EVALUATING DIFFERENCE-IN-DIFFERENCES MODELS UNDER DIFFERENT TREATMENT ASSIGNMENT MECHANISMS AND IN THE PRESENCE OF SPILLOVER EFFECTS

Guilherme Araújo Lima[1*], Igor Viveiros Melo Souza[2] and Mauro Sayar Ferreira[2]

[1*]Fundação Getúlio Vargas (FGV/EPGE), Rio de Janeiro, Brazil.
[2]Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, Brazil.

*Corresponding author(s). E-mail(s): guilherme.a.lima@hotmail.com;
Contributing authors: igorviveiros@gmail.com;
mferreira@cedeplar.ufmg.br;

Abstract

We conduct Monte Carlo experiments to evaluate the performance of different Difference-in-Differences estimators under treatment assignment mechanisms affected by shocks suffered by treated units and also in contexts where the treatment effect spills over to units in the control group. In particular, we compare the estimators proposed by Callaway and Sant'Anna (2021), Borusyak et al. (2021), and Sun and Abraham (2021), as well as the two-way fixed effects (TWFE) estimator. The results demonstrate that the treatment assignment mechanisms we design, and the presence of spillover effects can severely compromise the performance of the considered estimators, leading to bias and, even more importantly, inconsistency. Therefore, cautious for interpreting the results should be taken in applications where the environment studied resembles those we consider. The development of more robust estimators is a necessity and a prosperous research venue.

Keywords: Difference-in-Differences, Causal Inference, Treatment assignment mechanisms, Spillover effects

1 Introduction

Estimating the causal effect of a policy or intervention is both a major theme in Applied Economics and one of the greatest methodological challenges in Econometrics. In recent decades, in parallel to an increase in the amount of data available to researchers, there has been a significant increase in the quantity and quality of methods capable of estimating causal relationships. Angrist and Pischke (2010) called this phenomenon "The Credibility Revolution" within the field of Econometrics.

The Difference-in-Differences estimator is one of the main methods used to estimate causal effects with panel data. In the canonical version of the model, the researcher has data on two groups of units in two distinct periods in time. Between the first and second periods, one of the groups (the treatment group) is exposed to some intervention, while the other group (the control group) does not undergo any change during this time window. Under some assumptions, the Difference-in-Differences estimator provides a consistent way to construct a counterfactual scenario for the treated units. With this, one is able to estimate the average treatment effect among treated units (ATT), which is the parameter of interest in several empirical studies.

However, by construction, this canonical model is not able to handle some data structures that often appear in practical applications. For instance, in several occasions there are data for more than two periods, more than two groups of units, and there is still the possibility that the groups are treated at different periods. These circumstances generate additional challenges for estimation, since the assumptions under the canonical model and the estimation methods are particular to the context in which there are only two groups observed in only two periods.

As discussed by Roth et al. (2022) and de Chaisemartin and D'Haultfoeuille (2022), the traditional alternative that has emerged to deal with this more complex data structure consists of regressing the variable of interest against time and unit fixed effects, plus a binary variable indicating the treatment status. This strategy is known in the literature as two-way fixed effects regression (TWFE).

In recent years, however, several papers identified serious problems with the TWFE estimator, such as the fact that it may not estimate a convex combination of the average treatment effects among the treated subgroups of the sample, with the assigned weights possibly being negative (DE CHAISEMARTIN; D'HAULTFOEUILLE, 2020; SUN; ABRAHAM, 2021; GOODMAN-BACON, 2021). These articles conclude that this type problem arises in contexts where there is heterogeneity in the treatment effect over time or between groups of treated units. As an alternative, more robust estimators have been developed to incorporate heterogeneity in treatment effects (DE CHAISEMARTIN; D'HAULTFOEUILLE, 2020; SUN; ABRAHAM, 2021; CALLAWAY; SANT'ANNA, 2021; BORUSYAK et al., 2021; ATHEY; IMBENS, 2022; GARDNER, 2022).

In the current paper we investigate the properties of these estimators in environments that distances from two hypothesis under which they were developed. First, we consider a situation where the treatment assignment mechanism is a function of the shocks suffered by the units. An illustrative situation is a labor training program in which a negative shock to income determines whether a worker can receive the treatment. We also consider an environment where the Stable Unit Treatment Value Assumption

(SUTVA) is violated. In particular, we allow for spillovers of the treatment received by treated units to control units, which is a situatin commonly verified. For example, it is reasonable to expect that a credit policy directed to a specific municipality affects other municipalities close to it.
We verify the properties of the estimators developed by Callaway and Sant'Anna (2021), Sun and Abraham (2021) and Borusyak et al. (2021) by conducting two separate Monte Carlo experiments based on artificially generated datasets containing the treatment assignment mechanisms we intend to study. In particular, we compare the bias and the mean squared error of the estimators.
The remainder of the paper is organized as follows. In the next section we present a brief literature review. In section 3 we discuss details regarding the simulations whose results are presented in section 4. Section 5 concludes.

2 Literature Review

This section briefly reviews the Difference-in-Differences literature, focusing on the topics most related to the objective of this paper. Subsection 2.1 discusses the canonical Difference-in-Differences model developed for the case in which there are two groups observed in two periods, and where one of the groups suffers an intervention whose effect is to be estimated. Subsection 2.2 discusses the two-way fixed effects regression (TWFE), while subsection 2.3 summarizes the main results obtained by the most recent Difference-in-Differences literature on the problems to which the TWFE model is subject when the hypothesis of homogeneity of the treatment effect is absent.

2.1 The canonical Difference-in-Differences estimator

The canonical Difference-in-Differences model design consists of a set of units (indexed by $i = 1, 2, \ldots, N$) and two periods ($t = 1, 2$). The researcher wants to estimate the causal effect of some intervention between the two periods on an observed variable Y. We assign $D_i = 1$ for units treated between $t = 1$ and $t = 2$, and $D_i = 0$ for those in the control group. The researcher observes a panel with the variable of interest Y_{it} and with the binary variable indicating the treatment status D_i for all units in both periods. Let $Y_{it}(0)$ be unit $i's$ potential outcome in period t if it remains in the control group and $Y_{it}(1)$ if it is treated between the two time points. Due to the "fundamental problem of causal inference" described by Holland (1986), only one potential outcome per unit is observed. A more concise way of saying this is to write the realization of the variable Y for each unit at each period as

$$Y_{it} = D_i Y_{it}(1) + (1 - D_i) Y_{it}(0)$$

In this case, the parameter to be estimated is the average treatment effect on the treated (ATT) in period 2, that is,

$$\tau_2 = E[Y_{i2}(1) - Y_{i2}(0) | D_i = 1]$$

For the identification of this parameter, two main assumptions are made. The first is the absence of anticipation to the treatment, which basically means that the

intervention has no impact on units that will eventually be treated before it is actually implemented. This is the same as writing $Y_{i1}(1) = Y_{i1}(0)$ for every unit i in the treatment group. The second assumption is that of parallel trends, which requires that

$$E[Y_{i2}(0) - Y_{i1}(0)|D_i = 1] = E[Y_{i2}(0) - Y_{i1}(0)|D_i = 0]$$

This condition says that the evolution of the variable of interest between the two groups (control and treatment) should be the same in the absence of intervention. The parallel trends assumption is not directly testable, since $Y_{i2}(0)$ is not observed for units with $D_i = 1$. With these two assumptions it is possible to estimate τ_2, since

$$\tau_2 = E[Y_{i2}(1) - Y_{i2}(0)|D_i = 1] = E[Y_{i2}(1) - Y_{i1}(1) + Y_{i1}(1) - Y_{i2}(0)|D_i = 1] =$$

$$\begin{aligned}
&= E[Y_{i2}(1) - Y_{i1}(0)|D_i = 1] - E[Y_{i2}(0) - Y_{i1}(1)|D_i = 1] = \\
&= E[Y_{i2}(1) - Y_{i1}(0)|D_i = 1] - E[Y_{i2}(0) - Y_{i1}(0)|D_i = 1] = \\
&= E[Y_{i2}(1) - Y_{i1}(0)|D_i = 1] - E[Y_{i2}(0) - Y_{i1}(0)|D_i = 0] = \\
&= E[Y_{i2} - Y_{i1}|D_i = 1] - E[Y_{i2} - Y_{i1}|D_i = 0]
\end{aligned}$$

The latter two terms are easily estimated using their sample analog:

$$\hat{\tau}_2 = (\bar{Y}_{t=2,D=1} - \bar{Y}_{t=1,D=1}) - (\bar{Y}_{t=2,D=0} - \bar{Y}_{t=1,D=0})$$

In practice, however, it is often more convenient to estimate τ_2 by regression to facilitate inference. In this case, the estimator is the β coefficient of the following regression:

$$Y_{it} = \alpha + D_i\gamma + \mathbf{1}\{t = 2\}\gamma + (D_i \times \mathbf{1}\{t = 2\})\beta + \varepsilon_{it}$$

If D_i is orthogonal to ε_{it}, it is possible to show that this estimator of τ_2 is unbiased (WOOLDRIDGE, 2010). With a few more technical conditions, the OLS estimator of β is consistent and asymptotically normally distributed. Roth et al. (2022) discuss in more detail the inference procedure for this context.

