CESifo WORKING PAPERS

10532
2023

June 2023

Dynamic Causal Forests, with an Application to Payroll Tax Incidence in Norway

Evelina Gavrilova, Audun Langørgen, Floris T. Zoutman

CESifo

Impressum:

CESifo Working Papers
ISSN 2364-1428 (electronic version)
Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH
The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de
Editor: Clemens Fuest
https://www.cesifo.org/en/wp
An electronic version of the paper may be downloaded
· from the SSRN website: www.SSRN.com
· from the RePEc website: www.RePEc.org
· from the CESifo website: https://www.cesifo.org/en/wp

CESifo *Working Paper No. 10532*

Dynamic Causal Forests, with an Application to Payroll Tax Incidence in Norway

Abstract

This paper develops a machine-learning method that allows researchers to estimate heterogeneous treatment effects with panel data in a setting with many covariates. Our method, which we name the dynamic causal forest (DCF) method, extends the causal-forest method of Wager and Athey (2018) by allowing for the estimation of dynamic treatment effects in a difference-in-difference setting. Regular causal forests require conditional independence to consistently estimate heterogeneous treatment effects. In contrast, DCFs provide a consistent estimate for heterogeneous treatment effects under the weaker assumption of parallel trends. DCFs can be used to create event-study plots which aid in the inspection of pre-trends and treatment effect dynamics. We provide an empirical application, where DCFs are applied to estimate the incidence of payroll tax on wages paid to employees. We consider treatment effect heterogeneity associated with personal- and firm-level variables. We find that on average the incidence of the tax is shifted onto workers through incidental payments, rather than contracted wages. Heterogeneity is mainly explained by firm-and workforce-level variables. Firms with a large and heterogeneous workforce are most effective in passing on the incidence of the tax to workers.

JEL-Codes: C180, H220, J310, M540.

Keywords: causal forest, treatment effect heterogeneity, payroll tax incidence, administrative data.

Evelina Gavrilova
Department of Business and Management Science, NHH Norwegian School of Economics
Bergen / Norway
Evelina.Gavrilova-Zoutman@nhh.no

Audun Langørgen
Statistics Norway
Oslo / Norway
Audun.Langorgen@ssb.no

Floris T. Zoutman
Department of Business and Management Science
NHH Norwegian School of Economics
Bergen / Norway
floris.zoutman@nhh.no

June 27, 2023

For helpful discussions, comments and data access we would like to thank Astrid Kunze, Håkon Otneim, Jarkko Harju, Jonas Jessen, Krisztina Molnar, Oivind Anti Nilsen, Stefan Wager, Susan Athey, Tuomas Matikka, Ulrich Glogowsky and seminar and conference participants at IIPF Conference 2022, NHH Norwegian School of Economics, University of Bergen, University of Linz and the VATT Institute Helsinki. The authors declare that they have no competing interests.

1 Introduction

Policymakers often care about treatment effect heterogeneity. This allows them to target public interventions toward the groups that are most responsive, and to evaluate the effects of public policy on inequality. In addition, investigating heterogeneous effects can reveal information about underlying mechanisms that drive differences in outcomes. However, estimation of heterogeneous treatment effects is challenging in data-rich environments. The abundance of data offers empirical researchers too much flexibility in choosing specifications, which complicates inference (see e.g. Brodeur et al., 2016). One example of such a data-rich environment is the administrative data we exploit in our application, in which merging various administrative sources results in a large covariate-space.

As a response to such concerns, recent advances in causal machine learning have delivered the causal tree and forest algorithms (Athey and Imbens, 2016; Wager and Athey, 2018). Causal forests provide a data-driven approach to estimating treatment effect heterogeneity in the context of a cross-sectional experiment. A canny sample splitting technique, known as honesty, overcomes the issues generally associated with multiple hypothesis testing and overfitting.[1] For identification, causal forests rely on the strong assumption of conditional independence, which states that treatment status is independent of the outcome variable after conditioning on covariates.

Yet, there are limitations to the practical usefulness of causal forest methods. In many applied studies estimated on panel data, conditional independence is unlikely to be satisfied. In such a setting, causal forest estimates are not guaranteed to be consistent. Therefore, a recent overview paper Roth et al. (2022, p.42) calls for extending the causal forest method to a setting which assumes parallel trends but not (necessarily) conditional independence.

In this paper, we take on the challenge by developing a method that we name Dynamic Causal Forest (DCF). DCFs extend causal forests to a setting in which identification comes from parallel trends, rather than conditional independence. DCFs allow researchers to study heterogeneity of causal effects both between covariates and over time. Our main result shows that DCFs provide a consistent estimate for the average treatment effect on the treated, conditional on both covariates and time (CATT), provided that the outcome variable satisfies parallel trends conditional on covariates. Similar to event studies, DCFs can be used to examine whether pre-trends are parallel,

[1] The sample splitting technique splits samples into a training sample, on which a forest is trained, and an estimation sample, on which causal effects are estimated. (Wager and Athey, 2018) show that standard methods for inference apply in this setting for the estimation sample.

and to study treatment effect dynamics. Our method thus combines event study design with the strengths of causal forests, in which DCFs provide optimal ways of splitting the data to capture treatment effect heterogeneity.

To develop our method, we consider a set-up with panel data on an outcome variable of interest and a set of covariates that are assumed to vary by unit, but not over time. We assume a single treatment, occurring in period $b+1$ which divides the sample in a treatment and a control group (here period b signifies the base period).[2] The panel must be partially balanced in the sense that all units must be observed in the base-year period, and at least one other period. We assume that the outcome variable satisifies parallel trends after conditioning on covariates.

We use the following key insight to develop DCFs. Parallel trends on outcome variable y_{it} restrict the relationship between the base-year differenced outcome variable $z_{it} \equiv y_{it} - y_{ib}$ (henceforth the differenced outcome variable) and assignment to the treatment group. The parallel-trend assumption we make in this paper, implies that in the absence of treatment the expected value of z_{it} does not depend on treatment status after conditioning on covariates. We show that this assumption is sufficient to apply the main theorem of Wager and Athey (2018) (Theorem 11) to z_{it} for the treated group. Hence, under parallel trends, a causal forest estimated on z_{it} yields consistent estimates for the CATT.

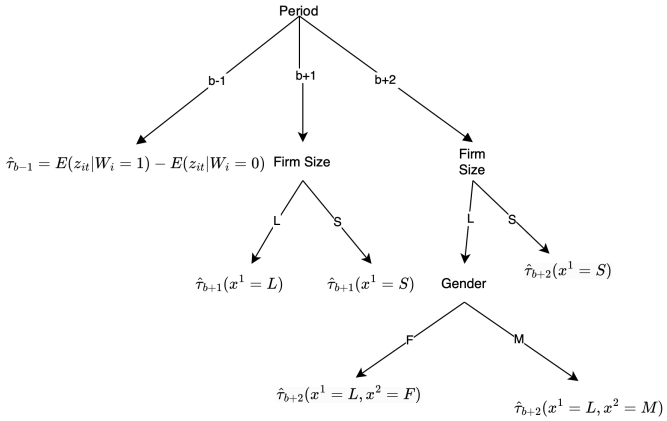
Leveraging this insight, we develop DCFs by i.) base-year differencing the outcome variable, ii.) applying Neyman-orthogonalization to the outcome variable of interest and the treatment variable to reduce the finite-sample bias inherent to machine-learning models (see e.g. Chernozhukov et al., 2018; Athey et al., 2019) and iii.) estimating a causal forest on the differenced outcome variable for each period $t \neq b$. Intuitively, the latter step is equivalent to estimating a forest on the full sample, but forcing each tree to first branch out by time, before considering splits along other covariates.[3]

Figure 1 provides a schematic of a tree in a DCF. Related to our application, the example illustrates an intervention that produces heterogeneous treatment effects in a sample of workers. There are 4 periods ranging from $b-1$ to $b+2$ and two covariates: the size of the firm, and the gender of the worker. All splits, except the initial time-split are data-driven and only occur if the algorithm finds significant treatment effect heterogeneity on the basis of the covariates on a training sample. Final splits of trees are known

[2] In the Appendix we extend our method to allow for staggered treatment.

[3] This third step is essential because the causal-forest estimator is consistent, but biased in finite samples. The bias that is most worrisome in a dynamic setting is the bias that follows when the sample is not split by time, and hence, pre- and post-reform are combined into a single estimate. Forcing the forest to first separate observations by time prevents this from occurring.

Figure 1: Example of a tree in a DCF



Notes: The Figure displays a simplified example tree in a DCF estimated on a sample of workers with covariates firm size (x^1) and gender of the worker (x^2). Trees always first split by period. There is no split for $t = b$, since the treatment effect is not identified in the base year. In period $b - 1$ the tree algorithm finds no treatment effect heterogeneity. Hence, the tree-estimate for the CATT is independent of the covariates, and given by the difference in the differenced outcome z_{it} between the treatment ($W_i = 1$) and control ($W_i = 0$) group. In period $b + 1$ the tree finds evidence of heterogeneity by firm size between large (L) and small firms (S). Here the CATT is estimated on the same difference in z_{it} but conditioned on x^1. In period $b + 2$ the algorithm finds evidence of additional heterogeneity by gender within large firms, and hence, the estimates are split accordingly.

as leaves. The average treatment effect in a particular leaf is the difference between treated and control units in the differenced outcome variable within the leaf. Hence, identification is based on the difference-in-difference (DiD) in the outcome variable of interest.

Proposition 1 of this paper formally shows that DCFs provide a consistent estimate for the treatment effect. This result relies on the following assumptions. First, the expected outcome variable must satisfy a parallel trend assumption in the absence of treatment, conditional on covariates. Second, there must be overlap between treatment and control group in all regions of the covariate-space. Third, we make a few regularity assumptions, which among others require that the CATT varies continuously with the covariates. All assumptions, apart from the parallel-trend assumption are equivalent to the assumptions taken in Wager and Athey (2018).

DCFs allows for standard inference, such as hypothesis testing of treatment effects between subpopulations or over time. In addition, the estimates can be used to create event-study diagrams to, for instance, study treatment dynamics for the average treatment effect, or for the treatment effect among a subsample of interest. DCFs also produce a variable-importance matrix for each period which allows researchers to identify

subsamples of interest. We demonstrate all of these features of DCFs in our empirical application.

Application In our empirical application of the DCF method, we study the incidence of payroll taxes on wages paid to workers in Norway. There is a large literature that studies the incidence of employer-paid payroll taxes on the wage of workers, but estimates of the pass-through coefficient vary strongly. For instance, Gruber (1997) find that all of the incidence is passed through to workers in Chile, whereas a more recent study by Ku et al. (2020) on Norwegian data find that most of the incidence remains with the employer. Research by Saez et al. (2012, 2019) provides evidence for rent sharing at the firm level, implying that the firm's exposure to the payroll tax is more important than the exposure of an individual worker. This wide variety of results implies that there can be significant treatment effect heterogeneity with respect to personal and firm characteristics.

At the same time, there exists a large literature in labor economics which finds evidence that firm-specific premiums are an important driver of wage inequality (e.g. Abowd et al., 1999; Card et al., 2013).[4] One possible channel for mediation of such inequality arises when some firms are better able to shift tax incidence to workers than others.

Nevertheless, studying the heterogeneous incidence of payroll taxes has so far received limited attention in the literature for two reasons. First, there is little employer-employee matched data that also cover a variety of wage concepts that can reveal different margins of adjustment to the tax. Second, there are many relevant dimensions of heterogeneity from a conceptual point of view, whilst theory offers little guidance concerning which dimensions are most important.

To address these key challenges, we exploit one of the main features of the forest algorithm, the variable importance matrix, as a data-driven method to determine the main margins of heterogeneity in the treatment effect in a panel data setting. Thus, the methodology allows us to identify the relative importance of firm vs worker-specific characteristics on payroll tax incidence.

