Inference in Difference-in-Differences with Few Treated Units and Spatial Correlation*

Bruno Ferman†

Sao Paulo School of Economics - FGV

First Draft: June 30th, 2020

This Draft: September 27th, 2021

Please click here for the most recent version

Abstract

We consider the problem of inference in Difference-in-Differences (DID) models when there are few treated units and errors are spatially correlated. We first show that, when there is a single treated unit, existing inference methods designed for settings with few treated and many control units remain asymptotically valid when errors are weakly dependent. However, these methods may be invalid with more than one treated unit. We propose asymptotically valid, though generally conservative, inference methods for settings with more than one treated unit. These alternatives are valid even when the relevant distance metric across units is unavailable. We also present an empirical application that highlights some common misunderstandings in the use of randomization inference in DID applications, and illustrates how our results can be used to provide proper inference.

Keywords: hypothesis testing; causal inference; randomization inference; permutation tests

JEL Codes: C12; C21; C23; C33

*I would like to thank Luis Alvarez and Chris Taber for comments and suggestions. Lucas Barros provided exceptional research assistance. I also thank Benjamin Sommers for useful discussions and for providing the county FIPS codes for the counties used in Sommers et al. (2014). I gratefully acknowledge financial support from FAPESP and CNPq.

†email: bruno.ferman@fgv.br; address: Sao Paulo School of Economics, FGV, Rua Itapeva no. 474, Sao Paulo - Brazil, 01332-000; telephone number: +55 11 3799-3350

1 Introduction

Difference-in-Differences (DID) models present a series of challenges for inference. There is a large number of inference methods for DID, but the effectiveness of different solutions depend crucially on the set of assumptions one is willing to make on the errors, and on many features of the empirical design, such as the number of treated and control units.[1] We consider a common setting in which a satisfactory solution is not yet available, when (i) there is a small number of treated units, (ii) the number of periods is fixed, and (iii) errors are possibly spatially correlated, but there is no available distance metric across units. Throughout, we refer to "spatial correlation" as any correlation in the cross section, not necessarily related to a geographical distance metric.

We first derive conditions in which the inference methods proposed by Conley and Taber (2011) (henceforth, CT) and Ferman and Pinto (2019) (henceforth, FP) remain asymptotically valid in the presence of spatial correlation when there is a single treated and many control units, even when a distance metric across units is not available. The main assumptions are that (i) the post-pre difference in average errors for each unit has the same marginal distribution for all units — we can relax this assumption by allowing for heteroskedasticity depending on observed covariates —, and (ii) the cross-section distribution of this post-pre difference in average errors is weakly dependent for the control units. Under these conditions, the asymptotic distribution of the DID estimator depends only on the post-pre difference in average errors of the treated unit (as originally shown by CT). Moreover, the residuals of the control units asymptotically recover the distribution of the errors of the treated unit even when there is spatial correlation (in contrast, CT and FP assume independence across units for this result or a spatial correlation depending on an observed distance metric).

However, when there is more than one treated unit, we show that the inference methods

[1] A non-exhaustive list of papers that proposed and/or analyzed different inference methods in the DID settings include Arellano (1987), Bertrand et al. (2004), Cameron et al. (2008), Brewer et al. (2017), Conley and Taber (2011), Donald and Lang (2007) Ferman and Pinto (2019), Canay et al. (2017), MacKinnon and Webb (2020), and Ferman (2021).

proposed by CT and FP may not be asymptotically valid if there is spatial correlation, and the applied researcher does not have information on the relevant distance metric. The intuition is clear: when we (mistakenly) assume errors are independent across clusters, we underestimate the volatility of the average of the errors for the treated units if errors are positively correlated across the treated units.

We propose two alternative inference methods that are asymptotically valid, though generally conservative, in this setting. The first one considers the worst-case scenario for the (unobserved) dependence between the errors of the treated units. While this guarantees a test that is asymptotically valid when the number of control units increases (with the number of treated units fixed), the cost of being robust to unobserved spatial correlation among the treated units is a potential loss in terms power.

The second alternative we propose applies to common settings in which the outcome variable is the aggregate of unit $\times$ time individual-level observations, and is valid whether we have access to the individual-level data or only to the aggregate-level data. In such settings, we are able to use information on the within-unit correlation to bound the spatial correlation between treated units. Under the assumption that two individuals within the same unit are relatively more correlated than two individuals in different unit, we derive an alternative inference method that can be substantially less conservative relative to the previous alternative, implying a relatively lower loss in terms of power.

We analyze a Monte Carlo simulation based on the spatial correlation structure of the American Community Survey (ACS). We then revisit the work by Sommers et al. (2014), who analyzed the effects of the Massachusetts 2006 health reform, in light of our results. This empirical application highlights some common misunderstandings in the use of randomization inference in DID applications, and illustrates how our results can be used to provide proper inference in this and other more complex settings.

2 Setting

Let $Y_{st}(0)$ ($Y_{st}(1)$) be the potential outcome of unit s at time t when this unit is untreated (treated) at this period. We consider that potential outcomes are given by

$$\begin{cases} Y_{st}(0) = \theta_s + \gamma_t + \eta_{st} \\ Y_{st}(1) = \alpha_{st} + Y_{st}(0), \end{cases} \tag{1}$$

where α_{st} is the (possibly heterogeneous) treatment effects for unit s at time t, θ_s are time-invariant unobserved effects, and γ_t are group-invariant unobserved effects. The error term η_{st} represents unobserved determinants of $Y_{st}(0)$ that are not captured by the fixed effects. This structure for the potential outcomes, together with Assumption 2.1(i) below, is similar to assuming parallel trends for all periods. We can extend our results to consider alternative parallel trends assumptions (Marcus and Sant'Anna, 2021) and alternative estimators.

We focus on the case in which treatment is non-reversible and starts for all treated units after date t^* (see Remark 3.5 for the case with variation in adoption time). There are N_1 treated units, N_0 control units, and T time periods. Let $\mathcal{I}_1$ ($\mathcal{I}_0$) be the set of indices for treated (control) units, while $\mathcal{T}_1$ ($\mathcal{T}_0$) be the set of indices for post- (pre-) treatment periods. We treat $\{\alpha_{st}\}_{s\in\mathcal{I}_1,t\in\mathcal{T}_1}$ as fixed parameters,[2] and define $\alpha = \frac{1}{N_1}\frac{1}{T_1}\sum_{s\in\mathcal{I}_1}\sum_{t\in\mathcal{T}_1}\alpha_{st}$ as the average treatment effects across treated units and treated periods. In this case, we have the following DID model

$$Y_{st} = \alpha d_{st} + \theta_s + \gamma_t + (\alpha_{st} - \alpha)d_{st} + \eta_{st}, \tag{2}$$

where Y_{st} is the observed outcome variable for unit s at time t, while d_{st} is an indicator

[2] CT and FP also consider a setting in which $\{\alpha_{st}\}_{s\in\mathcal{I}_1,t\in\mathcal{T}_1}$ is fixed. FP consider a setting in which treatment effects are not only fixed, but also homogeneous, which is stronger than what we assume. CT also consider in their Section III.A the case in which treatment effects may be heterogeneous across units and time. Since we consider a setting with N_1 fixed, it would not be possible to assess the distribution of α_{st}, if we considered this term as stochastic. Therefore, it would not be possible to provide valid inference if we consider α_{st} as stochastic and define the estimand of interest as its expected value.

variable equal to one if unit s is treated at time t, and zero otherwise.

Following FP, for a generic variable A_t, define $\nabla A = \frac{1}{T-t^*}\sum_{t\in\mathcal{T}_1} A_t - \frac{1}{t^*}\sum_{t\in\mathcal{T}_0} A_t$. In particular, we define $W_s = \nabla\eta_{st}$, which is the post-pre difference in average errors for each unit s. Note that by focusing on this linear combination of the errors, W_s, we do not impose any restriction on the serial correlation of the errors.

In this case in which treatment starts at the same period for all treated units, the DID estimator is numerically equivalent to the two-way fixed effects (TWFE) estimator of α,[3] which is given by

$$\widehat{\alpha} \;=\; \frac{1}{N_1}\sum_{s\in\mathcal{I}_1}\nabla Y_s - \frac{1}{N_0}\sum_{t\in\mathcal{I}_0}\nabla Y_s = \alpha + \frac{1}{N_1}\sum_{s\in\mathcal{I}_1} W_s - \frac{1}{N_0}\sum_{s\in\mathcal{I}_0} W_s. \tag{3}$$

We consider a repeated sampling framework over the distribution of $\{W_s\}_{s\in\mathcal{I}_0\cup\mathcal{I}_1}$, in which treatment assignment is pre-determined (alternatively, we can consider that the analysis is conditional on the vector of treatment assignment). We allow the distribution of W_s to vary across s according to an observed covariate Z_s, as considered by considered by FP.

Assumption 2.1 (i) $\mathbb{E}[W_s|Z_s] = 0$ for all $s \in \mathcal{I}_1 \cup \mathcal{I}_0$; (ii) $W_s = h(Z_s,\delta)\xi_s$, where ξ_s is independent of Z_s and has the same marginal cdf F_ξ for all $s \in \mathcal{I}_1 \cup \mathcal{I}_0$, with finite second moment and continuous distribution; (iii) $h(z,\delta)$ is a known function, where $\delta \in \mathbb{R}^m$ is an unknown parameter, and $h(z,\delta) \leq \bar{h}(\delta)$ for some function $\bar{h}(\delta)$; (iv) Z_s has the same marginal cdf F_z for all $s \in \mathcal{I}_0$; (v) let $Z \sim F_z$ and $\xi \sim F_\xi$ with $Z \perp \xi$. Then, for any $g(Z,\xi)$ such that $\mathbb{E}[|g(Z,\xi)|^2] < \infty$, $\frac{1}{N_0}\sum_{s\in\mathcal{I}_0} g(Z_s,\xi_s) \xrightarrow{p} \mathbb{E}[g(Z,\xi)]$.

Assumption 2.1(i) is a standard DID identification assumption. Assumptions 2.1(ii) to 2.1(iii) restrict the marginal distribution of the treated and control units. With this assumption, the residuals of the control units become informative about the distribution of the errors of the treated units, which is the main insight from CT. If we set $h(Z_s,\delta)$ constant,

[3] Our results are also relevant and can be extended for alternative estimators. For example, we can consider the first differences estimator, or the estimators proposed by de Chaisemartin and D'Haultfoeuille (2018), Sun and Abraham (2020) and Callaway and Sant'Anna (2018).

then these assumptions imply that the marginal distribution W_s is the same for both treated and control units, as considered by CT in their running model. More generally, we allow for scale changes in the distribution of W_s depending on observable variables Z_s. For example, FP consider in detail the case in which Y_{st} is the average of individual-level observations Y_{ist}. In this case, variation in units' sizes lead to heteroskedasticity, which can be captured by $h(Z_s, \delta)$. Note that we do not impose any restriction on the distribution of Z_s for the treated units, as we evaluate the properties of different inference methods conditional on the realization of $\{Z_s\}_{s \in \mathcal{I}_1}$. Therefore, in the example considered by FP, we can encompass cases in which treated units are systematically smaller or larger relative to the control units.

Finally, Assumption 2.1(v) allows for spatially correlated shocks, but restricts such dependence so that we can apply a law of large numbers when we consider the control units. This will be satisfied, for example, if we assume strong mixing conditions in the cross section (see Theorem 3 from Jenish and Prucha (2009)). For simplicity, we refer in the text to Assumption 2.1(v) as a weak dependence assumption. Importantly, Assumption 2.1 allows for unrestricted spatial correlation among the treated units.