2.2 The TWFE estimator

In several cases of interest, the dataset available to the researcher contains more than two periods and the units in the sample are treated in multiple distinct periods. In such contexts, using the panel data structure to estimate the causal effect of treatment is considerably more complicated than in the canonical Difference-in-Differences estimator.

The first step required to analyze the more general context is to modify the model framework and the required assumptions. Roth et al. (2022) describe the most common framework in the literature, which we reproduce here. There are T periods (indexed by $t = 1, 2, \ldots, T$) and treatment is considered an absorbing state, i.e., units treated at some point in the sample do not revert to the control group later. This type of treatment is known in the literature as *staggered*. As in the canonical case, we consider a variable D_{it} to indicate whether unit i is in the treatment group at time t, and

a variable G_i to denote the first period in which unit i enters the treatment group. If unit i remains in the control group throughout the sample, we consider $G_i = \infty$ with some abuse of notation. The fact that the treatment is staggered implies that the sequence of potential outcomes for each unit can be completely identified by the first period in which the treatment is implemented for it. Thus, for a given unit i, if $G_i = g$, its potential outcome at t is denoted by $Y_{it}(g)$. Similarly, assuming $G_i = \infty$, the potential outcome is $Y_{it}(\infty)$.
The realizations of variable Y are related to the potential outcomes by means of the following equality:

$$Y_{it} = Y_{it}(\infty) + \sum_{g=2}^{T} (Y_{it}(g) - Y_{it}(\infty))\mathbf{1}\{G_i = g\}$$

The natural extension (which is not the only one in the literature) of the parallel trends assumption to this general setup is to assume that for any $t \neq t^{'}$ and $g \neq g^{'}$, it holds that

$$E[Y_{it}(\infty) - Y_{it'}(\infty)|G_i = g] = E[Y_{it}(\infty) - Y_{it'}(\infty)|G_i = g^{'}]$$

Similarly, a natural extension of the no anticipation assumption is to assume that $Y_{it}(g) = Y_{it}(\infty)$ for any i, g and $t < g$.

Traditionally, as discussed by Roth et al. (2022), the TWFE regression is the main method used to estimate the causal effect of a treatment on the treated units in these cases, which consists of a natural extension of the regression used in the canonical case: the variable of interest is regressed against unit and time fixed effects and against a set of binary variables related to the treatment status for each unit at a given time. de Chaisemartin and D'Haultfoeuille (2020), from a survey of all applied papers published in the American Economic Review between 2010 and 2012, conclude that in about 20% of them a TWFE regression was used to estimate the effect of some treatment on a variable. Along the same line, de Chaisemartin and D'Haultfoeuille (2022) show that 26 of the 100 most cited articles in the American Economic Review between 2015 and 2019 estimate some TWFE-type regression. These two surveys highlight the popularity of TWFE regression for estimating causal effects.

Sun and Abraham (2021) divide TWFE regression into two cases. The first is the so-called static TWFE and has the following specification:

$$Y_{it} = \alpha_i + \lambda_t + \mu D_{it} + \varepsilon_{it}$$

where α_i is a unit fixed effect, λ_t is a time fixed effect and D_{it} indicates whether unit i was being treated in period t or not. In this case, μ is usually interpreted as being an estimator of the causal effect of the treatment on the treated.

The dynamic specification of TWFE is broader than the static one. To fix the notation, let E_i be the instant at which unit i enters the treatment group and K and G be positive integers chosen by the researcher. In this case, one estimates a regression with a specification similar to the following:

$$Y_{it} = \alpha_i + \lambda_t + \sum_{l=-K}^{-2} \mu_l \mathbf{1}\{t - E_i = l\} + \sum_{g=0}^{G} \mu_g \mathbf{1}\{t - E_i = g\} + \varepsilon_{it}$$

The coefficients μ_l are then used to test the parallel trends assumption, while μ_g is used to analyze the dynamic evolution of the average treatment effect across treated units.

2.3 The most recent Difference-in-Differences literature

For a long time, it was assumed, without an adequate theoretical basis, that the TWFE estimator worked in a similar way to the canonical Difference-in-Differences estimator. More recently however, several problems raised have served to contradict this thesis. de Chaisemartin and D'Haultfoeuille (2020) show, for the case of static TWFE, that the coefficient associated with the variable indicating the treatment status, which is commonly interpreted as a measure of the causal effect of the intervention, can be written as a weighted average of the average treatment effect for subgroups of the sample at that time, but the weights can be negative. In the limit, for example, it may happen that the average effect is positive for each subgroup, but the overall coefficient is negative. Sun and Abraham (2021) show, for the dynamic version of the TWFE estimator, that the coefficients μ_l and μ_g are contaminated by treatment effects associated with other periods relative to the beginning of the intervention, affecting the usual interpretation of these coefficients, specially the practice of using μ_l to test for parallel trends. Goodman-Bacon (2021), using another decomposition, shows that the TWFE estimator is a weighted average of all possible two-by-two Difference-in-Differences combinations between groups that change their treatment status and groups whose status remains constant. From this decomposition, the author explicitly shows that the TWFE estimator estimates a weighted average of coefficients related to the treatment effect for subgroups of the sample, but again the weights may be negative, so the combination of these parameters may not be convex. Negative weights are problematic in this context as they compromise the causal interpretation of the estimates obtained. In all these papers, the common factor explaining this problem was found to be heterogeneity in the treatment effect over time or between subgroups of treated units. As summarized by de Chaisemartin and D'Haultfoeuille (2022), the TWFE estimator requires an additional assumption (compared to the traditional Difference-in-Differences model) to estimate the average treatment effect in a non-biased way, namely that the impact of the intervention is constant over time and across units.

While the derivation of these results is complicated, their intuition is not difficult to rationalize. The specification of TWFE regressions, both static and dynamic, implicitly assumes the extent to which the treatment is allowed to vary along i or t. Taking the static specification as an example, the coefficient μ, which estimates the ATT, is by assumed to be the same for all units in all periods, so it is intuitive to think that systematic heterogeneities in the treatment effect pose a problem for making inference with this specification. Roth et al. (2022) and de Chaisemartin and D'Haultfoeuille (2022) discuss in more depth the main results found by this most recent literature.

As problems with the TWFE estimator have been identified, alternative estimators have been developed intending to increase robustness to contexts in which the treatment effect is heterogeneous over time or across units (DE CHAISEMARTIN; D'HAULTFOEUILLE, 2020; SUN; ABRAHAM, 2021; CALLAWAY; SANT'ANNA, 2021; BORUSYAK et al, 2021; LIU et al., 2021; ATHEY; IMBENS, 2022; GARDNER, 2022). The differences between them regard the hypotheses assumed, the contexts in which they should be applied, and the specific parameter they seek to estimate. The rest of the paper is dedicated to evaluating the performance of these new estimators when exposed to situations we may observe in real appliactions but are not considered among the hypotheses behind their development.

3 Simulations

We conduct two exercises, each considering a different type of problem one may face while working with real data. The first focuses on evaluating how treatment assignment mechanisms based on the shocks suffered by the units affect the estimates of the treatment effect. The second seeks to ascertain how the estimates are affected when the treatment effect on the treated units spills over to those in the control group.
The estimators proposed by Sun and Abraham (2021), Callaway and Sant'Anna (2021) and Borusyak et al. (2021) are evaluated, in addition to the TWFE estimator. This list does not exhaust all the estimators recently proposed, but, besides relying on similar assumptions, they have been the most adopted when it comes to estimate dynamic treatment effects (DE CHAISEMARTIN; D'HAULTFOEUILLE, 2022; ROTH et al., 2022).
In the case of the estimator proposed by Callaway and Sant'Anna (2021), two estimates are considered: one using as a control group the units that at a given time have not yet been treated, but eventually will be (the not yet treated units), and another considering the units that have not been treated at any time (the never treated units). The estimation method chosen was the doubly robust, which is the standard of the *did* package in *R*, developed by the same authors of the mentioned article.
Six different data structures are considered, four for the first exercise and two for the second. For each structure, panels with 50, 250 and 500 individuals are simulated, covering 20 periods. For each panel the treatment effect on treated units is estimated with the five estimators considered. For each combination of data structure and dimension, 1000 panels are generated, allowing us to estimate the same number of treatment effect. Empirical distribution of the estimates are then constructed, and the following descriptive statistics computed: mean, median, 2.5th and 97.5th percentile, and the mean squared error (MSE).
The panel size (i.e., the number of individuals multiplied by the number of periods) is denoted by NT in the remainder of the paper. Thus, we simulate datasets with NT equal to 1000, 5000 and 10000, allowing to assess the asymptotic properties of each estimator. The asymptotic analysis considers a fixed $T = 20$ but varies the number of units, following the tradition in the panel data literature.
On the simulated datasets, the treatment effect on treated units is assumed to be homogeneous, both across units and across periods relative to the start of treatment.

