

CPB Netherlands Bureau for Economic Policy Analysis

Causal forests with fixed effects for treatment effect heterogeneity in difference-in-differences

We modify the causal forest algorithm such that estimation of heterogeneous treatment effects becomes computationally feasible in the presence of many fixed effects. Our modification identifies treatment effects by partialling out fixed effects using group averages.

Simulation results suggest that our algorithm provides consistent estimates of the Conditional Average Treatment effect for the Treated in a (staggered) difference-indifferences research design. Our modification suggests that hourly wages fell by 3.7 percent in the first year of payrolling for a specific subgroup of workers only. We discover a wider range of heterogeneity than when we manually compare subgroups.

CPB Discussion Paper

Mark Kattenberg, Bas Scheer, Jurre Thiel

November 2023

Doi: https://doi.org/10.34932/216c-yz58

Causal forests with fixed effects for treatment effect heterogeneity in difference-in-differences^{*}

Mark A.C. Kattenberg^{a,b}, Bas J. Scheer^a, and Jurre H. Thiel^a

^aNetherlands Institute for Economic Policy Analysis (CPB) ^bCorresponding author: m.a.c.kattenberg@cpb.nl

November 27, 2023

Abstract

Recently developed heterogeneity-robust two-way fixed effects (TWFE) estimators do not quantify the full heterogeneity in treatment effects in a differencein-differences research design. We therefore present a computationally feasible algorithm to estimate heterogeneous treatment effects in the presence of many fixed effects using causal forests. Our modification identifies treatment effects by partialling out fixed effects using group averages. Simulation results suggest that our algorithm provides consistent estimates of the Conditional Average Treatment effect for the Treated in a (staggered) difference-in-differences research design. Finally, we use our method to document heterogeneity in the treatment effect of alternative work arrangements (payrolling) on hourly wages. We find evidence that wages fell by 3.7 percent in the first year of payrolling for a specific subgroup of workers only. Both conclusions did not appear in a conventional heterogeneity analysis using manual subgroups. The R-code of our algorithm is publicly available online.

Keywords: causal forest, difference-in-differences, fixed effects, treatment effect heterogeneity, alternative work arrangements, administrative data

JEL codes: C18; C23; C88

^{*}For this research we used simulations and microdata of Statistics Netherlands. The code to replicate the simulation results is available upon request. The data from Statistics Netherlands we used cannot be shared publicly, because it is highly confidential personal data. We will support getting access to the data-project of Statistics Netherlands. We are grateful for feedback from seminar participants at CPB and the NED. In particular we received helpful feedback from Falco Bargagli-Stoffi. All remaining errors are ours.

Declarations of interest: none.

1 Introduction

. Our modification indicates that the hourly wages of a particular subgroup of workers decreased by 3.7 percent in the first year of payrolling. We found a greater degree of diversity than when we manually examined the subgroup.

Difference-in-differences is a widely used research design (Currie et al., 2020), with fifteen percent of all papers published in the American Economic Review in 2022 utilizing it. Additionally, between 2015 and 2019, 26 out of the 100 most popular articles in that journal employed regressions with period and group fixed effects (De Chaisemartin and d'Haultfoeuille, 2023). However, the economics community has recently come to understand that estimation results with time and group fixed effects may be misleading when treatment effects differ between groups or over time. As a result, the past few years have seen the development of heterogeneity-robust estimators of the average treatment effect (see the review by De Chaisemartin and d'Haultfoeuille (2023) and references therein).

However, these estimators generally do not quantify the full heterogeneity in treatment effects, since they depend on the researcher's manual formation of subgroups. This approach carries two potential risks. First, the undetected heterogeneity problem might occur when researchers only consider their 'usual suspects' or groups that are easy to target by policymakers. Other groups, that might have treatment effects that differ substantially from the average, might simply not be considered. Second, there is the false discovery problem. Researchers might mistakenly present false discoveries as genuine effects, when they test many different interactions (or sample splits) without taking multiple hypothesis testing into account.¹

We mitigate these problems by developing a feasible algorithm to estimate a causal forest with high-dimensional fixed effects, which we call the causal forest with fixed effects. The flexibility of the causal forest estimator (Athey et al., 2019; Wager and Athey, 2018; Athey and Imbens, 2016) reduces the missed discovery problem, because the algorithm uses a potentially high-dimensional set of characteristics to form sub-groups for which the treatment effects differ from the average. At the same time, the false discovery problem is mitigated because the estimator is 'honest' (Athey et al., 2019; Athey and Imbens, 2016).² Despite the causal forest algorithm being only a few years old, it has become a popular tool to study treatment effect heterogeneity rapidly.³

However, causal forests cannot be applied directly to a standard difference-indifferences research design, because the treatment effect is not identified conditional on the inclusion of indicators for treatment group status and treatment period in its

¹See Cook et al. (2004) and Assmann et al. (2000) for an explanation. The comic by Munroe (2022) nicely illustrates the risk imposed by multiple hypothesis testing. Benjamini and Hochberg (1995) and Holm (1979) provide possible corrections.

²Honesty refers to the fact that the algorithm uses split samples to estimate treatment effects out of sample. This approach to prevent overfitting is similar in spirit to (two-fold) cross-validation. In section 3 we explain honesty in further detail.

³Causal forests are used to study heterogeneous responses in several fields, such as medical studies (Verstraete et al., 2023; Jawadekar et al., 2023; Bodory et al., 2022; Raghavan et al., 2022; Hermansson and Svensson, 2021), public economics (Shah et al., 2023; Hoffman and Mast, 2019), labor economics (Zheng and Yin, 2023; Davis and Heller, 2020), banking (Brock and De Haas, 2023; Gulen et al., 2020), business economics (Luo et al., 2019) and environmental economics (Murakami et al., 2022; Knittel and Stolper, 2021; Miller, 2020) including the impact of natural disasters (Shiba et al., 2021).

current implementation by Tibshirani et al. (2022).⁴

Our modification controls for various fixed effects using averages of the outcome and treatment indicator over the drivers of these fixed effects, such as individuals, time and/or event-time (Somaini and Wolak, 2016).⁵ This allows the use of causal forests in a difference-in-differences research design without affecting the statistical properties of causal forests described in Athey et al. (2019); Wager and Athey (2018). Our approach yields consistent estimates of the CATT and the ATT that are identified when the common-trend and no-anticipation assumptions hold at the subgroup level. In this paper we first consider the performance of the causal forest with fixed effects using simulated data for a single-event and for a staggered difference-in-differences research design. Second, we use our method to estimate the heterogeneous impact of alternative work arrangements (payrolling) on worker wages using Dutch employeremployee matched registration data, as in Goos et al. (2022).

Our paper provides two important conclusions. First, our simulation results show that a causal forest with fixed effects provides consistent estimates of heterogeneity in treatment effects. This is concluded when single event and staggered difference-indifferences research designs are simulated. Also, we show that a manually recentered causal forest provides inconsistent estimates, which highlights that our modification of the causal forest algorithm is necessary.

Second, we study the effect on payrolling on short term hourly wages. In particular, we compare a heterogeneity analysis based on manually formed subgroups against such an analysis based on a causal forest with fixed effects. Both methods yield similar estimates of the average treatment effect, \in -0.2 per hour (1.8 percent) for a causal forest with fixed effects and \notin -0.17 per hour (1.6 percent) for an OLS-regression. Yet, the pattern of heterogeneity detected is very different. When manual subgroups are formed, many worker characteristics lead to a significant treatment effect. For instance, the treatment effect for women is significant at \notin -0.27 per hour (2.5 percent), which suggests that hourly wages decrease after payrolling for about half of the population. Yet, when we group workers into deciles based on their Conditional Average Treatment Effect of the Treated (CATT) provided by our algorithm, we find that the CATT is only significant for the first two deciles and equal to about \notin -0.4 per hour (3.7 percent). In line with the manual subgroup analysis, we find that workers in this group are more often female, young, enrolled in education or workers with a first generation migration background.

Our paper is related to three fast developing fields within the econometrics literature. First, this paper is related to the recent econometrics literature on identification and bias of (staggered) difference-in-differences when individual and time fixed effects are present (De Chaisemartin and d'Haultfoeuille, 2023; Roth et al., 2023; Dube et al., 2023; Baker et al., 2022; Goodman-Bacon, 2021; De Chaisemartin and d'Haultfoeuille, 2020). These

⁴The use of difference-in-differences as a research design is well explained in Angrist and Pischke (2010) and Angrist and Pischke (2009). The assumptions underlying and the behavior of differencein-differences estimators are studied in many papers, including their performance when there are few treatment groups (Donald and Lang, 2007), when there is autocorrelation (Bertrand et al., 2004), semi-parametric or Double-Machine-Learning estimation of difference-in-differences research (Abadie, 2005; Chang, 2020; Knaus, 2022), non-linear difference-in-differences (Athey and Imbens, 2006) or the relation between synthetic control and DD (Doudchenko and Imbens, 2016; Arkhangelsky et al., 2021).

⁵The approach is extensively discussed in section 4.

papers detail why inclusion of time and group fixed effects might bias the estimation of the average treatment effect and discuss solutions to this problem. In particular, a naive twoway fixed effects difference-in-differences specification with binary treatment would make the forbidden comparison of (newly) treated observations against already treated observations in a staggered research design. In contrast, we extent the causal forest algorithm with twoway and threeway fixed effects. We demonstrate that this allows for estimation of heterogeneity in treatment effects in a difference-in-differences setting when treatment is binary and the estimation sample is restricted to exclude so-called 'forbidden comparisons' of treated against already-treated individuals. Importantly, our method provides an estimate of the average treatment effect that is robust to variation in treatment effects between groups and over time. Finally, we emphasize the advantages of using causal forests to predict treatment effects compared to the standard two-way fixed effects specification. The recursive partitioning of causal forests guarantees that treatment effects are estimated for treated and control units that are very similar in confounding characteristics.

Second, our paper is related to papers that use machine learning to estimate heterogeneity in treatment effects. Several machine learning methods have been used to estimate treatment effect heterogeneity, including Bayesian Additive Regression Trees (BART, see Chipman et al. (2010); Green and Kern (2012); Hill (2011) and Hill and Su (2013)), regression trees (Su et al., 2009; Zeileis et al., 2008), regression forests (Foster et al., 2011), Support Vector Machines (Imai and Ratkovic, 2013) and LASSO (Tian et al., 2014). Other papers estimate treatment effect heterogeneity using tree-based algorithms when treatment assignment is not randomly assigned (Bargagli-Stoffi et al., 2022; Bargagli Stoffi and Gnecco, 2020; Wang et al., 2017; Johnson et al., 2022; Hartford et al., 2016). Hatamyar et al. (2023) use the R-learner, a machine learning technique, to estimate treatment effect heterogeneity in a staggered difference-in-differences design. We focus on causal forests, because this estimator is designed to maximize treatment effect heterogeneity and econometric theory proves the method is consistent, also when irrelevant variables are used to look for treatment effect heterogeneity (Wager and Athey, 2018).

Third, our paper is related to studies that extent the application of causal forests to panel data and difference-in-differences. Panel data enables estimation conditional on individual fixed effects (Jens et al., 2021) or the estimation of effects that vary over time (Bodory et al., 2022; Miller, 2020). Knittel and Stolper (2021) and Gavrilova et al. (2023) use the regular causal forests algorithm in a difference-in-differences research design. Because the data used by Knittel and Stolper (2021) comes from three RCT treatment interventions, they do not have to account for individual fixed effects. Therefore they do not encounter the identification issue that the treatment indicator is not identified conditional on individual and time fixed effects in a causal forest. Gavrilova et al. (2023) modify the input data to overcome this issue. In particular, they subtract the value in the last period before treatment from the outcome variable and then estimate separate causal forests for each time period using the treatment group indicator. This approach allows for the estimation of heterogeneous treatment effects in the presence of individual and time fixed effects. To the best of our knowledge, we are the first to test the working of a dynamic causal forest using simulations. Results suggest our approach is more efficient, provided that the common trend assumption holds during the entire pre-treatment period.

The contributions of this paper are twofold. First, we modify the original causal forest algorithm by Tibshirani et al. (2022). Their algorithm uses the same set of observed time-invariant characteristics to control for confounding variables and to look for heterogeneity, whereas our modification can control for unobserved time-invariant characteristics (such as individual effects). Second, we show how our modification can be used in a (staggered) difference-in-differences research setting. We use simulations to compare the performance of our modification and two other types of causal forests that can be applied to difference-in-differences. We find that consistent results are only delivered by our method and the dynamic causal forest (Gavrilova et al., 2023). Our method is more efficient when time and individual fixed effects are present and the common trend assumption holds. The code of the causal forest with fixed effects is publicly available.⁶

This paper proceeds as follows. We review identifying assumptions underlying (staggered) difference-in-differences and different estimation strategies in section 2. We discuss the causal forest algorithm in section 3. In section 4 we explain how the causal forest with fixed effects, the manually recentered causal forest and the dynamic causal forest apply causal forests in a difference-in-differences research design. In section 5 we present how these methods perform when single-event difference-in-differences data is simulated, in section 6 we do this using simulated staggered difference-in-differences data. We illustrate the use of our method in section 7, by studying the effect of payrolling on worker wages using Dutch employer-employee matched registration data. Finally, section 8 concludes.

2 Difference-in-differences

This section reviews the estimation of difference-in-differences models and the key assumptions underlying them. This is done for the canonical two period, two groups difference-in-difference design (section 2.1) and the staggered design (section 2.2). In section 2.3 we present assumptions to identify the conditional average treatment effect in a staggered difference-in-differences design.

2.1 Difference-in-differences

The most basic form of a difference-in-differences research design involves a single treatment event, two treatment periods (before and after treatment) and two groups. Individuals are observed in both time periods and belong to either the control or to the treatment group. The idea behind difference-in-differences is that the treatment effect can be estimated empirically by subtracting the trend in outcomes for individuals in the control group (i.e, the difference over time) from the trend in outcomes for individuals in the treatment group. This idea provides the method its name.

The working of difference-in-differences can be explained formally using the potential outcomes framework (Rubin, 2005; Imbens and Rubin, 2015). Denote the value of the outcome for individuals *i* in period *t* as $Y_{it}(1)$ when the individual is treated and as $Y_{it}(0)$ when the individual is not. Then the average treatment effect of the treated (ATT) is defined as the difference $Y_{it}(1) - Y_{it}(0)$ averaged across the units receiving treatment.

⁶See https://github.com/MACKattenberg/cffe.