Since we focus on settings in which researchers do not have information on what generates the spatial correlation or they are not willing to assume such structure, we do not need to model in detail the sources of spatial correlation. As a concrete example, we can consider a setting in which the N_1 treated units are closely located geographically, but we have a larger number of control units in different locations. In this case, if spatial correlation goes to zero when geographical distance increases, then we would have Assumption 2.1(v) satisfied, even though we may have arbitrarily strong spatial correlation among the N_1 treated units. The spatial correlation may also arise from other mechanisms not related to geographical distance (for example, units with similar industry shares may have more correlated errors). See also Appendix A2 from Ferman (2019) for an example of a spatial correlation model in a DID setting that would satisfy this weak dependence condition. Importantly, we consider a setting in which the applied researcher may be unaware or may not have information on

the relevant distance metrics in the cross section, which is common in DID applications. We further discuss the validity of this assumption in Remark 3.4.

Moreover, while we consider a setting in which the treatment allocation is fixed, we can think about an underlying model in which treatment assignment and errors are stochastic, and units that experience correlated shocks are more likely to receive the same treatment status. In this case, once we condition on the treatment allocation, we would have a setting in which the errors of the treated units are correlated.

Under Assumption 2.1, it follows from equation (3) that the limiting distribution of $\widehat{\alpha}$ when N_1 is fixed and $N_0 \to \infty$ is given by $\widehat{\alpha} \xrightarrow{p} \alpha + \widetilde{W}$, where $\widetilde{W} \equiv {N_1}^{-1} \sum_{s \in \mathcal{I}_1} W_s$. This result comes essentially from Proposition 1 of CT, with the minor difference that we allow for scale changes in the distribution of W_s. The intuition is that, given the weak dependence condition (Assumption 2.1(v)), the average of the control errors converges in probability to zero as $N_0 \to \infty$. However, since the number of treated units remain fixed, the distribution of $\widehat{\alpha}$ will still depend on the errors of the treated units even asymptotically. In this case, the estimator is unbiased, but not consistent.

3 Inference with spatially correlated shocks

CT propose an interesting inference method in this setting by noting that the residuals $\widehat{W}_s$ of the control units may be informative about the distribution of W_s and, therefore, about the distribution of $\widehat{\alpha}$. In their running model, the main idea is to estimate the distribution of W_s using the empirical distribution of $\{\widehat{W}_s\}_{s \in \mathcal{I}_0}$. Assuming that the distribution of W_s is the same across units, this works because, as $N_0 \to \infty$, $\widehat{W}_s \xrightarrow{p} W_s$ for $s \in \mathcal{I}_0$.[4] Therefore, hypotheses tests and confidence intervals can be constructed based on the empirical distribution of $\{\widehat{W}_s\}_{s \in \mathcal{I}_0}$. FP consider a setting in which we may have scale changes in the distribution of W_s, as present in Assumption 2.1(ii). In this case, we would approximate the distribution of

[4] CT consider in their appendix other inference procedures that allow for some types of heteroskedasticity and spatial correlation (if a distance metric is available).

W_s for a treated unit s with the empirical distribution of $\{h(Z_s, \hat{\delta})\widehat{W}_q / h(Z_q, \hat{\delta})\}_{q \in \mathcal{I}_0}$. That is, we estimate the heteroskedasticity parameters δ, and used this estimated structure so that the residuals of the control units become informative about the distribution of the errors of the treated units.

In their running model, CT show validity of their inference method when W_s is independent across s (Assumption 2 in CT). They also provide in their appendix an example model that allows spatial dependence. However, in this case they require an observed distance metric across units, and correct specification of the spatial correlation structure. We consider instead a setting in which such distance metric is not available, as is common in DID applications. FP also assume in their main analysis that W_s is independent across s.[5]

It is easy to understand the importance of this independence assumption when $N_1 > 1$ for the standard methods proposed by CT and FP. Consider a simple case in which $N_1 = 2$, and $\{W_1, W_2\}$ is multivariate normally distributed with correlation ρ. For simplicity, consider the case in which $h(Z_s, \delta)$ is constant. Under Assumption 2.1, $\widehat{\alpha} \xrightarrow{p} \alpha + \widetilde{W}$, where the variance of $\widetilde{W}$ is given by $2^{-1}(1+\rho)\mathbb{V}(W_s)$. However, given Assumption 2.1(v), if we consider two random draws from $\{\widehat{W}_s\}_{s \in \mathcal{I}_0}$ to recover the distribution of $\widetilde{W}$, the correlation between these draws would converge to zero when $N_0 \to \infty$. As a consequence, the standard approach proposed by CT would recover a distribution for $\widetilde{W}$ that has a variance given by $2^{-1}\mathbb{V}(W_s)$. Therefore, the variance of $\widehat{\alpha}$ would be underestimated if $\rho > 0$, leading to over-rejection.[6] The same problem applies for the inference method proposed by FP.

3.1 Validity of existing methods when $N_1 = 1$

When $N_1 = 1$, however, we show that the inference methods proposed by CT and FP can remain valid even if we allow for spatial correlation, and even when we do not have information on the relevant distance metric. The main intuition is that, under Assumption

[5] In Section IV, they consider an alternative that allows for spatial correlation, but that relies instead on a large number of pre-treatment periods.

[6] If $\rho < 0$, then we would have under-rejection.

2.1, the asymptotic distribution of $\widehat{\alpha}$ depends only on W_1, and the distribution of W_1 can still be asymptotically approximated using the residuals from the controls.

Let $\widehat{F}_\xi(c) = N_0^{-1} \sum_{s \in \mathcal{I}_0} \mathbb{1}\{\widehat{W}_s / h(Z_s, \hat{\delta}) \leq c\}$, where $\hat{\delta}$ is an estimator for δ. We first show that $\widehat{F}_\xi(c)$ approximates the distribution of ξ_s if $\hat{\delta}$ is consistent.

Proposition 3.1 *Suppose Assumption 2.1 holds, and $\hat{\delta}$ is a consistent estimator for δ when $N_0 \to \infty$. Then, as $N_0 \to \infty$, $\widehat{F}_\xi(c)$ converges in probability to $F_\xi(c)$ uniformly on any compact subset of the support of ξ.*

We present details in Appendix A.1. In this setting with $N_1 = 1$, the inference method proposed by FP would approximate the asymptotic distribution of $\widehat{\alpha}$ with the empirical distribution of $\{h(Z_1, \hat{\delta})\widehat{W}_s / h(Z_s, \hat{\delta})\}_{s \in \mathcal{I}_0}$. It immediately follows from Proposition 3.1 that the inference method proposed by FP remains valid for the case with $N_1 = 1$, even when we may have spatial correlation. To formalize this result, let ϕ_{FP} be an indicator variable equal to one if we reject the null hypothesis $H_0 : \alpha = \alpha_0$ for a τ nominal level test using the inference method proposed by FP.

Corollary 3.1 *Consider the problem of testing the null hypothesis $H_0 : \alpha = \alpha_0$, for some $\alpha_0 \in \mathbb{R}$. Suppose Assumption 2.1 holds, and $\hat{\delta}$ is a consistent estimator for δ when $N_0 \to \infty$. Then, if $N_1 = 1$ and the null is true, $\mathbb{E}[\phi_{FP} | Z_1 = z] \to \tau$ when $N_0 \to \infty$.*

Therefore, if we consider the setting with a single treated unit, the inference method proposed by FP remains asymptotically valid, even when we condition on the Z_1 for the treated unit. Note that if we set $h(Z_s, \delta)$ constant, and $h(Z_s, \hat{\delta}) = 1$, then Proposition 3.1 and Corollary 3.1 imply that the standard procedure proposed by CT in their running model is also asymptotically valid when $N_1 = 1$ and $N_0 \to \infty$, even when we have spatial correlation. We show in Appendix A.1 that these results remain valid when we include covariates in model (2).

Remark 3.1 While a standard permutation test, as proposed by Fisher (1935), would be valid under any structure on the potential outcomes *if treatment is randomly assigned*, note

that Assumption 2.1 does not require random assignment. For example, we allow for a setting in which the error of the treated unit has larger variance than the error of the control units. In this case, we show that FP would be valid, while a standard permutation test would lead to over-rejection.

3.2 Alternatives when $N_1 > 1$

As explained above, however, the inference methods proposed by CT and FP would not be valid when $N_1 > 1$ and errors are spatially correlated (considering the case in which we do not have information on a distance metric). Given this potential problem, we consider the case in which the errors of all treated units are perfectly correlated as a worst-case scenario to provide an asymptotically valid, though possibly conservative, inference method in this setting. More specifically, from Proposition 3.1, $\widehat{F}_{\xi}(c)$ asymptotically recovers the marginal distribution of ξ_s, even when we allow for spatial correlation. Given that, instead of considering independent draws from this distribution to put in place of the treated units, we sample only one $\widehat{\xi}_s = \widehat{W}_s/h(Z_s, \hat{\delta})$ and place it for all treated units. Therefore, we approximate the asymptotic distribution of $\widehat{\alpha}$ with the empirical distribution $\widehat{H}(c) = \frac{1}{N_0}\sum_{h\in\mathcal{I}_0} \mathbb{1}\{N_1^{-1}\sum_{s\in\mathcal{I}_1} h(Z_s, \widehat{\delta})\widehat{\xi}_h < c\}$. This would recover the asymptotic distribution of $\widehat{\alpha}$ if $\{W_s\}_{s\in\mathcal{I}_1}$ were perfectly correlated. When treated units are not perfectly correlated, though, we would recover a distribution for $\widehat{\alpha}$ that has a *higher* variance relative to the true distribution of $\widehat{\alpha}$. Let $\mathbf{Z}$ be a matrix with information on Z_s for all treated units. To formalize this idea, we consider the following high-level assumption.

Assumption 3.1 *(regularity condition)* For any $\tau \in (0,1)$, let c_τ be the $\tau-$quantile of the distribution of $N_1^{-1}\sum_{s\in\mathcal{I}_1} h(Z_s,\delta)\xi$ conditional on $\mathbf{Z} = \mathbf{z}$, and define $\widetilde{W} = N_1^{-1}\sum_{s\in\mathcal{I}_1} W_s$. Then $Pr\left(\{\widetilde{W} < c_{\tau/2}\} \cup \{\widetilde{W} > c_{1-\tau/2}\}|\mathbf{Z} = \mathbf{z}\right) \leq \tau$.

If we consider the simpler case in which $h(Z_s, \delta)$ is constant, then this regularity condition simply means that, regardless of the spatial correlation among the treated units, the prob-

ability of having extreme values for the average of the treated units, $\widetilde{W}$, is weakly smaller than the probability of having extreme values for a single draw of W_s. We analyze this assumption in detail. We focus on the case in which $h(Z_s, \delta)$ is constant, but all arguments remain valid when we consider a setting in which we may have variation in $h(Z_s, \delta)$.

