ERASMUS UNIVERSITY ROTTERDAM

ERASMUS SCHOOL OF ECONOMICS

Estimating treatment effects with Local Linear Causal Forests: an application to microfinance services and subsidized entrepreneurship training programs

BACHELOR THESIS: DOUBLE BACHELOR BSC^2 IN ECONOMETRICS AND ECONOMICS

Name student: Mihaela ILOVA

Student ID number: 470812

Supervisor: Andrea NAGHI

Second assessor: Sebastiaan VERMEULEN

Date final version: July 4, 2021

The views stated in this thesis are those of the author and not necessarily those of the supervisor, second assessor, Erasmus School of Economics or Erasmus University Rotterdam.

Abstract

Random forests have demonstrated to be an effective method for high-dimensional non-parametric regression across many fields. One of their limitations, however, is their inability to exploit local trends. Linear regressions, on the contrary, fit smooth functions significantly well in low dimensions but they quickly weaken as the dimension increases. In this paper, we investigate the performance of *local linear random forests*, namely a novel estimation method that uses the weights generated by random forests to fit a local linear regression with a ridge penalty for regularization instead of fitting them to a local average. We provide empirical causal inference applications by revisiting two papers from the American Economic Journal that investigate the effectiveness of subsidized entrepreneurship training programs and microfinance services in encouraging entrepreneurship. Additionally, we conduct a simulation study showing the performance of local linear forests in different scenarios by analyzing their robustness against a variety of sources of bias that researchers often need to overcome when inferring causal relationships. Finally, we show the value added of orthogonalization and honesty, two features of generalized random forests that have proven to be effective at overcoming confounding and over-fitting issues.

Contents

1 Introduction

Entrepreneurship has long been seen as a driver for economic growth, competitiveness and job creation. In the era of COVID-19, which has sparked an economic crisis like no other, many people around the world have faced financial and employment hardships. As the world is adjusting to the virus, many countries need to adapt to a new reality and pave the way for economic recovery.

Entrepreneurship, as reflected in the creation, building and scaling of new firms is considered to be "the vaccine to revive economies" (Hwang, 2020). Its importance was for instance highlighted during the 2015 Global Entrepreneurship Summit (GES) held in Nairobi, where the former president of the United States (US) Barack Obama stated that "Entrepreneurship creates new jobs and new businesses, new ways to deliver basic services, new ways of seeing the world — it is the spark of prosperity" (Obama, 2015) .

Entrepreneurship and Small and Medium-sized Enterprises (SMEs) are considered to be the backbone of the economy as they have a sizable impact on employment and income, SMEs generating 45 percent of employment in emerging countries and 70 percent in countries that are part of the Organisation for Economic Co-operation and Development (OECD) (McKinsey, 2020). Entrepreneurship can be a powerful job generation tool, with the promotion of entrepreneurship being central to many developing countries such as Uganda that redesigned its education system to offer entrepreneurship courses in schools and colleges (Obonyo, 2016).

While the pivotal role SMEs play in the growth of economies has been acknowledged for many decades, the COVID-19 crisis recently shed light on the various challenges they are facing. SMEs have been more impacted than large firms by the crisis which has exposed their greater vulnerability. Such disproportionate impact being explained by their greater fragility arising from a lack of cash buffers, weaker supply chain capabilities and a lag in digital inertia (OECD, 2021). More importantly, the main challenges that impede the progress of entrepreneurship area is a lack of funds, mentorship and weak government support. As a consequence, helping SMEs in building resilience has been a rationale for a broad range of stimulus measures around the world, with many countries facilitating SMEs' access to urgent support (OECD, 2014).

Motivated by the central role SMEs play in our economy and the disproportionate effect the COVID-19 crisis has had on them, in this paper we seek to evaluate the effectiveness of two popular tools used by policymakers to tackle the aforementioned challenges, namely subsidized entrepreneurship training programs and microfinance services. Considering the above, the following research question is put forward:

How effective are microfinance services and subsidized entrepreneurship training programs at encouraging entrepreneurship and who do they benefit the most?

In addition to answering the research question mentioned above, we analyze the effectiveness of Local Linear Forests (LLF). Random Forests (RF) have demonstrated to be an effective method for non-parametric regression across many fields (Goldstein et al., 2010; Antipov & Pokryshevskaya, 2012; Arora & Kaur, 2020). A major flaw of random forests however, is their inability to exploit local trends. As a result, Friedberg et al. (2020) have suggested the use of local linear forests which fit smooth functions significantly well in high dimensions. As local linear forests have proven to perform substantially well in the presence of smooth signals, this paper compares the LLF algorithm to the RF algorithm by applying both methods to two data sets. The first data set is obtained from the Growing America through Entrepreneurship (GATE) project provided by Fairlie et al. (2015) who investigated the effectiveness of subsidized entrepreneurship training programs in the US while the second data set is obtained from an experiment on microfinance provided by Field et al. (2013), who investigated the effectiveness of grace periods given to women in low-income neighborhoods in the city of Kolkata, India. Considering the fact that the outcome variables of the first study are predominately binary while the outcome variables of the second study are continuous, the following research question is put forward:

At what extent can local linear forests deliver more accurate treatment effect estimates than the conventional random forests in the presence of smooth signals in the feature space?

Due to the novelty of LLF, only a few studies have shown the estimation improvements brought by the combination of local linear regressions and conventional random forests when drawing statistical inferences about causality. By revisiting the aforementioned studies, we show that LLF can improve the validity of causal analysis and can help detect heterogeneity in treatment effects that traditional methods fail to detect. Additionally, the paper written by Friedberg et al. (2020) is the only study to our knowledge that uses simulations to evaluate empirical errors of LLF applied to causal inferences. While Friedberg et al. (2020) evaluate how effective LLF is at learning smooth heterogeneous treatment effects in the presence of noise, we extend their research by assessing the robustness of LLF against a variety of sources of bias that researchers often need to overcome when performing causality analyses. By analyzing the performance of LLF in different scenarios, we show that LLF improves upon traditional machine learning estimation methods used for causal inferences. Additionally, we delve deeper into the value added of orthogonalization and honesty, two popular features of random forests developed by Athey et al. (2019).

Our paper is organized as follows. In Section 2, we conduct a literature overview on the topics of entrepreneurship and machine learning. Section 3 describes the data used for our empirical applications of random forests. Section 4 yields a comprehensive description of our methodology. Section 5 displays our results. Namely, in Section 5.1 and Section 5.2 we revisit the aforementioned studies with the use of random forests. In Section 5.3, we conduct a simulation study showing the performance of random forests in various scenarios. Finally, in Section 6, we draw a conclusion and suggest recommendations for future research.

2 Literature

2.1 Entrepreneurship

In the past decades, there has been a growing interest in analyzing entrepreneurship in relation to economic growth and as a consequence, economists have given various definitions of the term ”entrepreneur” (Acs et al., 2008; Audretsch et al., 2015; Dejardin et al., 2000). The term was first popularized by Richard Cantillon who identified entrepreneurs as risk-takers (Hébert & Link, 1989). Later, entrepreneurs were seen as economic entities shifting resources from sectors of low productivity to sectors of high productivity (Say & Schumpeter, n.d.). Finally, Schumpeter (1982) identified entrepreneurs as innovators that are key contributors to the economic growth of a region. Despite the various definitions, it is widely accepted that entrepreneurship is " a process that involves the discovery, evaluation, and exploitation of opportunities to introduce new products, services, processes, ways of organizing, or markets" (S. Shane & Venkataraman, 2000).

Researchers started realizing the importance of SMEs in the 1970s and 1980s. With more than 80 percent of all new jobs in the US being created by SMEs in the 1980s, for instance, it became apparent that economic activity has moved away from large firms to SMEs (Toma et al., 2014). As a result, many researchers abandoned their traditional theory that economic growth is driven by large companies and started recognizing SMEs and entrepreneurship as key contributors to job creation and income growth (S. A. Shane, 2007). A substantial amount of attention has been paid to the contribution of entrepreneurship to employment, productivity growth, the production of innovations, and the growth of cities (Van Praag & Versloot, 2007; Acs & Armington, 2006). According to Burns (2016), for instance, entrepreneurship encourages economic growth for three main reasons, namely it stimulates competition, it generates knowledge spillover, and it produces diversity amongst enterprises. Additionally, entrepreneurship has been proven to be an important instrument used to alleviate market inefficiencies (Baum et al., 2007). Today, policy makers are increasingly relying on SMEs to stimulate growth and seek for instruments to encourage entrepreneurship, whether these are microfinance services, entrepreneurship training programs, tax policies or other instruments (S. A. Shane, 2007; Kressel & Lento, 2012).

With lending being considered risky, a substantial amount of academic literature on entrepreneurship focuses on credit constraints. The risk is even higher when lending to entrepreneurs, with empirical evidence showing that entrepreneurs are more likely to engage in risky projects (Vereshchagina & Hopenhayn, 2009). Microfinance has evolved substantially since its emergence in Bangladesh in the 1970s and has, since then, been acknowledged to be an important instrument to support the impoverished (Daley-Harris & Laegreid, 2006). Lending to the impoverished is considered to be even riskier due to the lack of collateral to seize in case of default, nonetheless, evidence has shown that providing microfinance results in significantly higher repayment rates (Banerjee et al., 2015). Despite this success, there is growing evidence that microfinance does not significantly impact poverty and SMEs growth (Karlan & Zinman, 2011). This is notably surprising given that credit constraints have proven to restrict SMEs development (Banerjee & Duflo, 2014). As micro-entrepreneurs often lack the financial knowledge required to make business decisions, providing standard financial training has proven to be helpful at alleviating this issue (Drexler et al., 2014).

Indeed, governments invest billions in subsidized entrepreneurship training programs as evidence shows that training programs can help business owners to get acquainted with techniques that can facilitate their work (Nieuwenhuizen & Kroon, 2002). Van Vuuren & Nieman (1999), for instance, found evidence that entrepreneurial success depends on the product of motivation times entrepreneurial and business skills, hence emphasizing the importance of skills needed to succeed in the market. Additionally, researchers acknowledge the fact that the learning needs of entrepreneurs depend on the company's stages of development, highlighting the need for one-to-one counselling and small group classes. Gorman et al.

(1997) for instance, identified heterogeneity within teaching methods, learning strategies and courses design in their survey on entrepreneurship literature used in post-graduate courses. While most evidence points to the need of entrepreneurship training programs which often have the primary objective to stimulate entrepreneurial drive and to assist participants in identifying products and performing market research, Timmons et al. (1987) recognize the limited effectiveness of such programs and argue that the only way one can learn is through personal experiences. Despite the contradicting findings on the effectiveness of entrepreneurship training programs, Field et al. (2013) acknowledge the re-distributive effects subsidized training programs can have, namely the redistribution of frictions in insurance, labor, credit and human capital markets, resulting in greater equality among subpopulations.

2.2 Machine Learning

Literature on Machine Learning (ML) has been mounting with the increasing need to process larger amounts of complex information. Traditional statistical prediction models, such as regression models, have been widely used for hypothesis testing due to their simplicity and transparency (Zhu & Zhang, 2004). With data becoming more complex, however, researchers are increasingly recognizing the limitations of such traditional methods (Brnabic & Hess, 2021). Their lack of model flexibility and their weak estimation power in the presence of high-dimensional data have motivated researchers to use more data-driven estimation tools such as tree-based techniques (Vamathevan et al., 2019). Most early studies on ML focus on clinical prediction models as genomic and clinical data are often characterized by samples and features sets of considerable sizes (Gawehn et al., 2016). Health care professionals, for instance, have been using patient demographic and health characteristics to predict the likelihood of developing a variety of health complications for decades (Schnabel et al., 2009; D'Agostino et al., 1994; Menden et al., 2013). Although ML has been traditionally used in clinical studies, it is increasingly gaining recognition in the field of economics and social sciences as prediction by itself has become insufficient (Efron, 2020). The literature review shows that research questions are often causal and not predictive, highlighting the larger need to understand the underlying Data Generating Process (DGP). Consequently, this has led to the rise of a new field in ML, namely Causal Machine Learning (CML). As a result, a variety of algorithms such as Model-based Recursive Partitioning (MOB), Bayesian Additive Regression Trees (BART), and Generalized Random Forests (GRF) have been developed to estimate causal effects (Zeileis et al., 2008; Hill, 2011; Athey et al., 2019). CML has popularized itself due to its ability to unveil causal relationships that can be applied to new data sets, hence allowing for model generalizability (Pearl & Mackenzie, 2018). Additionally, CML has gained importance in a variety of business areas. Athey (2017), for instance, investigates the gaps between prediction and causality. Her study on airline price elasticity suggests that predictive models estimate high booking rates during periods of high airline ticket prices while causality unveils the opposite, namely that a rise in prices generates less ticket sales, emphasizing the importance of causal inferences for effective decision-making. CML has also been widely used to identify customers with the highest churn likelihood, a relevant concern for efficient resources allocation (Verhelst et al., 2019; Cook et al., 2004).

In addition to estimating average treatment effects, researchers are interested in understanding how the effect varies with subject characteristics (Knaus et al., 2021). In marketing, for instance, a popular application of heterogeneous treatment effect estimation is customized marketing recommendations and the evaluation of A/B tests (Allenby & Rossi, 1998; Malhotra et al., 1998). However, the most common application of heterogeneous treatment effect estimation is drug trials as it has proven to be particularly useful in assessing the effectiveness of drugs on subpopulations (Saunders et al., 2012; Devi & Scheltens, 2018). The traditional technique of estimating such effects consists of performing linear regressions with interaction terms between the variables and the treatment indicator functions. However, as the number of variables increases, the variable combinations soars and the statistical power of the method plunges. Another drawback of linear regressions is the lack of model flexibility and the imposition of linear relationships. ML algorithms, in contrast, have the advantage of handling high-dimensional data and allowing for nonlinear and interactive relationships between variables. In the past years, a number of algorithms have proven their effectiveness at estimating heterogeneous treatment effects. Athey et al. (2019), for instance, recently developed the Generalized Random Forests (GRF) used for heterogeneous treatment effect estimations, non-parametric quantile regressions and instrumental variable regressions. Greatly similar to random forests, what makes this algorithm relatively more powerful at estimating heterogeneous treatment effects is its main additional features, namely its splitting criterion and the use of honest trees. Although researchers have recognized the importance of capturing heterogeneous treatment effects, several concerns have been raised on the use of data-driven estimation techniques such as tree-based methods. With the aim to find a rationale to release a product or algorithm, researchers and engineers, for instance, are often prone to over-fitting to outliers and to erroneously detecting high treatment effects. In the presence of outliers with high outcome variables researchers are tempted to build a group around them. Although the group appears to have a high treatment effect, the results do not replicate on new data sets. This concern arises in drug trials, for instance, where one might be tempted to find subgroups on which the drug works relatively well as a rationale to approve the drug (Assmann et al., 2000; Cook et al., 2004). Since these post-mining practices generate spurious results that often do not replicate on new data sets, Wager & Athey (2018) emphasize the need to pre-specify research hypotheses before implementing algorithms. Additionally, treatment effect sizes are often small relative to noise, signals might therefore be hard to find for new treatments. Overall, the data-driven nature of ML estimation techniques is a potential source of overfitting and raises concerns with regards to reliability and replicability. Distinguishing between spurious results dependent on sampling variations and real heterogeneous treatment effects is hence necessary in order to reduce the risk of making erroneous causal inferences.