The empirical challenge is that individuals are either in the treatment group or in the control group. Therefore the researcher only observes $Y_{it}(1)$ or $Y_{it}(0)$, but not both. The difference-in-differences estimator solves this missing data problem by implicitly imputing the missing outcome values $Y_{it}(0)$ for treated units using information on units in the control group. The validity of the difference-in-differences estimator relies on the so-called 'parallel-trends assumption', the assumption that the observed trend in outcomes for the control group equals the trend in outcomes for the treatment group had they not received treatment (see equation 1). Also, there should be no anticipatory effect of treatment before treatment has started (equation 2). Under these assumptions the ATT equals the difference in outcomes before and after treatment Roth et al. (2023).

parallel trends assumption

$$E[Y_{i,1}(0) - Y_{i,0}(0)|W_i = 1] = E[Y_{i,1}(0) - Y_{i,0}(0)|W_i = 0]$$
(1)

no anticipation assumption

$$Y_{i,0}(1) = Y_{i,0}(0) \tag{2}$$

The formal derivation of the difference-in-differences estimator is shown in equation 3, where t = 1 denotes the time period after treatment and where $W_i = 1$ denotes that the unit *i* is in the treatment group. Underbraces are used to highlight which parts of the equation are unobserved. The first equality defines the ATT, the estimand of interest. The second equality follows from subtracting $E[Y_{i,0}(1)|W_i = 1]$ from both sides, which allows to write the ATT as the difference in trends for units in the treatment group when they receive treatment and would they not have received treatment. In the third equality the common trend assumption is used to replace the unobserved trend in outcomes for units in the treatment group would they not have received treatment by the observed trend in outcomes for units in the control group. Naturally, when the trend in outcomes for the control and treatment group differ (i.e. the common trend assumption does not hold), the difference-in-differences estimator is biased.

$$ATT = \underbrace{E[Y_{i,1}(1) - Y_{i,1}(0)|W_i = 1]}_{unobserved}$$

= $E[Y_{i,1}(1) - Y_{i,0}(1)|W_i = 1] - \underbrace{E[Y_{i,1}(0) - Y_{i,0}(0)|W_i = 1]}_{unobserved}$
= $E[Y_{i,1}(1) - Y_{i,0}(1)|W_i = 1] - E[Y_{i,1}(0) - Y_{i,0}(0)|W_i = 0]$ (3)

Equation 3 is typically derived by estimating equation 4 using OLS, which provides consistent estimates and asymptotically valid confidence intervals when observations are sampled independently and the parallel-trends and no-anticipation assumptions hold (Roth et al., 2023). In this equation i = 1, ..., N denotes units and t denotes time periods $t = 1, ..., \mathcal{T}$. y denotes the outcome variable of interest, c_i is an individual fixed effect and δ_t are time fixed effects. W_{it} is the treatment indicator, which is the product of the indicator whether observations are observed after treatment and an indicator whether individual i belongs to the treatment group ($W_i = 1$). Equation 4 has a general form. In particular the individual fixed effects absorb all time-invariant characteristics, including an indicator whether individual i belongs to the treatment group indicator. The time fixed effects absorb time-variant characteristics that are equal for all observations, including an indicator whether the treatment period has started.

$$y_{it} = c_i + \delta_t + \tau W_{it} + \varepsilon_{it} \tag{4}$$

2.2 Staggered difference-in-differences

Equation 4 has also been used to estimate 'staggered difference-in-differences', a more general specification where multiple treatment events define the treatment indicator W_{it} . Staggered difference-in-differences occur more frequently than single-event difference-in-differences and have been believed to be more stable, because the use of multiple treatment events provides more robust controls against confounding time trends.

Recent papers have highlighted the inconsistency of equation 4 for estimation of the ATT in a staggered difference-in-differences research designs. First, it has been shown that the estimate of τ in equation 4 is a weighted average of all possible twogroup/two-period DiD estimators in the data when the treatment effect is equal in all treatment events (Goodman-Bacon, 2021). When the treatment effect is dynamic, i.e. the treatment effect varies with treatment event, or –as we assume– treatment effects vary by group, the estimate of τ s will differ from the sample ATT (Baker et al., 2022; De Chaisemartin and d'Haultfoeuille, 2020).

The 'staggered' parallel-trends assumption requires additional notation, which we borrow from Roth et al. (2023). There are $t = 1, \ldots, T$ time periods and units can receive a binary treatment in any period t > 1. Once a unit is treated, they remain treated for the remainder of the panel. We denote by $D_{i,t}$ an indicator for whether unit *i* receives treatment in period *t* and we let G_i be the earliest period *t* at which unit *i* has received treatment. If a unit is never treated, then $G_i = \infty$. We use this notation to define potential outcomes as follows. Let $\mathbf{0}_s$ and $\mathbf{1}_s$ be a s-dimensional vector of zeros and ones, respectively. We denote unit *i*'s potential outcome in period *t* if they were first treated at time $G_i = g$ by $Y_{i,t}(\mathbf{0}_{g-1}, \mathbf{1}_{T-g-1})$, and we denote by $Y_{i,t}(\mathbf{0}_T)$ their 'never-treated' potential outcome. As units remain treated after receiving treatment for the first time, we can simplify notation using *g*. The notation for the potential outcomes for units that are first treated in period *g* becomes $Y_{i,t}(g)$, whereas the notation for never treated units becomes $Y_{i,t}(\infty)$. Note that this notation allows for dynamic treatment effects, treatment effects that vary by treatment event *g* and treatment effects that vary with individual characteristics.

The 'staggered' parallel-trends assumption states that if treatment had not taken place, the average outcomes for all treated groups would have evolved parallel as in equation 5. This assumption requires that trends are parallel between any two periods t and t'. Sun and Abraham (2021) consider a relaxation that states that requires that trends are parallel only between treatment period t and the last period before treatment $g_{\min} = g - 1$ (equation 6). It relaxes the parallel trends assumption, because trends need no longer be parallel during the entire period before treatment. The parallel trend assumption can also be relaxed by imposing parallel trends for a subset of groups. In particular, Callaway and Sant'Anna (2021) and Sun and Abraham (2021) consider a parallel trends assumption for units that are treated eventually. This assumption requires parallel trends between units that are eventually treated, but not between treated units and never-treated units.

The no-anticipation assumption and the random sampling assumption generalize straightforwardly to the staggered difference-in-difference setting. In particular, the no-anticipation assumption is given by equation 7.

parallel trends assumption for staggered design

$$\mathbb{E}\left[Y_{i,t}(\infty) - Y_{i,t'}(\infty)|G_i = g\right] = \mathbb{E}\left[Y_{i,t}(\infty) - Y_{i,t'}(\infty)|G_i = g'\right], \quad (5)$$
$$\forall t \neq t', g \neq g'$$

parallel trends assumption for staggered design – post treatment only

$$\mathbb{E}\left[Y_{i,t}(\infty) - Y_{i,t'}(\infty)|G_i = g\right] = \mathbb{E}\left[Y_{i,t}(\infty) - Y_{i,t'}(\infty)|G_i = g'\right], \quad (6)$$
$$\forall t \neq t', g \neq g', t, t' \ge g_{\min}$$

no anticipation assumption for staggered design

$$Y_{i,t}(g) = Y_{i,t}(\infty), \quad \forall i \text{ and } t < g.$$

$$\tag{7}$$

In order to account for dynamic treatment effects, a staggered version of equation 4 has been estimated that interacts each relative-time-after-treatment indicator $1[R_{i,t}]$ with the treatment indicator W_{it} (see equation 8). Here $R_{i,t}$ denote relative time after treatment $R_{i,t} = T - g_i + 1$ such that $R_{i,t} = 1$ denotes the first period after treatment. As a result, τ is estimated separately for each period after treatment. As demonstrated in Sun and Abraham (2021), this estimator produces valid results under the parallel trends assumption introduced above and when treated units do not anticipate treatment. However, they also point out that these estimates may be difficult to interpret when there is treatment heterogeneity across adopted cohorts, due to 'cross-contamination' and 'negative weighting' (Sun and Abraham, 2021; Roth et al., 2023).

$$y_{it} = c_i + \delta_t + \sum_{r \neq 0}^{\mathcal{T}^*} \mathbb{1}[R_{i,t} = r] \tau W_{it} + \varepsilon_{it}$$
(8)

Two type of estimators have been proposed to overcome these problems. The first type of estimators compare outcomes in year t for units i against their outcomes in the baseyear, typically the year before treatment commences $(t = g_i - 1)$. The control group can be never-treated or not-yet-treated individuals. This approach is followed by Callaway and Sant'Anna (2021). Because treatment effects might vary with time-since-treatment and by treatment cohort, Callaway and Sant'Anna (2021) estimate treatment effects separately for each combination of treatment event and relative-time-after-treatment. Note that because treatment effects are computed relative to the base year, results are consistent under the relaxed parallel trends assumption in equation 6. The second type of estimator imputes the counterfactual outcome of treated individuals using outcomes and characteristics of not-yet-treated individuals, an approach outlined in Borusyak et al. (2021) and implemented by Borusyak (2023). They fit a TWFE regression using all units and time periods that are not-yet-treated. This model is used to predict the potential outcome for treated individuals when they would not have received treatment. This prediction straightforwardly can be used to construct the average treatment effect.

Similar to Callaway and Sant'Anna (2021), this approaches yield valid estimates, provided the parallel trend assumption holds for all groups and time periods and there is no anticipation of treatment. As OLS is the best linear unbiased estimator under the Gauss-Markov conditions, this approach is more efficient than the linear estimator of Callaway and Sant'Anna (2021). Yet, the estimator proposed by Borusyak et al. (2021) also requires a stronger identifying assumption (De Chaisemartin and d'Haultfoeuille, 2023; Roth et al., 2023).

To see why, note that the main difference in estimation methods relies in the use of all periods before treatment as a comparison group (Borusyak et al., 2021), against the use of a specific base-period (Callaway and Sant'Anna, 2021). Consequently, the former requires that the common trend assumption is valid for all the periods considered (both before and after treatment), whereas the latter require it to hold only in between the base period b and period t. This makes the estimator by Borusyak (2023) more prone to bias when time-trends diverge over time, which occurs when group-specific linear trends are present. On the other hand, the estimator by Borusyak (2023) is more efficient when treated units anticipate the treatment before it has started (see Roth (2022)). Also, the estimator by Callaway and Sant'Anna (2021) is more robust to serial correlation in the error term, because it compares outcomes in two periods in time only.

2.3 Staggered difference-in-differences with individual specific treatment effects

Finally, we consider how the identifying assumptions in difference-in-differences change when treatment effects are a function of individual, time-invariant characteristics, as in equation 9. These characteristics can also affect the outcome variable, yet – obviously – they are not identified due to the inclusion of the individual fixed effects. Again, we assume that treatment is binary, that individuals that receive treatment remain treated in all following periods and that individuals differ in the moment they are first treated.

$$y_{it} = c_i + \delta_t + \tau(X_i)W_{it} + \varepsilon_{it} \tag{9}$$

Let S denote a subgroup of individuals with characteristics X that have the same treatment effect $\tau(X)$. We propose local variants of the parallel trend assumption and the no-anticipation assumption, that only hold for the subgroup of units $i \in S$. Consequently, these assumptions are given by equations 10 to 12. These assumptions are weaker than those described in Roth et al. (2023), because they identify the conditional average treatment effect of the treated only for individuals in S. local parallel trends assumption for staggered design

$$\mathbb{E}\left[Y_{i,t}(\infty) - Y_{i,t'}(\infty)|G_i = g\right] = \mathbb{E}\left[Y_{i,t}(\infty) - Y_{i,t'}(\infty)|G_i = g'\right], \quad (10)$$
$$\forall t \neq t', g \neq g', i \in S$$

local parallel trends assumption for staggered design – post treatment only

$$\mathbb{E}\left[Y_{i,t}(\infty) - Y_{i,t'}(\infty) | G_i = g\right] = \mathbb{E}\left[Y_{i,t}(\infty) - Y_{i,t'}(\infty) | G_i = g'\right], \qquad (11)$$
$$\forall t \neq t', g \neq g', t, t' \ge g_{\min}, i \in S$$

local no anticipation assumption for staggered design

$$Y_{i,t}(g) = Y_{i,t}(\infty), \quad \forall i \in S \text{ and } t < g$$
(12)

In principle, local conditional treatment effects can be estimated using the estimators developed by Sun and Abraham (2021) and Borusyak et al. (2021) by making subgroups S manually. In the next section we detail how causal forests can be used to make subgroups of the data and estimate their conditional average treatment effects. In section 4 we explain how to apply causal forests to difference-in-differences settings.

3 Causal forests

Causal forests are based on the random forest algorithm (Breiman, 2001).⁷ Random forests combine the predictions from regression trees (Breiman et al., 1984), where each tree is grown on a separate bootstrapped set of the data (a procedure referred to as bootstrap aggregation or 'bagging'). In a regression tree, a continuous outcome is predicted by consecutively splitting the data along one variable and assigning the mean value for the outcome within the group as the prediction.⁸ A final characteristic of random forests is that at each node in the tree, splits can be made on a random subset of variables only. Because of this random selection procedure, variables that are very predictive of the outcome are not always chosen. This procedure helps to prevent overfitting the training data as it reduces the correlation between the trees in the forest which in turn reduces the variance of the final prediction. The idea behind it is that the out-of-sample prediction of each tree is equal to the true value plus a random error. Since the final prediction is an average of a large number of weakly correlated trees, the final prediction equals the average of the true values plus the average of the errors. When errors are weakly correlated, the latter average tends to zero and predictive performance improves.

 $^{^7\}mathrm{We}$ provide a concise summary of random forests only. See Hastie et al. (2009) for an in-depth discussion.

⁸For example, at a point in the tree (a 'node'), the data might be split into two groups according to whether $x_{ik} \leq \bar{x_k}$ or $x_{ik} > \bar{x_k}$. The splits are chosen in a greedy way: at every node, the variable along which the split is made, x_{ik} and the cut-off, $\bar{x_k}$, are chosen to maximize the summed objective function across both groups. A often used objective function is the reduction in RMSE.

Since random forests combine the predictions of many regression trees, they are usually seen as an ensemble method (Hastie et al., 2009). However, they can also be interpreted as a nearest-neighbor estimator (Athey and Wager, 2019). The predictions of a random forest are essentially weighted averages of outcomes, where the weight between observation i and j depends on the number of times observation i ends up in the same terminal node as observation j. Observations far away in the covariate space receive small weights, whereas points nearby receive large weights.

Causal forests (Athey et al. (2019) and Wager and Athey (2018)) combine many causal trees (Athey and Imbens, 2016) to estimate heterogeneous causal effects. Causal trees differ from regression trees as they make splits that maximize the heterogeneity in treatment effects. Also, causal forests do not average the predictions of the conditional average treatment effect in the leaves of the causal forests. Instead they define observations that end up in the same leaf as observation i as being "similar to i" and define similarity weights (using all trees in the forest) to estimate the treatment effect. Thus causal forests can be interpreted as an adaptive nearest neighbor estimator that uses recursive binary splitting to determine similar observations.

The flexibility of the algorithm mitigates the missed discovery problem, whereas the false discovery problem is mitigated because the causal forest algorithm is 'honest'. Honesty refers to the fact that predictions are made realistically – out of sample – so overfitting is mitigated. In particular, the algorithm randomly splits the observations into two groups. One group of observations is used to generate the splits and similarity weights, the other is used to evaluate the treatment effects that result from these splits.⁹ Finally, the treatment effects estimated by a causal forest are piecewise consistent and are asymptotically normally distributed (Wager and Athey, 2018).

To see how a causal forest works, consider equation 13. Here y_i is the outcome variable, α is a constant, X_i is a matrix of control variables and ε_i is an error term assumed to be i.i.d. W_i is the treatment indicator and the parameter of interest is $\tau(X_i)$. Note that $\tau(X_i)$ is a function of X_i , hence it measures the Conditional Average Treatment Effect (CATE) conditional on X_i . A causal forest relies on the unconfoundedness assumption for identification of $\tau(X)$. This assumption is met when selection of individuals into treatment is as good as random conditional on X_i . More formally, identification of the causal forest algorithm requires that the conditional probabilities to receive treatment $P(W_i = 1|X_i, W_i = 1)$ and $P(W_i = 1|X_i, W_i = 0)$ overlap (Imbens and Rubin, 2015), and that they are bounded away from zero and one (Tibshirani et al., 2022).

$$y_i = \alpha + \beta X_i + \tau(X_i) W_i + \varepsilon_i \tag{13}$$

As mentioned, causal forests use recursive binary splitting to maximize the heterogeneity in treatment effects. For illustrative purposes, suppose $\tau(X_i)$ in equation 13 is estimated with a causal forest, that only contains the shallow causal tree in figure 1. The causal tree is grown to compute similarity weights as follows. Initially, all observations in the train data are in the root node. The root node is split into two

⁹As treatment effects are estimated out of sample, this approach resembles the use of cross-validation to prevent overfitting in many machine learning applications (Hastie et al., 2009). Honesty is a vital element in the literature that uses machine learning to estimate causal effects, see Athey et al. (2019), Chernozhukov et al. (2018), Chernozhukov et al. (2018), Chernozhukov et al. (2017) and Athey and Imbens (2016).

groups according to $x_3 \leq 5$. Those observations not meeting that condition are in the left group. These observations are further split down into groups according to $x_1 \leq 2$. Numbering leafs from left to right, observations that do not meet the second condition are in the first leaf. Those that do meet this condition are in the second leaf. Likewise, observations with $x_3 > 5$ are split according to whether $x_2 \leq 1$ and end up in the third or fourth leaf.