This is mainly for two reasons. The first is to focus on the specific problems we want to analyze. Under homogeneity of treatment effect, any convex combination of the treatment effect across subgroups of the sample will be equal to each other. In particular, any global parameter associated with the treatment effect will have the same value, which makes it easier to interpret and report the results. Secondly, the properties of the new DID estimators and the problems of TWFE regression when treatment effects are heterogeneous are already well documented in the literature.
In all simulated panels, the treatment will be staggered. In addition, without any loss of generality regarding the results obtained, it is always assumed that the treatment effect on the treated units is positive, also in order to facilitate interpretations. In simulations with spillover effect, we assume the indirect effect to positively affect units in the control group.

The simulations should begin with a specification for the potential outcome of the units in the absence of treatment (i.e., when $D_{i,t} = 0$). The general specification chosen has a form similar to

$$Y_{i,t}(0, D_{-i,t}) = Y_{i,t}(D_{i,t} = 0, D_{-i,t}) = \alpha_i + \lambda_t + \phi(D_{-i,t}) + \varepsilon_{it}$$

with $\varepsilon_{it} \sim N(0, 1)$. The coefficients α_i and λ_t represent unit and time fixed effects respectively. The term $D_{-i,t}$ is a vector with the treatment status of all units in period t except the i-th. The map ϕ, which assigns to each $D_{-i,t}$ a non-negative real number, represents the spillover effect. In simulations focused on selection mechanism, ϕ is the null function since spillover effects are absent. It is further assumed that $\phi(0) = 0$, meaning that there is no spillover without treatment of any unit. The α_i are drawn from a uniform distribution over the interval $[80, 120]$ and $\lambda_t = t$. These are arbitrary choices that do not affect the validity of the results since they do not interfere in the parallel trends assumption. This specification for the potential outcome is standard in the literature (except for the term associated with the spillover effect) and coincides with the one used by Borusyak et al. (2021) in their simulation exercises.

Considering first the case with no spillover effect, one can formally define the treatment effect at period t on unit i (treated at t) as

$$\tau_{i,t} = Y_{i,t}(D_{i,t} = 1) - Y_{i,t}(D_{i,t} = 0) = Y_{i,t} - Y_{i,t}(D_{i,t} = 0) = Y_{i,t} - Y_{i,t}(0)$$

That is, the observed value is compared with the counterfactual scenario in which the unit had not been treated at t. In the presence of spillover, the treatment effect at time t on unit i (treated at t) is analogously represented by

$$\tau_{i,t} = Y_{i,t}(D_{i,t} = 1, D_{-i,t}) - Y_{i,t}(D_{i,t} = 0, 0)$$

Now the comparison is made between the value actually observed for the unit and what would have been observed in the absence of treatment in any unit (so there is no spillover).

Although the binary variable D_{it} is observed by the econometrician, one may think of situations where D_{it} depends on other variables, such as α_i or ε_{it}. In the first set of simulations the relationship with ε_{it} is explored. As an example, one can think of measuring the impact of government labor training programs, which however tend to have large participation of workers who have suffered sharp and recent decline in their income. This situation is known in the policy evaluation and labor economics literature as *Ashenfelter's dip*, following Ashenfelter (1978). This possibility directly affects the properties of the DID estimators, since they depend on some extent on D_{it} to validate the parallel trend assumption. To see why, suppose for simplicity that there is no spillover effect and that a simple version of the parallel trend assumption is considered, requiring that for any $t, t^{'}$ and $g, g^{'}$, it holds that

$$E[Y_{it}(0) - Y_{it'}(0)|G_i = g] = E[Y_{it}(0) - Y_{it'}(0)|G_i = g^{'}]$$

where G_i is a variable indicating the first period at which unit i receives the treatment. Note that G_i is completely determined by D_{it}, after all, the period where unit i enters the treatment group is precisely the (only) period when D_{it} has a jump-type discontinuity, i.e., $G_i = g$ if and only if $D_{ig} - D_{i,g-1} = 1$. By the assumed specification for the potential outcome in the absence of treatment, this is equivalent to requiring that

$$E[\varepsilon_{it} - \varepsilon_{it'}|G_i = g] = E[\varepsilon_{it} - \varepsilon_{it'}|G_i = g^{'}]$$

From this expression, it is clear that the parallel trend assumption depends on D_{it}. Moreover, it becomes clear that the validity of this assumption depends on the relationship between D_{it} and the sequence of shocks $\{\varepsilon_{it}\}_t$. So, essentially, what we are doing by exploring different treatment selection mechanism is stressing the parallel trends assumption. In the case of exploring spillover effects, the stressed hypothesis is SUTVA.

In summary, in the simulations related to treatment assignment, different relationships between D_{it} and $\{\varepsilon_{it}\}_t$ are considered, while in the simulations involving spillover effects, different forms for the function ϕ are explored.

3.1 Treatment assignment mechanisms

This subsection is dedicated to simulations involving four different treatment assignment mechanisms. For each of them, we simulate 1000 independent panel datasets containing the untreated potential outcomes of the units over time. In each of them, we consider a situation where units that suffer a sufficiently negative shock to their potential outcome are assigned to the treatment, which has a positive effect (equal to 3) on the treated units. That is, $\tau_{i,t} = 3$ is the true value of the parameter that we wish to estimate. When a unit is treated, the actual observed value for it, once the treatment starts, is given by

$$Y_{i,t} = \alpha_i + \lambda_t + 3 + \varepsilon_{i,t} = Y_{i,t}(0) + 3$$

The first mechanism (mechanism A) assigns unit i to the treatment group if and only if $\varepsilon_{i,t} < -1.64$. As the shocks are drawn from a standard normal distribution, the value of -1.64 represents the 5th percentile of the distribution function.

The second mechanism (B) selects some unit i to the treatment group at time t if $\varepsilon_{i,t-4} < -1.64$. Now the treatment assignment is not instantaneous after the negative shock suffered, so there is a delay between the occurrence of the shock that causes a unit to be treated and the start of the treatment. This mechanism may be more realistic in the real world. For example, if a city suffers a shock (in health, violence, economy, etc.), it takes a while for the government to identify the problem and implement policies to mitigate it. Another example would be a person who suffers a significant and unexpected drop in her salary and waits a while before deciding whether to enroll in a labor qualification program.

The third mechanism (C) is analogous to the second but assigns a unit to the treatment 8 periods (instead of 4) after suffering a negative shock of less than -1.64. The aim of this round is to allow for a better understanding of the effects of the time lag between the shock and the start of the treatment on the estimators.

Finally, the fourth mechanism (D) assigns unit i to treatment in period t if and only if $\varepsilon_{i,t}/\alpha_i < -2\%$. Unlike previous mechanisms, now the magnitude of the shock is considered relative to the unit's fixed effect, rather than its magnitude in absolute terms. In the context of labor training programs, for example, such a mechanism could be rationalized by interpreting that the assignmente to treatmente is determined by the relative impact on individual's base wage, rather than in absolute terms. Another example would be the adoption of a specific anti-robbery policy that would take into account not simply the absolute variation in cases recorded in different localities, but rather this variation in per capita terms.

3.2 Spillover effects

This subsection details the strategy in the presence of spillover effects. Two situations are studied (spillover I and spillover II), both of them considering that all units belong to some (but only one) cluster and the units in the control group derive a positive effect from the treatment received by the units in the same cluster. Cities of a metropolitan area and neighboring areas in a city are good examples of clusters.

Regardless of the spillover effect, the general design of the experiments is as follows. In each of the 1000 repetitions we create panels with 20 periods containing 50, 250 and 500 individuals, and then generate their potential outcomes over time. These units are randomly divided into clusters of 10. For each cluster we randomly select a period for it to be treated. This means that at least one (but not necessarily all) unit of the cluster enters the treatment group in that period. From the moment the cluster is treated until the final period $T = 20$, 7 of the 10 units are randomly assigned to treatment over time. The potential outcome is given by

$$Y_{i,t}(D_{i,t}, D_{-i,t}) = \alpha_i + \lambda_t + 3D_{i,t} + 3\rho(1 - D_{i,t})\phi(D_{-i,t}) + \epsilon_{i,t}$$

where $D_{-i,t}$ represents the treatment status of the units in the same cluster as unit i, ϕ is a function that varies depending on the spillover effect (I or II), and $\rho > 0$

measures the intensity of the spillover effect. It is easy to see that the population treatment effect on a treated unit i in period t is equal to

$$\tau_{i,t} = Y_{i,t}(D_{i,t} = 1, D_{-i,t}) - Y_{i,t}(D_{i,t} = 0, 0) = 3$$

Under the spillover effect I, it is assumed that $\phi(D_{-i,t}) = \max_{j \neq i} D_{j,t}$ and $\rho = 10\%$. That is, once a cluster joins the treatment group, the units actually treated receive a benefit equal to 3, while the units remaining in the control group derive a benefit of 0.3. Over time, more units in the cluster derive the full effect of 3 as more of them receive the treatment. Under this design, we assume that the spillover on control units does not increase as more units in the same cluster become treated. As discussed in Butts (2021), this pattern of spillover can happen, for example, with the construction of a new library in some city in which there was none before. In this case, there may be a spillover effect for residents of surrounding towns (presumably smaller than the benefit of living in a town with a library, but still positive) as they also gain access to this new library. However, it should not make a difference whether there is one or more libraries they can go to, as long as they have access to one. One can rationalize this by thinking that the spillover effect to the control group is not additive in the number of units treated in the cluster, but binary.