3 Data

3.1 Data on the GATE project

Project Growing America through Entrepreneurship (GATE) is a study organized by the Small Business Administration (SBA) and the US Department of Labor (DOL) in which free entrepreneurship training was randomly provided to entities wishing to start their own company (Fairlie et al., 2015). The project was aimed at helping aspiring entrepreneurs to achieve the American dream of owning their own company. (DOL, 2005). More than 4000 entities applied at fourteen Small Business Development Centers (SBDCs) and nonprofit community-based organizations (CBOs) based across seven sites, SBDCs and CBOs being the dominant procurer of entrepreneurship training services in the US. Due to limited capacity, participants were randomly selected to participate in the project and were subsequently randomly assigned to treatment and control groups. Subjects in the treatment group were provided free training services whereas subjects from the control group were not offered any services. Surveys were subsequently sent to each participant at 6, 18 and 60 months after treatment assignment, giving the opportunity to study and compare a wide variety of outcomes at different points in time. The project's training providers, namely a mix of fourteen SBDCs and CBOs, were carefully selected from both rural and urban locations to accurately represent the US subsidized entrepreneurship training market. Throughout the experiment, several participants quit the study, with 3449 subjects having completed the first wave, 3038 having completed the second wave and 2450 having completed the third wave. A total of 29 covariates are used in this study which can be found in Table A.1 in the Appendix. The training was tailored to each entity, as in the classical subsidized market, with 64 percent of the treatment group receiving one-to-one consulting and 77 percent of the treatment group receiving group training. Introductory classes focused on topics such as business and marketing plans, intermediate classes focused on finances, law and business growth whereas advanced classes covered topics such as accounting, digital sales and information technology. The total value of the training was estimated to be 1,321$ per individual (Benus et al., 2009).

3.2 Data on the microfinance experiment

The second experiment that is analyzed in this paper is an experiment that was conducted by Village Financial Services (VFS), a microfinance institution (MFI) that delivers loans to women in low-income areas in the city of Kolkata, India. A total of 845 clients participated in the study where each client received an individual loan ranging from 4,000 rupees ($90) to 10,000 rupees ($225) (Field et al., 2013). The clients were randomly allocated to a control and treatment group with the control group being assigned to the regular VFS debt contract with repayment starting two weeks after loan delivery and the treatment group being assigned a comparable contract that included an additional grace period of two months. Data on socioeconomic conditions, demographic characteristics and business activities were gathered at three points in time. Namely, eight weeks after the subjects entered the study (survey 1), one year after loan delivery (survey 2), and three years after loan delivery (survey 3). A total of 20 covariates are used when performing analysis. A summary of the covariates can be found in Table A.3 in the Appendix.

4 Methodology

4.1 Instrumental Variables

The paper written by Fairlie et al. (2015) focuses on estimating the effects of receiving entrepreneurship training as opposed to estimating the effects of being offered free entrepreneurship training. Randomized controlled trials (RCT) are often exposed to non-compliance issues as subjects assigned to the treatment group might refuse to receive treatment while subjects assigned to the control group might seek treatment elsewhere (Sagarin et al., 2014). With regards to the GATE project for instance, on one hand, unmotivated subjects offered free training might lose interest, cease attending workshops and fail to schedule one-to-one consulting sessions. Analogously, motivated subjects that were not offered free training might seek training elsewhere and invest time and effort into self-education. As a result, three different groups of people can be identified, namely the "never takers", those that refuse treatment, the "always takers", those that receive treatment, and the "compliers", those that comply and receive treatment if assigned to a treatment group but do not receive treatment when assigned to the control group. In this respect, Fairlie et al. (2015) estimate the Treatment on the Treated (TOT) effect, namely the treatment effect on the compliers. Accordingly, they estimate the Local Average Treatment Effect (LATE) using an Instrumental Variables (IV) estimation method with the first-stage Ordinary Least Squares (OLS) regressions of the form

$$E_{it} = \omega + \gamma X_{ib} + \pi T_{ib} + u_{it} \tag{1}$$

where E_{it} measures whether subject i had obtained any training by wave t, X_{ib} contains the characteristics reported in Table A.1 in the Appendix, and T_{ib} is a an indicator variable that is equal to one if subject i is in the treatment group. The second-stage regressions for the outcome variable of interest y estimated for subject i at wave t are as follows

$$y_{it} = \alpha + \beta X_{ib} + \Delta \widehat{E}_{it} + \varepsilon_{it} \tag{2}$$

where $\widehat{E}_{it}$ is the predicted likelihood of subject i receiving training at wave t, Δ represents the LATE, and u_{it} and ε_{it} are error terms. When estimating heterogeneous treatment effects, interactions between the covariates and the treatment assignment are included in the second-stage regressions.

4.2 Ordinary Least Squares

The Ordinary Least Squares (OLS) estimation method is used by Field et al. (2013) to delve into the impact grace periods have on investment behavior and on micro-enterprise activity in the long-run. The following formula estimates the Average Treatment Effect (ATE) for client i in loan group g

$$y_{ig} = \beta G_g + B_g + \delta X_{ig} + \varepsilon_{ig} \tag{3}$$

where y_{ig} is the outcome of interest, G_g is an indicator variable that equals to one if the group was assigned to the grace period contract and X_{ig} is a vector containing the characteristics of subject i in group g found in Table A.3 in the Appendix. All regressions control for the stratification batch B_g and standard errors are clustered within the loan group g.

4.3 Random Forests

The Random Forests (RF) algorithm developed by Breiman (2001) has demonstrated to be an effective method for non-parametric regression conditional mean estimation. The Generalized Random Forests (GRF) algorithm, implemented by Athey et al. (2019), is an extension to the RF algorithm. While RF is used for conditional means estimation, GRF has more versatile uses and can be used to estimate any solution of local moment equations sets such as non-parametric quantile regressions and instrumental variable regressions. In this section we demonstrate how GRF can be used for non-parametric mean estimation where we observe covariates X_i and the outcome variable Y_i and where we aim to estimate the conditional mean function $\mu(x) = E[Y|X = x]$.

4.3.1 Cluster-Robust Random Forests

A particular feature of the GRF that we are interested in is the clustering of errors. In both studies that we revisit, we expect the outcome Y_i of subject i to be correlated to the site (for the GATE project) or to the loan group (for the microfinance project). Regarding the GATE project, for instance, there might exist heterogeneity across sites in terms of quality of services provided. As a result of the highly customizable nature of the services arising from the one-to-one consulting sessions and the small group workshops, one might expect some consulting teams to be more efficient at advising and helping the participants. Regarding the microfinance project, we expect the spending behavior of the clients to be correlated to the loaned amount. Since we wish our results to generalize outside of the seven sites and outside of the loan groups, we allow for the outcomes of the subjects from the same sites to be correlated within a site and the outcomes of the clients from the same loan group to be correlated within the loan group.

Accordingly, we let Y_i represent the outcome of subject i from site $A_i \in \{1, ..., J\}$ with $J = 7$ (for the GATE project) and the outcome of client i from loan group $A_i \in \{1, ..., J\}$ with $J = 6$ (for the microfinance project). The overall mean and standard error across sites or loan groups is defined as

$$\hat{\mu}_j = \frac{1}{n_j} \sum_{\{i:A_i=j\}} Y_i, \quad \hat{\mu} = \frac{1}{J} \sum_{j=1}^{J} \hat{\mu}_j, \quad \hat{\sigma}^2 = \frac{1}{J(J-1)} \sum_{j=1}^{J} (\hat{\mu}_j - \hat{\mu})^2 \tag{4}$$

where n_j represents the number of subjects in site or loan group j.

As suggested by Abadie et al. (2017), to avoid making assumptions about the distribution of the effect each site has on the subjects' outcome, we use cluster-robust random forests to predict $\mu(x)$. The following steps retrieved from Athey & Wager (2019) are undertaken:

(1) Rather than growing B trees and drawing a subsample $\mathcal{S}_b$ from each tree b=1,...,B, we draw a subsample $\mathcal{J}_b \subseteq$ {1,..., J} of clusters. Next, we take a sample $\mathcal{S}_b$ by taking k samples at random from each cluster $j \in \mathcal{J}_b$.

(2) We grow a tree on each subsample by splitting each subsample into subgroups based on several dichotomous characteristics.

(3) We make the following predictions

$$\hat{\mu}(x) = \frac{1}{B} \sum_{b=1}^{B} \sum_{i=1}^{n} \frac{Y_i 1\left(\{X_i \in L_b(x), i \in \mathcal{S}_b\}\right)}{|\{i : X_i \in L_b(x), i \in \mathcal{S}_b\}|}$$

$$= \sum_{i=1}^{n} Y_i \frac{1}{B} \sum_{b=1}^{B} \frac{1\left(\{X_i \in L_b(x), i \in \mathcal{S}_b\}\right)}{|\{i : X_i \in L_b(x), i \in \mathcal{S}_b\}|}$$

$$= \sum_{i=1}^{n} \alpha_i(x) Y_i \tag{5}$$

where $\text{Ł}_b(x)$ represents the leaf of the b-th tree in which the training sample x is contained. For out-of-bag predictions, namely predictions made for subjects which were not used in the building of the tree, we estimate $\hat{\mu}^{(-i)}(X_i)$ by only taking into account the trees b for which $i \notin \mathcal{S}_b$. In order to take into account the possible correlation between subjects from the same site or loan group, we consider subject i to be out-of-bag if its cluster was not used in step (1), $\mathcal{A}_i \notin \mathcal{J}_b$.

4.3.2 Local Linear Random Forests

Random forests are commonly seen as an ensemble method (Breiman, 2001). Namely, predictions are derived from the average of predictions made by individual trees. Given a point x, the traditional k-nearest neighbors method detects the k closest points to x with respect to Euclidean distance. Analogously, random forests predict using a weighted average of observations in a neighborhood where the neighborhood is defined according to a decision tree and the nearest points to x are those contained in the same leaf. Although, random forests are considered to be a form of nearest-neighbors estimator, they have the unique characteristics of being adaptive. In other words, random forests have the ability to ignore variables that weakly affect the outcome variables. The leaves can for instance, be wider in the direction of stable signals and narrower in the direction of fast-paced signals. As a result, each tree is split on the most influential covariates, allowing random forests to handle high-dimensional covariates spaces and to benefit from an important increase in estimation power. As shown in Athey et al. (2019), random forests can be seen as an adaptive kernel method. The predictions $\hat{\mu}(x)$ (5) made by the forest can hence be rewritten as

$$\hat{\mu}(x) = \sum_{i=1}^{n} \alpha_i(x) Y_i, \quad \alpha_i(x) = \frac{1}{B} \sum_{b=1}^{B} \frac{1\left(\{X_i \in L_b(x), i \in \mathcal{S}_b\}\right)}{|\{i : X_i \in L_b(x), i \in \mathcal{S}_b\}|} \tag{6}$$

where $\alpha_i(x)$ is a kernel that indicates the frequency at which the i-th subject from the training sample is in the same leaf as point x.

Although classical random forests perform very well for non-parametric estimations, they deteriorate in the presence of strong smooth effects. Linear regressions, on the other hand, fit smooth functions significantly well in low dimensions but they quickly weaken as the number of covariates increases. To improve our predictions and confidence intervals, instead of using the weights $\alpha_i(x)$ (6) to fit a local average at x_i , we use them to fit a local linear regression with a ridge penalty for regularization. As suggested by Friedberg et al. (2020) we take the forest weights $\alpha_i(x)$ (6) and we minimize the local average $\mu(x)$ and the slope $\theta(x)$ of the local trend in $X_i - x$, expressed as

$$\begin{pmatrix} \hat{\mu}(x_i) \\ \hat{\theta}(x_i) \end{pmatrix} = \text{argmin}_{\mu,\theta} \left\{ \sum_{i=1}^{n} \alpha_i(x_i) \left(Y_i - \mu(x_i) - (X_i - x_i)\,\theta(x_i)\right)^2 + \lambda \left\|\theta(x_i)\right\|_2^2 \right\} \tag{7}$$

where the ridge penalty $\lambda \left\|\theta(x_i)\right\|_2^2$ prevents over-fitting to the local trend.

4.4 Causal Inference

As mentioned in Section 4.3, the GRF algorithm has more versatile uses than the traditional random forest algorithm. In this section we explain how GRF can be used to make causal inferences.

4.4.1 Potential Outcomes Framework and Treatment Effects

Causal relationships are often analyzed in the potential outcomes framework of Splawa-Neyman et al. (1990) and Rubin (1974). Suppose we observe $\{X_i, Y_i, W_i\}_i^N$ where W_i represents whether subject i received free entrepreneurship training services or a grace period contract , Y_i represents an outcome variable, and X_i is a vector containing the covariates. Using the notation from the potential outcomes model from Imbens & Rubin (2015): $Y_i(1)$ represents the potential outcome of subject i if they received the treatment, and $Y_i(0)$ represents the potential outcome if they did not receive the treatment. To assess the overall effectiveness of the subsidized training and the grace period, we aim to estimate the corresponding individual treatment effect for subject i which can be written as the following: $Y_i(1)-Y_i(0)$. Regrettably, when estimating treatment effects, one can only observe one of these potential outcomes hence making it impossible to calculate the difference for each subject. Nonetheless, one can calculate the average treatment effect (ATE) defined as $\tau := E[Y_i(1) - Y_i(0)]$, with the corresponding conditional average treatment effect (CATE), namely $\tau(x) = E[Y_i(1) - Y_i(0)|X_i = x]$.

With respect to the GATE project, we deviate from the approach taken by Fairlie et al. (2015) in the estimation of treatment effects. While treatment assignment is random, non-compliance is not, hence a relationship between potential outcomes and the actual treatment received is likely to exist. As a result, the comparison of those actually treated with those untreated might lead to biased results (Ebenstein, 2009). Additionally, the estimation of LATE relies on several assumptions such as the exclusion restriction, the non-zero causal effect of the instrument on the treatment variable, and monotonicity (Frölich, 2007). In an attempt to undertake the bias issue, we calculate the Intent to Treat (ITT) effect, capturing the effect the treatment has on all subjects assigned to the treatment group. While the ITT approach gives an unbiased causal estimate, it is often a diluted effect and is likely to underestimate the true value (Angrist, 2006).

With respect to the first study we revisit, namely the paper written by Fairlie et al. (2015), the average impact of entrepreneurship training on business scale is investigated through the comparison of the ATE on business ownership, monthly business sales, hiring of employees and household earnings across the different waves, differentiating between short-run, medium-run and long-run effects represented by the 6 months, 12 months and 60 months follow-up surveys, respectively. A summary of the outcome variables used in this analysis are found in Table A.2 in the Appendix. With respect to the microfinance project studied by Field et al. (2013), we investigate in what manner grace periods influence loan use and we delve into their long-run effects on SME growth trough an analysis of the ATE on long-run weekly profits, income and capital. Table A.4 and Table A.5 in the Appendix show a summary of the corresponding outcome variables. It should be noted that the first paper makes use of 4 continuous variables to investigate the impact of grace periods whereas the first paper makes use of 5. Furthermore, the outcome variables of the first paper are predominately binary whereas the outcome variables of the second paper are entirely continuous.