First, note that Assumption 3.1 is satisfied if $\{W_s\}_{s \in \mathcal{I}_1}$ is multivariate normal. In this case, $\widetilde{W}$ would also be normally distributed, and we have that $\mathbb{V}(\widetilde{W}) \leq \mathbb{V}(W_s)$ irrespectively of the spatial correlation among the treated units. Therefore, $\widetilde{W}$ will be less likely to attain extreme values than W_s. This is valid for any value of the spatial correlation, even when treated units are negatively correlated. Moreover, note that in many DID applications we have that Y_{st} is the aggregate of individual-level observations Y_{ist}. If we can apply a central limit theorem to these averages, then the aggregate errors would asymptotically be multivariate normal (when the number of individuals in each unit $\times$ time cell increases). Therefore, in settings in which Y_{st} is the aggregate of many individual-level observations (for example, employment rate or average income in a state s at time t), we should also expect Assumption 3.1 to hold.

More generally, for any distribution of W_s (whether or not it is multivariate normal), we can always guarantee that $\mathbb{V}(\widetilde{W}) \leq \mathbb{V}(W_s)$. However, we may find some extreme examples in which Assumption 3.1 may not hold. To provide more intuition on this assumption, we present in Appendix A.3 an example in which this assumption may not hold. To engineer such example, we had to consider a distribution for W_s that is bimodal, and such that the two peaks are very far apart, which is not something we should expect in common empirical applications. In this example, when the two peaks are not very far apart, then Assumption 3.1 would still hold.

Importantly, we note that applied researchers can evaluate the marginal distribution of $\widehat{W}_s$ in their empirical applications. Therefore, if the marginal distribution does not exhibit such large peaks in the tails of the distribution, then they should have more confidence that Assumption 3.1 is valid. Moreover, from the discussion above, we note that we should not

expect to see distributions for the errors with such extreme feature when Y_{st} is the aggregate of many individual-level observations.

Finally, we also present in Appendix A.3 an exercise showing that this assumption would be satisfied if we consider the ACS data that we used to base our MC simulations. We consider either the case in which $\widetilde{W}$ is the average of two independent or of two highly correlated units. Therefore, Assumption 3.1 is reasonable for a dataset that is widely used in applied work.

It follows directly from Proposition 3.1 and Assumption 3.1 that this modified test asymptotically controls for size under Assumption 2.1. Let ϕ_{cons1} be an indicator variable equal to one if we reject the null hypothesis $H_0 : \alpha = \alpha_0$ for a τ nominal level test using this alternative inference method.

Proposition 3.2 *Consider testing the null hypothesis $H_0 : \alpha = \alpha_0$ with a significance level τ by contrasting $\widehat{\alpha} - \alpha_0$ with the $\tau/2$ and $1 - \tau/2$ quantiles of $\widehat{H}(c)$. Suppose Assumptions 2.1 and 3.1 hold, and we have an estimator $\widehat{\delta}$ that is consistent for δ when $N_0 \to \infty$. Then, if the null is true, $limsup_{N_0 \to \infty} \mathbb{E}[\phi_{cons1} | \mathbf{Z} = \mathbf{z}] \leq \tau$.*

Therefore, this conservative test provides a viable alternative in settings where $N_1 << N_0$ that is robust to weakly dependent spatial correlation. While we guarantee that this modified test does not over-reject (asymptotically), it will generally be conservative when $N_1 > 1$, unless the errors of treated units are perfectly correlated. In Appendix A.1, we extend this result for settings in which allow for covariates in model (2).

Remark 3.2 If a distance metric is available and the researcher is willing to assume such distance metric is the relevant one for the spatial correlation, then other available alternatives might present better power (for example, the inference method proposed in the appendix of CT). In this case, there would be a trade-off between a test that is more powerful but requires correct specification of the spatial correlation, versus a test that is less powerful,

but does not require correct specification of the spatial correlation. We also recall that the test we propose does not require observation of a distance metric in the cross section.

Remark 3.3 Related to Remark 3.2, another alternative to provide a more powerful test may be to infer about the spatial correlation using the time series. However, such alternatives would require a large time series,[7] while the alternatives we propose remain valid even when we have only one pre- and one post-treatment periods. Given the survey from Roth (2019), settings in which the time series dimension is short is prevalent in DID applications.

Remark 3.4 The assumption that W_s is weakly dependent would not be satisfied if there are unobserved shocks that affect all control units in the same way (and differently from how they affect the treated units, so they are not captured by the time fixed effects). In this case, the DID residuals $\widehat{W}_s$ would not capture these group-specific shocks, and $\widehat{F}_\xi(c)$ would underestimate the dispersion of the marginal distribution of ξ_s. As a consequence, even the conservative test may over-reject. Importantly, however, the over-rejection of the conservative test in this case would be no larger than the over-rejection of CT (or FP). In contrast, this weak dependence assumption may be reasonable when units closer in some distance metric have more correlated errors, but such correlation goes to zero when this distance increases. Such distance metric does not need to coincide with geographical distance (for example, we may have that units with similar industry shares have more correlated errors), and we do not even need to have information on the relevant distance metrics.

Remark 3.5 Constructing a conservative test becomes more complicated if treated units start treatment at different periods, as we discuss in Appendix A.4.

4 A more poweful test

While the conservative test presented in Section 3.2 is valid even when we have weakly dependent spatial correlation, such test may be too conservative, implying low power. We now

[7] For example, Vogelsang (2012), Ferman and Pinto (2019) (Section IV) and Chernozhukov et al. (2019).

show that it is possible to construct a test that is asymptotically valid and less conservative in common settings in which Y_{st} represents averages of M_{st} individual-level observations Y_{ist}. We can consider either the case in which the econometrician observes only unit aggregates Y_{st} or the individual-level data Y_{ist}.[8]

To illustrate this setting, we consider a simplified model in which the $N = N_1 + N_0$ units are partitioned into $F = N/2$ pairs, $\Lambda_1, \ldots, \Lambda_F$, and potential outcomes of individual i, in unit s, and period t, are given by

$$\begin{cases} Y_{ist}(0) = \theta_s + \gamma_t + \sum_{f=1}^{F} \omega_{ft} \mathbb{1}\{s \in \Lambda_f\} + \nu_{st} + \epsilon_{ist} \\ Y_{ist}(1) = \alpha + Y_{ist}(0), \end{cases} \tag{4}$$

where ω_{ft} generates spatial correlation across units in the same partition Λ_f, and ν_{st} generates spatial correlation for individual within the same unit. To simplify the exposition, we consider the case in which $\{\omega_{ft}\}_{t=1}^{T}$ is iid across f, $\{\nu_{st}\}_{t=1}^{T}$ is iid across s, and we observe a panel with individuals with $\{\epsilon_{ist}\}_{t=1}^{T}$ iid across i and s. Moreover, we assume that these variables are multivariate normal and mutually independent. Note that we do not impose any restriction on the serial correlation for these variables. We also assume for simplicity that treatment effects are homogeneous. We relax these simplifying assumptions below.

Consider the case in which $N_1 = 2$, with numbers of individual-level observations M_1 and M_2 that do not vary with t for the treated units. Since we focus on settings in which we may have relevant spatial correlation, but there is no available distance metric across unit, we consider the case in which (i) the econometrician does not have information on the partitions $\Lambda_1, \ldots, \Lambda_F$, and (ii) the two treated units belong to the same partition (so their errors are spatially correlated). Considering the case in which both treated units start treatment at the same period, the DID estimator using individual-level data is numerically equivalent to the DID estimator using aggregate data with population weights. Let $\widetilde{\alpha}$ be this weighted

[8] In case we observe individual-level data, then the idea in these approaches is to aggregate at the unit level to take within unit correlation into account.

DID estimator. Under Assumption 2.1, we have that

$$\widetilde{\alpha} \xrightarrow{p} \alpha + \frac{M_1}{M_1 + M_2} W_1 + \frac{M_2}{M_1 + M_2} W_2, \tag{5}$$

where

$$W_s = \sum_{f=1}^{F} \nabla\omega_f \mathbb{1}\{s \in \Lambda_f\} + \nabla\nu_s + \frac{1}{M_s} \sum_{i=1}^{M_s} \nabla\epsilon_{is}. \tag{6}$$

For a given unit s with number of observations $M_s = m$, we have that

$$\mathbb{V}(W_s | M_s = m) = \mathbb{V}(\nabla\omega_f) + \mathbb{V}(\nabla\nu_s) + \frac{1}{m}\mathbb{V}(\nabla\epsilon_{is}) = A + B/m, \tag{7}$$

where $A = \mathbb{V}(\nabla\omega_f) + \mathbb{V}(\nabla\nu_s)$ and $B = \mathbb{V}(\nabla\epsilon_{is})$.

The main idea from FP is to estimate A and B using the residuals from the OLS regression, and then re-scale the residuals from the control units, so that they become informative about the distribution of the errors of the treated units.[9] Instead of considering the empirical distribution of $\{\widehat{W}_q\}_{q \in \mathcal{I}_0}$ to approximate the distribution of W_s for a treated unit $s \in \mathcal{I}_1$, FP consider the empirical distribution of $\{\widehat{W}_q \times (\hat{A} + \hat{B}M_s^{-1})^{1/2} / (\hat{A} + \hat{B}M_q^{-1})^{1/2}\}_{q \in \mathcal{I}_0}$.

As we discussed in detail in Section 3, this procedure would not work well in this setting with $N_1 > 1$ and spatial correlation. The problem is that the asymptotic distribution of $\widetilde{\alpha} - \alpha$, conditional on $(M_1, M_2) = (m_1, m_2)$, is $N(0, \sigma^2_{\text{true}})$, where

$$\sigma^2_{\text{true}} \quad = \quad \mathbb{V}(\nabla\omega_f) + \frac{m_1^2 + m_2^2}{(m_1 + m_2)^2}\mathbb{V}(\nabla\nu_s) + \frac{1}{m_1 + m_2}\mathbb{V}(\nabla\epsilon_{is}). \tag{8}$$

In contrast, a naive approach suggested by FP, without taking into account spatial cor-

[9] More specifically, we regress $\widehat{W}_s^2$ on a constant and $1/M_s$.

relation, would asymptotically recover a normal distribution with variance

$$\sigma^2_{\text{naive}} = \frac{m_1^2 + m_2^2}{(m_1+m_2)^2}(\mathbb{V}(\nabla\omega_f) + \mathbb{V}(\nabla\nu_s)) + \frac{1}{m_1+m_2}\mathbb{V}(\nabla\epsilon_{is}) = \sigma^2_{\text{true}} - \frac{2m_1m_2}{(m_1+m_2)^2}\mathbb{V}(\nabla\omega_f), \quad (9)$$

which will under-estimate the true asymptotic variance of $\widetilde{\alpha}$ if $\mathbb{V}(\nabla\omega_f) > 0$, leading to over-rejection.

The conservative test presented in Section 3.2 would consider an extreme scenario for the spatial correlation between W_1 and W_2, asymptotically recovering a distribution for $\widetilde{\alpha}$ with variance given by

$$\sigma^2_{\text{cons1}} = \left[\left(\frac{m_1}{m_1+m_2}\right)\left(A + B{m_1}^{-1}\right)^{\frac{1}{2}} + \left(\frac{m_2}{m_1+m_2}\right)\left(A + B{m_2}^{-1}\right)^{\frac{1}{2}}\right]^2 \geq \sigma^2_{\text{true}}. \quad (10)$$

We refer henceforth to this test as "conservative test 1." As discussed in Section 3.2, a potential problem with this test is that it may be very conservative, leading to substantial loss in terms of power.