On the other hand, under spillover II, it is assumed $\rho = 2\%$ and $\phi(D_{-i,t}) = \langle D_{-i,t}, D_{-i,t} \rangle$, where $\langle \cdot, \cdot \rangle$ is the usual inner product for Euclidean spaces. Thus, now $\phi(D_{-i,t})$ indicates the number of treated units at time t belonging to the same cluster as unit i, and the spillover benefit increases with the number of treated units in the same cluster. According to Butts (2021), this spillover pattern may represent externalities associated with economies of agglomeration. For instance, one can think of a situation where residents of one neighborhood hire private security that also benefits surrounding neighborhoods, and this positive effect increases as residents from other close surrounding areas also start hiring private security.

4 Results

This section presents the results of the simulations. For each mechanism, spillover effect, and each sample size, there is a table with the descriptive statistics calculated from the simulations. Density plots of the estimates for all simulations are presented in the Appendix.

4.1 Treatment assignment mechanisms

Tables 1, 2, and 3 show the results of mechanism A, which considers an immediate assignment to the treatment units that suffer a high negative shock in absolute terms. The best-performing estimators by the criteria of average bias and MSE are the TWFE and the one proposed by Borusyak et al. (2021). Still, the results show that these models present considerable bias, overestimating the true value of the coefficient by more than 15% on average. Moreover, the true value of the treatment effect (3) is outside the estimated 95% confidence interval for all sample sizes and all five estimators.

The performance of the two estimators proposed by Callaway and Sant'Anna (2021) and the one proposed by Sun and Abraham (2021) is qualitatively similar. In addition to the overall MSE being very high compared to the other models, they have an estimated bias greater than 2, which is greater than 2/3 of the true value. As the sample size increases, the high bias and high MSE do not seem to disappear (Tables 2 and 3). In this sense, the estimators appear to become inconsistent under such a treatment assignment mechanism.

Table 1 Mechanism A / NT=1000

Estimator	**LB (95%)**	**UB (95%)**	**MSE**	**Mean**	**Median**
TWFE	3.2329	3.7141	0.2436	3.4780	3.4839
C&S-NV	4.8688	5.3046	4.3242	5.0764	5.0764
C&S-NY	4.9575	5.3889	4.7245	5.1707	5.1669
Borusyak	3.5329	4.2427	0.8031	3.8766	3.8703
Sun & A	4.8398	5.3001	4.2756	5.0643	5.0637

Note: The true treatment value is 3. "LB" and "UP" refer to the lower and upper bounds of the 95% confidence interval. "C&S-NV" and "C&S-NY" refer to the estimator proposed by Callaway and Sant'Anna (2021) using as a control group the never treated and the not yet treated units, respectively.

Table 2 Mechanism A / NT=5000

Estimator	**LB (95%)**	**UB (95%)**	**MSE**	**Mean**	**Median**
TWFE	3.3752	3.5778	0.2310	3.4778	3.4787
C&S-NV	4.9809	5.1656	4.2921	5.0712	5.0702
C&S-NY	5.0745	5.2651	4.6991	5.1672	5.1664
Borusyak	3.7304	4.0310	0.7811	3.8802	3.8822
Sun & A	4.9702	5.1556	5.0587	5.0587	5.0567

Note: The true treatment value is 3. "LB" and "UP" refer to the lower and upper bounds of the 95% confidence interval. "C&S-NV" and "C&S-NY" refer to the estimator proposed by Callaway and Sant'Anna (2021) using as a control group the never treated and the not yet treated units, respectively.

Table 3 Mechanism A / NT=10000

Estimator	**LB (95%)**	**UB (95%)**	**MSE**	**Mean**	**Median**
TWFE	3.4125	3.5519	0.2313	3.4795	3.4770
C&S-NV	5.0088	5.1410	4.2953	5.0722	5.0717
C&S-NY	5.1046	5.2359	4.7036	5.1685	5.1678
Borusyak	3.7712	3.9989	0.7854	3.8844	3.8836
Sun & A	4.9946	5.1259	4.2409	5.0590	5.0581

Note: The true treatment value is 3. "LB" and "UP" refer to the lower and upper bounds of the 95% confidence interval. "C&S-NV" and "C&S-NY" refer to the estimator proposed by Callaway and Sant'Anna (2021) using as a control group the never treated and the not yet treated units, respectively.

Tables 4 to 9 show the results of mechanisms B and C, which consider 4 and 8 lags between the negative shock and the entry to the treatment group. Unlike mechanism A, the performance of the models proposed by Callaway and Sant'Anna (2021) and Sun and Abraham (2021) is superior to the estimator of Borusyak et al. (2021) and TWFE when considering the MSE, the estimated confidence interval, and the two measures of centrality (mean and median). Just as in mechanism A, there does not seem to be a convergence to the true value of the treatment effect as the number of individuals in the panel grows, since the MSE decreases at a small rate. Thus, all estimators appear to be inconsistent under the scenario we analyze. Comparing mechanisms, there is very little qualitative change in the results whether the assignment to treatment occurs 4 or 8 periods following the shock (mechanism B and C, respectively). On the other hand, the magnitude of the distortions is considerably higher in the case where the assignment to the treatment occurs instantaneously with the shock (mechanism A).

Table 4 Mechanism B / NT=1000

Estimator	**LB (95%)**	**UB (95%)**	**MSE**	**Mean**	**Median**
TWFE	3.0589	3.5122	0.0970	3.2896	3.2888
C&S-NV	2.5625	3.5372	0.0612	3.0497	3.0461
C&S-NY	2.6041	3.5964	0.0675	3.0871	3.0820
Borusyak	3.1192	3.6322	0.1512	3.3675	3.3655
Sun & A	2.5250	3.5450	0.0716	3.0512	3.0615

Note: The true treatment value is 3. "LB" and "UP" refer to the lower and upper bounds of the 95% confidence interval. "C&S-NV" and "C&S-NY" refer to the estimator proposed by Callaway and Sant'Anna (2021) using as a control group the never treated and the not yet treated units, respectively.

Table 5 Mechanism B / NT=5000

Estimator	**LB (95%)**	**UB (95%)**	**MSE**	**Mean**	**Median**
TWFE	3.1919	3.3863	0.0884	3.2932	3.2948
C&S-NV	2.8328	3.2202	0.0110	3.0353	3.0393
C&S-NY	2.8711	3.2649	0.0152	3.0726	3.0756
Borusyak	3.2644	3.4787	0.1416	3.3722	3.3731
Sun & A	2.8321	3.2202	0.0110	3.0352	3.0393

Note: The true treatment value is 3. "LB" and "UP" refer to the lower and upper bounds of the 95% confidence interval. "C&S-NV" and "C&S-NY" refer to the estimator proposed by Callaway and Sant'Anna (2021) using as a control group the never treated and the not yet treated units, respectively.

Finally, Tables 10, 11 and 12 show the results of mechanism D, which assigns to treatment those units that suffer a high negative shock relative to their own average, instead of in absolute terms. The results are quite similar to mechanism A. The TWFE model and the one proposed by Borusyak et al. (2021) outperform those of Callaway and Sant'Anna (2021) and Sun and Abraham (2021). However, in all of them, for all panel dimensions, the true value of the treatment effect is not within the estimated confidence interval. Moreover, as in the previous cases, the estimators do not seem to be consistent.

Table 6 Mechanism B / NT=10000

Estimator	LB (95%)	UB (95%)	MSE	Mean	Median
TWFE	3.2250	3.3626	0.0858	3.2909	3.2910
C&S-NV	2.9003	3.1815	0.0069	3.0404	3.0412
C&S-NY	2.9405	3.2213	0.0113	3.0772	3.0775
Borusyak	3.2946	3.4448	0.1383	3.3699	3.3698
Sun & A	2.9003	3.1815	0.0069	3.0404	3.0412

Note: The true treatment value is 3. "LB" and "UP" refer to the lower and upper bounds of the 95% confidence interval. "C&S-NV" and "C&S-NY" refer to the estimator proposed by Callaway and Sant'Anna (2021) using as a control group the never treated and the not yet treated units, respectively.