Figure 1: Example of a shallow causal tree

The CATEs estimated by this causal tree can be reproduced using OLS when equation 14 is estimated on the same data. In this equation I() is an indicator function equal to one when the condition within brackets is true and is zero otherwise. In particular, the estimated CATE for the observations in leaf 1 equals γ_1 , whereas predictions for observations in leaves two, three and four are equal to $\gamma_1 + \gamma_2$, $\gamma_1 + \gamma_3$ and $\gamma_1 + \gamma_4$ respectively. In practice causal trees are generally deeper, and form more subgroups, than the one presented in figure 1 and equation 14. This illustrates the flexibility of the causal forest algorithm in finding treatment effect heterogeneity. Also, these subgroups are formed by the algorithm based on the data, and not by the researcher based on economic theory, research practice or results.

$$y_{i} = \alpha + \beta X_{i} + \gamma_{1} W_{i} + \gamma_{2} W_{i} * I(x_{3,i} > 5 \& x_{1,i} \le 2) + \gamma_{3} W_{i} * I(x_{3,i} \le 5 \& x_{1,i} > 2) + \gamma_{4} W_{i} * I(x_{3,i} \le 5 \& x_{1,i} \le 2) + \varepsilon$$
(14)

We end this section by considering the objective function a causal forest maximizes when making splits. The parameter τ in equation 13 is identified by the moment condition

$$\mathbb{E}\left[W_i(y_i - W_i\tau(x_i))|x_i\right] = 0 \tag{15}$$

By default $E[y_i|X_i]$ and $E[W_i|X_i]$ are determined using a (honest) regression forest, but researchers can provide alternative estimates, a feature of the algorithm we will discuss later in this paper. A causal forest estimates the moment condition (15) locally by overweighing "similar" observations, where splits in the causal trees in the forest maximize heterogeneity in τ . These splits are made as follows. In a given parent node P, the estimator corresponding to the moment condition (15) is simply the linear regression estimator

$$\hat{\tau}_P = \frac{\sum_{i \in P} W_i y_i}{\sum_{i \in P} W_i^2}.$$

The idea is to split the node P into two children (C_1, C_2) such that the heterogeneity between the estimated parameters, $\hat{\tau}_{C_1}$ and $\hat{\tau}_{C_2}$, increases as fast as possible. Rather than computing $(\hat{\tau}_{C_1}, \hat{\tau}_{C_2})$ for every possible split, Athey et al. (2019) propose a firstorder approximation:

$$\hat{\tau}_C \approx \hat{\tau}_P - \frac{\sum_{i \in C} W_i (y_i - W_i \hat{\tau}_P)}{\sum_{i \in C} W_i^2}$$

for $C \in \{C_1, C_2\}$. They then show that, given this approximation, the optimal split can be made in a way similar to a standard random forest.

After growing the causal forest, an estimate of the treatment effect $\tau(x)$ is given by reweighing the OLS-estimator of 15, where the weight is simply the fraction of trees where individual *i* is in the same final leaf as an individual with covariates *x*. Writing $\alpha_i(x)$ for the weight that individual *i* receives in estimating $\tau(x)$, the causal forest estimate for the CATE is then given by equation 16.

$$\hat{\tau}(x) = \frac{\sum_{i=1}^{N} \alpha_i(x) W_i y_i}{\sum_{i=1}^{N} \alpha_i(x) W_i^2}$$
(16)

In practice, the outcome and treatment variables are recentered locally to improve estimation quality (Athey et al., 2019). In terms of notation y_i and W_i in equations 15 and 16 are replaced by \tilde{y}_i and \tilde{W}_i , where \tilde{i} indicates the variable is residualized by subtracting its expectation conditional on X.

4 Using causal forests in difference-in-differences

We start this section with an explanation how subgroup analysis can be used to estimate conditional average treatment effects. Then, we detail the working of the causal forest algorithm with fixed effects and show under which conditions conditional average treatment effects estimated by a causal forest with fixed effects (CFFE) are identical to those using a manually formed subgroup. Then we describe a manually recentered causal forest (MRCF) and we detail why it provides inconsistent estimates of the conditional average treatment effect. We end this section with a description of a dynamic causal forest (DCF) (Gavrilova et al., 2023).

4.1 Estimating parameters using manually defined subgroups

Suppose a dataset consists of two subgroups g = 0, 1 of sizes N^0 and N^1 that have different parameters τ^g . We can obtain τ^g using a split sample analysis by estimating $y_{it} = \tau^g W_{it} + c_i + \delta_t + \varepsilon_{it}$ for each group separately. Alternatively, following the Frisch-Waugh-Lovell theorem (Frisch and Waugh, 1933; Lovell, 1963), we could have estimated equations 17a and 17b. In equation 17a the effect of individual and time fixed effects is partialled out using the conditional expectations of y_{it} and W_{it} . In equation 17b the within-transformation, indicated by ", is used to partial out time-invariant characteristics (Wooldridge, 2010). The average of within-transformed variables in period t partials out time-varying effects that are constant across individuals (Somaini and Wolak, 2016).

$$\tilde{y}_{it} = \tau^g \tilde{W}_{it} + \varepsilon_{it}, \quad \forall i \in g, \quad g = 0, 1$$

$$\tilde{y}_{it} = y_{it} - E[y_{it}|c_i, \delta_t], \qquad \tilde{W}_{it} = W_{it} - E[W_{it}|c_i, \delta_t]$$
(17a)

$$\begin{aligned} & \breve{y}_{it} = \tau^g \breve{W}_{it} + \varepsilon_{it}, \quad \forall i \in g, \quad g = 0, 1 \\ & \breve{y}_{it} = \ddot{y}_{it} - \frac{1}{N^g} \sum_{i \in N^g} \ddot{y}_{it}, \quad \breve{W}_{it} = \ddot{W}_{it} - \frac{1}{N^g} \sum_{i \in N^g} \ddot{W}_{it} \end{aligned} \tag{17b}$$

All methods provide identical estimates for τ^{g} . In particular, we can compute τ^{g} without relying on estimation methods using equation 18.

$$\hat{\tau}^{g} = \frac{\sum_{t=1}^{\mathcal{T}} \sum_{i \in g} \breve{W}_{it} \breve{y}_{it}}{\sum_{t=1}^{\mathcal{T}} \sum_{i \in g} \left(\breve{W}_{it}\right)^{2}}, \quad g = 0, 1$$
(18)

4.2 Causal forests with fixed effects

Causal forest are designed to estimate τ^g directly from the data y_{it} , X_i , W_{it} in equation 4 without requiring the researcher to specify the subgroups g. Unfortunately $\tau(x)$ cannot be estimated using a regular causal forest. Recall that a regular causal forest uses a regression forest to determine $E[W_{it}|c_i, \delta_t]$. Because W_{it} is an combination of c_i and δ_t , the estimated propensities to be treated are concentrated around zero and one, violating the identifying assumptions of a causal forest.

$$y_{i,t,t^*} = c_i + \delta_t + \delta_{t^*} + \tau(X_i)W_{it^*} + \varepsilon_{i,t,t^*}$$

$$\tag{19}$$

We therefore develop the CFFE, a computationally feasible causal forest that can be used when the data contains many fixed effects. Below we show how a CFFE partials out individual and time fixed effects using the within-transformation and their averages over individuals, analogous to equation 17b where this is done using a split sample analysis. It is trivial to extend the procedure to account for other fixed effects, such time-event fixed effects. Our estimator is numerically equivalent to the causal forest estimator in Athey et al. (2019) with fixed effect dummies included in the "treatment indicator" matrix. Hence, all theoretical results on the consistency and asymptotic distribution of (Athey et al., 2019) extend to our setting. However, this approach is not computationally feasible when the number of fixed effects is large, as a regression with a large number of covariates has to be undertaken in every leaf of the forest.

In particular, CFFE estimates the moment condition given in equation 20 by estimating 21 with OLS. The similarity indicator $\alpha_i(x)$ is crucial to the CFFE-estimator. Recall that $\alpha_i(x)$ is zero for 'dissimilar individuals' that are never in the same leaf as observation *i*. Therefore its incorporation in equation 21 essentially splits the estimation sample such that $\tau(x)$ is estimated using similar individuals only. Furthermore, $\alpha_i(x)$ is also included in the definition of \check{y}_{it}^* and \check{W}_{it}^* , where it guarantees that the average of the within transformed variables at each period is computed using similar individuals only. Finally, time-variant characteristics that are identical for all individuals cancel out because $\sum_{i}^{N} \alpha_i(x) = 1$.

$$\mathbb{E}\left[\left|\breve{W}_{it}^{*}(\breve{y}_{it}^{*}-\breve{W}_{it}^{*}\tau(x_{i}))\right|x_{i}\right]=0,$$

$$\breve{y}_{it}^{*}=\ddot{y}_{it}-\sum_{i=1}^{N}\alpha_{i}(x)\ddot{y}_{it},$$

$$\breve{W}_{it}^{*}=\ddot{W}_{it}-\sum_{i=1}^{N}\alpha_{i}(x)\ddot{W}_{it},$$
(20)

$$\hat{\tau}^{*}(x) = \frac{\sum_{t=1}^{\mathcal{T}} \sum_{i=1}^{N} \alpha_{i}(x) \breve{W}_{it}^{*} \breve{y}_{it}^{*}}{\sum_{t=1}^{\mathcal{T}} \sum_{i=1}^{N} \alpha_{i}(x) \left(\breve{W}_{it}^{*}\right)^{2}}$$
(21)

In general, the right-hand-side of equation 21 can differ from the right-hand-side of equation 18, because the weights α_i need not be equal for all observations in the same group g. This is because the weight α_i equals the average of one over the leaf size when observation j shares at least one leaf with observation i and is zero otherwise.¹⁰ Note that $\tau^*(i \in g)$ in equation 21 is exactly equal to τ^g in equation 18 in the special case that $\alpha_i(i \in g)$ equals $1/N^g$ when individual j is in the same group as i and is zero otherwise.

As identification in a difference-in-differences research design comes from the parallel trend assumption, $\hat{\tau}^*(x)$ identifies the CATT. In sections 5 and 6 we use simulations to investigate how well this parameter is estimated by CFFE in general. In the simulations, we will impose treatment effects that vary at the individual level and with time since treatment, so we estimate equation 21 for each period after treatment has begun. To do this, we combine the observations for individuals from a specific cohort and period with all their observations before treatment. This prevents that observations from different time periods end up in the same leaf of the causal forest, which would bias estimates when treatment effects are dynamic. Similar to the estimation strategy in Borusyak (2023) this approach takes advantage of all the observations before treatment.

4.3 Manually recentered causal forest

As Athey et al. (2019) note, in general generalized random forests remain consistent after recentering. In fact, local centering can improve the small sample properties of the estimator. This would seem to imply that we can demean our variables manually and use the general algorithm to compute generalized random forests.

Concretely, one might consider the following approach of a manually recentered causal forest (MRCF). In particular, $E[y_{it}|c_i, \delta_t]$ in the moment condition estimated by

¹⁰Causal forest use subsampling for honest estimation. Suppose observations j and k are in the same group as i. Even if the causal forests places j and k in the same leaf as i in every tree of the causal forest, the weight $\alpha_{i,j}$ will be different from $\alpha_{i,k}$ when i and j are sampled together more often than observations i and k or when sample sizes differ.

a MRCF (the panel version of equation 15) is replaced by $\ddot{y}_{it} - \frac{1}{N} \sum_{N} \ddot{y}_{it}$ and $E[W_{it}|c_i, \delta_t]$ is replaced by $\ddot{W}_{it} - \frac{1}{N} \sum_{N} \ddot{W}_{it}$. In this case, the estimator becomes

$$\hat{\tau}^{\dagger}(x) = \frac{\sum_{t=1}^{\mathcal{T}} \sum_{i=1}^{N} \alpha_i(x) \left(\ddot{W}_{it} - \frac{1}{N} \sum_N \ddot{W}_{it} \right) \left(\ddot{y}_{it} - \frac{1}{N} \sum_N \ddot{y}_{it} \right)}{\sum_{t=1}^{\mathcal{T}} \sum_{i=1}^{N} \alpha_i(x) \left(\ddot{W}_{it} - \frac{1}{N} \sum_N \ddot{W}_{it} \right)^2}$$
(22)

However, this estimator does not identify the CATT when a difference-in-differences research design is used. The reason is that this type of recentering is not local. Recentering works because if the CATT $\tau(x)$ is constant for some subset of S of individuals, then

$$\frac{\sum_{S} (y_{it} - \mathbb{E}[y_{it}|S])(W_{it} - \mathbb{E}[W_{it}|S])}{\sum_{S} (W_{it} - \mathbb{E}[W_{it}|S])^2}$$

also identifies the CATT. That is, since we can estimate the CATT using linear regression on a subset of individuals with the same treatment effect, we can do the same by locally recentering. In essence, the causal forest gives us the individuals that are likely to be similar.

From this, we can see that identification fails because the subset of individuals over which the MRCF centers does not have the same treatment effect. In particular, we use *all* individuals to remove the time dimension. Hence, the manually recentered forest only identifies the treatment effect when it is homogeneous.

To see this more clearly, consider the following stylized example. Assume that the causal forest makes optimal splits and that the weights in $\alpha_i (i \in g)$ are $1/N^g$ for similar observations and zero otherwise. Then $\hat{\tau}^{\dagger}(i \in g)$ becomes:

$$\hat{\tau}^{\dagger}(i \in g) = \frac{\sum_{t=1}^{\mathcal{T}} \sum_{i \in g} \left(\ddot{W}_{it} - \frac{1}{N} \sum_{N} \ddot{W}_{it} \right) \left(\ddot{y}_{it} - \frac{1}{N} \sum_{N} \ddot{y}_{it} \right)}{\sum_{t=1}^{\mathcal{T}} \sum_{i \in g} \left(\ddot{W}_{it} - \frac{1}{N} \sum_{N} \ddot{W}_{it} \right)^2}$$

This estimate is different from the one in equation 17b. It uses the average over all individuals at period t to control for time fixed effects, whereas equation 17b uses the group average. We note that these estimators do not converge to the same limit. In particular, $1/N \sum_i \ddot{y}_{it}$ converges to $\mathbb{E}[y_{it}]$, while $\sum_i \alpha_i(x)\ddot{y}_{it}$ converges to $\mathbb{E}[y_{it}|x]$. A similar divergence occurs for the average of the treatment indicator W_{it} . Since the CFFE is known to be consistent based on the results in Athey et al. (2019), the MRCF must be inconsistent even for asympotically valid forest weights. In particular, the bias is large for subgroups with characteristics x such that the subgroup average of the within-transformed outcome variable and the subgroup average of the within-transformed treatment indicator are very different from the sample average. For the same reasons as described in the previous section, we will estimate the treatment effects separately for each period after treatment has begun.

4.4 Dynamic causal forest

Recently, Gavrilova et al. (2023) demonstrated how a causal forest can be used in a difference-in-differences setting by transforming the data. Specifically, they defined ∇Y_{it} and ∇W_{it} as the differences in the outcome and the treatment indicator between treatment year t and a base period b, usually the last year before the treatment began. Note that ∇W_{it} is equivalent to the treatment group indicator T_i . Taking this difference eliminates the individual fixed effects. Additionally, since the authors estimate the treatment effect separately for each year t, the trend term $\nabla \delta_t = \delta_t - \delta_b$ is constant and does not affect the estimates for the treatment effect. In particular, their approach uses the set of variables X both for recentering of the transformed outcome and treatment indicator and to make the similarity weights $\alpha_i(x)$ as described in section 3 and therefore the treatment effect estimated by DCFs is simply defined by equation 23.

$$\hat{\tau}^{\ddagger}(x) = \frac{\sum_{t=1}^{\mathcal{T}} \sum_{i=1}^{N} \alpha_i(x) \tilde{\nabla} W_{it} \tilde{\nabla} y_{it}}{\sum_{t=1}^{\mathcal{T}} \sum_{i=1}^{N} \alpha_i(x) \left(\tilde{\nabla} W_{it}\right)^2}$$

$$\tilde{\nabla} y_{it} = \nabla y_{it} - E[\nabla y_{it}|X],$$

$$\nabla y_{it} = y_{it} - y_{ib},$$

$$\tilde{\nabla} W_{it} = \nabla W_{it} - E[\nabla W_{it}|X],$$

$$\nabla W_{it} = W_{it} - W_{ib} = T_i,$$

$$(23)$$

The information exploited by a DCF differs from that exploited by a CFFE, because a DCF compares outcomes in an event period to a base period, whereas a CFFE uses the average of all observations before treatment. As was discussed in section 2 the latter approach is more efficient because more information is used, but it requires more stringent conditions on the common trend and serial correlation of the errors.