We consider an alternative that is less conservative for this setting. We note that, given the structure on the errors considered in Model (4), the correlation between two individuals in different units is weakly smaller than the correlation between two individuals in the same unit. Therefore, a worst-case scenario for the spatial correlation in this setting would be one in which the correlation between individuals in different treated units is the same as the correlation between two individuals in the same unit. In this case, it would be as if we only had one treated unit, which is a combination of units 1 and 2.

Therefore, we propose an alternative inference method in which we consider that we only have a single treated unit with $M_1 + M_2$ observations. In this simpler case in which M_{st} does not vary with t, we can generate a single treated unit that is a weighted average of the treated units, $\widetilde{y}_t = \frac{M_1}{M_1+M_2}y_{1t} + \frac{M_2}{M_1+M_2}y_{2t}$, with a number of observations $\widetilde{M} = M_1 + M_2$. In this case, the DID estimator using this collapsed treated unit with $\widetilde{M}$ and $\{M_s\}_{s\in\mathcal{I}_0}$ as weights is numerically the same as the one using the original data with $\{M_s\}_{s\in\mathcal{I}_0\cup\mathcal{I}_1}$ as weights. Then

we can apply the inference method proposed by FP for this modified DID model with a single treated unit. We name this procedure as "conservative test 2." In this case, conditional on $(M_1, M_2) = (m_1, m_2)$, we would asymptotically recover a distribution with variance equal to

$$\sigma^2_{\text{cons2}} = A + B\left(m_1 + m_2\right)^{-1}. \tag{11}$$

Importantly, we have that $\sigma^2_{\text{cons2}} \leq \sigma^2_{\text{cons1}}$, so this alternative would be less conservative than the conservative test 1, considered in Section 3.2. Moreover, we have that $\sigma^2_{\text{cons2}} \geq \sigma^2_{\text{true}}$, so we would still be conservative to the true variance of the asymptotic distribution of $\widehat{\alpha}$. Intuitively, the main assumption behind this approach is that individuals in the same unit are more spatially correlated than individuals in different units. Therefore, we can use information on the within-unit correlation as an upper bound to the across-unit correlation.

Note that we would have equality $\sigma^2_{\text{cons2}} = \sigma^2_{\text{true}}$ when $\mathbb{V}(\nabla\nu_s) = 0$, meaning that all spatial correlation (if any) comes from the shocks ω_{ft}. In this case, this test would remain valid and it would not be conservative. In contrast, we would generally have $\sigma^2_{\text{cons1}} > \sigma^2_{\text{true}}$, so that the conservative test 1 would still be conservative in this case.

We formalize the conservative test 2 in Appendix A.2. Importantly, the exact structure on the errors presented in Model (4) is not crucial for this approach. Let $\mathbf{M} = (M_1, ..., M_{N_1})$, $\mathbf{m} = (m_1, ..., m_{N_1})$, and $\mathcal{W} = \sum_{s\in\mathcal{I}_1} \frac{M_s}{\sum_{s'\in\mathcal{I}_1} M_{s'}} W_s$. The key assumptions we need are that (i) $W_s = [A + B/M_s]^{1/2}\xi_s$, where ξ_s has the same marginal distribution for all $s \in \mathcal{I}_1 \cup \mathcal{I}_0$, and (ii) $Pr\left(\{\mathcal{W} < c_{\tau/2}\} \cup \{\mathcal{W} > c_{1-\tau/2}\}|\mathbf{M} = \mathbf{m}\right) \leq \tau$ for any $\tau \in (0, 1)$, where c_τ be the $\tau-$quantile of $[A + B/(\sum_{s'\in\mathcal{I}_1} m_{s'})]^{1/2}\xi_s$.

Condition (i) states that the marginal distribution of W_s is the same for all units, up to a scale parameter that depends on the number of observations in unit s. FP show that this parametric form for the conditional variance holds under a wide range of structures on the errors. In particular, we do not need to impose the restrictive structure from Model (4) that we assumed for expositional purposes.[10] While the condition that variation in the

[10]We can relax the assumption that the within-unit correlation comes from a unit-level shock that affects

number of observations across units generate only scale changes in the aggregate distribution of the errors can be guaranteed assuming that all shocks are multivariate normal, we can see this condition as an approximation in case we relax this multivariate normal condition. Simulations from FP and from a previous version of this paper (Ferman, 2020) using real datasets without imposing normality suggest that such approximation is reasonable. Moreover, Ferman and Pinto (2017) provide in their appendix a series of MC simulations in which errors are not normally distributed, and the approach proposed by FP also perform well.

Condition (ii) is equivalent to the regularity condition we present in Assumption 3.1. With multivariate normal shocks, this is valid under the condition discussed above that individuals within the same unit are more correlated than individuals across different units. Under this condition, regardless of the distribution of W_s, we guarantee that $\mathbb{V}(\mathcal{W}|\mathbf{M}=\mathbf{m}) \leq [A + B/(\sum_{s' \in \mathcal{I}_1} m_{s'})]$. Therefore, under multivariate normality, this implies that condition (ii) would be satisfied. More generally, this inequality does not guarantee condition (ii) when we relax the normality assumption. Still, as discussed in Section 3.2 and Appendix A.3, this is a reasonable condition in common applications.

Remark 4.1 We consider the case in which M_{st} varies across t in Appendix A.2.

5 Simulations with Real Datasets

We analyze the spatial correlation problem, and the proposed conservative tests, in simulations based on the ACS (Ruggles et al., 2015). We estimate a model for the spatial correlation in which the covariance between two PUMAs may depend on their distance, on whether they belong to the same state, and on their population sizes (details in Appendix

all individuals in the same unit $\times$ time in the same way. For example, we may have individuals in county c and state s, where individuals in the same state but different counties are spatially correlated, but individuals in the same county are relatively more spatially correlated. Importantly, such more complex within-unit dependence does *not* need to be observed nor modeled by the applied researcher. We can also have either a panel or a repeated cross-section of individual level observations in each unit. See FP for more details. Moreover, the structure for the across units correlation can be substantially more general.

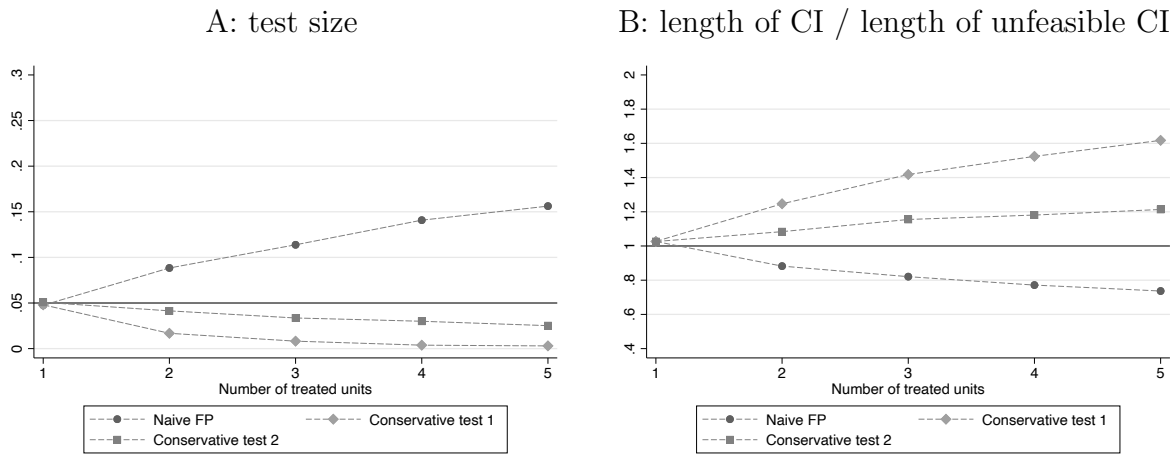
A.5). If we consider all pairs of PUMAs, the correlation between their errors in this estimated model is greater than 0.05 in only 0.42% of the cases. Therefore, the weak dependence condition we consider seems reasonable in this setting. Still, we have PUMAs with spatial correlation as strong as 0.30. Therefore, we may have a setting in which the N_1 treated PUMAs exhibit strong spatial correlation.

In line with our theoretical model, we consider a setting in which we fix $N_1 \in \{1, 2, 3, 4, 5\}$ PUMAs as the treated ones, and generate multivariate normal draws of W_s with this estimated spatial dependence. To illustrate issues related to spatial correlation, we choose the PUMA with the strongest spatial correlation with some other PUMA in our dataset to be treated, and then include the $N_1 - 1$ PUMAs that are most correlated with this one as the remaining treated PUMAs. We consider three different inference procedures: (i) the naive version of FP approach (which assumes errors are independent across PUMAs), (ii) the conservative test presented in Section 3.2 (which we refer to as "conservative test 1"), and (iii) the conservative test proposed in Section 4 (which we refer to as "conservative test 2"). In Appendix A.5, we show that other alternatives, such as inference based on cluster robust variance estimator (CRVE) and on wild cluster bootstrap (WCB) perform poorly in our setting, both because the number of treated units is very small, and because we have spatial correlation.[11]

Figure 1.A presents the size of these tests when we vary N_1, while Figure 1.B presents the ratio of length of the confidence intervals relative to the length of an unfeasible confidence interval using the distribution of $\widehat{\alpha}$ in the simulations.

[11] See, for example, MacKinnon and Webb (2017) and Djogbenou et al. (2019) for discussions on asymptotic approximations for CRVE and WCB.

Figure 1: **MC simulations**



Notes: Figure A presents rejection rates for the three difference inference methods we discuss in Section 5 for different values of N_1. Figure B presents information on the ratios between the average length of the CI relative to the length of the CI considering the MC distribution of $\widehat{\alpha}$.

Overall, these simulations highlight the main messages of the paper: (1) the inference methods proposed by CT and FP remain valid when we have a single treated unit, even when errors are spatially correlated; (2) with $N_1 > 1$, these inference methods over-reject when errors are spatially correlated, and the over-rejection is increasing with N_1; (3) the proposed conservative tests control for size, although they may be conservative when $N_1 > 1$; (4) we are able to provide a less conservative test exploiting the structure of common empirical applications in which units are aggregates of individual-level observations to construct a more powerful test (even if we do not have information on the individual-level observations).

Note that the length of the confidence intervals of the conservative test 2 is only 20% larger than the unfeasible confidence intervals when $N_1 = 5$ (in contrast, the conservative test 1 has a 60% larger confidence interval). Of course, if we had information on the relevant distance metrics in the cross section, then it would be possible to correct for spatial correlation with a more powerful test. Still, these simulations illustrate that it is possible to construct valid tests even when such information is unavailable or the researcher is not willing to impose a

structure on the spatial correlation.

6 Empirical Illustration

We illustrate our findings analyzing the effects of the Massachusetts 2006 health care reform. This reform was analyzed by Sommers et al. (2014) using a DID design comparing 14 Massachusetts counties with 513 control counties from 45 different states that were selected based on a propensity score to be more similar with the treated counties.[12] Sommers et al. (2014) find a reduction of 2.9%-4.2% in mortality in Massachusetts relative to the controls after the reform (depending on whether covariates are included). Their inference procedures were then re-analyzed by Kaestner (2016).

We revisit the use of different inference methods for this empirical application in light of our findings.[13] Table 1 presents p-values and the length of confidence intervals for a wide number of different inference methods.[14]

The first conclusion is that p-values vary wildly across inference methods, which reflects the fact that some inference methods are not appropriate for this empirical application. Since (i) we have only a single treated state, (ii) errors are possibly correlated across counties within the same state, and (iii) there is variation in population sizes across counties/states that may lead to heteroskedasticity, we consider FP at the state level as the most appropriate inference

[12]The propensity score used age distribution, sex, race/ethnicity, poverty rate, median income, unemployment, uninsured rate, and baseline annual mortality as predictors. We take this first selection step as given in our analysis. We find similar results if we consider a DID regression using all counties, so that there is no pre-selection of control counties.