Our empirical application utilizes a special survey panel of matched employer-employee data from Norway that contain information on various wage concepts, most notably, contracted wages, and wages inclusive of overtime payments and bonuses. We link our data to administrative data on individual- and firm-level tax returns, demographic data and education data, which provide a host of covariates that can potentially mediate treatment. We are interested in examining the margins along which we observe the largest

[4]See Card et al. (2018) for an overview.

heterogeneity in the pass-through of the payroll tax. Given that incidence is the result of bargaining between workers and employers, we are interested in variables that typically influence the bargaining position of workers. For example, variables such as gender and education, can play a major role in pass-through of the payroll tax, shedding light on the practical mechanisms of the incidence of the payroll tax.

To account for alternative channels, our DCF includes both firm and worker-level covariates. Further, for each worker-level covariate, we also include a firm-level aggregate. This allows us to, for example, distinguish between the case where heterogeneity is primarily driven by the union-status of the worker, or by the fraction of unionized workers within a firm. The DCF can distinguish between these cases, because the former would be observationally consistent with large within-firm heterogeneity in payroll-tax incidence, whereas the latter only drives between-firm heterogeneity.

To obtain causal evidence we employ the following reform. Norway has a system of regionally differentiated payroll tax rates that has undergone several reforms. We focus on a Norwegian payroll tax reform implemented in 2004. For workers in our Treatment Zone, the payroll tax rate increased permanently in 2004, whereas the tax rate remained unchanged in our Control Zone throughout the period of analysis. We use this reform as a quasi-experimental framework and apply both standard difference-in-difference and DCF methods. We focus on workers that can be observed throughout the sample period from 2002 to 2008.

With respect to the average treatment effect, we find suggestive evidence that some of the tax incidence is shifted to workers when considering contracted wages as outcome, but the response of contracted wages does not differ significantly from zero. However, this result changes when we examine as an outcome variable the full wage inclusive of overtime and bonus payments. In this specification, the regression event study has large confidence intervals which do not rule out full incidence on the firm, but also contain (close-to) full incidence on workers. DCF estimates are more precise, and the central estimate implies that incidence is fully shifted on the worker. The effect becomes apparent in the second year after the reform and it persists until 2008, the end of the sample period.

If we were confined to only using event study analysis we would be tempted to delve into the heterogeneity analysis by running multiple regressions, where for each splitting variable we would vary the splitting rule and the functional form. Instead, we rely on DCF method to decompose the long-term treatment effect heterogeneity of the payroll tax. We group explanatory variables into several groups, consistent with theoretical expectations on the impact of payroll taxes. We find that firm characteristics are the

most important in determining how full wages react to changes in the payroll tax. More specifically, important variables are firm size in terms of number of employees, the gender ratio and unionization rate at the firm level.

Overall, the emerging picture is that large firms, with a more heterogeneous workforce are most effective in passing through the incidence of the tax onto workers. The effect on firm size in terms of employees is monotonous, such that smaller firms are less likely to pass through the incidence. With the other variables, the effect is non-monotonic, such that firms with a more even gender ratio and average unionization rates tend to pass-through more of the incidence. These findings could be consistent with the idea that workers in larger firms with a more heterogeneous workforce find it more difficult to form a coalition when bargaining with management.

The rest of the paper is structured as follows. The Dynamic Causal Forest methodology is developed in Section 2. In Section 3, we provide the background on payroll tax reform in Norway and the data that are utilized in our application. The empirical results are discussed in Section 4, and Section 5 concludes. We review related literature in the main text. Methodological literature is discussed at the end of section 2, whereas literature related to our application is discussed at the end of section 4.

2 Methodology

2.1 Set-up

We consider a panel dataset which contains i.) the outcome variable of interest y_{it}, where i is the observational unit, and t denotes period of time, ii.) a treatment indicator W_i which equals 1 for the treated units, and zero for control units throughout time, and iii.) a set of d covariates denoted by X_i which are assumed to be constant over time. The panel is not required to be balanced. However, all units must be observed in a base year b and, at a minimum, one additional time period. The number of units observed in period t is expressed by N_t. Denote the set of outcomes in period t by $\mathcal{S}_t \equiv \{y_{it}\}_{\forall i \in \{1,\dots,N_t\}}$.

We are interested in the conditional average treatment effect on the treated (CATT). Formally, we define the horizon-specific CATT in period t as:

$$\tau_t(x) \equiv \mathbb{E}\left[y_{it}(1)|W_i = 1, X_i = x\right] - \mathbb{E}\left[y_{it}(0)|W_i = 1, X_i = x\right], \tag{1}$$

where $y_{it}(W)$ denotes the potential outcome variable under actual and counterfactual treatment status $W \in \{0, 1\}$ in period t.

Our aim is to evaluate the effects of a policy intervention that is implemented in

one step in year $b+1$ during our sample period.[5] Event studies are commonly used in such settings to account for dynamic treatment effects that may depend on the time elapsed since the onset of treatment. Although we are primarily interested in the post-treatment effects, this research design also accounts for the "treatment effects" prior to the intervention. The pre-treatment effects are used to inspect (deviations from) parallel trends in the outcomes within treatment versus control group. Thus, we adopt the convention that treated units are treated in all years t except for a base year b. For the treatment group, we observe $y_{it}(1)$ for all years $t \neq b$, whilst $y_{it}(0)$ is the counterfactual outcome, except in the base year. For the control group, we only observe the untreated state. Therefore, in equation (1), for all years $t \neq b$, the second term is unobservable.

2.2 Causal Forest Estimator

In this paper we introduce Dynamic Causal Forest (DCF) as a method for estimating the CATT in (1). DCFs share a large number of properties with regular (causal) forests, which we will shortly introduce here. Forests consist of trees. For future reference, we denote the k-th tree estimated on a sample from period t by Π_t^k. In (causal) forests each tree is "trained" on a random subsample of the data through recursive partitioning. Denote the sample used to train Π_t^k, by $\mathcal{J}_t^k$.

Trees split data along the covariates. The training algorithm creates a split in the data along the covariate that provides the best prediction of heterogeneity in the estimate of interest, in our case the treatment effect. Final splits of trees are known as leafs. Let x be a point of interest in the covariate space. Then the leaf around x consists of all observations in the tree Π_t^k which are contained in the same leaf as an observation with covariate-vector x. By construction, causal forest secures that each leaf contains both treated and control units.

With many, or continuously distributed covariates, the data can in principle be subdivided until each leaf contains a single treated observation. To prohibit overfitting, the training algorithm contains several stopping rules that can for instance require a minimum number of observations/groups in each leaf, or a minimum level of treatment effect heterogeneity between leafs.

Training and estimating trees on the same (sub)sample results in inconsistent estimates, because of dependence between the structure of the tree and the estimates produced by the tree. Therefore, Athey and Imbens (2016) introduce a concept known as "honesty" to the machine-learning literature, which relies on sample-splitting methods

[5]The Appendix sketches an approach to extend DCFs to staggered-treatment settings.

and separation between training and estimation sample. An honest tree estimates the average treatment effect in each leaf of the tree using the estimation sample $\mathcal{I}_t^k = \mathcal{S}_t \setminus \mathcal{J}_t^k$. By contrast, the structure of the tree is determined by using the training sample $\mathcal{J}_t^k$. As a result, the structure of the tree and the treatment effect estimates are independent, thereby overcoming the crucial challenge of producing consistent estimates with machine-learning methods.

To understand how trees can be used to estimate heterogeneous treatment effects consider a given tree Π_t^k that is estimated on a sample from period t. Let $\mathcal{I}_t^k(x, W)$ denote the subset of the estimation sample that falls in the leaf around x and has treatment status W. One way to estimate the treatment effect at x is to take the difference of the average value of the outcome variable in the leaf between the treatment and the control group. The expected value of such an estimator is given by:

$$\mathbb{E}[\Pi_t^k(x)|X_i, W_i] = \frac{\sum_{i \in \mathcal{I}_t^k(x,1)} \mathbb{E}[y_{it}(1)|X_i, W = 1]}{|\mathcal{I}_t^k(x, 1)|} - \frac{\sum_{i \in \mathcal{I}_t^k(x,0)} \mathbb{E}[y_{it}(0)|X_i, W = 0]}{|\mathcal{I}_t^k(x, 0)|}. \quad (2)$$

This estimator has desirable properties in the sense that it is consistent for the conditional average treatment effect (CATE) when treatment assignment is conditionally independent (Athey and Imbens, 2016; Wager and Athey, 2018). To understand this, note that as the size of the leaf shrinks, the partitioning on x becomes finer. In the limit, as the leaf size shrinks to zero, confounding variation that is correlated with X_i is fully conditioned out, and assuming conditional independence, the bias vanishes. Wager and Athey (2018) show that the leaf size shrinks with the number of observations, thus proving consistency of the causal-tree estimator.

To get from a tree estimate to a forest estimate one simply takes an average of the treatment effects over all trees in the forest. Forests typically outperform singular trees, since forests consists of multiple trees that each split the data in different ways. Therefore, in finite samples, forests offer an improvement in comparison to singular trees (e.g. Breiman, 2001).

In settings where the conditional independence assumption is valid, we may employ the estimator in (2) directly to obtain estimates for dynamic treatment effects. Since the estimator accounts for heterogeneous effects by time, it provides dynamic treatment effect estimates that are comparable to conventional event study designs. A concern with this method, however, is that conditional independence is a rather strong assumption which is rarely plausible in settings with observational data.

2.3 Dynamic Causal Forest

Therefore, our setting differs from Wager and Athey (2018) because we do not assume that treatment assignment is independent of potential outcomes conditional on covariates. Estimates based on (2) will not be consistent when there are systematic differences between treated and control units within the leaf. Instead, we employ a weaker assumption of parallel trends conditional on covariates. Formally:

Assumption 1 *Parallel trends*

$$\begin{aligned} &\mathbb{E}[y_{it}(0) - y_{ib}(0)|W_i = 1, X_i] = \\ &\mathbb{E}[y_{it}(0) - y_{ib}(0)|W_i = 0, X_i] \quad \forall \quad x, t. \end{aligned} \tag{3}$$

That is, in the untreated state, for all time periods, and conditional on covariates, the difference in the average outcome between period t and the base period b would have been the same in the treatment group as observed in the control group. Assumption 1 restricts the expected counterfactual outcomes for the treatment group $E[y_{it}(0)|W_i = 1, X_i = x]$, but places no restriction on the counterfactual outcome for the control group. Hence, with this assumption we can, at most, identify the treatment effect on the treated, which is therefore our focus.[6]

We also assume overlap for all covariates.

Assumption 2 *Overlap*

For each X_i the probability of treatment is bounded away from zero and one:

$$1 > E[W_i|X_i] > 0 \quad \forall \quad X_i. \tag{4}$$

If there are regions in the covariate-space where the probability of treatment is either zero or one, then the overlap assumption is violated, and conditional treatment effects are not identified everywhere. Note that from an applied perspective, this assumption may impose restrictions on the covariates included in the forest. For instance, in our application we use regional variation in payroll taxation. Covariates that measure the location of a worker violate the overlap assumption, since in some locations all workers

[6]To also estimate the treatment effect for the control group we would have to make the additional assumption that trends in the control group are parallel to trends in the treatment group in the counterfactual case where the control group is treated. Most of the DiD literature does not make this additional assumption (see e.g. Angrist and Pischke, 2008).

are treated, whereas in other locations all workers are untreated. These covariates can therefore not be included in the forest.

Under parallel-trend assumption 1 the CATT is identified through DiD conditional on X_i:

$$\tau_t(x) = \mathbb{E}\left[z_{it}(1)|W_i = 1, X_i = x\right] - \mathbb{E}\left[z_{it}(0)|W_i = 0, X_i = x\right], \tag{5}$$

where $z_{it} \equiv y_{it} - y_{ib}$ denotes the outcome variable differenced by the base year, and the equality follows from substituting (3) into (1). When x is high-dimensional and/or continuous, it is not possible to directly estimate (5), since fully conditioning out x would absorb all variation, unless the researcher imposes strong conditions on the functional form of $\tau(x)$, *and* restrictions on the conditioning set that result in parallel trends.