Table 7 Mechanism C / NT=1000

Estimator	LB (95%)	UB (95%)	MSE	Mean	Median
TWFE	2.9558	3.4481	0.0624	3.2148	3.2175
C&S-NV	2.5336	3.5322	0.0679	3.0522	3.0564
C&S-NY	2.4891	3.5022	0.0672	3.0151	3.0214
Borusyak	2.9661	3.4785	0.0708	3.2320	3.2346
Sun & A	2.4598	3.5759	0.0801	3.0533	3.0418

Note: The true treatment value is 3. "LB" and "UP" refer to the lower and upper bounds of the 95% confidence interval. "C&S-NV" and "C&S-NY" refer to the estimator proposed by Callaway and Sant'Anna (2021) using as a control group the never treated and the not yet treated units, respectively.

Table 8 Mechanism C / NT=5000

Estimator	LB (95%)	UB (95%)	MSE	Mean	Median
TWFE	3.1152	3.3373	0.0511	3.2195	3.2200
C&S-NV	2.8351	3.2909	0.0179	3.0632	3.0615
C&S-NY	2.8008	3.2526	0.0194	3.0262	3.0226
Borusyak	3.1300	3.3552	0.0595	3.2375	3.2379
Sun & A	2.8351	3.2909	0.0179	3.0632	3.0618

Note: The true treatment value is 3. "LB" and "UP" refer to the lower and upper bounds of the 95% confidence interval. "C&S-NV" and "C&S-NY" refer to the estimator proposed by Callaway and Sant'Anna (2021) using as a control group the never treated and the not yet treated units, respectively.

Overall, the results show that the performance of Difference-in-Differences models depends largely on the selection mechanism that assigns units to treatment. Mechanisms that depend on the shocks suffered by the units can cause a significant bias in the estimates, as well as inconsistency. If there is a large time lag between the shock and the start of treatment, these problems may decrease, but even with large time windows the problem tends to persist. In addition, the conclusions do not depend on whether we consider shocks to be large in absolute or relative terms. Moreover, it is important to note, for the specific case of TWFE, that part of its good performance may be due to the fact that the treatment effect was assumed to be homogeneous. In this way, Gauss-Markov conclusions can be applied, since the TWFE is estimated by OLS. In practical applications, where the assumption of homogeneity in the treatment

Table 9 Mechanism C / NT=10000

Estimator	LB (95%)	UB (95%)	MSE	Mean	Median
TWFE	3.1377	3.2967	0.0488	3.2173	3.2177
C&S-NV	2.8969	3.2162	0.0102	3.0603	3.0606
C&S-NY	2.8588	3.1803	0.0073	3.0230	3.0243
Borusyak	3.1542	3.3159	0.0572	3.2356	3.2352
Sun & A	2.8969	3.2162	0.0102	3.0603	3.0606

Note: The true treatment value is 3. "LB" and "UP" refer to the lower and upper bounds of the 95% confidence interval. "C&S-NV" and "C&S-NY" refer to the estimator proposed by Callaway and Sant'Anna (2021) using as a control group the never treated and the not yet treated units, respectively.

Table 10 Mechanism D / NT=1000

Estimator	LB (95%)	UB (95%)	MSE	Mean	Median
TWFE	3.0862	3.6468	0.1562	3.3689	3.3696
C&S-NV	5.0110	5.5609	5.2384	5.2842	5.2828
C&S-NY	5.0675	5.6233	5.4636	5.3330	5.3293
Borusyak	3.2838	4.2101	0.5708	3.7173	3.7162
Sun & A	4.9046	5.6297	5.1890	5.2698	5.2710

Note: The true treatment value is 3. "LB" and "UP" refer to the lower and upper bounds of the 95% confidence interval. "C&S-NV" and "C&S-NY" refer to the estimator proposed by Callaway and Sant'Anna (2021) using as a control group the never treated and the not yet treated units, respectively.

Table 11 Mechanism D / NT=5000

Estimator	LB (95%)	UB (95%)	MSE	Mean	Median
TWFE	3.2567	3.4905	0.1396	3.3688	3.3665
C&S-NV	5.1764	5.4031	5.2273	5.2855	5.2850
C&S-NY	5.2255	5.4516	5.4620	5.3363	5.3361
Borusyak	3.5400	3.9482	0.5467	3.7317	3.7336
Sun & A	5.1723	5.3982	5.2048	5.2806	5.2805

Note: The true treatment value is 3. "LB" and "UP" refer to the lower and upper bounds of the 95% confidence interval. "C&S-NV" and "C&S-NY" refer to the estimator proposed by Callaway and Sant'Anna (2021) using as a control group the never treated and the not yet treated units, respectively.

effect may be too strong, one may expect, based on results from this literature, that its performance will be impaired.

4.2 Spillover effect

Spillover Effects I and II

This subsection reports the results under spillover effects I and II. As shown in Tables 13, 14, and 15, the presence of a non-additive spillover pattern (spillover I) systematically underestimates the true value of the parameter. Only the estimator proposed by Sun and Abraham (2021) simulated with a shorter panel dimension ($NT = 1000$) delivers the true treatment effect ($\tau = 3$) within the estimated confidence interval at the 95% level.

Table 12 Mechanism D / NT=10000

Estimator	LB (95%)	UB (95%)	MSE	Mean	Median
TWFE	3.2804	3.4683	0.1413	3.3731	3.3724
C&S-NV	5.2050	5.3705	5.2368	5.2879	5.2873
C&S-NY	5.2541	5.4194	5.4702	5.3384	5.3384
Borusyak	3.5904	3.8896	0.5524	3.7391	3.7404
Sun & A	5.1997	5.3646	5.2140	5.2830	5.2828

Note: The true treatment value is 3. "LB" and "UP" refer to the lower and upper bounds of the 95% confidence interval. "C&S-NV" and "C&S-NY" refer to the estimator proposed by Callaway and Sant'Anna (2021) using as a control group the never treated and the not yet treated units, respectively.

The estimator of Callaway and Sant'Anna (2021) that uses the not yet treated units as a control group, but mainly the estimator of Sun and Abraham (2021), are the ones impacted to a lesser extent by the presence of the non-additive spillover. Still, on average, the estimates underestimated the true value of the treatment effect by more than 10%. The worst performer in this regard is the estimator of Borusyak et al. (2021), which on average underestimates the true value of the parameter by more than 40%, even in the largest panels. Moreover, for all models, the slow rate of decrease of the MSE as the panel dimension grows highlights that, in the presence of such a pattern of spillover effect I, the estimators studied are not only biased but also seem to be inconsistent.

Table 13 Spillover I (non-additive) / NT=1000

Estimator	LB (95%)	UB (95%)	MSE	Mean	Median
TWFE	1.3179	2.3146	1.2687	1.9016	1.9350
C&S-NV	0.9829	2.3472	1.5646	1.7965	1.8329
C&S-NY	1.2815	2.6786	0.9126	2.1096	2.1546
Borusyak	1.0156	2.1730	1.7527	1.7081	1.7495
Sun & A	1.7222	3.3475	0.3108	2.6408	2.6810

Note: The true treatment value is 3. "LB" and "UP" refer to the lower and upper bounds of a confidence interval at the 95% level. "C&S-NV" and "C&S-NY" refer to the estimator proposed by Callaway and Sant'Anna (2021) using as a control group the never treated and the not yet treated units, respectively.

Finally, Tables 16, 17, and 18 present the results under spillover effect II, which considers the possibility of an additive impact in non-treated groups as more units in the cluster receive the treatment. Again, all estimators underestimate the true value of the treatment effect, albeit by a smaller magnitude than under spillover I. On average, the underestimation is just under 10%. Furthermore, although in panels of dimension NT=1000, the true value of τ lies within the confidence interval of all estimators, these intervals shrink significantly to the point of no longer including the true value of 3 as the panel dimension increases.

Table 14 Spillover I (non-additive) / NT=5000

Estimator	LB (95%)	UB (95%)	MSE	Mean	Median
TWFE	1.7449	2.1602	1.0598	1.9760	1.9856
C&S-NV	1.4909	2.0962	1.3823	1.8349	1.8494
C&S-NY	1.7776	2.4033	0.7636	2.1410	2.1546
Borusyak	1.5003	2.0083	1.4920	1.7856	1.7993
Sun & A	2.2511	2.9330	0.1684	2.6295	2.6459

Note: The true treatment value is 3. "LB" and "UP" refer to the lower and upper bounds of a confidence interval at the 95% level. "C&S-NV" and "C&S-NY" refer to the estimator proposed by Callaway and Sant'Anna (2021) using as a control group the never treated and the not yet treated units, respectively.

Table 15 Spillover I (non-additive) / NT=10000

Estimator	LB (95%)	UB (95%)	MSE	Mean	Median
TWFE	1.8297	2.1243	1.0399	1.9831	1.9878
C&S-NV	1.6178	2.0352	1.3685	1.8351	1.8431
C&S-NY	1.9175	2.3343	0.7503	2.1407	2.1460
Borusyak	1.5983	1.9689	1.4582	1.7961	1.8020
Sun & A	2.3705	2.8359	0.1584	2.6214	2.6260

Note: The true treatment value is 3. "LB" and "UP" refer to the lower and upper bounds of a confidence interval at the 95% level. "C&S-NV" and "C&S-NY" refer to the estimator proposed by Callaway and Sant'Anna (2021) using as a control group the never treated and the not yet treated units, respectively.