[13]We consider a DID estimator based on OLS TWFE regression with no covariates. Point estimates are slightly different than reported in the original paper because we use the publicly available data set (so we do not have death counts for cells with fewer than 10 deaths), and because we weight observations by population mean across years to avoid problems with TWFE estimand highlighted in recent papers. See, for example, Goodman-Bacon (2018), de Chaisemartin and D'Haultfoeuille (2018), and Callaway and Sant'Anna (2018). We also restrict to counties with information non-missing information for all years. We end up with 485 control counties, as compared to 513 from the original study.

[14]The difference inference methods are, in order presented in Table 1 , (1) clustered standard errors at the state level, (2) WCB with the null imposed at the state level, (3) WCB without the null imposed at the state level, (4) clustered standard errors at the county level, (5) WCB with the null imposed at the county level, (6) WCB without the null imposed at the county level, (7) CT at the state level, (8) FP at the state level, (9) CT at the county level, (10) FP at the county level, (11) the conservative inference method proposed in Section 3.2, and (12) the conservative inference method proposed in Section 4.

method for this application, and use the p-values and CIs from this inference method as a benchmark.

However, such alternative would be unfeasible in case we did not have information on the relevant distance metric (in this application, information on the states), or if we had a more complex setting in which we had states with both treated and control counties.[15] Therefore, we analyze the performance of the conservative tests we propose in this scenario in which aggregating at the state level would not be feasible. The advantage of analyzing these tests in this empirical illustration is that we can contrast our findings with the conclusions based on FP at the state level, which we use as a benchmark.

The conservative tests we propose in Sections 3.2 and 4 present p-values very similar to the ones from FP at the state level. The fact that we find similar results to FP at the state level reinforces that these conservative tests would have worked well even if we did not have information on a distance metric, and despite the fact that we have spatial correlation across counties. When we consider the conservative test proposed in Section 4, the length of the confidence intervals is only 7.5% larger for one of the outcomes, and 23% larger for the other. Therefore, there is some loss in terms of power, but in settings in which a distance metric is not available, this may be a cost we would have to pay to provide a test that controls for size in the presence of spatial correlation.

The results from Table 1 also illustrate some common potential problems in the use of alternative inference methods in this setting. First, clustering at the state level substantially underestimates uncertainty, which was the main criticism of Kaestner (2016) to the inference procedures considered by Sommers et al. (2014). As Ferman (2021) reports, it is not uncommon that applied researchers rely on such inference procedure, despite leading to large over-rejection. WCB at the state level also perform poorly in this setting.[16]

Moreover, most of the inference methods that exploit across-county variation present

[15] Note that within-state variation in treatment assignment does *not* imply that spatial correlation is innocuous.

[16] In this case with a single treated cluster, WCB without the null imposed leads to large over-rejection, while when we impose the null the CIs are extremely wide (MacKinnon and Webb, 2017).

very low p-values, which is consistent with the presence of spatial correlation across counties within the same state.[17] In Appendix A.6 we propose a test for spatial correlation in this setting, which reinforces that spatial correlation is indeed relevant in this application.

Finally, in all cases the p-values based on FP are smaller than the ones based on CT. This is consistent with the fact that MA counties are relatively larger than the control counties. Therefore, following the rationale from FP, a standard implementation from CT (without taking heteroskedasticity into account) would be too conservative.

Note that one of the approaches considered by Kaestner (2016) to revisit the findings from Sommers et al. (2014) was a permutation test at the county level. We show that this approach may lead to over-rejection due to state-level shocks, but also to under-rejection due to variation in population sizes. When we correct for these two problems, we find no evidence that the policy led to statistically significant effects on mortality. However, it is interesting to note that if we corrected only for heteroskedasticity, using FP at the county level, then we would have found very low p-values due to spatial correlation.

[17] The only exception is when we consider CT at the county level. This happens because the MA counties are relatively larger, which leads to under-rejection. See discussion in the next paragraphs.

Table 1: **Alternative inference methods for Sommers et al. (2014)**

	All cause mortality (deaths per 100,000 adults)		Health care-amenable mortality (deaths per 100,000 adults)	
	p-value	length of CI	p-value	length of CI
	(1)	(2)	(3)	(4)
CRVE and WCB at state level				
CRVE	0.025	22.34	0.001	11.61
WCB with null imposed	0.419	399.02	0.433	300.73
WCB without null imposed	0.000	11.36	0.000	8.03
CRVE and WCB at county level				
CRVE	0.002	16.08	0.001	11.36
WCB with null imposed	0.007	20.98	0.022	16.50
WCB without null imposed	0.009	21.62	0.016	16.31
CT and FP at state level				
CT	0.434	113.75	0.357	63.98
FP	0.344	58.94	0.210	32.04
CT and FP at county level				
CT	0.304	54.13	0.266	38.81
FP	0.062	26.72	0.013	14.83
Conservative tests				
Conservative test 1	0.463	79.24	0.319	41.47
Conservative test 2	0.393	72.60	0.229	34.47

Notes: This table presents p-values and lengths of confidence intervals for a series of alternative inference methods for the application from Sommers et al. (2014). The point estimates are -12.77 for all cause mortality and -9.94 for health care-amenable mortality.

7 Conclusion

We consider the problem of inference in DID models when there are few treated units and errors are spatially correlated. We first show that, when there is a single treated unit, the inference methods proposed by CT and FP, which were designed for settings with few treated and many control units, remain asymptotically valid when errors are weakly dependent. This extends the set of possible applications in which the tests proposed by CT and FP can be reliably used when there is only a single treated unit. However, these methods can lead to over-rejection with more than one treated unit. We propose two alternative inference methods that are asymptotically valid, though generally conservative, in the presence of spatial correlation. These tests provide interesting alternatives when spatial correlation is likely relevant, but the researcher does not have information on a distance metric or is not willing to impose a structure on the spatial correlation.

References

Arellano, M. (1987). Computing robust standard errors for within-groups estimators. *Oxford Bulletin of Economics and Statistics*, 49(4):431–434.

Bertrand, M., Duflo, E., and Mullainathan, S. (2004). How much should we trust differences-in-differences estimates? *Quarterly Journal of Economics*, page 24975.

Brewer, M., Crossley, T. F., and Joyce, R. (2017). Inference with difference-in-differences revisited. *Journal of Econometric Methods*, 7(1).

Callaway, B. and Sant'Anna, P. H. C. (2018). Difference-in-Differences with Multiple Time Periods and an Application on the Minimum Wage and Employment. Working Paper, arXiv:1803.09015 .

Cameron, A., Gelbach, J., and Miller, D. (2008). Bootstrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics*, 90(3):414–427.

Canay, I. A., Romano, J. P., and Shaikh, A. M. (2017). Randomization tests under an approximate symmetry assumption. *Econometrica*, 85(3):1013–1030.

Chernozhukov, V., Wuthrich, K., and Zhu, Y. (2019). An Exact and Robust Conformal Inference Method for Counterfactual and Synthetic Controls. Papers 1712.09089, arXiv.org.

Conley, T. G. and Taber, C. R. (2011). Inference with Difference in Differences with a Small Number of Policy Changes. *The Review of Economics and Statistics*, 93(1):113–125.

de Chaisemartin, C. and D'Haultfoeuille, X. (2018). Two-way fixed effects estimators with heterogeneous treatment effects.

Djogbenou, A., MacKinnon, J., and Nielsen, M. (2019). Asymptotic theory and wild bootstrap inference with clustered errors. *Journal of Econometrics*, 212(2):393–412.

Donald, S. G. and Lang, K. (2007). Inference with Difference-in-Differences and Other Panel Data. *The Review of Economics and Statistics*, 89(2):221–233.

Ferman, B. (2019). Inference in Differences-in-Differences: How Much Should We Trust in Independent Clusters? *arXiv e-prints*, page arXiv:1909.01782.

Ferman, B. (2020). Inference in Differences-in-Differences with Few Treated Units and Spatial Correlation. *arXiv e-prints*, page arXiv:2006.16997.

Ferman, B. (2021). Assessing Inference Methods. *arXiv e-prints*, page arXiv:1912.08772.

Ferman, B. and Pinto, C. (2017). Inference in differences-in-differences with few treated groups and heteroskedasticity. Mpra paper, University Library of Munich, Germany.

Ferman, B. and Pinto, C. (2019). Inference in differences-in-differences with few treated groups and heteroskedasticity. *The Review of Economics and Statistics*, 0(ja):null.

Fisher, R. (1935). *The design of experiments. 1935.* Oliver and Boyd, Edinburgh.

Goodman-Bacon, A. (2018). Difference-in-differences with variation in treatment timing. Working Paper 25018, National Bureau of Economic Research.

Jenish, N. and Prucha, I. R. (2009). Central limit theorems and uniform laws of large numbers for arrays of random fields. *Journal of Econometrics*, 150(1):86 – 98.

Kaestner, R. (2016). Did Massachusetts health care reform lower mortality? no according to randomization inference. *Statistics and Public Policy*, 3:1 – 6.

MacKinnon, J. G., Nielsen, M. O., and Webb, M. D. (2020). Testing for the appropriate level of clustering in linear regression models. Working Paper 1428, Economics Department, Queen's University.

MacKinnon, J. G. and Webb, M. D. (2017). Wild bootstrap inference for wildly different cluster sizes. *Journal of Applied Econometrics*, 32(2):233–254.

MacKinnon, J. G. and Webb, M. D. (2020). Randomization inference for difference-in-differences with few treated clusters. *Journal of Econometrics*, 218(2):435–450.

Marcus, M. and Sant'Anna, P. H. C. (2021). The role of parallel trends in event study settings: An application to environmental economics. *Journal of the Association of Environmental and Resource Economists*, 0(ja):null.

Newey, W. K. and McFadden, D. (1994). Chapter 36 large sample estimation and hypothesis testing. volume 4 of *Handbook of Econometrics*, pages 2111 – 2245. Elsevier.

Roth, J. (2019). Pre-test with caution: Event-study estimates after testing for parallel trends.

Ruggles, S., Genadek, K., Goeken, R., Grover, J., and Sobek, M. (2015). Integrated Public Use Microdata Series: Version 6.0 [Machine-readable database].

Sommers, B. D., Long, S. K., and Baicker, K. (2014). Changes in mortality after Massachusetts health care reform. *Annals of Internal Medicine*, 160(9):585–593. PMID: 24798521.

Sun, L. and Abraham, S. (2020). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*.

van der Vaart, A. W. (1998). *Asymptotic statistics*. Cambridge series in statistical and probabilistic mathematics. Cambridge University Press, Cambridge (UK), New York (N.Y.).

Vogelsang, T. J. (2012). Heteroskedasticity, autocorrelation, and spatial correlation robust inference in linear panel models with fixed-effects. *Journal of Econometrics*, 166(2):303 – 319.

