Three Essays on Applied Econometrics with Applications in Urban Economics

by

Kyle Butts

B.S., University of California Santa Barbara, 2017

M.S., University of Colorado Boulder, 2020

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Department of Economics

2024

Committee Members:

Taylor Jaworski, Chair

Adam McCloskey

Brian Cadena

Stephen Billings

Richard Mansfield

Butts, Kyle (Ph.D., Economics)

Three Essays on Applied Econometrics with Applications in Urban Economics

Thesis directed by Prof. Taylor Jaworski

This dissertation contains three chapters that develop econometric tools for practical problems faced in the field of urban economics.

The first chapter ‘Dynamic Treatment Effect Estimation with Interactive Fixed Effects and Short Panels’ studies inference on dynamic average treatment effect parameters for staggered interventions when parallel trends are only valid conditional on unobserved interactive fixed effects. Interactive fixed effects allow for the common setting where units are assigned to treatment based on their exposure to national macroeconomic shocks. Our identification strategy allows for any first stage set of moments that estimate the column space of the macroeconomic shocks including principal components, common correlated effects, quasi-differencing, and more. This allows us to estimate and control for the confounding common macroeconomics shocks. This result applies to data sets with either many or few pre-treatment time periods. We investigate the effect of Walmart openings on local economic conditions and demonstrate that our methods ameliorate pre-trend violations commonly found in the literature.

The second chapter ‘Difference-in-Differences with Spatial Spillovers’ addresses a common setting where treatment is assigned following geographic boundaries but the effects of the policies extend over space. Potential outcomes are modelled as a function of a unit’s treatment status as well as their exposure to treatment status of other units. The paper highlights biases in standard estimators due to spillover effects. Then identification and estimation of treatment effects are discussed including in settings with staggered treatment adoption. The proposed methods are applied to reanalyze three empirical applications highlighting the impact of spillovers on policy evaluations.

The final chapter ‘Difference-in-Differences with Geocoded Microdata’ formalizes a commonly-used estimator for the effects of spatially-targeted treatment with geocoded microdata. This estimator compares units immediately next to treatment to units slightly further away. I introduce intuitive identifying assumptions for the average treatment effect among affected units and illustrate problems when these assumptions fail. Since one of these assumptions requires knowledge of exactly how far treatment effects are experienced, I propose a new method that allows for nonparametric estimation following methods introduced in Cattaneo et al. (2019b). Since treatment effects can change with distance, the proposed estimator improves estimation by estimating a **treatment effect curve**.

Dedication

I dedicate this dissertation to my family: the queen Astrid, my boys Huckleberry and Archibald, and my best friend and teammate Hannah. You bring joy in my life every single day.

Acknowledgements

I extend my heartfelt gratitude to my family for their unwavering love and support throughout my life. A special thank you to my mom for being my first phone call for every challenge and to Star for bringing brightness into my life.

Thank you Rivka for helping me apply to programs and Phoenix for moving to Colorado and enjoying the mountains with me. I want to express my appreciation to all my classmates, including Brach, Jim, Alex, Payne, and Nick, for our first-year study sessions, research talk, and trips to the Downer. Thank you all for your valuable input and camaraderie. I am grateful to my coauthor Nicholas Brown for his collaboration and dedication.

Thank you to my advisor, Taylor Jaworski, for allowing me to be creative and follow my interests; encouraging me to think broader and with more rigor; and for guiding me through challenges and supporting my growth every step of the way. Thank you to my committee and the fantastic economists I've met in graduate school for lending eyes on my work and giving encouragement.

Contents

Chapter

Appendix

Tables

Table

Figures

Figure

Chapter 1

Dynamic Treatment Effect Estimation with Interactive Fixed Effects and Short Panels

Authors: Kyle Butts and Nicholas Brown

1.1 Introduction

There are two common approaches for estimating the dynamic effects of a binary treatment in linear panel data models. Difference-in-differences estimators are one of the most popular causal inference tools and computationally simple but rely on strong parallel-trends assumptions. In many empirical settings, treatment is assigned non-randomly based on trends in economic variables, rendering this method unusable. For example, in urban economics, place-based policies target places with worsening labor markets (Neumark and Simpson, 2015), new apartments are built in appreciating neighborhoods (Asquith et al., 2021a; Pennington, 2021), and firms opening new stores in growing economies (Basker, 2005; Neumark et al., 2008). Estimation of treatment effects in this setting is confounded by the pre-existing economic trends. It is common, though, that the causes of these trends are due to larger economic forces and not location-specific shocks. Continuing our examples, the national decline of manufacturing caused targeted manufacturing hubs to be declining, consumer trends for walkable neighborhoods cause certain neighborhoods to become increasingly demanded, and national changes to employment rates benefit certain counties.

A recent growing literature models these kind of parallel trends deviations using an interactive fixed effects, but requires long panels for consistent estimation. While interactive fixed effects relax the parallel trends assumptions relied on by difference-in-differences, estimators requiring long panels are often impractical because of (i) lack of data for many years of outcomes, (ii) strong assumptions like serially uncorrelated outcomes, or (iii) the presence of structural breaks, e.g. recessions or structural changes to the macroeconomy, rendering previous time periods uninformative about the current economy. This paper proposes a treatment effect estimator under the more general interactive fixed effect model that is robust to certain violations of parallel trends while remaining consistent in short panels and under heterogeneous treatment effects.

We model untreated potential outcomes, $y_{it}(\infty)$, as an interactive fixed effect model

$$y_{it}(\infty) = \boldsymbol{f}_t'\gamma_i + u_{it}, \tag{1.1}$$

where $\boldsymbol{F}_t$ is a $p \times 1$ vector of unobservable factors, $\boldsymbol{\gamma}_i$ is a $p \times 1$ vector of unobservable factor loadings, and $\mathbb{E}[u_{it}] = 0$ for all (i, t).[1] We can view, as we did in the above examples, the factors $\boldsymbol{F}_t$ as macroeconomic shocks with factor loadings $\boldsymbol{\gamma}_i$ denoting a unit's exposure to the shocks. Another possibility lets the $\boldsymbol{\gamma}_i$ represent time-invariant characteristics with a marginal effect on the outcome $\boldsymbol{F}_t$ that changes over time.[2] Note that this model nests the standard two-way error model when $\boldsymbol{F}_t' = (\lambda_t, 1)$ and $\boldsymbol{\gamma}_i' = (1, \mu_i)$; that is, $\boldsymbol{F}_t'\gamma_i = \lambda_t + \mu_i$. The interactive structure allows for more general patterns of unobserved heterogeneity. Importantly, we allow for treatment to be correlated with a unit's exposure to macroeconomic shocks via their factor loadings $\boldsymbol{\gamma}_i$.

For a concrete example, our empirical application focuses on estimating the effect of Walmart store openings on county-level employment. Estimation of a standard two-way fixed effect event-study model suggests that Walmart opened stores in counties that had higher retail employment growth prior to the opening (e.g. Neumark et al. (2008)). In Figure 1.2, we present an event-study graph and overlay a line of best fit on the pre-treatment estimates. That the line is positive sloping and the estimates are different from zero at the 5% level suggests that estimated positive impacts are due to pre-existing trends rather than the effect of Walmart per se. However, there seems to be a discrete jump when the Walmart opened. The goal then is to remove these pre-existing trends to

[1] We follow Callaway and Sant'Anna (2021) and define the state of not receiving treatment in the sample as '∞'. This is useful in settings with staggered treatment timing where potential outcomes are denoted by the period where a unit start treatments.

[2] Ahn et al. (2013) suggest a wage equation where $\boldsymbol{\gamma}_i$ are unobserved worker characteristics of an individual and $\boldsymbol{F}_t$ are their time-varying prices or returns to those characteristics. See Bai (2009) for a collection of economic examples that justify the inclusion of a factor structure.

isolate the treatment effect. It is plausible to assume that during their period of mass expansion, Walmart selected appealing locations based on their local demographic background and national economic trends, while ignoring transitory local economic shocks. Our framework allows this type of selection mechanism and effectively 'controls' for these pre-existing trends in outcome.

Our main treatment effect identification result only requires fixed-T consistent estimates of the column space of $\boldsymbol{F}_t$. Using the estimated factors, we compute a matrix that projects the pre-treatment outcomes onto the estimated post-treatment factors, imputing the untreated potential outcome for treated units. Averaging over the difference between the post-treatment observed outcomes and the estimated untreated potential outcomes gives a consistent estimator of average treatment effects. In specifications that include the two-way error model, we show how to explicitly remove the additive fixed effects with a double-demeaning transformation that maintains the common factor structure across treated groups and the never-treated group.

There are two major benefits of our general identification argument. First, fixed-T consistent estimation of $\boldsymbol{F}_t$ is possible through a variety of approaches, which primarily includes quasi-differencing (Ahn et al., 2001, 2013; Callaway and Karami, 2023) and cross-sectional averages (Westerlund et al., 2019; Juodis and Sarafidis, 2022b,a; Brown et al., 2023b) as a non-exhaustive list. These techniques allow the user to tailor their factor estimator to the specific data and problem under consideration, including how many pre-treatment time periods are available. Our identification result provides a recipe for using any consistent estimator of the factors to estimate treatment effects, opening up the large factor-model literature for causal inference methods. Second, our imputation method allows researchers to graph the estimated counterfactual untreated potential outcomes and the observed outcomes for treated units as a visual check for the parallel trends assumption, similar to a synthetic control plot.

We derive asymptotic properties of an imputation estimator with factor proxies that contain the true unobserved factors in their column space. The resulting estimator takes the form of a generalized method of moments (GMM) estimator, which allows estimation and inference via common statistical software. It is also consistent when the number of pre-treatment time periods is small.[3] One advantage of this estimator is that we can form statistical tests for the consistency of the two-way fixed effects (TWFE) estimator. These tests are practically useful since difference-in-differences is simple to implement.

1.1.1 *

Relation to Literature

Recent work has proposed ‘imputation’ estimators for treatment effects using non-treated and pre-treatment observations to ‘impute’ the untreated potential outcomes for the post-treatment observations (e.g. Borusyak et al., 2024; Gardner, 2022; Wooldridge, 2021). However, these approaches only allow for level fixed effects and preclude interactions like in equation (1.1). Borusyak et al. (2024) allow a structure similar to equation (1.1) but requires the the factors $\boldsymbol{F}_t$ be observed. We generalize these techniques by proposing an estimator that imputes the untreated potential outcomes under the more general (1.1) with unobserved interactive effects.

Current estimators that allow for selection based on a factor model either require (i) the number of time periods available is large, e.g. synthetic control (Abadie, 2021), factor-model imputation (Xu, 2017; Gobillon and Magnac, 2016), and the matrix completion method (Athey et al., 2021; Fernández-Val et al., 2021); or (ii) that an individual's error term u_{it} is uncorrelated over time (Imbens et al., 2021).[4] Both of these restrictions are non-realistic in many applied microeconomic data sets where the number of time periods is much smaller than the number of units and serial correlation of shocks is expected. Further, large-T estimators often place restrictions on the dynamic heterogeneity of treatment. Our method requires neither large T nor error term restrictions, but can still accommodate large-T and unit-heterogeneous estimation strategies.

Our work contributes to an emerging literature on adjusting for parallel trends violations in short panels. Freyaldenhoven et al. (2019) propose a similar instrumental variable type estimator in

[3] Deriving the asymptotic distribution of treatment effects using large-T factor estimators is left for future work.

[4] Imbens et al. (2021) allow correlation within the post- and pre-treatment sets of the idiosyncratic errors, but assume independence between the two sets. This assumption is still strong in a static modeling context.

the presence of time-varying confounders. Their results rely importantly on homogeneous treatment effects. Their simulations show that heterogeneous treatment effects bias their estimates severely, while our estimator allows for arbitrary time heterogeneity. The most similar paper to our current approach is Callaway and Karami (2023), who also allow for heterogeneous effects in short panels. They prove identification using a similar strategy to QLD and instrumental variables and derive asymptotic normality assuming the number of time periods is fixed. They require time-invariant instruments whose effects on the outcome are constant over time. Their instruments are valid for the QLD estimator in our application, but we also allow for time-varying covariates as instruments. They do not provide a general identification scheme like ours and so their results do not readily extend to other estimators like principal components or common correlated effects.

The rest of the paper is divided into the following sections: Section 1.2 describes the theory behind our methods and presents identification results of the group-specific dynamic treatment effect parameters. Section 1.3 provides the main asymptotic theory for a particular QLD estimator. We also discuss practical concerns for practitioners. We include a small Monte Carlo experiment in Section 1.4 to examine the finite-sample performance of our estimator. Finally, Section 1.5 contains our application and Section 2.5 leaves with some concluding remarks.

1.2 Model and Identification

We assume a panel data set with units $i = 1, \ldots, N$ and periods $t = 1, \ldots, T$. Treatment turns on in different periods for different units; we denote these groups by the period they start treatment. For each unit, we define G_i to be unit i's group with possible values $\{g_1, \ldots, g_G\} \equiv \mathcal{G} \subseteq \{2, \ldots, T\}$. We follow Callaway and Sant'Anna (2021) and denote $G_i = \infty$ for units that never receive treatment. We assume that $0 < P(G_i = g) < 1$ for all $g \in \mathcal{G} \cup \{\infty\}$, so that the number of individuals in each group and the never-treated group grow with N. Treated potential outcomes are a function of group-timing, which we denote $y_{it}(g)$. For treatment indicators, we define the vector of treatment statuses $\boldsymbol{d}_i = (d_{i1}, ..., d_{iT})$ where $d_{it} = \mathbf{1}(t \geq G_i)$ and the indicator $D_{ig} = \mathbf{1}(G_i = g)$ if unit i is a member of group g. Let $T_0 = \min_j\{g_j\} - 1$ be the last period before the earliest treatment adoption.

Following Callaway and Sant'Anna (2021), we aim to estimate group-time average treatment effects on the treated:

$$\text{ATT}(g,t) = \tau_{gt} \equiv \mathbb{E}[y_{it}(g) \mid G_i = g] - \mathbb{E}[y_{it}(\infty) \mid G_i = g] \tag{1.2}$$

These quantities represent the average effect of treatment at time t for units that start treatment in period g for $t \geq g$. It is trivial to estimate other averages as well in our framework, including averaging over all post-treatment observations to estimate an overall ATT, and averaging over (i,t) where $t - G_i = \ell$ to estimate event-study estimands ATT$^{\ell}$'s. We discuss these and other extensions from Callaway and Sant'Anna (2021) in Section 1.3.

We now state our main identifying assumptions.

Assumption 1.1 (**Sampling**). *The random vectors* $\{(\boldsymbol{d}_i, \boldsymbol{\gamma}_i, \boldsymbol{u}_i)\}$ *are randomly sampled from an infinite population and has finite moments up to the fourth order.* ■

Assumption 1.2 (**Untreated potential outcomes**). *The untreated potential outcomes take the form*

$$y_{it}(\infty) = \boldsymbol{F}_t' \boldsymbol{\gamma}_i + u_{it}$$

where $\mathbb{E}[u_{it} \mid \boldsymbol{d}_i, \boldsymbol{\gamma}_i] = 0$ *for* $t = 1, ..., T$. ■

Assumption 1.3 (**No anticipation**). *For all units* i *and groups* $g \in \mathcal{G}$, $y_{it} = y_{it}(\infty)$ *for* $t < g$. ■

Assumption 1.2 imposes a factor-model for the untreated potential outcomes. The Online Appendix discusses the inclusion of covariates and the subsequent relaxation of assumption 1.2. We allow for heterogeneous and dynamic treatment effects of any form, i.e. $y_{it}(g) = \tau_{igt} + y_{it}(\infty)$. We also allow arbitrary serial correlation among the idiosyncratic errors.[5] We assume the common factors $\boldsymbol{F}_t$ are nonrandom parameters and the number of factors p is fixed in the asymptotic analysis.

Assumption 1.2 is more general than the standard difference-in-differences parallel trend assumption since we include the factor structure in our potential outcome model. In particular, it assumes that the error term is uncorrelated with treatment status *after* controlling for the factor

[5] This condition may need to be strengthened for inference when $T \to \infty$.

loadings. Treatment can still be correlated with contemporaneous shocks so long as the shocks, but not necessarily the exposure to them, are 'common' across the sample. For example, our identification strategy is valid if workers select into a job training program based on their exposure (or adaptability) to macroeconomic productivity shocks.

The two-way error model cannot generally accommodate differential exposure.[6] In the more general factor model and Assumption 1.2, changes in untreated potential outcomes are given by

$$\mathbb{E}[y_{it}(\infty) - y_{it-1}(\infty) \mid G_i = g] = \lambda_t + (\boldsymbol{F}_t - \boldsymbol{F}_{t-1})' \, \mathbb{E}[\boldsymbol{\gamma}_i \mid G_i = g]$$

Unless either (i) the factor loadings have the same mean across treatment groups, $\mathbb{E}[\boldsymbol{\gamma}_i \mid G_i = g] = \mathbb{E}[\boldsymbol{\gamma}_i]$, or (ii) the factors are time-invariant, then the standard parallel trends assumption that the group g and the never-treated group follow common trends would not hold. If either of the two cases hold for all g and t, the two-way error model is correctly specified.[7] However, these are knives edge cases which are not the focus of the paper. Our Assumption 1.2 allows for the factor loadings to be correlated with treatment timing and opens up treatment effect estimation for a much broader set of empirical questions.

The key econometric challenge lies in that we do not observe $y_{it}(\infty)$ whenever $d_{it} = 1$. Our goal is to consistently estimate $\mathbb{E}[y_{it}(\infty) \mid G_i = g]$ under equation (1.1) to consistently estimate group-time average treatment effects. Gardner (2022), Wooldridge (2021), and Borusyak et al. (2024) implicitly rely on this insight in studying the two-way error model.

Prior attempts at estimating average treatment effects in a factor-model setting focus on finding conditions that allow for estimation of $\boldsymbol{\gamma}_i$ and $\boldsymbol{F}_t$ jointly as in Gobillon and Magnac (2016) and Xu (2017), or a generalized version of a factor model as in Arkhangelsky et al. (2021). These techniques require the number of pre-treatment periods to grow to infinity and often place restrictions on both the dynamics of the treatment effects' distribution and the serial dependence among the idiosyncratic errors. Instead, we pursue identification noting that

$$\mathbb{E}[y_{it}(\infty) \mid G_i = g] = \boldsymbol{F}_t' \, \mathbb{E}[\boldsymbol{\gamma}_i \mid G_i = g] \tag{1.3}$$

[6] The following derivation is also shown in Callaway and Karami (2023), but we are repeating it here for exposition.

[7] We explicitly prove this result later.

Therefore, we only need to estimate the *average* of the factor loadings among a treatment group, which we can always do even with a small number of post-treatment time periods. We can then accommodate either a large or small number of pre-treatment periods and allow for estimation using a broad range of known strategies.

1.2.1 ATT(g, t) Identification

We begin by describing the intuition behind our identification result. Consider a unit subject to treatment at time g. Define $\boldsymbol{y}_{i,t<g}$ and $\boldsymbol{y}_{i,t\geq g}$ as respectively the first $(g-1)$ and last $(T-g+1)$ outcomes for unit i. Define $\boldsymbol{F}$ to be the matrix of factor shocks with rows given by $\boldsymbol{F}_t$. We similarly define $\boldsymbol{F}_{t<g}$ and $\boldsymbol{F}_{t\geq g}$ as the first and last rows of matrix $\boldsymbol{F}$. Equation (1.3) implies

$$\mathbb{E}[\boldsymbol{y}_{i,t<g}(\infty) \mid \boldsymbol{G}_i = g] = \boldsymbol{F}_{t<g}\, \mathbb{E}[\gamma_i \mid \boldsymbol{G}_i = g] \tag{1.4}$$

If the factors were observed, we could consistently estimate the mean values of the p-vector of average factor loadings for treated group $G_i = g$. More formally, if $\text{Rank}(\boldsymbol{F}_{t<g}) = p$, the coefficient from the population regression of $\mathbb{E}[y_{i,t<g}(\infty) \mid G_i = g]$ on $\boldsymbol{F}_{t<g}$ is $\mathbb{E}[\gamma_i \mid G_i = g]$. Equation (1.3) also gives us

$$\mathbb{E}[\boldsymbol{y}_{i,t\geq g}(\infty) \mid \boldsymbol{G}_i = g] = \boldsymbol{F}_{t\geq g}\, \mathbb{E}[\gamma_i \mid \boldsymbol{G}_i = g] \tag{1.5}$$

for the post-treated outcomes. Assuming $\boldsymbol{F}$ is known (for now), we can predict $\mathbb{E}[\boldsymbol{y}_{i,t} \mid G_i = g]$ for $t \geq g$ by multiplying $\boldsymbol{F}_t$ by the OLS estimate from the prior regression. We then obtain $\mathbb{E}[y_{it}(\infty) \mid G_i = g]$ for the post-treatment outcomes, which we can subtract from y_{it} and average over the respective sample to obtain ATT(g, t).