4.4.2 Assumptions

A natural technique to estimate treatment effects would be to compare the ATE of the control group to the ATE of the treatment group. This is a valid method in randomized experiments where the treatment is randomly assigned. In observational studies, however, treatment assignment is often not random due to unobservable confounding variables that affect both the outcome variable Y_i and the treatment assignment W_i. In order to use the characteristics X_i in our analysis, we first make the assumption of unconfoundedness that can be written as the following

Assumption 1 (Unconfoundedness)

$$Y_i(1), Y_i(0) \perp W_i | X_i \tag{8}$$

Unconfoundedness indicates that the treatment is randomly assigned within each subgroup indexed by $X_i = x$ (Rosenbaum & Rubin, 1983). Namely, once the characteristics of subject i are known, knowing about their corresponding treatment does not give additional information on the potential outcome.

Second, we assume that the probability of a subject being assigned to the treatment given the set of characteristics is bounded between zero and one.

Assumption 2 (Overlap)

$$0 < P[W_i | X_i = x] < 1 \tag{9}$$

This probability is denoted by $e(x)$ and is known as the propensity score. In other words, we assume that every individual in the population can be assigned to the treatment.

4.4.3 Causal Forests

Causal Forests (CF), developed by (Athey et al., 2019), is a treatment effect estimation method used to make causal inferences. As mentioned in Section 4.3, GRF extends on the random forest algorithm displayed in Section 4.3.1 by applying it to treatment effect estimation. The first step of estimating the treatment effect $\tau(x)$ consists of estimating the propensity score $e(x)$ expressed as follows

$$e(x) = P[W_i | X_i = x] \tag{10}$$

and the conditional expected outcome $m(x)$ expressed as follows

$$m(x) = P[Y_i | X_i = x] \tag{11}$$

As mentioned in Section 4.4.2, when performing causal analysis we make the assumption of unconfoundedness (8). In observational studies, however, the propensity score $e(x)$ is often correlated to the subject's characteristics. As a result, k nearest-neighbor matching and other local methods are inconsistent estimators of $\tau(x)$. According to Athey & Wager (2019), $\tau(x)$ can be written as the following

$$\tau(x) = E\left[Y_i \left(\frac{W_i}{e(x)} - \frac{1-W_i}{1-e(x)}\right) \mid X_i = x\right] \tag{12}$$

implying that knowing $e(x)$ allows researchers to have an unbiased estimator of $\tau(x)$. Estimating $e(x)$ and $\tau(x)$ through random forests requires trees to be split on variables that both affect $\tau(x)$ and $e(x)$. According to Athey et al. (2019), however, this is a wasteful practice as modeling propensities does not contribute to heterogeneity estimation. As a solution, Athey et al. (2019) suggests orthogonalization.

Orthogonalization: Namely, $e(x)$ and $m(x)$ are first estimated by using the random forests algorithm found in Section 4.3, producing the respective treatment estimate $\hat{e}(x)$ and outcome estimate $\hat{m}(x)$. Next, the residual treatment $\tilde{W}_i = W_i - \hat{w}^{(-i)}(X_i)$ and the residual outcome $\tilde{Y}_i = Y_i - \hat{y}^{(-i)}(X_i)$ are computed where $\hat{w}^{(-i)}(X_i)$ and $\hat{y}^{(-i)}(X_i)$ are estimated with random forests by performing out-of-bag predictions. Namely, the $i - th$ observation is excluded from the forest and the forest is implemented using the centered treatment $\tilde{W}_i$ and the centered outcome $\tilde{Y}_i$. The weights $\alpha_i(x)$ obtained from the forest are used to estimate τ as follows (Chernozhukov, Chetverikov, et al., 2018; Robinson, 1988)

$$\hat{\tau} = \frac{\sum_{i=1}^{n} \alpha_i(x) \left(\tilde{Y}_i - \hat{m}^{(-i)}(X_i)\right)\left(\tilde{W}_i - \hat{e}^{(-i)}(X_i)\right)}{\sum_{i=1}^{n} \alpha_i(x) \left(\tilde{W}_i - \hat{e}^{(-i)}(X_i)\right)^2} \tag{13}$$

Honesty: As mentioned in Section 2.2, in ML we often worry about spurious results as researchers are often tempted to over-fit to outliers with high outcome values and to erroneously detect high treatment effects. If an observation has a significantly high outcome, for instance, the algorithm might erroneously find subjects with similar characteristics in the control group with average treatment effects. As a result, the group will appear to have a high treatment effect. When applied to another sample, however, the results will not replicate. Even with large data sets, the danger of over-fitting to the original training data set is high. In high-dimensional covariates spaces, for instance, the forest has a lot of ways to partition the trees. As a consequence, in the presence of noise, ML algorithms might model the noise rather than the true treatment effects. To prevent over-fitting, Athey & Imbens (2016) relies on honesty. Honesty refers to using one part of the data for model selection and the other part for model estimation. Namely, the training sample is split into a sample used to construct the trees and a sample used to estimate the effects. By using cross-validation, groups are hence defined in the first sample and hypotheses are tested in the second sample. Honesty is particularly important as it gives a theoretical guarantee to researchers. Athey et al. (2019), for instance, demonstrates that honesty allows researchers to have consistent and asymptotically Gaussian distributed effect estimates and to construct valid confidence intervals with coverage rates that do not deteriorate as the DGP becomes more complex or as more covariates are included (Athey et al., 2019). As a result, researchers can make valid and accurate statistical inferences by making no assumptions on the DGP other than unconfoundedness (8) and overlap (9). The cost of this practice, however, is shallower trees and less personalized predictions.

4.4.4 Heterogeneity

After having assessed the overall effectiveness of the subsidized training and the grace period, we investigate whether treatment heterogeneity is present. The CATE estimates evidently differ across groups. In order to avoid producing spurious results, one might therefore want to test whether the out-of-bag predictions perform better at estimating the CATE than the estimated overall ATE. Consequently, we test whether the heterogeneity found in the CATE of the out-of-bag sample is associated with the heterogeneity in the CATE. In order to test for heterogeneity, we use two approaches.

Differential ATE approach: We group subjects according to their out-of bag CATE estimates. Namely, subjects with out-of-bag CATE estimates above the median CATE estimate are placed in one group and subjects with out-of-bag CATE estimates below the median CATE estimate are placed in another group. Subsequently, the ATE is estimated separately in the two subgroups. The differential ATE is computed by subtracting the ATE of the group with below-median CATE from the ATE of the group with above-median CATE. According to Athey & Wager (2019), however, this method produces weak results about the strength of heterogeneity. As a result, we use a second method to analyze and test the presence of heterogeneity, namely the "Best Linear Predictor" approach of Chernozhukov, Demirer, et al. (2018).

Best Linear Predictor approach: We create two predictors, namely $C_i = \overline{\tau}(W_i - \hat{e}^{(-i)}(X_i))$ and $D_i = (\hat{\tau}^{(-i)}(X_i) - \overline{\tau})(W_i - \hat{e}^{(-i)}(X_i))$, with $\overline{\tau}$ representing the average of the out-of-bag treatment effect estimates. Subsequently, we perform a regression with dependent variable $Y_i - \hat{m}^{(-i)}(X_i)$ and independent variables C_i and D_i. If the estimated coefficient of D_i is significantly close to one, the treatment heterogeneity estimates are well calibrated. The corresponding p-values of the coefficients of D_i are used to test the hypothesis that the heterogeneity found by the causal forest is truthful and non-spurious. If the estimated coefficient of C_i is significantly different from zero, the out-of-bag predictions of the causal forests are correct.

4.4.5 Hypotheses Testing

As mentioned in Section 2.2, post-mining practices generate spurious results that often do not replicate on new data sets. As a result, Wager & Athey (2018) emphasizes the need to pre-specify research hypotheses before implementing algorithms. Consequently, we investigate whether treatment effects deviate across various subpopulations that we specify in this section. Namely for each hypothesis, the sample is divided into two groups, the sub-population of interest and the rest of the population. The differential ATE represents the ATE on the sub-population of interest subtracted by the ATE on the rest of the population. To formally test for heterogeneity, t-tests are performed for each outcome variable.

With regard to the first study we revisit, similar to Fairlie et al. (2015), we test several hypotheses with respect to heterogeneous treatment effects to explore whether training benefits deviate across different subgroups. Consequently, the following four hypothesis are tested:

H1: A rationale for training subsidies is credit constraints since credit and liquidity constraints might inhibit aspiring entrepreneurs from pursuing entrepreneurial activities. However, the GATE project aims to alleviate credit issues trough the provision of assistance and information on credit and liquidity. Additionally, the training might allow participants to be granted access to other sources of funding and financing. As a result, training has a significantly strong positive effect on participants having declared a bad credit history.

H2: Another rationale for training subsidies is labor market discrimination since women and minorities are more likely to be discriminated against in the labor market. Subsidizing the training and making it more affordable can therefore have an important re-distributive effect. As a result, training has a significantly strong positive effect on participants that might be subject to labor market discrimination, such as a) females and b) minorities.

H3: Human and managerial capital constraints are another rationale for the subsidized training since the training might help those lacking human capital components correlated with high business performance. As a result, training has a significantly strong positive effect on subjects with human capital constraints such as a) no college education, b) no previous managerial experience, c) no previous working experience in family business, and d) no previous experience owning a business.

H4: The most popular rationale for subsidized training is reducing unemployment since entrepreneurship training is expected to be significantly useful in motivating unemployed individuals in working and generating a job for themselves. As a result, training has a significantly strong positive effect on participants with unemployment insurance frictions.

With regard to the second study we revisit, similar to Field et al. (2013), we test several hypotheses with respect to heterogeneous treatment effects to explore whether grace period benefits deviate across different subgroups. Consequently, we test the following three hypotheses:

H1: Assuming high return investments are illiquid and risky, the effect of a grace period on clients for whom the risk reduction provided by the grace period is valuable, namely a) risk-averse individuals, and b) those lacking other forms of income-smoothing instruments to buffer against short-run income fluctuations (e.g., savings account), is significantly positive.

H2: Grace periods are inefficiently utilized by impatient individuals that give relatively more importance to the present. As a result, the grace period has a significantly weak effect on present-biased individuals.

H3: Business owners and entrepreneurs have skills required to succeed in entrepreneurial activities. The effect of grace periods is significantly lower for clients lacking those skills.

4.5 Simulations

Evaluating the performance of random forests empirically is difficult, hence we complement our paper with several simulations. Since causal forests are considered to be an adaptive version of nearest neighbor estimators, we compare their performance to a non-adaptive nearest neighbor estimator. Namely, we compare it to the standard k nearest neighbors (k-NN) matching algorithm which estimates the treatment effects as follows

$$\hat{\tau}_{\mathrm{KNN}}(x)=\frac{1}{k} \sum_{i \in S_1(x)} Y_i-\frac{1}{k} \sum_{i \in S_0(x)} Y_i \tag{14}$$

Where S_1 and S_0 are the k nearest neighbors to x in the treatment ($W = 1$) and control ($W = 0$) groups respectively.

The aim of these simulations is to assess the performance of random forests and to evaluate their robustness against a variety of sources of bias that researchers often need to overcome when performing causality analyses. We compare 15-NN, 30-NN, Causal Forest (CF) and Local Linear Causal Forest (LLCF). For each simulation we use training and testing samples of size $n = 1000$ and we vary the dimension such that $d = 2, 4, 6, 10, 15, 20, 25$. We assess the relative performance of the algorithms by comparing the Root Mean Square Error (RMSE), the absolute bias and the coverage of the treatment effect $\tau(x)$ which are averaged over 50 repetitions of each simulation. We conduct our simulations in R, using the packages grf (Athey et al., 2019) for building forests and FNN (Beygelzimer et al., 2013) for k-NN regressions. We use the publicly available code that replicates Table 4 from the paper written by Friedberg et al. (2020) as a base for our analysis . While this code is used to compare the RMSE of the X-Bart estimation method to CF and LLCF for a fixed dimension d, we adjust it to compute the absolute bias and the coverage for various values of d and to produce figures. We use one model structure for all simulations with elements that we modify and adjust according to our research needs. Namely, we let $X \sim U([0,1]^d)$ and use the following outcome model

$$Y=m(x)+W_i . \tau(x)+\epsilon \tag{15}$$

with the following functions

$$\begin{aligned}
&\text{main effect: } m(x)=2^{-1} E\left[Y^{(0)}+Y^{(1)} \mid X=x\right] \\
&\text{treatment effect: } \tau(x)=E\left[Y^{(1)}-Y^{(0)} \mid X=x\right] \\
&\text{treatment propensity: } e(x)=P[W=1 \mid X=x]
\end{aligned}$$

The treatment W_i is drawn from a Bernoulli distribution such that $W_i \sim B(e(x))$ and ϵ is drawn from a normal distribution such that $\epsilon \sim N(0, \sigma)$. The main effect $m(x)$, the treatment propensity $e(x)$ and the standard error σ are adjusted for each simulation.

Simulation 1 - Heterogeneity: First, treatment effects are frequently not constant across population subgroups. Accurate identification of neighborhoods over which the treatment effect is steady is therefore crucial for correct heterogeneity estimation. For our first experiment, we assess the ability of our algorithms to capture heterogeneity by simulating strong smooth heterogeneity along the X_1 and X_2 variables. Similar to Wager & Athey (2018) and Friedberg et al. (2020), we set $\tau(x)$ as follows

$$\tau(x)=\zeta\left(X_1\right) \zeta\left(X_2\right), \quad \zeta(x)=1+\frac{1}{1+\exp (-20(x-1 / 3))} \tag{16}$$

We set $m(x) = 0$ and we assume that we are in a randomized experiment by assuming unconfoundedness and by fixing the propensity score such that $e(x) = 0.5$. Additionally, we draw ϵ form a standard normal distribution such that $\epsilon \sim N(0, 1)$.

Simulation 2 - Unconfoundedness : Second, we investigate how robust our algorithms are against the presence of confounding factors. As mentioned in Section 4.4.2 unconfoundedness (8) states that the treatment is randomly assigned. In observational studies, however, this is often not true. As a result, researchers run the risk of falsely making causal links between treatment and outcome variables leading to distorted and spurious causal claims (Skelly et al., 2012). Motivated by this, we hold the treatment effect fixed at $\tau(x) = 0$ and create an interaction between the main effect $m(x)$ and the propensity score $e(x)$. Accordingly, we set

$$e(x) = \frac{1}{4}\left(1 + \beta_{2,4}\left(X_1\right)\right), \quad m(x) = 2X_1 \tag{17}$$

where $\beta_{2,4}$ is the β-density with shape parameters 2 and 4. In other words, subjects with $X_1 \in [0.1; 0.5]$ are more likely to be assigned to the treatment group. Additionally, we draw ϵ from a standard normal distribution such that $\epsilon \sim N(0, 1)$. In addition to comparing the performance of the aforementioned k-NN and forest algorithms, we assess the value added of orthogonalization by disabling local centering on the CF. As mentioned in Section 4.4.3, the correlation between the treatment assignment and the potential outcomes generates a bias that is adjusted for trough orthogonalization.

Simulation 3 - Noise : Third, we investigate how robust our algorithms are against the presence of noise. Similar to the first simulation, we set $m(x) = 0$ and $e(x) = 0.5$ and we generate heterogeneity along the X_1 and X_2 variables by setting $\tau(x)$ equal to formula (16). To assess the algorithms' ability to estimate heterogeneity in the presence of noise, we set $\sigma = 5$. In addition to comparing the performance of the aforementioned algorithms, we assess the value of honesty by disabling it and running the analog adaptive CF that does not implement data splitting. As mentioned in Section 4.4.3, honesty reduces the risk of modelling noise.