A Appendix

A.1 Proof of main results

We consider a setting including covariates, which is more general than the setting we considered in the main text. We have that

$$Y_{st} = \alpha d_{st} + X_{st}'\beta + \theta_s + \gamma_t + (\alpha_{st} - \alpha)d_{st} + \eta_{st}, \tag{12}$$

where X_{st} is a $(p \times 1)$ vector of covariates. As in the main text, we consider $W_s = \nabla\eta_{st}$, which is the post-pre difference in average errors for each unit s. Moreover, we assume that we observe a $(q \times 1)$ vector Z_s that potentially generates heteroskedasticity on W_s. We do not impose any restriction on whether X_{st} and Z_s share common elements or not. Propositions 3.1 and 3.2 and Corollary 3.1 follow by considering $\beta = 0$ and that we do not observe X_{st}. We considered the case without covariates in the main text to simplify the exposition of the main results. We consider the following assumption instead of Assumption 2.1.

Assumption A.1 (i) $(\eta_{s1}, ..., \eta_{st})$ have expectation zero conditional on $(Z_s, X_{s1}, ..., X_{st})$, for both $s \in \mathcal{I}_0$ and $s \in \mathcal{I}_1$, (ii) $W_s = h(Z_s, \delta)\xi_s$, where ξ_s is independent of $(Z_s, X_{s1}, ..., X_{st})$ and has the same marginal cdf F_ξ for all $s \in \mathcal{I}_1 \cup \mathcal{I}_0$, with finite second moment and continuous distribution; (iii) $h(z, \delta)$ is a known function, where $\delta \in \mathbb{R}^m$ is an unknown parameter, and $h(z, \delta) \leq \bar{h}(\delta)$ for some function $\bar{h}(\delta)$; (iv) $(Z_s, \nabla X_s)$ has the same marginal cdf's $F_{z,x}$ for all $s \in \mathcal{I}_0$; (v) X_{st} has finite second moments, and the residuals of X_{st} and d_{st} after the projection on time and unit fixed effects are linearly independent and those residuals of X_{st} have variation in the limit; (v) let $(Z, \nabla X) \sim F_{z,x}$ and $\xi \sim F_\xi$, with $(Z, \nabla X) \perp \xi$. Then, for any $g(Z, \nabla X, \xi)$ such that $\mathbb{E}[|g(Z, \nabla X, \xi)|^2] < \infty$, $\frac{1}{N_0}\sum_{s\in\mathcal{I}_0} g(Z_s, \nabla X_s, \xi_s) \xrightarrow{p} \mathbb{E}[g(Z, \nabla X, \xi)]$.

We need to impose Assumption A.1.(i) on the η_{st}'s rather than on the linear combination W_s, and the extra Assumption A.1.(v), because of the covariates X_{st}. Under these assumptions, we have conditions such that β is consistently estimated when $N_0 \to \infty$. Instead

of considering these assumptions, it would have been possible to simply add a high-level assumption stating that we have enough variation in X_{st} so that β is consistently estimated. Finally, note that we can consider the homoskedastic case with covariates by setting $h(Z_s, \delta)$ constant.

First, under Assumption A.1, we can still apply Proposition 1 from CT, so $\widehat{\alpha} \xrightarrow{p} \alpha + {N_1}^{-1} \sum_{s \in \mathcal{I}_1} W_s$, where $W_s = \nabla \eta_{st}$. Now let $\dot{A}_s = A_{st} - \overline{A}_s - \overline{A}_t + \overline{A}$, where $\overline{A}_s = \frac{1}{T} \sum_{t \in \mathcal{T}_0 \cup \mathcal{I}_1} A_{st}$, $\overline{A}_t = \frac{1}{N_0+N_1} \sum_{s \in \mathcal{I}_0 \cup \mathcal{I}_1} A_{st}$, and $\overline{A} = \frac{1}{T} \frac{1}{N_0+N_1} \sum_{t \in \mathcal{T}_0 \cup \mathcal{T}_1} \sum_{s \in \mathcal{I}_0 \cup \mathcal{I}_1} A_{st}$. From Frisch–Waugh–Lovell theorem, it follows that $\widehat{\eta}_{st} = (\alpha - \widehat{\alpha}) \dot{d}_{st} + \dot{X}'_{st}(\beta - \widehat{\beta}) + \dot{\omega}_{st} + \dot{\eta}_{st}$, where we define $\omega_{st} = (\alpha_{st} - \alpha) d_{st}$. Therefore, for $s \in \mathcal{I}_0$,

$$\begin{aligned} \widehat{W}_s &= (\widehat{\alpha} - \alpha) \left(\frac{T - t^*}{T} \right) \left(\frac{N_1}{N_0 + N_1} \right) + \left(\nabla X_s - \nabla \overline{X} \right)' (\beta - \widehat{\beta}) + W_s - \overline{W} \qquad (13) \\ &= W_s + \widehat{v} + \left(\nabla X_s - \nabla \overline{X} \right)' (\beta - \widehat{\beta}), \qquad (14) \end{aligned}$$

where $\widehat{v} = (\widehat{\alpha} - \alpha) \left(\frac{T-t^*}{T} \right) \left(\frac{N_1}{N_0+N_1} \right) - \overline{W}$. We use the fact that $\nabla \dot{A}_s = \nabla A_s - \nabla \overline{A}$, where $\nabla \overline{A} = \frac{1}{T-t^*} \sum_{t \in \mathcal{T}_1} \overline{A}_t - \frac{1}{t^*} \sum_{t \in \mathcal{T}_0} \overline{A}_t$. Note that $\nabla \dot{\omega}_s = 0$ for $s \in \mathcal{I}_0$.

We assume we have a consistent estimator for δ, $\widehat{\delta}$. Note that this implicitly impose restrictions on the function $h()$ and on the distribution of Z_s. See FP for an example in which heteroskedasticity comes from variation in the number of observations per unit, and we have a consistent estimator of the parameters of this function $h()$.

We consider an estimator for the cdf of ξ_s given by the empirical distribution of $\widehat{\xi}_s = h(Z_s, \widehat{\delta})^{-1} \widehat{W}_s$ of the control units. Let $\widehat{F}_\xi(c) = {N_0}^{-1} \sum_{s \in \mathcal{I}_0} \mathbb{1}\{\widehat{\xi}_s \leq c\}$, and $F_\xi(c)$ be the cdf of ξ_s.

Proposition A.1 *Suppose Assumption A.1 holds, and we have an estimator $\widehat{\delta}$ that is consistent for δ when $N_0 \to \infty$ and N_1 is fixed. Then, as $N_1 \to \infty$, $\widehat{F}_\xi(c)$ converges in probability to $F_\xi(c)$ uniformly on any compact subset of the support of ξ.*

Proof. This proof follows similar steps from the proof of Proposition 2 from CT, but correcting for heteroskedasticity as FP do, and allowing for spatial correlation.

First, let

$$\widehat{\lambda}_s(c; a_1, a_2, a_3, a_4) = \mathbb{1}\left\{h(Z_s, a_1)^{-1}\left(W_s + a_2 + (\nabla X_s - a_3)'(\beta - a_4)\right) \leq c\right\}, \tag{15}$$

and

$$\lambda(c; a_1, a_2, a_3, a_4) = Pr\left(h(Z, a_1)^{-1}\left(W + a_2 + (\nabla X - a_3)'(\beta - a_4)\right) \leq c\right). \tag{16}$$

Note that $\lambda(c; a_1, a_2, a_3, a_4) = \mathbb{E}[\widehat{\lambda}_s(c; a_1, a_2, a_3, a_4)]$, and that $\lambda(c; \delta, 0, \mathbb{E}[\nabla X], \beta) = F_\xi(c)$. Moreover, from the characterization of $\widehat{W}_s$ above, we have that $\widehat{\lambda}_s(c; \widehat{\delta}, \widehat{v}, \nabla\overline{X}, \widehat{\beta}) = \mathbb{1}\{\widehat{\xi}_s \leq c\}$.

For any compact subspace Ω of the support of ξ_s,

$$\sup_{c\in\Omega}\left|\widehat{F}_\xi(c) - F_\xi(c)\right| = \sup_{c\in\Omega}\left|\frac{1}{N_0}\sum_{s\in\mathcal{I}_0}\widehat{\lambda}_s\left(c; \widehat{\delta}, \widehat{v}, \nabla\overline{X}, \widehat{\beta}\right) - \lambda(c; \delta, 0, \mathbb{E}[\nabla X], \beta)\right| \tag{17}$$

$$\leq \sup_{c\in\Omega}\left|\frac{1}{N_0}\sum_{s\in\mathcal{I}_0}\widehat{\lambda}_s\left(c; \widehat{\delta}, \widehat{v}, \nabla\overline{X}, \widehat{\beta}\right) - \lambda\left(c; \widehat{\delta}, \widehat{v}, \nabla\overline{X}, \widehat{\beta}\right)\right| + \tag{18}$$

$$+\sup_{c\in\Omega}\left|\lambda\left(c; \widehat{\delta}, \widehat{v}, \nabla\overline{X}, \widehat{\beta}\right) - \lambda(c; \delta, 0, \mathbb{E}[\nabla X], \beta)\right| \tag{19}$$

Now consider a compact subset of the parameter space of (a_1, a_2, a_3, a_4), Θ, where $(\delta, 0, \mathbb{E}[\nabla X], \beta)$ is an interior point. Then,

$$\sup_{c\in\Omega}\left|\frac{1}{N_0}\sum_{s\in\mathcal{I}_0}\widehat{\lambda}_s\left(c; \widehat{\delta}, \widehat{v}, \nabla\overline{X}, \widehat{\beta}\right) - \lambda\left(c; \widehat{\delta}, \widehat{v}, \nabla\overline{X}, \widehat{\beta}\right)\right| \leq \mathbb{1}\left\{(\widehat{\delta}, \widehat{v}, \nabla\overline{X}, \widehat{\beta}) \notin \Theta\right\} \tag{20}$$

$$+ \sup_{c\in\Omega, (a_1,a_2,a_3,a_4)\in\Theta}\left|\frac{1}{N_0}\sum_{s\in\mathcal{I}_0}\widehat{\lambda}_s\left(c; a_1, a_2, a_3, a_4\right) - \lambda\left(c; a_1, a_2, a_3, a_4\right)\right| \tag{21}$$

From Proposition 1 of CT we have that $\widehat{\beta} \xrightarrow{p} \beta$, and we assume $\widehat{\delta} \xrightarrow{p} \delta$. Since $\widehat{\alpha} = O_p(1)$ and $N_1(N_1+N_0)^{-1} = o(1)$, combined with Assumption A.1, we have $\widehat{v} = o_p(1)$. Finally, from Assumption A.1 $\nabla\overline{X} \xrightarrow{p} \mathbb{E}[\nabla X]$. Therefore, $(\delta, \widehat{v}, \nabla\overline{X}, \widehat{\beta}) \xrightarrow{p} (\delta, 0, \mathbb{E}[\nabla X], \beta)$, which implies

that $\mathbb{1}\{(\widehat{\delta}, \widehat{v}, \nabla\overline{X}, \widehat{\beta}) \notin \Theta\} = o_p(1)$.

For the other term, note that $\mathbb{E}[\widehat{\lambda}_s(c; a_1, a_2, a_3, a_4)] = \lambda(c; a_1, a_2, a_3, a_4)$. From Assumption A.1, we have that $N_0^{-1} \sum_{s \in \mathcal{I}_0} \widehat{\lambda}_s(c; a_1, a_2, a_3, a_4) \xrightarrow{p} \lambda(c; a_1, a_2, a_3, a_4)$. Moreover, $\widehat{\lambda}_s(c; a_1, a_2, a_3, a_4)$ is continuous with probability one and bounded by 1 for any c and (a_1, a_2, a_3, a_4), and Ω and Θ are compact. A simple adaptation from Lemma 2.4 of Newey and McFadden (1994) to allow for spatial dependence satisfying Assumption A.1(v) instead of iid sampling implies that the first term is also $o_p(1)$.

