees, socio-economic status and the interaction of the both. After splitting the overall sample into quartiles according to individual Condition Average Treatment Effect (CATE), I found that the lower quartile has an effect of -29% in the time in employment, while the higher quartile has an effect of -19%. The difference is wider for the average wage, as the most penalized quartile has an effect of -72% of the average wage on the three years following the dropping out, while the less penalized quartile exhibits a positive effect.

While vocational degree dropouts are less penalized in terms of time in employment, they are far more penalized regarding their average wage, thus questioning the actual distribution of resources allocated to dropout policies in France. Vocational degree dropouts have a penalty of around -24% in time in employment, while it is -28% for general degree dropouts. However, vocational track dropouts have an effect of -36% on the average wage, while University dropouts don't have a significant effect of dropping out after around three years on the labor market.

Finally, considering the socio-economic status of the parents within each degree reveals another layer of heterogeneity, especially for university dropouts. Students from low SES who drop out from vocational track have a negative effect of -40% on their average wage over three years, while these same students don't exhibit any significant effect when they drop out from the university track. Students from high SES backgrounds are more penalized regarding the time in employment when they drop out from the University than from vocational degree, while this effect is almost homogeneous conditional on the SES for vocational degrees dropouts.

This paper sheds new light on the integration of french tertiary education dropouts in the labor market. The main contribution is the application of machine learning techniques that helps to account for individuals' characteristics and unfold the heterogeneous structure of the dropout effect on labor market outcomes. These results allow to allocate more accurately the dropout policies' resources for higher education system with many tracks, and to understand better the path of vocational and general degrees dropouts. Finally, this paper considers the heterogeneous effect of dropping out conditional on high-dimension and non linear interactions in order to account for the diverse background of students attending higher education.

2 Data

To identify the effect of dropping out on former students' labor market outcomes, I use "Génération 2010", a longitudinal survey provided by the CEREQ (Centre d'Etudes et de Recherches sur les Qualifications) [4]. This survey is conducted on individuals who have finished their education in 2010 (between October 2009 and October 2010), without any interruption before. Individuals are surveyed in 2013, three years after they left the educational system. The resulting database consists of a panel gathering information about former students' background, education, and a detailed schedule of employment from 2010 to 2013. The survey covers 33547 individuals with a wide range of education and social background variables and their profes-

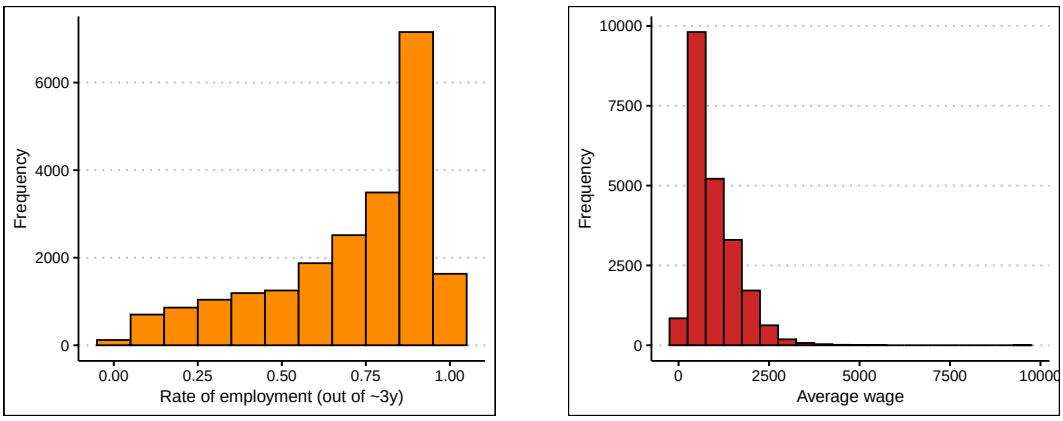
[4]Génération 2010 – Interrogation à 3 ans – 2013 (2013, CEREQ)

sional records. I restrain this data set to individuals who at least, tried to obtain a higher education degree. This subset of individuals goes from high school diploma holders who tried one year of higher education to Ph.D. graduates.

The chosen methodology relies on the common support in dropout probability for every individuals, leading to discard around 42 % of the sample with too low or too high propensity of dropping out. The final data set counts around 12600 observations (see section 4.1), and its descriptive statistics are presented in appendix, section A. Since the first stage is implemented on the initial data set (before the discard step), the descriptive statistics will be presented on the 21829 individuals included in the database, in table A1.

I create indicators variables for dropout, the number of months worked as a rate, and the average wages. Dropping out is defined here as not having validated a diploma in 2010, or exiting the educational system before the last year of said diploma. For example, if a student didn't graduate from her Master 2 because she didn't pass the exams, she will be considered as a dropout. A student who interrupted her study in the second year of undergraduate university degree, out of the three required years will also be considered as a dropout. According to this definition, the database consists of 4923 individuals who dropped out, and 16906 who didn't (23% of dropouts).

Each individual's work curriculum is entered in a side database where employment and unemployment periods are filled in. For each working sequence, the beginning and ending salaries are specified, as the duration in months. The first outcome variable, *Rate of employment* (*roe*), consists in the number of months worked over the period spent on the labor market. I use this definition since not every degree finish at the same time of the year, or student who drop out spend more time on the labor market. The second outcome variable, *Average Wages* (*aw*), is an average of the wages on the whole labor market period observed. Then, if an former student works 12 months with a salary of 1200€ while spending 36 months on the labor market, her *Average Wages* will be equal to 400€. In order to work on percentage differential between individuals, and not percentage points or monetary difference, I use the logarithm of the rate of employment and average wage as dependent variables. The distribution of *roe* and *aw* are presented in figures 1.



(a) Time in Employment (b) Average wages

Figure 1: Distribution of dependent variables

The following variables are used for the estimation : highest diploma tried on 6 levels, if the individuals has done a foreign study travel or an internship during her higher education period, the geographical location in 6th grade, of the higher education establishment, and when the individual left the education system. I also keep the gender of the individual, the professional occupations and diplomas of both parents, and information about past education such as the discretized grade of the high school diploma, the type of high school diploma (general, technical or professional). The descriptive statistics are presented in table A1. For commodity reasons, the social origin is presented only for the highest among the both parents.

	Dropout rate	Frequency	Percentage
Gender			
Male	27.3 %	10092	46.2%
Female	18.5 %	11737	53.8%
Highest diploma tried			
Bac +2 (STS/IUT)	32.3%	10095	46.2%
Bac +3 (university)	21.4%	3262	14.9%
Bac +4 (university)	51.3%	943	4.3%
Bac +5 (university/Grande Ecole)	4.7%	5051	23.1%
PhD	9.9%	2478	11.4%
Parents' highest social category			
Disadvantaged	26.1%	3132	14.3%
Intermediate	25.3%	5718	26.2%
Advantaged	23.2%	4277	19.6%
Highly Advantaged	19.1%	8702	39.9%
Parents' highest diploma			
No diploma	28.2 %	3819	17.5%
Bac or below	25.4%	8119	37.2%
Short degree	19.8%	5915	27.1%
Long degree	15.3%	3976	18.2%
Other			
Foreign trip (= yes)	8.8%	4513	20.7%
Internship (= yes)	15.0%	14717	67.4%

Table 1: Summary statistics by dropout status

While females represent around 54% (and thus the majority) of the sample, they also drop out less than males, with a dropout rate of 18.5%. It has been shown that women tend to undertake more often higher education, and we can observe that female also tend to drop out less in secondary education.

Bac (or *Baccalauréat*) corresponds to the High School Diploma, and is the reference for the time spent in higher education. The time needed to acquire a diploma is counted as "+y" : Bac +2 corresponds to two years of study after the HSD, and correspond here to vocational degree (labeled STS/IUT), which lead to a precise field, and are considered as "professional degrees". The Bac +3, obtained at the university, are general diploma which lead to a broad array of jobs, and are organized around field (such as STEM, law, economics, management). STS/IUT students show a dropout rate of 32% while 21% of Bac +3 students are concerned by dropout. The high dropout level for Bac +4 (Master 1) can be explained by the fact that most student which start a Master's usually undertake the full program, in two years, and not only the first one. The dropout rate in Master 2 is very low, as for most higher education path, it is the last year of studying. Finally, the PhD students present a dropout rate of 11.4%, which is quite high for the longest degree possible.

Concerning parents' occupation, the levels are defined using their socioeconomic status. Disadvantaged social category corresponds to factory worker and unemployed individuals. The intermediate category gathers employee and farmer, the advantaged category gathers intermediary profession, craftsman and independent while the highly advantaged gathers CEO, managers and executives. The parents' diploma are self explanatory : the long degree category gathers parents with 5 years or more of higher education, and short degree those with less than 5 years of higher education. I use the maximum of these variables among the both parents in order to account for the global family environment, and not only the father's or the mother's background. The dropout rate is decreasing with the increase of the parents' highest social category or highest diploma. These results are fully in line with the literature documenting the heterogeneity of the dropout rate among different social origin.

Finally, I base my instrumental variable setting on the distance to the closest higher education institution from the student's 6th-grade city. This distance is computed using the GPS coordinates and the distance between both points on the geodesic[5]. The geographical unit is the *zone d'emploi*, dividing France into around 310 areas. If there is a university or a school in the *zone d'emploi* of the 6th-grade city, the distance is then 0. The density function of this variable is presented in figure 2. I use the square of the distance as an instrument, and the distance between *zone d'emploi* is computed using the centroid of those areas.