Regarding centrality measures and MSE, the estimator TWFE (to a greater extent) and the one proposed by Borusyak et al. (2021) are less affected by the additive spillover effect, opposing the pattern verified under spillover I.

Similar to all previous simulations, the slow rate of decay of the MSE suggests inconsistency of the estimator also under spillover effect II.

Table 16 Spillover II (additive) / NT=1000

Estimator	LB (95%)	UB (95%)	MSE	Mean	Median
TWFE	2.5637	3.0262	0.0584	2.7915	2.7964
C&S-NV	2.2969	3.1990	0.1192	2.7468	2.7623
C&S-NY	2.2952	3.2157	0.1146	2.7568	2.7696
Borusyak	2.4963	3.0197	0.0758	2.7622	2.7635
Sun & A	2.2611	3.2035	0.1298	2.7446	2.7444

Note: The true treatment value is 3. "LB" and "UP" refer to the lower and upper bounds of a confidence interval at the 95% level. "C&S-NV" and "C&S-NY" refer to the estimator proposed by Callaway and Sant'Anna (2021) using as a control group the never treated and the not yet treated units, respectively.

Table 17 Spillover II (additive) / NT=5000

Estimator	LB (95%)	UB (95%)	MSE	Mean	Median
TWFE	2.7056	2.9084	0.0412	2.8037	2.8004
C&S-NV	2.5569	2.9486	0.0693	2.7572	2.7560
C&S-NY	2.5563	2.9689	0.0644	2.7684	2.7666
Borusyak	2.6579	2.8989	0.0529	2.7777	2.7783
Sun & A	2.5543	2.9520	0.0701	2.7559	2.7460

Note: The true treatment value is 3. "LB" and "UP" refer to the lower and upper bounds of a confidence interval at the 95% level. "C&S-NV" and "C&S-NY" refer to the estimator proposed by Callaway and Sant'Anna (2021) using as a control group the never treated and the not yet treated units, respectively.

Table 18 Spillover II (additive) / NT=10000

Estimator	LB (95%)	UB (95%)	MSE	Mean	Median
TWFE	2.7231	2.8739	0.0403	2.8027	2.8038
C&S-NV	2.6110	2.8907	0.0657	2.7535	2.7545
C&S-NY	2.6193	2.9026	0.0607	2.7641	2.7644
Borusyak	2.6889	2.8739	0.0523	2.7757	2.7763
Sun & A	2.6107	2.8892	0.0659	2.7530	2.7533

Note: The true treatment value is 3. "LB" and "UP" refer to the lower and upper bounds of a confidence interval at the 95% level. "C&S-NV" and "C&S-NY" refer to the estimator proposed by Callaway and Sant'Anna (2021) using as a control group the never treated and the not yet treated units, respectively.

In general, the presence of the spillover effects we study systematically compromises the performance of Difference-in-Differences estimators that become biased and inconsistent. Considering that spillovers are evident in several empirical analyses, our results suggest exercising caution when interpreting the outcomes in real-world applications. In addition, despite the methodological difficulties involved, the importance of developing estimators that are robust to the circumstances consider becomes evident and should prove to be an important research venue.

5 Conclusion

The goal of this paper was to conduct Monte Carlo experiments to evaluate the performance of commonly used Difference-in-Differences estimators under scenarios in which the treatment selection mechanism is affected by shocks suffered by the units, and under situations where treatment effects spillover to non-treated units that are in the control group. Specifically, we compare the estimators proposed by Callaway and Sant'Anna (2021), Borusyak et al. (2021) and Sun and Abraham (2021), in addition to the TWFE estimator.

Importantly, the scenarios we consider, although commonly observed in real applications, imply in violation of some assumptions behind the derivation of these estimators. In particular, the treatment selection mechanism influenced by shocks violates the

parallel trends assumption, while the spillover effect on units belonging to the control group violates the SUTVA hypothesis.
Our results indicate that the presence both situations can greatly compromise the performance of the estimators responsible for measuring the effect of the treatment on treated units: they become biased and may not even be consistent, since the bias is not eliminated when using panels with dimensions enlarged by the addition of extra units.
In practical applications, our results suggest cautious with the estimates obtained by this class of estimators when there is a suspicion that the context examined is similar to those we study. In addition, especially in the case of spillover effects, we emphasize the importance of developing estimators capable of dealing adequately with this type of structure. In this sense, the paper illuminates a possible direction for further methodological research on the Difference-in-Differences literature.
Finally, we highlight the importance of conducting similar research, both analytical and computational, but exploring other contexts, such as the presence of anticipation to the treatment effect, new treatment assignment mechanisms, new functional forms for potential outcomes, and the interaction of these factors with the heterogeneity of the treatment effect among the treated units and over time.

References

ANGRIST, Joshua D.; PISCHKE, Jörn-Steffen. The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. Journal of Economic Perspectives, v. 24, n. 2, p. 3-30, 2010.
ASHENFELTER, Orley. Estimating the effect of training programs on earnings. The Review of Economics and Statistics, p. 47-57, 1978.
ATHEY, Susan; IMBENS, Guido W. Design-based analysis in difference-in-differences settings with staggered adoption. Journal of Econometrics, v. 226, n. 1, p. 62-79, 2022.
BORUSYAK, Kirill; JARAVEL, Xavier; SPIESS, Jann. Revisiting event study designs: Robust and efficient estimation. arXiv preprint arXiv:2108.12419, 2021.
BUTTS, Kyle. "Difference-in-differences estimation with spatial spillovers." arXiv preprint arXiv:2105.03737, 2021.
CALLAWAY, Brantly; SANT'ANNA, Pedro HC. Difference-in-differences with multiple time periods. Journal of Econometrics, v. 225, n. 2, p. 200-230, 2021.
DE CHAISEMARTIN, Clément; D'HAULTFOEUILLE, Xavier. Two-way fixed effects estimators with heterogeneous treatment effects. American Economic Review, v. 110, n. 9, p. 2964-96, 2020.
DE CHAISEMARTIN, Clément; D'HAULTFOEUILLE, Xavier. Two-way fixed effects and differences-in-differences with heterogeneous treatment effects: A survey. National Bureau of Economic Research, 2022.
GARDNER, John. Two-stage differences in differences. arXiv preprint arXiv:2207.05943, 2022.
GHANEM, Dalia; SANT'ANNA, Pedro HC; WÜTHRICH, Kaspar. Selection and parallel trends. arXiv preprint arXiv:2203.09001, 2022.
GOODMAN-BACON, Andrew. Difference-in-differences with variation in treatment timing. Journal of Econometrics, v. 225, n. 2, p. 254-277, 2021.
HECKMAN, James J.; SMITH, Jeffrey A. The pre-programme earnings dip and the

determinants of participation in a social programme. Implications for simple programme evaluation strategies. The Economic Journal, v. 109, n. 457, p. 313-348, 1999.
HOLLAND, Paul W. Statistics and causal inference. Journal of the American Statistical Association, v. 81, n. 396, p. 945-960, 1986.
LECHNER, Michael et al. The estimation of causal effects by difference-in-difference methods. Foundations and Trends in Econometrics, v. 4, n. 3, p. 165-224, 2011.
LIU, Licheng; WANG, Ye; XU, Yiqing. A practical guide to counterfactual estimators for causal inference with time-series cross-sectional data. arXiv preprint arXiv:2107.00856, 2021.
MARCUS, Michelle; SANT'ANNA, Pedro HC. The role of parallel trends in event study settings: an application to environmental economics. Journal of the Association of Environmental and Resource Economists, v. 8, n. 2, p. 235-275, 2021.
ROBINS, James. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. Mathematical Modelling, v. 7, n. 9-12, p. 1393-1512, 1986.
ROTH, Jonathan et al. What's Trending in Difference-in-Differences? A Synthesis of the Recent Econometrics Literature. arXiv preprint arXiv:2201.01194, 2022.
RUBIN, Donald B. Estimating causal effects of treatments in randomized and non-randomized studies. Journal of Educational Psychology, v. 66, n. 5, p. 688, 1974.
SUN, Liyang; ABRAHAM, Sarah. Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. Journal of Econometrics, v. 225, n. 2, p. 175-199, 2021.
WOOLDRIDGE, Jeffrey M. Econometric analysis of cross section and panel data. MIT press, 2010.