Finally, note that $h(Z, \hat{\delta})^{-1} \left(W + \widehat{v} + (\nabla X - \nabla\overline{X})'(\beta - \widehat{\beta}) \right) \xrightarrow{d} \xi \sim F_\xi$. Therefore, from Lemma 2.11 from van der Vaart (1998), the term in line 19 is $o_p(1)$, which completes the proof. ■

A.2 Conservative test 2

We consider a more general setting than the one considered in Model (4). Consider the following model

$$Y_{st} = \alpha d_{st} + \theta_s + \gamma_t + (\alpha_{st} - \alpha) d_{st} + \eta_{st}, \tag{22}$$

where we also observe M_s, which is the (time-invariant) number of individual-level observations that generate the aggregates Y_{st}.[18] We consider below the case in which M_{st} varies with t. With some abuse of notation because we used a different definition of α in the main text, we define $\alpha = \frac{1}{N_1} \frac{1}{T_1} \sum_{s \in \mathcal{I}_\infty} \sum_{t \in \mathcal{T}_1} m_s \alpha_{st}$, where $\{m_s\}_{s \in \mathcal{I}_1}$ are the realized number of observations for the treated units. Note that all of our analyses in this section are conditional on $\mathbf{M} = (M_1, \ldots, M_{N_1})$ being equal to $\mathbf{m} = (m_1, \ldots, m_{N_1})$. The following assumption is similar to Assumption 2.1.

Assumption A.2 (i) $\mathbb{E}[W_s | M_s] = 0$ for all $s \in \mathcal{I}_1 \cup \mathcal{I}_0$; (ii) for some parameters A and B, $W_s = (A + B/M_s)^{1/2} \xi_s$, where ξ_s is independent of M_s and has the same marginal

[18] If we observe the individual-level data, then the idea in these approaches is to aggregate at the unit level to take within-unit correlation into account.

cdf F_ξ for all $s \in \mathcal{I}_1 \cup \mathcal{I}_0$, with finite second moment and continuous distribution; (iii) M_s is uniformly bounded and has the same marginal cdf F_m with positive variance for all $s \in \mathcal{I}_0$; (iv) if $M \sim F_m$ and $\xi \sim F_\xi$, then, for any $g(M, \xi)$ such that $\mathbb{E}[|g(M,\xi)|^2] < \infty$, $\frac{1}{N_0}\sum_{s\in\mathcal{I}_0} g(M_s, \xi_s) \xrightarrow{p} \mathbb{E}[g(M,\xi)]$.

We consider in this case the DID estimator using M_s as weights, $\widetilde{\alpha}$. We have that

$$\widetilde{\alpha} = \alpha + \frac{1}{\sum_{s\in\mathcal{I}_1} M_s}\sum_{s\in\mathcal{I}_1} M_s W_s - \frac{1}{\sum_{s\in\mathcal{I}_0} M_s}\sum_{s\in\mathcal{I}_0} M_s W_s \tag{23}$$

Given Assumption A.2, we have that $\frac{1}{\sum_{s\in\mathcal{I}_0} M_s}\sum_{s\in\mathcal{I}_0} M_s W_s = o_p(1)$. Therefore, conditional on $(M_1, \ldots, M_{N_1}) = (m_1, \ldots, m_{N_1})$, we have that $\widetilde{\alpha} \xrightarrow{p} \alpha + \mathcal{W}$, where $\mathcal{W} = \frac{1}{\sum_{s\in\mathcal{I}_1} m_s}\sum_{s\in\mathcal{I}_1} m_s W_s$.

The following assumption is a regularity condition similar to Assumption 3.1.

Assumption A.3 *(regularity condition)* For any $\tau \in (0,1)$, let c_τ be the $\tau-$quantile of the distribution of $[A + B/(\sum_{s'\in\mathcal{I}_1} m_{s'})]^{1/2}\xi_s$. Then $Pr\left(\{\mathcal{W} < c_{\tau/2}\} \cup \{\mathcal{W} > c_{1-\tau/2}\} | \mathbf{M} = \mathbf{m}\right) \leq \tau$.

Let $\widehat{P}(c) = \frac{1}{N_0}\sum_{s\in\mathcal{I}_0} \mathbb{1}\{[\widehat{A} + \widehat{B}/(\sum_{s'\in\mathcal{I}_1} m_{s'})]^{1/2}\widehat{\xi}_s < c\}$. Also, let ϕ_{cons2} be an indicator variable equal to one if we reject the null hypothesis $H_0 : \alpha = \alpha_0$ for a τ nominal level test that contrasts $\widehat{\alpha} - \alpha_0$ with the $\tau/2$ and $1 - \tau/2$ quantiles of $\widehat{P}(c)$

Proposition A.2 *Consider testing the null hypothesis $H_0 : \alpha = \alpha_0$ with a significance level τ by contrasting $\widetilde{\alpha} - \alpha_0$ with the $\tau/2$ and $1 - \tau/2$ quantiles of $\widehat{P}(c)$. Suppose Assumptions A.2 and A.3 hold. Then, if the null is true, $limsup_{N_0\to\infty}\mathbb{E}[\phi_{cons2}|\mathbf{M} = \mathbf{m}] \leq \tau$.*

If M_{st} varies with t, then the original DID estimator would not be numerically the same as the one using an aggregate treated unit $\widetilde{y}_t$. In this case, we recommend first aggregating the data at the unit $\times$ time level, and then estimating the DID estimator using sampling weights $M_s^{\min} = \min_{t\in\mathcal{T}_0\cup\mathcal{T}_1}\{M_{st}\}$ (or $M_s^{\text{mean}} = T^{-1}\sum_{t=1}^{T} M_{st}$). Then we can conduct the conservative

inference method proposed above based on this alternative DID estimator. The advantage of using $M_s^{\min}$ is that we guarantee a conservative estimator for the variance of $\mathcal{W}$, which in this case will be a weighted average of W_s of the treated units using $M_s^{\min}$. The reason is that for each t we will calculate $\widetilde{y}_t = \sum_{s\in\mathcal{I}_1} \frac{M_s^{\min}}{\sum_{s'\in\mathcal{I}_1} M_{s'}^{\min}} Y_{st}$, where $M_{st} \geq M_s^{\min}$ for all $t \in \mathcal{T}_0 \cup \mathcal{T}_1$ and $s \in \mathcal{I}_1$. Note that this way we would also have that the estimand is a well-defined weighted average of the heterogeneous treatment effects, avoiding the problems highlighted by de Chaisemartin and D'Haultfoeuille (2018) for the TWFE estimator in settings in which we have variation in the number of observations per cell. Moreover, if M_{st} does not vary much on t, then the estimand and the resulting estimator should be close to the original DID estimator using individual-level data.

A.3 On the validity of Assumption 3.1

Consider first the case in which $h(Z,\delta)$ is constant. As discussed in the paper, Assumption 3.1 is satisfied if $\{W_s\}_{s\in\mathcal{I}_1}$ is multivariate normal. In this case, $\widetilde{W} \equiv \frac{1}{N_1}\sum_{s\in\mathcal{I}_1} W_s$ will also be normally distributed, and we have $\mathbb{V}(\widetilde{W}) \leq \mathbb{V}(W_s)$, implying that Assumption 3.1 is satisfied. The condition that W_s is normally distributed can be justified in settings in which Y_{st} is the average of many individual-level observations in unit s at time t, and we can apply a central limit theorem for these averages. In this case, $\{W_s\}_{s\in\mathcal{I}_1}$ would approximately be multivariate normal.

Overall, if W_s has the same marginal distribution, we always have that $\mathbb{V}(\widetilde{W}) \leq \mathbb{V}(W_s)$, regardless of the distribution of W_s. This happens because

$$\mathbb{V}\left(\widetilde{W}\right) = \frac{1}{N_1^2}\sum_s\sum_j \text{cov}(W_s, W_j) \leq \frac{1}{N_1^2}\sum_s\sum_j \mathbb{V}(W_s) = \mathbb{V}(W_s). \tag{24}$$

However, having a lower variance does not directly imply that the probability of having extreme outcomes is lower for $\widetilde{W}$ if we consider a generic distribution for W_s. One case in which this assumption may not be valid for some values of τ is when the distribution of W_s

is bimodal, with the two peaks very far apart. To illustrate that, consider the case in which W_s is $N(-4.167, 1)$ with probability 0.96 and $N(100, 1)$ with probability 0.04. Therefore, we have that $\mathbb{E}[W_s] = 0$. If we consider $\tau = 0.1$, then $c_{0.05} \approx -5.792$ and $c_{0.95} \approx -1.857$. If we use these critical values to calculate $\tilde{p}$ when we have $N_1 = 5$, then we have $\tilde{p} \approx 0.18$. This happens because the probability that at least one of the observations come from the distribution with mean 100 is approximately 18% and, conditional on that, the probability that $\widetilde{W} > c_{0.95}$ is close to one.

We stress, however, that this is a distribution for the errors that we engineered to provide an example of a distribution in which Assumption 3.1 would not hold, and that this is not a kind of distribution for the errors we should expect in common applications. To construct this example, we need not only a bimodal distribution, but also the distance between that peaks to be very large. If instead we consider a setting in which W_s is $N(-0.042, 1)$ with probability 0.96 and $N(1, 1)$ with probability 0.04 (so, again, $\mathbb{E}[W_s] = 0$), then we would have $\tilde{p} \approx 0$. If we had instead that the mean of the second peak is 10, then we would have $\tilde{p} = 0.066 < 0.1 = \tau$. Overall, this example highlights that we need very extreme and unrealistic distributions for the errors so that Assumption 3.1 fails. Note also that in empirical applications it would be possible to observe the marginal distribution of $\widehat{W}_s$, and check whether it has this kind a shape that could invalidate Assumption 3.1. If we find evidence that this marginal distribution does not have peaks or large mass on the tails, then we should be more comfortable with Assumption 3.1.

In order to provide even more evidence in favor of the validity of this assumption in common empirical applications, we provide evidence that this assumption is valid when we consider the ACS data that we base the MC simulations presented in Section 5. As presented in detail in Appendix A.5, we calculate the residuals $\widehat{W}_s$ at the PUMA level with the ACS data, and then estimate a model for the spatial correlation across PUMAs. In order to assess the validity of Assumption 3.1, for a given $\tau \in (0, 1)$, we first calculate the $\tau/2$ and $1 - \tau/2$ quantiles of the distribution of $\{\widehat{W}_s\}$, $c_{\tau/2}$ and $c_{1-\tau/2}$. Then, we use the estimated

correlation matrix to identify for each PUMA the other PUMA that is most correlated with it, and calculate the proportion of pairs in which the average of these two PUMAs is smaller than $c_{\tau/2}$ or greater than $c_{1-\tau/2}$. For all $\tau \in \{0.01, 0.02, \ldots, 0.99\}$, we find that this proportion is smaller than τ. We reach the same conclusion if we consider the average of two random PUMAs. Therefore, Assumption 3.1 is reasonable in the application we consider for our MC simulations.