To apply causal forests to a setting with parallel trends, i.e. to develop a DCF, we make three adjustements to the causal-forest algorithm. First, we use the differenced outcome variable z_{it} as an input to the causal forest. Intuitively, the idea is that causal forests identify the treatment effect based on the difference between the treatment and the control group. Using the differenced outcome variable z_{it} as an input to a causal forest, implies that we are effectively identifying the treatment effect through DiD.

Second, we separately estimate a causal forest for each time period $t \neq b$. This approach is essential to ensure that the treatment effect for each period is identified separately. To see this, consider an alternative approach in which a causal forest is estimated on the full sample, but time t is included as a regular covariate. In this case, through random subsampling, there will almost surely exist trees with leafs that contain observations from different time periods. This method leads to bias because the estimated treatment effect in period t will partly depend on treatment effects for other periods. To manually separate the data by time before estimating causal forests is equivalent to estimating a forest on the full sample, but forcing each tree in the forest to first split on time periods, before considering other splits. Hence, our method secures that observations from different time periods are never combined into a single estimate.

Third, we only predict the treatment effect for the treated units, since the parallel-trend Assumption 1 is insufficient to identify the treatment effect for the control units. Below we outline the algorithm for an DCF estimated on outcome variable y_{it} with treatment assignment W_i, covariates X_i and base period b.

Algorithm 1 *1. Transform the outcome variable $z_{it} = y_{it} - y_{ib}$*

2. For each period $t \neq b$, estimate a regular causal forest using z_{it} as the outcome

variable, W_i as the treatment variable and X_i as the covariates to obtain the DCF estimate $\hat{\tau}_t(x)$. Here we apply the causal-forest algorithm and code developed in Athey et al. (2019).

3. *Predict the treatment effect for the treated units:*

$$\hat{\tau}_{it} = \hat{\tau}_t^{-i}(X_i) \quad \forall \quad i \in W_i = 1 \tag{6}$$

where the superscript $-i$ denotes the fact that we predict the treatment effect for unit i on those trees in the forest for which example i is not in the training samples, i.e. trees that satsify $y_{it} \notin \mathcal{J}_t^k$ consistent with the honesty-property defined above. For future reference, we refer to this as out-of-bag estimates.

4. *Use the variable-importance matrices to identify subgroups that exhibit strong heterogeneity in the treatment effect of interest (see subsection 2.5 for more details).*

The structure of our algorithm is represented in Figure 1. For each tree in the forest, data is first separated by time. Subsequent splits are made by the algorithm based on the training sample. The CATT is estimated on the difference in the differenced-outcome variable (i.e. the DiD) in each leaf.

In order to prove that this algorithm provides a consistent estimate for the CATT, we require that the data-generating process satisfies a number of regularity assumptions:

Assumption 3 *Regularity assumptions*

1. *The covariates X_i are continuously distributed in the unit-hypercube $(0,1)^d$*
2. *$\mathbb{E}[z_{it}(W)|X_i = x]$ and $\mathbb{E}[(z_{it}(W))^2|X_i = x]$ are Lipschitz continuous for all x, t, W*
3. *$var[z_{it}(W)|X_i = x] > 0$ and $\mathbb{E}[|z_{it}(W) - E[z_{it}(W)|X_i = x]|^{2+\delta}|X_i = x] \leq M$ for some constants δ, M uniformly over all x, t, W.*
4. *The distribution of $\mathcal{Z}_{it} \equiv (x_i, w_i, z_{it})$ is independent between units*

These assumptions play the following role in the proof. First, it is essential that leaf size shrinks as the number of observations increases. Part 1 of Assumption 3 ensures that covariates follow a continuous distribution in a precisely defined space. This allows for a clear definition of what it means for a leaf to shrink. Part 2 assumes that the expected value of the transformed outcome variable for the treated and control population, and hence, the treatment effects, exhibit sufficient continuity. Part 3 secures that the second

moments of z_{it} are bounded. Finally, Part 4 introduces a relatively strong independence assumption, however at the end of this section we generalize this by considering cluster-robust inference.

We now arrive at our main proposition:

Proposition 1 *Under Assumptions 1-3 the estimator $\hat{\tau}_t(x)$ that is derived by estimating a dynamic causal forest on (x_i, W_i, y_{it}) with base year b converges to:*

$$\frac{\hat{\tau}_t(x) - \tau_t(x)}{\sqrt{var(\hat{\tau}_t(x))}} \rightarrow \mathcal{N}(0,1), \tag{7}$$

for all t as the number of units approaches infinity, $N_t \rightarrow \infty$ for all $t \neq b$. The variance can be estimated through the infinitesimal jackknife estimator developed in Wager and Athey (2018).

Proof. Through Algorithm 1 a DCF estimated on (X_i, W_i, Y_{it}) corresponds to a causal forest estimated on $\mathcal{Z}_{it}$. Therefore, our aim is to show that Assumptions 1-3 are jointly sufficient to apply Theorem 11 in Wager and Athey (2018) to $\mathcal{Z}_{it}$ which proves consistency for the causal-forest estimator.

To see that this is indeed the case, note that Theorem 11 in Wager and Athey (2018) requires that:

1. $\mathcal{Z}_{it}$ satisfies regularity conditions equivalent to Assumption 3
2. W_i satisfies an overlap assumption equivalent to Assumption 2
3. The tree-estimator (2) applied to $\mathcal{Z}_{it}$ converges to $\tau_t(x)$ as the size of the leaf shrinks, and the number of observations per leaf increases (this corresponds to equation (25) in Wager and Athey (2018) with the exception that our version concerns the CATT rather than the CATE.

Expanding on the latter point, applying equation (2) to $\mathcal{Z}_{it}$ we arrive at:

$$\mathbb{E}[\Pi_t^k(x)|X_i, W_i] \quad = \quad \frac{\sum_{i \in \mathcal{I}_t^k(x,1)} \mathbb{E}[z_{it}(1)|X_i, W=1]}{|\mathcal{I}_t^k(x,1)|} - \frac{\sum_{i \in \mathcal{I}_t^k(x,0)} \mathbb{E}[z_{it}(0)|X_i, W=0]}{|\mathcal{I}_t^k(x,0)|},$$

which converges to the right-hand side of (5) as the size of the leaf shrinks. What remains to be shown is that the size of the leaf indeed shrinks to zero as the number of units increases. However, this is already shown in the proof to Theorem 1, 11 in Wager and Athey (2018) to which we refer for further details. ■

It is useful to discuss the intuition behind Proposition 1 in more detail. Specifically, it is important to know in which circumstances a method aimed at estimating (heterogeneous) treatment effects for cross-sectional data can be extended to panel data by simply differencing the outcome variable. In the spirit of the Rubin causal model, Rubin (1974), cross-sectional methods like causal forests typically assume conditional independence between treatment assignment and the potential outcomes:

$$(y_{it}(w) \perp\!\!\!\perp W_i) \,|X_i \quad \forall \quad w \in \{0, 1\}. \tag{8}$$

It is easy to see that any method that derives consistent estimates under this assumption can be extended to panel settings when the differenced outcome variable satisfies:

$$(y_{it}(0) - y_{ib}(0) \perp\!\!\!\perp W_i) \,|X_i = (z_{it}(0) \perp\!\!\!\perp W_i) \,|X_i. \tag{9}$$

Equation (9) is a particularly strong parallel-trend assumption which states that, in the absence of treatment, trends are conditionally independent from treatment assignment. In settings where assumption (9) is appropriate one can essentially extend all cross-sectional methods to a panel setting with the algorithm:

Algorithm 2 *1. Derive z_{it} by differencing the outcome variable*

2. Use a cross-sectional method of your choice to estimate the CATT.

However, it is uncommon to assume conditional independence on trends, (9), since it imposes the strong restriction that the full distribution of z_{it} is independent of treatment assignment. Testing such an assumption is non-trivial even in the pre-reform period. By contrast, the parallel-trend Assumption 1 only restricts the first moment of z_{it} (see Roth and Sant'Anna, 2020 for a further discussion on the difference between these parallel-trend assumptions).

Hence, in practice extending a cross-sectional method to a panel setting using our approach requires that the cross-sectional estimator retains its properties of consistency when only the first moment of the trends are restricted. For causal forests Wager and Athey (2018) impose conditional independence, assumption (8). However, the proof to their main result only requires a restriction on the first moment of y_{it}.[7] Hence, conditional independence is sufficient, but not necessary. This is the property we exploit in Proposition 1. We conjecture that a similar extension applies to many other cross-sectional methods for estimating heterogeneous treatment effects (e.g. Lee et al., 2017;

[7] Equation (25) in Wager and Athey (2018).

Chernozhukov et al., 2020) but leave the verification of this conjecture to future research.

2.4 Neyman orthogonalization and clustering

Proposition 1 proves consistency, but nevertheless $\hat{\tau}_t(x)$ is biased in finite samples. The source of this bias is that parallel trends are only assumed to hold after conditioning on X_i. However, in finite samples, trees cannot fully condition out all covariates. In other words, leafs can potentially contain units that are on different trends paths.

A practice that reduces this bias is Neyman orthogonalization, which was recently introduced in a machine-learning context by Chernozhukov et al. (2018). Neyman orthogonalization is the process of explicitly modelling the probability of treatment (i.e. the propensity score) and the outcome variable as a function of the covariates. This first stage typically applies machine-learning methods such as random forests. Intuitively, by first predicting out heterogeneity in the probability of treatment and the outcome variable that does not relate to treatment, the training algorithm of the causal forest becomes more sensitive to variation that drives treatment effect heterogeneity. Neyman orthogonalization has been shown to have a number of attractive properties. Specifically, Athey et al. (2019) show that orthogonalization reduces the bias in generalized random forests when there exists confounding variation.

In our application of DCFs we apply Neyman orthogonalization to the pair z_{it}, W_{it}:

$$z_{it} = f_t(x_i) + \epsilon_{it}, \tag{10}$$

$$W_{it} = g(x_i) + \delta_i, \tag{11}$$

where the functions $f_t(x_i)$ and $g(x_i)$ are estimated through a random forest. Note that $f_t(x_i)$ varies by time, such that a different forest is estimated for each period. On the other hand, $g(x_i)$ is time-invariant, since treatment status does not vary by time in our set-up. The causal forest is then estimated on the out-of-bag estimated residuals $\hat{\epsilon}_{it} = z_{it} - \hat{f}_t^{-i}(x_i)$, and $\hat{\hat{\delta}}_i = W_i - \hat{g}^{-i}(x_i)$. These residuals are called centered outcomes in Athey et al., 2019 .

Intuitively, explicitly predicting out differences in trends via (10) and systematic differences between the treatment and control group (11) implies that the remaining residuals fed to the causal forest are more similar in trends and characteristics. As such, when estimated on centered outcomes the causal forest is better able to select on covariates that modify the treatment effects, rather than effects of confounding variables that drive differences in trends.

To implement Neyman orthogonalization in our algorithm we apply the generalized

random forest (grf) package in R in step 2 of Algorithm 1 to estimate the causal forest, which automatically performs the local centering transformation to achieve orthogonality. The package also extends inference to allow for clustered standard errors, and we use this extension in our application as well (see Athey et al., 2019; Athey and Wager, 2019 for more details).

2.5 Variable-Importance Matrices in DCFs and heterogeneous treatment effects

One of the questions that is particularly relevant when analyzing treatment effects is how well the covariates predict treatment effect heterogeneity. DCFs provide a data-driven approach to this question in the form of covariate-importance matrices. The covariate-importance matrix provides information on the fraction of trees in a particular forest that splits the sample by a particular covariate. For instance, if a large percentage of trees splits by gender, this indicates that gender is an important predictor of treatment heterogeneity, and it is useful to further inspect how the treatment effect varies by gender.