We now define a useful matrix function for a more formal derivation of our main result. Given matrices $\boldsymbol{X}_1$ and $\boldsymbol{X}_0$ that are respectively $n \times k$ and $m \times k$, suppose $\text{Rank}(\boldsymbol{X}_0) = k$. We define the *imputation matrix*

$$\boldsymbol{P}(\boldsymbol{X}_1, \boldsymbol{X}_0) \equiv \boldsymbol{X}_1(\boldsymbol{X}_0'\boldsymbol{X}_0)^{-1}\boldsymbol{X}_0' \tag{1.6}$$

This matrix takes a similar form to a projection matrix but "imputes" the fitted values from regressing on $\boldsymbol{X}_0$ onto a different matrix $\boldsymbol{X}_1$. Gardner (2022) and Borusyak et al. (2024) implicitly uses the imputation matrix for an additive error model where $\boldsymbol{X}_1$ is the matrix of unit and time fixed effects and $\boldsymbol{X}_0$ is $\boldsymbol{X}_1$ with rows of zero whenever $d_{it} = 1$. When applying this matrix of factors to our outcomes, the post-treatment factors are multiplied by the factor loadings from the pre-treatment observations. In particular, we impute $y_{it}(\infty)$ by $\boldsymbol{P}(\boldsymbol{F}_t', \boldsymbol{F}_{t<g})\boldsymbol{y}_{i,t<g}$ for $G_i = g$, similar to the bridge function identification scheme in Imbens et al. (2021). However, because we only need a conditional mean assumption, we can allow arbitrary correlation between the idiosyncratic errors.

The next theorem provides our main identification result:

Theorem 1.1. *Suppose $\boldsymbol{F}$ is known and Rank$(\boldsymbol{F}_{t\leq T_0}) = p$. Under Assumptions 1.1, 1.2, and 1.3 for all $g \in \mathcal{G}$,*

$$ATT(g,t) = \mathbb{E}\big[y_{it} - \boldsymbol{P}(\boldsymbol{F}_t', \boldsymbol{F}_{t<g})\boldsymbol{y}_{i,t<g} \mid G_i = g\big] \tag{1.7}$$

for $t \geq g$.

Moreover, let $\boldsymbol{F}^*$ be a full rank $T \times m$ matrix where $m < T_0$ and $\boldsymbol{F} \in$ col$(\boldsymbol{F}^*)$, the column space of $\boldsymbol{F}^*$. Then the imputation matrix is invariant to $\boldsymbol{F}^*$

$$\boldsymbol{P}(\boldsymbol{F}_t^{*'}, \boldsymbol{F}_{t<g}^*)\boldsymbol{F}_{t<g}\boldsymbol{\gamma}_i = \boldsymbol{F}_t'\boldsymbol{\gamma}_i \tag{1.8}$$

■

All proofs are contained in the Online Appendix. Theorem 1.1 shows that we can identify the ATTs if we know the factor matrix. The second part of the theorem suggests that any rotation of the true factor matrix, $\boldsymbol{F}$, can be used in the imputation matrix. This is important because it is well understood that $\boldsymbol{F}_t$ and $\boldsymbol{\gamma}_i$ are not separately identified (Ahn et al., 2013; Xu, 2017). All of the estimators discussed so far can at best approximate the column space of the factors because both $\boldsymbol{F}_t$ and $\boldsymbol{\gamma}_i$ are unobserved. The second part of the theorem shows that our identification scheme allows for this class of estimators.

Theorem 1.1 shows we can apply these conclusions to any estimator that achieves fixed-T consistency by asymptotically spanning the factor space. The most popular classes of these estimators examples include the common correlated effects (Pesaran, 2006) and quasi-differencing

(Ahn et al., 2013). Westerlund (2020) shows that principal components can also fit in this class, so long as strong restrictions are met. As long as the column space of the factors are consistently estimated using the control sample, dynamic ATTs are identified as in Theorem 1.1, regardless of the normalization used for estimation.

To present a general framework for the estimation of the factors, we formally present the identifying assumptions needed for factor space estimators:

Assumption 1.4. *There exists a $q \times 1$ vector of parameters $\boldsymbol{\theta}$ and a $T \times m$ function $\boldsymbol{F}(\boldsymbol{\theta})$ such that the following conditions hold:*

(i) For some full-rank matrix $\boldsymbol{A}$, $\boldsymbol{F}(\boldsymbol{\theta})\boldsymbol{A} = \boldsymbol{F}$ where Rank$(\boldsymbol{F}(\boldsymbol{\theta})) = m < T_0$

(ii) There is a $s \times 1$ vector of moment functions $\boldsymbol{g}_{i\infty}(\boldsymbol{\theta})$ such that

$$\mathbb{E}[\boldsymbol{g}_{i\infty}(\boldsymbol{\theta}) \mid G_i = \infty] = \mathbf{0} \tag{1.9}$$

(iii) Let $\boldsymbol{D}_\infty = \mathbb{E}[\nabla_{\boldsymbol{\theta}} \boldsymbol{g}_{i\infty}(\boldsymbol{\theta}) \mid G_i = \infty]$. Then Rank$(\boldsymbol{D}_\infty) = q$.

(iv) $\mathbb{E}[\boldsymbol{g}_{i\infty}(\boldsymbol{\theta})\boldsymbol{g}_{i\infty}(\boldsymbol{\theta})' \mid G_i = \infty]$ is positive definite.

Part (i) implies that the estimated factors can be reduced to a finite dimension of estimable parameters. The matrix $\boldsymbol{A}$ is the full rank linear rotation that turns $\boldsymbol{F}(\boldsymbol{\theta})$ into $\boldsymbol{F}$. For the example estimators expressed above, $\boldsymbol{F}(\boldsymbol{\theta})$ asymptotically spans the unknown factors, $\boldsymbol{F}$. Parts (ii)-(iv) imply the parameters $\boldsymbol{\theta}$ are identified and consistently estimable. The parameters $\boldsymbol{\theta}$ themselves are often the result of an underlying normalization like in Ahn et al. (2001, 2013), Juodis and Sarafidis (2022b,a), and Callaway and Karami (2023). Sometimes they are population moments estimated by cross-sectional averages like in Westerlund et al. (2019) and Brown et al. (2023b).

Assumption 1.4 is written with fixed-T estimation and inference in mind. As mentioned before, accommodating general principal components estimation requires additional restrictions, as well as a large time series in the pre-treatment periods. However, the general identification result is the same and our estimator is still valid for estimating dynamic effects in the post-treatment period.

Chan and Kwok (2022) study a principal components estimator for unit-specific treatment effects. Further research should formally derive the large-T properties of our estimator using principal components in the first stage.

Remark 1.1 (Quasi-Long-Differencing)**.** *A leading example of a set of moment equations for factor-space estimation is the quasi-long differencing (QLD) estimator of Ahn et al. (2013). They propose a QLD transformation given by*

$$\boldsymbol{H}(\boldsymbol{\theta}) = (\boldsymbol{I}_{T-p}, \boldsymbol{\Theta}) \tag{1.10}$$

where $\boldsymbol{\Theta}$ is a $(T-p) \times p$ matrix of unrestricted parameters and $\boldsymbol{\theta} = vec(\boldsymbol{\theta})^8$. They normalize the factors as

$$\boldsymbol{F}(\boldsymbol{\theta}) = \begin{pmatrix} \boldsymbol{\Theta} \\ \\ -\boldsymbol{I}_p \end{pmatrix} \tag{1.11}$$

so that $\boldsymbol{H}(\boldsymbol{\theta})\boldsymbol{F}\boldsymbol{\gamma}_i = \boldsymbol{0}$ by construction. We modify their proposed moment conditions to use just the never-treated group:

$$\mathbb{E}[g_{i\infty}(\boldsymbol{\theta}) \mid D_{i\infty} = 1] = \mathbb{E}[\boldsymbol{H}(\boldsymbol{\theta})\boldsymbol{y}_i \otimes \boldsymbol{w}_i \mid D_{i\infty} = 1] = \boldsymbol{0} \tag{1.12}$$

where $\boldsymbol{w}_i$ is a vector of instruments that are exogenous with respect to the idiosyncratic error in Assumption 1.2 but correlated with $\boldsymbol{\gamma}_i$. We discuss the choice of instruments $\boldsymbol{w}_i$ in more practical terms in section 1.5.

While both approaches are valid in the first stage of our setting, we use the Ahn et al. (2013) estimator because it is more general than Callaway and Karami (2023). For one, they allow for a larger set of instruments. One identification strategy proposed by Callaway and Karami (2023) requires time-invariant covariates whose effects on y_{it} are independent of time, meaning the researcher must decide which of the time-invariant observables have constant effects on the outcome. Ahn et al. (2013) can allow for arbitrary time effects on covariates while still using those covariates as instruments. Ahn et al. (2013) also give a road map to estimation based on weakly exogenous

[8] We reuse the "$\boldsymbol{\theta}$" notation throughout the remainder of the text.

covariates that allows for dynamic modeling. This aspect of the estimator is left for future research. ■

1.2.2 Two-Way Error Model

We now demonstrate how to explicitly nest the standard two-way error model. While this structure is a special case of the factor model studied above, we consider the special case for two main reasons. First, eliminating the additive effects saves degrees of freedom to estimate the factor models; it may also provide efficiency by reducing the burden on first-stage factor estimators. Second, a thorough study of the additive model will provide insight into the link between TWFE estimation and more complicated and computationally involved factor model estimation. It will also allow us to show when TWFE estimation is consistent in the presence of interactive fixed effects.

We first note that care must be taken when eliminating additive effects so that the overall factor structure is preserved. The methods in Borusyak et al. (2024), Gardner (2022), and Wooldridge (2021) that estimate the additive effects using the untreated sample will not maintain a common factor structure. For example, consider the first order conditions from the regression of $(1 - d_{it})y_{it}$ on unit and time effects. The estimators for the unit effect of a unit treated at time g and a never-treated unit respectively satisfy

$$\sum_{t=1}^{g-1}(y_{it} - \widehat{\lambda}_t - \widehat{\mu}_i) = 0 \tag{1.13}$$

$$\sum_{t=1}^{T}(y_{it} - \widehat{\lambda}_t - \widehat{\mu}_i) = 0 \tag{1.14}$$

The control sample will remove more time averages than in every treated sample, meaning the factors are demeaned using different subsamples. As such, the transformed factors are not equal across groups and so we cannot then use the control sample to estimate the factors for the treated samples.

We first define the following averages for the purpose of removing the additive effects:

$$\overline{y}_{\infty,t} = \frac{1}{N_\infty}\sum_{i=1}^{N} D_{i\infty} y_{it} \tag{1.15}$$

$$\overline{y}_{i,t\leq T_0} = \frac{1}{T_0}\sum_{t=1}^{T_0} y_{it} \tag{1.16}$$

$$\overline{y}_{\infty,t<T_0} = \frac{1}{N_\infty T_0}\sum_{i=1}^{N}\sum_{t=1}^{T_0} D_{i\infty} y_{it} \tag{1.17}$$

where $\overline{y}_{\infty,t}$ is the cross-sectional averages of the never-treated units for period t, $\overline{y}_{i,t\leq T_0}$ is the time-averages of unit i before any group is treated, and $\overline{y}_{\infty,t<T_0}$ is the total average of the never-treated units before any group is treated.

We then perform all estimation on the residuals $\tilde{y}_{it} \equiv y_{it} - \overline{y}_{\infty,t} - \overline{y}_{i,t<T_0} + \overline{y}_{\infty,t<T_0}$. These residuals are reminiscent of the usual TWFE residuals, except we carefully select this transformation to accomplish two things. First, this transformation leaves the treatment dummy variables unaffected to prevent problems with negative weighting when aggregating heterogeneous treatment effects (Goodman-Bacon, 2021; Borusyak et al., 2024). Second, it preserves a common factor structure for all units and time periods[9] . The TWFE imputation estimator of Gardner (2022), Wooldridge (2021), and Borusyak et al. (2024) would not share this property because they estimate μ_i and λ_t based on the full sample $d_{it} = 0$, while we use a specific subsample.

This result is summarized in the following lemma:

Lemma 1.1. $\mathbb{E}[\tilde{y}_{it} \mid G_i = g] = \mathbb{E}\big[d_{it}\tau_{it} + (\boldsymbol{F}_t - \overline{\boldsymbol{F}}_{t<T_0})'(\gamma_i - \overline{\gamma}_\infty) \mid G_i = g\big]$ *for* $t = 1, ..., T$ *and* $g \in \mathcal{G} \cup \{\infty\}$ *where* $\overline{\boldsymbol{F}}_{t<T_0}$ *is the average of* $\boldsymbol{F}_t$ *in the pre-treatment periods and* $\overline{\gamma}_\infty$ *is the average of* $\boldsymbol{\gamma}_i$ *among the control units.* ■

Lemma 1.1 demonstrates how to explicitly nest the two-way error model model while allowing for a general common factor structure. Since we are not interested in inference on the factors themselves, this form will suffice for the imputation process. The transformed outcomes take the

[9] Such a transformation should not be used when considering the common correlated effects estimator because it would violate the CCE rank condition. See Brown et al. (2023a).

form

$$\tilde{y}_{it} = d_{it}\tau_{it} + (\boldsymbol{F}_t - \overline{\boldsymbol{F}}_{t<T_0})'(\boldsymbol{\gamma}_i - \overline{\boldsymbol{\gamma}}_\infty) + \tilde{u}_{it}. \tag{1.18}$$

For ease of exposition, we rewrite the above equation as:

$$\tilde{y}_{it} = d_{it}\tau_{it} + \tilde{\boldsymbol{F}}_t'\tilde{\boldsymbol{\gamma}}_i + \tilde{u}_{it}. \tag{1.19}$$

Lemma 1.1 has the added benefit of showing us when the ATTs are identified by our TWFE transformation alone.

Corollary 1.1. *Under Assumptions 1.1-1.3, $ATT(g, t)$ is identified by the fixed effects imputation transformation if $\mathbb{E}[\gamma_i \mid G_i = g] = \mathbb{E}[\gamma_i]$ for all $g \in \mathcal{G} \cup \{\infty\}$.* ■

This result is an immediate consequence of Assumptions 1.1 – 1.3 as $\mathbb{E}[\boldsymbol{\gamma}_j \mid G_i = g] = \mathbb{E}[\boldsymbol{\gamma}_i]$ for $j \neq i$ under random sampling. Corollary 1.1 tells us that TWFE imputation is sufficient to estimate the ATTs, even when the factor structure exists, so long as the average factor loadings do not differ systemically with treatment status. Asymptotic normality of our imputation procedure under a two-way error model is studied in the Online Appendix. We also provide simple tests for mean independence of the factor loadings in Remark 1.5, i.e. consistency of the TWFE estimator. However, if the researcher believes a TWFE estimator is sufficient, they should use one of the other techniques mentioned above. Our method sacrifices potential efficiency by not using all observations to eliminate the additive effects in order to allow for additional interactive effects.

1.3 Estimation and Inference

This section considers estimation of the group-time average treatment effects. A major benefit of our approach is the simplicity of inference while allowing for a large number of possible estimation techniques in the first stage. Our moment conditions lead to a simple GMM estimator for which inference is standard and can be computed via routine packages in standard statistical software. Further, we can use the moment conditions to test the fundamental features of the model.

1.3.1 Asymptotic Normality

Equations (1.7) and (1.9) provide us with the necessary moment conditions to estimate the ATTs. We collect them here in their unconditional form:

$$\mathbb{E}\left[\frac{D_{i\infty}}{\mathbb{P}(D_{i\infty}=1)}\boldsymbol{g}_{i\infty}(\boldsymbol{\theta})\right]=\mathbf{0}$$

$$\mathbb{E}[\boldsymbol{g}_{ig_G}(\boldsymbol{\theta},\boldsymbol{\tau}_{g_G})]=\mathbb{E}\left[\frac{D_{ig_G}}{\mathbb{P}(D_{ig_G}=1)}\left(\boldsymbol{y}_{i,t\geq g_G}-\boldsymbol{P}(\boldsymbol{F}_{t\geq g_G}(\boldsymbol{\theta}),\boldsymbol{F}_{t<g_G}(\boldsymbol{\theta}))\boldsymbol{y}_{i,t<g_G}-\boldsymbol{\tau}_{g_G}\right)\right]=\mathbf{0}$$

$$\vdots$$

$$\mathbb{E}[\boldsymbol{g}_{i1}(\boldsymbol{\theta},\boldsymbol{\tau}_{g_1})]=\mathbb{E}\left[\frac{D_{ig_1}}{\mathbb{P}(D_{ig_1}=1)}\left(\boldsymbol{y}_{i,t\geq g_1}-\boldsymbol{P}(\boldsymbol{F}_{t\geq g_1}(\boldsymbol{\theta}),\boldsymbol{F}_{t<g_1}(\boldsymbol{\theta}))\boldsymbol{y}_{i,t<g_1}-\boldsymbol{\tau}_{g_1}\right)\right]=\mathbf{0}$$

where $\boldsymbol{\tau}_g=(\tau_{gg},...,\tau_{gT})'$ is the vector of post-treatment treatment effects. We stack these over g as $\boldsymbol{\tau}=(\boldsymbol{\tau}_{g_1}',...,\boldsymbol{\tau}_{g_G}')'$. The first set of moment conditions identify the factor space by Assumption 1.4 and the remaining moments identify the τ_{gt} via our imputation method.[10] Implementation requires replacing $\mathbb{P}(D_{ig}=1)$ with its sample counterpart N_g/N. This setting can also accommodate cases as in Hahn et al. (2018) where the factor structure is estimated nonparametrically in the first stage but the parametric estimator in the second stage is still $O_p(N^{-1/2})$. We leave this case for future study.

We need one final assumption to implement the asymptotically efficient GMM estimator:

Assumption 1.5. $\mathbb{E}[\boldsymbol{g}_{ig}(\boldsymbol{\theta},\boldsymbol{\tau}_g)\boldsymbol{g}_{ig}(\boldsymbol{\theta},\boldsymbol{\tau}_g)']$ *is positive definite for each* $g\in\mathcal{G}$. ■

We collect the moment functions into the vector

$$\boldsymbol{g}_i(\boldsymbol{\theta},\boldsymbol{\tau})=(\boldsymbol{g}_{i\infty}(\boldsymbol{\theta})',\boldsymbol{g}_{ig_G}(\boldsymbol{\theta},\boldsymbol{\tau}_{g_G})',...,\boldsymbol{g}_{ig_1}(\boldsymbol{\theta},\boldsymbol{\tau}_{g_1})')'.$$

In an abuse of notation, we assume $\boldsymbol{g}_{i\infty}$ is the moment function from equation (1.9) but scaled by $D_{i\infty}/P(D_{i\infty}=1)$. We define $\boldsymbol{\Delta}=\mathbb{E}[\boldsymbol{g}_i(\boldsymbol{\theta},\boldsymbol{\tau})\boldsymbol{g}_i(\boldsymbol{\theta},\boldsymbol{\tau})']$ which is positive definite by Assumptions 1.4 and 1.5. Then our GMM estimator $(\widehat{\boldsymbol{\theta}}',\widehat{\boldsymbol{\tau}}')'$ solves

$$\min_{\boldsymbol{\theta},\boldsymbol{\tau}}\left(\sum_{i=1}^N\boldsymbol{g}_i(\boldsymbol{\theta},\boldsymbol{\tau})\right)'\widehat{\boldsymbol{\Delta}}^{-1}\left(\sum_{i=1}^N\boldsymbol{g}_i(\boldsymbol{\theta},\boldsymbol{\tau})\right)\tag{1.20}$$

where $\widehat{\boldsymbol{\Delta}}\,\mathrm{plim}\,\boldsymbol{\Delta}$ uses an initial consistent estimator of $(\boldsymbol{\theta}',\boldsymbol{\tau}')'$.

[10] We implicitly assume $\mathbb{P}(D_{ig_h}=1)$ is strictly between 0 and 1 for every $g_h\in\mathcal{G}\cup\{\infty\}$.