5 Simulation single-event difference-in-differences

In this section we present simulation results when a CFFE is used to estimate treatment effect heterogeneity using a difference-in-differences research design. We consider a difference-in-differences setup with one treatment period. The CFFE is provided a vector of outcomes, a vector of treatment indicators and two vectors indicating individual and time fixed effects. We compare this estimator to a MRCF. This estimator is provided a vector of outcomes and treatment indicators as well. In addition, the MRCF receives a vector of recentered outcomes and recentered treatment indicators, that are computed as described in section 4.3. Both the MRCF and the CFFE use the ten x-variables to look for treatment effect heterogeneity using splits. Both estimators use the default settings from the GRF-package, except for the fact that observations are clustered at the individual level to account for the use of panel data Wager and Athey (2018); Athey and Wager (2019).

5.1 Simulation

Equation 24 describes the data generating process (DGP) for the single event simulation. In short, we sample 1500 individuals (indexed *i*) who are observed in four periods (t = 1, 2, 3, 4), where the last two periods define the treatment period. Besides the outcome variable *y* we observe ten time-invariant individual characteristics $(x_{k,i}, k = 1, ..., 10)$. c_i denote unobserved individual characteristics. People in the treatment group $(W_i = 1)$ have structurally higher outcomes than people in the control group $(W_i = 0)$ when $\kappa > 0$. About half of the individuals is treated. x_1 is a pivotal variable in the simulated data as it affects a) the size of the treatment effect, b) the propensity to be treated and c) the outcome variable. In particular, the treatment effect is nonlinear in x_1 and equals $max(0, x_1t^*)$, where t^* indicates the number of periods elapsed since treatment has started. Thus treatment effects vary at the individual level and they are dynamic. Treatment status is positively correlated with x_1 and negatively with x_2 as $P(W_i = 1) = exp(x_{i1})/(exp(x_{i1}) + exp(x_{i2}))$. We vary the difference in outcomes (in absence of treatment) between the treatment and control group by setting $\kappa = 0, 5$. Thus, we consider results when all relevant individual fixed effects are observed ($\kappa = 0$) and results when unobserved heterogeneity at the individual level influences the outcome ($\kappa = 5$). We consider a DGP with and without time fixed effects by choosing $\lambda = 5$ or $\lambda = 0$. Finally, all x variables and the error term follows a uniform distribution.

$$y_{it}^{\kappa,\lambda} = \kappa(W_i + c_i) - \lambda t + \tau(X_i)W_{it} + x_{i1}^2 + x_{i2}^2 + \varepsilon_{it},$$
(24)

$$\kappa = 0, 5, \lambda = 0, 5$$

$$P(W_{it} = 1) = exp(x_{i1})/(exp(x_{i1}) + exp(x_{i2})),$$

$$W_{it} = post_t * W_i,$$

$$\tau(X_i) = max(x_1t^*, 0),$$

$$c_i, X_{ik} \sim U(-1, 1), k = 1, \dots, 10$$

Figure 2 shows the simulated data (when $\kappa = 5$ and $\lambda = 5$). In particular the outcome variable is larger for people in the treatment group in all periods. The difference between the treatment and control group in abcence of treatment is constant before *and* after the treatment period has started. Due to the treatment effect, the average outcome for the treated when treated is larger than the average outcome when not treated. Note that the initial difference in outcomes between the control and treatment group and the time fixed effects are large compared to the size of treatment effect.



Figure 2: Illustration of the DGP

Average outcome for the control group (solid red line) and the treatment group (solid magenta line). The dashed line indicates the average outcome for the treatment group in the abcence of treatment.

5.2 Simulation results single-event difference-in-differences

The simulation results demonstrate that the CFFE-estimator produces estimates that are consistent. Figure 3 shows that the bias of the estimator is centered around zero. The expected bias, indicated by the dashed line, is zero, despite the fact that there are some outliers in which our method severly underpredicts. The right panel of the figure reveals that the predicted CATEs are concentrated around the true CATEs. The estimator picks up the kink at around $X_1 = 0$ and then increasing with the true CATE. However, for large outlier values of the forcing variable $x_1 * (t-2)$ the method predicts too low values. This occurs because forest type models underpredict (in absolute terms) at the boundary space of the features it uses (Athey and Imbens, 2019).



Figure 3: causal forest with fixed effects

Estimation results when a causal forest with fixed effects is used to estimate conditional average treatment effects in a single event difference-in-differences research design. Panel (a) shows the bias of the estimated treatment effect $(\hat{\tau} - \tau)$ when $\kappa = 5$ and $\lambda = 5$. Panel (b) shows the estimated values of the treatment effect $(\hat{\tau})$ and the true treatment effect (τ) as a function of the forcing variable, x_1 , when $\kappa = 5$ and $\lambda = 5$. Dots indicate the estimated treatment effects. The true treatment effect is indicated by the solid line.

In contrast, the CATEs estimated by a MRCF are inconsistent. Figure 4 shows that most of the estimated treatment effects are biased upwards and therefore the bias of this estimator is not concentrated around zero (see Figure 4a). Indeed, this upward bias also appears when we plot the estimated CATEs against the forcing variable $(x_1*(t-2))$. The upward bias becomes clearly visible when the forcing variable exceeds one in absolute terms (Figure 4b). Also the estimates by the MRCF are more noisy than those from a CFFE, especially when the true treatment effect is zero. This conclusion follows from a comparison of the MRCF in Figure 4b to a CFFE in Figure 3b. Clearly, the estimates by the MRCF are less well concentrated around the true CATE.



Figure 4: manually recentered causal forest

Estimation results when a manually recentered causal forest is used to estimate conditional average treatment effects in a single event difference-in-differences research design. Panel (a) shows the bias of the estimated treatment effect $(\hat{\tau} - \tau)$ when $\kappa = 5$ and $\lambda = 5$. Panel (b) shows the estimated values of the treatment effect $(\hat{\tau})$ and the true treatment effect (τ) as a function of the forcing variable, x_1 , when $\kappa = 5$ and $\lambda = 5$. Dots indicate the estimated treatment effects. The true treatment effect is indicated by the solid line.

Figure 5 shows that the estimates from a DCF are similar to those provided by a CFFE. The plot in the left panel shows a large mass around zero, although the top of the distribution is slightly at the right of zero. Again, there are some outlier observations with a negative bias (figure 5a). The right plot shows that the estimates of the DCF are concentrated around the true CATE. In particular, the method succeeds in estimating the kink at x - 1 = 0 and the estimated treatment effect increases linearly afterwards. Again, panels 5a and 5b illustrate that also a DCF predicts too low values for outlier observations of the forcing variable $x_1 * (t - 2)$.



Figure 5: dynamic causal forest

Estimation results when a dynamic causal forest is used to estimate conditional average treatment effects in a single event difference-in-differences research design. Panel (a) shows the bias of the estimated treatment effect $(\hat{\tau} - \tau)$ when $\kappa = 5$ and $\lambda = 5$. Panel (b) shows the estimated values of the treatment effect $(\hat{\tau})$ and the true treatment effect (τ) as a function of the forcing variable, x_1 , when $\kappa = 5$ and $\lambda = 5$. Dots indicate the estimated treatment effects. The true treatment effect is indicated by the solid line.

The predictions of a CFFE are superior to those of a MRCF in all variants we consider as it has less bias and a lower RMSE. This conclusion also holds compared to a DCF when individual and/or time fixed effects are present. We present results on the bias in table 1. It shows that the average bias of a CFFE is about two percent when there are no individual or time fixed effects. The bias is virtually zero when individual and time fixed effects are present. This constrasts strongly to the average bias of a MRCF which is about -0.48 regardless of fixed effects being present. The bias of the DCF is slightly higher than a CFFE when individual and/or time fixed effects are present, between two and five percent, but the bias is of the DCF is lower than that of the CFFE (about one and about two percent) when there are no individual or time fixed effect ($\kappa = 0$ and $\lambda = 1$). However, these numbers are not statistically different as the bias of the one is within the 95 percent confidence interval of the other.

The efficient of a CFFE is higher than that of MRCF and DCF, as is shown in table 2. As τ is known, we compute the RMSE as $\left(\frac{1}{N}\sum_{i}^{N}(\tau_{i}-\hat{\tau}_{i})^{2}\right)^{0.5}$. The RMSE of the CFFE ranges between 0.17 and 0.24. In contrasts, the RMSEs of the MRCF range between 0.81 and 0.85, and those of a DCF between 0.24 and 0.29. Note that the size of the RMSEs from the CFFE are only about 75 percent of those from a DCF and that the RMSE of the DCF is always outside the confidence interval of the CFFE. This shows the gains in efficiency from a CFFE as measured by the RMSE is substantial.

We investigate to what extent the better performance of the CFFE compared to the DCF is due to the additional observations it uses, by re-estimating both estimators using observations in periods two and three only. This rules out differences in performance

due to differences in the information sets applied. First, the absolute value of the bias of this CFFE is slightly higher than that of a DCF when there are no time fixed effects $(\lambda = 0)$, see table 1. When time fixed effects are present, the bias of the CFFE is lower than that of a DCF.

Table 2 show that the RMSE of the CFFE is lower than that of the DCF in three out of four simulations that we consider. But as the relative difference between the two estimators is smaller when the CFFE uses all periods before treatment started, we conclude from this that the CFFE is more efficient than the DCF mainly because it can exploit information contained in additional observations.

estimator	κ	λ	bias	cfi
CFFE	0	0	0.02	(0.0148 - 0.0275)
CFFE	5	0	0.00	(-0.0094 - 0.0085)
CFFE	0	5	0.00	(-0.006 - 0.0089)
CFFE	5	5	0.00	(-0.0057 - 0.0069)
MRCF	0	0	-0.48	(-0.50060.4538)
MRCF	5	0	-0.49	(-0.51350.4647)
MRCF	0	5	-0.50	(-0.52240.4752)
MRCF	5	5	-0.49	(-0.51590.4667)
DCF	0	0	0.01	(0.0034 - 0.0221)
DCF	5	0	0.02	(0.0123 - 0.033)
DCF	0	5	0.05	(0.0429 - 0.063)
DCF	5	5	0.04	(0.0287 - 0.0457)
CFFE 2P	0	0	-0.05	(-0.05950.0407)
CFFE 2P	5	0	-0.02	(-0.03550.0133)
CFFE 2P	0	5	-0.00	(-0.0117 - 0.0109)
CFFE 2P	5	5	0.01	(0.0074 - 0.0168)
DCF $2P$	0	0	-0.02	(-0.03520.0137)
DCF 2P	5	0	0.01	(-0.0076 - 0.0182)
DCF 2P	0	5	0.02	(0.0089 - 0.0295)
DCF $2P$	5	5	0.05	(0.0443 - 0.0584)

Table 1: bias simulation 1

The table shows the average of the bias when a causal forest with fixed effects (cffe) is used to estimate heterogeneous treatment effects and when this is done using a manually recentered causal forest (mrcf) or a dynamic causal forest (dcf). We also present results when a CFFE and DCF are trained on periods 2 and 3 only (cffe 2p and dcf 2p). Each row presents the average of the bias for various levels of the individual effects vary ($\kappa = 0, 5$) and time fixed effects ($\lambda = 0, 5$). The reported value is computed as $\frac{1}{N} \sum_{i}^{N} (\hat{\tau}_{i} - \tau_{i})$, where τ_{i} indicates the true treatment effect for individual i and $\hat{\tau}_{i}$ indicates the estimated treatment effect for individual i. The numbers within brackets denote the bootstrapped 95 percent confidence interval based on 500 replications.

Finally, the simulations demonstrate that only the estimates of the average treatment effect provided by the CFFE contain the true value of this parameter in the data (see Table 3). Recall that Athey et al. (2019)'s causal forest algorithm provides consistent estimates for the conditional average treatment effect and its variance. We therefore simulated 500 draws from a normal distribution with mean equal to $\hat{\tau}_i$ and variance equal to the predicted variance $\hat{\sigma}_i$. The average treatment effect was then calculated

estimator	κ	λ	rmse	cfi
CFFE	0	0	0.18	(0.1671 - 0.1839)
CFFE	5	0	0.24	(0.2275 - 0.2463)
CFFE	0	5	0.21	(0.1998 - 0.2184)
CFFE	5	5	0.17	(0.1622 - 0.176)
MRCF	0	0	0.81	(0.7805 - 0.8332)
MRCF	5	0	0.84	(0.8147 - 0.8621)
MRCF	0	5	0.85	(0.8264 - 0.8787)
MRCF	5	5	0.85	(0.8186 - 0.8759)
DCF	0	0	0.26	(0.2412 - 0.2691)
DCF	5	0	0.29	(0.2756 - 0.3101)
DCF	0	5	0.28	(0.2607 - 0.296)
DCF	5	5	0.24	(0.2224 - 0.254)
CFFE 2P	0	0	0.19	(0.1835 - 0.1954)
CFFE 2P	5	0	0.23	(0.2138 - 0.2379)
CFFE 2P	0	5	0.23	(0.2222 - 0.2419)
CFFE 2P	5	5	0.10	(0.0944 - 0.1044)
DCF $2P$	0	0	0.22	(0.2039 - 0.2287)
DCF 2P	5	0	0.26	(0.242 - 0.2706)
DCF 2P	0	5	0.20	(0.1931 - 0.2172)
DCF 2P	5	5	0.14	(0.1356 - 0.1539)

Table 2: RMSE simulation 1

The table shows the RMSE when a causal forest with fixed effects (cffe) is used to estimate heterogeneous treatment effects and when this is done using a manually recentered causal forest (mrcf) or a dynamic causal forest (dcf). We also present results when a CFFE and DCF are trained on periods 2 and 3 only (cffe 2p and dcf 2p). Each row presents the RMSE for various levels of the individual effects vary ($\kappa = 0, 5$) and time fixed effects ($\lambda = 0, 5$). The RMSE is computed as $\left(\frac{1}{N}\sum_{i}^{N}(\tau_{i}-\hat{\tau}_{i})^{2}\right)^{0.5}$, where τ_{i} indicates the true treatment effect for individual i and $\hat{\tau}_{i}$ indicates the estimated treatment effect for individual i. The numbers within brackets denote the bootstrapped 95 percent confidence interval based on 500 replications.

as the mean of the simulated treatment effects for all observations, both treated and untreated individuals, in the treatment period. Table 3 shows that the mean of these simulations for the CFFE is close or identical to the value of the parameter in the data after rounding. In contrast, the mean of the simulated estimates from the DCF is lower than the value of the average treatment in the data and the true average treatment effect is outside its confidence bound. This suggests that the additional information exploited by the CFFE improves the estimation of the average treatment effect. We should mention that the difference between the average treatment effect computed by a DCF and the true value of this parameter is small, especially when compared to the upward biased estimate of the average treatment effect provided by a MRCF.

Overall, we conclude that the performance of the CFFE and the DCF in our first simulation is very similar, although the performance of the former is better in terms of average bias and RMSE. In appendix A we visually show the performance of the CFFE, MRCF and DCF estimators for other values of κ and λ . Appendix B summarizes the bias, RMSE and ATE when the error term is normally distributed. Results are

var	κ	λ	mean	lb	ub
truth			1.84		
CFFE	0	0	1.82	1.81	1.82
CFFE	0	5	1.84	1.83	1.84
CFFE	5	0	1.84	1.83	1.84
CFFE	5	5	1.84	1.83	1.84
MRCF	0	0	2.31	2.30	2.32
MRCF	0	5	2.34	2.32	2.35
MRCF	5	0	2.32	2.31	2.34
MRCF	5	5	2.33	2.31	2.34
DCF	0	0	1.82	1.82	1.83
DCF	0	5	1.78	1.78	1.79
DCF	5	0	1.81	1.81	1.82
DCF	5	5	1.80	1.79	1.81

qualitatively similar to the ones presented here.