[5]For computation methodology, see : C.F.F. Karney, 2013. Algorithms for geodesics, J. Geodesy 87: 43-55. doi: 10.1007/s00190-012-0578-z. Addenda: https://geographiclib.sourceforge.io/geod-addenda.html. Also, see https://geographiclib.sourceforge.io/

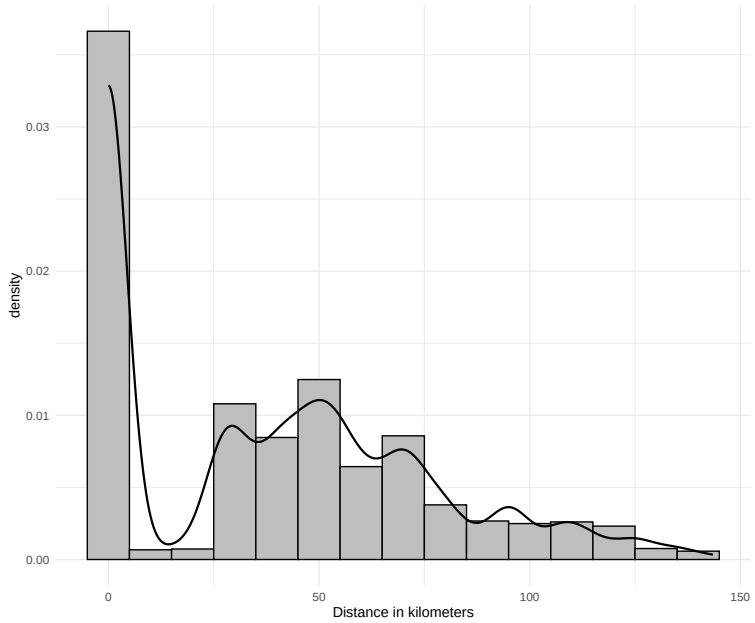


Figure 2: Distribution of the distance from 6th grade home to the closest university

While already used as an instrument for educational attainment and duration, the distance to the closest university can also be used as an instrument for dropout or delay in graduation. The distance to the closest university affects the dropout probability in two main ways. The considered distance captures either the cost of education (or the effort produced to acquire education), but also a part of the sunk costs in case of dropout. For students who have a university in their surrounding ($distance = 0$), the cost of acquiring education are lower that for those who have to commute to the university, reducing the sunk cost in case of dropout, and then increasing the probability of dropping out. For students who either have to commute to the university, or to live in another city, the potential effects of distance are plurals. The distance is increasing the cost of acquiring degrees, thus increasing the probability of dropping out. However, for students having to live in another city, the sunk cost of housing and transportation will acts negatively on the dropout probability. Thus, I will include a polynomial characterization of the distance to the closest university in order to predict the probability of dropping out.

3 Methodology

The objective of this analysis is to identify subgroups with different treatment effects of dropping out, conditional on a vector of covariates X, in order to estimate the effect of dropping out for heterogeneous sub-groups : dropouts from different degrees, or socioeconomic backgrounds. As defined by Rubin (1974), the treatment effect of dropping out is computed as the individual difference in potential outcome $\tau_i = Y_i(1) - Y_i(0)$ with $Y_i(W_i)$ the outcome depending on the treatment status W_i. Since we do not observe both $Y_i(1)$ and $Y_i(0)$, we focus on the estimation of the Conditional Average Treatment Effects (CATE) defined as $\tau(x) = \mathbb{E}[Y_i(1) - Y_i(0) \mid X_i = x]$.

This estimator is constructed as a subsample average treatment effect on the individuals sharing $X_i = x$. Thus, for a combination of the vector $X_i = x$, we will be able to compute the treatment effect on this combination $X = x$, corresponding to individuals showing similar characteristics with i.

If we want to test for every interaction that the vector X allows, the number of interaction terms could be gigantic and will obviously detect spurious correlation. To avoid this pitfall, I rely on the Generalized Random Forest developed by Athey et al., 2019, and use the data structure to identify heterogeneous treatment effects. This method allows us to compute Conditional Average Treatment Effect (CATE), the individual treatment effect, and the corresponding standard error, and then to average these effects on selected partitions of the population as Average (Conditional) Local Average Treatment Effects (A(C)LATE). The Generalized Random Forest relies on regression trees to estimate the individual CATE, and then average the estimated treatment effects across all trees. The Random Forest (Breiman, 2001) was developed to account for a large possibilities of non linear interactions between covariates, without risking over-fitting. To avoid leveraging spurious correlations due to using the similar data to find heterogeneous sub-partition in the dataset and to estimate the corresponding treatment effects, Athey and Imbens (2016) rely on the honest methodology. Finally, I use an adapted instrumental setting in order to tackle the endogeneity of dropping out to unobserved characteristics. In this section, we will develop the Generalized Random Forest algorithm, the instrumental variable setting, and then the Average (Conditional) Local Average Treatment Effect estimation.

In this section, I will avoid to use too much technical explanations and try to focus on the general idea of the methodology. See Athey et al., 2019 for all the technical details of the Generalized Random Forest.

3.1 The Generalized Random Forest algorithm

(All the notation are taken from either Hastie et al. (2009), Athey and Imbens (2016), Wager and Athey (2018) or Athey et al. (2019)).

The GRF is based on the regression tree algorithm developed by Breiman et al. (1983) (called the CART for Classification and Regression Trees) and adapted as causal honest tree by Athey and Imbens (2016). I will proceed by first describe the regression tree, the adapted honest causal tree and finaly the Generalized Random Forest.

In the initial paper by Breiman et al. (1983), the classification and regression (CART) trees use a training sample $\mathcal{S}^{tr}$ for which we know (Y_i, X_i) with Y_i the outcome and X_i a vector of covariates, and a target sample for which we know only X_i. By fitting a tree model on $\mathcal{S}^{tr}$, the objective is to predict correctly the outcomes for the target sample. To do so, the algorithm first search for a splitting point s on a splitting variable X_j in order to create two subsample

$R_1(j,s) = [X \mid X_j \leq s]$ and $R_2(j,s) = [X \mid X_j > s]$. In this setting, s is found by minimizing the mean squared error defined as :

$$MSE = \left[\sum_{x_i \in R_1} (y_i - \bar{y_1}(j,s))^2 + \sum_{x_i \in R_2} (y_i - \bar{y_2}(j,s))^2 \right] \tag{1}$$

Finally, the algorithm repeat this method until a stopping point (usually a minimum number of individuals in the sub-samples, or a maximum number of sub-samples). To compute the predictions for another sample $\mathcal{S}^{pred}$, the CART fit new observations into their corresponding subsamples, and then assign the mean of this subsample as the predicted outcome $\hat{Y}_i$. Compared to linear regression or similar methods, the CART allows to account for high dimensional and non linear interactions between all covariates in X_i and help to build strong predictive models.

If the CART is efficient to produce prediction on a target sample, it is not yet suitable to estimate CATE. For this aim, we need two modifications of the original algorithm : introduce an "honest" design and use an modified splitting rule.

The honest design, firstly applied to regression tree by Athey and Imbens (2016), allow to solve the over fitting problem. Over fitting arise when a model match too closely the data and then present no generalization power. Indeed, if we use the same sample to build the regression tree and to estimate the CATE in every created sub-samples, wee will obtain completely biased estimators, as both sample are no independants. In the honest design, we use two different and randomly drawn sub-samples to build the tree with the first one, and then to estimate effect in the sub-samples build by the regression tree in the second one.

The objective of an honest causal tree is to create subgroups in the population on which the Conditional Average Treatment Effects (CATE) are evaluated. For a given dataset, we observe (Y_i, W_i, X_i), for $i = 1, ..., N$, with Y_i the outcome, X_i a vector of covariates and W_i the treatment status. In our example, if the considered individual has dropped out, she shows $W_i = 1$ and $W_i = 0$ if she didn't.

In order to account for the second stage estimations, we need to adapt the objective function. We will focus on the Expected Mean Square Error, an adapted estimator of the Mean Squared Error.

We introduce here the estimated Conditional Average Treatment Effect, the empirical expression of the CATE presented below. With $\hat{\mu}$ the conditional mean of a subsample, it is defined as :

$$\hat{\tau}(x; \mathcal{S}) = \hat{\mu}(w_i = 1, x, \mathcal{S}) - \hat{\mu}(w_i = 0, x, \mathcal{S})$$

This expression estimates the CATE on individuals with $X_i = x$ as the difference between the both treated and non treated conditional mean on the given subsample. The objective function, the Expected Mean Squared Error (EMSE) is then design using the estimated CATE.

With N^{tr} the size of the training sample (made equal to the size of the estimation sample), l a subsample, $S^2_{\mathcal{S}^{tr}}(l)$ the subsample estimated variance of $\hat{\tau}$ and p the probability of being treated , the adapted expected Mean Squared Error is defined as :

$$\widehat{EMSE}_\tau(\mathcal{S}^{tr}) = \frac{1}{N^{tr}} \sum_{i \in \mathcal{S}^{tr}} \hat{\tau}^2(X_i; \mathcal{S}^{tr}) - \frac{2}{N^{tr}} \sum_l \left(\frac{S^2_{\mathcal{S}^{tr}_{treated}}(l)}{p} + \frac{S^2_{\mathcal{S}^{tr}_{control}}(l)}{1-p} \right) \tag{2}$$

This estimator of the Expected Mean Squared Error is almost composed as the MSE, but add a negative effect of within subsample variance of the CATE. This allows the algorithm to take into account that finer partition generate greater variances. Then, with this objective function, the algorithm will search for split that maximize treatment heterogeneity in treatment effect while avoid generating too much in-partition variance. For more details on the construction of this objective function, please refer to Athey and Imbens, 2016.

Since we have an efficient splitting criterion, one problem remain : due to the honest design, the built tree will greatly depend of the initial random splitting. To solve this issue, I use the Random Forest algorithm developed by Wager and Athey (2018). The objective of the causal Random Forest is to create causal honest trees on sub-samples of the whole population. For example, we draw a partition α of the initial population, and build the honest causal tree on this partition as described below. Then, the algorithm average all the individual CATE given by all trees to compute the individual CATE. This method provide unbiased estimates of individuals treatment effects with the associated standard error. One of the main assumption of this model is the unconfoundedness i.e $W_i \perp (Y_i(0)), Y_i(1), X_i)$. This assumption is satisfied in a random treatment assignment setting such as Random Control Trials. Since it is almost impossible to randomize the dropout, I have to include a instrumental variable setting in the framework.

The Generalized Random Forest developed by Athey et al. (2019) propose a general framework to estimate CATE with methods such as causal Random Forest and Instrumental Forest. The main divergence from the initial causal Random Forest come from the usage of a gradient-based loss criterion rather than the exact loss criterion (2). The gradient-base criterion is an approximation of (2) build with gradient-based approximations of $\hat{\tau}$ for each sub-samples. This method, designed as a general framework for estimation in non-linear setting, help to use IV and is less costly in computation.

3.2 The forests construction

In this paper, I use the GRF to build individual CATE by using the following variables : the highest diploma tried on 6 levels, if the student made internship or international travel, the higher education institution region, the type of high school diploma (general, technical or professional) on three variables, a categorical variable for the grade at the high school diploma, the highest professional occupation and diploma of both parents, and the gender. I build two

forests : the first one dedicated to estimate the heterogeneous effect of dropping out on the rate of employment, and the second on the average wage.

The forest is constructed using the following parameters : each tree is built using 50% of the sample, of this half, 70% is dedicated to build the tree and 30% to estimate the treatment effect in each sub-samples. The maximum imbalance between two splits is 88% to 12% for both built forest. All the variables are dichotomized, and represents 26 binary variables. At each splitting try, 10 variables are tried out of the 26 variables. Finally, I build 5000 trees in each forest which is enough to obtain a stable estimation of the CATE.