Appendix A

Fig. A1 Mechanism A / NT = 1000

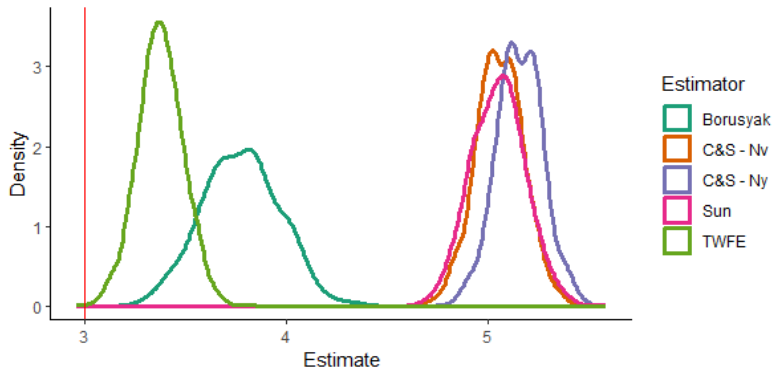

Fig. A2 Mechanism A / NT = 5000

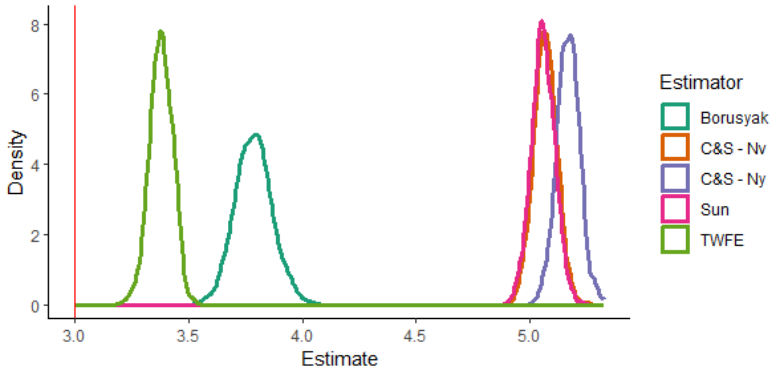

Fig. A3 Mechanism A / NT = 10000

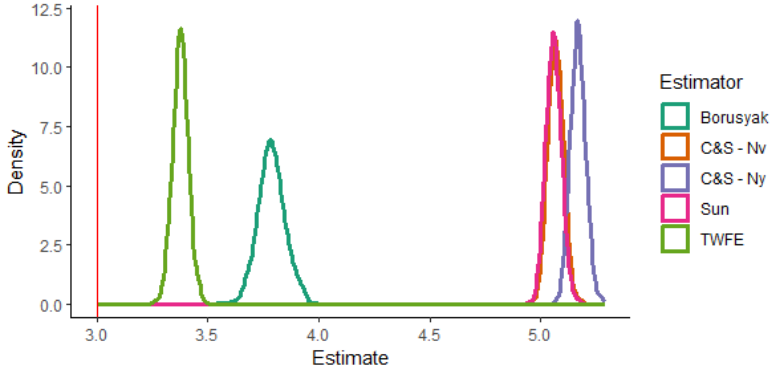

Fig. A4 Mechanism B / NT = 1000

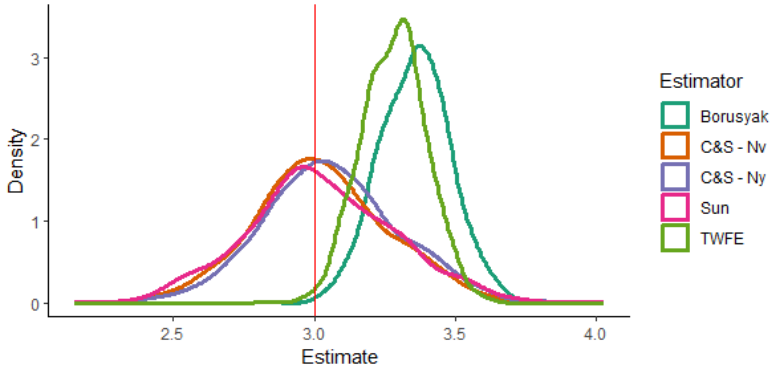

Fig. A5 Mechanism B / NT = 5000

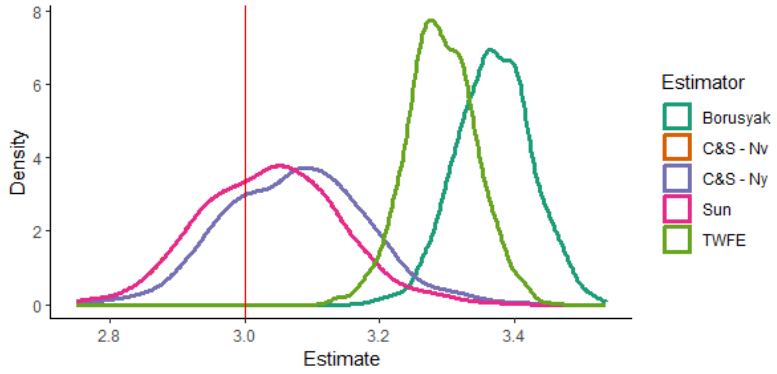

Fig. A6 Mechanism B / NT = 10000

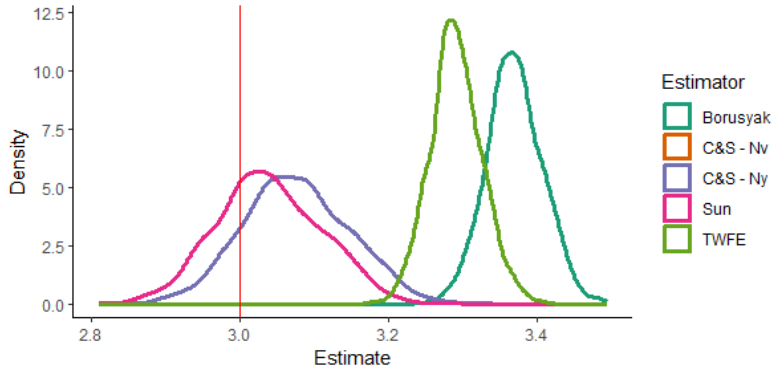

Fig. A7 Mechanism C / NT = 1000

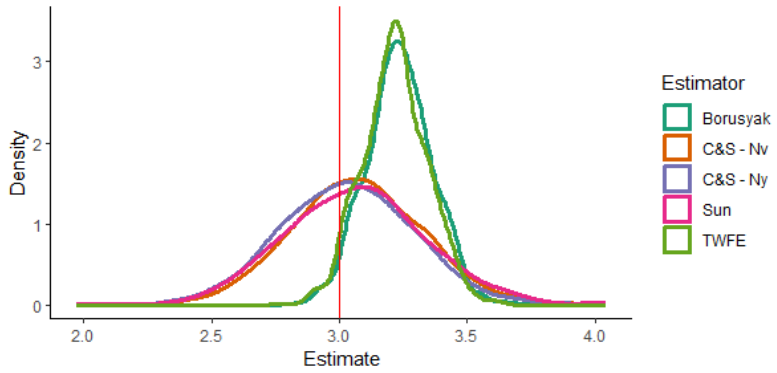

Fig. A8 Mechanism C / NT = 5000

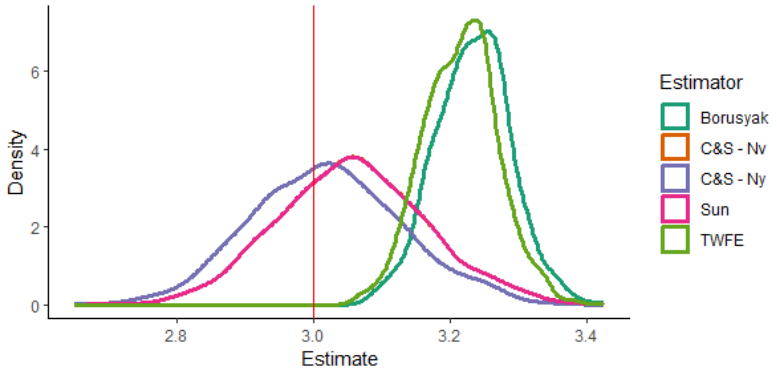

Fig. A9 Mechanism C / NT = 10000

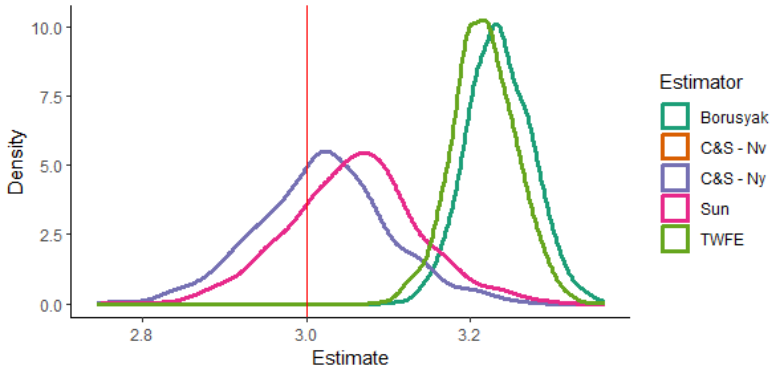

Fig. A10 Mechanism D / NT = 1000