Since a DCF estimates separate causal forests for each time period, it is possible to separately identify and distinguish between variables that predict long-term heterogeneity versus variables that predict short-term heterogeneity. Pre-reform heterogeneity may also be of interest. In particular, when parallel trends hold, the treatment effect should not vary by covariates in the pre-reform period. Therefore, the variable-importance matrix for pre-reform periods can be exploited as a diagnostic tool to investigate potential violations of the parallel trend assumption.

2.6 Review of related methodological literature

There exists a rapidly growing literature that combines insights from machine-learning with causal methods. Our main contribution to the literature is to devise a method to estimate heterogeneous treatment effects when identification relies on parallel trends. This contrasts with the original causal-forest method of Athey and Imbens (2016); Wager and Athey (2018); Athey et al. (2019), and other methods for estimating conditional average treatment effects (e.g. Lee et al., 2017; Chernozhukov et al., 2020) which require that the identifying assumption of conditional independence between the outcome variable and treatment assignment is satisfied.

There are some other recent papers that extend the causal-forest method. Gulen et al. (2020) marries causal forest to a regression discontinuity design. Miller (2020) also

uses causal-forest methods on dynamic data, but in his setting identification continues to rely on conditional independence, rather than parallel trends.

There exists a burgeoning literature that deals with issues related to traditional DiD estimators (see Roth et al., 2022 for an overview). The traditional method in the literature for estimating heterogeneous treatment effects under parallel trend assumption is through a two-way fixed effects (TWFE) model on the following form:

$$y_{it} = \alpha_i + \gamma_t + \tau_t W_i + \epsilon_{it} \quad \forall \quad X_i \in \mathcal{X}, \tag{12}$$

where α_i and γ_t denotes a full set of fixed effects for observational units and periods of time, and $\mathcal{X}$ is a subgroup of interest that is defined by the researcher. The parameter of interest is the treatment effect τ_t.[8] Equation (12) is typically estimated with OLS. The estimates derived through TWFE models are unbiased if, within the subgroup $\mathcal{X}$, trends are parallel (e.g. Roth et al., 2022).

DCFs offer various advantages over TWFE models. First, the covariate-importance matrices provide further insight into the responses of heterogeneous subgroups of interest. In particular, this can be helpful in settings where theory is vague and previous empirical work is ambiguous regarding the relative importance of treatment effect modifiers. Second, DCFs estimates remain consistent even if trends are not parallel within a particular subgroup $\mathcal{X}$, but are parallel when conditioning on the full set of observable covariates. The reason is that DCFs only makes comparisons between treatment and control units that fall in the same leaf. In the limit, as these leafs shrink, DCF-estimates therefore completely condition out confounding variation related to observable covariates. Thus DCFs provide a clear distinction between heterogeneity in trends and heterogeneity in the treatment effect, which TWFE models do not. Third, because the model of heterogeneity is specified more explicitly, DCF estimates are likely to be more precise. For instance, in our application, we find that standard errors are typically smaller with DCFs than with TWFE models.

The methodological literature on DiD has recently grown extensively (e.g. Schmidheiny and Siegloch, 2019; Roth and Sant'Anna, 2020; Borusyak et al., 2021; Goodman-Bacon, 2021; Sant'Anna and Zhao, 2020; Chang, 2020; Callaway and Sant'Anna, 2021; Wooldridge, 2021. The main estimate of interest in this literature is the average treatment effect on the treated (ATT), rather than the CATT. For this reason the overview article by Roth et al. (2022, p.42) considers extending heterogeneous treatment effect

[8] In the context of our application in payroll taxation some recent examples of this methodology are Saez et al. (2019); Ku et al. (2020); Benzarti and Harju (2021a).

estimators like causal forests to a DiD setting a "promising area of research". Nevertheless, DCFs can also be used to provide an estimate for the ATT, by simply averaging the CATT over all units. Therefore it is useful to make a comparison in the context of the ATT.

The recent DiD literature has uncovered various shortcomings of the TWFE model. The first is that TWFE estimators of the ATT are only consistent if trends are parallel unconditionally. Related to our set-up, (Abadie, 2005; Sant'Anna and Zhao, 2020; Callaway and Sant'Anna, 2021) extend DiD to a setting where trends are parallel after conditioning on X_i. Similar to these approaches, Proposition 1 shows that DCF-estimates remain consistent when trends are parallel conditional on X_i. The second shortcoming is that TWFE-estimators are biased with staggered treatment timing when the treatment effect is heterogeneous between cohorts. The reason is that TWFE-models, estimated through OLS, make "forbidden comparisons" between units that are treated early, and units that are treated later. Borusyak et al. (2021); Goodman-Bacon (2021); Callaway and Sant'Anna (2021) make progress by deriving estimators of the treatment effect that are more explicit with respect to the comparisons that drive identification. In the Appendix we apply these insights to extend DCFs to a setting with staggered treatment timing where we only allow for comparisons between treated units and never-treated units.

3 Application: Norwegian Payroll Tax

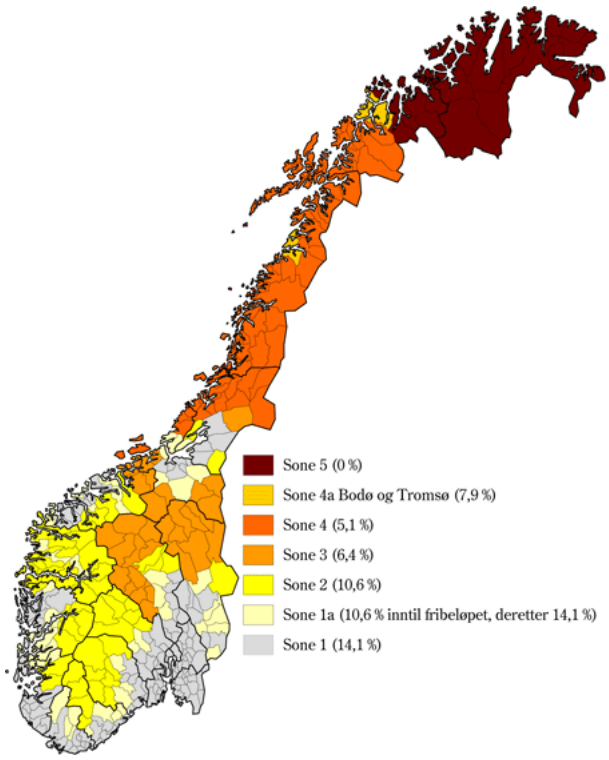
3.1 Background

In Norway, payroll taxes are paid by employers and collected by the central government. Revenues from payroll taxes are used to fund state pensions. It is important to note that there is no linkage between pension benefits and payroll-tax payments, in the sense that workers in all payroll tax zones have the same pension rights. The payroll tax rate is geographically differentiated since 1975. The official objective is to counter negative trends of depopulation and underemployment in peripheral and rural regions of the country. Over the years, the number of zones and the tax rates have undergone several reforms. The system has seen multiple reforms in the 00's which were mandated by interpretations of EU law. Specifically, regulations of unfair competitive tax advantages in the EU market were found to require changes in the Norwegian payroll tax system.[9]

[9]Although Norway is not a member of the EU, it must still abide by most EU regulations due to its membership of the European Economic Area (EEA).

As a result, a reform was imposed in 2004 whereby most zones saw an increase in their payroll tax rates. However, since the 2004 reform was contested between Norway and EU, it was partly reversed in 2007. In Figure 2 we show a map of Norway with the (current) distribution of payroll tax zones since 2007.

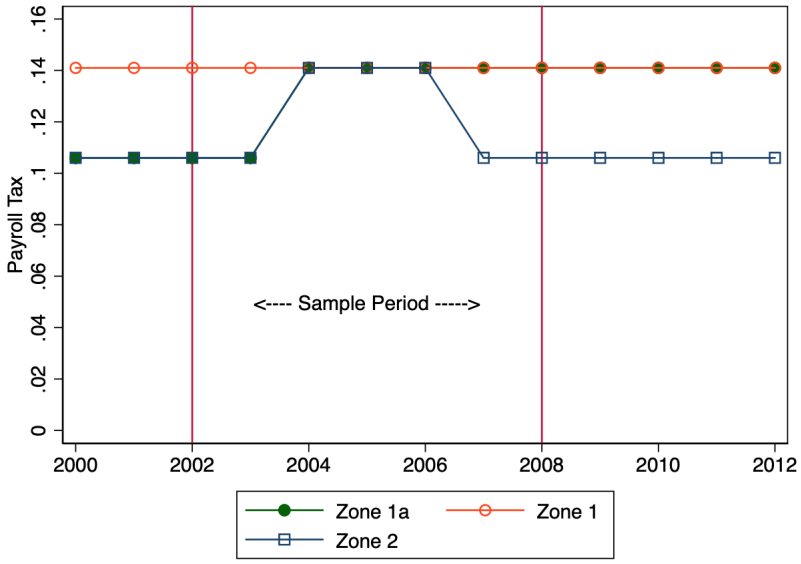
Figure 2: Geographically Differentiated Payroll Tax for Norway



Notes: The Figure plots payroll-tax zones and corresponding tax rates in Norway at the end of our sample-period.

Our study exploits features of the reform in 2004 that became permanent. More precisely, since reforms in Zones 3 and 4 and parts of Zone 2 were reversed in 2007, our focus will be on the reform implemented in Zone 1a when using Zone 1 as control group. In Figure 3 we show the evolution of the payroll tax rates for Zones 1, 1a and 2. The payroll tax rate in Zone 1 remained unchanged at 14.1 percent throughout our period of analysis. In 2004, the rate in Zone 2 increased from the lower rate of 10.6 percent in pre-reform years to the same level as in Zone 1. Three years later, Zone 2 was divided in two parts; Zone 1a which kept the same tax rate as Zone 1, and the "new" residual of Zone 2, where the tax rate was cut back to the pre-reform level, which means that the reform became transitory. By contrast, Zone 1a was exposed to a permanent tax rate

Figure 3: Payroll Tax Rate in Zones 1 and 1a in Norway



Notes: The Figure plots the evolution of the payroll tax rate over time for the Control group (zone 1) and the treatment group (zone 1A). Zone 2, which initially was combined with zone 1A, is plotted for comparison.

increase that closed the earlier gap in comparison to Zone 1.

This provides us with i.) a clean split in treatment and control regions, and ii.) a reform implemented in a single step in 2004 which clearly defines pre- and post-reform periods. Thus, the comparison of Zone 1a and 1 allows us to study longer-term effects in a setting that is appropriate for DiD and event study designs.[10]

Following the financial crisis, there was a recession and European debt crisis in the years 2008-2010. Notice that the effect of these crises on the Norwegian economy was relatively muted. For instance, annual unemployment decreased from 4.3 percent in 2005 to 2.5 percent in 2007, and later there was only a small increase during the height of the financial crises (3.5 percent in 2010).

[10]The 2007 reform also affected the way the payroll tax rate is determined. Prior to 2007, the payroll tax rate depended on the residence of the worker. After 2007, the payroll tax rate depends on location of the firm in which the worker is employed. To circumvent potential confounding variation related to this, we focus on firms and workers that are located in the same tax zone throughout our sample period.

3.2 Data

Our main sample comes from the wage register that is maintained by Statistics Norway. The wage register is a matched employer-employee sample based on an annual survey of firms which is taken yearly around September/October. For each job contract, sampled firms are required to provide detailed information on (the compensation of) their employees. Most importantly for our study, the wage register contains information on i.) contracted monthly wages at the time of the survey, ii.) average monthly overtime payments during the year, iii.) average monthly surcharges due to working in weekends/nights, and iv.) bonus payments.