Theorem 1.2. *Under Assumptions 1.1-1.5, $\sqrt{N}\big((\widehat{\boldsymbol{\theta}}', \widehat{\boldsymbol{\tau}}')' - (\boldsymbol{\theta}', \boldsymbol{\tau}')'\big)$ is jointly asymptotically normal as $N \to \infty$ and*

$$\sqrt{N}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} N\left(\mathbf{0}, \left(\boldsymbol{D}_\infty' \boldsymbol{\Delta}_\infty^{-1} \boldsymbol{D}_\infty\right)^{-1}\right)$$

$$\sqrt{N}(\widehat{\boldsymbol{\tau}}_{g_G} - \boldsymbol{\tau}_{g_G}) \xrightarrow{d} N\left(\mathbf{0}, \boldsymbol{\Delta}_{g_G} + \boldsymbol{D}_{g_G}\left(\boldsymbol{D}_\infty' \boldsymbol{\Delta}_\infty^{-1} \boldsymbol{D}_\infty\right)^{-1} \boldsymbol{D}_{g_G}'\right)$$

$$\vdots$$

$$\sqrt{N}(\widehat{\boldsymbol{\tau}}_{g_1} - \boldsymbol{\tau}_{g_1}) \xrightarrow{d} N\left(\mathbf{0}, \boldsymbol{\Delta}_{g_1} + \boldsymbol{D}_{g_1}\left(\boldsymbol{D}_\infty' \boldsymbol{\Delta}_\infty^{-1} \boldsymbol{D}_\infty\right)^{-1} \boldsymbol{D}_{g_1}'\right)$$

where $\boldsymbol{D}_g$ is the gradient of group g's moment function with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\Delta}_g$ is the variance of group g's moment function. Further, the asymptotic covariance between $\sqrt{N}(\widehat{\boldsymbol{\tau}}_{g_h} - \boldsymbol{\tau}_{g_h})$ and $\sqrt{N}(\widehat{\boldsymbol{\tau}}_{g_k} - \boldsymbol{\tau}_{g_k})$ is given by $\boldsymbol{D}_{g_h}(\boldsymbol{D}_\infty' \boldsymbol{\Delta}_\infty^{-1} \boldsymbol{D}_\infty)^{-1} \boldsymbol{D}_{g_k}'$. We derive the functional forms of the various matrices for the QLD estimator in the Appendix. ■

Valid inference is easy to obtain because we use a GMM framework. Analytic standard errors are computed and reported by most routine statistical packages implementing GMM estimation. We achieve the desired $\sqrt{N}$-convergence rate because we assume a $\sqrt{N}$-convergent estimator of the factor proxies. Examples include quasi-differencing, common correlated effects, and even principal components, though the latter also requires T to go to infinity for the asymptotic results to hold[11] . Because we have proved asymptotic normality, one can also use the usual nonparametric panel bootstrap. We derive an asymptotically linear representation of the ATT estimates in the Appendix that also allow for the multiplier bootstrap as in Callaway and Karami (2023). We also contain a discussion on deriving standard errors when the factors are not $\sqrt{N}$-convergent.

The asymptotic distribution of $\sqrt{N}(\widehat{\boldsymbol{\tau}}_g - \boldsymbol{\tau}_g)$ generally depends on the estimation of $\boldsymbol{\theta}$ in the first stage by the term $D_g(D_\infty' \Delta_\infty^{-1} D_\infty)^{-1} D_g'$. We can see directly from Theorem 1.2 that a smaller Avar($\sqrt{N}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})$) leads to a smaller Avar($\sqrt{N}(\widehat{\boldsymbol{\tau}}_g - \boldsymbol{\tau}_g)$) (in the matrix sense), strictly so when $\boldsymbol{D}_g$ has full rank. This result also suggests that more efficient estimation of the factors is an important avenue of future work and demonstrates why our general identification result is so powerful: we can

[11] Ahn et al. (2013) and Westerlund et al. (2019) prove fixed-T, $\sqrt{N}$-asymptotic normality of the factor proxies for quasi-long-differencing and common correlated effects, respectively.

use different estimators of the factors if we believe we can achieve substantial efficiency gains.

Estimation of $\boldsymbol{\tau}_g$ is not dependent on the first stage estimation of $\boldsymbol{\theta}$ when $\boldsymbol{D}_g = \mathbf{0}$. A sufficient condition for this equality occurs when the transformed factor loadings for group g center about zero. The fixed-T common correlated effects analysis of Westerlund et al. (2019) implies such a condition. We may also think this condition holds in certain applications where the factor model is relevant. For example, suppose γ_i is exposure to an information shock f_t such that $\gamma_i \in [0,1]$ with probability one. If non-institutional investors of a given asset do not have access to privately held limited information, we would expect $\gamma_i \approx 0$ for units in said group. When the gradient $\boldsymbol{D}_g = \mathbf{0}$ for a given g, the asymptotic variance of $\sqrt{N}(\widehat{\boldsymbol{\tau}}_g - \boldsymbol{\tau}_g)$ is just $\boldsymbol{\Delta}_g$. This quantity is simple to estimate via a nonparametric variance estimator. Let

$$\widehat{\boldsymbol{\Delta}}_g = \frac{1}{N_g - 1}\sum_{i=1}^{N} D_{ig}\left(\widehat{\boldsymbol{\Delta}}_{ig} - \widehat{\boldsymbol{\tau}}_{gG}\right)\left(\widehat{\boldsymbol{\Delta}}_{ig} - \widehat{\boldsymbol{\tau}}_{gG}\right)' \tag{1.21}$$

where $\widehat{\boldsymbol{\Delta}}_{ig} = \boldsymbol{y}_{i,t\geq g} - \boldsymbol{P}(\boldsymbol{F}_{t\geq g}(\widehat{\boldsymbol{\theta}}), \boldsymbol{F}_{t<g}(\widehat{\boldsymbol{\theta}}))\boldsymbol{y}_{i,t<g}$. This estimator is sufficient to generate valid standard errors whenever $\boldsymbol{D}_g = \mathbf{0}$.

Theorem 1.3. *Under Assumptions 1.1-1.5,* $\widehat{\boldsymbol{\Delta}}_g^{-1}\,\mathrm{plim}\,\boldsymbol{\Delta}_g^{-1}$.

1.3.2 Extensions

We conclude this section with a few extensions of our estimator to highlight the flexibility of our approach.

Remark 1.2 (Limited Anticipation)**.** *We can relax the limited anticipation assumption by simply redefining the last pre-treatment period as* $q_g - 1$ *and incorporate the additional* $g - q_g$ *periods into the moment conditions, so long as there are still enough pre-treatment periods to construct the imputation matrix. Then* $\boldsymbol{\tau}_g$ *is a* $T - q_g + 1$ *vector that makes treatment anticipation a testable hypothesis:*

$$H_0 : \tau_{g,q_g} = ... = \tau_{g,g-1} = 0 \tag{1.22}$$

■

Remark 1.3 (Other Aggregate Treatment Effects)**.** *Our estimation method can handle other aggregations of $y_{it} - \hat{y}_{it}(\infty)$. For example, one could aggregate over all post-treatment (i, t) to estimate an overall ATT or over event-time indicators to estimate aggregate event-study estimates.*[12] *Researchers can perform heterogeneity analyses by aggregating for units with different values of X_i like gender, race, or age to estimate a conditional ATT. All one needs to do to estimate such aggregate effects is to correctly specify the unconditional treatment effect moment conditions. If there are a priori restrictions on treatment effects as in Borusyak et al. (2024), these can be imposed on the moment conditions as well.*

We can also derive pre-treatment "placebo" effects by estimating a coefficient on the pre-treatment time periods. The imputation matrix that carries out this estimation is the usual projection matrix $\boldsymbol{P}(\boldsymbol{F}_{t\leq g}, \boldsymbol{F}_{t\leq g})$. Under the no anticipation assumption,

$$\mathbb{E}[(\boldsymbol{I}_g - \boldsymbol{P}(\boldsymbol{F}_{t\leq g}, \boldsymbol{F}_{t\leq g}))\, \boldsymbol{y}_{i,t\leq g} \mid G_i = g] = \boldsymbol{0} \tag{1.23}$$

so that the properly standardized vector of pre-treatment residuals is asymptotically normal. ■

Remark 1.4 (Plotting Estimates)**.** *The proposed estimator can be used to produce estimates for $y_{it}(\infty)$ in all periods for the treated observations:*

$$\hat{y}_{it}(\infty) = \boldsymbol{P}(\boldsymbol{F}_t, \boldsymbol{F}_{t<g})\boldsymbol{y}_{i,t<g} + \overline{y}_{\infty,t} + \overline{y}_{i,t<T_0} - \overline{y}_{\infty,t<T_0} \tag{1.24}$$

where the first term on the right-hand side imputes $\hat{y}_{it}(\infty)$ and the last three terms in the sum 'undo' the within-transformation[13] *. In the pre-treatment periods, our estimates $\hat{y}_{it}(\infty)$ should be approximately equal to the observed y_{it} under our assumptions. Similar to synthetic control estimators, comparing the imputed values to the true value can validate the 'fit' of our model. However, since we have many treated units, doing so unit by unit is not practical. There are two complementary ways to aggregate treated units that will prove useful.*

First, one can aggregate over a group and plot the average of y_{it} and the average of $\hat{y}_{it}(\infty)$ separately for each group $g \in \mathcal{G}$. This will create a set of 'synthetic-control' like plots. To produce

[12] Alternatively, we allow for aggregation of ATT(g, t) estimates as in Callaway and Sant'Anna (2021) by deriving the influence function in the Online Appendix.

[13] Leave this part out if you do not remove the additive effects by hand.

an 'overall' plot, the observed outcome y_{it} and the estimated untreated potential outcome $\hat{y}_{it}(\infty)$ should be 'recentered' to event-time, i.e. reindex time to $e = t - G_i$, so that treatment is centered at event-time 0. Then y_{ie} and $\hat{y}_{ie}(\infty)$ can be aggregated for each value of event-time e. We produce such a plot in our empirical example.

Remark 1.5 (TWFE Specification Testing)**.** *This paper is motivated by the fact that the two-way error model's generality is suspicious in practice. Therefore, we think a test of the two-way error structure versus a more complicated interactive effects model is of practical importance. Ahn et al. (2013) discuss consistent estimation of p. Their tests have a new interpretation under this null hypothesis when testing for p on the residuals $\tilde{y}_{it}$.*

Theorem 1.4. *If Assumption 1.1 and 1.2 hold with $\boldsymbol{F}_t'\boldsymbol{\gamma}_i = \mathbf{0}$ almost surely, then $p = 0$.* ■

If the null hypothesis is true, the more computationally burdensome QLD procedure is unnecessary for estimating the ATTs.[14] *Even if the two-way error model is unrepresentative of the factor structure, Corollary 1.1 shows that mean independence of the factor loadings with respect to treatment timing is sufficient for consistency of TWFE. See the Online Appendix for an additional test of the equality of the factor loadings' conditional means.* ■

1.4 Simulations

We present a brief simulation study to compare our estimator to different TWFE specifications. We specifically study the quasi-differencing factor estimation approach of Ahn et al. (2013) in the first stage because it is used in our empirical example. See Brown et al. (2023a) for simulation evidence for common correlated effects as the first stage estimator. We consider the setting where $T = 8$ and treatment turns on starting in period 6 implying $T_0 = 5$. We draw $N = 200$ observations, which is a relatively small number for a nonlinear estimation problem.

We generate untreated potential outcomes following equation (1.1). We consider the setting with one factor that we generate as a time-trend $f_t = t$.[15] We generate the time fixed effects as

[14] Even if TWFE is consistent, it is not necessarily more efficient than our procedure. See Section 1.4 for example.

[15] In this particular case, if the researcher knew that f_t took this form, then including unit-specific time-trends

$\zeta_t = 0.75 * \zeta_{t-1} + \nu_t$ where $\nu_t \sim N(0,1)$. We generate the unit fixed effects as iid with $\mu_i \sim N(0,4)$ and the factor loadings to be correlated with the unit fixed effects by drawing from $\gamma_i \sim N(\mu_i, 1)$. The error term is generated as an $AR(1)$ process with correlation coefficient 0.75 and is uncorrelated with treatment status. We generate individual-level treatment effect heterogeneity by defining individual treatment effects $\tau_{i\ell}$ to be τ_ℓ times the unit fixed effect but then re-scale the individual effects to have mean equal to $\tau_6 = 1$, $\tau_7 = 2$, and $\tau_8 = 3$ and for the variance of $\tau_{i\ell}$ to be one. For example, $\tau_{i6} = (\mu_i + 2)/2$.

We generate a covariate $w_i = \gamma_i + \xi_i$ where ξ_i is white-noise measurement error. w_i will be used as a covariate in some TWFE specifications and as our instrument for our factor-model estimation. In the baseline simulation, we consider the case where $\xi_i \sim N(0,1)$, which creates a signal-to-noise ratio for the instrument of 1/2. In a set of simulations, we vary the level of noise to see how the instrument strength affects estimates. These results will allow us to compare our methods to those that use noisy measurements of unobserved heterogeneity.

We consider three data-generating processes. First, we consider the true two-way error model where there are no interactive effects. In this case, the two-way fixed effects estimator should be unbiased. Second, we generate outcomes with the factor model described above. Treatment is then assigned completely randomly with probability of treatment at 50% for all units. This implies that the factor loadings are uncorrelated with treatment status, which Corollary 1.1 shows is sufficient for the TWFE imputation procedure to be consistent. Third, we generate treatment with probability increasing in the factor loading such that parallel trends fail (since treated units are more exposed to the time-trend in f_t). In particular, we form the term

$$\pi_i = 0.5 + \frac{\gamma_i}{\max_i \gamma_i - \min_i \gamma_i} \tag{1.25}$$

We normalize this term by the mean of π_i so that the unconditional probability of treatment stays at 50%.

We estimate event-study treatment effects using four estimators. First, we estimate the

would fix this problem. However, we emphasize that f_t is generally not observable. We include this simple form of f_t so that the expected bias of TWFE is easy to compute: $t * (\mathbb{E}[\gamma_i \mid D_i = 1] - \mathbb{E}[\gamma_i \mid D_i = 0])$.

classical two-way error model using ordinary least squares (OLS), i.e. the TWFE estimator. Second, we estimate the two-way error model using the imputation estimator proposed by Borusyak et al. (2024) and Gardner (2022).[16] Third, we augment the two-way error model by including a noisy measure of the factor loadings. This is sometimes done by applied researchers in an attempt to control for confounders. That is, they model outcomes as

$$y_{it} = \mu_i + \lambda_t + w_i \beta_t + u_{it} \tag{1.26}$$

where w_i is a time-invariant covariate and β_t allows for trends to vary based on w_i. In the case where $w_i = \gamma_i$, i.e. the factors are observable, this model is correctly specified. However, when $\text{Var}(\xi_i) > 0$, i.e. the covariates are noisy measures for the underlying factor loadings, model (1.26) will only partially absorb the factor model. We compare this method to our estimator using the QLD transformation of Ahn et al. (2013) to estimate the factors.[17] The covariate w_i is our instrument in the first stage to estimate the QLD parameters. See Remark 1.1.

Results are presented in table 1.1. Each panel presents results from each of the three data-generating processes described above. For each estimate, we present the average bias for the estimate as well as the mean-squared error. For Panel A where the outcomes are generated under the two-way fixed effect model (i.e. without a factor structure), all estimators are unbiased for the treatment effects, but the more robust factor imputation pays an efficiency cost with larger mean-squared error. However, this flips in Panel B where outcomes are generated under a factor model but with parallel trends holding for the two-way error model. In this case, all estimators are still unbiased but the factor imputation estimator is the most efficient because it absorbs the factor-structure that is present in the error term for the two-way error model.

Turning to where parallel trends does not hold in Panel C, we see that only our factor-imputation estimator is unbiased. This result emphasizes that our estimator is robust for parallel trend violations coming from differential exposure to macroeconomic factors. The magnitude of bias present in the two-way error models is growing from τ_6 to τ_8 due to the factor being a linear

[16] We use the R package `did2s` (Butts and Gardner, 2022) for estimation.

[17] We use the `Optim.jl` package for GMM estimation (Mogensen and Riseth, 2018).

Table 1.1: Monte Carlo Simulation

Panel A: Two-way error model.

	Bias $(\hat{\tau}_6)$	MSE $(\hat{\tau}_6)$	Bias $(\hat{\tau}_7)$	MSE $(\hat{\tau}_7)$	Bias $(\hat{\tau}_8)$	MSE $(\hat{\tau}_8)$
TWFE	0.00	0.01	-0.00	0.02	0.00	0.02
TWFE Imputation	0.01	0.01	0.00	0.02	0.01	0.02
TWFE Imputation with Covariates	0.01	0.01	0.00	0.02	0.01	0.02
Factor Imputation	-0.00	0.04	-0.01	0.11	-0.01	0.24

Panel B: Factor Model. Parallel Trends Hold

	Bias $(\hat{\tau}_6)$	MSE $(\hat{\tau}_6)$	Bias $(\hat{\tau}_7)$	MSE $(\hat{\tau}_7)$	Bias $(\hat{\tau}_8)$	MSE $(\hat{\tau}_8)$
TWFE	0.00	0.11	0.00	0.43	0.01	0.95
TWFE Imputation	0.00	0.94	0.00	1.67	0.01	2.60
TWFE Imputation with Covariates	0.00	0.17	0.00	0.29	0.01	0.44
Factor Imputation	-0.00	0.02	-0.00	0.03	0.00	0.05

Panel C: Factor Model. Parallel Trends Do Not Hold

	Bias $(\hat{\tau}_6)$	MSE $(\hat{\tau}_6)$	Bias $(\hat{\tau}_7)$	MSE $(\hat{\tau}_7)$	Bias $(\hat{\tau}_8)$	MSE $(\hat{\tau}_8)$
TWFE	-1.63	2.77	-3.27	11.05	-4.90	24.84
TWFE Imputation	-4.90	24.81	-6.53	44.12	-8.16	68.93
TWFE Imputation with Covariates	-0.92	1.06	-1.22	1.88	-1.53	2.93
Factor Imputation	0.01	0.03	0.01	0.05	0.02	0.09

Notes. This table presents a set of simulations with 10000 simulations. Each panel contains one of three data-generating processes described in the text. Each row in a panel consists of one of the four treatment effect estimators as described in the text. The columns report average bias and mean-squared error for the three post-treatment treatment effects.

time-trend, implying parallel trend deviations grow worse over time.

It is worth noting that while including $w_i\beta_t$ in the model does remove some bias, the estimates still perform worse than our imputation procedure due to w_i being a noisy measure. To highlight the problems with noisy proxies for factor loadings, figure 1.1 presents a set of simulation results where the covariate w_i has different amount of noise added in. In particular, we choose different values of $\text{Var}(\xi_i)$ to have different signal-to-noise measures. The signal-to-noise definition is

$$\text{signal to noise ratio} = \frac{\text{Var}(\gamma_i)}{\text{Var}(\gamma_i) + \text{Var}(\xi_i)} \tag{1.27}$$

For each signal to noise ratio, we estimate the TWFE imputation estimator with covariates and the factor model imputation estimator. Figure 1.1 presents the results of estimates for τ_8. At one extreme, where the signal to noise ratio is approximately 0, i.e. ξ_i is white noise, the estimated bias for the TWFE imputation estimator is the same as the TWFE imputation estimator that does not include covariates. At the other extreme, where the signal to noise ratio is approximately 1, i.e. $w_i = \gamma_i$, the bias is completely removed. Regardless, the factor model imputation estimator is unbiased in all cases. This experiment echos the results of Kejriwal et al. (2021). However, we note that our results are still generous to estimators that use such noisy measure because we generate w_i as an unbiased estimator of γ_i. The instrument requirement for QLD estimation does not require unbiased estimation of $\boldsymbol{\gamma}_i$ for identification of the normalized parameters.