3.3 Orthogonalization

The Generalized Random Forest rely on an orthogonalization step which regress out the effect of the covariates X on Z the instrument, W the treatment indicator and Y the outcome. The objective of this step is to obtain accurate treatment effect estimation, and to increase the efficiency of the learning sequence of the forest. By regressing out the effect of X, the forest is trained on an dependent variable vector $\tilde{Y}$ which doesn't depend on X, thus concentrating the learning on the heterogeneity actually depending on W, and not on X. The same procedure is applied to Z by regressing it on X.

To do so, the conditional marginal expectations of Y, W and Z are computed and used to obtain the conditionally centered outcomes :

$$\tilde{Y}_i = Y_i - \hat{y}^{(-1)} \quad \text{with} \quad \hat{y}^{(-1)} = \mathbb{E}[Y_i|X = x] \tag{3}$$

$$\tilde{W}_i = W_i - \hat{w}^{(-1)} \quad \text{with} \quad \hat{w}^{(-1)} = \mathbb{E}[W_i|X = x] \tag{4}$$

$$\tilde{Z}_i = Z_i - \hat{z}^{(-1)} \quad \text{with} \quad \hat{z}^{(-1)} = \mathbb{E}[Z_i|X = x] \tag{5}$$

$$\tag{6}$$

The forest is trained using the set of transformed outcomes $(\tilde{Y}_i, \tilde{W}_i)$ (also called the centered outcomes). This step also helps to reduce the training time dedicated to estimating the propensity of treatment conditional on X, since it is already sorted out with this step Athey et al., 2019. The structure of transformed outcomes conditional on certain covariates are presented in section 4.1.

3.4 The instrumental variable setting

As explained before, we cannot consider students who drop out as randomly selected into treatment, even conditionally on coviates, due to unobserved characteristics such as the individual ability. Thus, we need to rely on an instrumental variable setting to identify the unbiased effect of dropping out on labor market outcomes. As precised in the section 2, the used instrumental variable is the distance to the closest university at 6th grade. I don't use the distance between the high school and the closest university, since some students move from their initial

high school to a better one, usually based on performance or merit, thus inducing endogeneity between the distance and the labor market outcomes (Brodaty et al., 2008).

In order to obtain stable and consistent estimate of the effects of dropout on labor market outcomes and to include a polynomial expression of the distance, I adapt the methodology proposed by Adams et al., 2009 : a four step instrumental variable process. The steps are :

1. Estimate $Pr(w = 1|x_c, z) = \phi(\gamma_0 + \gamma z + \theta x_c)$, with ϕ a cumulative distribution function (here the logistic cumulative distribution function)
2. Compute the fitted probability $\hat{w_1}$ using the precedent step estimation
3. Estimate $w_i = \theta_0 + \eta \hat{w_1} + \theta x + \epsilon_i$, this linear model is estimated with an OLS
4. Compute the fitted probabilities $\hat{w_2}$ using the precedent step estimation, and use these fitted probabilities as the instrumental variable

With z the vector of instrumental variable, x_c the first step vector of control, x the vector of covariates used to build the forest. The instrumental variables include a dichotomous variable indicating if an individual has an university in her area (distance to closest university = 0), $dist_0$, and a polynomial expression of the distance : $z = (dist_0, dist, dist^2)$. The vector of control x_c consists in fixed effects for all the french regions.

In the presence of multiple instrumental variable and control, this methodology has advantages compared to the pseudo-IV methods : it can smoothly include many instrumental variable, and doesn't need the first step to be correctly specified. The only requisite for the first step is for the instrumental variables to be correlated with the dropout indicator, and for $\hat{w_1}$ to keep a strong correlation with this indicator in step 3. The results of step 1 and 3 are presented in section 4.2.

In the case of the GRF algorithm, estimating a Instrumental Forest is equivalent to apply the Wald formula for individuals with $X_i = x$. The interactions terms generated by the GRF, change for every tree and then help us to account for high dimension heterogeneity. Since there exists instruments z satisfying all the IV assumptions, the dropout effect can be estimated as :

$$\tau(x) = \frac{Cov[Y_i, Z_i \mid X_i = x]}{Cov[W_i, Z_i \mid X_i = x]} \tag{7}$$

In this setting, the IV can be implemented with a binary or continuous instrumental variable. However, since the A(C)LATE is used to estimate treatment effect on group of former students, a dichotomous instrumental variable is needed. The identification strategy rely on the estimation of doubly robust scores, as proposed by Athey and Wager (2020), and average them over sub-samples to get the unbiased A(C)LATE. As precised in Athey and Wager

(2020), we need a binary instrument to compute the doubly robust scores. I finally proceed to a dichotomization of the fitted probability $\hat{w_2}$:

$$\begin{cases} \tilde{w_2} = 1 & \text{if} \quad \hat{w_2} > p(\alpha) \\ \tilde{w_2} = 0 & \text{if} \quad \hat{w_2} \leq p(\alpha) \end{cases}$$

With $p(\alpha)$ the value corresponding to the α^{th} percentile and $\hat{w_2}$ the fitted probability of dropping out computed at step 3. My choice of α is motivated by the A(C)LATE estimation step. The A(C)LATE is the average treatment effect on the compliers i.e individual who respond positively to the instrument. Since the A(C)LATE is computed by averaging the treatment effect times a weighting function which is divided by product of compliance scores, we need to keep the compliance scores as high as possible. The compliance score is defined as the individual propensity to dropout conditional on (x, z). The threshold which maximize the product of the scores is around $(\alpha) = 0.80$. After this step, the dichotomized instrument is equal to 1 for 4278 observations, and equal to 0 for 8435 observations. The distribution of the instrument is not 20% positive because a part of the sample is dropped due to too low propensity score (see section 4.1).

It is possible that certain students with high effect of dropping out are below the $(\alpha) = 0.80$ threshold, thus potentially overturning the results of this estimation. Thus, I perform the same analysis with $(\alpha) = 0.60$ as a robustness check. The results can be found in section E.

3.5 Doubly robust estimation and Average Conditional Local Average Treatment Effect

The instrumental forest described previously generate individual Conditional Average Treatment Effect (A(C)LATE), formally $\tau(X) = \frac{Cov[Y,Z|X=x]}{Cov[W,Z|X=x]}$. In their paper, Athey and Wager (2020) propose a method inspired from Chernozhukov et al. (2022) to estimate doubly robust score of $\tau(X)$. To assess potential heterogeneity in the estimated treatment effects, we average the doubly robust scores to obtain the Average Conditional Local Average Treatment Effect. The A(C)LATE is asymptotically normally distributed, thus we can interpret it as an estimator of the doubly robust treatment effect on the compliers for a chosen subgroup.

The method chosen to assess CATE heterogeneity is to used the estimated treatment effect value generated by the Instrumental Forest built with the GRF methodology, to split the sample around the median of estimated CATE and to compute the A(C)LATE on each subsample. Since the A(C)LATE is a asymptotically normal, we can test if each subsample groups individuals with a treatment effect significantly different from 0 , and if the difference between the both groups A(C)LATE is significant.

The doubly robust score is computed as the average of the estimate CATE by the Instrumental Forest and the multiplication of the Y residuals multiplied by a debiasing weight

:

$$\Gamma = \tau(X) + g(X, Z)\left(Y - \mathbb{E}[Y|X] - (W - \mathbb{P}[W = 1|X])\tau(X)\right) \tag{8}$$

With $g(X, Z)$ the vector of debiasing weight :

$$g(X, Z) = \frac{1}{\Delta(X)} \frac{Z - \mathbb{P}[Z = 1|X]}{\mathbb{P}[Z = 1|X](1 - \mathbb{P}[Z = 1|X])} \tag{9}$$

In (6), $\Delta(X)$ is the vector of compliance score : $\mathbb{P}[W|Z = 1, X]$. It represent the propensity of an individual to dropout if the instrument is positive. The compliance score are computed using a causal forest (see Arcidiacono et al., 2010 for detailed explanation around the compliance score). For the practical way of estimating the doubly robust score, see Athey and Wager (2020).

Finally, the A(C)LATE is estimated as the average of all doubly robust scores. The A(C)LATE are computed on each subsample divided around the median of the CATE, as well as in quartiles and per degree and socioeconomics status.

4 Results

4.1 Preliminary steps : overlapping and orthogonalization of the outcomes

Before building the Generalized Random Forest, I need to address the question of common support and then show the preliminary step of the GRF defined as orthogonalization (or residualisation) defined in section 3.3.

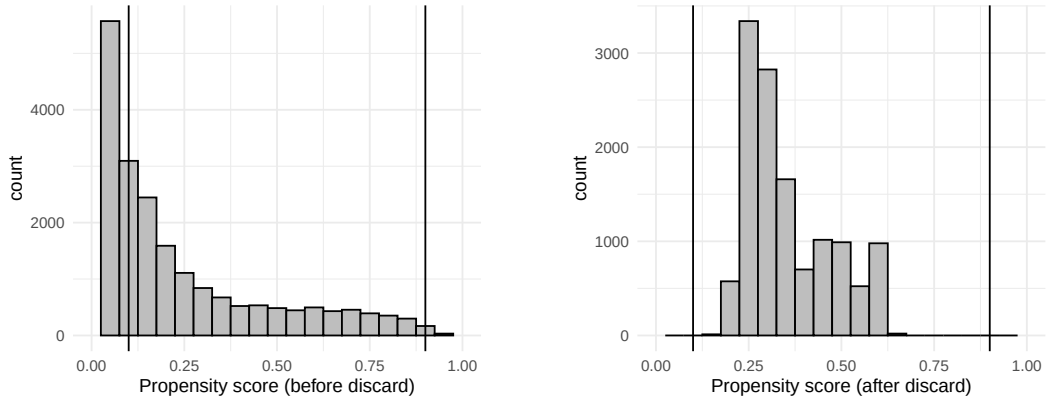
Overlapping

The Average (Conditional) Local Average Treatment Effect (A(C)LATE) proposed by Athey et al., 2019 relies (among others) on propensity score to estimate the treatment effect. The propensity score, defined as $p_s = \mathbb{E}[W_i = 1|X_i = x]$, is estimated using a regression forest[6] in order to account for the highly dimensional predictable power of X on W. However, having propensity scores close to 0 or 1 leads to unstable estimator of the A(C)LATE.

In order to solve this issue, I follow Crump et al., 2009 and discard observations with a propensity score such that $p_s \notin [0.1; 0.9]$. The source of extreme propensity scores are plurals, for example certain covariates' levels exhibit very low dropout rate, leading to very low propensity score. Having too few observations per class can leads to biased estimation of the treatment effect, however these levels (mostly Bac +4 and Bac +5) do not represent the core target of this work.