5 Results

5.1 GATE project

5.1.1 Average Treatment Effect

The first paper that we revisit delves into the re-distributive effects of subsidized entrepreneurship training programs and their impact on micro-enterprise growth. Accordingly, the first question brought by Fairlie et al. (2015) asks about the overall effect of subsidized entrepreneurship training programs on business scale. Column 1 of Table 1 (a replication of Table 4 from the original paper) displays the LATEs on business ownership, monthly business sales, hiring of employees and household earnings. The original analysis indicates that the average impact of entrepreneurship training on business ownership at wave 1 is positive and statistically significant with a value equal to 0.134. This suggests that subjects having received training are 13.4 percent more likely to start a new business at wave 1. At wave 2 and 3 , however the effects are smaller and no longer statistically significant. Overall the results from the IV estimation method suggest that entrepreneurship training has positive short-term effects on business ownership. Our ATE estimates and corresponding 95 percent confidence intervals using RF and LLF, displayed in columns 2,3 4 and 5 of Table 1 confirm these findings. Indeed, the ATE confidence interval at wave 1 is positive while the confidence intervals at wave 2 and wave 3 are centered around zero for both RF and LLF. Interestingly, the LATE on monthly sales indicates a negative yet statistically insignificant treatment effect across all waves. In line with this negative relationship, the RF and LLF detect similar effects. Finally, the bottom of Table 1 which examines the impact training has on the hiring of employees and on income indicates no significant effects at any horizon for all three estimation methods, suggesting that the new businesses created at wave 1 had low levels of sales and did not hire employees.

Although the IV results and causal forests results lead to identical conclusions, namely that entrepreneurship training has a significant short-term effect on business ownership, two important differences can be observed. First, the ATE estimates are relatively smaller than the LATE estimates. The IV estimate of the LATE on business ownership at wave 1, for instance, is equal to 0.134, while the RF and LLF estimates of the ATE are equal to 0.056 and 0.054 respectively. Second, the use of RF and LLF results in more accurate treatment effect estimates. The standard error of the IV estimate of the LATE on business ownership at wave 1, for instance, is equal to 0.040 while the RF and LLF standard errors are equal to 0.016 and 0.014 respectively. Although radical differences can be observed between the IV estimation method and the causal forests method, the equivalent cannot be claimed between the RF and LLF estimation methods as they produce similar treatment effect estimates with no observable differences in accuracy.

Table 1: LATEs and ATEs on business ownership, sales, employees and income for the GATE project

	IV	RF		LLF	
Dependent variables	LATE	ATE	Confidence interval	ATE	Confidence interval
Business owner at W1 survey date	0.134 (0.040)	0.054 (0.015)	[0.025 0.083]	0.055 (0.015)	[0.024 0.085]
Business owner at W2 survey date	0.069 (0.057)	0.025 (0.028)	[-0.029 0.080]	0.026 (0.029)	[-0.030 0.082]
Business owner at W3 survey date	0.011 (0.081)	0.002 (0.038)	[-0.073 0.076]	0.001 (0.039)	[-0.076 0.078]
Monthly business sales at W1 survey date (000s)	-0.94 (0.734)	-0.052 (0.033)	[-0.117 0.012]	-0.052 (0.035)	[-0.121 0.017]
Monthly business sales at W2 survey date (000s)	-0.441 (1.115)	-0.003 (0.039)	[-0.080 0.074]	-0.003 (0.038)	[-0.077 0.071]
Monthly business sales at W3 survey date (000s)	-2.552 (2.289)	-0.044 (0.026)	[-0.095 0.008]	-0.044 (0.028)	[-0.098 0.011]
Has any employees at W1 survey date	0.036 (0.025)	0.017 (0.011)	[-0.004 0.038]	0.017 (0.010)	[-0.003 0.037]
Has any employees at W2 survey date	0.007 (0.036)	0.007 (0.018)	[-0.028 0.042]	0.007 (0.019)	[-0.030 0.044]
Has any employees at W3 survey date	-0.087 (0.053)	-0.019 (0.007)	[-0.034 -0.004]	-0.019 (0.007)	[-0.033 -0.004]
log household income at W1 survey date	-0.022 (0.064)	0.002 (0.049)	[-0.095 0.098]	0.002 (0.054)	[-0.103 0.107]
log household income at W2 survey date	0.064 (0.095)	0.029 (0.063)	[-0.094 0.152]	0.028 (0.061)	[-0.091 0.147]
log household income at W3 survey date	0.092 (0.149)	0.032 (0.039)	[-0.045 0.108]	0.031 (0.034)	[-0.035 0.097]

Note: The standard errors are in parenthesis.

5.1.2 Heterogeneity

The next question pertain to treatment heterogeneity. As mentioned in Section 4.4.4 we use two approaches to test the presence of heterogeneity. The first approach to testing heterogeneity consists of grouping clients according to whether their out-of-bag CATE estimates are above or below the median CATE estimate and calculating the difference in ATE between the two groups. The results displayed in columns 1, 2, 6 and 7 of Table 2 vaguely give a first impression of the existence of heterogeneity. Heterogeneity seems to be present solely for business ownership at wave 1. In order to determine whether the causal forests have identified heterogeneity correctly, we use the second approach motivated by the "Best Linear Predictor". The synthetic predictors C_i and D_i together with their standard errors are displayed in columns 3, 4, 8 and 9 of Table 2. The asterisks corresponding to the D_i coefficients indicate that the causal forests successfully capture heterogeneity for business ownership at wave 1, monthly business sales at wave 3 and employees hiring at wave 3. Reassuringly, the C_i coefficients corresponding to the aforementioned variables are statistically different from zero indicating that the "out-of-bag" predictions of the causal forests are correct. To delve deeper into the possible causes of heterogeneity, we test several hypotheses in the following section.

Table 2: Heterogeneity tests: Differential ATEs and calibration tests for the GATE project

	RF				LLF			
Dependent variables	ATE	Confidence interval	C_i	D_i	ATE	Confidence interval	C_i	D_i
Business owner at W1 survey date	-0.012 (0.061)	[-0.131 0.107]	0.836*** (0.201)	0.088** (0.334)	0.000 (0.056)	[-0.109 0.110]	0.900*** (0.234)	0.157** (0.355)
Business owner at W2 survey date	0.053 (0.072)	[-0.087 0.194]	1.228 (1.660)	0.521 (0.284)	0.042 (0.072)	[-0.099 0.182]	1.060 (1.417)	0.326 (0.230)
Business owner at W3 survey date	0.011 (0.073)	[-0.132 0.154]	-0.146 (2.732)	-0.123 (0.421)	-0.056 (0.109)	[-0.269 0.158]	-0.097 (2.449)	-0.121 (0.443)
Monthly business sales at W1 survey date (000s)	0.080 (0.085)	[-0.087 0.246]	0.891 (0.756)	0.720 (0.459)	-0.005 (0.103)	[-0.208 0.198]	1.147 (0.898)	0.757 (0.554)
Monthly business sales at W2 survey date (000s)	-0.006 (0.101)	[-0.204 0.191]	0.061 (2.365)	-0.100 (0.343)	0.022 (0.086)	[-0.146 0.190]	-0.079 (4.591)	-0.310 (0.406)
Monthly business sales at W3 survey date (000s)	0.010 (0.073)	[-0.133 0.153]	0.736* (0.400)	-0.225* (0.838)	0.077 (0.052)	[-0.024 0.178]	0.674* (0.349)	0.082* (0.794)
Has any employees at W1 survey date	0.018 (0.036)	[-0.053 0.088]	1.039 (0.818)	0.178 (0.222)	0.033 (0.035)	[-0.036 0.102]	1.059 (0.835)	0.312 (0.307)
Has any employees at W2 survey date	-0.024 (0.054)	[-0.129 0.081]	-7.106 (13.897)	-0.139 (0.240)	-0.003 (0.051)	[-0.103 0.097]	-2.662 (5.180)	-0.172 (0.215)
Has any employees at W3 survey date	0.004 (0.042)	[-0.079 0.086]	0.955** (0.357)	0.099** (0.341)	0.023 (0.035)	[-0.046 0.092]	0.895** (0.356)	0.232** (0.313)
log household income at W1 survey date	-0.093 (0.156)	[-0.399 0.213]	0.042 (1.911)	-0.025 (0.283)	-0.058 (0.139)	[-0.332 0.215]	0.103 (1.748)	-0.016 (0.327)
log household income at W2 survey date	-0.076 (0.135)	[-0.342 0.189]	0.493 (0.971)	-0.329 (0.195)	-0.031 (0.128)	[-0.282 0.220]	0.513 (1.019)	-0.320 (0.222)
log household income at W3 survey date	-0.127 (0.131)	[-0.384 0.129]	0.838 (0.667)	-0.478 (0.100)	-0.080 (0.116)	[-0.307 0.147]	1.080 (0.760)	-0.356 (0.083)

Note: The standard errors are in parenthesis. The asterisk $*$ corresponds to the significance level of the t-statistic with $^{*}p<0.05$; $^{**}p<0.01$; $^{***}p<0.001$.

5.1.3 Hypotheses testing

Although the aforementioned heterogeneity tests give mixed results, heterogeneity along the participants' characteristics can still be present. In the following section we investigate whether training benefits deviate across the subpopulations pre-specified in section 4.4.5. Table 3, a replication of Table 8B from the paper written by Fairlie et al. (2015), shows the differential LATEs for each subgroup.

Table 3: IV estimates of differential LATEs for various sub-groups for the GATE project

	Credit Constraint	Discrimination		Human capital constraints				UI frictions
Dependent variables	Bad credit	Minority	Female	No college	No manager. exp.	Did not work in fam. bus.	No prior business exp.	Unemployed
Business owner at W1 survey date	0.121 (0.059)	0.085 (0.065)	0.032 (0.063)	0.132 (0.052)	0.081 (0.069)	0.143 (0.049)	0.161 (0.051)	0.228 (0.057)
Business owner at W2 survey date	0.026 (0.068)	-0.017 (0.074)	-0.051 (0.07)	0.062 (0.059)	0.110 (0.078)	0.073 (0.055)	0.045 (0.058)	0.104 (0.063)
Business owner at W3 survey date	-0.085 (0.08)	-0.057 (0.09)	0.006 (0.081)	-0.027 (0.069)	0.129 (0.091)	0.021 (0.062)	0.038 (0.065)	0.001 (0.071)
Monthly business sales at W1 survey date (000s)	-0.442 (0.982)	0.180 (1.03)	-1.049 (0.996)	-0.788 (1.268)	-0.607 (1.365)	-1.804 (1.079)	-0.629 (1.055)	0.127 (1.067)
Monthly business sales at W2 survey date (000s)	2.402 (1.93)	0.729 (1.048)	-1.854 (1.300)	-0.678 (1.405)	1.139 (1.694)	0.693 (1.447)	-0.927 (1.09)	-0.181 (1.229)
Monthly business sales at W3 survey date (000s)	1.281 (1.489)	0.995 (2.019)	-0.250 (1.080)	-0.037 (2.200)	-0.894 (1.866)	-0.581 (1.812)	-1.236 (1.327)	0.920 (1.695)
Has any employees at W1 survey date	0.043 (0.045)	0.078 (0.046)	0.008 (0.046)	0.037 (0.041)	0.100 (0.049)	0.079 (0.036)	0.048 (0.031)	0.044 (0.037)
Has any employees at W2 survey date	0.044 (0.051)	0.011 (0.055)	-0.025 (0.049)	0.013 (0.046)	0.091 (0.055)	0.010 (0.039)	0.003 (0.036)	0.010 (0.041)
Has any employees at W3 survey date	-0.034 (0.053)	-0.021 (0.061)	-0.043 (0.051)	0.015 (0.048)	0.073 (0.061)	-0.013 (0.041)	-0.006 (0.039)	-0.018 (0.045)
log household income at W1 survey date	0.084 (0.141)	0.099 (0.174)	-0.047 (0.136)	-0.051 (0.120)	-0.185 (0.161)	-0.197 (0.100)	-0.226 (0.102)	-0.250 (0.111)
log household income at W2 survey date	0.117 (0.159)	-0.066 (0.175)	0.265 (0.147)	0.023 (0.122)	0.084 (0.178)	-0.030 (0.108)	-0.008 (0.109)	-0.036 (0.122)
log household income at W3 survey date	0.083 (0.178)	0.120 (0.204)	0.055 (0.169)	0.104 (0.146)	0.261 (0.186)	0.128 (0.124)	0.081 (0.135)	0.153 (0.140)

Note: The standard errors are in parenthesis.

Our first hypothesis states that training has a significantly strong positive effect on participants having declared a bad credit history. Table 3 indicates that training has a significantly short-term positive effect on the business ownership of the credit constrained individuals at wave 1 and no heterogeneity evidence is found on the rest of the outcome variables. Table 4, however, corresponding to the RF results, indicates that training has a significantly long-lasting negative effect on the business ownership of the credit-constrained individuals across all waves, a long-term negative effect on the sales at wave 3 and a short term positive effect on income at wave 1 and 2.

Our second hypothesis states that training has a significantly strong effect on participants that might be subject to labor market discrimination, such as females and minorities. The results of the original analysis displayed in Table 3 indicate no strong effects for minorities or females. In reality, the estimates for business ownership are negative for women at wave 1 and wave 2. Our results displayed in Table 4 are consistent with the negative effect the training has on females. Additionally, the results uncover a significantly negative long-run effect on employees for females at wave 3. Similar conclusions are drawn for the minorities with the point estimates of business ownership being for instance negative and significant in the long-run at wave 2 and 3. Overall, our findings suggest that the training discouraged females and minorities to start their own business.