Large firms, and public sector organizations are sampled every year. Medium private sector firms are sampled each year with probability 1/2. Small private sector firms are sampled with probability 1/10. The sampling procedure makes use of industry-specific thresholds to categorize firms according to size. In most cases, firms with more than 300 employees are considered as large firms which are always included. Firms with less than 5 employees are excluded from the sample. The sampling unit is firm-by-industry, which means that a sampled unit will report information about employees in the plants that belong to a given firm and industry (by main division). Besides inclusion of large firms, there is complete coverage of firms with membership in a selection of Employer's Associations. The latter coverage is aimed at producing relevant wage statistics for the participating organizations, but also contributes to increase the overall sample size.

We combine information in the wage register with several administrative registers that we can link through unique individual identifiers and firm/plant identifiers. The data sets include tax and social security registers, employer-employee register, education registers, household register, corporate accounting register and firm/plant unit register that contain records for Norwegian individuals and/or firms for the years 2002 to 2008. The databases provide extensive demographic and socioeconomic information for individuals in combination with their wages and characteristics of the firms where they are employed. Municipality identifiers allow us to observe the location of firms and residents by payroll tax zone.

The treated observations are workers who are employed by firms located in any municipality in Zone 1a, whereas control observation are employed in a subset of municipalities in Zone 1. Since Zone 1a does not contain municipalities with higher levels of centrality, the excluded municipalities in Zone 1 are more central according to the official classification made by Statistics Norway. From the treatment and control municipalities, we make the following selection of our sample. First, we drop all workers with more than

2 jobs in the Wage register. Second, we focus on the private sector, and hence, drop workers in the public sector and in healthcare, since the payroll tax only applies in the private sector. Third, we remove observations with negative income from labor on tax returns. Fourth, we keep workers that have been employed at a sampled firm for the whole sample period, implicitly removing retirees, new entrants and job switchers who switch to unsampled firms. Fifth, we focus on worker-firm pairs in which both the firm and the worker are in payroll-tax zone 1/1a. Sixth, we drop workers who move from zone 1 to 1a or vice versa after the reform. We are left with 2970 firm-year observations and 44 510 worker-year observations.

To analyze the impact of an increase in the payroll tax on wages we use two outcome variables in order to capture different margins of responses to the reforms in the payroll tax. We show the summary statistics for these variables in Table 1. First, we present the Monthly Full Wage, which is the contracted monthly wage plus overtime and bonus payments. Second, we present the contracted monthly wage (*grunnlonn*) denoted as Contracted Wage in Table 1. This variables captures responses in the contracted wage, which is affected by changes in hours worked per month as well as wages per hour. As a robustness check, we also make use of contracted wage standardized by the contracted hours per month as a proportion of regular full-time position. This is a measure of the full-time equivalent of the contracted wage, denoted Full-time Equivalent Wage. For example, if a worker was half-time employed, then this variable would contain their compensation as if they were full-time employed and their actual compensation would be multiplied by two. Full-Time Equivalent Wage differs from Contracted Wage for part-time workers, but not for full-time workers.

Table 1 displays four columns. The first two columns show the average values for observations in the control Zone 1. The second two columns show the average values for units in the treatment Zone 1a. Each set of columns is divided into a Before and After period, where 2003 is the base year. On average for the two wage concepts, we find that wages in the sample are 5-7 percent higher in Zone 1 than in Zone 1a. Between wage concepts the Full Wage is around 10 percent higher than the Contracted Wage, on average. Finally, in Table 1 we present the treatment variable in the dynamic causal forests, which is set to 1 for the treated zone 1a throughout the whole time period.

In Table 2 we present the covariates in our analysis. These variables are measured in the pre-reform period 2002 and 2003 and do not vary by time. The covariates fall in two groups - depending on whether they are measured at the firm level or at the individual/household level. All variables are further described in Appendix Section B.

From the firm level we measure the number of employees and codes in the NACE

industry classification. In addition, we make use of variables that are based on firm balance sheets, such as the capital-to-labour ratio, earned capital, liquidity and cash holdings, investments and dividends. Capital-to-labour ratio is defined as fixed tangible assets divided by the number of employees, whereas other variables from the balance sheets are standardized by total assets of the firm.

From the worker level we observe age, gender, household characteristics, social assistance, (years of) education, indicator for labor union membership, and indicator for workers who received bonus payments in the pre-reform period. Out of these, social assistance recipiency is observed in the tax register and it includes child and other subsidies. The household status is given by the following variables: Couple in the household denotes whether the worker is married or part of a cohabiting couple. Number of adults in the household denotes how many adults above 18 years old are part of a given household. This variable counts both adults in a couple and any other adults. Number of children in the household denotes the number of individuals who are 18 years or younger in the household.

In order to distinguish between effects driven by worker-specific characteristics, and effects driven by workforce composition, we also aggregate each of the worker-level variables to the firm-level. We aggregate the worker variables as follows. For binary variables we take the mean at the firm level. For continuous variables we take the median at the firm level. For categorical variables we take the mode.

4 Results

4.1 Different Wage Concepts

This section presents results on how different wage concepts are affected by changes in the payroll tax. Figure 4 displays event-study plots for estimates of the incidence of the payroll tax on different measures of worker compensation. Estimates are normalized such that a coefficient of 0 or -1 indicates zero versus full incidence on the worker, respectively. Each plot reports estimates from the dynamic causal forest in black and conventional regression event study estimates in red. In panel (a), the dependent variable is the contracted wage, which captures the compensation for contracted hours. Although confidence intervals are too wide to draw any firm conclusions, the TWFE regression suggest that there is around 50 percent payroll tax incidence on contracted wages. Point estimates for short-term incidence are larger through the lens of DCF than with TWFE

regression.[11]

Next, in panel (b) we examine the notion that the employer could shift the incidence of the payroll tax to workers through a reduction in non-contracted payments. The dependent variable is the monthly full wage, inclusive of bonus and overtime payments. In panel (b), estimates from DCF indicate that most of the incidence is on workers when accounting for non-contracted compensation in the wage concept. By contrast, the TWFE regression finds smaller point estimates which imply that the incidence remains mostly with the firm.

Overall, through a TWFE regression event study we find consistent evidence in all wage concepts that the employer retains most of the incidence of the increase in the payroll tax. However, when we more closely control for firm and worker heterogeneity via a DCF we find evidence for shifting of the incidence through incidental payments. An important difference between regression-event studies and DCFs that could explain this difference is that the treatment variable, and the outcome variable in the DCF are first Neyman-orthogonalized. This implies that identification in the DCF comes from firm-worker pairs that are similar in terms of observable covariates, but that nevertheless receive different treatment. In contrast TWFE regression does not match on observables. The implication is that the wage of comparable worker-firm pairs in the treatment zone has declined relative to the control group consistent with full incidence on the worker. In the next section we focus on disentangling the heterogeneity in the treatment effect.

4.2 Heterogeneity

One of the main outputs of the causal forest is the variable importance matrix, which ranks the variables according to the proportion of splits along dimensions that modify the treatment effects. In Table 3 we present the most important variables in the last causal forest for 2008, effectively disentangling the sources of the long-run incidence on the full monthly wage. Panel A presents the variables with importance above 5 percent. The top variable, Firm Nr. Employees, accounts for 11 percent of the heterogeneity in the treatment effect. The next variables in the ranking are the gender ratio at the firm, the rate of unionization at the firm, the proportion of married employees in the firm and an indicator for the firm paying out bonuses.

Panel B of Table 3 reports summary statistics for the variables in panel A. We divide

[11] Contracted wages ignore the possibility that incidence of the payroll tax could be shifted to employees by changing hours worked. In Figure 9 b in the appendix, we consider the contracted monthly wage that is standardized by contracted hours per month. Estimation results for this wage definition does not differ materially from the results in Figure 4a.

the sample into quartiles by covariate and put together the second and third quartile. We observe that on average, small firms in quartile 1 have 31 employees, while medium-sized firms have 215 employees and large firms in the sample have on average more than five thousand employees. We also note that in all variables the values show that there are sizable differences between firms in different parts of the distribution, which may constitute a precondition for finding significant heterogeneity in the treatment effect.

Figure 5 provides a graphical representation of heterogeneity in treatment effects through event-study plots. Panel (a) shows estimates for the treatment effect for three subsamples of firms, divided based on quartiles in the distribution of the pre-reform number of employees. We find that incidence on workers is mainly mediated through large companies. Moreover, the estimated effects are monotonic: smaller companies display smaller treatment effect, larger companies have the strongest treatment effect.

In panel (b) in Figure 5 we present estimates showing a non-monotonic treatment effect. In this figure, the two quartiles in the middle of the firm gender distribution shift more of the incidence to workers. In companies with predominantly female employees the incidence remains with the firm, whereas the incidence is male-dominated firms is shared between worker and employer. In the longer term, firms with a mixed gender composition in the workforce are more effective at passing through the incidence of the tax onto workers. Notice, however, that *individual* gender *per se* plays only a small role in modifying the treatment effect (the total value in the variable importance matrix is presented in Figure 8, which is discussed further below.).

In panel (c) in Figure 5 there is another instance of a non-monotonic effect. We observe that firms in the middle of the unionization distribution are more likely to bear the incidence of the payroll tax. In this case we could expect that within firms, individual workers who are non-unionized will tend to bear the incidence, but again the individual characteristic accounts for little heterogeneity in the response of wages.

In panel (d) we show the treatment effect on subsamples of firms with different proportions of workers in a household couple. If anything, incidence on workers is larger in firms with a high proportion of single employees, but the difference in the long-run estimate is very small.

In Panel (e) of Figure 5, we find that firms which pay out bonuses display full incidence of the payroll tax on employees. By contrast, the point estimate for wage response in the first quartile of bonus payments is close to zero. These results are consistent with the finding from Figure 4, indicating that more incidence is passed onto full wages in comparison to contracted wages.

Overall, our findings, especially 5 a-c are consistent with the idea that tax incidence

on workers is larger in larger firms that have a more heterogeneous workforce.

In Figure 6 we present results from interacting the first two variables in the variable-importance matrix: Firm Nr. Employees and firm gender ratio. We construct nine separate groups that are defined by the intersections between 3 x 3 quantile groups along the two dimensions. Next, we create an event-study plot that accounts for payroll tax incidence on workers in each of the nine groups. We find that in most of the subgroups, the response of wages to increased payroll taxes does not differ significantly from zero.

The main takeaway from Figure 6 is that heterogeneous effects are mostly found among large firms, whereas there is less variation in treatment effects among small and medium sized firms. Among large firms, incidence is on the employer in firms with a high proportion of female workers, whilst incidence is on workers in firms with a mixed workforce composition by gender or with a predominantly male workforce. Hence, workforce composition in terms of gender, appears to be important in large firms, but not in smaller firms.

4.3 Variable Importance and Margins of Heterogeneity

An interesting question is what types of covariates are most important in modifying the treatment effect. Such information may provide suggestive evidence about the channels through which treatment effects are mediated. In Figure 7 we present the importance of the main five variables in each forest. In 2008, as shown in Table 3 the most important variable was the firm number of employees. In 2004 and 2005 the variables have similar importance, while in 2006 the gender ratio at the firm level gains prominence as the top variable. Looking at panel b in Figure 5 we observe that firms with female-dominated workforce deviate in trend from the other parts of the distribution, which increases the heterogeneity in the treatment effect and results in the relatively high importance of the firm gender ratio in 2006.

In Figure 8 we demonstrate a use of the variable importance matrix output. We sum the importance of the control variables into several possible channels of influence. To this end, we classify the covariates by level of aggregation and type of characteristic. The first basic division we make is between firm vs. individual level variables. Next, we add up the variable importance by type of covariate, where the combined importance expresses the contribution of firm vs. individual level variables to uncover differential treatment effects. We find that firm level variables account for 82.4 percent of heterogeneity, which means that individual level variables play a minor role for the estimated differences in tax incidence.