[6]The regression forest is a causal random forest estimating the heterogeneous conditional mean $\mu(X) = \mathbb{E}[W|X]$

This discard phase leads to drop 9116 observations, and let 12713 observations in the data set. The result of this step is presented in figure 3.



(a) Before discard (b) After discard

Figure 3: Propensity score $\mathbb{E}[W_i = 1|X_i = x]$

The distribution between discarded and non discarded students by variables is presented in table 2. This step leads to drop 9116 observations, including 2945 Bac +2 and 177 Bac +3. While dropping more Bac +2 than Bac +3, the final database include a reasonable number of observations for each levels, and should not biased the estimation of the A(C)LATE for each type of degrees. The final database counts 7150 Bac +2 and 3085 Bac +3.

	Frequency (before discard)	Discarded	Non-discarded
	Gender		
Male	10092	3552 (35.2%)	6540 (64.8%)
Female	11737	5564 (47.4%)	6173 (52.6%)
	Highest diploma tried		
Bac +2 (STS/IUT)	10095	2945 (29.2%)	7150 (70.8%)
Bac +3 (university)	3262	177 (5.4%)	3085 (94.6%)
Bac +4 (university)	943	0 (0%)	943 (100%)
Bac +5 (university/Grande Ecole)	5051	4449 (88.1%)	602 (11.9%)
PhD	2478	1545 (62.3%)	933 (60.4%)
	Parents' highest social category		
Disadvantaged	3132	1069 (32.1%)	2063 (65.9%)
Intermediate	5718	2116 (37.0%)	360 (63.0%)
Advantaged	4277	1605 (37.5%)	2672 (62.5%)
Highly Advantaged	8702	4326 (49.7%)	4376 (50.3%)
	Parents' highest diploma		
No diploma	3819	1199 (31.4%)	2620 (68.6%)
Bac or below	8119	2867 (35.3%)	5252 (64.7%)
Short degree	5915	2771 (46.8%)	3144 (53.2%)
Long degree	3976	2279 (58.8%)	1697 (42.7%)

Table 2: Summary statistics by discard status

Orthogonalization

As described in section 3.3, the outcomes Z, W and Y are centered (or orthogonalized) by subtracting the expected outcomes conditional on X : $\mathbb{E}[Z_i|X = x], \mathbb{E}[W_i|X = x], and \mathbb{E}[Y_i|X =$

x]. This preliminary step, except from ensuring the efficiency of the forest building part, also focus the splitting part (or the tree building) on the treatment effect conditional on X, and not on splits related to the direct effect of X on Y, W or Z. The orthogonalized outcomes are obtains using :

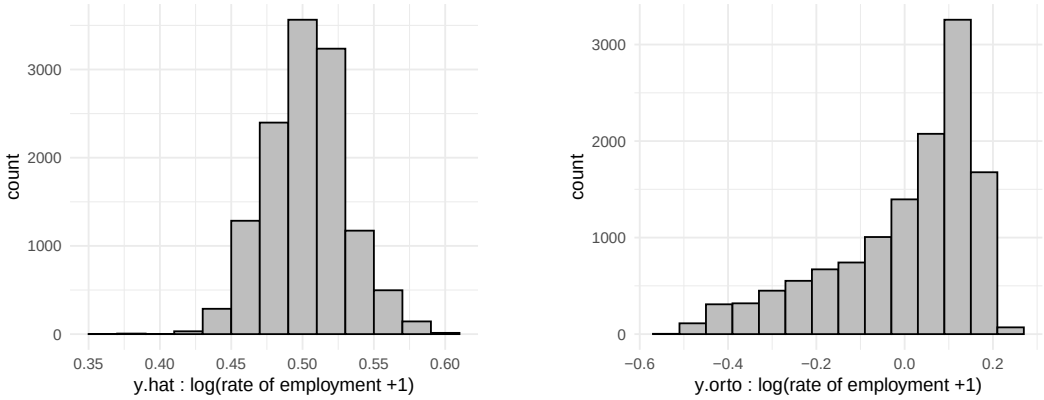
$$\tilde{Y}_i = Y_i - \hat{y}^{(-1)} \quad \text{with} \quad \hat{y}^{(-1)} = \mathbb{E}[Y_i|X = x] \tag{10}$$

$$\tilde{W}_i = W_i - \hat{w}^{(-1)} \quad \text{with} \quad \hat{w}^{(-1)} = \mathbb{E}[W_i|X = x] \tag{11}$$

$$\tilde{Z}_i = Z_i - \hat{z}^{(-1)} \quad \text{with} \quad \hat{z}^{(-1)} = \mathbb{E}[Z_i|X = x] \tag{12}$$

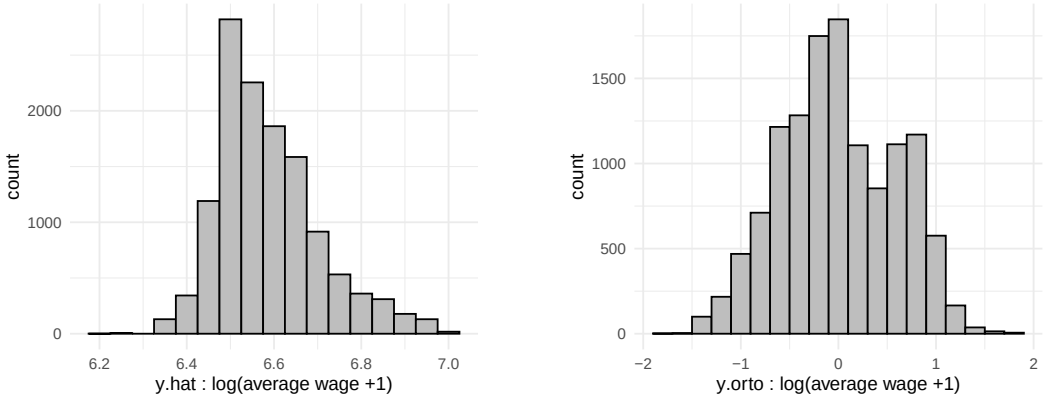
$$\tag{13}$$

In this section are presented the predicted and orthogonalized outcomes Y for the rate of employment and the average wage, and the orthogonalized outcomes for the average wage conditional on two levels of the maximum social category of the parents.



(a) $\hat{Y}$ (b) Y_{ortho}

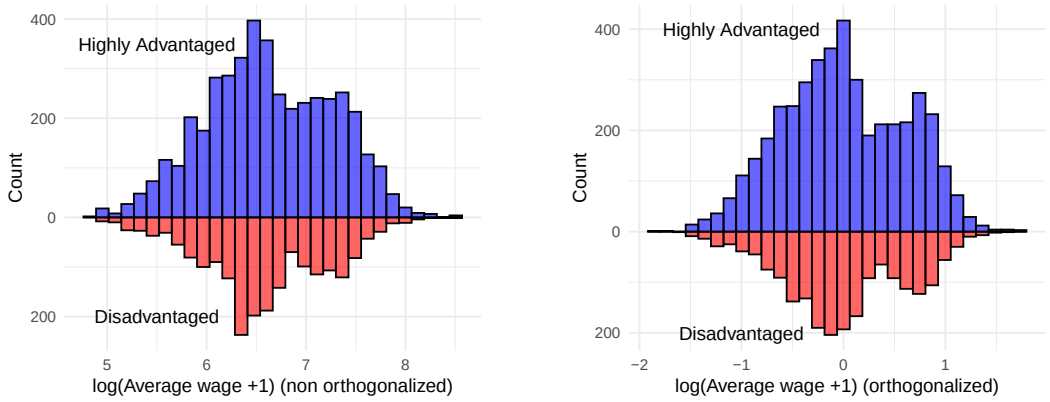
Figure 4: Predicted and orthogonalized outcome Y : rate of employment



(a) $\hat{Y}$ (b) Y_{ortho}

Figure 5: Predicted and orthogonalized outcome Y : average wage

We can observe that the orthogonalization center the Y distribution on 0, but keep the overall shape of the distribution.



(a) Non orthogonalized (b) Orthogonalized

Figure 6: Orthogonalized outcome Y conditional on the social category of the parents : average wage

	Min	Q1	Median	Mean	Q3	Max
$[Y\|X = \text{Disadvantaged}]$	4.95	6.18	6.53	6.57	7.05	8.46
$[Y\|X = \text{Highly advantaged}]$	4.86	6.18	6.58	6.63	7.15	8.54
$[Y_{ortho}\|X = \text{Disadvantaged}]$	-1.31	-0.39	-0.05	0.00	0.45	1.63
$[Y_{ortho}\|X = \text{Highly advantaged}]$	-1.77	-0.42	-0.05	0.00	0.47	1.82

Table 3: Distribution of the initial and orthogonalized outcome conditional on the social origin : average wage

In figure 6 and table 3, we can observe that the orthogonalization step helps to smooth the distribution and to close the median and mean difference between disadvantaged and highly advantaged background student. The step generate the same effect for every variables included in X, and the interaction generated by the regression forest used to estimate $\mathbb{E}[Y_i|X = x]$. Thus, the built forest will be around the heterogeneous treatment effect of dropping out conditional on the vector of covariables X.

4.2 First stage regression

As described before, dropping out is not randomly distributed in the students population and thus in the used sample. We need to rely on an instrument that is correlated with the dropout indicator, but excluded from the outcome equation. In this section, I present the results from the step 1 and 3 described in section 3.4. The first step is to estimate a logit model of the probability of dropping out conditional on a polynomial of the distance to the closest higher education establishment in 6th grade : $z = (dist_0, dist, dist^2)$.

	Dependent variable:	
	dropout	
	(1)	(2)
Constant	-1.447^{***}	-1.496^{***}
	(0.108)	(0.114)
$dist_0$	0.179	0.308^{**}
	(0.112)	(0.125)
$dist$	1.049^{***}	1.281^{***}
	(0.339)	(0.380)
$dist^2$	-0.009^{***}	-0.009^{***}
	(0.002)	(0.003)
Regional FE	No	Yes
Observations	21,829	21,829
Log Likelihood	-11,638.390	-11,520.570
Akaike Inf. Crit.	23,284.790	23,091.150
Note:	*p<0.1; **p<0.05; ***p<0.01	

I regress the polynomial expression of the distance on the dropout indicators, with and without regional fixed effects. For the readability of the parameters, the distance and distance squared have been divided by 100. The regional fixed effects include 22 indicators.