Table 4: RF estimates of differential ATEs for various sub-groups for the GATE project

	Credit Constraint	Descrimination		Human capital constraints				UI frictions
Dependent variables	Bad credit	Minority	Female	No college	No manager. exp.	Did not work in fam. bus.	No prior business exp.	Unemployed
Business owner at W1 survey date	-0.016*** (0.047)	-0.050 (0.032)	-0.098 (0.043)	0.031*** (0.053)	-0.033 (0.040)	-0.009 (0.029)	0.016 (0.033)	0.100*** (0.057)
Business owner at W2 survey date	-0.009*** (0.051)	-0.065 (0.050)	-0.096*** (0.056)	0.068*** (0.063)	0.031 (0.042)	0.014* (0.047)	-0.027*** (0.043)	0.092*** (0.055)
Business owner at W3 survey date	-0.071*** (0.084)	-0.071*** (0.067)	-0.022 (0.071)	0.030*** (0.067)	0.072*** (0.082)	-0.017 (0.059)	0.036 (0.058)	0.018*** (0.060)
Monthly business sales at W1 survey date (000s)	0.021 (0.072)	0.028 (0.050)	0.083 (0.096)	0.003 (0.063)	0.064* (0.073)	-0.062 (0.090)	0.015 (0.123)	0.089 (0.095)
Monthly business sales at W2 survey date (000s)	0.118** (0.091)	0.027 (0.062)	-0.075*** (0.067)	-0.062*** (0.112)	0.026* (0.066)	0.048* (0.105)	-0.045*** (0.077)	0.131 (0.109)
Monthly business sales at W3 survey date (000s)	0.047** (0.047)	0.129 (0.107)	0.016 (0.042)	-0.015*** (0.082)	0.015** (0.058)	-0.014*** (0.081)	0.027 (0.056)	0.092 (0.066)
Has any employees at W1 survey date	0.015 (0.016)	0.010 (0.020)	-0.005* (0.022)	0.010*** (0.018)	0.019 (0.017)	0.046*** (0.022)	-0.010*** (0.022)	0.011 (0.016)
Has any employees at W2 survey date	-0.003*** (0.031)	-0.010*** (0.037)	-0.011*** (0.039)	0.003*** (0.028)	0.018* (0.029)	0.011*** (0.026)	-0.007 (0.026)	0.026 (0.030)
Has any employees at W3 survey date	-0.018 (0.014)	-0.012 (0.015)	0.014*** (0.045)	0.023*** (0.030)	0.030 (0.020)	0.021 (0.019)	0.062*** (0.034)	-0.010 (0.026)
log household income at W1 survey date	0.069** (0.090)	0.104*** (0.101)	0.059 (0.084)	-0.061*** (0.104)	-0.093 (0.086)	-0.028** (0.085)	-0.039* (0.089)	-0.035 (0.085)
log household income at W2 survey date	0.063 (0.100)	-0.088 (0.103)	0.000 (0.102)	0.010 (0.096)	0.011 (0.125)	-0.036 (0.090)	0.008 (0.100)	0.010 (0.092)
log household income at W3 survey date	-0.031*** (0.100)	0.011 (0.100)	-0.002 (0.061)	0.079*** (0.055)	0.115 (0.080)	0.041 (0.071)	0.032 (0.070)	0.024 (0.071)

Note: The standard errors are in parenthesis. The asterisk $*$ corresponds to the significance level of the t-statistic with *p<0.05; **p<0.01; ***p<0.001.

Our next hypothesis states that training has a significantly strong positive effect on participants with human capital constraints and with a) no college education, b) no previous managerial experience, c) no previous working experience in family business, and d) no previous experience owning a business. The original analysis fails to uncover any significant heterogeneity. Strikingly, Table 4, corresponding to the RF estimates, indicates the presence of a positive treatment effect for subjects with no college education at a significance level of 0.1 percent on nine out of 12 of the outcome variables. This strong effect indicates that the training is a relatively good substitute to college and provides elementary knowledge required to be a successful entrepreneur. The results on subjects with no business and managerial experience give somewhat mixed results with largely medium-run strong negative effects that dissipate over time.

Our last hypothesis states that training has a significantly strong effect on participants with unemployment insurance frictions. Table 3 indicates that those unemployed at baseline are more likely to have a business at wave 1. This effect disappears at later waves and no significant effects are found in the longer-run. Our results found in Table 4, however, suggest that training significantly encourages unemployed subjects to create a job for themselves. Although this effect decreases over time, it is significant at a 0.1 percent level across all waves.

Overall, the contradicting treatment effects results obtained from the RF compared to the IV can be explained by several reasons. First, the LATE is often a biased estimate that overestimates or underestimates the true treatment effect , depending on the sizes of the unobserved ”always takers” population and ”never takers” population (Ebenstein, 2009). By not distinguishing between compliers and non-compliers, we seek to give an unbiased estimate of the true treatment effect. Second, causal forests are data-driven estimation methods that estimate treatment effects in a more systematic way than traditional approaches (Baiardi & Naghi, 2021). Causal forests retain the factors that contribute the most to causality and neglect less influential factors. This is particularly useful in larger covariates dimensions where selecting a limited number of covariates significantly reduces noise and standard errors. Furthermore, heterogeneous treatment effects are less likely to be overlooked by the causal forests as a result of their greater ability to dissect highly dimensional relationships (Baiardi & Naghi, 2021).

Although we find radical differences between RF and IV, we find no observable distinctions between RF and LLF, suggesting a lack of smooth signals and a lack of strong local trends as can be observed in Table 4 and Table 5 . Overall, our analysis shows that LLF does not improve on the traditional RF in the presence of covariates and outcome variables that are predominantly binary.

Table 5: LLF estimates of differential ATEs for various sub-groups for the GATE project

	Credit Constraint	Descrimination		Human capital constraints				UI frictions
Dependent variables	Bad credit	Minority	Female	No college	No manager. exp.	Did not work in fam. bus.	No prior business exp.	Unemployed
Business owner at W1 survey date	-0.016*** (0.049)	-0.052 (0.032)	-0.098 (0.042)	0.034*** (0.056)	-0.033 (0.039)	-0.008 (0.029)	0.015 (0.034)	0.098*** (0.059)
Business owner at W2 survey date	-0.010*** (0.053)	-0.066 (0.051)	-0.095*** (0.056)	0.070*** (0.067)	0.031 (0.045)	0.018* (0.049)	-0.028*** (0.046)	0.092** (0.059)
Business owner at W3 survey date	-0.070*** (0.083)	-0.070*** (0.066)	-0.024 (0.070)	0.034*** (0.064)	0.073** (0.081)	-0.014 (0.058)	0.038 (0.059)	0.018*** (0.060)
Monthly business sales at W1 survey date (000s)	0.020 (0.074)	0.028 (0.052)	0.083 (0.100)	-0.001*** (0.091)	0.065*** (0.072)	-0.067 (0.091)	0.022 (0.123)	0.092 (0.095)
Monthly business sales at W2 survey date (000s)	0.116 (0.093)	0.027* (0.070)	-0.078*** (0.070)	-0.056 (0.093)	0.025 (0.069)	0.047 (0.108)	-0.041*** (0.081)	0.131 (0.109)
Monthly business sales at W3 survey date (000s)	0.048*** (0.044)	0.129* (0.103)	0.014 (0.040)	-0.015*** (0.084)	0.013*** (0.057)	-0.015 (0.079)	0.026 (0.055)	0.089* (0.064)
Has any employees at W1 survey date	0.015 (0.016)	0.009 (0.018)	-0.005*** (0.023)	0.011*** (0.020)	0.018 (0.017)	0.047*** (0.023)	-0.010*** (0.024)	0.010 (0.016)
Has any employees at W2 survey date	-0.003*** (0.032)	-0.012*** (0.041)	-0.013 (0.031)	0.003*** (0.030)	0.020*** (0.031)	0.011* (0.028)	-0.006** (0.027)	0.026*** (0.031)
Has any employees at W3 survey date	-0.021 (0.014)	-0.013 (0.015)	0.015*** (0.047)	0.022*** (0.031)	0.032 (0.020)	0.022 (0.018)	0.064*** (0.035)	-0.010 (0.025)
log household income at W1 survey date	0.069** (0.089)	0.102*** (0.099)	0.060 (0.083)	-0.061*** (0.099)	-0.100 (0.088)	-0.032 (0.083)	-0.039 (0.087)	-0.031 (0.082)
log household income at W2 survey date	0.059 (0.100)	-0.087 (0.103)	0.001 (0.103)	0.006 (0.093)	0.010 (0.125)	-0.040 (0.091)	0.009 (0.100)	0.008 (0.093)
log household income at W3 survey date	-0.028*** (0.099)	0.003*** (0.100)	0.001*** (0.058)	0.069 (0.058)	0.122* (0.083)	0.044 (0.065)	0.022 (0.065)	0.018* (0.065)

*Note:*The standard errors are in parenthesis. The asterisk ∗ corresponds to the significance level of the t-statistic with *p<0.05; **p<0.01; ***p<0.001.

5.2 Microfinance experiment

5.2.1 Average Treatment Effect

The second paper that we revisit examines the existence of credit market failures in the entrepreneurship world and whether clients are faced with liquidity and credit constraints that inhibit microentreprise growth. Accordingly, Field et al. (2013) investigate the effect grace periods have on investment behaviour and microentreprise growth.

The first question brought by Field et al. (2013) asks about the effect grace periods have on loan use. For this aim, clients were asked to categorise the use of the borrowed money into several business and non-business related expenditures. Table 6 (representing a replication of Table 1 from the original paper), shows the ATEs and corresponding 95 percent confidence intervals on category-wise spending using OLS. In rows 2 through 4, the business spending is separated into expenditure on inventory and raw materials, business equipment and other operating costs whereas in rows 6 trough 11, the non-business spending is divided into spending on human capital, money for re-lending, savings, food and durable consumption.

In Panel A of Table 6, the original analysis indicates that grace period clients invest significantly more of their loan on business (roughly 6 percent more, 364.9 Rs more) and that this increase in total business spending is mostly caused by higher expenditures on inventory and raw materials which face an increase of roughly 8 percent (corresponding to 367.6 Rs more). From these results we speculate that on average, grace period clients finance additional business investments out of money that would have otherwise be set aside for loan repayment. Additionally, we speculate that grace periods allow clients to expand their investments and take advantage of economies of scale through access to larger wholesale discounts. Our results using RF confirm these findings, although the estimated effects are considerably higher. In column 1 of Panel A of Table 7, we observe an ATE equal to 0.244 on business spending and an ATE equal to 0.193 for inventory and raw materials. However, while OLS detects significance for the ATE on business spending but fails to detect it for the ATE on inventory and raw materials, the RF achieves the opposite, indicated by the fact that the 95 percent confidence interval for the inventory and raw materials estimate does not contain zero.

Table 6: OLS estimates of the ATEs on category-wise spending of the loan for the micro-finance project

		Coefficient on grace period dummy	
Dependent variables	Control group mean	OLS (no controls)	OLS (with controls)
Panel A. Total business spending	6142.4 (162.4)	364.9* (180.1)	383.9* (185.2)
Inventory and raw materials	4521.4 (226.3)	337.1 (279.9)	367.6 (272.8)
Business equipment	1536.5 (172.4)	8.786 (234.1)	-14.4 (227.1)
Operating costs	84.46 (36.91)	19.01 (48.37)	30.75 (49.38)
Panel B. Total nonbusiness spending	1149.1 (149.1)	-356.1* (172.4)	-371.6* (178.7)
Home repairs	557.2 (116)	208.8* (105.1)	-222.1* (110.4)
Utilities, taxes and rent	25.95 (15.66)	8.214 (19.9)	-9.657 (20.66)
Human capital	237.9 (76.88)	34.97 (90.26)	33.06 (91.99)
Money for relending	197.6 (56.74)	-27.42 (70.61)	-30.13 (69.51)
Savings	131.6 (35.97)	15.02 (47.12)	-10.75 (47.48)
Food and durable consumption	151 (76.21)	91.79 (94.11)	-94.73 (97.86)
Panel C. New business	0.02 (0.006)	0.0268* (0.014)	0.0258 (0.014)

Note: The standard errors are in parenthesis. *p<0.05; **p<0.01; ***p<0.001.

In Panel B of Table 6, the original analysis indicates that grace period clients invest significantly less of their loan on non-business spending (roughly 32 percent less, 371.6 Rs less) which is mostly driven by a significant decline in expenditure on home repairs of roughly 40 percent. According to Field et al. (2013), the purchase of construction materials is considered to be a form of informal savings practice amongst the impoverished. Housing materials, which typically consists of bricks and bags of concretes are harder to steal than money, making it a safer investment than cash. Additionally, housing materials are highly liquid and highly divisible due to the fact that unused materials can be readily liquidated and bags of concrete are usually sold individually. Overall the results are consistent with the hypothesis that grace period clients reduce their investments in zero-interest safe assets and increase investment in risky and illiquid assets. As a result, grace periods encourage experimentation with business opportunities and increase the willingness to take on entrepreneurial risk. Our results displayed in Table 7 give rise to

Table 7: RF and LLF estimates of the ATEs on category-wise spending of the loan for the micro-finance project

	RF		LLF	
Dependent variables	ATE	Confidence Interval	ATE	Confidence Interval
Panel A. Total business spending	0.244 (0.173)	[-0.095 0.584]	0.233 (0.142)	[-0.045 0.511]
Inventory and raw materials	0.193 (0.085)	[0.027 0.359]	0.197 (0.084)	[0.033 0.361]
Business equipment	-0.016 (0.035)	[-0.085 0.054]	-0.015 (0.034)	[-0.083 0.052]
Operating costs	-0.013 (0.198)	[-0.401 0.374]	-0.012 (0.166)	[-0.336 0.313]
Panel B. Total nonbusiness spending	-0.114 (0.156)	[-0.420 0.192]	-0.112 (0.157)	[-0.419 0.196]
Home repairs	-0.117 (0.201)	[-0.512 0.278]	-0.118 (0.236)	[-0.581 0.345]
Utilities, taxes and rent	-0.013 (0.199)	[-0.402 0.377]	-0.017 (0.207)	[-0.423 0.390]
Human capital	-0.036 (0.116)	[-0.263 0.190]	-0.034 (0.109)	[-0.247 0.179]
Money for relending	-0.003 (0.183)	[-0.361 0.355]	-0.002 (0.190)	[-0.375 0.371]
Savings	-0.019 (0.101)	[-0.218 0.179]	-0.020 (0.099)	[-0.214 0.174]
Food and durable consumption	-0.021 (0.273)	[-0.555 0.514]	-0.024 (0.241)	[-0.497 0.449]
Panel C. New business	0.142 (0.292)	[-0.431 0.715]	0.139 (0.287)	[-0.425 0.702]

Note: The standard errors are in parenthesis.

equivalent spending patterns. The RF, for instance, detects negative ATEs on all non-business spending categories. However, the negative effect is considerably lower in magnitude and not significant at a 5 percent level. Our results demonstrate that grace periods, on average, lead to a statistically insignificant shift away from safe investments.

In Panel C of Table 6, the original analysis indicates that grace period clients are 129 percent more likely to start a new business. This is indicative of the fact that grace periods not only affected category-wise spending but also the associated risk. Indeed, new ventures and businesses are generally considered risky investments, indicating that grace periods encourage higher risk-taking amongst the clients. In line with the aforementioned positive effect on new businesses, the RF estimates a positive ATE equal to 0.142, a significantly lower and statistically insignificant value.

To summarise, in line with the original analysis, we observe negative effects on non-business spending and positive effects on business spending and risky investments. Although we identify similar relationships, our results suggest stronger effects on business spending and weaker effects on non-business spending and risky investments.

Overall, the RF and LLF produce similar estimates. Contrary to our revisited study of the GATE project, however, we observe a notable difference. Namely the LLF estimates have lower standard errors and smaller confidence intervals for ten out of the twelve dependent variables. These results are particularly insightful as they demonstrate that the use of regression adjustments improves the preciseness of the ATE estimates and suggest the presence of smooth signals in the covariates space.

The second question brought by Field et al. (2013) asks about the effect grace periods have on micro-enterprise activity in the long run. Table 8 (a replication of Panel A of Table 2 from the original paper) shows the estimated effects grace periods have on three measures of long-run business profitability and size, namely monthly profits, log of monthly household income, and capital. As indicated, the OLS results indicate that grace periods cause significantly higher profits and business growth for micro-enterprises. Grace period clients, for instance, report roughly 56 percent higher profits and the difference is statistically significant at a 5 percent level.