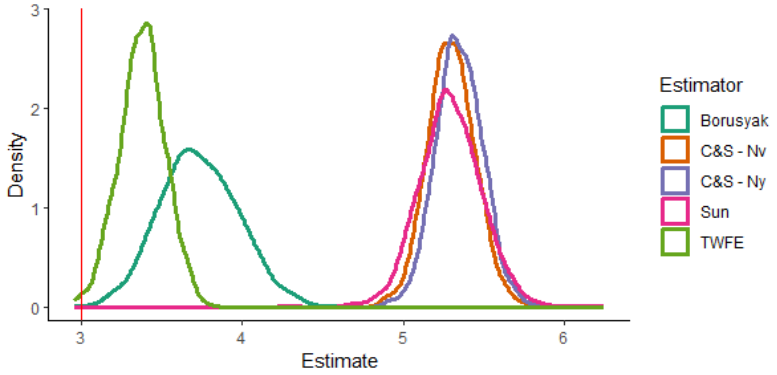

Fig. A11 Mechanism D / NT = 5000

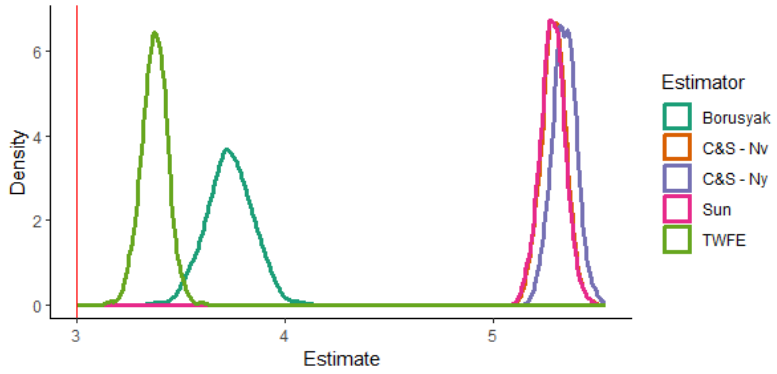

Fig. A12 Mechanism D / NT = 10000

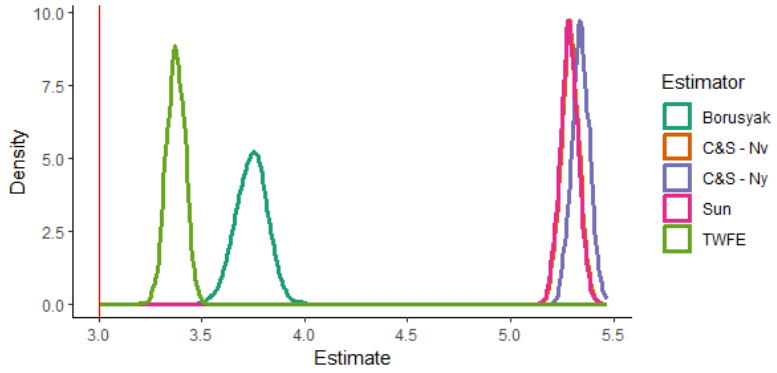

Fig. A13 Spillover I (non-additive) / NT=1000

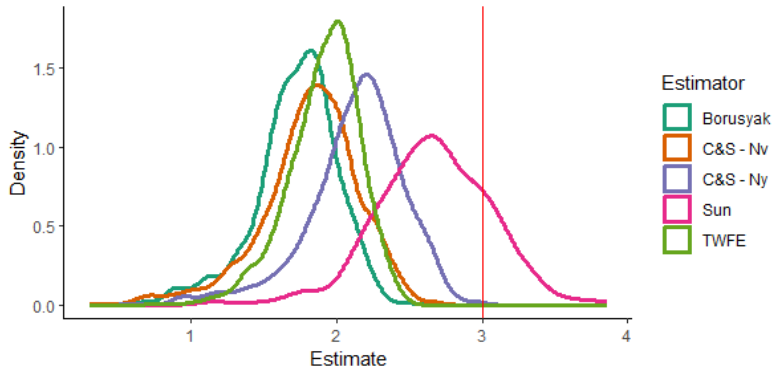

Fig. A14 Spillover I (non-additive) / NT=1000

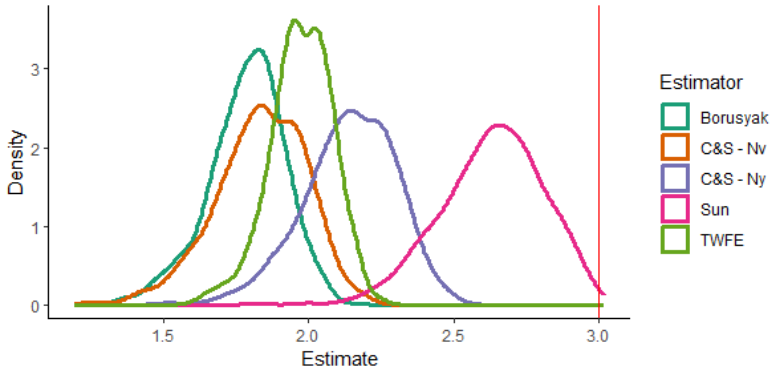

Fig. A15 Spillover I (non-additive) / NT=1000

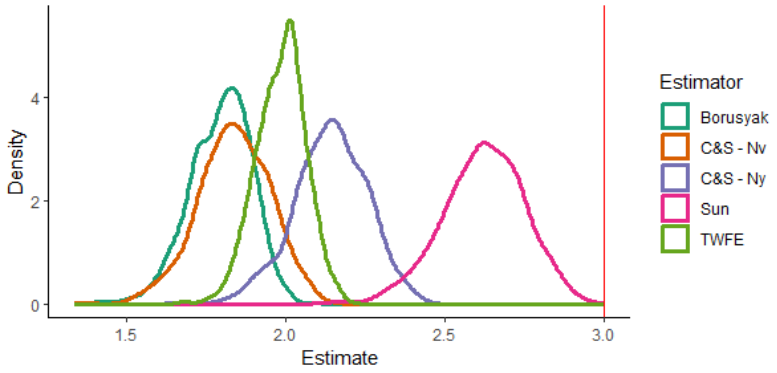

Fig. A16 Spillover II (additive) / NT=1000

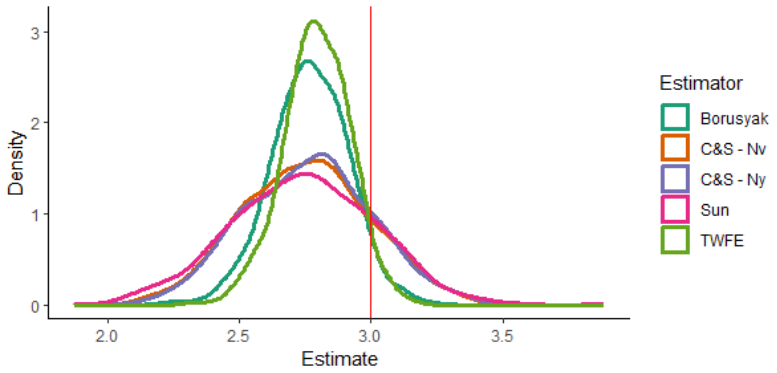

Fig. A17 Spillover II (additive) / NT=1000

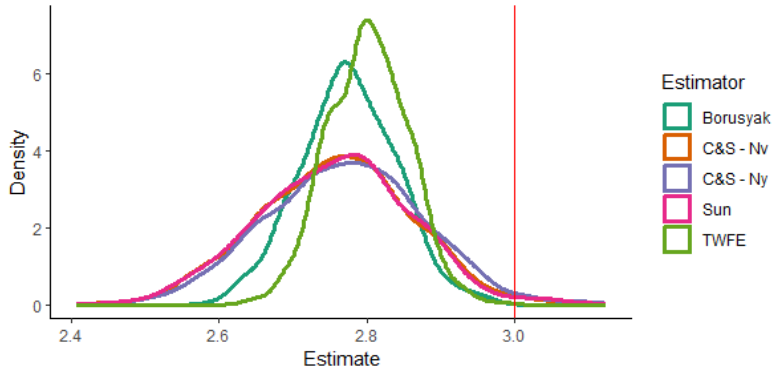
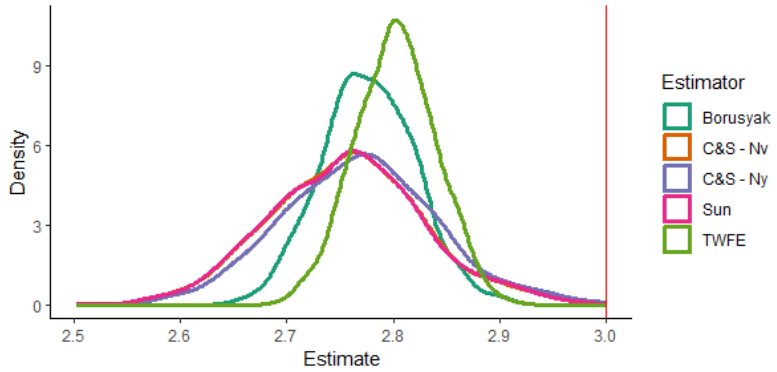
Fig. A18 Spillover II (additive) / NT=1000