Table 4: First stage LOGIT regression

The results table 4 indicate a strong correlation of the three part of the instrumental variable with the dropout indicator. When the regional fixed effects are included, living in an area hosting a university impacts positively the dropout propensity, while the distance to the closest university follow a concave parabolic curve. This indicates a threshold of the distance effect : under a certain distance, living far from a university increases the likelihood of dropping out, while after this threshold, the distance tends to decrease the probability of dropping out. Since regional fixed effect are included in the second model, and every region have a university, this definition of distance measure the effect within each region, cleaning out the potential heterogeneity proper to each region. The fitted probabilities of the stage 2 and 4 are presented in figure 7.

The stage 3 results are presented in table 5. When all the controls are included, the first stage fitted probabilities are still very significant and the t-stat of the instrument : $t_{stat} = 7.62 \Leftrightarrow F_{stat} = 58$ is high enough to rule out the weak instrument bias. The controls included

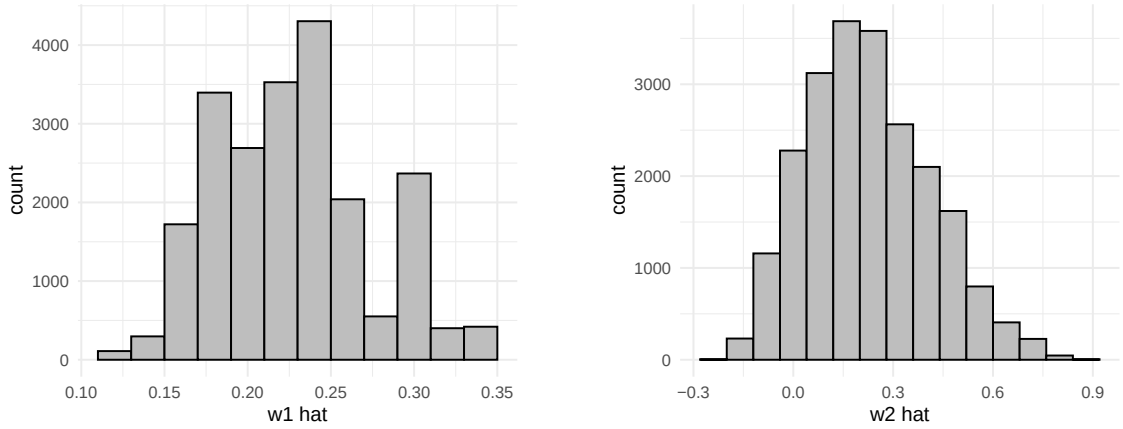
	Dependent variable:	
	dropout	*dropout*
	(1)	(2)
Constant	−0.0002	0.248***
	(0.014)	(0.021)
$\hat{w_1}$	1.001***	0.572***
	(0.061)	(0.075)
Controls	No	Yes
Observations	21,829	21,829
R^2	0.012	0.199
Adjusted R^2	0.012	0.197
Residual Std. Error	0.415 (df = 21827)	0.374 (df = 21787)
F Statistic	269.573*** (df = 1; 21827)	131.912*** (df = 41; 21787)
Note:		*p<0.1; **p<0.05; ***p<0.01

$\hat{w_1}$ correspond to the predicted probability of dropping out using the polynomials expression of the distance with regional fixed effect. In this table, I regress $\hat{w_1}$ and controls on the dropout indicator. The controls include : Gender, highest diploma tried (6 levels), Min/Max SES, Min/Max parents' diploma, type of HSD, HSD grade, region of the university, went to study abroad, did an internship

Table 5: First stage linear regression

are those used to build the forest : the gender, the highest diploma tried (6 levels), the maximum socio-economic status of the parents, the maximum parents' diploma, the type of HSD and the distinctions at the final exam (proxy of the grade), the region of the university the individual went at, if she went to study abroad and if she did an internship.

The final step is to dichotomize the obtain instrument $\hat{w_2}$: individuals in the top 20% of $\hat{w_2}$ are assigned a $Z_i = 1$, the others 0. I also perform the same analysis with $Z_i = 1$ for the top 40% of the distribution; the results can be found in section E of the appendix.



(a) Stage 2 : fitted probability of dropping out ($\hat{w_1}$)

(b) Stage 4 : fitted probability of dropping out ($\hat{w_2}$)

Figure 7: Distribution of the fitted probabilities

4.3 Instrumental variable regression

In order to understand how the highly dimensional setting used by the GRF is useful to obtain an accurate estimation of the effect of dropping out, I present the estimation of the local average treatment effect using a Two Stage Least Square approach. The TSLS estimation is preceded by the first stage treatment fitted probabilities estimation described in section 3.4 and 4.2. This methodology is also described in Adams et al. (2009).

An emphasis should be put on why the LATE obtained with TSLS is different from the A(C)LATE obtained with GRF, as they estimate the same effect. While TSLS estimate the effect of dropping out everything else equal, the GRF algorithm average all the individuals CATE for $X = x$ obtained by estimating the treatment effect in each terminal nodes of the trees, thus using the highly dimensional setting proposed by the tree structure to estimate the conditional treatment effects. Then, the $LATE_{GRF}$ can differs from $LATE_{TSLS}$ mainly because of the tree structure of the CATE estimation. The results of both TSLS estimations are presented in table 6, the full tables can be found in section D. The included controls are : the highest diploma tried (and eventually obtained) on 6 levels, the maximum SES of the parents, the maximum diploma of the parents, the discretized grade obtained at the high school diploma, if the students did an internship, if she did an foreign study trip, and an indicators of the higher education institution's region on 22 levels.

The $LATE_{TSLS}$ of dropping out on the rate of employment is significantly different from 0 and show an effect of $e^{-0.192} - 1 = -17.5\%$. The weak instrument and Wu-Hausman test are both significantly different from 0, assuring that we can safely rule out the weak instrument bias eventuality.

The $LATE_{TSLS}$ of dropping out on the average wage exhibits an effect of $e^{-0.369} - 1 = -30.9\%$ of dropping out on the average wage, and is significantly different from 0 at the 1% level. The weak instrument and Wu-Hausman test are again significantly different from 0.

	Dependent variable:	
	log(Rate of employment + 1)	log(Average wage + 1)
	(1)	(2)
Constant	0.570***	6.819***
	(0.011)	(0.039)
dropout	−0.192***	−0.369***
	(0.025)	(0.087)
Weak instruments	234.97***	234.97***
Wu-Hausman	23.35***	5.422**
Observations	12,713	12,713
R^2	-0.017	0.095
Adjusted R^2	-0.021	0.092
Residual Std. Error (df = 12671)	0.172	0.595

Note: *p<0.1; **p<0.05; ***p<0.01

This table presents the results for the TSLS estimation of the effect of dropping out on the rate of employment and the average wage. The instrument is a polynomial expression of the distance to the closest university in 6th grade, and is included following the four step methodology described in Adams et al., 2009. The variables included in the regression are the gender, highest diploma tried (6 levels), Min/Max SES, Min/Max parents' diploma, type of HSD, HSD grade, region of the university, went to study abroad, did an internship.

Table 6: TSLS estimation of dropout LATE

Those estimated effects are in line with the aforementioned literature, in their magnitude and significance. The TSLS estimation of dropping out will be used as the baseline result and compared to the GRF estimation in the following section.

4.4 Assessing the dropout effect heterogeneity

As described in the methodology part, the constructed distributions of CATE are the results of a data-mining process used to discover heterogeneity in the effect of dropping out on labor market outcomes. Since the whole process is made to maximize the heterogeneity in $\hat{\tau}$, the CATE, it is possible that the built distributions are the results of noise in the data. Then, we need to rely on the doubly unbiased estimation of the A(C)LATE and its statistical properties to test for the presence of actual heterogeneity in the distribution.

In section 4.4.1, the A(C)LATE obtained with GRF for the whole sample is compared to the LATE obtain with TSLS in section 4.3 in order to test if the highly dimensional structure used by GRF helps to catch previously uncovered effect. Then, the sample is split around the CATE median and quartiles, and the A(C)LATE are computed on each sub-samples (median and quartiles). I compute the t-statistic for the difference between each A(C)LATE of the sub-samples to assess if the effect heterogeneity between each part is generated by the data structure or by noise.

In section 4.4.2, the A(C)LATE per diploma, socio-economic status and the interaction of both variables are computed to assess which sub-samples is the most penalized when dropping out. The heterogeneous effect per diploma answer the question of the legitimacy of targeting mostly the university in dropout policy, while the interaction of both covariates provides evidence to solve the resources allocation trade-off in a multi-track higher education system, and which students are the optimal target for these policies.

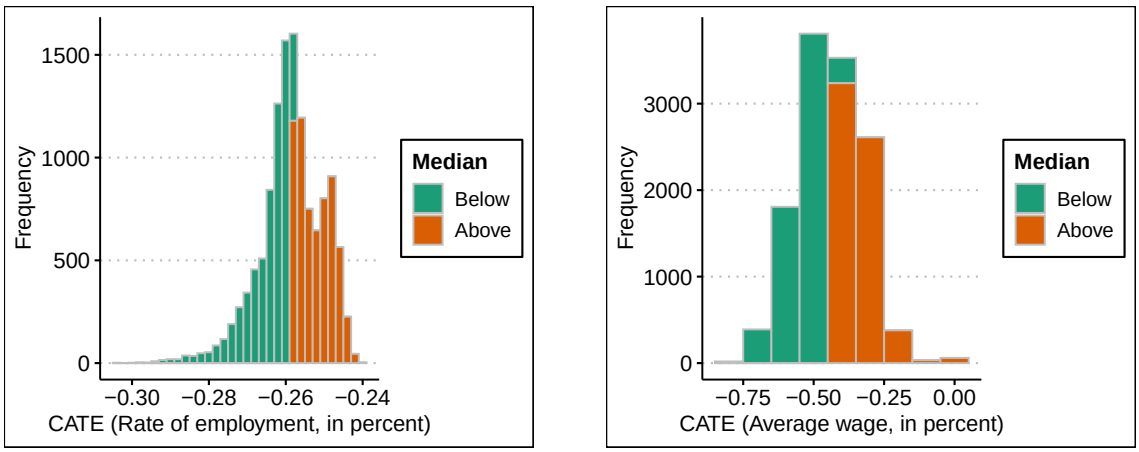
4.4.1 LATE estimation with GRF

The estimated Conditional Average Treatment Effect (CATE) follows the distributions showed in figure 8 and table 7. Since the variation measured by the CATE or the A(C)LATE are important, the log transformation doesn't measure correctly the variation in percentage, thus I apply the $e^{\text{CATE or LATE}} - 1$ transformation for every results presented in this section.