In line with these results, our results displayed in Table 9 indicate positive ATEs on both monthly

Table 8: OLS estimates of the ATEs on profit, income and capital for the micro-finance project

	Monthly profit		log of monthly HH income		Capital	
	OLS (no controls)	OLS (with controls)	OLS (no controls)	OLS (with controls)	OLS (no controls)	OLS (with controls)
Grace period	906.6* (373.8)	902.9* (370.2)	0.195* (0.0805)	0.199* (0.0782)	28,770.2* (11,291)	35,733.1** (13020.6)
Control mean	1,586.8 (121.8)	1,586.8 (121.8)	20,172.71 (55,972.25)	20,172.71 (55,972.25)	35,730.2 (5,056)	35,730.2 (5,056)

Note: The standard errors are in parenthesis. *p<0.05; **p<0.01; ***p<0.001.

profits and log of monthly household income, although significance is only detected for the log of monthly household income. The corresponding ATEs for monthly profits are 0.242 and 0.231 for RF and LLF respectively, and 0.201 and 0.195 for log of monthly household income. In conclusion, although on average we observe a positive effect on profits and income, contrary to the analysis of Field et al. (2013), we do not find significant evidence that grace periods cause higher business performance.

Table 9: RF and LLF estimates of the ATEs on profit, income and capital for the micro-finance project

	RF		LLF	
Dependent variables	ATE	Confidence Interval	ATE	Confidence Interval
Monthly profit	0.242 (0.149)	[-0.050 0.533]	0.231 (0.124)	[-0.011 0.473]
log of monthly HH income	0.201 (0.096)	[0.013 0.389]	0.195 (0.085)	[0.028 0.363]
Capital	-0.015 (0.037)	[-0.087 0.057]	-0.017 (0.035)	[-0.086 0.052]

Note: The standard errors are in parenthesis.

5.2.2 Heterogeneity

Next, we investigate whether the treatment effects are homogeneous across all subgroups. As mentioned in Section 4.4.4, the first approach to testing heterogeneity consists of grouping clients according to whether their out-of-bag CATE estimates are above or below the median CATE estimate and calculating the difference in ATE between the two groups. The results using this approach are displayed in columns 1 and 2 of Table 10. The second approach is motivated by the "best linear predictor". The synthetic predictors C_i and D_i, together with their standard errors are displayed in columns 3 and 4 of Table 10. The results from the first approach indicate that the difference in ATEs is statistically insignificant for all three dependent variables using either algorithm, indicated by the fact that the 95 percent confidence intervals are centered around zero. However, the "best linear predictor" approach detects heterogeneity for profits and capital which is indicated by the fact that the corresponding D_i coefficients are statistically significant at a 5 percent level. To delve deeper into the possible causes of heterogeneity, we test several hypotheses in the following section.

Table 10: Heterogeneity test: Differential ATEs and calibration tests for the micro-finance project

Algorithm	Dependent variables	Differential ATE	Confidence interval	C_i	D_i
RF	Monthly profit	0.177 (0.122)	[-0.062 0.416]	1.405* (0.813)	0.634* (0.348)
	log of monthly HH income	0.074 (0.184)	[-0.286 0.434]	0.463 (0.609)	-0.066 (1.432)
	Capital	0.125 (0.161)	[-0.191 0.440]	0.707* (0.318)	0.462* (0.825)
LLF	Monthly profit	0.255 (0.158)	[-0.055 0.565]	1.727 (1.290)	0.690 (0.599)
	log of monthly HH income	0.015 (0.176)	[-0.330 0.360]	0.433 (0.504)	0.188 (0.843)
	Capital	0.209 (0.166)	[-0.116 0.534]	0.695* (0.378)	0.451* (0.774)

Note: The standard errors are in parenthesis. *p<0.05; **p<0.01; ***p<0.001.

5.2.3 Hypotheses testing

Although the aforementioned heterogeneity tests give mixed results, heterogeneity according to client characteristics can still be present. Table 11 and Table 12 indicate that a more targeted and focused heterogeneity analysis uncovers significant heterogeneity on monthly profits along several variables.

Table 11: OLS estimates of differential ATEs for various sub-groups for the micro-finance project

Characteristic	Risk loving	Savings account	Impatient	No household business	Wage earner
Grace period	1,733.1* (710.9)	1,097.1* (710.10)	579.9 (710.11)	1,547.5* (710.12)	906.2* (710.13)
Characteristic x grace period	-1,557.9 (795.4)	-1,420.3 (1071.5)	-1,004.3 (512.9)	-1,208.8 (794.4)	-1,187.4* (568.9)
Characteristic	660.5* (284.1)	1,247.6 (633.1)	34.13 (384.7)	103.4 (247.9)	-1,366.1** (284.6)
Treatment effect evaluated at characteristic	175.2 (304.8)	-323.3 (822.6)	-424.3 (384.4)	338.6 (286.9)	-281.2 (427.9)

Note: The standard errors are in parenthesis. *p<0.05; **p<0.01; ***p<0.001.

Our first hypothesis states that high-return investments are illiquid and risky. The effect of a grace period should hence be more pronounced amongst clients for whom the risk reduction provided by the grace period is particularly valuable, namely clients that are more risk averse and clients that lack alternative income smoothing instruments, such as savings accounts, to buffer against short-term income fluctuations. In line with the hypothesis, the OLS results found in Table 11 (a replication of Panel A of Table 5 from the original paper) uncover negative ATEs for both risk-loving clients and clients with a savings account. Although both types of clients have higher profits on average, namely 660.5 Rs for the risk-loving clients and 1247.6 Rs for the clients with savings accounts, the interaction with the grace period contract leads to lower profits. This is indicated by the negative coefficients of the interaction of the grace period dummy with the characteristics. On average, risk loving clients having received a grace period have 1557.9 less Rs in profits whereas clients with a savings account have 1420.3 less Rs. Although the ATEs are negative, they are statistically insignificant at a 5 percent significance level. As a result, OLS fails to detect heterogeneity along the aforementioned variables. Contrary to the original analysis, our results uncover significant heterogeneity along both characteristics. The estimated ATEs for risk-loving individuals are significant at a 5 percent level and equal to -0.253 and -0.245 for RF and LLF respectively. Likewise, for clients with a savings accounts, Table 12 indicates significance at a 5 percent level for the RF and significance at a 1 percent level for the LLF with corresponding ATE estimates equal to -0.180 and -0.210 accordingly. These findings are insightful as they indicate the possible presence of smooth signals causing heterogeneity that the RF fails to detect. Similarly, for clients with savings accounts, LLF detects stronger significance than RF.

Our second hypothesis states that the grace periods are inefficiently utilized by impatient clients. As a result the effect should be higher for clients that are less present-biased. In line with the hypothesis, the OLS estimation method uncovers negative yet statistically insignificant ATEs for impatient clients. On average, impatient clients having received the grace period earn 14004.3 less Rs than the control group. Our results found in Table 12 are consistent with these findings. The RF, for instance, indicates that impatient clients having received the grace period earn 16.2 percent less in profits whereas clients with a savings accounts earn 18 percent less. Although the ATEs are negative, the performed t-tests fail to detect significance.

Our third and final hypothesis states that the treatment effect should be higher for clients that have

Table 12: RF and LLF estimates of differential ATEs for various sub-groups for the micro-finance project

Algorithm	Dependent variables	Risk loving	Savings account	Impatient	No household business	Wage earner
RF	Monthly profit	-0.253* (0.275)	-0.180* (0.123)	-0.162 (0.187)	-0.184 (0.576)	-0.200 (0.147)
	log of monthly HH income	-0.189*** (0.314)	-0.180* (0.172)	0.099 (0.142)	-0.569 (0.483)	-0.048** (0.122)
	Capital	-0.365** (0.353)	0.200 (0.165)	-0.198 (0.180)	-0.194 (0.241)	-0.217 (0.149)
LLF	Monthly profit	-0.245* (0.284)	-0.210** (0.145)	-0.160 (0.207)	-0.178 (0.458)	-0.176 (0.161)
	log of monthly HH income	-0.163*** (0.303)	-0.132** (0.167)	0.101 (0.154)	-0.626 (0.481)	-0.053** (0.139)
	Capital	-0.370** (0.355)	0.199 (0.163)	-0.206 (0.186)	-0.204 (0.242)	-0.223 (0.148)

Note: The standard errors are in parenthesis. The asterisk $*$ corresponds to the significance level of the t-statistic with *p<0.05; **p<0.01; ***p<0.001..

skills required to succeed in entrepreneurship (proxied by business ownership and past entrepreneurship experience). Consistent with the hypothesis, OLS estimates negative ATEs for both wage earners and clients with no household businesses and detects significance at a 5 percent level for clients with no household business. Contrary to these findings, the causal forests do not detect significant heterogeneity with the exception of heterogeneity of the log of monthly household income along the "wage earner" variable.

To summarize, we draw similar conclusions to Field et al. (2013). Our causal forests results suggests that on average, grace periods encourage entrepreneurs to take more risks and to invest in more illiquid and risky assets. Contrary to the original analysis, however, we detect significant heterogeneity along several variables. We unveil a strong positive heterogeneous effect on micro-enterprise growth among risk-averse clients and clients without savings account, confirming the hypothesis that high-return investments are illiquid and risky. These findings are insightful as they indicate failures in the credit markets and suggest that clients face borrowing constraints that inhibit them from investing in riskier and less liquid assets yielding higher returns.

5.3 Simulations

Evaluating the ability of random forests to learn heterogeneous effects empirically is difficult, hence in this section we analyze the simulations introduced in Section 4.5.

Simulation 1 - Heterogeneity : In our fist set-up aimed at assessing how effective our algorithms are at learning heterogeneity, we first fix the dimension d such that $d = 25$ and we vary the sample size such that $n = 300, 600, 900$. The results are indicated in Table 13. We observe that causal forests present a remarkable improvement over the k-NN matching algorithm indicated by the fact that the RMSE is substantially lower and the coverage is substantially higher for both CF and LLCF. LLCF performs the best in terms of both coverage and RMSE, although the LLCF does not exceed the 80 percent coverage. However, the poor coverage we observe across both forest algorithms is likely because the confidence intervals are built on asymptotic results. We hence expect the 95 percent coverage to not apply for n relatively low. Additionally, we observe that the RMSE of the LLCF decreases at a higher rate than the RMSE of the k-NN matching as n increases. The RMSE of the 30-NN for instance is equal to 1.460 for $n = 300$ and 1.356 for $n = 900$ (corresponding to a 7 percent decrease), while the RMSE of the LLCF is equal to 0.484 for $n = 300$ and 0.286 for $n = 900$ (corresponding to a 40 percent decrease). This is indicative of the fact that one of the strengths of the CF and LLCF lies in their ability to unveil relationships in settings were n and d are large.

Table 13: Simulation 1- RMSE, bias and coverage of $\tau(x)$ averaged over 50 simulation repetitions with d = 25 and $\epsilon \sim N(0,1)$

	15-NN			30-NN			CRF			LLCF		
n	RMSE	Bias	Coverage	RMSE	Bias	Coverage	RMSE	Bias	Coverage	RMSE	Bias	Coverage
300	1.440	0.162	0.380	1.460	0.151	0.151	0.525	**0.095**	0.776	**0.484**	0.117	**0.796**
600	1.381	0.13	0.407	1.394	0.127	0.167	0.385	**0.075**	0.694	**0.356**	0.092	**0.748**
900	1.353	0.119	0.422	1.356	0.112	0.178	0.333	**0.052**	0.685	**0.286**	0.056	**0.787**

Note: Maximizing coverage rates and minimizing RMSE and bias are in bold.

Table 14: Simulation 1- RMSE, bias and coverage of $\tau(x)$ averaged over 50 simulation repetitions with n = 1000 and $\epsilon \sim N(0,1)$

	15-NN			30-NN			CRF			LLCF		
d	RMSE	Bias	Coverage	RMSE	Bias	Coverage	RMSE	Bias	Coverage	RMSE	Bias	Coverage
2	0.398	**0.055**	**0.925**	0.333	0.058	0.893	0.265	**0.055**	0.849	**0.264**	0.058	0.842
4	0.597	0.066	**0.849**	0.652	0.096	0.655	0.271	**0.05**	0.829	**0.266**	0.054	0.835
6	0.824	0.098	0.734	0.874	0.122	0.481	0.288	**0.046**	0.801	**0.277**	0.053	**0.819**
10	1.056	0.150	0.599	1.089	0.156	0.314	0.298	**0.049**	0.764	**0.275**	0.053	**0.799**
15	1.205	0.128	0.510	1.221	0.124	0.238	0.308	**0.052**	0.720	**0.283**	0.064	**0.776**
20	1.283	0.130	0.459	1.292	0.124	0.200	0.322	**0.046**	0.717	**0.287**	0.058	**0.786**
25	1.353	0.144	0.414	1.355	0.140	0.174	0.325	**0.043**	0.679	**0.284**	0.056	**0.769**

Note: Maximizing coverage rates and minimizing RMSE and bias are in bold.

Next, we investigate how our algorithms perform as the dimension increases. We fix n such that $n = 1000$ and we vary d as indicated in Table 14. Figure 1, showing the RMSE plotted against the dimension, indicates that the algorithms perform similarly when the dimension is small ($d = 2$). However, as the dimension increases, the performance of the k-NN matching algorithm deteriorates at an exponential rate, indicated by the concave RMSE curve, while the forest algorithms are robust to positive dimension increments, indicated by the flat RMSE curve. The RMSE of the 30-NN, for instance, increases by 300 percent while the RMSE of the LLCF increases by only 7.5 percent as d increases from 2 to 25. This event is explained by Breiman (2001) and Hastie et al. (2009) who state that the variance of a forest is linked to the product of the correlation between the trees times the variance of each tree. As d increases, the trees have more ways to partition their leafs which reduces the correlation between trees and consequently reduces the overall variance. Similar to Table 13, we observe that the LLCF performs the best in terms of both coverage and RMSE. This is indicative of the fact that the LLCF fits $\tau(x)$ using a better shape than the CF and as a result, is able to better fit strong smooth signals.

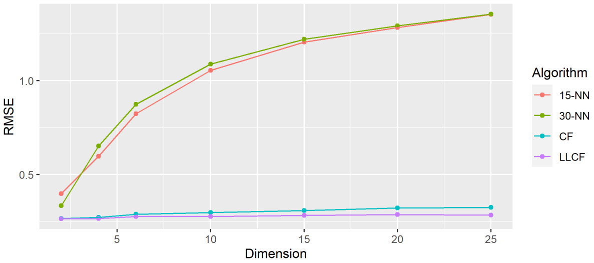

Figure 1: Simulation 1 - RMSE plotted against the dimension with n = 1000

Simulation 2 - Confounding : In our second set-up we asses the ability of our algorithms to overcome the presence of confounding variables. The results are displayed in Table 15. Strikingly, both the CF and LLCF achieve a coverage higher than 95 percent for d bigger than 6. Interestingly, the LLCF performs worse than the CF in terms of RMSE. Additionally, in Figure 2 displaying a plot of the RMSE against the dimension, we observe that the CF without local centering performs as badly as the k-NN algorithms which is explained by the relatively high bias indicated in Table 15. Indeed, the bias of CF without local centering ranges between 0.112 and 0.327 while the bias of the CF with local centering ranges between 0.043 and 0.059. Figure 3, which displays a plot of the estimated $\tau(x)$ against X_1 shows that the CF with local centering estimates and the LLCF estimates are clustered around the true effect which is equal to zero. On the contrary, the CF without local centering estimates together with the k-NN estimates are clustered around -0.3. Without local centering, the estimates are significantly biased, particularly as the dimension becomes larger. Table 15 indicates that as d gets larger, the CF without local centering is dominated by bias, leading to poor coverage rates since the confidence intervals are not centered.