1.5 Application

We revisit the literature on estimating local labor market effects of Walmart store openings (Basker, 2005; Neumark et al., 2008; Volpe and Boland, 2022). The primary identification concern is that Walmart targets where to open stores based on local economic trajectories (Neumark et al., 2008). For instance, if Walmart targeted areas with positive underlying economic fundamentals in anticipation of their growing consumptive expenditures, then the non-treated counties would fail to be a valid counterfactual group in the two-way error model. Indeed, we observe significant differences in both employment trends for treated counties in our data. Volpe and Boland (2022)

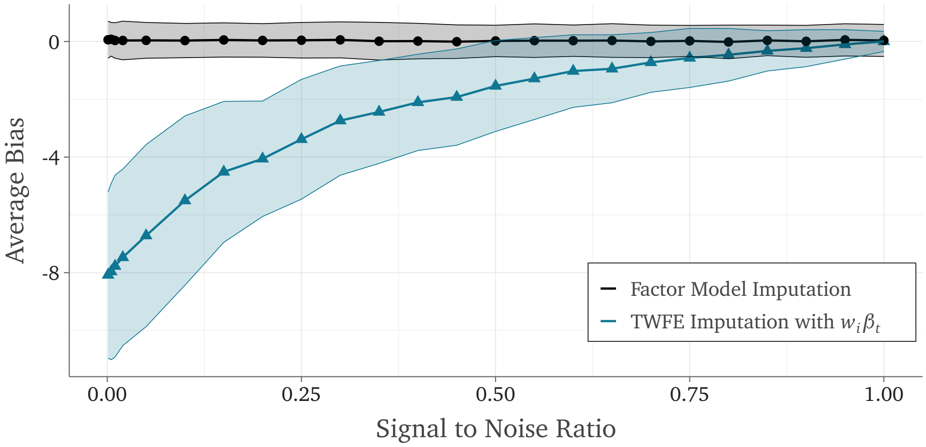


Figure 1.1: Bias of TWFE Imputation with Covariates

Notes. This figure plots the average and empirical 95% confidence intervals for treatment effect estimates in the final period, $\hat{\tau}_8$. We estimate the TWFE imputation estimator that includes $w_i \beta_t$ linearly in the model and our the factor imputation we propose using w_i instead as an instrument. We vary the signal to noise ratios of w_i to make it a better or worse measure for the factor loading. For each signal to noise ratio, we run 5000 simulations.

point to conflicting results on retail employment with two leading papers finding effects of opposite signs. Employing different instrumental variable strategies, Basker (2005) finds positive effects on retail employment while Neumark et al. (2008) finds negative effects. For this reason, we revisit this question with an alternative strategy to answer this question.

We construct a dataset following the description in Basker (2005). In particular, we use the County Business Patterns dataset from 1964 and 1977-1999, subsetting to counties that (i) had more than 1500 employees overall in 1964 and (ii) had non-negative aggregate employment growth between 1964 and 1977.[18] We use a geocoded data set of Walmart openings from Arcidiacono et al. (2020) to construct our treatment variable. Our treatment dummy is equal to one if the county has any Walmart in that year and our group variable denotes the year of entrance for the *first* Walmart in the county. [19] We drop any county that was treated with $g \leq T_0 = 1985$ so that we we have 9 pre-periods to use when estimating the factor model. Our remaining sample consists of 1274 counties (about 500 fewer than the sample used in Basker (2005) since we drop units treated between 1977 and 1985). We estimate impacts on retail and wholesale employment.[20] Walmart is a more vertically integrated business, so we expect Walmart to compete in the retail and the wholesale sectors (Basker, 2005).

First, we estimate the two-way fixed effect imputation estimator proposed by Borusyak et al. (2024) and estimate event-study effects on (log) retail and wholesale employment. In particular, we use the following model

$$\log(y_{it}) = \mu_i + \lambda_t + \sum_{\ell=-22}^{13} \tau^{\ell} d_{it}^{\ell} + u_{it} \tag{1.28}$$

where i denotes county, t denotes year, y_{it} is either retail or wholesale employment, and $d_{it}^{\ell} = 1(t - g_i = \ell)$ are indicator variables denoting event-time. Results of the event-study estimates are presented in panel (a) of figure 1.2 and figure 1.3.

For both retail and wholesale employment, counties receiving Walmarts had faster employment

[18] We use the 1977-1999 dataset with imputed values from Eckert et al. (2021).

[19] For our sample 82.4% of our counties receive ≤ 1 Walmart and another 10.4% receive two Walmarts in the sample, alleviating some concerns of making the treatment binary.

[20] Retail employment corresponds with NAICS 2-digit codes 44 and 45 and wholesale employment corresponds to NAICS 2-digit code 42.

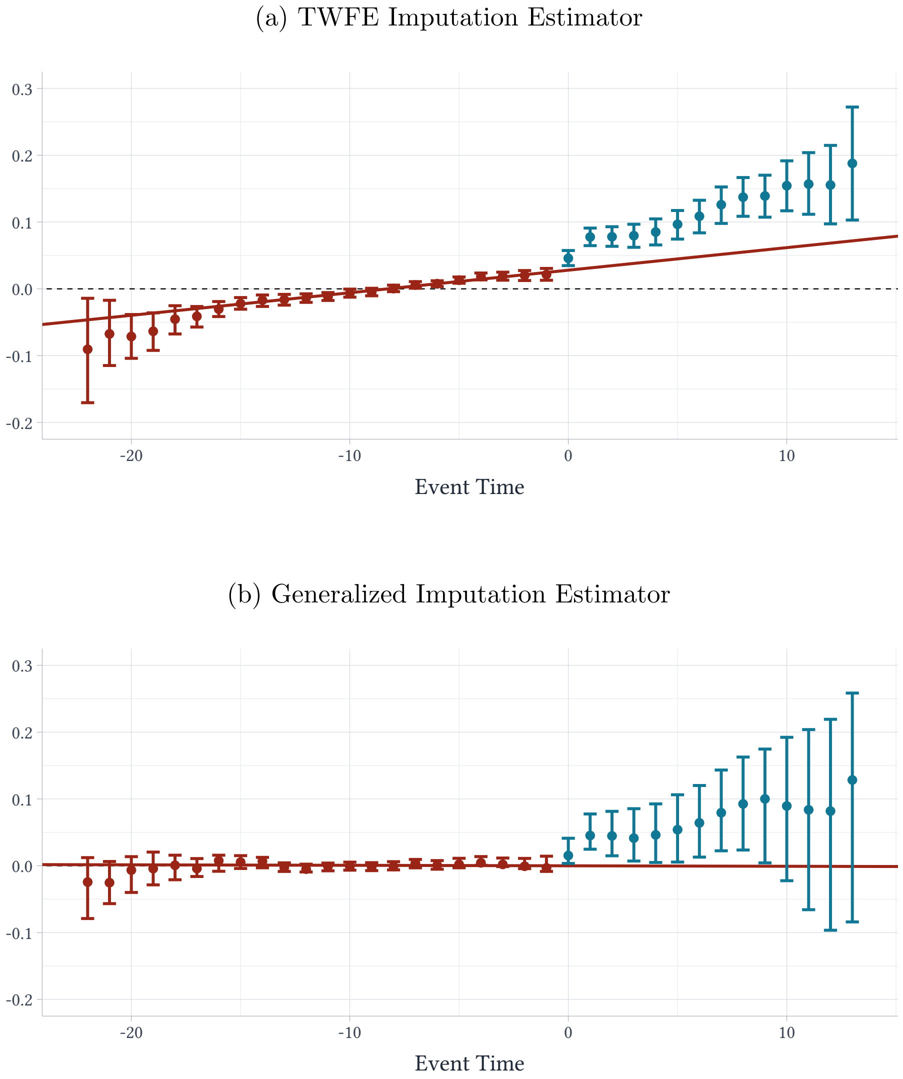
growth relative to the control counties, emphasizing our concern over endogenous opening decisions. In the spirit of Freyaldenhoven et al. (ming) and Rambachan and Roth (2023), we draw the line of best fit for the 15 most-recent pre-treatment estimates ($\hat{\tau}^{\ell}$ for $-15 \leq \ell < 0$) and extend it into the post-treatment estimates. For both retail and wholesale employment, the pre-trend lines would suggest that a large portion of the estimated effect is a continuation of already existing trends. However, there still appears to be positive effects on retail employment (if the pre-trend violations were indeed linear in the post-treatment period).

We use the QLD estimator of Ahn et al. (2013) to estimate the factors as described in remark 1.1. For this factor estimator, we need a set of instruments that satisfy the two standard instrument requirements: relevancy and exclusion. Intuitively, the relevancy restriction requires that the instruments are correlated with the full vector of factor-loadings. That is, the instruments should be selected as 'proxies' for the kinds of economic factor-loadings that the researcher is concerned of. The exclusion restriction requires that the instrument values are uncorrelated with location-specific idiosyncratic shocks. For this reason, we use baseline covariate values as instruments to avoid shocks to the covariates that are correlated with shocks to the outcome variable.

We select instruments that we suspect are driven by the general macroeconomic trends that cause differential retail employment growth in the 1980s and 1990s. For example, retail employment is likely driven by consumptive expenditures which in turn are reflective of local labor market trends. Therefore, we use instruments that we think proxy for characteristics that determine local labor market trends. Specifically, we use the 1980 baseline values of the following variables as instruments: share of population employed in manufacturing, shares of population below and above the poverty line; shares of population employed in the private-sector and by the government, and shares of population with high-school and college degrees.[21] We use baseline shares to prevent our instruments from picking up on contemporaneous economic shocks that could be correlated with Walmart opening, i.e. to avoid violations of the exclusion restriction. Note that instead of estimating ATT(g, t), we estimate ATT$^{\ell}$ pooling across (i, t) with $\ell = t - g_i$ as described after Theorem 1.2.

[21] All of these values are obtained from 1980 Census Tables accessed from Manson (2020).

Figure 1.2: Effect of Walmart on County log Retail Employment

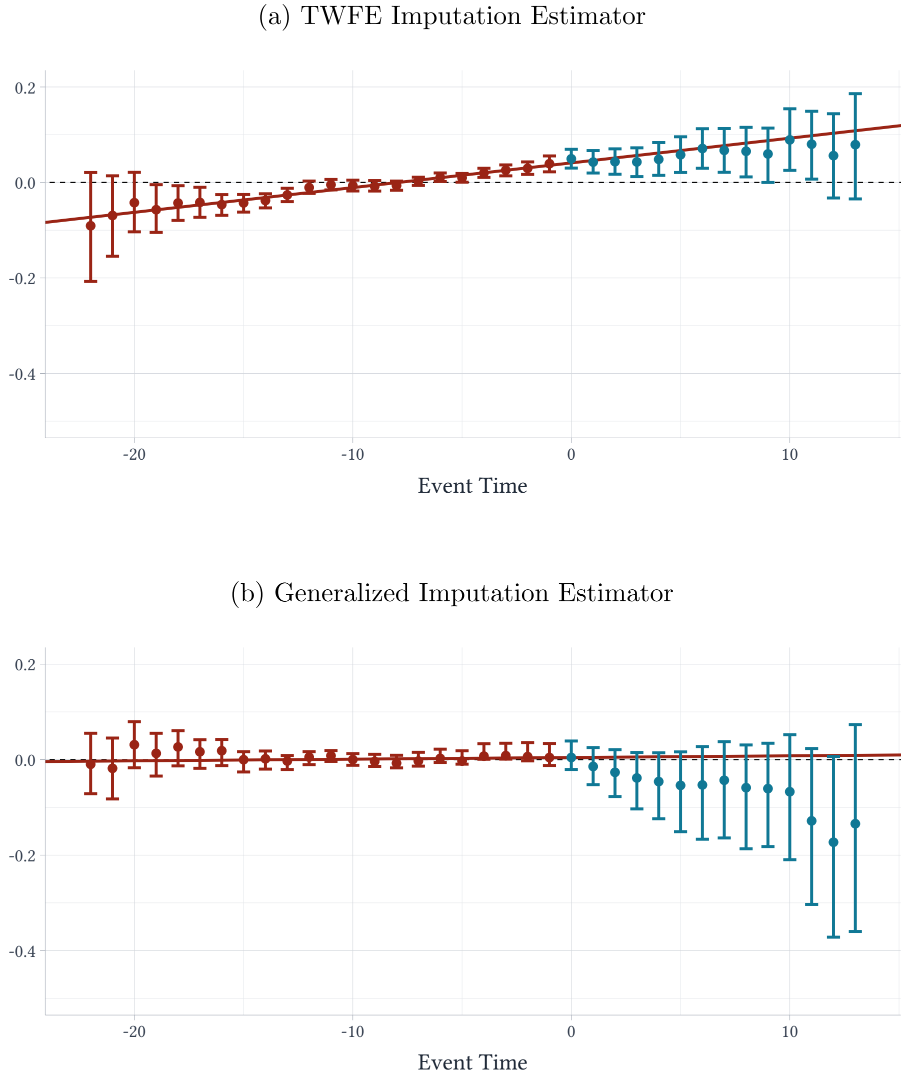


(a) TWFE Imputation Estimator

(b) Generalized Imputation Estimator

Notes. This figure plots point estimates and bootstrapped 95% confidence intervals for event-study treatment effects on log retail employment. Panel (a) estimates effects using the TWFE imputation estimator proposed in Borusyak et al. (2024). Panel (b) estimates effects using the generalized imputation estimator we propose in Section 1.3 with $p = 2$ and using the following instruments: 1980 share of population employed in manufacturing, 1980 shares of population below and above poverty line; 1980 shares of population employed in private-sector and by the government, 1980 shares of population with high-school degree and college degree. The red lines correspond to a linear estimate of pre-treatment point estimates for event time -15 to -1 and is extended into the post-treatment periods.

Figure 1.3: Effect of Walmart on County log wholesale Employment



(a) TWFE Imputation Estimator

(b) Generalized Imputation Estimator

Notes. This figure plots point estimates and bootstrapped 95% confidence intervals for event-study treatment effects on log wholesale employment. Panel (a) estimates effects using the TWFE imputation estimator proposed in Borusyak et al. (2024). Panel (b) estimates effects using the generalized imputation estimator we propose in Section 1.3 with $p = 1$ and using the following instruments: 1980 share of population employed in manufacturing, 1980 shares of population below and above poverty line; 1980 shares of population employed in private-sector and by the government, 1980 shares of population with high-school degree and college degree. The red lines correspond to a linear estimate of pre-treatment point estimates for event time -15 to -1 and is extended into the post-treatment periods.

The results of our estimator are presented in panel (b) of figure 1.2 and figure 1.3.[22] For retail employment, there is basically no pre-trend violations with the pre-treatment point estimates centered on zero. After removing the pre-existing economic trends, the point estimates are smaller than estimated by the two-way error model with an estimated effect on employment of around 6% on average in the post-treatment periods. Evaluated at the median baseline retail employment of 1417 employees, this would imply an increase in about 85 jobs, which is in line with the estimates of Basker (2005) and Stapp (2014) who use alternative instrumental variables strategies. It is important to note that post-treatment estimates are noisier than the TWFE estimates largely due to estimating the factor proxies in the first stage. This problem is at its worst for the furthest event-times due to very few counties being averaged over in the last few bins. We view this as a worthy trade-off since the point estimates are much less likely to be biased.

Turning to wholesale employment, we see a similar story with our estimator removing most of the pre-trend violations. In this case, however, the estimated effects flip signs with an estimated effect of around -6%, although they are not statistically significant at the 5% level. Evaluated at the 1977 median wholesale employment of 410, this suggests a decrease of about 25 jobs, which is similar to what Basker (2005) finds. Overall, we find effects very much in line with those reported in Basker (2005).

Our estimator allows for any root-N consistent estimator of the factor's column space to be 'plugged-in' and used for estimation of treatment effects. To show the versatility of the method, we use three different factor estimators in figure 1.4. First, we use our original quasi-differencing estimator from figure 1.2. Second, we use the common correlated effects (CCE) estimator originally proposed in Pesaran (2006). This estimator uses a set of covariates, $\boldsymbol{X}$, which are generated by the same factors, $\boldsymbol{F}$, as the outcome variable:

$$X_{it} = \boldsymbol{\alpha}_i' \boldsymbol{F}_t + \nu_{it}. \tag{1.29}$$

[22] We carry out the test to determine the correct number of factors p following the discussion in Ahn et al. (2013). For retail, the p-value of the over-identification test were as follows: p = 0 with a p-value of 1.56e-5; p = 1 with a p-value of 0.001; p = 2 with a p-value of 0.133. Since $p = 2$ is the first value where we fail to reject the null at a 10% level, we set $p = 2$. Similarly, we selected $p = 1$ for wholesale since the p-values were: p = 0 with a p-value of 0.049; and p = 1 with a p-value of 0.40.

Under this assumption, the cross-sectional averages of X (averaged over the never-treated group) consistently span the column space of $\boldsymbol{F}$. In our application, we use log employment for the manufacturing, construction, agriculture, and healthcare 2-digit NAICS codes. The choice of these covariates is plausible if the same sort of national shocks that affect retail employment also affect these other sectors. We more formally analyze this estimator in Brown et al. (2023a), which derives the asymptotic distribution of the estimates. One advantage of this factor estimator is that it allows decomposition of treatment effects into direct effects and mediated effects that operate through the covariates, X_{it}.

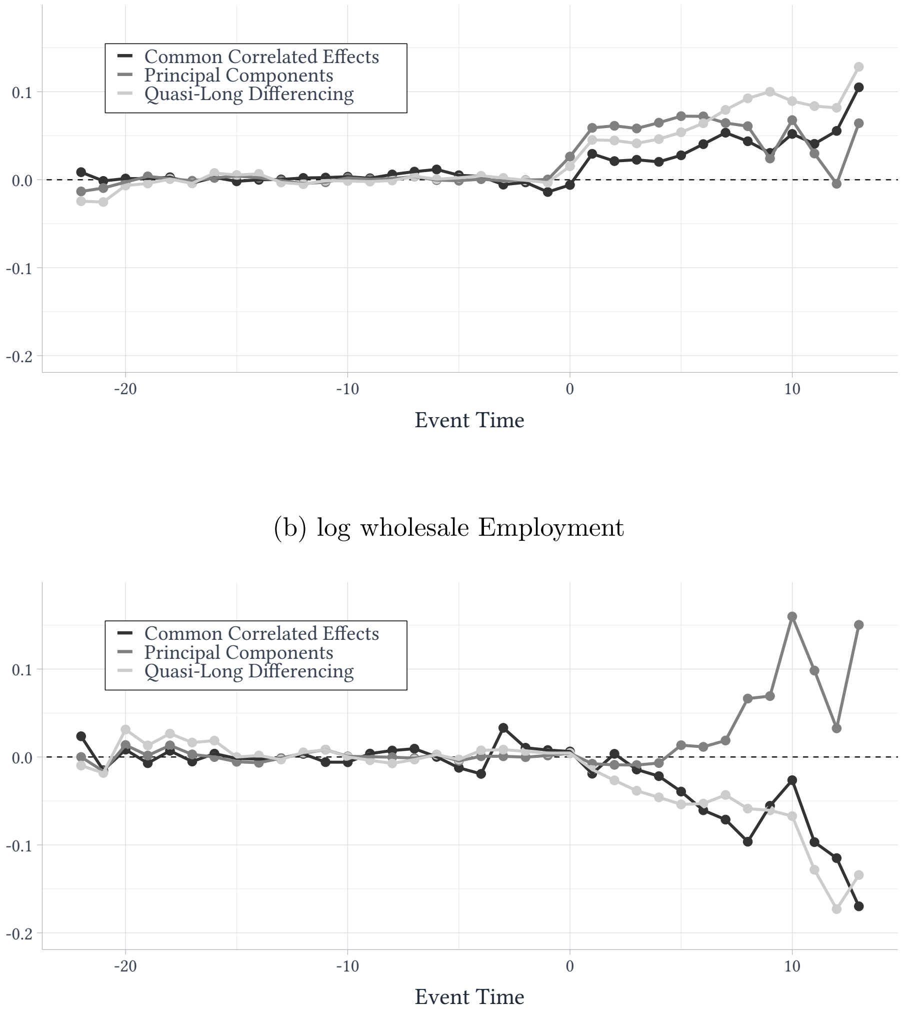
Last, we use the principal components estimator originally proposed in Bai (2009). This estimator uses the eigenvectors of the matrix $\boldsymbol{YY'}$ with the p largest eigenvalues as estimates for $\boldsymbol{F}$.[23] The advantage of this estimator is that no instrument or additional covariates are required. However this comes at the cost of requiring long panels, which may be infeasible to assume in our application.

The results of each estimator are presented in figure 1.4. All three estimators are effective at removing underlying trends that the treated counties experienced. Moreover, the estimated effects are similar between estimators suggesting that all three are doing a good job at estimating the underlying factors. This figure highlights the broad applicability of our identification results, allowing the factor estimator of choice to be tailored to the research context at hand. In panel (b), we use log wholesale employment as an outcome. The CCE and the quasi-differencing estimators produce very similar results, while the principal components estimator suggests positive growth in employment outcomes in later years. Corresponding confidence intervals are very large, suggesting that these results are too noisy to draw any meaningful conclusions. This could be due to wholesale employment being too auto-correlated for the factor estimates to be consistent, or because we do not have a large enough time series to get a meaningful asymptotic approximation of the factors.

To highlight the importance of the uncertainty from estimation of the factors in the first stage,

[23] This imputation estimator is proposed by Xu (2017) in the context of large panels. The author uses an alternative identification strategy that fails to work in short panels.