	Min	Q1	Median	Mean	Q3	Max
Rate of Employment	-30%	-26%	-26%	-26%	-25%	-24%
Average Wages	-79%	-51%	-43%	-42%	-33%	0.01%

Table 7: Distribution of CATE of dropout on the Rate of employment and the Average wage

The CATE of dropping out on the rate of employment ranges from -30% to -24% of the time spent on the labor market at the end of the education period. The individual estimated



(a) Time in Employment

(b) log(Average wages)

Figure 8: Conditional Average Treatment Effect of dropout

effect of dropping out on the time in employment is concentrated with only a 6 percentage pointss difference between the strongest and the smallest estimated CATE. The average (and the median) of the CATE is equal to -26%, which for a 36 months period represent a reduction of 9 months of the time in employment for dropouts.

The CATE on the average wage range from -79% to almost null effect, with the median at -43% and the mean at -42 % on monthly wage. The CATE distribution exhibits a strong heterogeneity in individual conditional average treatment effect, far greater than the distribution of CATE for the rate of employment. With an average wage of 877€ per month, an average decrease of 42% is equivalent to a reduction of 385€ per month.

As explained in section 3.5, even if the CATE distribution can be indicative of the dropout heterogeneous effect, we need to rely on the doubly robust score average (the A(C)LATE) to estimate the unbiased effect of dropping out on the labor market outcomes, conditional on the partioning. First, I compare the estimated LATE obtained with TSLS and GRF and test for their difference. Then, as proposed by Athey and Wager, 2019, the individuals are split following their individual CATE around the median. I define $\tau_1 \leq Median(\tau_i)$ the group of individuals showing a CATE equal or below the median of the estimated CATE and $\tau_2 > Median(\tau_i)$ the group of individuals showing a CATE above the median of the estimated CATE. The τ_1 corresponds to a group showing a strong negative effect of dropout, while the τ_2 correspond to a group showing a less negative effect. I also present the A(C)LATE for the four quartile (Q_1, Q_2, Q_3, Q_4). A student test for the difference between the A(C)LATE of both groups is performed and the T-Statistic is given for each group effect comparison, as well as the statistical significance of the difference between Q_1, Q_2, Q_3 and Q_4.

The A(C)LATE on each subgroups are presented in table 8 for A(C)LATE on time in employment and in table 9 for the A(C)LATE on average wages.

The first interesting result is that the GRF estimate a A(C)LATE of dropping out of -26%

	$TSLS$	GRF_1	GRF_2	GRF_3	t-stat
Overall sample	-0.187***	-0.256***			-2.23**
	(0.025)	(0.017)			
$\tau_1 \leq Median(\tau_i)$			-0.287***		-1.88*
			(0.025)		
$\tau_2 > Median(\tau_i)$			-0.224***		
			(0.023)		
Q_1				-0.292***	-2.49**
				(0.027)	
Q_2				-0.282***	-1.76*
				(0.042)	
Q_3				-0.254***	-1.36
				(0.034)	
Q_4				-0.192***	
				(0.029)	
Note:					*p<0.1; **p<0.05; ***p<0.01

The standard error is precised in parenthesis. The two first columns indicate the LATE obtain with TSLS and GRF, while the last column shows the t-statistics testing the difference between the both estimators. The third column shows the A(C)LATE for the 50% lower than the median and for the top 50%. The fourth column indicates the A(C)LATE per quartiles. For the effect per quartiles, the t-test is computed with respect to Q_4, with the smaller effect. $LATE \times 100$ gives the estimated effect in percentage.

Table 8: A(C)LATE of dropping out on the rate of employment

of the rate of employment on the whole sample, and that this effect is significantly different from 0, showing that the chosen methodology succeed to replicate the dropout penalty identified previously in the literature. By comparing the LATE obtained with TSLS and GRF, it appears that using a linear instrumental variable method to estimate the effect of dropping out on the rate of employment lead us to underestimate the effect by 7 percentage points. The difference between the TSLS and GRF estimators is significant at the 5% level, comforting the use of a highly dimensional method to estimate correctly the dropout effect.

The A(C)LATE on half sample is -29% for the most penalized group, and -22% for the less penalized group. Both A(C)LATE are significantly different from 0 but their difference is only significant at the 10% level, indicating a very small heterogeneity of the dropout effect on the rate of employment, which is in line with the CATE distribution presented in figure 8.

The A(C)LATE estimation per quartiles ranges from -29% for the lowest quartile to -19% for the highest quartile. While all these effects are significantly different from 0, only the difference between Q_1 and Q_4 is significant at the 5% level, reinforcing the fact that dropout show a strong significant effect on the rate of employment only between the extremity of the distribution. On approximately 36 months, a difference of 10 percentage points for the most penalized students compared to the less penalized correspond to a increase of 52% of the dropout penalty. This

	$TSLS$	GRF_1	GRF_2	GRF_3	t-stat
Overall sample	-0.367***	-0.408***			0.69
	(0.087)	(0.058)			
$\tau_1 \leq Median(\tau_i)$			-0.609***		-4.38***
			(0.091)		
$\tau_2 > Median(\tau_i)$			-0.103**		
			(0.071)		
Q_1				-0.720***	-5.08***
				(0.143)	
Q_2				-0.455***	-4.14***
				(0.111)	
Q_3				-0.307***	-3.27***
				(0.104)	
Q_4				0.160*	-
				(0.098)	
Note:					*p<0.1; **p<0.05; ***p<0.01

The standard error is precised in parenthesis. The two first columns indicate the LATE obtain with TSLS and GRF, while the last column shows the t-statistics testing the difference between the both estimators. The third column shows the A(C)LATE for the 50% lower than the median and for the top 50%. The fourth column indicates the A(C)LATE per quartiles. For the effect per quartiles, the t-test is computed with respect to Q_4, with the smaller effect. $LATE \times 100$ gives the estimated effect in percentage.

Table 9: A(C)LATE of dropping out on the average wage

reduction in employment time for the most penalized dropout is strong and has long lasting impact on the labor market path, as well as mechanically on the average wages over the period.

The GRF estimates an overall LATE of -41% of dropping out on the average wage. This effect is significantly different from 0 at the 0.1% level but not from the $LATE_{TSLS}$. However, using only TSLS to estimate the overall effect of dropping out on average wage could lead to underestimation of 4 percentage point of the effect of dropping out on the average wage.

The lowest 50% of the CATE distribution exhibits an effect of -61% while the top 50% shows an effect -10%. The difference between the effects of these two sub-samples is strongly significant, with a t-statistic of -4.38. The results indicates that the overall distribution of the CATE presents actual heterogeneity, and the effect for both 50% is highly different with a gap of 50 percentage point.

The LATE of each quartile go from -72% for Q_1 to a positive effect for Q_4, but only significant at the 10% level. Quartiles Q_1 to Q_3 have a statistically different effect from Q_4 at the 1% level. This indicates that, while 25% of the sample has a decrease of their monthly wage of more than 50%, the top 25% of the sample doesn't show any negative effect, and could even be advantaged when dropping out. We can observe that Q_2 and Q_3 have estimated effect

closer of each others than Q_1 and Q_2, or Q_3 and Q_4, indicating that while some parts of the distribution show extreme effect with a penalty of more than 72%, most of the distribution has an effect centered around the overall A(C)LATE estimated with GRF_1.

This section led us to understand that using highly dimensional estimation methodology is essential to study dropout penalty on labor market outcomes. Dropping out can take many social realities depending on the concerned individuals and only relying to linear estimation is not enough to cover all these cases. We also understood that the effect of dropping out is highly heterogeneous, and studying more closely the social determinants of this heterogeneity is needed to understand the social structure of this phenomenon.

4.4.2 Dropout heterogeneity per diploma, SES and Dip x SES

The objective of the paper is to assess the heterogeneity of the effect of dropping out from higher education, and then to explore which population is the most penalized conditional on its characteristics, either in terms of educational achievements, geographical or social origin. We have seen that, in France, most of the effort is targeted toward the university dropouts, while the vocational section (STS/IUT) dropouts could be as penalized as the university's ones but do not gather as much resources. Finally, I emitted the hypothesis that using a multidimensional understanding of student's profile (and not only their degree) in order to target dropout policies can be beneficial.

In this section, I present the A(C)LATE estimations of dropping out with respect to the highest diploma tried (or obtained) by the student, her socio-economic status and finally the interaction of the both. It is important to recall that, as explained in section 4.1, the structure of the database varied due to a low common support for certain former students. Thus, the estimated results are for individuals who have a non negligible probability of dropping out, and not an overall average effect. For example, 88% of the Bac+5 were discarded due to low probabilities of dropping out for this category of students, and the results presented in table 10 will be valid only for the students who had a chance to drop out during their degree.

The diploma indicator is divided in 5 levels : Bac +2 (STS/IUT), Bac +3 (university), Bac +4, Bac +5 and PhD. The first two levels are those of interest for the analysis, with the Bac +3 level concentrating most of the resources in France. I present the results for the other levels to get a broad understanding of the dropout effect in the french higher education. The highest professional occupation among both parents (noted SES) is defined as follow : disadvantaged corresponds to factory worker and unemployed parents. The intermediate category gathers employee and farmer, the advantaged category gathers intermediary profession, craftsman and independent while the highly advantaged gathers CEO, managers and executives. For the interaction $Diploma \times SES$, the SES are grouped as low and high SES, with disadvantaged and intermediate in the low level, and the rest in the high level.

The main results are highlighted in table 10. We can observe that both subgroups (STS/IUT

		Rate of employment	Average wage
Diploma	Bac +2 (STS/IUT)	-0.242***	-0.362***
		(0.021)	(0.072)
	Bac +3 (university)	-0.282***	-0.178
		(0.047)	(0.131)
	Bac +4	-0.239***	-0.354
		(0.067)	(0.292)
	Bac +5	-0.285***	-0.727***
		(0.049)	(0.198)
	PhD	-0.274	-0.839***
		(0.040)	(0.148)
SES	Disadvantaged	-0.284***	-0.454***
		(0.063)	(0.164)
	Intermediate	-0.220***	-0.384***
		(0.027)	(0.093)
	Advantaged	-0.263***	-0.321***
		(0.034)	(0.117)
	Highly advantaged	-0.267***	-0.452***
		(0.028)	(0.105)
DIP × SES	Bac +2 × Low SES	-0.234***	-0.401***
		(0.031)	(0.102)
	Bac +2 × High SES	-0.250***	-0.324***
		(0.028)	(0.100)
	Bac +3 × Low SES	-0.234**	0.068
		(0.090)	(0.188)
	Bac +3 × High SES	-0.313***	-0.313*
		(0.048)	(0.180)
Note:			*p<0.1; **p<0.05; ***p<0.01

The standard error is precised in parenthesis. The first two lines show the Average Conditional Local Average Treatment Effect of dropping out for the vocational and general undergraduate degrees on their rate of employment and average wage. The following lines indicates the treatment effect of dropping out for students coming from mentioned socioeconomic status, and then the interaction of the degree and the ses background.