Table	3:	ATE	simul	lation	1
-------	----	-----	-------	--------	---

The table shows the averages of the true and estimated treatment effects by estimator for various levels of κ and λ . "estimator = cffe" indicates the average of the causal forest with fixed effects and "estimator = mrcf" indicates the average of a manually recentered causal forest. The table reports the average (fourth column) as well as the lower and upper bound of the 95 percent confidence interval (columns five and six) of the ATE. These bounds have been computed by simulating 500 draws of the estimated treatment effect from a normal distribution with mean $\hat{\tau}_i$ and variance $\hat{\sigma}_i$. Then the average treatment effect is computed for each draws and the lower and upper bound are the 2.5th and 97.5th percentile.

6 Simulation staggered difference-in-differences

In this section we present simulation results when a CFFE is used to estimate treatment effect heterogeneity using a staggered difference-in-differences research design. Again, the CFFE is provided a vector of outcomes and a vector of treatment indicators. Yet it also has access to three vectors that indicate individual fixed effects, time fixed effects and event time fixed effects respectively. Again, we compare this estimator to a MRCF and DCF. The MRCF uses the same vector of outcomes and treatment indicators as the CFFE, but it uses vectors of recenter outcomes and recentered treatment indicators, that are computed as described in section 4.3, to partial out individual, time and event time fixed effects. The DCF transforms the outcome vector as described in section 4.4. All estimators use the ten x-variables to look for treatment effect heterogeneity. They use the default settings from the GRF-package, except for the fact that observations are clustered at the individual level.

6.1 Simulation

We consider a staggered difference-in-differences research design described by equation 25. In particular, we assume we observe individuals for eight periods and we consider

two treatment events (indicated with superscript $\mu = 1, 2$), one starting in period three and the other in period seven. The propensity to receive treatment is the same in both treatment events, but we allow for (substantial) heterogeneity in treatment effects across individuals and over time. In particular the treatment effect in the first event equals $max(0, x_1t^*)$, but the treatment effect in the second event is twice as large. Again, t^* denotes the time since treatment has started. Following the conventions of a stacked difference-in-differences research design, we discard individuals that are treated in both treatment events and randomly assign never treated observations to the control group for the first or second treatment event. In order to compare estimation results to the first simulation, we ensure that fifteen hundred individuals are sampled after these selections are made.

$$y_{it}^{\kappa,\lambda,\mu} = \kappa(W_i + c_i) - \lambda t + \tau^{\mu}(X_i)W_{it} + x_{i1}^2 + x_{i2}^2 + \varepsilon_{it},$$
(25)

$$\kappa = 0, 5, \lambda = 0, 5, \mu = 1, 2$$

$$W_{it} = post_t * W_i,$$

$$P(W_{it} = 1)(X_i) = exp(x_{i1})/(exp(x_{i1}) + exp(x_{i2}))$$

$$\tau^{\mu}(X_i) = \mu * max(x_1t^*, 0)$$

$$c_i, X_{ik} \sim U(-1, 1), k = 1, \dots, 10$$

Figure 6 plots the mean outcome for the treatment and control groups as a function of calender time t. It clearly shows that conditioning on time fixed effects (δ_t) and event time fixed effects (δ_{t^*}) is required to identify the treatment effect. In particular, the simulated data is characterized by substantial differences between the control group and the treatment groups in absence of treatment $(\kappa = 5)$ and by substantial time fixed effects $(\lambda = 5)$ compared to the size of both treatment effects. The figure also shows that all observations are only observed two periods before and after treatment.



Figure 6: Illustration of the DGP (staggered design)

Illustration of the data simulated in the staggered design. The data shows the average outcomes for the treatment group and the matched control group for the first treatment event (panel 6a and the second event (panel 6b). In these panels, the average outcomes for the control group is indicated by the solid red line. The observed outcomes for the treatment group are indicated by the solid magenta line. The average outcomes for the treatment group in the absence of treatment is indicated by the dashed line.

6.2 Simulation results staggered difference-in-differences

The simulation results for a staggered difference-in-differences research design demonstrate that the estimates from a CFFE are consistent and more efficient than those from a DCF. Additionally, we conclude that the estimates from a MRCF are inconsistent. Figure 7 shows that the estimates of a CFFE are unbiased. The bias from a CFFE is small and centered close to zero in both treatment events (panels 7a and 7b). Examining the distribution of the estimated treatment effects around their true value, it is evident that the size difference in treatment effects between the first and second event is captured by the estimator, as seen in panels 7c and 7d. This is as expected, since treatment effects are estimated separately for each year and treatment event. Moreover, the kink at around x_1t^* is accurately identified by the estimator in both events. Again, the CFFE provides a too low estimate at the boundary space of the forcing variable x_1t^* , as we have see in the previous simulation as well.



Figure 7: causal forest with fixed effects (staggered design)

Estimation results when a causal forest with fixed effects is used to estimate conditional average treatment effects in a staggered difference-in-differences research design. Panel (a) and (b) show the bias of the estimated treatment effect $(\hat{\tau} - \tau)$ when $\kappa = 5$ and $\lambda = 5$. Panel (a) shows results for the first treatment event, panel (b) for the second. Panels (c) and (d) show the estimated values of the treatment effect $(\hat{\tau})$ and the true treatment effect (τ) as a function of the forcing variable, x_1 , when $\kappa = 5$ and $\lambda = 5$. Dots indicate the estimated treatment effects. The true treatment effect is indicated by the solid line. Panel (c) shows results for the first treatment event, panel (d) for the second.

The shortcomings of a MRCF in a single event difference-in-differences research design also appear when a staggered difference-in-differences research design is used. Panels 8a and 8b show that its estimates are biased upwards severely. This picture is confirmed when we compare the predicted treatment effects to the true treatment effects in panels 8c and 8d. Indeed, estimated treatment effects are structurally above the true treatment effects for both treatment events. Again we conclude that estimates are biased upwards both when treatment effects should be zero (which is the case for observations with negative values of x_1) and when treatment effects should be large. On the bright side, the MRCF successfully detects the heterogeneity in treatment effect over treatment events, albeit with bias. Also, the estimated effects for the second treatment event are about two times as large as those estimated for the first treatment event, as they should be.



Figure 8: manually recentered causal forest (staggered design)

Estimation results when a manually recentered causal forest is used to estimate conditional average treatment effects in a staggered difference-in-differences research design. Panel (a) and (b) show the bias of the estimated treatment effect $(\hat{\tau} - \tau)$ when $\kappa = 5$ and $\lambda = 5$. Panel (a) shows results for the first treatment event, panel (b) for the second. Panels (c) and (d) show the estimated values of the treatment effect $(\hat{\tau})$ and the true treatment effect (τ) as a function of the forcing variable, x_1 , when $\kappa = 5$ and $\lambda = 5$. Dots indicate the estimated treatment effects. The true treatment effect is indicated by the solid line. Panel (c) shows results for the first treatment event, panel (d) for the second.

The results from a DCF are by and large comparable to those from a CFFE also when estimating treatment effects for a staggered difference-in-differences design. Figures 9a and 9b demonstrate that the bias of the DCF predictions is generally centered around zero, although there are some outliers with a negative bias especially in the first treatment. This bias is mainly caused by observations with high values of the forcing variable x_1t^* driving the size of the treatment effect (see figures 9c and 9d). A comparison of figures 7 and 9 suggests that in this simulation the DCF underestimates more for large values of the forcing variable than a CFFE. The DCF captures the kink in the treatment effects when $x_1t^* = 0$, and the estimates for the second event are approximately double those of the first.

Our simulation suggests that the CFFE is more accurate in predicting the CATT than the DCF, although both estimators have a similar estimation quality in terms of bias when compared to that of the MRCF. Table 4 shows that the bias of the CFFE (about 0.04 when $\kappa = 5$ and $\lambda = 5$) is lower than that of the MRCF (about -1.34) and the DCF (about 0.11). That the bias of the MRCF is substantially higher than that of the CFFE and DCF is in line with our theoretical results discussed in section 4.3.



Figure 9: dynamic causal forest (staggered design)

Estimation results when a dynamic causal forest is used to estimate conditional average treatment effects in a staggered difference-in-differences research design. Panel (a) and (b) show the bias of the estimated treatment effect $(\hat{\tau} - \tau)$ when $\kappa = 5$ and $\lambda = 5$. Panel (a) shows results for the first treatment event, panel (b) for the second. Panels (c) and (d) show the estimated values of the treatment effect $(\hat{\tau})$ and the true treatment effect (τ) as a function of the forcing variable, x_1 , when $\kappa = 5$ and $\lambda = 5$. Dots indicate the estimated treatment effects. The true treatment effect is indicated by the solid line. Panel (c) shows results for the first treatment event, panel (d) for the second.

The performance of the CFFE is superior to that of the MRCF and the DCF also in terms of efficiency. We derive this conclusion from table 5 which shows the RMSE for the three causal forest estimators for different values of the individual and time fixed effects. The RMSE for the CFFE is quite similar regardless of the considered values for

estimator	κ	λ	bias	cfi
CFFE	0	0	0.03	(0.0089 - 0.0475)
CFFE	5	0	0.02	(0.0043 - 0.0409)
CFFE	0	5	-0.01	(-0.0281 - 0.0095)
CFFE	5	5	0.04	(0.0173 - 0.0589)
MRCF	0	0	-1.36	(-1.43861.2871)
MRCF	5	0	-1.35	(-1.42231.2726)
MRCF	0	5	-1.39	(-1.46271.3135)
MRCF	5	5	-1.34	(-1.41681.2656)
DCF	0	0	0.13	(0.0931 - 0.1697)
DCF	5	0	0.13	(0.0942 - 0.1719)
DCF	0	5	0.11	(0.0657 - 0.1479)
DCF	5	5	0.11	(0.073 - 0.1539)
CFFE 2P	0	0	0.02	(-0.0023 - 0.0431)
CFFE 2P	5	0	0.03	(0.0094 - 0.0478)
CFFE 2P	0	5	-0.03	(-0.0540.0089)
CFFE 2P	5	5	-0.01	(-0.0352 - 0.0145)
DCF 2P	0	0	0.08	(0.0411 - 0.1161)
DCF 2P	5	0	0.10	(0.0672 - 0.1386)
DCF $2P$	0	5	0.04	(0.0031 - 0.0786)
DCF $2P$	5	5	0.07	(0.0295 - 0.1065)

Table 4: bias simulation 2

The table shows the average of the bias when a causal forest with fixed effects (cffe) is used to estimate heterogeneous treatment effects and when this is done using a manually recentered causal forest (mrcf) or a dynamic causal forest (dcf). We also present results when a CFFE and DCF are trained on periods 2 and 3 only (cffe 2p and dcf 2p). Each row presents the average of the bias for various levels of the individual effects vary ($\kappa = 0, 5$) and time fixed effects ($\lambda = 0, 5$). The reported value is computed as $\frac{1}{N} \sum_{i}^{N} (\hat{\tau}_{i} - \tau_{i})$, where τ_{i} indicates the true treatment effect for individual i and $\hat{\tau}_{i}$ indicates the estimated treatment effect for individual i. The numbers within brackets denote the bootstrapped 95 percent confidence interval based on 500 replications.

the individual and time fixed effects and equal to about 0.59 when $\kappa = 5$ and $\lambda = 5$. The RMSE of the MRCF and DCF are substantially higher with values of about 2.53 and 1.20 respectively. Note that these values are outside the bootstrapped confidence bound for the CFFE estimator. These findings confirm the conclusion in De Chaisemartin and d'Haultfoeuille (2023); Roth et al. (2023) that TWFE-estimators comparing treatment-periods to one base-period are less efficient than those that compare treatment-periods to all periods before treatment, provided the common trend assumption holds in all periods used for estimation (as is the case in our simulation).

Again, we estimate a CFFE and a DCF on identical sets of observations, the observations just before and just after treatment, to investigate to what extent the superior predictive performance of the CFFE is explained by the use of additional observations before treatment. Table 4 shows that the bias of the CFFE 2P estimator is lower than that of the DCF 2P estimator, although both numbers are low. The same picture arises when we look at the RMSE: this value is always lower for the CFFE 2P-estimator compared to the DCF 2P-estimator. However the relative performance of CFFE 2P against

DCF 2P is not as good as the relative performance of CFFE against DCF, which leads us to conclude that the additional information used by the CFFE is the main driver of the better performance of the CFFE estimator.

estimator	κ	λ	rmse	cfi
CFFE	0	0	0.52	(0.4719 - 0.5781)
CFFE	5	0	0.51	(0.4553 - 0.562)
CFFE	0	5	0.52	(0.4681 - 0.5737)
CFFE	5	5	0.58	(0.5273 - 0.6327)
MRCF	0	0	2.53	(2.4204 - 2.6456)
MRCF	5	0	2.53	(2.4234 - 2.6406)
MRCF	0	5	2.56	(2.4528 - 2.6815)
MRCF	5	5	2.53	(2.4184 - 2.6445)
DCF	0	0	1.16	(1.0675 - 1.2426)
DCF	5	0	1.13	(1.0385 - 1.2103)
DCF	0	5	1.21	(1.1211 - 1.2986)
DCF	5	5	1.20	(1.106 - 1.2812)
CFFE 2P	0	0	0.48	(0.4336 - 0.5326)
CFFE 2P	5	0	0.38	(0.3306 - 0.4342)
CFFE 2P	0	5	0.45	(0.4 - 0.5004)
CFFE 2P	5	5	0.50	(0.4578 - 0.5549)
DCF $2P$	0	0	0.76	(0.6881 - 0.8272)
DCF $2P$	5	0	0.70	(0.6331 - 0.7663)
DCF $2P$	0	5	0.77	(0.7039 - 0.8411)
DCF $2P$	5	5	0.77	(0.7041 - 0.8477)

Table 5: RMSE simulation 2

The table shows the RMSE when a causal forest with fixed effects (cffe) is used to estimate heterogeneous treatment effects and when this is done using a manually recentered causal forest (mrcf) or a dynamic causal forest (dcf). We also present results when a CFFE and DCF are trained on periods 2 and 3 only (cffe 2p and dcf 2p). Each row presents the RMSE for various levels of the individual effects vary ($\kappa = 0, 5$) and time fixed effects ($\lambda = 0, 5$). The RMSE is computed as $\left(\frac{1}{N}\sum_{i}^{N}(\tau_{i}-\hat{\tau}_{i})^{2}\right)^{0.5}$, where τ_{i} indicates the true treatment effect for individual i and $\hat{\tau}_{i}$ indicates the estimated treatment effect for individual i. The numbers within brackets denote the bootstrapped 95 percent confidence interval based on 500 replications.

Finally, we turn to estimates of the ATE and we conclude that its estimates by the CFFE are closest to the average treatment effect in the data. Table 6 shows the estimate for the ATE by the CFFE (about 4.22), is very close to the simulated average treatment effect of 4.24. Although the relative difference is less than one percent, the 95 percent confidence interval for the CFFE is so narrow that it does not contain the true parameter except when kappa = 0 and $\lambda = 5$. The DCF estimates an average treatment effect of about 4.12, with a relative difference from the true parameter value of less than three percent. In contrast, the average treatment effect estimated by a MRCF is about 5.6, which is substantially higher than the true parameter value (a deviance equal to more than thirty percent of the true parameter). Thus, the results of Table 6 suggest that the CFFE and DCF can be used to accurately estimate the

var	κ	λ	mean	lb	$\mathbf{u}\mathbf{b}$
truth			4.24		
CFFE	0	0	4.21	4.20	4.23
CFFE	0	5	4.25	4.24	4.26
CFFE	5	0	4.22	4.20	4.23
CFFE	5	5	4.20	4.19	4.21
MRCF	0	0	5.60	5.54	5.64
MRCF	0	5	5.62	5.58	5.66
MRCF	5	0	5.58	5.54	5.64
MRCF	5	5	5.58	5.54	5.63
DCF	0	0	4.11	4.09	4.13
DCF	0	5	4.13	4.12	4.15
DCF	5	0	4.11	4.09	4.12
DCF	5	5	4.13	4.11	4.14

average treatment effect in the data, while also revealing the full range of heterogeneity in CATT (as shown by figures 7 and 9).