Table 15: Simulation 2 - RMSE, bias and coverage of $\tau(x)$ averaged over 50 simulation repetitions with n $= 1000$ and $\epsilon \sim N(0, 1)$

	15-NN			30-NN			CF without local centering			CF with local centering			LLCF		
d	RMSE	Bias	Coverage	RMSE	Bias	Coverage	RMSE	Bias	Coverage	RMSE	Bias	Coverage	RMSE	Bias	Coverage
2	0.368	0.061	**0.940**	0.268	0.061	0.934	0.198	0.112	0.878	**0.063**	**0.059**	0.868	0.069	**0.059**	0.884
4	0.393	0.082	0.931	0.309	0.117	0.907	0.265	0.209	0.841	**0.060**	**0.053**	**0.967**	0.079	**0.053**	0.938
6	0.418	0.134	0.917	0.337	0.175	0.881	0.291	0.252	0.794	**0.064**	**0.051**	**0.969**	0.075	0.053	0.955
10	0.452	0.205	0.897	0.372	0.234	0.847	0.307	0.28	0.79	**0.064**	**0.053**	**0.963**	0.072	**0.053**	0.957
15	0.47	0.241	0.892	0.389	0.262	0.835	0.315	0.295	0.757	**0.072**	**0.047**	**0.988**	0.084	**0.047**	0.983
20	0.484	0.268	0.882	0.401	0.283	0.822	0.326	0.31	0.713	**0.063**	0.043	0.967	0.072	**0.042**	**0.978**
25	0.507	0.299	0.868	0.423	0.31	0.797	0.342	0.327	0.661	**0.077**	**0.058**	**0.984**	0.091	**0.058**	0.975

Note: Maximizing coverage rates and minimizing RMSE and bias are in bold.

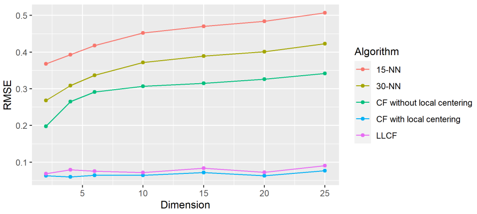

Figure 2: Simulation 2 - RMSE plotted against the dimension with n $= 1000$

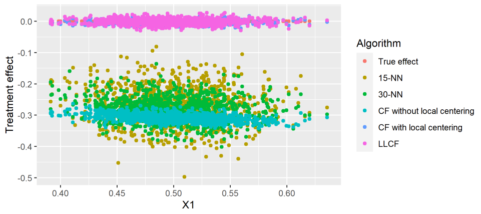

Figure 3: Simulation 2 - $\tau(x)$ plotted against X_1 for n $= 1000$ and d $= 25$

Simulation 3 - Noise : In our last set-up we assess the ability of our algorithms to overcome the presence of noise by setting $\sigma = 5$. The results are displayed in Table 16. First, we observe that the CF and LLCF perform relatively better than the k-NN algorithms, additionally indicated by Figure 4. Next, we observe that the honest CF performs relatively better than the adaptive CF in terms of both RMSE and coverage. As explained in Section 4.4.3, the honest forest splits the training sample into two and uses the first subsample to construct the model and the second subsample to perform estimation while the adaptive analog uses the entire sample for both model construction and estimation. Figure 4 shows that the risk of modelling the noise is relatively lower for honest forests. This is illustrative of the fact that the honest CF does not model noise, but rather captures the overall shape of the treatment effect.

Interestingly, we observe a limitation of the LLCF. Indeed, the LLCF performs worse than the honest CF in terms of both RMSE and coverage for d larger than 15. This is particularly insightful as it shows that the strength of LLCF which consists of modelling heterogeneity more precisely becomes a weakness in the presence of noise. As the dimension gets larger, the LLCF tends to overfit to the data and models the noise rather than the true heterogeneous affect, leading to a higher RMSE.

Table 16: Simulation 3 - RMSE, bias and coverage of $\tau(x)$ averaged over 50 simulation repetitions with n $= 1000$ and $\epsilon \sim N(0, 5)$

	15-NN			30-NN			Adaptive CF			Honest CF			LLCF		
d	RMSE	Bias	Coverage	RMSE	Bias	Coverage	RMSE	Bias	Coverage	RMSE	Bias	Coverage	RMSE	Bias	Coverage
2	1.796	0.232	0.944	1.258	0.234	**0.950**	0.716	0.229	0.736	0.704	**0.228**	0.706	**0.668**	0.237	0.798
4	1.857	0.296	0.935	1.335	0.287	**0.937**	0.868	**0.261**	0.639	0.864	0.270	0.620	**0.808**	0.268	0.776
6	1.865	0.242	**0.936**	1.349	0.247	0.935	0.879	0.252	0.589	0.872	0.251	0.669	**0.841**	**0.242**	0.792
10	1.904	0.308	**0.930**	1.423	0.317	0.920	0.954	0.262	0.625	**0.902**	**0.272**	0.685	0.928	0.277	0.744
15	1.931	0.285	**0.926**	1.457	0.279	0.912	1.009	0.239	0.626	0.929	0.235	0.715	**0.970**	**0.232**	0.782
20	1.935	0.287	**0.927**	1.451	0.280	0.915	0.978	**0.231**	0.608	**0.920**	0.232	0.731	0.941	0.235	0.766
25	1.959	0.401	**0.923**	1.479	0.388	0.909	1.049	**0.289**	0.705	**0.953**	0.294	0.746	0.979	0.292	0.789

Note: Maximizing coverage rates and minimizing RMSE and bias are in bold.

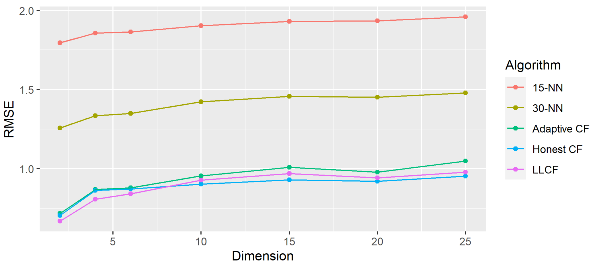

Figure 4: Simulation 3 - RMSE plotted against the dimension with n $= 1000$ and $\epsilon \sim N(0, 5)$

6 Conclusion

In conclusion, we provided evidence that local linear forests are a powerful tool for treatment effect estimation. By revisiting two studies from the American Economic Journal, we provided insightful information on the effectiveness and on the re-distributive effects of subsidized entrepreneurship training programs and microfinance services. Our results suggest that entrepreneurship training programs have a significant positive short-term effect on business ownership that gradually dissipates over time. Although the GATE project was aimed at destroying the barriers hindering entrepreneurship amongst the constrained individuals, it discouraged females and credit-constrained individuals yet stimulated and helped the unemployed and the less educated individuals. Our results on grace period contracts suggest that postponing the repayment of loans encourages entrepreneurs to finance additional business investments out of money that would have otherwise be set aside for initial loan repayment. Additionally, we provided evidence that grace period contracts significantly benefit risk-averse clients and clients without savings account, suggesting the existence of failures in the credit markets that inhibit entrepreneurs from investing in riskier and less liquid assets yielding higher returns. Moreover, we showed that causal forests estimate treatment effects in a more systematic way than traditional approaches. As a result, we provided more precise treatment effect estimates and we detected heterogeneity that traditional estimation methods overlooked. Our empirical applications, finally, showed that local linear forests are a more precise estimation method when applied to continuous variables yet do not improve upon the traditional random forests when applied to binary variables. Interestingly, our simulation studies suggest that they are less robust against the presence of confounding factors and noise although they are substantially more accurate when learning smooth heterogeneous effects in randomized experiments with relatively low levels of noise. Finally, we showed that orthogonalization and honesty are powerful features of the random forests that contribute to their robustness against potential sources of bias.

Since we are the first to suggest limitations to local linear forests, future research could further explore their performance in the presence of noise and confounding variables. Although we showed that the strength of local linear forests lies in their higher predictive accuracy in the presence of strong smooth signals, our findings suggest that they are less suitable for observational studies. Hence, it would be fruitful to further develop and confirm these initial findings. Moreover, our findings indicate that these limitations appear in the presence of high-dimensional covariates spaces, consistent with the idea that tree-based methods suffer from data overfitting as the dimension increases and as the trees have more splitting alternatives. Therefore, future studies could explore the role of the ridge penalty used to prevent over-fitting to the local trend. Considering we used the default settings of the grf package, it would be of considerable interest to experiment with various ridge penalty values and to develop a better understanding of the value added of such penalty.

References

Abadie, A., Athey, S., Imbens, G. W., & Wooldridge, J. (2017). *When should you adjust standard errors for clustering?* (Tech. Rep.). National Bureau of Economic Research.

Acs, Z. J., & Armington, C. (2006). *Entrepreneurship, geography, and american economic growth.* Cambridge University Press.

Acs, Z. J., Desai, S., & Hessels, J. (2008). Entrepreneurship, economic development and institutions. *Small business economics*, *31*(3), 219–234.

Allenby, G. M., & Rossi, P. E. (1998). Marketing models of consumer heterogeneity. *Journal of econometrics*, *89*(1-2), 57–78.

Angrist, J. D. (2006). Instrumental variables methods in experimental criminological research: what, why and how. *Journal of Experimental Criminology*, *2*(1), 23–44.

Antipov, E. A., & Pokryshevskaya, E. B. (2012). Mass appraisal of residential apartments: An application of random forest for valuation and a cart-based approach for model diagnostics. *Expert Systems with Applications*, *39*(2), 1772–1778.

Arora, N., & Kaur, P. D. (2020). A bolasso based consistent feature selection enabled random forest classification algorithm: An application to credit risk assessment. *Applied Soft Computing*, *86*, 105936.

Assmann, S. F., Pocock, S. J., Enos, L. E., & Kasten, L. E. (2000). Subgroup analysis and other (mis) uses of baseline data in clinical trials. *The Lancet*, *355*(9209), 1064–1069.

Athey, S. (2017). Beyond prediction: Using big data for policy problems. *Science*, *355*(6324), 483–485.

Athey, S., & Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, *113*(27), 7353–7360.

Athey, S., Tibshirani, J., Wager, S., et al. (2019). Generalized random forests. *Annals of Statistics*, *47*(2), 1148–1178.

Athey, S., & Wager, S. (2019). Estimating treatment effects with causal forests: An application. *arXiv preprint arXiv:1902.07409*.

Audretsch, D. B., Belitski, M., & Desai, S. (2015). Entrepreneurship and economic development in cities. *The Annals of Regional Science*, *55*(1), 33–60.

Baiardi, A., & Naghi, A. A. (2021). The value added of machine learning to causal inference: Evidence from revisited studies. *arXiv preprint arXiv:2101.00878*.

Banerjee, A., & Duflo, E. (2014). Do firms want to borrow more? testing credit constraints using a directed lending program. *Review of Economic Studies*, *81*(2), 572–607.

Banerjee, A., Duflo, E., Glennerster, R., & Kinnan, C. (2015). The miracle of microfinance? evidence from a randomized evaluation. *American Economic Journal: Applied Economics*, *7*(1), 22–53.

Baum, J. R., Frese, M., Baron, R. A., & Katz, J. A. (2007). Entrepreneurship as an area of psychology study: An introduction. *The psychology of entrepreneurship*, *1*, 18.

Benus, J., Shen, T., Zhang, S., Chan, M., & Hansen, B. (2009). Growing america through entrepreneurship: Final evaluation of project gate. *IMPAQ International*, 1–243.

Beygelzimer, A., Kakadet, S., Langford, Arya, S., Mount, D., & Li, S. (2013). Fnn: Fastnearestneighbor search algorithms and applications. *R package*.

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

Brnabic, A., & Hess, L. M. (2021). Systematic literature review of machine learning methods used in the analysis of real-world data for patient-provider decision making. *BMC medical informatics and decision making*, *21*(1), 1–19.

Burns, P. (2016). *Entrepreneurship and small business.* Palgrave Macmillan Limited.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). *Double/debiased machine learning for treatment and structural parameters.* Oxford University Press Oxford, UK.

Chernozhukov, V., Demirer, M., Duflo, E., & Fernandez-Val, I. (2018). *Generic machine learning inference on heterogenous treatment effects in randomized experiments* (Tech. Rep.). National Bureau of Economic Research.

Cook, D. I., Gebski, V. J., & Keech, A. C. (2004). Subgroup analysis in clinical trials. *Medical Journal of Australia*, *180*(6), 289.

D'Agostino, R. B., Wolf, P. A., Belanger, A. J., & Kannel, W. B. (1994). Stroke risk profile: adjustment for antihypertensive medication. the framingham study. *Stroke*, *25*(1), 40–43.

Daley-Harris, S., & Laegreid, L. (2006). *State of the microcredit summit campaign: report 2006.* Microcredit Summit Campaign Washington, DC.

Dejardin, M., et al. (2000). Entrepreneurship and economic growth: An obvious conjunction. *Institute for Development Strategies*.

Devi, G., & Scheltens, P. (2018). Heterogeneity of alzheimer's disease: consequence for drug trials? *Alzheimer's research & therapy*, *10*(1), 1–3.

DOL. (2005). *Project gate final evaluation dataset.* Retrieved from `https://www.dol.gov/agencies/eta/reports/project-gate-dataset`

Drexler, A., Fischer, G., & Schoar, A. (2014). Keeping it simple: Financial literacy and rules of thumb. *American Economic Journal: Applied Economics*, *6*(2), 1–31.

Ebenstein, A. (2009). When is the local average treatment close to the average? evidence from fertility and labor supply. *Journal of Human Resources*, *44*(4), 955–975.

Efron, B. (2020). Prediction, estimation, and attribution. *International Statistical Review*, *88*, S28–S59.

Fairlie, R. W., Karlan, D., & Zinman, J. (2015). Behind the gate experiment: Evidence on effects of and rationales for subsidized entrepreneurship training. *American Economic Journal: Economic Policy*, *7*(2), 125–61.

Field, E., Pande, R., Papp, J., & Rigol, N. (2013). Does the classic microfinance model discourage entrepreneurship among the poor? experimental evidence from india. *American Economic Review*, *103*(6), 2196–2226.

Friedberg, R., Tibshirani, J., Athey, S., & Wager, S. (2020). Local linear forests. *Journal of Computational and Graphical Statistics*, 1–15.

Frölich, M. (2007). Nonparametric iv estimation of local average treatment effects with covariates. *Journal of Econometrics*, *139*(1), 35–75.

Gawehn, E., Hiss, J. A., & Schneider, G. (2016). Deep learning in drug discovery. *Molecular informatics*, *35*(1), 3–14.

Goldstein, B. A., Hubbard, A. E., Cutler, A., & Barcellos, L. F. (2010). An application of random forests to a genome-wide association dataset: methodological considerations & new findings. *BMC genetics*, *11*(1), 1–13.

Gorman, G., Hanlon, D., & King, W. (1997). Some research perspectives on entrepreneurship education, enterprise education and education for small business management: a ten-year literature review. *International small business journal*, *15*(3), 56–77.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learnin. *Cited on*, 33.