Figure 1.4: Generalized Imputation Estimator for Effect of Walmart on County Employment with Different Factor Estimators

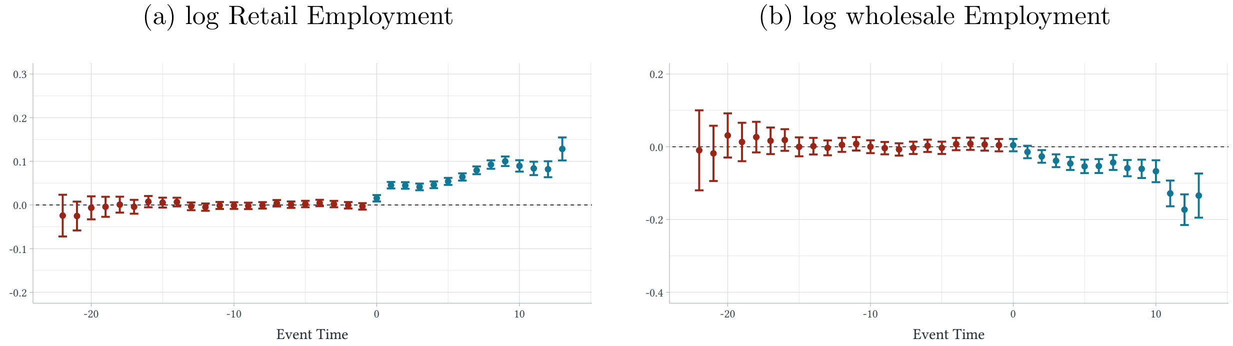


(a) log Retail Employment

(b) log wholesale Employment

Notes. This figure presents estimated treatment effects of Walmart entry on county-level log retail employment using the generalized imputation procedure proposed in section 1.2.1. The factor estimation procedures include the principal components estimator proposed in Bai (2009), the common correlated effects estimator proposed in Pesaran (2006), and the quasi-differencing estimator proposed in Ahn et al. (2013). Details of the estimation procedures appear in the text.

Figure 1.5: Generalized Imputation Estimator for Effect of Walmart on County Employment with Naive Standard Errors



(a) log Retail Employment

(b) log wholesale Employment

Notes. This figure recreates estimates from panel (b) of figure 1.2 and figure 1.3 with confidence intervals formed ignoring the uncertainty deriving from first-stage estimates of θ.

we recreate confidence intervals from our generalized imputation estimator with the QLD first stage using the nonparametric standard errors that are derived in Theorem 1.3. Results are given in figure 1.5. The standard errors on point estimates are far smaller, with estimates becoming strongly significant in wholesale Employment. This result shows an important step for future research in finding more efficient estimates of the factors. For instance, we consider the common correlated effects estimator in a follow-up paper. The CCE model generally implies that the nonparametric standard errors are valid when there is a common factor model for time-varying covariates.

1.6 Conclusions

We consider identification and inference of functions of heterogeneous treatment effects in a linear panel data model. We show how to relax the usual parallel trends assumption by introducing a linear factor model in the error. Our main identification result shows that a consistent estimator of the unobserved factors is all that one needs to estimate the dynamic treatment effect coefficients. This result is general and can be implemented by a number of modern interactive fixed effects estimators, such as quasi-long-differencing, internally generated instruments, common correlated effects, or principal components, allowing for both large and small numbers of pre-treatment time periods. Further work can demonstrate both theoretical and finite-sample properties of these various estimators of the factors and how they affect to ATT estimation, especially for larger time series. The GMM imputation framework should also be examined in the context of unbalanced panels as in Rai (2023).

While a factor model nests the usual two-way error structure, we explicitly model the level fixed effects in addition to the factors. This setting allows us to provide useful tests for the consistency of the TWFE estimator. We also show that one must remove the unit and time fixed effects in a particular way so as to preserve the common factor structure in all time periods for all individuals. We provide such a transformation and prove a novel identification result for TWFE imputation estimators of ATTs.

We implement the QLD estimator of Ahn et al. (2013) in a study of the local impact of Walmart openings. We demonstrate findings consistent with the IV estimation strategy of Basker (2005). Our estimator is shown to remove pre-trends that bias the usual TWFE estimates. Similar results are found using common correlated effects in the first stage. A principal components estimator is also explored, but performs suspiciously for the given problem. The QLD identification scheme can also allow sequentially exogenous outcomes like those generated by dynamic models. We leave this possibility for future study.

Chapter 2

Difference-in-Differences with Spatial Spillovers

2.1 Introduction

Empirical work in economics often considers settings where a policy targets units grouped by geographic boundaries but the effect of treatment spills over onto 'nearby' units.[1] For example, individuals in the surrounding area can travel across borders to receive treatment (e.g. a new hospital serves nearby residents); or shocks to a labor market can affect nearby areas (e.g. a new factory increases service sector spending in the entire commuting zone). In these settings, a common approach involves using a structural model to account and control for general equilibrium effects. For example, trade models generate a market access term that effectively controls for such effects (Donaldson and Hornbeck, 2016), or network models suggest a linear-in-means model controlling for the average characteristics of a unit's peers, e.g. proportion of peers treated (Manski, 1993; Goldsmith-Pinkham and Imbens, 2013; Miguel and Kremer, 2004).

However, researchers often wish to remain agnostic to structural assumptions, preferring minimal and transparent restrictions that, when made, allow for identification of treatment effects. This paper uses the potential outcomes framework proposed by Vazquez-Bare (2023) to consider non-parametric identification in difference-in-differences settings under the presence of spatial spillovers. In this framework, potential outcomes are now characterized by a unit's treatment status as well as the 'exposure' to the treatment status of other units. Identification arguments are then made by

[1] The framework of this paper applies to any setting with a well-defined measure of distance, e.g. geographic distance, economic distance such as supply chains, node distance in a graph, or social relationships in schools or cities.

imposing assumptions on the potential outcomes.

Under the more general potential outcome framework, there are many potential treatment effects that can be formed. I consider identification of two relevant treatment effects that are common in the literature on interference (Sävje et al., 2021). First, there is the 'switching effect' which is the relevant parameter for local policymakers who want to answer, "What is the effect of switching my treatment status, holding fixed all other units' treatment?". I show that identification of the switching effect is a difficult problem which in a difference-in-differences framework would require the researcher to identify treated and control units that have the same level of exposure.[2]

The second effect of interest is the 'total effect', which is the relevant parameter for national policymakers who want to answer, "What is the average effect of implementing the entire treatment regime?". This treatment effect is useful in post-hoc analyses analyzing the effect of implementing the observed treatment vector. This paper identifies simple and interpretable assumptions that allow for identification of the total effect. In settings where researchers are willing to impose an assumption that spillovers are 'local', i.e. spillovers only impact units within a certain distance of treatment, I show the total effect can be identified with a parallel trends assumptions between treated and non-affected 'far-away' units. The local spillovers assumption is reasonable in settings where travel is the primary driver of spillovers (e.g. people in nearby jurisdictions travel to the treated location). However, this assumption may not hold in the case of large general equilibrium shocks that affect units far away from the treatment area, such as when New York's economy impacts San Francisco.

In this framework, I evaluate common practices in empirical work. First, I show that the standard difference-in-differences estimator produces biased estimates for the total effect. This bias arises from the fact that untreated units that are "close" to treated units experience treatment effects, thereby failing to identify the counterfactual trend. The difference-in-differences estimate averages the spillover onto the "close" control units into the untreated units' change in outcomes. As a result, the spillover is subtracted from the estimated treatment effect, introducing bias in the

[2] Xu (2023) takes this approach and discusses estimation and inference about this estimate.

opposite sign of the spillover effect.

This problem is generally well understood, and it is common to either drop or include a dummy for nearby units in the standard two-way fixed effect model. This method consistently estimates the total effect under the assumptions I present in this paper. In this sense, this paper formalizes and clarifies the assumptions needed for this method.

To show the importance of considering spillover effects and the utility of my estimators, I revisit analyses of place-based policies in urban economics in section 2.3.1. I revisit the analysis of the Tennessee Valley Authority by Kline and Moretti (2014). The Tennessee Valley Authority was a large-scale New Deal program that lowered the cost of power for industrial firms (Kitchens, 2014). The scale of federal investment in the region was large and the pro-manufacturing benefits likely spread further than the Authority's boundary due to the electrification infrastructure and agglomeration economies (Severnini, 2023). I show that estimation by difference-in-differences fails to account for these spillovers and therefore the authors obtain biased estimates of the total effect of the Tennessee Valley Authority. For agriculture employment, I find that the long-run spillovers cause the original estimates to be about 40 percent too small for agriculture employment and 40 percent too large for manufacturing employment.

Last, in section 2.4, I extend estimation of treatment and spillover effects into settings with staggered-treatment timing by extending the work of Gardner (2022); Borusyak et al. (2024).[3] The proposed two-stage estimator first estimates unit and time fixed-effects using untreated/not-yet-treated observations. Since some control/not-yet-treated units can be affected by spillovers, these units must be removed to consistently estimate the unit and time fixed-effects. Then these estimated unit and time fixed-effects are subtracted from the observed outcome *in the full sample.* The resulting differenced outcomes are then regressed on the treatment and spillover variables to estimate treatment and spillover effects. In the appendix, I demonstrate the method by revisiting the analysis by Bailey and Goodman-Bacon (2015) of Community Health Centers which provided low-cost primary care to impoverished areas.

[3] I also briefly discuss how to adapt the estimation strategy of (Callaway and Sant'Anna, 2020).

This paper contributes to the literature that focuses on estimation of treatment effects with spatial spillovers using the difference-in-differences framework. Most work derives results for specific spillover mappings (Clarke, 2017; Berg and Streitz, 2019; Verbitsky-Savitz and Raudenbush, 2012a; Delgado and Florax, 2015). My paper is the first to consider non-parametric identification in terms of general potential outcomes. My paper also advances the literature by considering estimation of treatment effects and spillover effects in settings with staggered-treatment timing. If I assume the particular functional forms for potential outcomes, I arrive at the same bias equation as theirs. Xu (2023) complements this paper well, focusing on identification and inference on the 'switching effect'.

This paper relates to a broad literature on spillover effect estimation in randomized experiments.[4] There are two main strands of this literature. First, there is a large literature on the estimation of treatment effects in the presence of spillovers using a 'partial identification' framework where units are in distinct treatment clusters and outcomes depend on the treatment status within the observation's cluster only.[5] Estimation compares units in the partially treated clusters with control units in completely untreated clusters which do not receive spillover effects. This allows standard difference-in-differences estimation of both the total effect (treatment effect on the treated) and spillover effects (treatment effect on the untreated in the treated clusters). However, my proposed estimator focuses on a setting without distinct clusters.

There is also a nascent literature exploring estimation of treatment and spillover effects which does not require a completely untreated cluster in experimental settings (Sävje et al. (2021); Vazquez-Bare (2023); Hu et al. (2021); Yu et al. (2022)). Those papers' identification results rely on *design-based* assumptions around the treatment-assignment mechanism. Difference-in-differences, however, relies on *model-based* assumptions on the potential outcomes (i.e. parallel-trends assumption) for identification in non-experimental settings. I contribute to this literature by formalizing identification results of different treatment effects in non-experimental settings. Wang et al. (2020) consider

[4] See Sobel (2006) and Hudgens and Halloran (2008) for early work. See Hu et al. (2021), Sävje et al. (2021), and Vazquez-Bare (2023) and references therein for recent work on this.

[5] Angelucci and Di Maro (2016) provides an overview of estimation of treatment effects in the presence of "within-group" spillovers. Empirical examples include Halloran and Struchiner (1995); Sobel (2006); Miguel and Kremer (2004); Angrist (2014).

estimation of treatment effects in spatial settings under design-based settings and make a similar 'local' spillovers assumption as this paper.

2.2 Theory

There are a set of units $i \in \{1, \dots, N\}$ observed for periods $t \in \{0, 1\}$ and treatment turns on between periods for some of the units. The framework is extended to staggered treatment adoption below. Let D_i denote an indicator for unit i being treated and the vector of treatment assignments as $\boldsymbol{D} = (D_1, \dots, D_n)'$. Since units can be impacted by the vector of treatment assignments, the general potential outcome notation is given by $Y_{it}(\boldsymbol{D})$. To simplify the high-dimensional set of potential outcomes, we follow Aronow and Samii (2017); Vazquez-Bare (2023), and introduce an 'exposure mapping' which summarizes how a unit is exposed to spillovers: $h_i(\boldsymbol{D}) : \{0, 1\}^N \to \mathcal{H}$ for some space $\mathcal{H}$ with $\dim(\mathcal{H}) \leq N$. For example, researchers might think $h_i(\boldsymbol{D})$ to be an indicator variable equal to one if any contiguous units are treated. The exposure mapping simplifies the potential outcome notation by imposing that if two treatment vectors $\boldsymbol{D}$ and $\tilde{\boldsymbol{D}}$ have $h_i(\boldsymbol{D}) = h_i(\tilde{\boldsymbol{D}})$, then the potential outcomes are also equal. Potential outcomes therefore can be written as $Y_{it}(D_i, h_i(\boldsymbol{D}))$. In the counterfactual world without treatment, we write the potential outcome as $Y_{it}(0, \boldsymbol{0})$ where $\boldsymbol{0} \in \mathcal{H}$ is defined as zero-exposure.

Often, researchers will try to parameterize potential outcomes, i.e. selecting a function $h_i(\cdot)$, by reasoning through the nature of the spillovers and subsequent estimation will rely on the functional form assumptions made by researchers. This paper will take an alternative approach to identification and estimation which will allow for estimation of treatment effects without requiring direct knowledge of the functional form of the exposure mapping/potential outcomes.

Following Borusyak et al. (2024) and De Chaisemartin and d'Haultfoeuille (2024), our analysis views our panel and the treatment vector $\boldsymbol{D}$ (and hence exposure mappings) as fixed. Therefore, the uncertainty in the framework comes from stochastic potential outcomes.[6] In most of the related

[6] See appendix section A.2 of Borusyak et al. (2024) for further discussion on fixed-design. Xu (2023) focuses on inference in a similar finite population setting.

literature on interference, the treatment design and hence the distribution of exposure mappings is known (given h) so expectations can be taken without conditioning on the vector of exposure mappings.[7] The focus of this article is on difference-in-differences where identification comes from assumptions on the potential outcomes.

2.2.1 Treatment and Spillover Effects

In setups without spillovers, the treatment effect for an individual unit is well defined as $\tau_i \equiv Y_{i1}(1) - Y_{i1}(0)$. In the presence of spillovers, multiple treatment effects can be defined in this setting. This subsection will define commonly used estimands and discuss their interpretation. In the following sections, we will discuss identification strategies.

The natural analogue to the above treatment effect which I will label the 'switching effect' is:

$$\tau_{i,switch}(\boldsymbol{h}) \equiv Y_{i1}(1, \boldsymbol{h}) - Y_{i1}(0, \boldsymbol{h}).$$

This is the effect of changing only unit i's treatment status while keeping their exposure to spillovers constant at some value $\boldsymbol{h}$. This treatment effect is policy-relevant as it summarizes what would happen if a 'local' policymakers decide to enact the policy for unit i (implicitly keeping $\boldsymbol{h}$ constant).[8] It is important to note that the switching effect can depend on the level of exposure. For instance, consider the construction of libraries in towns (Berkes and Nencka, 2021). The effect of a new library in a town that is far away from any other library (exposure of $\mathbf{0}$) is different than a town that is very close to a neighboring town's library (large $\boldsymbol{h}$), hence the dependence of $\tau_{i,switch}$ on $\boldsymbol{h}$. Since the size of the switching effect depends on a unit's exposure level, we will consider an average switching effect at each value of exposure,

$$\tau_{\text{switch}}(\tilde{\boldsymbol{h}}) \equiv \mathbb{E}\Big[Y_{i1}\big(1, \tilde{\boldsymbol{h}}\big) - Y_{i1}\big(0, \tilde{\boldsymbol{h}}\big) \mid D_i = 1, h_i(\boldsymbol{D}) = \tilde{\boldsymbol{h}}\Big].$$

The second policy-relevant parameter is the 'total effect':

$$\tau_{i,\text{total}} \equiv Y_{i1}(1, h_i(\boldsymbol{D})) - Y_{i1}(0, \mathbf{0}).$$

[7] See Sävje et al. (2021) and Pollmann (2020) for discussion of estimation and inference when the treatment design is known.

[8] This is what Sävje et al. (2021) call the 'assignment-conditional unit-level treatment effect'.

As opposed to the switching effect, which keeps exposure constant, the total effect looks at turning on treatment *and* going from zero exposure to $h_i(\boldsymbol{D})$ simultaneously. This can be thought of as going from the world with a *complete* absence of treatment, $\mathbf{0}$ to the current treatment vector $\boldsymbol{D}$. Individual effects can be averaged over treated units to the 'total effect of treatment on the treated':

$$\tau_{\text{total}} \equiv \mathbb{E}[Y_{i1}(1, h_i(\boldsymbol{D})) - Y_{i1}(0, \mathbf{0}) \mid D_i = 1].$$

This treatment effect is helpful for 'national' policymakers to evaluate what *were* the effects of the entire vector of *enacted policies*.[9]

We formalize the 'spillover effect' as the difference in potential outcomes between being exposed to the observed spillover exposure and not being exposed:

$$\tau_{i,\text{spill}}(d) \equiv Y_{i1}(d, h_i(\boldsymbol{D})) - Y_{i1}(d, \mathbf{0}).$$

This effect can differ based on treatment status as the magnitude or even the causal mechanisms of spillovers might differ between treated and untreated units. For example, consider a targeted place-based policy that creates tax incentives to create businesses in designated census tracts. Nearby untreated census tracts could lose out on new business formation from this policy while nearby treated census tracts might benefit from agglomeration forces from a cluster of designated tracts.

It is often also of interest to estimate the average spillover effects on subsets of units (e.g. all units experiencing non-zero exposure). Since it shows up in our below decomposition of the difference-in-differences estimand, we define the average spillover onto all the control units as

$$\tau_{\text{spill}}(0) = \mathbb{E}[Y_{i1}(d, h_i(\boldsymbol{D})) - Y_{i1}(d, \mathbf{0}) \mid D_i = 0].$$

Estimation of spillover effects for different groups of units will be discussed in section 2.2.3.3.

[9] Yu et al. (2022) label the 'total effect' to be $\mathbb{E}[Y_{i1}(1, h_i(\boldsymbol{D})) - Y_{i1}(0, \mathbf{0})]$ where the average is over all units including control units. Identification of this effect relies on assumptions about experimental-design so this paper does not pursuit identification of this effect.

2.2.2 What Does Difference-in-Differences Estimate?

With treatment effects properly defined, I will first derive what the standard difference-in-differences estimand identifies under a modified parallel trends assumption.

Assumption 2.1 (Parallel Counterfactual Trends)**.** *Counterfactual trends do not depend on D_i:*

$$\mathbb{E}[Y_{i1}(0,\mathbf{0}) - Y_{i0}(0,\mathbf{0}) \mid D_i = 1] = \mathbb{E}[Y_{i1}(0,\mathbf{0}) - Y_{i0}(0,\mathbf{0}) \mid D_i = 0]$$

This assumption states that in the absence of treatment and with zero exposure (not just the absence of individual i's treatment), the change in potential outcomes from period 0 to 1 do not depend on treatment status. When the stable-unit treatment value assumption (SUTVA) is satisfied, all units have zero exposure and this assumption generalizes to the classic parallel counterfactual trends assumption. Second, I make the standard assumption 'no anticipation' assumption that units do not adjust their actions in period 0 from knowledge of future treatment:

Assumption 2.2 (No Anticipation)**.** $\mathbb{E}[Y_{i0}(D, \boldsymbol{h}(\boldsymbol{D}))] = \mathbb{E}[Y_{i0}(0,\mathbf{0})]$.

With these assumptions, we derive what the standard difference-in-difference estimand identifies.

Proposition 2.1 (Decomposition of Difference-in-Differences Estimand)**.**
If assumptions 2.1 and 2.2, the population difference-in-differences estimand can be decomposed as follows:

$$\mathbb{E}[Y_{i1} - Y_{i0} \mid D_i = 1] - \mathbb{E}[Y_{i1} - Y_{i0} \mid D_i = 0] = \tau_{total} - \tau_{spill}(0) \tag{2.1}$$

The proof is given in Appendix B.1, but the intuition is as follows. The change in outcomes among control units includes both the parallel counterfactual trend and the average spillover effect onto control units. Since $\hat{\tau}$ is found by subtracting the change in outcomes among the control units, we subtract the average spillover effect onto the control, $\tau_{\text{spill}}(0)$.