Table 6: ATE simulation 2

The table shows the averages of the true and estimated treatment effects by estimator for various levels of κ and λ . "estimator = cffe" indicates the average of the causal forest with fixed effects and "estimator = mrcf" indicates the average of a manually recentered causal forest. The table reports the average (fourth column) as well as the lower and upper bound of the 95 percent confidence interval (columns five and six) of the ATE. These bounds have been computed by simulating 500 draws of the estimated treatment effect from a normal distribution with mean $\hat{\tau}_i$ and variance $\hat{\sigma}_i$. Then the average treatment effect is computed for each draws and the lower and upper bound are the 2.5th and 97.5th percentile.

7 The heterogeneous effects of alternative work arrangements

Over the last two decades the share of the workforce engaged in 'flexible' alternative work arrangements has increased strongly in many OECD countries, including Italy, the Netherlands, the United Kingdom and the United States (Goos et al., 2022; Boeri et al., 2020; Katz and Krueger, 2019). People employed in alternative work arrangements work as temporary help agency workers, contract workers and independent contractor or freelancers. The growing use of alternative work arrangements has raised concerns about its effect on individual worker outcomes and about the need to change social protection to support these workers.

Unfortunately, the causal effect of alternative work arrangements on worker outcomes is not well understood. The main concern is that workers in alternative work arrangements could be negatively selected. That is to say, it is difficult to disentangle whether workers are paid less *because* they are employed in such arrangements, or lower-paid workers are simply more likely to be employed in these arrangements. Indeed, Drenik et al. (2023) show that workers at temp agencies have lower worker fixed effects, as measured by Abowd et al. (1999)-style models. Additionally, Katz and Krueger (2019) find that after controlling for observables, the wage penalty associated with temp agency work declines.

Moreover, most studies rely on surveys that typically do not capture well certain key aspects of alternative work arrangements, such as working many small jobs (Abraham et al., 2017; Abraham and Amaya, 2019; Katz and Krueger, 2019). Administrative matched employer-employee data sets do measure multiple small jobs of workers, but, often the employer-employee match is not well registered. In case of temporary work, the temporary work agency (the *de jure* employer) is registered as the employer, and not the firm where the work is done (the *de facto* employer). Suppose we observe that hourly wages decrease for workers who switch from a regular contract to an alternative work arrangement. Then we do not know whether this is due to the type of contract or whether this is due to a change in work environment and tasks.¹¹

Goos et al. (2022) (hereafter: GMSSB) study the case of payrolling in the Netherlands. Payrolling is a legal work arrangement whereby workers hired by one firm are placed on the payroll of another firm while continuing their job duties at the original firm. Using administrative matched employer-employee data, they compare individuals that are payrolled to individuals that work at firms that will payroll workers in the future using a staggered difference-in-differences research design. The authors find that, following a switch to payrolling, workers experience worse labor market outcomes, including lower hourly wage growth, a lower incidence of permanent contracts, lower employment probability, and lower pension contributions.

GMSSB report the conditional treatment effect for 29 subgroups. They conclude that the estimated overall impacts mask large heterogeneity between groups. The impact on hourly wages, for example, is negative for female workers, young (aged 18-24) and older workers (aged 46 - 60 years) and students. In contrast, for first generation migrants and higher educated workers the impacts are smaller and insignificant. In this

¹¹Some studies do observe the location of workers using alternative work contracts, see Drenik et al. (2023); Goldschmidt and Schmieder (2017).

paper we compare the heterogeneity analysis in GMSSB to a heterogeneity analysis based on a CFFE.

7.1 Data and methodology

We use the same estimation data as in GMSSB, with the exception of the estimation window and the number of matched control workers. Due to computational constraints, we set the maximum control workers for each treatment worker to 1 and limit the estimation window to 4 quarters before and 4 quarters after the payrolling event. In order to compare the CFFE outcomes with those in the original paper we re-estimate the original two-way fixed effect estimation on our restricted data set. After setting these restrictions, we have 39,880 treated workers and 39,865 matched control workers. We observe the same worker and job characteristics as in GMSSB's heterogeneity analysis: age, migration background, contract type, job tenure, enrollment in education and attained education level.

We apply a CFFE in which person, calendar time, and event time fixed effects are partialled out. The covariates that are provided to the CFFE to explore heterogeneity are: estimation quarter (from 2009Q1 to 2016Q1), gender, age at time of payrolling (18-24, 25-34, 35-44, 45-60), migration background (native, first generation migrant, second generation migrant), contract type before payrolling (temporary contract, permanent contract), job tenure before payrolling (0-1 years, 1-2 years, 2+ years), enrollment in education, and education level attained (low, middle, high).

We estimate the average treatment effect by estimating a specification that explains hourly wage using an indicator for being outsourced and indicators for individual, time and time event fixed effects.

7.2 Heterogeneity analysis using manually formed subgroups

First we repeat the main heterogeneity analysis by GMSSB for our data. Payrolling decreases the hourly wage in the first year of treatment with \in -0.17 / hour or nearly 1.6 percent.¹² This estimate is virtually similar to the effect reported by GMSSB in the first four quarters after payrolling occurred (see GMSSB Figure 9).

We conclude with GMSSB that the average treatment effect masks substantial heterogeneity between groups. We derive this conclusion after we compare the average treatment effect to the conditional average treatment effect for the 29 subgroups studied in GMSSB. Results are summarized in Figure 10. The figure reports that 23 out of 30 reported coefficients are significantly different from zero at the conventional significance level. The 95 percent confidence intervals of significant effects are often relatively wide. Most striking is the positive effect on wages of about $\in 0.3$ / hour for first generation migration workers, although here the confidence bound is also relatively large. This effect is economically relevant as about 12 percent of workers in the estimation sample has first generation background, see GMSSB.¹³ Overall, Figure 10 indicates substantial

¹²The standard error of the parameter is ≤ 0.043 / hour. The average hourly wage of payrolled workers is ≤ 10.835 euro, see GMSSB.

¹³The share of workers with a secondary migration background in the estimation sample is about 12 percent as well, see GMSSB.

variation by migration background, contract group, educational degree, being enrolled in education, estimation year and gender.

The effect of payrolling on hourly wages is always negative or insignificant when we correct for multiple hypothesis testing (Figure 10). Only 12 out of thirty estimated effects remain significant after we apply the Holm-Bonferroni correction. Importantly, the positive coefficient on first generation migration workers is not estimated precisely enough to reject the null hypothesis. The conclusion that this null hypothesis should be rejected is strengthened by the fact that GMSSB estimate a lower, but highly insignificant effect for this group as well. The figure suggest that the negative effect is driven by those workers that are young workers (aged 18 - 24), that are enrolled in education, that hold a middle or high educational degree, female workers, native Dutch workers and those workers with less than a year of tenure.



Figure 10: Heterogeneity in effects of payrolling on the hourly wage

Subgroup analysis for subgroups reported in GMSSB. The circles indicate the conditional average treatment effect of payrolling on hourly wages for subgroups mentioned on the y-axis. The horizontal lines indicate the 95 percent confidence bound. When subgroup parameters are indicated by a circle and a cross, the null hypothesis is rejected after adjusting for multiple hypothesis testing using the Holm-Bonferroni method.

7.3 Heterogeneity analysis using a causal forest

We now repeat the exercise above using the CFFE estimator. In line with the large confidence bounds shown in figure 10, we find that many conditional average treatment effects are positive. This conclusion is derived from Figure 11, which shows the distribution of the estimated conditional average treatment effect by our CFFE. The solid line indicates the sample mean of the conditional average treatment effect of payrolling on the hourly wage, which equals \in -0.20 / hour (1.8 percent of the average hourly wage of payrolled workers), which is similar to the average effect estimated using OLS. If there would not be heterogeneity in treatment effects, the distribution would be clustered around the sample mean. Although most of the mass is indeed around the sample average, the spread is substantial as indicated by the treatment effects of workers that

form the 2.5th and 97.5th percentiles. Payrolling results in a decrease in hourly wages of $\in 1.6 (15\%)$ for the first, but an increase in hourly wages by $\in 1.2 (11\%)$ for the latter.



Figure 11: Histogram of estimated conditional average treatment effects

Density plot for the conditional average treatment effect of payrolling on the hourly wage estimated by a causal forest with fixed effects. The solid vertical line indicates the sample mean of the conditional average treatment effect equal to -0.20. To improve readability, we only plot values larger than -2.5 and smaller than 2.5 omitting 1235 out of 318.980 observations from the plot. Fixed effects are computed for individual, time and event time. Variables used to look for heterogeneity are: gender, age group at time of payrolling, migration background, contract type before payrolling, job tenure before payrolling, enrollment in education, and education level attained. treatment effects are compute by every combination of quarter sinze payrolling occurred and quarter of the payrolling-event.

We find formal evidence suggesting heterogeneity in the effect of payrolling on hourly wages as shown by the Area Under the Targeting Operaror Characteristic (AUTOC) curve (Tibshirani et al., 2023; Yadlowsky et al., 2021). Here, individuals are ranked according to their average estimated treatment effect. Next, we take the q percent of individuals with the highest (in our case: the most positive) treatment effect and we compute the difference between the ATT for this subgroup and the overall ATT. Figure 12 shows this difference when q ranges from 0.1 to 1, when the difference is zero. The difference between the subgroup and overall average treatment effects steadily decreases in q and the difference remains significant even when q rises to 0.5, although the confidence interval almost contains zero. Formally, we test whether the area under the AUTOC is different from zero. This area, also defined as the Rank-weighted Average Treatment Effect (RATE) is 0.018. Bootstrapped confidence bounds based on 500 replications suggest that the RATE differs significantly from zero. This forms additional evidence that our CFFE detects treatment effect heterogeneity.



Figure 12: Subgroup analyses based on causal forest estimates (GATES)

Workers are ranked according to the estimated conditional average treatment effect by a causal forest with fixed effects. Then the difference between the overall ATT and the ATT for the subgroup with the q percent most affected individuals is estimated. The causal forest with fixed effect used to explain heterogeneity in the effect on hourly wage makes splits using estimation quarter, gender, age group at time of payrolling, migration background, contract type before payrolling, job tenure before payrolling, enrollment in education, and education level attained.

Next, we turn to the estimation of the ATT for subgroups and we conclude that the average treatment effect differs significantly from zero for the first two deciles only. To derive this conclusion we group workers into deciles according to the average of their estimated CATE from the CFFE and plot the estimated treatment effects. The treatment effect of payrolling is significant and about \in -0.4 per hour (3.7 percent) for the first two deciles only. This effect is substantially larger than the decrease in wages of at most \in 0.32 per hour (3 percent) that was found in the manual subgroup analysis (for highly educated workers and for workers with a temporary contract). The treatment effects for the other deciles are insignificant and substantially smaller, see figure 13.

Overall, the conclusion from this CFFE-based analysis differs substantially from the manual subgroup analysis. Even after correcting for multiple hypothesis testing, the latter suggests evidence for heterogeneity in many dimensions including gender, age, migration background and educational degree. In particular, the manual subgroup analysis suggests that many workers are negatively affected by payrolling. For instance, the negative effect for female workers of \in -0.27 per hour (2.5 percent) suggest that payrolling negatively effects worker outcomes for about half of the population. In contrast, the analysis from a CFFE indicates that negative treatment effects are concentrated among a specific subgroup of workers that forms about twenty percent of the population.

In agreement with the manual subgroup analysis, this specific subgroup is typically composed of female, young, educationally enrolled, or first-generation migrant workers. This last conclusion is derived from figure 14, that considers the characteristics of workers by decile of treatment effect. Surprisingly, workers with low education are underrepresented in the first decile, just as workers with a temporary contract.



Figure 13: dynamic ATE for the first four deciles

Dynamic ATE estimation results when a causal forest with fixed effects is used to estimate conditional average treatment effects in a staggered difference-in-differences research design. Next workers have been grouped into deciles according to their estimated treatment effect.



Figure 14: Classification analysis (CLAN)

Workers are grouped into deciles according to the estimated conditional average treatment effect predicted by a causal forest. Cells show the relative difference in worker characteristics by decile against the average worker. Thus in decile ten, the share of second generation migration workers is more than 100 percent higher than the average share of second generation migration workers (second row, final column).

8 Conclusion

Scholars studying the difference-in-differences methodology have extensively examined the estimation of the ATT, leading to the development of new estimators for the ATT when treatment effects are dynamic or vary with treatment event. This paper takes a different approach by concentrating on the CATT, and investigates the potential of using causal forests to determine how treatment effects vary in relation to covariate variables, an area which has been identified as a promising area for future research by Roth et al. (2023).

We present the causal forest with fixed effects (CFFE), a modification of the original causal forest algorithm by Tibshirani et al. (2022) that allows to estimate a causal forest conditional on a large number of fixed effects. Our modification uses averages at the individual level and over time to partial out individual and time fixed effects. We show our modification can be applied to estimate CATT in a (staggered) difference-in-differences setting. As such, this paper is closely related to the dynamic causal forest (DCF) method developed by Gavrilova et al. (2023), where the data is transformed to partial out individual and time fixed effects. We use simulations to compare the performance of our CFFE to their DCF and a naive manually recentered causal forest (MRCF).

We find that a CFFE provides consistent estimates of the true heterogeneity in treatment effects. This conclusion follows from the simulations for a single event and staggered difference-in-differences research design. Also the average of the estimated treatment effects in the data is nearly similar to true average of treatment effects. This suggests that our modification is capable of accurately estimating the conditional and average treatment effects, even when treatment effects are dynamic and when they vary depending on the characteristics of the treated individuals. Comparing our CFFE-estimator to the DCF-estimator, we find that both provide unbiased and consistent results. However, the CFFE-estimator is more efficient, both in the single event as in the staggered difference-in-differences simulation. This is mostly due to the fact that it uses more observations from the periods before treatment, which allows it to better distinguish signal from noise in the data, and because the common trend assumption holds in all periods in our simulation (De Chaisemartin and d'Haultfoeuille, 2023; Roth et al., 2023).

We use a CFFE to describe the heterogeneous effects of alternative work arrangements on worker outcomes. When subgroups are formed manually, we document substantial heterogeneity in the effect of payrolling on hourly wages in the short run. Many of the investigated subgroups show a significant decrease in hourly wages after payrolling, which suggests that payrolling has negative consequences for workers accros the board. Yet estimates of these effects are relatively inefficient as reflected by the relatively wide confidence bounds. In particular, evidence that workers with a first generation background is significant, but not robust to the Holm-Bonferroni adjustment to control for multiple hypothesis testing. As this conclusion was not found in previous work with similar data (Goos et al., 2022), this illustrates the risk that some significant results in manual subgroup analyses might be spurious.

In contrast, a CFFE allows us to group workers into deciles according to their conditional average treatment effect. Then we document that most of the treatment effects are concentrated around the sample average of $\in -0.2$ per hour (a decrease of

1.8 percent), although treatment effects can be substantially larger in the tails. We group workers according to the estimated treatment effect and we find evidence that the negative effects of payrolling are concentrated among workers in the first two deciles only. The treatment effects for these deciles is about \in -0.4 per hour (3.7 percent), which is substantially lower than results appearing from the manual subgroup analysis. For the remaining eight deciles, we do not find evidence that payrolling changes hourly wage on average. This reveals a causal forest can document heterogeneity in conditional average treatment effects that was not picked up by heterogeneity analysis using manually formed subgroups.

Finally, we see several developments that could improve the use of causal forest in future research. We did not develop a doubly robust estimator for the average treatment effect, although we acknowledge that such an estimator could improve the use of causal forests (CFFE or DCF) in future empirical studies. We demonstrate that our algorithm is successful in detecting heterogeneity in treatment effects, but we remain largely unaware of the individual characteristics that explain them. Future research, could enhance the interpretability of the heterogeneous treatment effects that our algorithm provides, for instance using *fit-of-fit* approaches (Bargagli-Stoffi et al., 2020) or explainable machine learning methods (Molnar, 2020). Finally, we note that further research could study the robustness of our approach to estimate the CATT, for instance when treatment anticipation effects are present or when the common trend assumption is mildly violated due to group specific time trends.