Hébert, R. F., & Link, A. N. (1989). In search of the meaning of entrepreneurship. *Small business economics*, *1*(1), 39–49.

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

Hwang, V. W. (2020). *Entrepreneurship is the vaccine for urban economies.* Retrieved from `https://www.bloomberg.com/news/articles/2020-10-21/entrepreneurship-will-accelerate-coronavirus-recovery`

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

Karlan, D., & Zinman, J. (2011). Microcredit in theory and practice: Using randomized credit scoring for impact evaluation. *Science*, *332*(6035), 1278–1284.

Knaus, M. C., Lechner, M., & Strittmatter, A. (2021). Machine learning estimation of heterogeneous causal effects: Empirical monte carlo evidence. *The Econometrics Journal*, *24*(1), 134–161.

Kressel, H., & Lento, T. V. (2012). *Entrepreneurship in the global economy: Engine for economic growth.* Cambridge University Press.

Malhotra, N. K., Agarwal, J., & Baalbaki, I. (1998). Heterogeneity of regional trading blocs and global marketing strategies. *International Marketing Review*.

McKinsey. (2020). *Setting up small and medium-size enterprises for restart and recovery.*

Menden, M. P., Iorio, F., Garnett, M., McDermott, U., Benes, C. H., Ballester, P. J., & Saez-Rodriguez, J. (2013). Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties. *PLoS one*, *8*(4), e61318.

Nieuwenhuizen, C., & Kroon, J. (2002). Identification of entrepreneurial success factors to determine the content of entrepreneurship subjects: Research in higher education. *South African Journal of Higher Education*, *16*(3), 157–166.

Obama, B. (2015). *Remarks by president obama at the global entrepreneurship summit.* Retrieved from `https://obamawhitehouse.archives.gov/the-press-office/2015/07/25/remarks-president-obama-global-entrepreneurship-summit`

Obonyo, R. (2016). *Africa looks to its entrepreneurs a useful strategy in the toolbox to reduce youth unemployment.* Retrieved from `https://www.un.org/africarenewal/magazine/april-2016/africa-looks-its-entrepreneurs`

OECD. (2014). *Small businesses continue to face finance constraints despite economic recovery.* Retrieved from `https://www.oecd.org/industry/small-businesses-continue-to-face-finance-constraints-despite-economic-recovery.htm`

OECD. (2021). *One year of sme and entrepreneurship policy responses to covid-19: Lessons learned to "build back better".* Retrieved from `https://www.oecd.org/coronavirus/policy-responses/one-year-of-sme-and-entrepreneurship-policy-responses-to-covid-19-lessons-learned-to-build-back-better-9a230220/`

Pearl, J., & Mackenzie, D. (2018). Ai can't reason why. *Wall Street Journal*.

Robinson, P. M. (1988). Root-n-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society*, 931–954.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, *66*(5), 688.

Sagarin, B. J., West, S. G., Ratnikov, A., Homan, W. K., Ritchie, T. D., & Hansen, E. J. (2014). Treatment noncompliance in randomized experiments: Statistical approaches and design issues. *Psychological methods*, *19*(3), 317.

Saunders, N. A., Simpson, F., Thompson, E. W., Hill, M. M., Endo-Munoz, L., Leggatt, G., … Guminski, A. (2012). Role of intratumoural heterogeneity in cancer drug resistance: molecular and clinical perspectives. *EMBO molecular medicine*, *4*(8), 675–684.

Say, J.-B., & Schumpeter, J. (n.d.). Théories de l'entrepreneur.

Schnabel, R. B., Sullivan, L. M., Levy, D., Pencina, M. J., Massaro, J. M., D'Agostino Sr, R. B., … others (2009). Development of a risk score for atrial fibrillation (framingham heart study): a community-based cohort study. *The Lancet*, *373*(9665), 739–745.

Schumpeter, J. A. (1982). The theory of economic development: An inquiry into profits, capital, credit, interest, and the business cycle (1912/1934). *Transaction Publishers.–1982.–January*, *1*, 244.

Shane, S., & Venkataraman, S. (2000). The promise of entrepreneurship as a field of research. *Academy of management review*, *25*(1), 217–226.

Shane, S. A. (2007). *Economic development through entrepreneurship: Government, university and business linkages.* Edward Elgar Publishing.

Skelly, A. C., Dettori, J. R., & Brodt, E. D. (2012). Assessing bias: the importance of considering confounding. *Evidence-based spine-care journal*, *3*(01), 9–12.

Splawa-Neyman, J., Dabrowska, D. M., & Speed, T. (1990). On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, 465–472.

Timmons, J. A., Muzyka, D. F., Stevenson, H. H., & Bygrave, W. D. (1987). Opportunity recognition: The core of entrepreneurship. *Frontiers of entrepreneurship research*, *7*(2), 109–123.

Toma, S.-G., Grigore, A.-M., & Marinescu, P. (2014). Economic development and entrepreneurship. *Procedia Economics and Finance*, *8*, 436–443.

Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., … others (2019). Applications of machine learning in drug discovery and development. *Nature Reviews Drug Discovery*, *18*(6), 463–477.

Van Praag, C. M., & Versloot, P. H. (2007). What is the value of entrepreneurship? a review of recent research. *Small business economics*, *29*(4), 351–382.

Van Vuuren, J., & Nieman, G. (1999). Entrepreneurship education and training: A model for syllabi/curriculum development. In *44th icsb world conference proceedings.*

Vereshchagina, G., & Hopenhayn, H. A. (2009). Risk taking by entrepreneurs. *American Economic Review*, *99*(5), 1808–30.

Verhelst, T., Caelen, O., Dewitte, J.-C., Lebichot, B., & Bontempi, G. (2019). Understanding telecom customer churn with machine learning: From prediction to causal inference. In *Artificial intelligence and machine learning* (pp. 182–200). Springer.

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.

Zeileis, A., Hothorn, T., & Hornik, K. (2008). Model-based recursive partitioning. *Journal of Computational and Graphical Statistics*, *17*(2), 492–514.

Zhu, H.-T., & Zhang, H. (2004). Hypothesis testing in mixture regression models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *66*(1), 3–16.

Appendix

Table A.1: Treatment variable and covariates used for the GATE project

W	treatment	The treatment variable which equals to 1 if the subject was given access to free entrepreneurship training, 0 otherwise
X_1	site1	Dummy variable which equals to 1 if the subject is from the Philadelphia site, 0 otherwise
X_2	site2	Dummy variable which equals to 1 if the subject is from the PiPh pattsburgh site, 0 otherwise
X_3	site3	Dummy variable which equals to 1 if the subject is from the Minneapolis-St.Paul site, 0 otherwise
X_4	site4	Dummy variable which equals to 1 if the subject is from the Duluth site, 0 otherwise
X_5	site5	Dummy variable which equals to 1 if the subject is from the Maine site, 0 otherwise
X_6	female	Dummy variable which equals to 1 if the subject is female, 0 otherwise
X_7	black	Dummy variable which equals to 1 if the subject is black, 0 otherwise
X_8	latino	Dummy variable which equals to 1 if the subject is latino, 0 otherwise
X_9	asian	Dummy variable which equals to 1 if the subject is asian, 0 otherwise
X_{10}	other	Dummy variable which equals to 1 if the subject has race ”other”, 0 otherwise
X_{11}	notusborn	Dummy variable which equals to 1 if the subject is not born in the US, 0 otherwise
X_{12}	age	Age of the subject
X_{13}	married	Dummy variable which equals to 1 if the subject is married, 0 otherwise
X_{14}	children	Dummy variable which equals to 1 if the subject has children, 0 otherwise
X_{15}	grade	Highest grade in highschool completed
X_{16}	hhinclt25	Dummy variable which equals to 1 if the household income in the last 12 months has been under 25,000$, 0 otherwise
X_{17}	hhinc25_49	Dummy variable which equals to 1 if the household income in the last 12 months has been between 25,000$ and 49,999$, 0 otherwise
X_{18}	hhinc50_74	Dummy variable which equals to 1 if the household income in the last 12 months has been between 50,000$ and 74,999$, 0 otherwise
X_{19}	hhinc75_99	Dummy variable which equals to 1 if the household income in the last 12 months has been between 75,000$ and 99,999$, 0 otherwise
X_{20}	hhincgt100	Dummy variable which equals to 1 if the household income in the last 12 months has been above 100,000$, 0 otherwise
X_{21}	se_app	Dummy variable which equals to 1 if the subject is self-employed, 0 otherwise
X_{22}	healthprob	Dummy variable which equals to 1 if the subject is currently receiving health benefits, 0 otherwise
X_{23}	worked_for_relatives_friends_se	Dummy variable which equals to 1 if the subject has ever worked for a businesss owned by relatives or friends, 0 otherwise
X_{24}	badcredit	Dummy variable which equals to 1 if the subject has a bad credit history, 0 otherwise
X_{25}	currently_receiving_ui_benefits	Dummy variable which equals to 1 if the subject is currently receiving UI benefits, 0 otherwise
X_{26}	emphealthins	Dummy variable which equals to 1 if the subject has health insurance, 0 otherwise
X_{27}	autonomy	Autonomy index
X_{28}	risk_tolerance_combined	Risk aversity index
X_{29}	notemp_app	Dummy variable which equals to 1 if the subject is unemployed, 0 otherwise

Table A.2: Set of dependent variables used for the GATE project

Y_1	se_w1	Dummy variable which equals to 1 if the subject is a business owner at W1 survey date, 0 otherwise
Y_2	se_w2	Dummy variable which equals to 1 if the subject is a business owner at W2 survey date, 0 otherwise
Y_3	se_w3	Dummy variable which equals to 1 if the subject is a business owner at W3 survey date, 0 otherwise
Y_4	salesun_w1	Monthly business sales at W1 survey date (000s)
Y_5	salesun_w2	Monthly business sales at W2 survey date (000s)
Y_6	salesun_w3	Monthly business sales at W3 survey date (000s)
Y_7	anyempsun_w1	Dummy variable which equals to 1 if the subject has any employees at W1 survey date, 0 otherwise
Y_8	anyempsun_w2	Dummy variable which equals to 1 if the subject has any employees at W2 survey date, 0 otherwise
Y_9	anyempsun_w3	Dummy variable which equals to 1 if the subject has any employees at W3 survey date, 0 otherwise
Y_{10}	lhhincome_w1	Log household income at W1 survey date
Y_{11}	lhhincome_w2	Log household income at W2 survey date
Y_{12}	lhhincome_w3	Log household income at W3 survey date

Table A.3: Treatment variable and covariates used for the microfinance project

W	sec_treat	The ”grace period” treatment, which equals 1 if the client received a two-month grace period, 0 otherwise
X_1	sec_loanamount	Amount loaned
X_2	Age_C	Age
X_3	Married_C	Dummy which equals 1 if the client is married, 0 otherwise
X_4	Literate_C	Dummy which equals 1 if the client is literate, 0 otherwise
X_5	Muslim_C	Dummy which equals 1 if the client is muslim, 0 otherwise
X_6	HH_Size_C	Household size
X_7	SEI	A principal component analysis index of whether the household had owned a radio, refrigerator, washing machine, heater or television for longer than 1 year, 0 otherwise
X_8	Years_Education_C	Years of education
X_9	shock_any_C	Household shock: dummy which equals 1 if a birth, death, heavy rain or flood orrcured in the last 30 days, 0 otherwise
X_{10}	Has_Business_C	Dummy which equals 1 if the clinet owns a business at baseline, 0 otherwise
X_{11}	Financial_Control_C	Dummy which equals 1 if the client would be able to lend money to a close relative who is sick and in need of money, 0 otherwise
X_{12}	homeowner_C	Dummy which equals 1 if the client owns the home she lives in, 0 otherwise
X_{13}	Risk_Loving	Dummy equal to 1 if the client is risk-loving, 0 otherwise
X_{14}	Has_Savings_Acc_Apr24	Dummy equal to 1 if the client owns a savings account, 0 otherwise
X_{15}	Stratification_Dummies	Dummy corresponding to the stratification batch of the client
X_{16}	Impatient	Dummy equal to 1 if the client is impatient, 0 otherwise
X_{17}	Has_Business_C	Dummy equal to 1 if the client does not own a business at baseline, 0 otherwise
X_{18}	Earns_Wage	Dummy equal to 1 if any household member earned wages at the time of the survey, 0 otherwise

Table A.4: First set of dependent variables used for the microfinance project

Y_1	Business_Expenditures	Total Business Expenditures from Loan Use
Y_2	Inventory_Raw_Mat	Expenditures on Inventory and Raw Materials from Loan Use
Y_3	Equipment	Expenditures on Equipment from Loan Use
Y_4	Other_Bus_Cost	Expenditures on Other Business Costs from Loan Use
Y_5	Non_Business_Exp	Total Non Business Expenditures from Loan Use
Y_6	Repairs_Repair_Only	Expenditures on Household Repairs from Loan Use
Y_7	Utilities_Taxes_Rent	Expenditures on Utilities Taxes and Rent from Loan Use
Y_8	Human_Capital	Expenditures on Schooling Education and Health from Loan Use
Y_9	Re_Lent	Expenditures on Relending from Loan Use
Y_{10}	Savings	Expenditures on Savings from Loan Use
Y_{11}	Food_And_Durable	Expenditures on Food and Durables from Loan Use

Table A.5: Second set of dependent variables used for the microfinance project

Y_{12}	Profit	Average monthly profits
Y_{13}	$\text{ln}_Q 50$	Log of monthly household income earned in the past 30 days
Y_{14}	Capital	Value of raw materials, inventory and equipment

The attached zip-file contains the following files:

GATE project

Folder GATE-Data

application.sas7bdat

wave1.sas7bdat

wave2.sas7bdat

wave3.sas7bdat

crdata_v17.sas: Program that reads the four GATE datasets and creates a working dataset used in all of the analyses of the GATE project

Folder GATE-Replication codes

rtreat regs bus outcomes iv_v5.: Replicates column 1 of Table 1 from the thesis (Table 4 from the original paper)

rtreat subgroups subsamples iv_v3.: Replicates Table 3 from the thesis (Table 8B from the original paper)

Folder GATE-Personal codes

GATE_ATE: Replicates columns 2, 3, 4, 5 of Table 1 from the thesis

GATE_HET: Replicates Table 2 from the thesis

GATE_groups: Replicates Table 4 and Table 5 from the thesis

Microfinance experiment

Folder Micro-Data

Grace-Period-Data.dta

Folder Micro-Replication codes

Table-1–Loan-Use.do : Replicates Table 6 from the thesis (Table 1 from the original paper)

Table-2–Profits-and-Income.do : Replicates Table 8 from the thesis (Table 2 from the original paper)

Table-5–Heterogeneity.do : Replicates Table 11 from the thesis (Table 5 from the original paper)

Folder Micro-Personal codes

Micro_ATE: Replicates Table 7 and Table 9 from the thesis

Micro_HET: Replicates Table 10 from the thesis

Micro_groups: Replicates Table 12 from the thesis

Simulations

Folder Simulation-1

heterogeneity1: Replicates Table 14 from the thesis

heterogeneity2: Replicates Table 15 and Figure 1 from the thesis

Folder Simulation-2

confounding1: Replicates Table 15 and Figure 2 from the thesis

confounding2: Replicates Figure 3 form the thesis

Folder Simulation-3

noise1: Replicates Table 16 and Figure 4 from the thesis