The firm variables can be roughly classified into balance sheet variables and characteristics of the workforce. The balance sheet variables account for 25 percent of the variation, while the leftover 58 percent is arguably accounted for by workforce characteristics. This was partially anticipated by the importance of composition of the workforce by gender, unionization and fraction of employees in a couple from Table 3.

Finally, we separate some workforce characteristics into age, gender and household characteristics, which are summed between firm and personal level variables. From this partition of variables, household characteristics hold the most importance.

4.4 Related Findings

There is an extremely large literature on payroll tax incidence. We here limit our discussion to comparing our approach and results to recent literature from the Nordics, which are more similar in terms of institutional setting (see Bozio et al., 2020 for a full overview).

Most closely related are papers that use regional variation in the payroll tax. Using various reforms in the 00s in Norway the central point estimates in Dale-Olsen (2018); Ku et al. (2020); Stokke (2021) are consistent with firms shifting between 15-30 percent of the payroll tax burden to workers. Similarly, using a reform in the 2002 in Sweden Bennmarker et al. (2009) find that around 25 percent of the payroll tax is borne by workers. In Finland, using variation around the financial crisis Korkeamäki and Uusitalo (2009) find that around 50 percent percent of the payroll tax is shifted onto workers. However, using parts of the same reform, a more recent study by Benzarti and Harju (2021a) finds no evidence that payroll taxes are shifted onto workers. This is consistent with findings in another Finnish paper by Benzarti and Harju (2021b) which uses firm-level variation, rather than regional variation. Our central estimates on contracted wages are roughly consistent with findings in the literature with point estimates ranging between around 20-50 percent, albeit with large standard errors that cannot rule out no shifting. However, we find significantly stronger evidence of shifting when considering the full wage payment which includes bonus and overtime payments, especially once we better control for heterogeneity through a DCF.

Using age-based variation in the payroll tax in Sweden Saez et al. (2019) finds that firms which employ young workers that face a lower payroll tax, increase the wage of all workers (both young and old) consistent with rent sharing. Given the regional nature of our reform, we cannot distinguish between the rent-sharing channel, and more traditional tax incidence. However, our heterogeneity analysis does appear to be more consistent

with rent sharing than with traditional tax incidence, in the sense that heterogeneity is mostly driven by firm-level variables, rather than worker-level variables. This is difficult to explain through traditional supply-demand channels, but easier to explain in a setting in which some firms are more prone to share rents with workers than others.[12]

With respect to heterogeneity, previous literature has considered heterogeneity in a large number of covariates such as worker-level variables of gender and education of the worker (e.g. Dale-Olsen, 2018; Saez et al., 2019; Stokke, 2021; Benzarti and Harju, 2021b, and firm-level balance-sheet variables (Saez et al., 2019; Benzarti and Harju, 2021b). Here we contribute to this literature by exploring heterogeneity with a data-driven approach. We establish that firm-level variables, particularly the size and composition of the workforce are more important determinants of payroll-tax incidence heterogeneity than workers' individual characteristics. Hence, heterogeneity in payroll-tax incidence between firms may be one of the mechanisms that drives firm-specific wage premiums.

5 Conclusion

In this paper we build on previous literature on causal machine learning. Athey and Wager (2019) develop causal forests as a way to derive data-driven heterogeneity estimates in a cross-sectional setting. We present the DCF method, which extends causal forests to a dynamic setting. Identification in Athey et al. (2019) is based on the strong assumption of random assignment, conditional on covariates. Instead, DCFs rely on the assumption of parallel trends. This allows for the DCFs to be implemented in research where difference-in-difference designs are applicable.

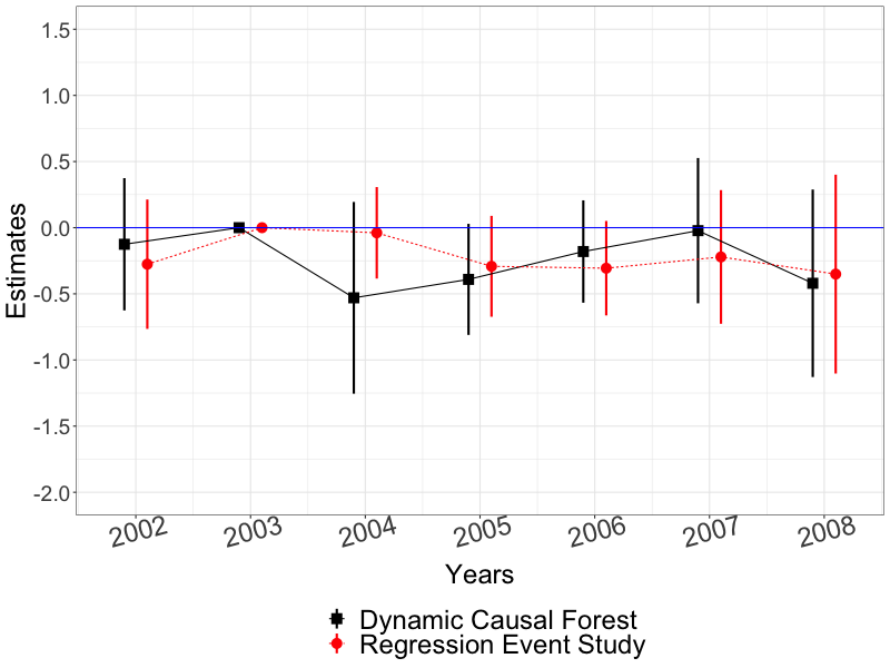
We demonstrate the DCF methodology through an application on the payroll tax in Norway. We find that incidence of the increase in the payroll tax is shifted onto employees through a reduction in bonus payments. We use the variable importance matrix to derive the margins along which we can observe the greatest heterogeneity. We estimate heterogeneous treatment effects along the number of employees, the firm gender ratio, firm unionization and other firm variables. We show an example of a double interaction, where we explore heterogeneity for firms of different sizes in terms of numbers of employees and different gender ratios. We demonstrate how the importance of these margins of heterogeneity develop over time. Our results imply that in the Norwegian (Scandinavian) institutional setting, firm-level variables and between-firm variation is

[12]Barth et al. (2020) also find evidence of rent sharing in a Norwegian setting, particularly in firms that exhibit a large degree of unionization.

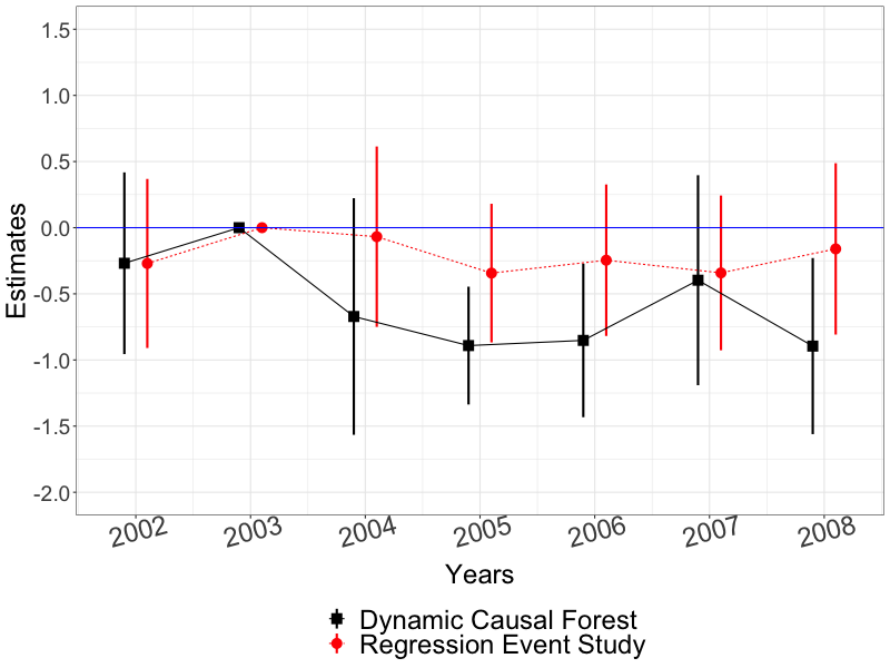
instrumental in understanding disparities in payroll tax incidence.

Our methodology has implications for the literature extending the use of difference-in-difference designs. In the Appendix we provide an extension of DCF to staggered difference-in-difference designs. A fruitful point for future research could be to extend the application of the DCF to shift-share designs (Adao et al., 2019; Goldsmith-Pinkham et al., 2020; Borusyak et al., 2022). Intuitively, it could be possible to difference out the base year in the shift-share variable and use it as a continuous treatment variable in a dynamic causal forest.

Figure 4: Event Study of the Effect of the Payroll Tax Reform on Different Wage Concepts



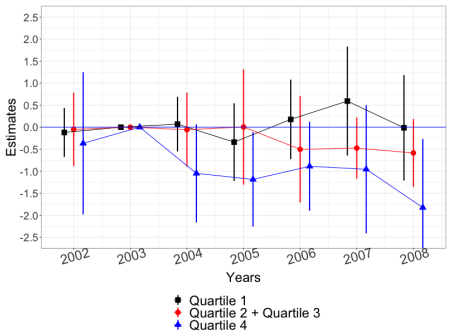
a. Monthly Contracted Wage



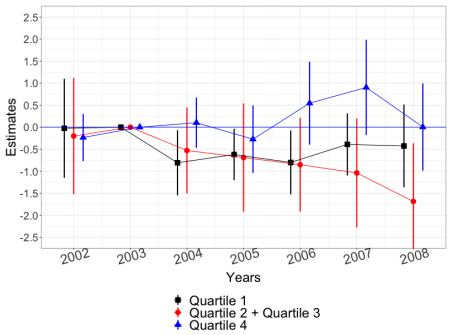
b. Monthly Full Wage

Notes: The Figure plots the outcome variable as listed in the caption, where Full Wage is the wage inclusive of overtime and bonus payment. In each plot we present estimates of a TWFE Regression Event Study and a DCF. Estimates are normalized by the change in the tax rate, such that a coefficient of 0 indicates no change in the wage rate (full incidince on the firm), and a coefficient of -1 indicates full incidence on the worker. Standard errors are clustered at the firm level.