References

- Abadie, A. (2005). Semiparametric difference-in-differences estimators. The Review of Economic Studies 72(1), 1–19.
- Abowd, J. M., F. Kramarz, and D. N. Margolis (1999). High wage workers and high wage firms. *Econometrica* 67(2), 251–333.
- Abraham, K., J. Haltiwanger, K. Sandusky, and J. Spletzer (2017). Measuring and accounting for innovation in the 21st century. In C. Corrado, J. Haskel, J. Miranda, and D. Sichel (Eds.), *Measuring the gig economy: Current knowledge and open issues*. University of Chicago Press.
- Abraham, K. G. and A. Amaya (2019). Probing for informal work activity. Journal of Official Statistics 35(3), 487–508.
- Angrist, J. D. and J.-S. Pischke (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.
- Angrist, J. D. and J.-S. Pischke (2010). The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of economic perspectives* 24(2), 3–30.
- Arkhangelsky, D., S. Athey, D. A. Hirshberg, G. W. Imbens, and S. Wager (2021). Synthetic difference-in-differences. American Economic Review 111(12), 4088–4118.

- Assmann, S. F., S. J. Pocock, L. E. Enos, and L. E. Kasten (2000). Subgroup analysis and other (mis) uses of baseline data in clinical trials. *The Lancet* 355(9209), 1064– 1069.
- Athey, S. and G. Imbens (2016). Recursive partitioning for heterogeneous causal effects. Proceedings of the National Academy of Sciences 113(27), 7353–7360.
- Athey, S. and G. Imbens (2019). Machine learning methods economists should know about. *Annual Review of Economics* 11, 685–725.
- Athey, S. and G. W. Imbens (2006). Identification and inference in nonlinear differencein-differences models. *Econometrica* 74(2), 431–497.
- Athey, S., J. Tibshirani, and S. Wager (2019). Generalized random forests. The Annals of Statistics 47(2), 1148–1178.
- Athey, S. and S. Wager (2019). Estimating treatment effects with causal forests: An application. *Observational Studies* 5(2), 37–51.
- Baker, A. C., D. F. Larcker, and C. C. Wang (2022). How much should we trust staggered difference-in-differences estimates? *Journal of Financial Economics* 144(2), 370–395.
- Bargagli-Stoffi, F. J., R. Cadei, K. Lee, and F. Dominici (2020). Causal rule ensemble: Interpretable discovery and inference of heterogeneous causal effects. arXiv preprint arXiv:2009.09036.
- Bargagli-Stoffi, F. J., K. De Witte, and G. Gnecco (2022). Heterogeneous causal effects with imperfect compliance: A bayesian machine learning approach. *The Annals of Applied Statistics* 16(3), 1986–2009.
- Bargagli Stoffi, F. J. and G. Gnecco (2020). Causal tree with instrumental variable: an extension of the causal tree framework to irregular assignment mechanisms. *Interna*tional Journal of Data Science and Analytics 9(3), 315–337.
- Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society:* series B (Methodological) 57(1), 289–300.
- Bertrand, M., E. Duflo, and S. Mullainathan (2004). How much should we trust differences-in-differences estimates? The Quarterly journal of economics 119(1), 249–275.
- Bodory, H., H. Busshoff, and M. Lechner (2022). High resolution treatment effects estimation: Uncovering effect heterogeneities with the modified causal forest. *Entropy* 24(8), 1039.
- Boeri, T., G. Giupponi, A. B. Krueger, and S. Machin (2020). Solo self-employment and alternative work arrangements: A cross-country perspective on the changing composition of jobs. *Journal of Economic Perspectives* 34(1), 170–195.

- Borusyak, K. (2023). Did_imputation: Stata module to perform treatment effect estimation and pre-trend testing in event studies.
- Borusyak, K., X. Jaravel, and J. Spiess (2021). Revisiting event study designs: Robust and efficient estimation. arXiv preprint arXiv:2108.12419.
- Breiman, L. (2001). Random forests. Machine learning 45(1), 5–32.
- Breiman, L., J. Friedman, R. Olshen, and C. Stone (1984). *Classification and Regression Trees.* Belmont, CA: Wadsworth.
- Brock, J. M. and R. De Haas (2023). Discriminatory lending: Evidence from bankers in the lab. *American Economic Journal: Applied Economics* 15(2), 31–68.
- Callaway, B. and P. H. Sant'Anna (2021). Difference-in-differences with multiple time periods. *Journal of econometrics* 225(2), 200–230.
- Chang, N.-C. (2020). Double/debiased machine learning for difference-in-differences models. *The Econometrics Journal* 23(2), 177–191.
- Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, and W. Newey (2017). Double/debiased/neyman machine learning of treatment effects. *American Economic Review* 107(5), 261–65.
- Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018). Double/debiased machine learning for treatment and structural parameters.
- Chernozhukov, V., M. Demirer, E. Duflo, and I. Fernandez-Val (2018). Generic machine learning inference on heterogeneous treatment effects in randomized experiments, with an application to immunization in india. Technical report, National Bureau of Economic Research.
- Chipman, H. A., E. I. George, and R. E. McCulloch (2010). Bart: Bayesian additive regression trees. *The Annals of Applied Statistics* 4(1), 266–298.
- Cook, D. I., V. J. Gebski, and A. C. Keech (2004). Subgroup analysis in clinical trials. Medical Journal of Australia 180(6), 289.
- Currie, J., H. Kleven, and E. Zwiers (2020). Technology and big data are changing economics: Mining text to track methods. *AEA Papers and Proceedings* 110, 42–48.
- Davis, J. M. and S. B. Heller (2020). Rethinking the benefits of youth employment programs: The heterogeneous effects of summer jobs. *Review of Economics and Statistics* 102(4), 664–677.
- De Chaisemartin, C. and X. d'Haultfoeuille (2020). Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review* 110(9), 2964–96.
- De Chaisemartin, C. and X. d'Haultfoeuille (2023). Two-way fixed effects and differences-in-differences with heterogeneous treatment effects: A survey. *The Econometrics Journal* 26(3), C1–C30.

- Donald, S. G. and K. Lang (2007). Inference with difference-in-differences and other panel data. *The review of Economics and Statistics* 89(2), 221–233.
- Doudchenko, N. and G. W. Imbens (2016). Balancing, regression, difference-indifferences and synthetic control methods: A synthesis. Technical report, National Bureau of Economic Research.
- Drenik, A., S. Jäger, P. Plotkin, and B. Schoefer (2023). Paying outsourced labor: Direct evidence from linked temp agency-worker-client data. *Review of Economics* and Statistics 105(1), 206–216.
- Dube, A., D. Girardi, O. Jorda, and A. Taylor (2023). A local projections approach to difference-in-differences event studies. Technical report, National Bureau of Economic Research.
- Foster, J. C., J. M. Taylor, and S. J. Ruberg (2011). Subgroup identification from randomized clinical trial data. *Statistics in medicine* 30(24), 2867–2880.
- Frisch, R. and F. V. Waugh (1933). Partial time regressions as compared with individual trends. *Econometrica: Journal of the Econometric Society* 1(4), 387–401.
- Gavrilova, E., A. Langørgen, and F. Zoutman (2023). Dynamic causal forests, with an application to payroll tax incidence in norway.
- Goldschmidt, D. and J. F. Schmieder (2017). The rise of domestic outsourcing and the evolution of the german wage structure. *The Quarterly Journal of Economics* 132(3), 1165–1217.
- Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. Journal of Econometrics 225(2), 254–277.
- Goos, M., A. Manning, A. Salomons, B. Scheer, and W. van den Berge (2022). Alternative work arrangements and worker outcomes: Evidence from payrolling. Technical report, CPB Netherlands Bureau for Economic Policy Analysis.
- Green, D. P. and H. L. Kern (2012). Modeling heterogeneous treatment effects in survey experiments with bayesian additive regression trees. *Public opinion quarterly* 76(3), 491–511.
- Gulen, H., C. Jens, and T. B. Page (2020). An application of causal forest in corporate finance: How does financing affect investment? Technical report, Texas A&M University.
- Hartford, J., G. Lewis, K. Leyton-Brown, and M. Taddy (2016). Counterfactual prediction with deep instrumental variables networks. *arXiv preprint arXiv:1612.09596*.
- Hastie, T., R. Tibshirani, J. H. Friedman, and J. H. Friedman (2009). *The elements of statistical learning: data mining, inference, and prediction*, Volume 2. Springer.
- Hatamyar, J., N. Kreif, R. Rocha, and M. Huber (2023). Machine learning for staggered difference-in-differences and dynamic treatment effect heterogeneity. arXiv preprint arXiv:2310.11962.

- Hermansson, E. and D. Svensson (2021). On discovering treatment-effect modifiers using virtual twins and causal forest ml in the presence of prognostic biomarkers. In *International Conference on Computational Science and Its Applications*, pp. 624– 640. Springer.
- Hill, J. and Y.-S. Su (2013). Assessing lack of common support in causal inference using bayesian nonparametrics: Implications for evaluating the effect of breastfeeding on children's cognitive outcomes. *The Annals of Applied Statistics* 7(3), 1386–1420.
- Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. Journal of Computational and Graphical Statistics 20(1), 217–240.
- Hoffman, I. and E. Mast (2019). Heterogeneity in the effect of federal spending on local crime: Evidence from causal forests. *Regional Science and Urban Economics* 78, 103463.
- Holm, S. (1979). A simple sequentially rejective multiple test procedure. Scandinavian journal of statistics 6(2), 65–70.
- Imai, K. and M. Ratkovic (2013). Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics* 7(1), 443–470.
- Imbens, G. W. and D. B. Rubin (2015). *Causal inference in statistics, social, and biomedical sciences.* Cambridge University Press.
- Jawadekar, N., K. Kezios, M. C. Odden, J. A. Stingone, S. Calonico, K. Rudolph, and A. Z. Al Hazzouri (2023). Practical guide to honest causal forests for identifying heterogeneous treatment effects. *American Journal of Epidemiology*, kwad043.
- Jens, C., T. B. Page, and J. C. Reeder III (2021). Controlling for group-level heterogeneity in causal forest.
- Johnson, M., J. Cao, and H. Kang (2022). Detecting heterogeneous treatment effects with instrumental variables and application to the oregon health insurance experiment. *The Annals of Applied Statistics* 16(2), 1111–1129.
- Katz, L. F. and A. B. Krueger (2019). The rise and nature of alternative work arrangements in the united states, 1995–2015. *ILR review* 72(2), 382–416.
- Knaus, M. C. (2022). Double machine learning-based programme evaluation under unconfoundedness. The Econometrics Journal 25(3), 602–627.
- Knittel, C. R. and S. Stolper (2021). Machine learning about treatment effect heterogeneity: The case of household energy use. In AEA Papers and Proceedings, Volume 111, pp. 440–44.
- Lovell, M. C. (1963). Seasonal adjustment of economic time series and multiple regression analysis. *Journal of the American Statistical Association* 58(304), 993–1010.

- Luo, X., X. Lu, and J. Li (2019). When and how to leverage e-commerce cart targeting: the relative and moderated effects of scarcity and price incentives with a two-stage field experiment and causal forest optimization. *Information Systems Research* 30(4), 1203–1227.
- Miller, S. (2020). Causal forest estimation of heterogeneous and time-varying environmental policy effects. Journal of Environmental Economics and Management 103, 102337.
- Molnar, C. (2020). Interpretable machine learning. Lulu. com.
- Munroe, R. (2022). Significant. https://xkcd.com/882/, accessed = 2022-11-09.
- Murakami, K., H. Shimada, Y. Ushifusa, and T. Ida (2022). Heterogeneous treatment effects of nudge and rebate: Causal machine learning in a field experiment on electricity conservation. *International Economic Review* 63(4), 1779–1803.
- Raghavan, S., K. Josey, G. Bahn, D. Reda, S. Basu, S. A. Berkowitz, N. Emanuele, P. Reaven, and D. Ghosh (2022). Generalizability of heterogeneous treatment effects based on causal forests applied to two randomized clinical trials of intensive glycemic control. Annals of Epidemiology 65, 101–108.
- Roth, J. (2022). Pretest with caution: Event-study estimates after testing for parallel trends. *American Economic Review: Insights* 4(3), 305–322.
- Roth, J., P. H. Sant'Anna, A. Bilinski, and J. Poe (2023). What's trending in differencein-differences? a synthesis of the recent econometrics literature. *Journal of Econometrics*.
- Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. Journal of the American Statistical Association 100(469), 322–331.
- Shah, A. M., M. Osborne, J. Lefkowitz Kalter, A. Fertig, A. Fishbane, and D. Soman (2023). Identifying heterogeneity using recursive partitioning: Evidence from sms nudges encouraging voluntary retirement savings in mexico. *PNAS Nexus*, pgad058.
- Shiba, K., A. Daoud, H. Hikichi, A. Yazawa, J. Aida, K. Kondo, and I. Kawachi (2021). Heterogeneity in cognitive disability after a major disaster: A natural experiment study. *Science advances* 7(40), eabj2610.
- Somaini, P. and F. A. Wolak (2016). An algorithm to estimate the two-way fixed effects model. *Journal of Econometric Methods* 5(1), 143–152.
- Su, X., C.-L. Tsai, H. Wang, D. M. Nickerson, and B. Li (2009). Subgroup analysis via recursive partitioning. *Journal of Machine Learning Research* 10(2), 141–158.
- Sun, L. and S. Abraham (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics* 225(2), 175–199.
- Tian, L., A. A. Alizadeh, A. J. Gentles, and R. Tibshirani (2014). A simple method for estimating interactions between a treatment and a large number of covariates. *Journal of the American Statistical Association 109*(508), 1517–1532.

- Tibshirani, J., S. Athey, E. Sverdrup, and S. Wager (2022). The grf algorithm. https://grf-labs.github.io/grf/REFERENCE.html, Accessed 2022-11-09.
- Tibshirani, J., S. Athey, E. Sverdrup, and S. Wager (2023). Assessing heterogeneity with rate. https://grf-labs.github.io/grf/articles/rate.html, Accessed 2023-11-14.
- Verstraete, K., I. Gyselinck, H. Huts, N. Das, M. Topalovic, M. De Vos, and W. Janssens (2023). Estimating individual treatment effects on copd exacerbations by causal machine learning on randomised controlled trials. *thorax*, 1–7.
- Wager, S. and S. Athey (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association* 113(523), 1228–1242.
- Wang, T., C. Rudin, F. Doshi-Velez, Y. Liu, E. Klampfl, and P. MacNeille (2017). A bayesian framework for learning rule sets for interpretable classification. *The Journal* of Machine Learning Research 18(1), 2357–2393.
- Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. MIT press.
- Yadlowsky, S., S. Fleming, N. Shah, E. Brunskill, and S. Wager (2021). Evaluating treatment prioritization rules via rank-weighted average treatment effects. arXiv preprint arXiv:2111.07966.
- Zeileis, A., T. Hothorn, and K. Hornik (2008). Model-based recursive partitioning. Journal of Computational and Graphical Statistics 17(2), 492–514.
- Zheng, L. and W. Yin (2023). Estimating and evaluating treatment effect heterogeneity: A causal forests approach. Research & Politics 10(1), 20531680231153080.

A All simulation results

In the main text we have visualized estimation results when $\kappa = 5$ and $\lambda = 5$. Here we report all results, thus those when $\kappa = 0, 5$ and $\lambda = 0, 5$ and results when $\kappa = 5$ and $\lambda = 0$. Results are nearly identical to those presented in the main test, which illustrates a CFFE effectively estimates the CATT when individual and/or time fixed effects are not present in the data.



A.1 Causal forest with fixed effects: simulation 1

Figure 15: causal forest with fixed effects

Estimation results when a causal forest with fixed effects is used to estimate conditional average treatment effects in a single event difference-in-differences research design. Panel (a) shows the bias of the estimated treatment effect $(\hat{\tau} - \tau)$ when $\kappa = 0$ and $\lambda = 0$. Panel (b) shows the estimated values of the treatment effect $(\hat{\tau})$ and the true treatment effect (τ) as a function of the forcing variable, x_1 , when $\kappa = 0$ and $\lambda = 0$. Dots indicate the estimated treatment effects. The true treatment effect is indicated by the solid line. Panels (c) to (f) show these plots for different values of κ and λ .