Table 10: A(C)LATE for Diploma, SES and Diploma × SES (Rate of employment and Average wage)

and university) have a negative and significant effect of dropping out on the rate of employment. However, the effect is slightly smaller for the STS/IUT, sustaining the historical hypothesis of STS/IUT students being less penalized when they drop out. The STS dropouts have a negative effect of -24% of time in employment, while the university dropouts have an effet of -28%. Both groups of interest present a strong effect of dropping out, while the university dropouts are more penalized by 4 percentage pointss, this slight difference justify to question the actual

distribution of resources.

Regarding the dropout effect on the average wage, STS/IUT dropouts exhibit a negative effect of -36% on the monthly average wage, while the university dropouts show a non significant effect. This results indicate that Bac +2 dropout are highly penalized when dropping out on both labor market oucomes, while Bac +3 dropout are only penalized on their time in employment, but not on their average wage.

While university students may undergo more time unemployed when they enter the labor market after dropping out, they are only penalized on their time in employment, and not on their average wage, as after three years on the labor market, the dropouts don't show any significant difference in earnings with the students who completed the same degree. The STS/IUT students however are strongly penalized when they drop out, on both the rate of employment and the average wage. After three years in the labor market, they exhibit lower time in employment and lower wage than the non-dropouts, which is not the case for university dropouts. The 38% penalty on the average wage for the vocational track dropout is stark and can not be considered as the result of a difficult insertion on the labor market for these students but as a long-lasting effect of the signal sent by dropping out. This setting can indicate that dropout policies actually implemented in France are potentially misdirected, and STS/IUT students should be the focus of more resources and more intense dropout policies.

The other objective of the paper is to understand if it could be efficient for policy makers to rely on other variables than the higher education diploma to design dropout policies. I present results regarding the social origin (SES) of the dropout and the interaction of the higher education degree and the SES in table 10

The LATE of dropping out on the rate of employment of SES levels indicate that the most penalized dropouts are those from either disadvantaged or highly advantaged background. Students coming from advantaged background are also very penalized compared to those from intermediate background who have an effect of -22% of dropping out on their rate of employment. This result, while quite surprising, can be explained by the difference in budget constraint of these students : individuals coming from highly advantaged SES have on the average more parental resources to choose their path after dropping out, thus allowing less time in employment to sustain a correct living level. However, the strongest negative effect is, in line with the vast literature on the topic, still for dropouts of disadvantaged background with an effect of -28%.

Regarding the effect of dropping out on average wage conditional on the SES of the dropout, the structure of the LATE follows the same pattern as the one observed for the rate of employment : the most penalized categories are disadvantaged and highly advantaged. The disadvantaged student can face a penalty of -45% of their average wage while those from highly advantaged background can undergo a penalty of -45% too. Students from advantaged background are the less penalized when dropping out with an effect of -32% on the average wage.

While it is easy to target dropout policy on certain degrees, it is more complicated to target students only based on their SES. However, it is possible to target students from a certain SES inside a given degree, as the SES is most of the time known by the higher education institution. Thus, I estimate the LATE of dropping out for Bac +2 and Bac +3, for low and high SES.

The A(C)LATE for every levels are between -23% to -31%, the most penalized category being the Bac +3 × high SES. We can observe that the negative average effect for the whole category Bac +3 is mostly driven by this level, as the A(C)LATE for Bac +3 is -28%. The effect of dropping out for Bac +2 is sensibly similar between both SES levels. Regarding the effect on the average wage, the most penalized level is Bac +2 × low SES with an effect of -40%, followed by Bac +2 × high SES with an effect of -32% on the average wage. The difference between low and high SES for Bac +2 dropouts is of 8 percentage points, indicating that it is beneficial to consider the dropout as a multidimensional process, and especially targeting from lower SES in Bac +2 would generate higher social benefit. It is worth noticing that individuals dropping out of the university before the master (bac +3) but coming from advantaged background (high SES) are also very penalized with a -31% effect. The more probable explanation relies on the budget constraint idea of dropouts, conditional and proportional to their parents' SES. Following the targeting reasoning, the first population to target is individuals coming from low SES and dropping out of a vocational track degree. However, this classification grid is very restrictive, and it is possible to pursue this analysis with available data for researchers and higher education institutions.

5 Conclusion

The first objective of this paper was to answer if directing most of the resources dedicated to reduce dropout on one of the main two tracks in the higher education system in France was legitimated by poorer performances on the labor market for those dropouts. To answer this question, I estimated the heterogeneous effect of dropping out for students of vocational and general undergraduate degrees using a causal random forest methodology in order to account for the heterogeneous composition of the individuals attending those degrees. The second objective was to state if using multidimensional consideration to structure dropout policy, based on the diploma and the social origin, can helps to create targeting categories to increase the efficiency of the resources allocated to fight the drop-out. The effect of dropping out was estimated on the rate of employment and the average wage, for a period of around three years following the exit of the education system.

To answer those questions, I applied the Generalized Random Forest algorithm with an instrumental variable setting to estimate individual Conditional Average Treatment Effects. Then, observations were grouped conditional either on the CATE or other characteristics such as the highest diploma tried or the SES of the parents, and the average conditional local average treatment effect was computed on each subgroups.

I proceeded in three steps. First, I considered if the dropout effect on labor market outcomes exhibited actual heterogeneity, and if this heterogeneity was statistically significant for different sub-samples (subs-sampling around the median and in quartiles). The distributions of the dropout effect are effectively heterogeneous for both considered labor market outcomes, thus justifying to test for the presence of heterogeneity conditional on more precise classes of students. Second, I tested if Bac +2 (STS/IUT) dropouts were actually less penalized in the long run, compared to Bac +3 university dropouts, which could justify the focus of dropout policy on the laters. I found that it was not the case, as Bac +2 dropouts were more penalized either on the average wage, over a period of around three years on the labor market, but not on the rate of employment. Finally, I estimated the effect of dropping out conditional on the highest diploma tried and the SES of the parents, and found that including this second layer in the targeting of dropout policies can helps to generate more useful categories to target, instead of only structuring the dropout policies around the degree.

This paper opens the discussion about the necessity of considering the potential heterogeneity of educational event such as the dropout, the accumulation of delay or even re-orientation on educational and labor market outcomes of students, to design adapted and efficient educational policies. It also enlight the need to account for the heterogeneous social and individual composition of degrees' population to design national or state policy. While the considered layers of analysis in this paper are reduced, it already shows that taking into account already available information when targeting dropout policy can helps to better understand students

paths, behaviors and potential issues in the higher education system.

List of Figures

List of Tables

References

Adams, R., Almeida, H., & Ferreira, D. (2009). Understanding the relationship between founder–CEOs and firm performance. *Journal of Empirical Finance*, *16*(1), 136–150. https://doi.org/10.1016/j.jempfin.2008.05.002

Agarwal, L., Brunello, G., & Rocco, L. (2021). The pathways to college. *Journal of Human Capital*, *15*(4), 554–595. https://doi.org/10.1086/716343

Aina, C., Baici, E., Casalone, G., & Pastore, F. (2018). The economics of university dropouts and delayed graduation: A survey. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3153385

Aina, C., Baici, E., Casalone, G., & Pastore, F. (2022). The determinants of university dropout: A review of the socio-economic literature. *Socio-Economic Planning Sciences*, *79*, 101102. https://doi.org/10.1016/j.seps.2021.101102

Angrist, J. D., & Krueger, A. B. (1991). Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics*, *106*(4), 979–1014. https://doi.org/10.2307/2937954

Arcidiacono, P., Bayer, P., & Hizmo, A. (2010). Beyond signaling and human capital: Education and the revelation of ability. *American Economic Journal: Applied Economics*, *2*(4), 76–104. https://doi.org/10.1257/app.2.4.76

Athey, S., & Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, *113*(27), 7353–7360. https://doi.org/10.1073/pnas.1510489113

Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, *47*(2). https://doi.org/10.1214/18-AOS1709

Athey, S., & Wager, S. (2019). Estimating treatment effects with causal forests: An application. *arXiv:1902.07409 [stat]*. Retrieved April 28, 2021, from http://arxiv.org/abs/1902.07409

Athey, S., & Wager, S. (2020). Policy learning with observational data. *arXiv:1702.02896 [cs, econ, math, stat]*. Retrieved April 25, 2021, from http://arxiv.org/abs/1702.02896

Becker, G. S. (1993). *Human capital: A theoretical and empirical analysis, with special reference to education* (3rd ed). The University of Chicago Press.

Behr, A., Giese, M., Teguim Kamdjou, H. D., & Theune, K. (2020). Dropping out of university: A literature review. *Review of Education*, *8*(2), 614–652. https://doi.org/10.1002/rev3.3202

Belskaya, V., Sabirianova Peter, K., & Posso, C. M. (2020). Heterogeneity in the effect of college expansion policy on wages: Evidence from the russian labor market. *Journal of Human Capital*, *14*(1), 84–121. https://doi.org/10.1086/706484

Bjerk, D. (2012). Re-examining the impact of dropping out on criminal and labor outcomes in early adulthood. *Economics of Education Review*, *31*(1), 110–122. https://doi.org/10.1016/j.econedurev.2011.09.003

Breiman, L., Friedman, J., Olshen, R., & Stone, C. J. (1983). Classification and regression trees.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Brodaty, T., Gary-Bobo, R., & Prieto, A. (2008). Does speed signal ability? the impact of grade repetitions on employment and wages. *C.E.P.R. Discussion Papers, CEPR Discussion Papers*.