2.2.3 Identification of Treatment Effects

2.2.3.1 Switching Effect

It is difficult to identify the switching effect in general as identifying the unobserved counterfactual outcome $Y_{i1}(0, h_i(\boldsymbol{D}))$ requires knowledge of which control units have the same level of exposure. The reason for this can be seen by rewriting this in terms of a difference-in-differences style estimand:

$$\begin{aligned}\tau_{\text{switch}}(\tilde{\boldsymbol{h}}) = \; & \mathbb{E}\Big[Y_{i1}(1, \tilde{\boldsymbol{h}}) - Y_{i0}(0, \boldsymbol{0}) \mid D_i = 1, \boldsymbol{h}(\boldsymbol{D}) = \tilde{\boldsymbol{h}}\Big] \\ & - \mathbb{E}\Big[Y_{i1}(0, \tilde{\boldsymbol{h}}) - Y_{i0}(0, \boldsymbol{0}) \mid D_i = 1, \boldsymbol{h}(\boldsymbol{D}) = \tilde{\boldsymbol{h}}\Big].\end{aligned}$$

The first term is identified by the observed outcomes of the treated units. The second term would typically be estimated using the control units with exposure $\tilde{\boldsymbol{h}}$ under a parallel trend assumption. This is difficult as it requires knowledge of the unobserved exposure mapping to identify control units with $h_i(\boldsymbol{D}) = \tilde{\boldsymbol{h}}$. In related work, Xu (2023) take the approach of parameterizing the exposure mapping and discusses doubly-robust estimates of the switching effect.

Since spillover effects show up in the second term, estimation of the average switching effect requires an additional assumption that spillover effects on the control units be on average the same as the treated units. As an example, this assumption could fail if units that would receive the largest negative spillover effects select into treatment to avoid them. The problem of effects showing up in the above decomposition is similar to those highlighted by Callaway et al. (2021) for difference-in-differences with continuous treatment. See their discussion on 'strong parallel trends'

Remark 2.1. *Researchers may be tempted to introduce the parameterized exposure mapping interacted with a post dummy variable linearly in a two-way fixed effects model to "compare individuals at the same level $\boldsymbol{h}$". However, this assumes additive separability between the treatment dummy and the exposure mapping imposing homogeneity of switching effects for treated and untreated units. If the exposure mapping is continuous, then this also imposes switching effect grows linearly with $\boldsymbol{h}$. An alternative would be running difference-in-differences on subsets of treated and control units with* *the same value of* $\boldsymbol{h}$.

2.2.3.2 Total Effect

To identify the total effect, additional assumptions on the nature of spillover effects are needed. In particular, I will formalize the idea that spillovers are 'local' in that units are only affected by treatment if they are near treatment. For example, if treatment has to be accessed in person (e.g. access to abortion clinics), then it is natural that further away places are not affected by treatment. However, this assumption may fail under general equilibrium shocks that do not necessarily decay over distance.

To formalize this assumption, let $d(i,j)$ be a function that measures the distance between units i and j.[10] For a given distance $\bar{d}$, let $S_i(\bar{d}) = \min_{j:\ D_j=1} d(i,j) \leq \bar{d}$ be an indicator equal to one if unit i is within $\bar{d}$ miles of the closest treated unit.

Assumption 2.3 (Spillovers Are Local)**.** *There exists a distance* $\bar{d}$ *such that*

(i) For all units i*,*

$$S_i(\bar{d}) = 0 \implies h_i(\boldsymbol{D}) = \boldsymbol{0}.$$

(ii) There exists treated and control units with $S_i(\bar{d}) = 0$.

Part (i) of assumption (2.3) requires that spillovers are 'local' in that units are no longer exposed to spillovers after some maximum distance $\bar{d}$. Part (ii) of Assumption (2.3) requires that there exists control and treated units with no exposure. This assumption is far less strict than the identifying assumption for the switching effect in that we only need to identify which units have non-zero exposure and don't need to parameterize the exposure mapping any further.

The intuition behind identifying the total effect requires that we use control units experiencing no spillover effects to identify the counterfactual trend. However, this changes the necessary parallel trends assumption:

[10] While the prose generally focuses on geographic distance, this could be any reasonable metric. For example, a measure of the connectedness between businesses where we assume larger values are more distant firms.

Assumption 2.4 (Total Effect Parallel Trends)**.** *For a given $\bar{d}$, counterfactual trends are equal for the treated group and the far-away control group:*

$$\mathbb{E}[Y_{i1}(0, \mathbf{0}) - Y_{i0}(0, \mathbf{0}) \mid D_i = 1] = \mathbb{E}\big[Y_{i1}(0, \mathbf{0}) - Y_{i0}(0, \mathbf{0}) \mid D_i = 0, S_i(\bar{d}) = 0\big]$$

With this assumption, we have the following identification result for the total effects:

Proposition 2.2 (Identification of Total Effects)**.** *Suppose that assumptions 2.2, 2.3, and 2.4 hold. Then,*

$$\tau_{total} = \mathbb{E}[Y_{i1} - Y_{i0} \mid D_i = 1] - \mathbb{E}[Y_{i1} - Y_{i0} \mid D_i = 0, S_i = 0]. \tag{2.2}$$

Proofs are given in Appendix B.1. This proposition shows that if a researcher believes that spillover effects are 'local', then estimation of the total effect on treated units only requires identifying which units are exposed to spillovers. The identifying assumption is that the control units are a valid comparison group *after* removing the (potentially) exposed units. To estimate the total effect, we could replace the terms in proposition 2.2 with their sample analogs.

Remark 2.2 (Far-away control units and parallel trends)**.** *It is worth remarking on Total Effect Parallel Trends in terms of applied work. Often, researchers use only a subsample of observations that are close to treated observations. The idea is that if unobservable confounders evolve smoothly over space, then close units are more likely to be on similar counterfactual trends than units that are further away. The identification result and its corresponding estimator are formed by using further away units (i.e. units with $S_i = 0$). In this case, if parallel trends only hold for nearby control units, then focusing on further away units can fix the spillover bias but introduce bias from non-parallel trends. See appendix section 2.3.2 for further discussion and an empirical example evaluating the 2017 U.S. Opportunity Zones using different identification strategies.*

Remark 2.3 (Selection of $\bar{d}$)**.** *A natural question is why wouldn't researchers make the distance large to guarantee they have an unbiased estimate of τ_{total}? Equation (2.2) should make this problem clear. Since these estimates rely on averages among units with $S_i = 0$, as $\bar{d}$ increases, the number of units with $S_i = 0$ decreases which yields more variable estimates. On the other hand, having units* *that experience spillovers not included in the S_i indicator will leave some bias in the estimate.*[11] *Therefore there is a bias-variance trade-off in extending the extent of S_i that should be balanced by researchers guided by their particular economic context. Additionally, each value of $\bar{d}$ corresponds to a different effective control group and hence a different parallel trends assumption, rendering comparisons across $\bar{d}$ uninformative.*

2.2.3.3 Spillover Effects

This section turns to identification of averages of spillover effects. To do so, it is typical to compare nearby control units to further away units. This requires an additional parallel trends assumption that the far-away control units and the nearby control units are on the same trends:

Assumption 2.5 (Spillover Effect Parallel Trends)**.** *For a given $\bar{d}$, we have:*

$$\mathbb{E}\big[Y_{i1}(0,\mathbf{0}) - Y_{i0}(0,\mathbf{0}) \mid D_i = 0, S_i(\bar{d}) = 1\big] = \mathbb{E}\big[Y_{i1}(0,\mathbf{0}) - Y_{i0}(0,\mathbf{0}) \mid D_i = 0, S_i(\bar{d}) = 0\big].$$

In this case, comparing control units with $S_i(\bar{d}) = 1$ to far-away control units will identify the average spillover effect on control units with $S_i(\bar{d}) = 1$:

$$\mathbb{E}\big[Y_{i1}(d, h_i(\boldsymbol{D})) - Y_{i1}(d,\mathbf{0}) \mid S_i(\bar{d}) = 1, D_i = 0\big]$$

Joint estimation of the total effect on the treated and the spillover effect on the nearby control units can be done using the following regression specification on the full sample:

$$Y_{it} = \tau D_i \, 1_{t=1} + \gamma_0 (1 - D_i) S_i(\bar{d}) \, 1_{t=1} + \mu_i + \lambda_t + \varepsilon_{it}. \tag{2.3}$$

Note that since some units with $S_i(\bar{d}) = 1$ may not have positive exposure ($\boldsymbol{h}_i(\vec{D}) = \mathbf{0}$), this is not the average spillover effect for units with non-zero exposure. For example, if only units really close to treatment receive spillover effects and $S_i(\bar{d})$ contains a lot of units that experience no spillover effects, then $\hat{\tau}_{\text{spill, control}}$ could be estimated near zero even though some units experience substantial spillover effects. For example, reanalysis of Linden and Rockoff (2008a) in Butts (2023)

[11] Although, spillover effects typically will grow weaker over distance and if distant units are mistakenly treated as if they have zero exposure, bias should be small.

finds that the arrival of sex offenders to a neighborhood has large impacts on homes within 0.05 miles and much smaller impacts for homes between 0.05 and 0.10 miles. Selecting $\bar{d}$ to be 0.1 estimated an effect that was about half as large as the estimated impact on the very close homes.

To better understand how spillover effects vary over distances, applied researchers will sometimes split up $S_i(\bar{d})$ into a set of J concentric 'rings' (e.g 0-20 miles, 20-40 miles, and 40-60 miles from the nearest treated unit). Equation (2.3) can be modified as

$$y_{it} = \tau D_i \, 1_{t=1} + \sum_{j=1}^{J} \gamma_j (1 - D_i) \text{Ring}_i^j \, 1_{t=1} + \mu_i + \lambda_t + \varepsilon_{it}, \tag{2.4}$$

where Ring_i^j is an indicator for a control unit i being in ring j with $\sum_j \text{Ring}_i^j = S_i(\bar{d})$. The estimate $\hat{\tau}$ is still consistent for the total effect since the effective control group remains the same. However, interpretation of each $\hat{\gamma}_j$ as causal requires a yet stronger parallel trends assumption. Namely, units in each ring j must be on the same parallel counterfactual trend as the $S_i(\bar{d}) = 0$ control group. If only assumption 2.5 is assumed, it is not possible separate variation in spillover effects from variation in counterfactual trends across rings. Butts (2023) shows that in some settings where you are willing to assume that parallel trends holds for all distances from treatment, a non-parametric estimator is available for average spillover effects as a function of distance. This assumption is more likely in very local settings where units do not vary in terms of temporal shocks.

2.3 Applications in Place-Based Policy Analysis

More generally, my framework provides important insights into identification strategies when analyzing the effects of place-based policies. There are two ways I contribute. First, Baum-Snow and Ferreira (2015) recognize the problem of spatial spillovers causing problems in identification and point to aggregation of units as a way to alleviate the problem (e.g. aggregating census tracts to metropolitan areas). However, this approach combines the treatment and spillover effects, which might each be of interest, into a singular aggregate effect. This method averages over heterogeneous effects (in magnitude or in nature) that might be of independent interest. The methods propose in this paper provide a strategy for disentangling treatment effect estimates with non-aggregated data.

Subsection 2.3.1 will revisit the analysis of a large-scale targeted policy on regional development. In this setting, spillover effects on nearby counties, via lost jobs, are of a very different nature than the direct impact on the targeted region and hence separate estimation if of practical interest.

Second, my framework provides insight into different identification strategies often used in the literature. Since place-based policies are targeted to specific distressed areas, comparison units are often hard to find. Researchers have developed many identification strategies to find comparison units that are similar in terms of unobservables. First, researchers use border discontinuities to compare treated units to units just on the other side of the border. Other times, researchers compare approved applicants to narrowly rejected applicants. Both strategies aim to balance unobservables between treated and control units. A key difference in these identification strategies is the distance comparison units are from the treated area and therefore the former is more prone to bias due to spillovers than the latter. Subsection 2.3.2 tries to highlight the differences in these identification strategies in the context of the U.S. Opportunity Zones created in 2017.

2.3.1 The Tennessee Valley Authority

To illustrate the importance of accounting for spatial spillovers in the estimation of treatment effects, I revisit the analysis of the Tennessee Valley Authority (TVA) in Kline and Moretti (2014). The TVA program was a large-scale federal investment started in 1934 that focused on the construction of dams and transportation canals in an attempt to modernize the Tennessee Valley's economy. By the end of WWII, the TVA became the largest single power supplier in the country and significantly lowered the cost of wholesale energy for manufacturers.[12] With over $20 Billion (in 2000 dollars) spent which is hundreds of dollars transferred per person in the Authority, the impacts are very likely to extend past the authority's borders. Kline and Moretti (2014) analyze a range of outcome variables, but I focus on agricultural and manufacturing employment.

The analysis in Kline and Moretti (2014) begins by comparing changes in county-level outcomes

[12] More details on the program are found in Kline and Moretti (2014). The effects on wholesale electricity are discussed in Kitchens (2014).

from 1940 to either 1960 (short run effects) or 2000 (long run effects) between treated counties in the Authority and control counties outside. The primary specification is

$$y_{c,t} - y_{c,1940} = \alpha + \text{TVA}_c \tau + X_{c,1940}\beta + (\varepsilon_{c,t} - \varepsilon_{c,1940}), \tag{2.5}$$

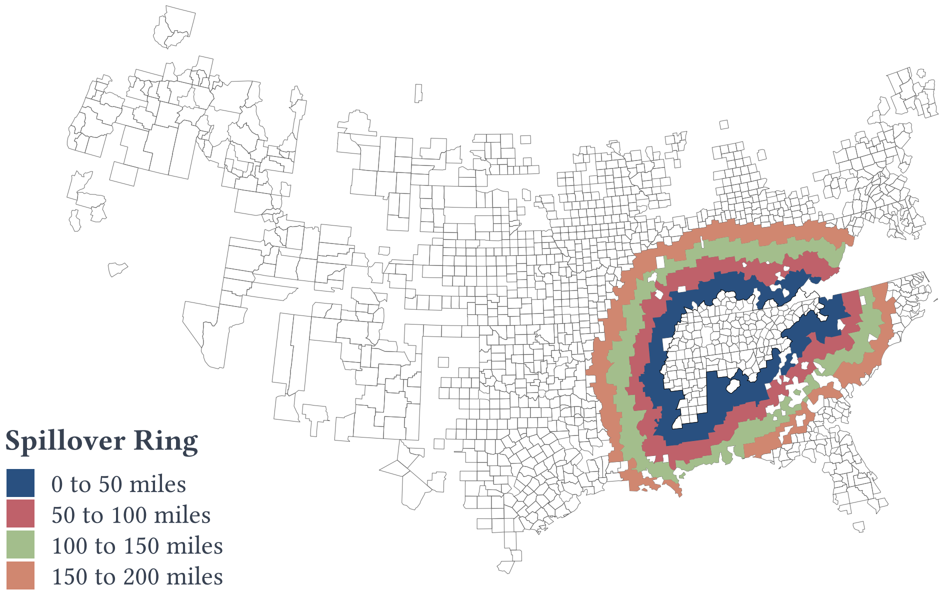
where c denotes county, TVA_c is an indicator variable for being in the Authority, and y is a set of outcome variables (in logs).[13] Pre-treatment control variables, $X_{c,1940}$, are included to allow for places to be on different long-term trends.[14] To improve the likelihood of the parallel trends assumption, they run a logistic regression to predict being in the TVA based on their set of control variables $X_{c,1940}$ and keep only observations in the top 75% of predicted probability. The subsample produced from this exercise is presented in Figure 2.1.

In the paper, the authors discuss the nature of spillovers that they think can occur. For agriculture employment, the authors argue that improved wages in the TVA will draw agriculture workers out of nearby counties producing negative spillovers. For manufacturing, there are two countervailing forces (Cuberes et al., 2021). Positive spillovers from 'urban access' would occur if electrification brought cheap power and agglomeration economies to the neighboring areas. The countervailing 'urban shadow' would cause negative spillovers if firms chose to locate in the Authority that would have, in the absence of the program, decided to locate in nearby counties. My methodology will allow me to empirically test for these two forces in the data over time while removing their bias from the total effect estimates.

Kline and Moretti (2014) estimate (2.5) to identify what I am calling the 'total effect on the treated'. However, their point estimates compare, in part, changes in outcomes in TVA counties with changes in outcomes for neighboring counties that likely were impacted by the large-scale program. The authors do recognize the problem of spillover effects and the majority of the paper uses structural models to estimate the general equilibrium effects of the TVA. In this light, the re-analysis here is complementary to their work.

[13] The two-period difference-in-differences regression is equivalent to a first-difference regression. The authors use an Oaxaca-Blinder estimator on the first differences and the results of Kline (2011) show that this estimator is equivalent to a weighted difference-in-differences estimate. Their estimator does not differ much from the standard

Figure 2.1: TVA Effective Sample and Spillover Variables



Notes: The above figure plots all the counties used in the estimation. Counties that fall within the distance intervals $\{(0,50],(50,100],(100,150],(150,200]\}$ measured in miles are colored by their respective bin.

I extend their analysis to control for spatial spillovers in the difference-in-differences specification. To parameterize the exposure mapping, I use a set of rings as described in section 2.2.3.3. Specifically, the specification with spillovers is given as follows:

$$y_{i,t} - y_{i,1940} = \alpha + \text{TVA}_i \tau + \sum_{j \in \text{Dist}} \text{Ring}_i^j \gamma_j + X_{i,1940}\beta + (\varepsilon_{i,t} - \varepsilon_{i,1940}), \tag{2.6}$$

where $\text{Dist} = \{(0,50],(50,100],(100,150],(150,200]\}$ measured in miles and define Ring_i^j as an indicator for being within the interval $d \in \text{Dist}$ away from the Authority and $t \in \{1960, 2000\}$. Figure 2.1 displays the four spillover variables by filling in each distance bin in a different color. The coefficients γ_j estimate the average spillover effect onto control units for each of these distance bins. It's important to note that estimation of (2.6) changes the comparison group to counties further away than 200 miles from the TVA.

difference-in-differences results since the weights are not that different from uniform weights.

[14] See footnote 8 in Kline and Moretti (2014) for a full listing of control variables.

Table 2.1: Effects of Tennessee Valley Authority on Decadal Growth

	Diff-in-Diff	**Diff-in-Diff with Spillovers**				
			TVA between	TVA between	TVA between	TVA between
	TVA	TVA	0-50 mi.	50-100 mi.	100-150 mi.	150-200 mi.
Dependent Var.	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: 1940-2000						
Agricultural employment	-0.0514^{***}	-0.0739^{***}	-0.0371^{**}	-0.0164	-0.0298^{***}	-0.0157^{*}
	(0.0135)	(0.0163)	(0.0149)	(0.0114)	(0.0104)	(0.0095)
Manufacturing employment	0.0560^{***}	0.0350	-0.0203	-0.0245	-0.0331	-0.0296^{*}
	(0.0196)	(0.0267)	(0.0274)	(0.0338)	(0.0227)	(0.0166)
Panel B: 1940-1960						
Agricultural employment	0.0940^{***}	0.0856^{*}	-0.0062	-0.0042	-0.0303	-0.0039
	(0.0309)	(0.0473)	(0.0507)	(0.0487)	(0.0471)	(0.0339)
Manufacturing employment	0.0894^{**}	0.0993^{**}	0.0228	0.0225	-0.0055	-0.0066
	(0.0348)	(0.0473)	(0.0554)	(0.0630)	(0.0399)	(0.0292)

Notes. Each row corresponds to an outcome variable. Each cell is the point estimate and the standard error for the variable described in the column title. All standard errors are Conley standard errors with a correlation cutoff of 200 miles following Conley (1999). The column labeled 'Diff-in-Diff' estimates (2.5) by OLS and is similar to the estimate reported in Kline and Moretti (2014). The final four columns labeled 'Diff-in-Diff with Spillovers' are estimates from (2.6). $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

The results of the long run analysis from 1940 to 2000 are presented in Panel A of Table 2.1. Each row contains the results for a different outcome variable (measured in logs). Columns (1)-(5) contain point estimates for τ and γ_j's in different specifications. The point estimates can be interpreted as decadal growth rates in outcomes. The column labeled difference-in-differences estimates equation (2.5). This estimate finds a decline in agricultural employment of about 5.1% per decade and an increase in manufacturing employment of about 5.6% per decade.

Turning to specification (2.6) which includes spillovers, column (2) contains a point estimate for τ and columns (3)-(6) contain point estimates of the spillover effects γ_j. For agricultural employment, the point estimates show there was a decline in agriculture employment in control units near the Authority. For control units between 0 and 50 miles, column (4) indicates a decline in agricultural employment of 3.7% per decade. Between 50 and 100 miles the point estimate is -1.6% per decade, between 100 and 150 miles the point estimate is -3% per decade, and between 150 and 2000 miles the point estimate is -1.6% per decade. This is likely because higher-paying manufacturing jobs within the Authority drew farm-worker migrants from nearby counties. Because the spillovers onto the control counties are negative, the original difference-in-differences estimator was positively biased. The new point estimate indicates a decline of agricultural employment of about 7.4% per decade compared to 5.1% in the standard difference-in-difference specification.