A.2 Causal forest with fixed effects: simulation 2

Figure 16: causal forest with fixed effects

Estimation results when a causal forest with fixed effects is used to estimate conditional average treatment effects in a single event difference-in-differences research design. Panel (a) shows the bias of the estimated treatment effect $(\hat{\tau} - \tau)$ when $\kappa = 0$ and $\lambda = 0$. Panel (b) does this for event 2. Panel (c) shows the estimated values of the treatment effect $(\hat{\tau})$ and the true treatment effect (τ) as a function of the forcing variable, x_1 times time since the event, when $\kappa = 0$ and $\lambda = 0$ for event 1. Panel (d) does this for event 2. The other panels shows these plots when $\kappa = 0$ and $\lambda = 5$ (panels (e) to (h)) or when $\kappa = 5$ and $\lambda = 0$ (panels (i) to (l)).

A.3 Manually recentered causal forest: simulation 1



Figure 17: manually recentered causal forest

Estimation results when a manually recentered causal forest is used to estimate conditional average treatment effects in a single event difference-in-differences research design. Panel (a) shows the bias of the estimated treatment effect $(\hat{\tau} - \tau)$ when $\kappa = 0$ and $\lambda = 0$. Panel (b) shows the estimated values of the treatment effect $(\hat{\tau})$ and the true treatment effect (τ) as a function of the forcing variable, x_1 , when $\kappa = 0$ and $\lambda = 0$. Dots indicate the estimated treatment effects. The true treatment effect is indicated by the solid line. Panels (c) to (f) show these plots for different values of κ and λ .

A.4 Manually recentered causal forest: simulation 2



Figure 18: manually recentered causal forest

Estimation results when a manually recentered causal forest is used to estimate conditional average treatment effects in a single event difference-in-differences research design. Panel (a) shows the bias of the estimated treatment effect $(\hat{\tau} - \tau)$ when $\kappa = 0$ and $\lambda = 0$. Panel (b) does this for event 2. Panel (c) shows the estimated values of the treatment effect $(\hat{\tau})$ and the true treatment effect (τ) as a function of the forcing variable, x_1 times time since the event, when $\kappa = 0$ and $\lambda = 0$ for event 1. Panel (d) does this for event 2. The other panels shows these plots when $\kappa = 0$ and $\lambda = 5$ (panels (e) to (h)) or when $\kappa = 5$ and $\lambda = 0$ (panels (i) to (l)).



A.5 Dynamic causal forest: simulation 1

(e) bias

(f) estimated values

Figure 19: dynamic causal forest

Estimation results when a dynamic causal forest is used to estimate conditional average treatment effects in a single event difference-in-differences research design. Panel (a) shows the bias of the estimated treatment effect $(\hat{\tau} - \tau)$ when $\kappa = 0$ and $\lambda = 0$. Panel (b) shows the estimated values of the treatment effect $(\hat{\tau})$ and the true treatment effect (τ) as a function of the forcing variable, x_1 , when $\kappa = 0$ and $\lambda = 0$. Dots indicate the estimated treatment effects. The true treatment effect is indicated by the solid line. Panels (c) to (f) show these plots for different values of κ and λ .

A.6 Dynamic causal forest: simulation 2



Figure 20: dynamic causal forest

Estimation results when a manually recentered causal forest is used to estimate conditional average treatment effects in a single event difference-in-differences research design. Panel (a) shows the bias of the estimated treatment effect $(\hat{\tau} - \tau)$ when $\kappa = 0$ and $\lambda = 0$. Panel (b) does this for event 2. Panel (c) shows the estimated values of the treatment effect $(\hat{\tau})$ and the true treatment effect (τ) as a function of the forcing variable, x_1 times time since the event, when $\kappa = 0$ and $\lambda = 0$ for event 1. Panel (d) does this for event 2. The other panels shows these plots when $\kappa = 0$ and $\lambda = 5$ (panels (e) to (h)) or when $\kappa = 5$ and $\lambda = 0$ (panels (i) to (l)).

B Simulation results using alternatively distributed data

The simulated data in the paper follows a uniform distribution. Here we show how results change when the data is simulated using a normal distribution. Overall, results are quite similar regardless whether the data is simulated using the uniform or normal distribution.

B.1 Tables simulation 1

Tables 7 to 9 show the bias, RMSE and the ATE of the estimators. In particular, the bias of a CFFE and DCF are similar, and different from that of a MRCF (Table 7). Also, we conclude that the estimates provided by a CFFE are more precise than those of a MRCF and DCF (Table 8). Also, the CFFE correctly estimates the average treatment effect (Table 9).

estimator	κ	λ	bias	cfi
CFFE	0	0	-0.05	(-0.0730.0327)
CFFE	5	0	0.01	(-0.0135 - 0.0328)
CFFE	0	5	-0.05	(-0.07220.0229)
CFFE	5	5	-0.06	(-0.07890.0341)
MRCF	0	0	-1.44	(-1.51061.3758)
MRCF	5	0	-1.39	(-1.46721.3265)
MRCF	0	5	-1.43	(-1.50021.3605)
MRCF	5	5	-1.49	(-1.56741.4295)
DCF	0	0	-0.04	(-0.07850.0076)
DCF	5	0	0.03	(-0.0038 - 0.0743)
DCF	0	5	-0.04	(-0.0738 - 0.0049)
DCF	5	5	0.02	(-0.0157 - 0.0593)
CFFE 2P	0	0	-0.11	(-0.1250.0854)
CFFE 2P	5	0	-0.01	(-0.0337 - 0.0249)
CFFE 2P	0	5	-0.12	(-0.15230.0928)
CFFE 2P	5	5	-0.07	(-0.09380.051)
DCF 2P	0	0	-0.06	(-0.08940.0231)
DCF $2P$	5	0	0.03	(-0.0114 - 0.0677)
DCF $2P$	0	5	-0.09	(-0.12150.0518)
DCF 2P	5	5	0.00	(-0.0308 - 0.0373)

Table 7: bias simulation 1

The table shows the average of the bias when a causal forest with fixed effects (cffe) is used to estimate heterogeneous treatment effects and when this is done using a manually recentered causal forest (mrcf) or a dynamic causal forest (dcf). We also present results when a CFFE and DCF are trained on periods 2 and 3 only (cffe 2p and dcf 2p). Each row presents the average of the bias for various levels of the individual effects vary ($\kappa = 0, 5$) and time fixed effects ($\lambda = 0, 5$). The reported value is computed as $\frac{1}{N} \sum_{i}^{N} (\hat{\tau}_{i} - \tau_{i})$, where τ_{i} indicates the true treatment effect for individual i and $\hat{\tau}_{i}$ indicates the estimated treatment effect for individual i. The numbers within brackets denote the bootstrapped 95 percent confidence interval based on 500 replications.

estimator	κ	λ	rmse	cfi
CFFE	0	0	0.59	(0.3907 - 0.8036)
CFFE	5	0	0.63	(0.4395 - 0.8441)
CFFE	0	5	0.63	(0.4554 - 0.8283)
CFFE	5	5	0.59	(0.4007 - 0.8147)
MRCF	0	0	2.39	(2.2967 - 2.4893)
MRCF	5	0	2.37	(2.283 - 2.4783)
MRCF	0	5	2.40	(2.3163 - 2.5062)
MRCF	5	5	2.44	(2.3419 - 2.5394)
DCF	0	0	1.00	(0.7933 - 1.2194)
DCF	5	0	1.08	(0.8723 - 1.2959)
DCF	0	5	1.08	(0.8649 - 1.2967)
DCF	5	5	1.02	(0.8162 - 1.2387)
CFFE 2P	0	0	0.43	(0.3 - 0.5641)
CFFE 2P	5	0	0.59	(0.4494 - 0.7373)
CFFE 2P	0	5	0.60	(0.4867 - 0.7419)
CFFE 2P	5	5	0.43	(0.2892 - 0.5702)
DCF 2P	0	0	0.68	(0.5276 - 0.8352)
DCF 2P	5	0	0.81	(0.6706 - 0.9642)
DCF $2P$	0	5	0.75	(0.6082 - 0.9091)
DCF $2P$	5	5	0.69	(0.5405 - 0.8433)

Table 8: RMSE simulation 1

The table shows the RMSE when a causal forest with fixed effects (cffe) is used to estimate heterogeneous treatment effects and when this is done using a manually recentered causal forest (mrcf) or a dynamic causal forest (dcf). We also present results when a CFFE and DCF are trained on periods 2 and 3 only (cffe 2p and dcf 2p). Each row presents the RMSE for various levels of the individual effects vary ($\kappa = 0, 5$) and time fixed effects ($\lambda = 0, 5$). The RMSE is computed as $\left(\frac{1}{N}\sum_{i}^{N}(\tau_{i} - \hat{\tau}_{i})^{2}\right)^{0.5}$, where τ_{i} indicates the true treatment effect for individual i and $\hat{\tau}_{i}$ indicates the estimated treatment effect for individual i. The numbers within brackets denote the bootstrapped 95 percent confidence interval based on 500 replications.

var	κ	λ	mean	lb	ub
truth			2.97		
CFFE	0	0	3.03	3.01	3.04
CFFE	0	5	3.02	3.01	3.04
CFFE	5	0	2.96	2.95	2.98
CFFE	5	5	3.03	3.02	3.04
MRCF	0	0	4.41	4.39	4.44
MRCF	0	5	4.40	4.37	4.42
MRCF	5	0	4.37	4.34	4.39
MRCF	5	5	4.47	4.44	4.49
DCF	0	0	3.02	3.00	3.03
DCF	0	5	3.01	2.99	3.02
DCF	5	0	2.94	2.92	2.95
DCF	5	5	2.95	2.94	2.97

Table 9: ATE simulation 1

The table shows the averages of the true and estimated treatment effects by estimator for various levels of κ and λ . "estimator = cffe" indicates the average of the causal forest with fixed effects and "estimator = mrcf" indicates the average of a manually recentered causal forest. The table reports the average (fourth column) as well as the lower and upper bound of the 95 percent confidence interval (columns five and six) of the ATE. These bounds have been computed by simulating 500 draws of the estimated treatment effect from a normal distribution with mean $\hat{\tau}_i$ and variance $\hat{\sigma}_i$. Then the average treatment effect is computed for each draws and the lower and upper bound are the 2.5th and 97.5th percentile.

B.2 Tables simulation 2

Tables 10 to 12 show the bias, RMSE and the ATE of the estimators. In particular, the bias of a CFFE and DCF are similar, and different from that of a MRCF (Table 10). Also, we conclude that the estimates provided by a CFFE are more precise than those of a MRCF and DCF (Table 11). Also, the CFFE correctly estimates the average treatment effect (Table 12).

estimator	κ	λ	bias	cfi
CFFE	0	0	-0.13	(-0.18560.0572)
CFFE	5	0	-0.11	(-0.17290.0389)
CFFE	0	5	-0.15	(-0.20740.0732)
CFFE	5	5	-0.02	(-0.0753 - 0.0528)
MRCF	0	0	-2.00	(-2.12361.8852)
MRCF	5	0	-2.00	(-2.11641.8861)
MRCF	0	5	-2.03	(-2.15281.9231)
MRCF	5	5	-1.89	(-2.01571.7861)
DCF	0	0	-0.07	(-0.1653 - 0.0397)
DCF	5	0	-0.08	(-0.1765 - 0.0331)
DCF	0	5	-0.10	(-0.191 - 0.0153)
DCF	5	5	-0.07	(-0.1627 - 0.034)
CFFE 2P	0	0	-0.14	(-0.19820.0687)
CFFE 2P	5	0	-0.07	(-0.149 - 0.0096)
CFFE 2P	0	5	-0.15	(-0.20830.0726)
CFFE 2P	5	5	-0.08	(-0.14580.0182)
DCF 2P	0	0	-0.10	(-0.19930.0108)
DCF 2P	5	0	-0.03	(-0.1343 - 0.0777)
DCF 2P	0	5	-0.09	(-0.1899 - 0.0053)
DCF 2P	5	5	-0.08	(-0.1754 - 0.0151)

Table 10: Bias simulation 2

The table shows the RMSE when a causal forest with fixed effects (cffe) is used to estimate heterogeneous treatment effects and when this is done using a manually recentered causal forest (mrcf) or a dynamic causal forest (dcf). We also present results when a CFFE and DCF are trained on periods 2 and 3 only (cffe 2p and dcf 2p). Each row presents the RMSE for various levels of the individual effects vary ($\kappa = 0, 5$) and time fixed effects ($\lambda = 0, 5$). The RMSE is computed as $\left(\frac{1}{N}\sum_{i}^{N}(\tau_{i} - \hat{\tau}_{i})^{2}\right)^{0.5}$, where τ_{i} indicates the true treatment effect for individual i and $\hat{\tau}_{i}$ indicates the estimated treatment effect for individual i. The numbers within brackets denote the bootstrapped 95 percent confidence interval based on 500 replications.

estimator	κ	λ	rmse	cfi
CFFE	0	0	1.83	(1.4342 - 2.3108)
CFFE	5	0	1.89	(1.4725 - 2.3909)
CFFE	0	5	1.88	(1.4682 - 2.3773)
CFFE	5	5	1.85	(1.4476 - 2.3469)
MRCF	0	0	4.04	(3.8503 - 4.2527)
MRCF	5	0	3.91	(3.7302 - 4.1086)
MRCF	0	5	3.95	(3.7746 - 4.1505)
MRCF	5	5	3.87	(3.6829 - 4.0825)
DCF	0	0	2.84	(2.3843 - 3.421)
DCF	5	0	2.95	(2.4866 - 3.5323)
DCF	0	5	2.87	(2.4092 - 3.4659)
DCF	5	5	2.81	(2.3548 - 3.3943)
CFFE 2P	0	0	1.30	(0.9113 - 1.6589)
CFFE 2P	5	0	1.45	(1.044 - 1.8402)
CFFE 2P	0	5	1.33	(0.9331 - 1.6967)
CFFE 2P	5	5	1.23	(0.8501 - 1.5738)
DCF 2P	0	0	1.79	(1.3646 - 2.2024)
DCF 2P	5	0	1.93	(1.4824 - 2.3514)
DCF 2P	0	5	1.84	(1.3973 - 2.2573)
DCF 2P	5	5	1.77	(1.3493 - 2.1693)

Table 11: RMSE simulation 2

The table shows the RMSE when a causal forest with fixed effects (cffe) is used to estimate heterogeneous treatment effects and when this is done using a manually recentered causal forest (mrcf) or a dynamic causal forest (dcf). We also present results when a CFFE and DCF are trained on periods 2 and 3 only (cffe 2p and dcf 2p). Each row presents the RMSE for various levels of the individual effects vary ($\kappa = 0, 5$) and time fixed effects ($\lambda = 0, 5$). The RMSE is computed as $\left(\frac{1}{N}\sum_{i}^{N}(\tau_{i} - \hat{\tau}_{i})^{2}\right)^{0.5}$, where τ_{i} indicates the true treatment effect for individual i and $\hat{\tau}_{i}$ indicates the estimated treatment effect for individual i. The numbers within brackets denote the bootstrapped 95 percent confidence interval based on 500 replications.

var	κ	λ	mean	lb	ub
truth			5.47		
CFFE	0	0	5.60	5.57	5.62
CFFE	0	5	5.62	5.59	5.65
CFFE	5	0	5.58	5.55	5.61
CFFE	5	5	5.49	5.46	5.51
MRCF	0	0	7.46	7.38	7.53
MRCF	0	5	7.51	7.43	7.59
MRCF	5	0	7.46	7.40	7.54
MRCF	5	5	7.36	7.29	7.43
DCF	0	0	5.54	5.51	5.57
DCF	0	5	5.57	5.53	5.60
DCF	5	0	5.55	5.52	5.58
DCF	5	5	5.54	5.50	5.57

Table 12: ATE simulation 2

The table shows the averages of the true and estimated treatment effects by estimator for various levels of κ and λ . "estimator = cffe" indicates the average of the causal forest with fixed effects and "estimator = mrcf" indicates the average of a manually recentered causal forest. The table reports the average (fourth column) as well as the lower and upper bound of the 95 percent confidence interval (columns five and six) of the ATE. These bounds have been computed by simulating 500 draws of the estimated treatment effect from a normal distribution with mean $\hat{\tau}_i$ and variance $\hat{\sigma}_i$. Then the average treatment effect is computed for each draws and the lower and upper bound are the 2.5th and 97.5th percentile.