Card, D. (1993). *Using geographic variation in college proximity to estimate the return to schooling* (NBER Working Papers No. 4483). National Bureau of Economic Research, Inc. https://EconPapers.repec.org/RePEc:nbr:nberwo:4483

Chernozhukov, V., Escanciano, J. C., Ichimura, H., Newey, W. K., & Robins, J. M. (2022). Locally robust semiparametric estimation. *Econometrica*, *90*(4), 1501–1535. https://doi.org/10.3982/ECTA16294

Courtioux, P., Gregoir, S., & Houeto, D. (2014). Modelling the distribution of returns on higher education: A microsimulation approach. *Economic Modelling*, *38*, 328–340. https://doi.org/10.1016/j.econmod.2014.01.010

Crump, R. K., Hotz, V. J., Imbens, G. W., & Mitnik, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, *96*(1), 187–199. https://doi.org/10.1093/biomet/asn055

Fox, J., & Monette, G. (1992). Generalized collinearity diagnostics. *Journal of the American Statistical Association*, *87*(417), 178–183. https://doi.org/10.1080/01621459.1992.10475190

Gury, N. (2011). Dropping out of higher education in france: A micro-economic approach using survival analysis. *Education Economics*, *19*(1), 51–64. https://doi.org/10.1080/09645290902796357

Hällsten, M. (2017). Is education a risky investment? the scarring effect of university dropout in sweden. *European Sociological Review*, jcw053. https://doi.org/10.1093/esr/jcw053

Hanushek, E. A., Schwerdt, G., Woessmann, L., & Zhang, L. (2017). General education, vocational education, and labor-market outcomes over the lifecycle. *Journal of Human Resources*, *52*(1), 48–87. https://doi.org/10.3368/jhr.52.1.0415-7074R

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning*. Springer New York. https://doi.org/10.1007/978-0-387-84858-7

Heigle, J., & Pfeiffer, F. (2019). *An analysis of selected labor market outcomes of college dropouts in germany: A machine learning estimation approach. research report*

(ZEW-Gutachten und Forschungsberichte). ZEW - Leibniz-Zentrum für Europäische Wirtschaftsforschung. Mannheim. http://hdl.handle.net/10419/222378

Mahjoub, M.-B. (2017). The treatment effect of grade repetitions. *Education Economics*, *25*(4), 418–432. https://doi.org/10.1080/09645292.2017.1283006

Matkovic, T., & Kogan, I. (2012). All or nothing? the consequences of tertiary education non-completion in croatia and serbia. *European Sociological Review*, *28*(6), 755–770. https://doi.org/10.1093/esr/jcr111

McNamara, S. (2020). Returns to higher education and dropouts: A double machine learning approach. https://doi.org/10.2139/ssrn.3766733

Ménard, B. (2018). Le décrochage dans l'enseignement supérieur à l'aune de l'approche par les capabilités. *Formation emploi*, (142), 119–141. https://doi.org/10.4000/formationemploi.5684

Morlaix, S., & Perret, C. (2013). L'évaluation du plan réussite en licence : Quelles actions pour quels effets ? analyse sur les résultats des étudiants en première année universitaire. *Recherches en éducation*, (15). https://doi.org/10.4000/ree.7390

Oreopoulos, P., & Petronijevic, U. (2013, May). Making college worth it: A review of research on the returns to higher education. https://doi.org/10.3386/w19053

Powell, J. J., & Solga, H. (2010). Analyzing the nexus of higher education and vocational training in europe: A comparative-institutional framework. *Studies in Higher Education*, *35*(6), 705–721. https://doi.org/10.1080/03075070903295829

Psacharopoulos, G., & Patrinos, H. A. (2018). Returns to investment in education: A decennial review of the global literature. *Education Economics*, *26*(5), 445–458. https://doi.org/10.1080/09645292.2018.1484426

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688–701. https://doi.org/10.1037/h0037350

Schnepf, S. V. (2014). *Do tertiary dropout students really not succeed in european labour markets?* (IZA Discussion Papers No. 8015). Institute for the Study of Labor (IZA). Bonn. http://hdl.handle.net/10419/96694

Scholten, M., & Tieben, N. (2017). Vocational qualification as safety-net? education-to-work transitions of higher education dropouts in germany. *Empirical Research in Vocational Education and Training*, *9*. https://doi.org/10.1186/s40461-017-0050-7

Spence, M. (1973). Job market signaling. *The Quarterly Journal of Economics*, *87*(3), 355. https://doi.org/10.2307/1882010

Vignoles, A. F., & Powdthavee, N. (2009). The socioeconomic gap in university dropouts. *The B.E. Journal of Economic Analysis & Policy*, *9*(1). https://doi.org/doi:10.2202/1935-1682.2051

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. https://doi.org/10.1080/01621459.2017.1319839

Appendices

A Descriptive statistics database after discard

	Dropout rate	Frequency	Percentage
Gender			
Male	39.1%	6472	51.5%
Female	32.1%	6098	48.5%
Highest diploma tried			
Bac +2 (STS)	43.0%	7065	56.2
Bac +3 (university)	22.6%	3061	24.4%
Bac +4 (university)	51.3%	943	7.5%
Bac +5 (university/Grande Ecole)	16.5%	588	4.7%
PhD	19.0%	913.00	7.3%
Parents' highest social category			
Disadvantaged	35.8%	2045	16.3%
Intermediate	38.2%	3527	28.1%
Advantaged	34.6%	2661	21.2%
Highly Advantaged	34.2%	4337	34.5%
Parents' highest diploma			
No diploma	36.8%	2606	20.7%
Bac or below	37.4%	5197	41.3%
Short degree	33.9%	3108	24.7%
Long degree	31.8%	1659	13.2%
Other			
Foreign trip (= yes)	42.2%	6166	49.1%
Internship (= yes)	29.3%	6404	51.0%

Table A1: Summary statistics by dropout status, after discard

B GVIF : Assessing potential multicolinearity

	GVIF	Df	GVIF^(1/(2*Df))	(1/(2*Df))^2
Gender	1.46	1.00	1.21	1.46
Highest diploma tried (6 levels)	4.70	4.00	1.21	1.47
SES max parents	1.47	3.00	1.07	1.14
Diploma max parents	1.50	3.00	1.07	1.14
Type of HSD	2.40	2.00	1.24	1.55
HSD grade discretize	1.70	3.00	1.09	1.19
Region higher education institution	1.82	22.00	1.01	1.03
Foreign travel	1.31	1.00	1.14	1.31
Internship	2.54	1.00	1.60	2.54

As suggested in Fox and Monette, 1992, using GVIF^(1/(2*Df)) allows to compare the value of GVIF across different number of parameters. I elevate this measure to the square to use the standard rule of thumb of GVIF. Here, no GVIF goes above 2, so I can safely include and interpret all the parameters in the TSLS model.

C Compliance score : distribution

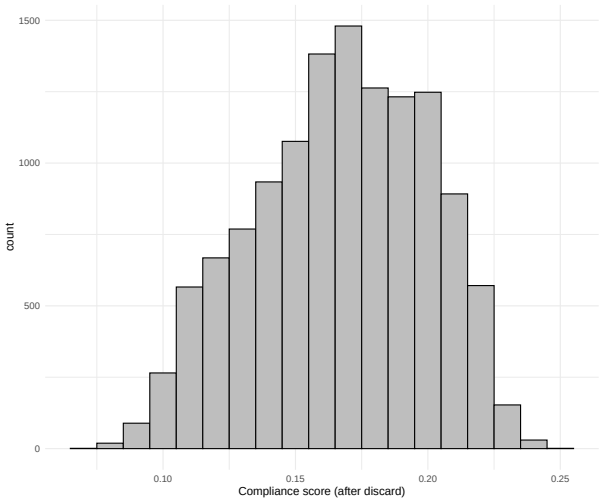


Figure 9: Distribution of the compliance score

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
compliance score	0.06	0.13	0.16	0.16	0.18	0.24

Table A2: Distribution of the compliance score $\mathbb{E}[W_i|Z_i = 1, X_i = x]$

D Full tables TSLS

	Dependent variable:	
	log(roe + 1)	log(av wage + 1)
	(1)	(2)
Constant	0.570^{***}	6.819^{***}
	(0.011)	(0.039)
dropout	-0.192^{***}	-0.369^{***}
	(0.025)	(0.087)
gender2	-0.018^{***}	-0.075^{***}
	(0.004)	(0.013)
nisor_6Bac +4	0.026^{***}	0.138^{***}
	(0.010)	(0.034)
nisor_6Bac +5	0.029^{***}	0.230^{***}
	(0.008)	(0.028)
nisor_6Bac+2	0.020^{***}	0.026
	(0.006)	(0.019)
nisor_6PhD	0.038^{***}	0.449^{***}
	(0.007)	(0.023)
cat_max_tDisadvantaged	0.004	-0.027
	(0.005)	(0.018)
cat_max_tHighly Advantaged	0.018^{***}	-0.010
	(0.005)	(0.016)
cat_max_tIntermediate	0.003	-0.003
	(0.004)	(0.015)
dip_max_tLong degree	-0.026^{***}	0.060^{***}
	(0.005)	(0.019)
dip_max_tNo diploma	-0.010^{**}	0.043^{***}
	(0.004)	(0.015)
dip_max_tShort degree	-0.012^{***}	-0.005
	(0.004)	(0.014)
typeBAC2	0.005	-0.067^{***}
	(0.004)	(0.013)
typeBAC3	0.039^{***}	-0.122^{***}
	(0.007)	(0.025)
mentionBAC2	0.003	0.037^{***}
	(0.004)	(0.014)
mentionBAC3	0.017^{***}	0.104^{***}
	(0.006)	(0.022)
mentionBAC4	0.035^{***}	0.129^{***}
	(0.013)	(0.044)
foreign_travel1	-0.028^{***}	-0.050^{**}
	(0.006)	(0.020)
internship1	-0.017^{***}	-0.028^{*}
	(0.005)	(0.017)
Observations	12,713	12,713
R^2	-0.017	0.095
Adjusted R^2	-0.021	0.092
Residual Std. Error (df = 12671)	0.172	0.595
Note:		*p<0.1; **p<0.05; ***p<0.01

Table A3: TSLS estimation of dropout LATE

E Robustness check Δ IV threshold discretization

After lowering the threshold $\alpha = 0.60$, the number of observations with $Z_i = 1$ is 8201, while the instrument is equal to 0 for 4512 observations. This unbalance is due to the discard phase needed to avoid too low propensity score. We don't observe a major change in the structure of the results, which is encouraging about how chosing $\alpha = 0.60$ is not a determinant choice to obtain the results.

Rate of employment

	GRF_1	GRF_2	t-stat
Overall sample	-0.20.6***		
	(0.0146)		
$\tau_1 \leq Median(\tau_i)$		-0.245***	
		(0.024)	
$\tau_2 > Median(\tau_i)$		-0.216***	
		(0.019)	
Note:	*p<0.1; **p<0.05; ***p<0.01		

Hello

Table A4: Estimation of dropout LATE (rate of employment) with $\alpha = 0.6$

Average wage

	GRF_1	GRF_2	t-stat
Overall sample	-0.520***		
	(0.066)		
$\tau_1 \leq Median(\tau_i)$		-0.466***	
		(0.090)	-2.07
$\tau_2 > Median(\tau_i)$		-0.057	
		(0.087)	
Note:	*p<0.1; **p<0.05; ***p<0.01		

Hello

Table A5: Estimation of dropout LATE (average wage) with $\alpha = 0.6$