For manufacturing, our point estimates for spillovers are consistently negative though imprecisely estimated. The spillover estimates suggest that neighboring counties potentially experienced negative spillover effects in the long run. Since there are negative spillover effects present, the new point estimate in column (2) of 3.5% is smaller than the original estimate of 5.6%. The spillover estimates are evidence that in the long run, 'urban shadow' (stealing manufacturing firms) forces dominate the benefits of 'urban access' (agglomeration effects).

To see how spillover effects from a large-scale place-based policy develop over time, Panel B in Table 2.1 presents results for the effects of the Tennessee Valley in the short run using outcome data in 1960. Unlike in the long run, areas near the Tennessee Valley did not experience significant declines in agricultural employment in the short run. Since our long run analysis finds significant increases in high-paying manufacturing employment in the Tennessee Valley, this result is consistent with long run migration costs being lower than short run costs.

For manufacturing, there are potentially positive increases in manufacturing employment within 100 miles of the Tennessee Valley Authority and near-zero effects between 100 and 200 miles. In the short run, it appears that the effects of urban access and the cheap wholesale electricity dominated the effects of urban shadow. The effect of urban shadows can potentially be smaller in the short run if operating firms are unlikely to relocate. long run effects can be larger as entrant firms change their location decision and operating firms are slowly replaced which is what the long run spillover effect estimates suggest.

These results show that including spillovers in the estimation of treatment effects is potentially important and can lead to *significant* differences in treatment effect estimates. Analysis of place-based policies that do not account for the fact that treatment effects can spill beyond the borders of treated areas can potentially be biased. More, the results suggest that the spillover effects caused by place-based policies change over time as frictions can create delays in re-optimizing behavior.

2.3.2 United States Opportunity Zones

In this subsection, I revisit the analysis of Chen et al. (2023) on the 2017 Opportunity Zone program which created tax incentives for capital investment in targeted Census Tracts.[15] The impacts of these kinds of programs is contentious in the literature. For example, there are a set of conflicting results on the impacts of the Empowerment Zones in the late 1990s with some papers suggesting that the Empowerment Zones reduced poverty rates while others finding near-zero effects.[16] The analysis of the 2017 Opportunity Zones creates a nice setting to consider the differences between the strategies since which tracts were eligible but ultimately were not selected is public information.

To measure the affects of the opportunity zone program on housing prices, the authors collect a panel of census tracts with a measure of housing prices from the Federal Housing Finance Agency (FHFA) from 2014-2019.[17] They produce estimates using two different identification strategies. First, they compare census tracts that were selected as Opportunity Zones to eligible, but ultimately not selected, census tracts. This estimation strategy relies on the assumption that since these census tracts are similar in nature to the treated census tracts (both meeting the program's criteria), it is likely that home prices would continue on similar trajectories in the absence of the program. The authors run a standard two-way fixed effect specification on the subsample of eligible census tracts:

$$Y_{it} = \mu_i + \mu_t + \tau d_{it} + \varepsilon_{it}, \tag{2.7}$$

where Y_{it} is the annual change in the home price index, μ_i are tract fixed-effects, μ_t are time fixed effects, and d_{it} is an indicator for treatment.

[15] Concurrent work by Arefeva et al. (2021) also estimates spillover to nearby opportunity zones using a different dataset and finds positive spillover effects on nearby census tracts as well.

[16] See Table 1 of Neumark and Young (2019) for a summary of the various treatment effect estimates in the literature. Busso et al. (2013) compare census tracts in Empowerment Zones to census tracts that qualified and were rejected from the program. The rejected tracts are not typically geographically near accepted Empowerment Zones and they find large significant reductions in poverty rates. Meanwhile, Neumark and Kolko (2010) compare census tracts in Empowerment Zones to census tracts within 1,000 feet of the Zone. These control counties are likely the ones that experience the largest spillover effects and they find near-zero effects on poverty. My paper suggests that the former is the preferred strategy in the presence of significant spillovers onto nearby control units.

[17] The index tries to create a consistent price index that captures for changes in the composition of homes that are sold over time.

The second identification strategy relies on comparing selected census tracts to geographically neighboring census tracts. This estimation strategy relies on the assumption that proximity of census tracts would face similar economic shocks and hence home prices would likely evolve in parallel. For each treated census tract, they find the nearest non-treated census tract to form a pair, $(i, \tilde{i}) \equiv \nu$. They then estimate the following equation:

$$Y_{it} - Y_{\tilde{i}t} = \mu_\nu + \tau d_{it} + u_{it}, \tag{2.8}$$

where Y_{it} is the annual change in the home price index, μ_ν are pair fixed-effects and d_{it} is an indicator for treatment.

A valid concern is that not-selected tracts were not selected because they were viewed as having better economic prospects than selected tracts, implying that a parallel trends assumption like in 2.4 is unlikely to hold. The authors provide evidence of parallel 'pre-trends' for both the nearby tracts and the eligible but not-selected tracts, alleviating concerns in this context. The results of both estimation strategies are shown in Table 2.2. Column (1) and (2) show the results of estimating Equation (2.7) and (2.8) respectively. The 'not-selected' estimate finds a marginally significant effect of an increase in home prices of about 0.3% annually while the 'neighboring' estimate finds a strongly significant effect twice as large of about 0.65%.[18]

What is driving the differences in these estimates? As I proposed above, the differences could be due to the fact that neighboring units could be experiencing effects from the Opportunity Zones. To test this, I use the estimation strategy proposed in subsection 2.2.3 to modify (2.7). For the subsample of eligible census tracts, I run the following specification:

$$Y_{it} = \mu_i + \mu_t + \tau d_{it} + \gamma_1 \text{Within 1/2mi.} + \gamma_2 \text{Within 1mi.} + \varepsilon_{it}, \tag{2.9}$$

where Within are indicators for being within 1/2mi. and being between 1/2 and 1mi. from an Opportunity Zone. This estimation strategy uses non-neighboring census tracts as the effective control group and estimates effects for census tracts within or close to Opportunity Zones. The

[18] The authors use this estimate to rule out effects larger than $\approx 0.65 + 2 * 0.25 = 1.15\%$. As shown before, these estimates are biased upwards and hence the upper bound of effect size can be lowered to about $\approx 0.65\%$.

Table 2.2: Effects of Opportunity Zones on Annual Home Price Growth

	(1)	(2)	(3)
Treat × Post	0.3033^{*}	0.6478^{***}	0.1788
	(0.1661)	(0.2457)	(0.1692)
< 1/2mi. × Post			-1.057^{***}
			(0.3618)
< 1mi. × Post			-0.7430^{***}
			(0.1922)
Control Group:	Not-Selected	Neighboring	Not-Selected

Notes This table contains estimates of models (2.7), (2.8), and (2.9) using the sample from Chen et al. (2023). $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

results are estimated in Table 2.2 in Column (3). These estimated treatment effect decreased in this specification to 0.17% and reveals that census tracts just outside of the Opportunity Zone experience negative and significant declines in home prices of about 1%. This result explains why the estimated effect using the 'neighbor' specification is twice as large as the 'eligible' specification.

2.4 Estimating Event Study Specifications with Spillovers

Now, I turn to the staggered treatment adoption setting where treatment turns on at different times for different units. The intuition from how spillovers cause biases in estimates of the total effect of treatment on the treated extends into the staggered adoption setting. In this setting, the literature has shown that two-way fixed-effect estimates can be viewed as a weighted sum of 2 × 2 difference-in-differences estimates.[19] Therefore the bias terms will be an identically weighted sum of the bias term(s) from the 2 × 2 estimates, assuming that the parallel counterfactual trends assumption (2.1) holds for all groups. However, since weights on some of the 2 × 2 estimates can be negative, the sign of the spillover effects does not determine the sign of the weighted average of the spillover effects. This makes the bias from spillovers much more difficult to sign.

[19] Various forms for these weights are described in Goodman-Bacon (2021), Sun and Abraham (2021), and De Chaisemartin and d'Haultfoeuille (2020). I do not re-characterize the weights in this article and guide interested readers to the source articles themselves.

To estimate treatment and spillover effects in the presence of spillovers and staggered treatment timing, I will propose an estimation strategy that follows the 'imputation-based' approach proposed in concurrent work by Gardner (2022) and Borusyak et al. (2024) by incorporating spillovers directly into estimation.[20] The imputation-based method relies on a model-based assumption for untreated and unexposed potential outcomes to formalize parallel trends:

Assumption 2.6 (Parallel Counterfactual Trends (Staggered))**.** *For all units and all periods, the untreated and unexposed outcome is given by*

$$Y_{it}(0, \mathbf{0}) = \mu_i + \lambda_t + \varepsilon_{it}, \tag{2.10}$$

where $\mathbb{E}[\varepsilon_{it}] = 0$*.*

This formalizes parallel trend by imposing that there is a common time-trend in the absence of treatment, λ_t. Note that assumption 2.6 does not impose any structure on the effects of treatment and exposure $(d_{it}, h_i(\boldsymbol{d}_t))$ where now the exposure mapping is a function of the current period's treatment vector $\boldsymbol{d}_t \equiv \{d_{1t}, \dots, d_{nt}\}$. I also assume that units do not have any anticipatory effects before treatment starts:

Assumption 2.7 (No Anticipation (Staggered))**.** *For all observations with* $d_{it} = 0$ *and* $h_i(\boldsymbol{d}_t) = 0$*,* $Y_{it} = Y_{it}(0, \mathbf{0})$*.*

With a model for $Y_{it}(0, \mathbf{0})$, stochastic treatment effects for individual i and time t under treatment status d_{it} and exposure $h_i(\boldsymbol{D})$ are given by $\tau_{it}(d_{it}, h_i(\boldsymbol{D})) = Y_{it}(d_{it}, h_i(\boldsymbol{D})) - Y_{it}(0, \mathbf{0})$. Then, the total effect is formed similar to above:

$$\tau_{\text{total}} \equiv \mathbb{E}[\tau_{it}(1, h_i(\boldsymbol{d}_t)) \mid d_{it} = 1],$$

averaging over all the post-treatment observations. It is also common in event-study analyses to allow heterogeneity in effects by the number of years that a unit experiences treatment. Let K_{it}

[20] Alternatively, we could propose an estimator similar to Callaway and Sant'Anna (2020) using far-away control units (at time t) as the comparison units. One advantage of this is that covariates can be included flexibly to allow for parallel-trends to hold conditionally on a set of covariates.

denote the number of years since treatment turns on (-1 for the year prior, 0 for the initial year, and so on). Then we can define 'relative year' dummy variables $d_{it}^k \equiv D_i \, 1_{K_{it}=k}$ and estimate average total effects for each relative period:

$$\tau_{\text{total}}^k \equiv \mathbb{E}\Big[\tau_{it}(1, h_i(\boldsymbol{d}_t)) \mid d_{it}^k = 1\Big].$$

Our identification argument follows from noting that under Parallel Counterfactual Trends (Staggered),

$$\mathbb{E}[Y_{it}(d_{it}, h_i(\boldsymbol{d}_t)) - \mu_i - \lambda_t \mid \Omega] = \mathbb{E}[\tau_{it}(d_{it}, h_i(\boldsymbol{d}_t)) \mid \Omega],$$

where Ω is a set of post-treatment observations (i, t). Estimation of τ_{it} is the difference between the observed Y_{it} and the unobserved outcome $Y_{it}(0, \mathbf{0})$. Similar to the previous section, under assumption (2.3) we define s_{it} to be an indicator equal to one if unit i is within $\bar{d}$ distance from the nearest treated unit from the set of units treated in period t.

Following the procedure of Gardner (2022); Borusyak et al. (2024), I propose a modified version of the two-stage difference-in-differences estimator:

(1) Estimate $Y_{it} = \mu_i + \lambda_t + u_{it}$ for observations with $d_{it} = 0$ and $s_{it} = 0$ to estimate the common trend λ_t and unit fixed-effects μ_i. For all observations, residualize $\tilde{Y}_{it} \equiv Y_{it} - \hat{\mu}_i - \hat{\lambda}_t$ which is our estimate for τ_{it}.

(2) Regress $\tilde{Y}_{it}$ on treatment dummies (d_{it} or d_{it}^k) and possibly spillover dummies (s_{it} or s_{it}^k).

The first stage uses untreated and unexposed observations to estimate model (2.10). Then, averaging $\tilde{Y}_{it}$ in the second state will be unbiased and consistent for corresponding averages of τ_{it} under assumptions 2.6 and 2.7.[21] Identification is discussed in appendix section B.1.3. Note that recent work by Goldsmith-Pinkham et al. (2022) highlights problem with standard ordinary-least squares estimates with multiple mutually-exclusive treatment variables. It is important to note the imputation estimator does not face these problems.

[21] Recent work by Goldsmith-Pinkham et al. (2022) suggests that multiple mutually exclusive treatment variables can create contamination bias. This is not a concern in this procedure because the above imputation strategy avoids residualizing the treatment variables.

As discussed in Gardner (2022), inference must account for the fact that the regressand in the second stage is estimated in the first stage. Gardner (2022) propose reframing the two-stage process as a two-stage GMM estimator and discusses how to perform valid inference in this setting. This procedure is implemented in the R/Stata package `did2s` (Butts, 2021a).

2.4.1 Application on Community Health Centers

As an application of the above methods, I extend the analysis of Bailey and Goodman-Bacon (2015). The authors study the creation of federal community health centers between 1965 and 1974 that provided *primary* care to low-income communities. They test the hypothesis if access to low-/no-cost health care services decreased mortality rates on the treated population. To answer this question, the authors use a common event-study framework to compare outcomes in treated counties to all other US counties by estimating the following specification

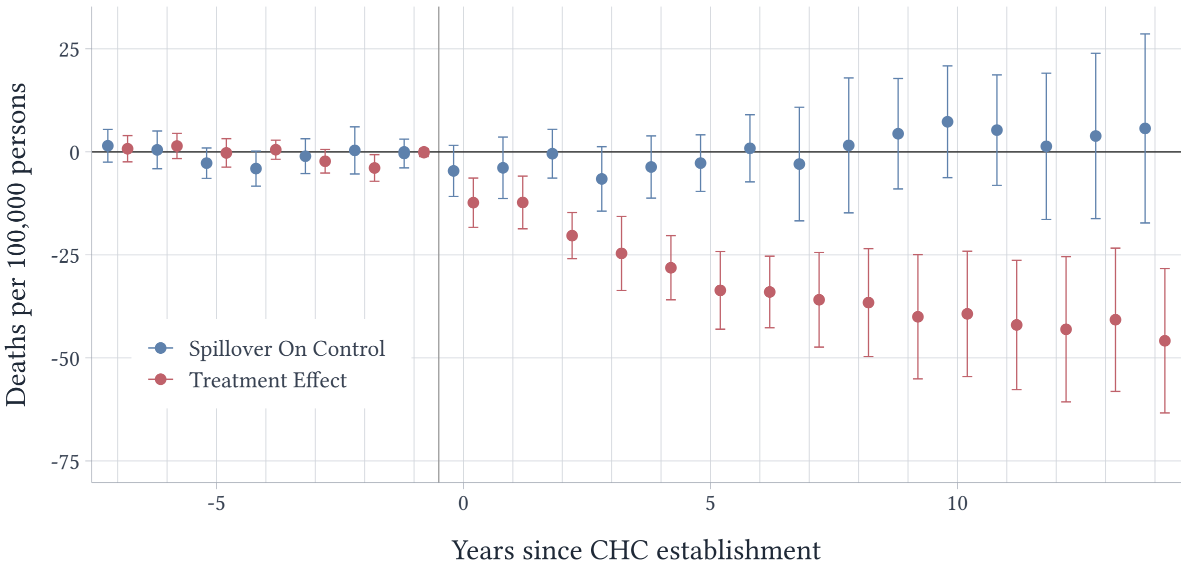
$$Y_{it} = \mu_i + X_{it}\beta + \sum_{k=-7}^{-2} \pi_y d_{it}^k + \sum_{k=0}^{15} \tau_y d_{it}^k + \varepsilon_{it}, \tag{2.11}$$

where d_{it}^k is the typical 'event study' indicator for being treated for k years, μ_i is county fixed-effects, and X_{it} contains a set of controls.[22] The coefficient's π_y can be interpreted as tests of parallel pre-trends and τ_y can be interpreted as the treatment effect of a community health center y years after establishment. In years following the establishment of the community health centers, the authors find a reduction of between 15-30 deaths per 100,000 residents compared to a baseline adjusted mortality rate of 929 deaths per 100,000 residents.

There are theoretical reasons to think spillovers may or may not exist in this context. On the one hand, individuals outside the county can potentially travel to the community health centers to receive care. This would create a negative spillover effect on mortality rates in nearby counties which would bias their estimates towards zero. On the other hand, Bailey and Goodman-Bacon (2015) document evidence that their estimated effects are not due to emergencies but rather primary care services. In this case, it is less likely low-income individuals would travel very far to receive

[22] Controls include 1960 county characteristic trends, state-year fixed-effects, and urban-group fixed-effects. A full list can be found on page 1080 of Bailey and Goodman-Bacon (2015).

Figure 2.2: Total and Spillover Effects of Community Health Centers



Notes: This figure plots event study estimates for the total effect and the spillover effect on control units within 25 miles of treatment at different periods relative to establishment year. The estimates are generated using the 'did2s' package (Butts, 2021a).

primary care, and hence the spillover effects are potentially near zero. My methodology can provide an answer to the question of how far do individuals travel for low-/no-cost primary care.

As in the method detailed above, I use an indicator for being within 25 miles of a treated county as the spillover variable. The results are presented in Figure 2.2. The confidence intervals labeled with circles represent point estimates for the average spillover effect on control units within 25 miles. No spillover effect is estimated to be significantly different from zero which suggests that the effects of community health centers are very local. Since there are near zero spillover effects, the total effect estimates marked in Figure 2.2 as diamonds maintain the same shape as the author's original estimates with estimates between 15-30 fewer deaths per 100,000 persons.

The spillover effects results provide evidence that low-income individuals will not travel far to receive primary care. Practically, this suggests that community health centers should be targeted to be as accessible as possible for poor individuals as they are unable to travel far to access the services.

2.5 Discussion

This paper has considered the common environment where treatment is assigned to groups of units while the effects of treatment spread across these borders. I model this phenomenon using a potential outcomes framework and show why standard difference-in-differences estimation does not identify treatment effects of interest. Next, I discuss identification of two well-defined treatment effects. First, the average local effect of a single unit changing their treatment status is hard to identify without imposing assumptions on the spillover mapping. The second effect, the post-hoc average total effect on the treated is more simply identified under the assumption that spillovers are local. Effects can be identified by using far-away control units, but the practical usefullness of this result depends on whether far-away control units satisfy the parallel counterfactual trends. Last, I show how to extend the estimation strategy proposed in the paper to settings with staggered treatment-timing, a recent concern in the literature.

I use a set of empirical examples to show that in settings with spatial spillovers, estimates can change significantly. In particular, place-based interventions change the nature of agglomeration in the local and surrounding area–that is cause spillovers–local effects of these policies can be misestimated without controlling for general equilibrium effects. I also show the importance of considering spillovers in weighing the pros and cons of various identification strategies. Identification strategies based on geographic continuity of unobservables can magnify the bias from spillovers as they restrict the comparison group to observations experiencing the largest spillover effects.

Chapter 3

Difference-in-Differences with Geocoded Microdata

3.1 Introduction

The rise of geocoded microdata has allowed researchers to begin answering questions about the effects of spatially-targeted treatments at a very granular level. How do local pollutants affect child health?[1] Does living within walking distance to a new bus stop improve labor market outcomes?[2] How far do neighborhood shocks, such as foreclosures or new construction spread?[3] A standard method of evaluating the effects of the treatment is to compare changes in outcomes for units that are close to treatment to those slightly further away – what I label the 'ring method' as illustrated in Figure 3.1. This paper formalizes the assumptions required for identification in the ring method, highlighting potential pitfalls of the currently-used estimator, and proposes an improved estimator.

First, I formalize the necessary assumptions for unbiased estimates of the average treatment effect on the affected units.[4] The first assumption is the well understood parallel trends assumption for the treated and control units. This requires that the *average* change in (counterfactual) untreated outcomes in the treated ring is equal to the *average* change in the control ring. The rings method is typically motivated by the fact that since the treated and control units are all very close in physical location, e.g. having access to the same labor market and consumptive amenities, shocks over time should be common across units in the neighborhood.[5]

[1] See, e.g., Currie et al. (2015) and Marcus (2021).