When we consider the setting in which $h(Z, \delta)$ is not constant, we also have that $\mathbb{V}(\widetilde{W}|\mathbf{Z} = \mathbf{z}) \leq \mathbb{V}(N_1^{-1} \sum_{s \in \mathcal{I}_1} h(z_s, \delta)\xi)$. In this case,

$$\begin{aligned} \mathbb{V}\left(\widetilde{W}|\mathbf{Z} = \mathbf{z}\right) &= \frac{1}{N_1^2} \sum_s \sum_j \text{cov}(W_s, W_j | Z_s = z_s, Z_j = z_j) \leq \frac{1}{N_1^2} \sum_s \sum_j h(Z_s, \delta) h(z_j, \delta) \mathbb{V}(\xi) \\ &= \mathbb{V}(\xi) \frac{1}{N_1^2} \left(\sum_s h(z_s, \delta) \right)^2 = \mathbb{V}\left(N_1^{-1} \sum_{s \in \mathcal{I}_1} h(z_s, \delta)\xi \right). \end{aligned} \tag{25}$$

Again, while this result guarantees that Assumption 3.1 holds if ξ is normally distributed, it would be possible to engineer examples in which it does not hold. As before, however, such examples would generally consider extreme and unrealistic distributions that we should not expect to see in common empirical applications.

A.4 Variation in treatment timing

Constructing a conservative test as proposed in Section 3 becomes more complicated if treated units start treatment at different periods. For example, consider that unit 1 starts treatment after t_1, while unit 2 starts treatment after t_2. In this case, the asymptotic distribution of $\widehat{\alpha}$ would depend on the linear combinations of the errors $W_1(1) = \frac{1}{T-t_1} \sum_{t=t_1+1}^{T} \eta_{1t} - \frac{1}{t_1} \sum_{t=1}^{t_1} \eta_{1t}$, and $W_2(2) = \frac{1}{T-t_2} \sum_{t=t_2+1}^{T} \eta_{2t} - \frac{1}{t_2} \sum_{t=1}^{t_2} \eta_{2t}$. We can still consistently estimate the marginal distributions of $W_1(1)$ and $W_2(2)$ by considering the appropriate linear combination of the residuals from the control units. This is what CT and FP do, and works well in a setting in which $W_1(1)$ and $W_2(2)$ are independent. When we allow for spatial dependence,

however, it becomes harder to define a worst case scenario. The worst case scenario will generally *not* be such that $\text{corr}(\eta_{1t}, \eta_{2t}) = 1$. To see that, suppose there are 3 time periods, with $t_1 = 1$ and $t_2 = 2$. In this case, $\widehat{\alpha} - \alpha \xrightarrow{p} 0.5(0.5\eta_{13} + 0.5\eta_{12} - \eta_{11}) + 0.5(\eta_{23} - 0.5\eta_{22} - 0.5\eta_{21})$. Therefore, assuming $\text{corr}(\eta_{1t}, \eta_{2t}) = 1$ will lead to a *lower* variance relative to the case with no spatial dependence if $\mathbb{V}(\eta_{i2})$ is substantially larger than the variance at the two other periods. A possible alternative in this case could be to first estimate the variance/covariance matrix of the marginal distribution of $(\eta_{s1}, ..., \eta_{st})$ for the treated units using the residuals of the control units. These estimated marginal distributions would be the same for all treated units under the stationarity assumption from CT, or will vary depending on the estimated heteroskedascity as considered by FP. Then we can calculate the spatial correlation parameters among the treated units (which will be $N_1(N_1 - 1)$ symmetric $T \times T$ matrices) that maximizes the variance of $\widehat{\alpha}$ to construct a worst case scenario. To implement that, we could follow a similar strategy as the one considered by CT to deal with spatial correlation in their appendix. The difference is that, since we cannot estimate the spatial dependence because there is no distance metric in our setting, we would have to do a grid search over the possible variance-covariance matrices that are consistent with the marginal distributions of the errors, and consider the worst case scenario.

A.5 Details on the MC simulation

We construct our MC simulations based on the American Community Survey (ACS). We restrict the sample to women between the ages 25 and 50, and consider log wages as outcome variable. With data from 2015 to 2018, we estimate the residuals $\widehat{W}_s$ at the PUMA level, considering a setting in which that treatment starts after 2016. More specifically, we compute the residuals from a PUMA $\times$ year regression using PUMA and time fixed effects, and then calculate $\widehat{W}_s = (\hat{e}_{s,2018} + \hat{e}_{s,2017})/2 - (\hat{e}_{s,2016} + \hat{e}_{s,2015})/2$.

Given $\widehat{W}_s$ for all PUMAs, we estimate a model for the spatial correlation in which the covariance between two PUMAs may depend on their distance, on whether they belong to

the same state, and on their population sizes. More specifically, let Z_s be the centroid of PUMA s, and S_s be the state of PUMA s. We assume that $(W_1, ..., W_N)$ is multivariate normally distributed with mean zero and variance/covariance matrix

$$\text{cov}(W_s, W_h) = \begin{cases} \sigma_S^2 + \sigma_P^2 + \sigma_{dist}^2 + \sigma_\epsilon^2/M_s & \text{if } s = h \\ \sigma_S^2 + \sigma_{dist}^2 \exp(-\alpha d(Z_s, Z_h)) & \text{if } s \neq h \text{ and } S_s = S_h \\ \sigma_{dist}^2 \exp(-\alpha d(Z_s, Z_h)) & \text{if } s \neq h \text{ and } S_s \neq S_h \end{cases}, \tag{26}$$

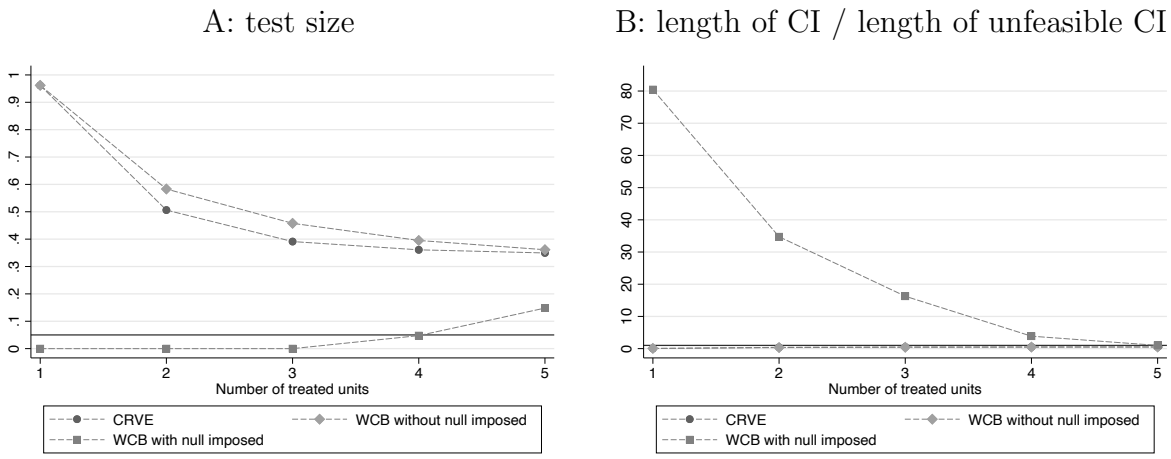
where $d(Z_s, Z_h)$ is the distance between the centroids of PUMAs s and h in kilometers. We impose the restrictions that the parameters $(\sigma_P^2, \sigma_\epsilon^2, \sigma_S^2, \sigma_{dist}^2, \alpha)$ are non-negative. These parameters are estimated by maximum likelihood. We find

$$(\hat{\sigma}_P^2, \hat{\sigma}_\epsilon^2, \hat{\sigma}_S^2, \hat{\sigma}_{dist}^2, \hat{\alpha}) = (0.000373, 0.918491, 0.000155, 0.001127, 0.088673).$$

For the MC simulations presented in Section 5 we first selected a PUMA that exhibits a strong spatial correlation with some other PUMAs to be treated. Then we selected the $N_1 - 1$ PUMAs that are most correlated with this one to be the remaining treated PUMAs. Given this treatment assignment, we draw random normal variables $(W_1, ..., W_N)$ from the estimated DGP.

We present in the main text the results for a naive FP, and for conservative tests 1 and 2. We present in Appendix Figure A.1 the results using alternative inference methods. For inference based on CRVE and WCB without the null imposed we have over-rejection for two reasons: because we have few treated clusters and because we have spatial correlation. This is why rejection rates do not become close to 5% even when N_1 increases. For WCB with the null imposed we have over-rejection due to spatial correlation and under-rejection due to a small number of treated clusters. When $N_1 = 5$ the distortions due to spatial correlation dominates, and the test over-rejects.

Figure A.1: **MC simulations - alternative inference methods**



Notes: This figure replicates Figure 1 for alternative inference methods. We consider inference based on CRVE, WCB without the null imposed, and WCB with the null imposed.

A.6 Testing for Spatial Correlation

It is common in empirical applications to have settings in which, for example, there are only a few states that are treated, but we observe subgroups (e.g., counties) within each state. Consider a setting in which there is only one treated state, and assume spatial correlation is restricted within states. In this case, the researcher could decide between using CRVE at the county level or a method such as CT and FP at the aggregate level. There is a trade off in terms of assumptions for these different type of methods. The first one relies on no spatial correlation within states, while the former relies on homoskedasticity (or heteroskedasticity with a structure as considered by FP). See Ferman (2021) for a more thorough comparison among different inference methods.

In this setting, we can consider a simple way to assess whether spatial correlation poses a problem for CRVE at the county level. Let $\widehat{W}_{cs}$ be the post-pre different in the residuals of county c in state s. Under the assumption that errors are independent across counties, and given that $\widehat{W}_{cs} \xrightarrow{p} W_{cs}$, we should expect that the proportion of residuals $\widehat{W}_{cs}$ that is

positive to be uniform across states (if W_{cs} is symmetric, then we should expect that to be approximately 1/2 for all states). In contrast, if there are state shocks, then we should expect positive residuals concentrated in some states, and negative residuals concentrated in others.

This suggests the following test. We calculate the proportion of positive residuals for each state ($\hat{p}_s$), and construct a test statistic $t = \sum_{i=1}^{S}(\hat{p}_s - \bar{p})^2$, where S is the number of states and $\bar{p} = S^{-1}\sum_{s=1}^{S}\hat{p}_s$. Then we can consider permutations of $\widehat{W}_{cs}$, and reconstruct $\hat{p}_s^b$. With that, we then calculate t^b for each permutation, and check whether t is extreme in the distribution of permutations. The distribution of the permutations would recover the discrepancies between $\hat{p}_s$ and $\bar{p}$ that we should expect when errors are independent, given that we have only a finite number of counties per state. If t is extreme in this distribution, then this would indicate relevant within-state spatial correlation.

This idea is similar in spirit to the test for appropriate level of clustering proposed by MacKinnon et al. (2020). Note, however, that CRVE at the state level would not be consistent in our setting, so their test would not be valid. Following the work by Roth (2019), we consider such test with caution, as we may fail to reject the null of no spatial correlation due to sampling variation. Still, such test may be very informative about when we *cannot* rely on CRVE and, therefore, should focus on alternatives such as CT and FP. If we consider this test for the empirical illustration in Section 6, the p-value is virtually equal to zero, providing strong evidence that spatial correlation is problematic in this setting.

When we consider this test for the empirical application considered in Section 6, we find a p-value of zero. This provides strong evidence that spatial correlation is important in this setting.