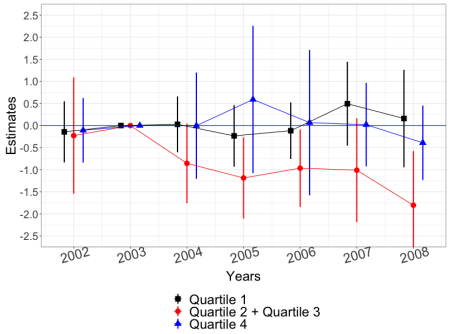
Figure 5: Heterogeneity Effects in Full Monthly Wage



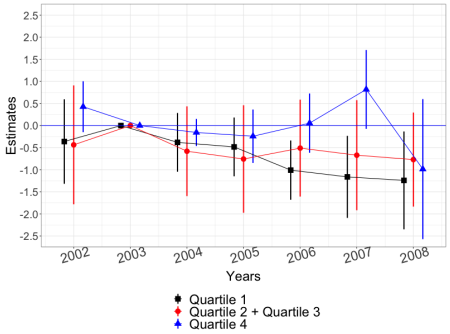
a. Firm Nr. Employees



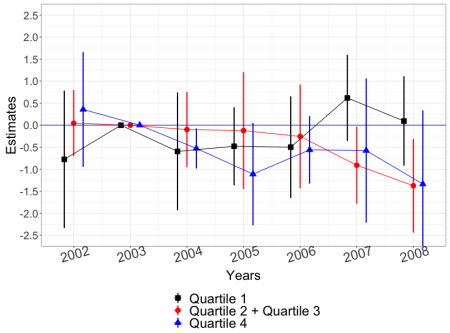
b. Firm Gender Ratio



c. Firm Unionization



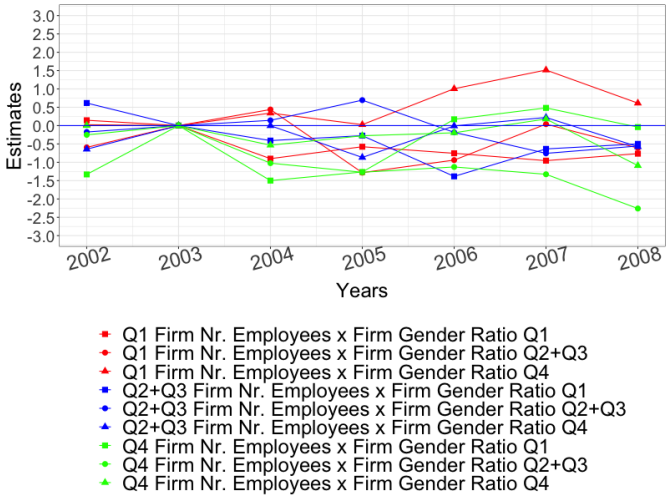
d. Firm HH with a Couple



e. Firm Paying Out Bonuses

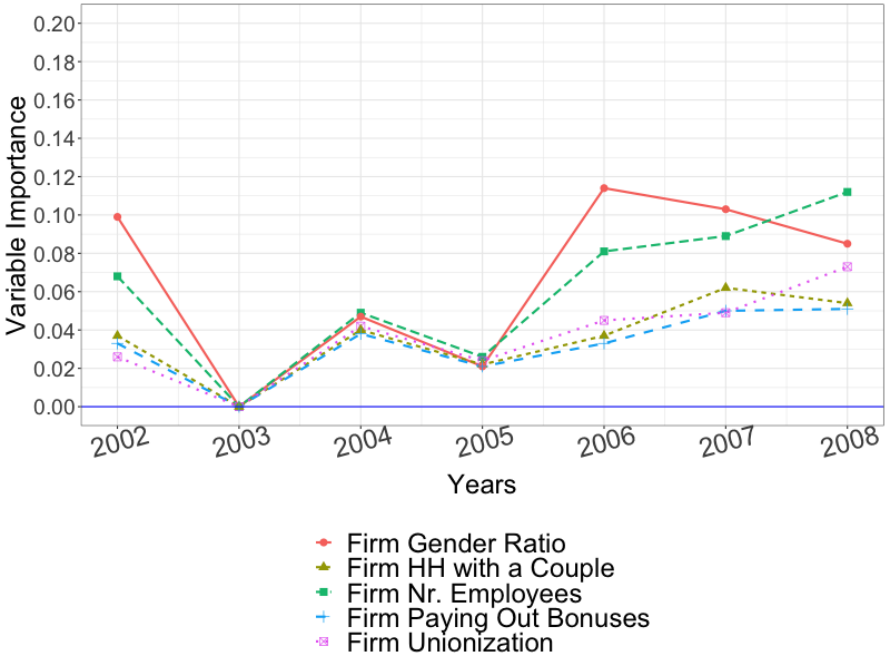
Notes: The Figure plots the heterogeneity in Full Monthly Wage by subsamples on the variable listed in the caption. In each plot we present estimates of a Dynamic Causal Forest. Estimates are normalized by the change in the tax rate, such that a coefficient of 0 indicates no change in the wage rate (full incidence on the firm), and a coefficient of -1 indicates full incidence on the worker. See Table 3 for reference values. Standard errors are clustered at the firm level.

Figure 6: Heterogeneity Effects in Full Monthly Wage - Interaction between Firm Nr. Employees and Firm Gender Ratio



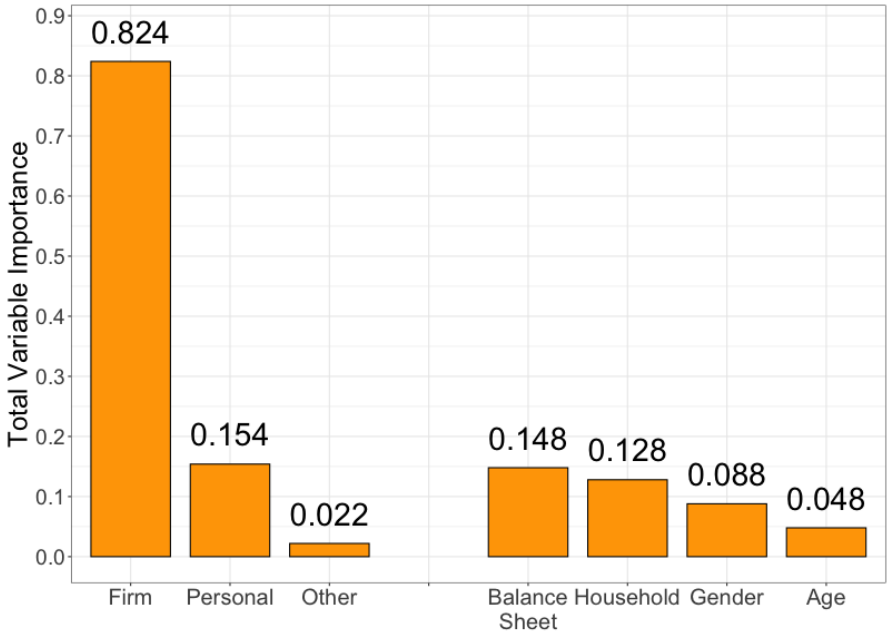
Notes: The Figure plots the outcome variable full monthly wage inclusive of over time and bonus payments in different subsets of the data. In each plot we present estimates of a Dynamic Causal Forest. Estimates are normalized by the change in the tax rate, such that a coefficient of 0 indicates no change in the wage rate (full incidence on the firm), and a coefficient of -1 indicates full incidence on the worker. See Table 3 for reference values. For firm employees: quartile one is colored in red, quartile two and three are in blue, quartile four is in green. For firm gender ratio: quartile one is with square nodes, quartile two and threee with circles, quartile four with triangles. Standard errors are clustered at the firm level.

Figure 7: Event Plot on Variable Importance



Notes: The Figure plots the outcome variable as listed in the caption. In each plot we present estimates of a Dynamic Causal Forest. Estimates are normalized by the change in the tax rate, such that a coefficient of 0 indicates no change in the wage rate (full incidence on the firm), and a coefficient of -1 indicates full incidence on the worker. Standard errors are clustered at the firm level.

Figure 8: Importance of Different Channels



Notes: The Figure plots the sum of variables in the variable importance matrix of the dynamic causal forest. For e.g. "Personal" sums the importance of all variables measured at the employee level. "Firm" is the sum of all variables measured at the firm level. "Other" is the residual category, which in this case includes only the measure of centrality at the municipality level. "Balance Sheet" sums all variables measured in the balance sheet of the company, except capital-labor ratio. "Household" sums the importance both employee and firm household indicators of household characteristics.

Table 1: Summary Statistics

	Zone 1		Zone 1 a	
Variable	Before	After	Before	After
Contracted Wage	26053.99	28584.93	24745.73	27128.371
	(8417.83)	(9624.3252)	(8509.85)	(9501.7471)
Full Wage	29045	31154.81	27216.9	29115.246
	(9486.17)	(10752.573)	(9284.96)	(10215.122)
Treatment	0	0	1	1
	(0)	(0)	(0)	(0)
Observations	15951	34885	2767	6031

Notes: Standard Deviations in parenthesis. Monthly Full Wage contains all bonus and overtime payments. The monthly wage is based on regular hours worked. The Equivalent Wage takes into account the percentage of employment and converts remuneration on hours worked into a full time equivalent wage. The Annual Taxable Income is based on data from the Tax Administration. Employees is the number of full-time employees per firm. Percent is the percentage of employment. Zone 1a is denoted as treated during the whole period 2002-2010. Before refers to the period before 2003 inclusive. After refers to the period after 2003. In the main analysis 2003 is the base year. The number of observations is 44 510 workers for all variables, except for employees where the unit of analysis is the firm - 2970 firm-year observations.

References

Abadie, Alberto (2005) 'Semiparametric difference-in-differences estimators.' *The Review of Economic Studies* 72(1), 1–19

Abowd, John M, Francis Kramarz, and David N Margolis (1999) 'High wage workers and high wage firms.' *Econometrica* 67(2), 251–333

Adao, Rodrigo, Michal Kolesár, and Eduardo Morales (2019) 'Shift-share designs: Theory and inference.' *The Quarterly Journal of Economics* 134(4), 1949–2010

Angrist, Joshua D, and Jörn-Steffen Pischke (2008) *Mostly harmless econometrics: An empiricist's companion* (Princeton university press)

Athey, Susan, and Guido Imbens (2016) 'Recursive partitioning for heterogeneous causal effects.' *Proceedings of the National Academy of Sciences* 113(27), 7353–7360

Athey, Susan, and Stefan Wager (2019) 'Estimating treatment effects with causal forests: An application.' *Observational Studies* 5(2), 37–51

Table 2: Control Variables in the Causal Forest

Variable	Employee level	Firm level
Wage	Yes	Yes
Employee Age	Yes	Yes
Employee Gender	Yes	Yes
Nr. Adults in the Household	Yes	Yes
Nr. Children in the Household	Yes	Yes
Couple in the Household	Yes	Yes
Social Assistance	Yes	Yes
Employee Education	Yes	Yes
Employee Union Membership	Yes	Yes
Nr. Employees	No	Yes
Capital-Labor Ratio	No	Yes
Earned Capital	No	Yes
Liquidity and Cash holdings	No	Yes
ROA	No	Yes
Investment	No	Yes
Dividents	No	Yes
NACE Industry Classification	No	Yes

Notes: All control variables are taken in base year values from 2003. "Yes" at the Firm level denotes that the individual variable has been aggregated at the firm level and used as an input in the analysis. Additional variables in the analysis include dummy variables tagging workers who have switched jobs before the base year, workers that have received bonuses before the base year, firm dividends for 2002 and for 2003 and a centrality index at the municipality level.

Table 3: Variable Importance Matrices for Causal Forests with Dependent Variable Full Wage

Panel A. Variable Importance

	Dynamic CF
Firm Nr. Employees	0.112
Firm Gender Ratio	0.085
Firm Unionization	0.073
Firm HH with a Couple	0.054
Firm Paying Out Bonuses	0.051

Panel B. Quartiles

	Quartile 1	Quartile 2 & 3	Quartile 4
Firm Nr Employees	31.542	215.393	5,291.606
Firm Gender Ratio	0.040	0.276	0.742
Firm Union	0.374	0.883	0.973
Firm HH with a Couple	0.536	0.741	0.893
Firm Paying Out Bonuses	0.495	0.878	0.990

Notes: CF stands for Causal Forest. All variables are taken in the pre-reform period 2002-2003. Firm Capital-Labor Ratio is defined as assets on the balance sheet (denominated in 1000s NOK) divided by firm employees. Firm Employees are logged. Firm HH Married stands for fraction of employees that belong to a household with a partnership, within a given firm.

Athey, Susan, Julie Tibshirani, Stefan Wager et al. (2019) 'Generalized random forests.' *The Annals of Statistics* 47(2), 1148–1178

Barth, Erling, Alex Bryson, Harald Dale-Olsen et al. (2020) 'Union density effects on productivity and wages.' *Economic Journal* 130(631), 1898–1936

Bennmarker, Helge, Erik Mellander, and Björn Öckert (2009) 'Do regional payroll tax reductions boost employment?' *Labour Economics* 16(5), 480–489

Benzarti, Youssef, and Jarkko Harju (2021a) 'Can payroll tax cuts help firms during recessions?' *Journal of Public Economics* 200, 104472

_ (2021b) 'Using payroll tax variation to unpack the black box of firm-level production.' *Journal of the European Economic Association* 19(5), 2737–2764

Borusyak, Kirill, Peter Hull, and Xavier Jaravel (2022) 'Quasi-experimental shift-share research designs.' *The Review of Economic Studies* 89(1), 181–213

Borusyak, Kirill, Xavier Jaravel, and Jann Spiess (2021) 'Revisiting event study designs: Robust and efficient estimation.' arXiv preprint arXiv:2108.12419

Bozio, Antoine, Thomas Breda, and Malka Guillot (2020) 'The contribution of payroll taxation to wage inequality in france.' IZA Discussion Paper No 13317

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

Brodeur, Abel, Mathias Lé, Marc Sangnier, and Yanos Zylberberg (2016) 'Star wars: The empirics strike back.' *American Economic Journal: Applied Economics* 8(1), 1–32

Callaway, Brantly, and Pedro HC Sant'Anna (2021) 'Difference-in-differences with multiple time periods.' *Journal of Econometrics* 225(2), 200–230

Card, David, Ana Rute Cardoso, Joerg Heining, and Patrick Kline (2018) 'Firms and labor market inequality: Evidence and some theory.' *Journal of Labor Economics* 36(S1), S13–S70

Card, David, Jörg Heining, and Patrick Kline (2013) 'Workplace heterogeneity and the rise of west german wage inequality.' *The Quarterly journal of economics* 128(3), 967–1015

Chang, Neng-Chieh (2020) 'Double/debiased machine learning for difference-in-differences models.' *The Econometrics Journal* 23(2), 177–191

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins (2018) 'Double/debiased machine learning for treatment and structural parameters: Double/debiased machine learning.' *The Econometrics Journal*

Chernozhukov, Victor, Mert Demirer, Esther Duflo, and Ivan Fernandez-Val (2020) 'Generic machine learning inference on heterogeneous treatment effects in randomized experiments, with an application to immunization in india.' NBER Working Paper No. 24678

Dale-Olsen, Harald (2018) 'Labour demand and supply changes in norway following an imposed harmonization of geographically differentiated payroll-tax rates.' *Labour* 32(2), 261–291

Goldsmith-Pinkham, Paul, Isaac Sorkin, and Henry Swift (2020) 'Bartik instruments: What, when, why, and how.' *American Economic Review* 110(8), 2586–2624

Goodman-Bacon, Andrew (2021) 'Difference-in-differences with variation in treatment timing.' *Journal of Econometrics* 225(2), 254–277

Gruber, Jonathan (1997) 'The incidence of payroll taxation: Evidence from chile.' *Journal of Labor Economics* 15(S3), S72–S101

Gulen, Huseyin, Candace Jens, and T Beau Page (2020) 'An application of causal forest in corporate finance: How does financing affect investment?'

Korkeamäki, Ossi, and Roope Uusitalo (2009) 'Employment and wage effects of a payroll-tax cut-evidence from a regional experiment.' *International Tax and Public Finance* 16(6), 753–772

Ku, Hyejin, Uta Schönberg, and Ragnhild C Schreiner (2020) 'Do place-based tax incentives create jobs?' *Journal of Public Economics* 191, 104105

Lee, Sokbae, Ryo Okui, and Yoon-Jae Whang (2017) 'Doubly robust uniform confidence band for the conditional average treatment effect function.' *Journal of Applied Econometrics* 32(7), 1207–1225

Miller, Steve (2020) 'Causal forest estimation of heterogeneous and time-varying environmental policy effects.' *Journal of Environmental Economics and Management* 103, 102337

Roth, Jonathan, and Pedro HC Sant'Anna (2020) 'When is parallel trends sensitive to functional form?' *arXiv preprint arXiv:2010.04814*

Roth, Jonathan, Pedro HC Sant'Anna, Alyssa Bilinski, and John Poe (2022) 'What's trending in difference-in-differences? a synthesis of the recent econometrics literature.' arXiv preprint arXiv:2201.01194

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

Saez, Emmanuel, Benjamin Schoefer, and David Seim (2019) 'Payroll taxes, firm behavior, and rent sharing: Evidence from a young workers' tax cut in sweden.' *American Economic Review* 109(5), 1717–63

Saez, Emmanuel, Manos Matsaganis, and Panos Tsakloglou (2012) 'Earnings determination and taxes: Evidence from a cohort-based payroll tax reform in Greece.' *Quarterly Journal of Economics*

Sant'Anna, Pedro HC, and Jun Zhao (2020) 'Doubly robust difference-in-differences estimators.' *Journal of Econometrics* 219(1), 101–122

Schmidheiny, Kurt, and Sebastian Siegloch (2019) 'On event study designs and distributed-lag models: Equivalence, generalization and practical implications.' CESifo Working Paper No. 7481 Munich

Stokke, Hildegunn E (2021) 'Regional payroll tax cuts and individual wages: heterogeneous effects of worker ability and firm productivity.' *International Tax and Public Finance* 28(6), 1360–1384

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

Wooldridge, Jeffrey M (2021) 'Two-way fixed effects, the two-way mundlak regression, and difference-in-differences estimators.' SSRN Working Paper 3906345

Appendix For Online Publication

A Extension to staggered treatment

In the main text we introduce the DCF in the context of simultaneous adoption of treatment in period $b+1$. However, in many relevant cases researchers have access to data where the period of adoption varies by unit, which is called staggered treatment. Here we sketch one approach that extends the DCF method to staggered settings. For simplicity, we assume that treatment is an absorbing state. Then we may divide the data into treatment cohorts, and separately estimate the treatment effect for each. Thus, we are interested in the CATT within subgroups that are partitioned by cohort and period. The key assumption we make is that the data contains never-treated units. The approach we outline below only makes comparisons between treated units and never-treated units.

Let superscript j denote the treatment cohort. For instance, $j=1$ can denote the cohort that is treated first, and so on. Never-treated units are denoted $j=\infty$. The outcome variable is thus expressed by y_{igt}^{j} and covariates by x_{ig}^{j}. As in the standard DCF method $W_i^j = 1$ for all treated units ($j < \infty$), independent of when they are treated.

We first estimate DCFs by cohort and period. The inputs to DCF j are as follows. The vector of outcome variables is given by $[y^j, y^\infty]$, where y^j denotes the vector of outcome variables for treatment cohort j. Correspondingly, the treatment variable is given by $[W^j, W^\infty]$, and the covariates by the matrix $[X^j, X^\infty]$. The cohort-specific base period b^j is the last pre-treatment period before adoption of treatment by cohort j.

The result of such a DCF is an estimate of the CATT by period and cohort, $\hat{\tau}_t^j(x)$. To make the estimates comparable between cohorts, these estimates are converted to treatment event time h, where $h=0$ corresponds to period b^j+1 when treatment is turned on for cohort j. We can do this by applying the transformation:

$$\hat{\tau}_h^j(x) = \hat{\tau}_{t-b^j-1}^j(x).$$

In event studies, a common estimation target is the average treatment effect h periods since treatment for a given horizon $h \geq 0$. Denote the estimate of the average horizon-specific CATT by $\hat{\tau}_h(x)$, which is interpreted as the average dynamic treatment effect at event time h. To estimate this parameter, the researcher can take a (weighted) cross-cohort average of $\hat{\tau}_h^j(x)$. See Callaway and Sant'Anna, 2021; Goodman-Bacon, 2021 for further discussion of appropriate weights and inference in this setting.

Note that cohorts which are treated late provide no identification for $\hat{\tau}_h(x)$ when h is large. In some settings it may therefore be appropriate to only study $\hat{\tau}_h(x)$ in a range of event time where multiple treatment cohorts provide identification. Alternatively, to purge the estimand for compositional differences, one may consider the average treatment effect at horizon h only for the subset of units that is also observed at horizon h'.

B Control Variables

Measurement of covariates in our study is based on the following administrative registers: Wage register, matched employer-employee register, tax and social security income register, family and household register, education register, firm balance sheets and firm unit register. All covariates are measured in the base years 2002 and 2003. For individual and household characteristics, we also include average values (mean or median) at the firm level. See for reference Table 2. The register data includes unique identifiers for individual, family, household, firm, plant and the municipalities where residents and firms are located. Here, we provide an overview of covariates and their definitions.

Individual and household characteristics of employees:

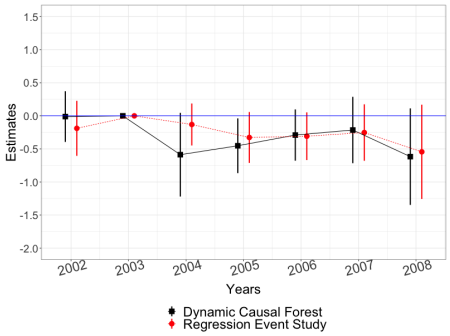
- Wage: Average monthly wage in the pre-reform years
- Age : Age in the year 2003
- Gender: Indicator for female gender
- Nr. Adults in the Household: Number of household members aged 18 or above
- Nr. of Children in the Household: Number of household members aged below 18
- Couple in the Household: Presence of a married or cohabiting couple in household
- Social assistance recipient: Indicator for individuals who are recipients of social assistance benefits in the social security register
- Educational attainment: Years of schooling based on educational attainment in the education register data
- Labor union membership: Indicator for employees with tax reported deduction for membership fee in labor union
- Indicator for having switched the job in the pre-reform years
- Indicator for having received bonuses in the pre-reform years
- Percentage employment: denotes at how many percent a person is employes on average

Firm characteristics of employer:

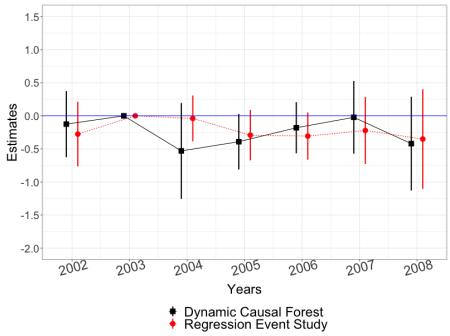
- Wage: Average monthly wage in the pre-reform years

- Age: Median age of employees
- Social Assistance: Median recipient status for social assistance of all workers
- Education variables at the firm level: Median of employee education
- Household variables at the firm level: Mode of individual variables
- Dividents disbursed in 2002 and 2003, included as separate variables
- Number of employees: Number of employees in the firm
- Capital - labor ratio: Total assets reported in firm balance sheet divided by number of employees
- Percent earned capital: Accumulated retained earnings (undistributed profits) in percent of total assets
- Percent liquid assets: Holdings of cash and liquid assets in percent of total assets
- Return on assets (ROA): After-tax profits in percent of total assets
- Investment: Total investments in percent of total assets
- NACE Industry Classification: Code in the NACE-classification of industries (Statistical Classification of Economic Activities in the European Community) at the 1 digit level
- Centrality Index: denotes the degree of centrality of the municipality

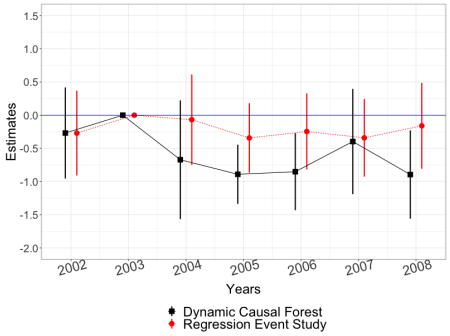
Figure 9: Event Study of the Effect of the Payroll Tax Reform on Different Wage Concepts



a. Monthly Contracted Wage (FTE)



b. Monthly Contracted Wage



c. Monthly Full Wage

Notes: The Figure plots the outcome variable as listed in the caption, where FTE refers to the Full-time Equivalent Wage, and Full Wage is the wage inclusive of overtime and bonus payment. In each plot we present estimates of a TWFE regression event study and a DCF. Estimates are normalized by the change in the tax rate, such that a coefficient of 0 indicates no change in the wage rate (full incidence on the firm), and a coefficient of -1 indicates full incidence on the worker. Standard errors are clustered at the firm level.

C Additional Results