[2] See, e.g., Gibbons and Machin (2005) and Billings (2011).

[3] See, e.g., Asquith et al. (2021b); Cui and Walsh (2015); Gerardi et al. (2015) and Campbell et al. (2011).

[4] This generalizes the treatment effect on the treated in the case where treatment is not assigned to specific units.

[5] For example, Asquith et al. (2021b) write "The idea is that within a small area, developers have few sites that are

Instead of *fully* leveraging the common neighborhood trends assumption that motivates the ring method, researchers rely on a second assumption which requires correct identification of how far treatment effects are experienced (the inner ring). This is a very strict assumption that when not satisfied, causes biased estimates of the treatment effect. If the treated ring is too narrow, then units in the control ring experience effects of treatment and the change among 'control' units would no longer identify the counterfactual trend. On the other hand, if the treated ring is too wide, then the zero treatment effect of some unaffected units are averaged into the change among 'treated' units. Therefore, results will be attenuated towards zero.

I propose an alternative estimator that fully leverages the 'common neighborhood trends' assumption using a nonparametric partitioning-based least square estimator (Cattaneo et al., 2019a,b). My proposed methodology estimates the treatment effect curve as a function of distance by using many rings rather than trying to estimate the average treatment effect with one inner ring. This method requires that treatment effects become zero somewhere between the distance of 0 and the outer ring *without* the need to specify the *exact* distance at which it does. To do this, the estimator leverages the stronger 'common neighborhood trends' assumption that the counterfactual trend is constant across distance. This new assumption is more strict in that the standard method only requires that parallel trends holds *on average* in each ring. However, researchers typically motivate the identification strategy by saying within a small distance from treatment that units are subject to a common set of shocks which implies the more strict assumption. While this assumption is not directly testable, my proposed estimator creates a set of point estimates of treatment effects that can be used to visually inspect the plausability of the assumption. If after some distance, treatment effects become centered at zero, this suggests that common trends hold, akin to the pre-trends test in event study regressions.

The nonparametric estimator provides many benefits. First, the estimator allows the researcher to get a more complete picture of how the intervention affects units at various distances rather

available and properly zoned, leading to hyper-local variation in the location of new construction that is not related to future price changes."

than estimating an “overall effect”. For example, a new bus-stop potentially creates net costs to immediate neighbors while providing net benefits for homes slightly further away. Estimation of the treatment effect curve can illustrate these different effects that a single ring may average out to near-zero effects. Additionally, the proposed nonparametric estimator selects the rings in a *completely data-driven and optimal manner* removing incentives for specification searching or ‘pre-testing’ the data to determine the ‘correct’ rings (McCloskey and Michaillat, 2020; Andrews and Kasy, 2019).

This paper contributes to the literature in many ways. First, I contribute to a literature that discuss identification using the rings method. Sullivan (2017) and Gerardi et al. (2015) discuss the identification assumption of identifying the maximum treatment distance. My paper formalizes all the necessary assumptions for identification in the ring method. Other researchers have recognized that estimating a single average treatment effect is less informative than a treatment effect curve, instead using multiple rings to estimate treatment effects at different distances (e.g. Alexander et al. (2019); Casey et al. (2018); Di Tella and Schargrodsky (2004)). However, this approach selects multiple rings in an ad-hoc manner, requires treatment effects to become zero after the outer-most treatment ring, and is prone to problems of specification searching. The current study’s proposed estimator selects the number and location of rings in a data-driven way and does not require correct specification of where treatment effects become zero.

Diamond and McQuade (2019) propose an alternative nonparametric estimator aimed at estimating a treatment effect surface. Similar to this paper, they leverage the ”common local trends” assumption to estimate a *two-dimensional* treatment effect curve. Their procedure allows “location” fixed effects to differ along two-dimensions, latitude and longitude (e.g. the mean home price is higher in the north-west than the south-east) which in cross-sectional data can absorb more variation in the data yielding smaller standard errors than my method, though both are consistent. In the case of panel data, this does not matter since first-differencing remove this systematic variation. To produce noise estimates/standard errors, their method requires integrating over the 2D curve for arbitrarily chosen rings. In cases where researchers chose these aggregation ranges based on the results of the estimation procedure, standard inference tools are not valid even when bootstrapping the standard errors (Leeb and Pötscher, 2005). As discussed above, my proposed estimator provides valid inference after selection of aggregation regions and choses these aggregation regions in a data-driven manner.

This paper also contributes to a small literature on difference-in-differences estimators from a spatial lens (Butts, 2021b; Clarke, 2017; Berg and Streitz, 2019; Verbitsky-Savitz and Raudenbush, 2012b; Delgado and Florax, 2015). These papers address instances where treatment is well defined by administrative boundaries but spillovers cause problems of defining who is 'treated' and at what level of exposure. This paper contributes to this literature by considering the setting where treatment occurs at a point in space. Butts (2021b) and Clarke (2017) both recommend a method of using many rings to estimate a treatment effect curve, but are not able to provide a data-driven way to select the rings. Since this paper focuses on local shocks where constant parallel trends are plausible, I am able to provide a data-driven approach to choosing rings.

3.2 Example of Problem

To illustrate the methodological difficulties in this method, suppose that an overgrown empty lot in a high-poverty neighborhood is cleaned up by the city and the outcome of interest is home prices. The researcher observes a panel of home sales before and after the lot is cleaned. Figure 3.2 shows a plot of simulated data from this example. The black line is treatment effect at different distances from the empty lot and the grey line is the underlying (*constant*) counterfactual change in home prices, normalized to 0. Panel (a) of Figure 3.2 shows the best-case scenario where the treated ring is correctly specified. The two horizontal lines show the average change in outcome in the treated ring and the control ring and our treatment effect estimate, $\hat{\tau}$, is the difference between these two averages. However, this singular number masks over a large amount of treatment effect heterogeneity with units experiencing treatment effects as little as half and as large as double the magnitude that of $\hat{\tau}$. Later in this paper, I recommend nonparametrically estimating the treatment effect curve as a function of distance rather than using average effect.

However, the researcher does not typically know the distance at which treatment effects stop. Panel (b) shows when the 'treatment' ring is too wide. In this case, some of the units in the treatment ring receive no effect from treatment and therefore makes the average treament effect among units in the treatment ring smaller. Panel (c) of Figure 3.2 shows the opposite case, where the treated ring is too narrow. In this case, there are some units in the 'control' ring that experience treatment effects and the coutnerfactual trend estimate is biased upward. The treatment effect estimate grows since the new 'treated' group is closer and hence experiences larger treatment effects. Panel (d) of Figure 3.2 shows an example of why using different rings as a 'robustness check' is a problematic practice. Using Panel (d) as a robustness check for the researchers' original specification of Panel (c), a researcher would be quite confident in their results even though the estimate is too large in both cases.

From these three examples, it's clear that the estimation strategy requires researchers to know the exact distance at which treatment effects become zero. Since this is a very demanding assumption, I propose an improved estimator in section 3.4 that relaxes this assumption by leveraging the constant common trends assumption.

3.3 Theory

Now, I develop econometric theory to formalize the intuition developed in the previous section. A researcher observes panel data of a random sample of units i at times $t = 0, 1$ located in space at point $\theta_i = (x_i, y_i)$. Treatment occurs at a location $\bar{\theta} = (\bar{x}, \bar{y})$ between periods. Therefore, units differ in their distance to treatment, defined by $\text{Dist}_i \equiv d(\theta_i, \bar{\theta})$ for some distance metric d (e.g. Euclidean distance) with a distribution function F. Outcomes are given by

$$Y_{it} = \mu_i + \tau_i \, 1_{t=1} + \lambda_i \, 1_{t=1} + u_{it}, \tag{3.1}$$

where μ_i is unit-specific time-invariant factors, λ_i is the change in outcomes due to non-treatment shocks in period 1, τ_i is unit i's treatment effect. Both λ and τ can be split into a systematic function of distance $z(\text{Dist}_i)$ and an idiosyncratic term $\tilde{z}_i \equiv z_i - z(\text{Dist}_i)$ with z being τ and λ.

$\tau(d)$ is the average effect of treatment at a given distance and $\lambda(d)$ summarizes how covariates and shocks change over distance.

Therefore, we could rewrite our model as

$$Y_{it} = \mu_i + \tau(\text{Dist}_i)\, 1_{t=1} + \lambda(\text{Dist}_i)\, 1_{t=1} + \varepsilon_{it}, \tag{3.2}$$

where $\varepsilon = u_{it} + \tilde{\tau}_i + \tilde{\lambda}_i$ which is uncorrelated with distance to treatment. Researchers are trying to identify the average treatment effect on units experiencing treatment effects, i.e. $\bar{\tau} = \mathbb{E}[\tau_i \mid \tau(\text{Dist}_i) > 0]$.

Assumption 3.1 (Random Sampling)**.** *The observed data consists of* $\{Y_{i1}, Y_{i0}, Dist_i\}$ *which is independent and identically distributed.*

Taking first-differences of our model, we have $\Delta Y_{it} = \tau(\text{Dist}_i) + \lambda(\text{Dist}_i) + \Delta\varepsilon_{it}$. It is clear that $\tau(\text{Dist}_i)$ and $\lambda(\text{Dist}_i)$ are not seperately identified unless additional assumptions are imposed. The central identifying assumption that researchers claim when using the ring method is that counterfactual trends likely evolve smoothly over distance, so that $\lambda(\text{Dist}_i)$ is approximately constant within a small distance from treatment. This is formalized in the context of our outcome model by the following assumption.

Assumption 3.2 (Local Parallel Trends)**.** *For a distance* $\bar{d}$*, we say that 'local parallel trends' hold if for all positive* $d, d' \leq \bar{d}$*, then* $\lambda(d) = \lambda(d')$*.*

This assumption requires that, in the absence of treatment, outcomes would evolve the same at every distance from treatment within a certain maximum distance, $\bar{d}$. To clarify the assumption, it is helpful to think of ways that it can fail. First, if treatment location is targeted based on trends *within a small-area/neighborhood*, then trends would not be constant within the control ring. Second, if units sort either towards or away from treatment in a way that is systematically correlated with the outcome variable, then the compositional change can cause a violation in trends over time. Note that Local Parallel Trends implies the standard assumption that parallel trends holds *on average* between the treated and control rings:

Assumption 3.3 (Average Parallel Trends)**.** *For a pair of distances d_t and d_c, we say that 'average parallel trends' hold if* $\mathbb{E}[\lambda_d \mid 0 \leq d \leq d_t] = \mathbb{E}[\lambda_d \mid d_t < d \leq d_c]$.

If Local Parallel Trends holds for some d_c, then our first-difference equation can be simplified to $\Delta Y_{it} = \tau(\text{Dist}_i) + \lambda + \Delta\varepsilon_{it}$ where λ is some constant for units in the subsample $\mathcal{D} \equiv \{i \ : \ \text{Dist}_i \leq d_c\}$. Therefore, the treatment effect curve $\tau(\text{Dist}_i)$ is identifiable up to a constant under Assumption 3.2. To identify $\tau(\text{Dist}_i)$ seperately from the constant, researchers will often claim that treatment effects stop occuring before some distance $d_t < d_c$. This is formalized in the following assumption.

Assumption 3.4 (Correct d_t)**.** *A distance d_t satisfies this assumption if (i) for all $d \leq d_t$, $\tau(d) > 0$ and for all $d > d_t$, $\tau(d) = 0$ and (ii) $F(d_c) - F(d_t) > 0$.*

With this assumption, the first difference equation simplifies to $\Delta Y_{it} = \lambda + \Delta\varepsilon_{it}$ for units with $d_t < \text{Dist}_i < d_c$. These units therefore identify λ. The 'ring method' is the following procedure. Researchers select a pair of distances $d_t < d_c$ which define the "treated" and "control" groups. These groups are defined by $\mathcal{D}_t \equiv \{i : 0 \leq \text{Dist}_i \leq d_t\}$ and $\mathcal{D}_c \equiv \{i : d_t < \text{Dist}_i \leq d_c\}$. On the subsample of observations defined by $\mathcal{D} \equiv \mathcal{D}_t \cup \mathcal{D}_c$, they estimate the following regression:

$$\Delta Y_{it} = \beta_0 + \beta_1 \, 1_{i \in \mathcal{D}_t} + u_{it}. \tag{3.3}$$

From standard results for regressions involving only indicators, difference-in-differences estimator is $\hat{\beta}_1$ with the following expectation:

$$\mathbb{E}\left[\hat{\beta}_1\right] = \mathbb{E}[\Delta Y_{it} \mid \mathcal{D}_t] - \mathbb{E}[\Delta Y_{it} \mid \mathcal{D}_c].$$

This estimate is decomposed in the following proposition.[6]

Proposition 3.1 (Decomposition of Ring Estimate)**.** *Given that units follow model (3.2),*

[6] A similar derivation of part (i) is found in Sullivan (2017) but does not include difference in parallel trends.

(i) The estimate of β_1 in (3.3) has the following expectation:

$$\begin{aligned}\mathbb{E}\left[\hat{\beta}_1\right] &= \mathbb{E}[\Delta Y_{it} \mid \mathcal{D}_t] - \mathbb{E}[\Delta Y_{it} \mid \mathcal{D}_c] \\ &= \underbrace{\mathbb{E}[\tau(Dist) \mid \mathcal{D}_t] - \mathbb{E}[\tau(Dist) \mid \mathcal{D}_c]}_{\textit{Difference in Treatment Effect}} + \underbrace{\mathbb{E}[\lambda(Dist) \mid \mathcal{D}_t] - \mathbb{E}[\lambda(Dist) \mid \mathcal{D}_c]}_{\textit{Difference in Trends}}.\end{aligned}$$

(ii) If d_c satisfies Local Parallel Trends or, more weakly, if d_t and d_c satisfy Average Parallel Trends, then

$$\mathbb{E}\left[\hat{\beta}_1\right] = \underbrace{\mathbb{E}[\tau(Dist) \mid \mathcal{D}_t] - \mathbb{E}[\tau(Dist) \mid \mathcal{D}_c]}_{\textit{Difference in Treatment Effect}}.$$

(iii) If d_c satisfies Local Parallel Trends and d_t satisfies Assumption 3.4, then

$$\mathbb{E}\left[\hat{\beta}_1\right] = \bar{\tau}.$$

Part (i) of this proposition shows that the estimate is the sum of two differences. The first difference is the difference in average treatment effect among units in the treated ring and units in the control ring. The second difference is the difference in counterfactual trends between the treated and control rings. This presents two possible problems. If some units in the control group experience effects from treatment, the average of these effects will be subtracted from the estimate. Second, since treatment can be targeted, the treated ring could be on a different trend than units further away and hence control units do not serve as a good counterfactual for treated units.

Part (ii) says that if d_c satisfies Local Parallel Trends, then the difference in trends from part (i) is equal to 0. As discussed above, the decomposition in part (ii) of Proposition 3.1 is not necessarily unbiased estimate for $\bar{\tau}$. First, if d_t is *too wide*, then $\mathcal{D}_t$ contain units that are not affected by treatment. In this case, $\hat{\beta}_1$ will be biased towards zero from the inclusion of unaffected units from d_t being too wide. Second, if d_t is *too narrow* then the $\mathcal{D}_c$ will contain units that experience treatment effects. It is not clear in this case, though, whether $\hat{\beta}_1$ will grow or shrink without knowledge of the τ(Dist) curve, but typically $\hat{\beta}_1$ will not be an unbiased estimate for $\bar{\tau}$. See the previous section for an example.

Part (iii) of Proposition 3.1 shows that if d_t is correctly specified as the maximum distance that receives treatment effect, then $\hat{\beta}_1$ will be an unbiased estimate for the average treatment effect among the units affected by treatment. However, Assumption 3.4 is a very demanding assumption and unlikely to be known by the researcher unless there are *a priori* theory dictating d_t.[7] The following section will improve estimation by allowing consistent nonparametric estimation of the entire $\tau(\text{Dist})$ function. An estimate of $\tau(\text{Dist})$ can then be numerically integrated to for an estimate of $\bar{\tau}$.

3.4 Nonparametric Estimation of the Treatment Effect Curve

In this section, I propose an estimation strategy that nonparametrically identifies the treatment effect curve $\tau(\text{Dist}_i)$ using partitioning-based least squares estimation and inference methods developed in Cattaneo et al. (2019a,b). Partition-based estimators seperate the support of a covariate, Dist_i, into a set of quantile-spaced intervals (e.g. 0-25th percentiles of Dist_i, 25-50th, 50-75th, and 75-100th). Then the conditional $\mathbb{E}[Y_i \mid \text{Dist}_i]$ is estimated seperately within each interval as a k-degree polynomial of the covariate Dist_i.

For a given outer-ring distance, d_c, our sample is given by $\mathcal{D} = \{i : \text{Dist}_i \leq d_c\}$. We will split the sample into L intervals based on quantiles of the distance variable. For a given $j \in \{1, \dots, L\}$, I denote the j^{th} interval as $\mathcal{D}_j \equiv \{i : F_n^{-1}(\frac{j-1}{L}) \leq \text{Dist}_i < F_n^{-1}(\frac{j}{L})\}$ where F_n is the empirical distribution of Dist. Let $\{\mathcal{D}_1, \dots, \mathcal{D}_L\}$ be the collection of the L intervals. This paper will use a 0-degree polynomial for each interval, predicting ΔY_{it} with a constant within each interval.[8] These averages are defined as

$$\overline{\Delta Y}_j \equiv \frac{1}{n_j} \sum_{i \in \mathcal{D}_j} \Delta Y_{it},$$

[7] As an example, Currie et al. (2015) uses results from scientific research on the maximum spread of local pollutants and Marcus (2021) use the plume length of petroleum smoke.

[8] Approximation can be made arbitrarily close to the true conditional expectation function by *either* increasing the number of intervals *or* by increasing the polynomial order to infinity, so setting $k = 0$ does not impose any asymptotic cost.

where the number of units in bin $\mathcal{D}_j$ is $n_j \approx n/L$. Our estimator for $\mathbb{E}[\Delta Y_{it} \mid \text{Dist}_i]$ is then given by

$$\widehat{\Delta Y_{it}} = \sum_{j=1}^{L} 1_{i \in \mathcal{D}_j} \, \overline{\Delta Y}_j$$

As the number of intervals approach infinity, this estimate will approach $\mathbb{E}[\Delta Y_{it} \mid \text{Dist} = d]$ uniformly. Under Local Parallel Trends, $\mathbb{E}[\Delta Y_{it} \mid \text{Dist} = d] \equiv \mathbb{E}[\tau(\text{Dist}) \mid \text{Dist} = d] + \lambda$. To remove λ, we require a less-strict version of assumption 3.4.

Assumption 3.5 (d_t is within d_c). *A distance d_c satisfies this assumption if there exists a distance d_t with $0 < d_t < d_c$ such that (i) Assumption 3.4 holds and (ii) $F(d_c) - F(d_t) > 0$.*

If a distance d_c satisfies Local Parallel Trends and (3.5), the mean within the last ring $\mathcal{D}_k$ will estimate λ as the number of bins $L \to \infty$. The reason for this is simple, as $L \to \infty$, the last bin will have a left end-point $> d_t$ and therefore $\tau(\text{Dist}) = 0$ in $\mathcal{D}_L$. Under local parallel trends, the last ring will therefore estimate λ. Therefore, estimates of $\tau(\text{Dist}_i)$ can be formed for each interval as $\hat{\tau}_j \equiv \overline{\Delta Y}_j - \overline{\Delta Y}_L$. These $\hat{\tau}_j$ can be plotted over their intervals to provide a graphical estimate of the treatment effect curve (see below for an example).

Proposition 3.2 (Consistency of Nonparametric Estimator). *Given that units follow model (3.2) and d_c satisfies Local Parallel Trends and assumption (3.5), as n and $L \to \infty$*

$$\hat{\tau} \equiv \sum_{i=1}^{L} \hat{\tau}_j 1_{i \in \mathcal{D}_j} \to^{unif} \tau(Dist)$$

under regularity conditions given by Assumption (1) of Cattaneo et al. (2019b), $L \log(L)/n \to 0$ and $nL^{-5} \to 0$.[9]

As discussed in section 3.2, specifying d_t correctly is important to identify the average treatment effect among the affected in the parametric estimator. The nonparametric estimator only requires that treatment effects become zero before d_c, i.e. that such a d_t exists. However, the estimator comes at a cost, namely it would no longer identify the treatment effect curve under the

[9] See section C.1 for the proof.

milder Average Parallel Trends assumption. Therefore, a researcher should justify explicity the assumption that, within the d_c ring, every unit is subject to the same trend. This is most likely to be satisfied on a very local level and not very plausible in the case of larger units, e.g. counties.

The nonparametric approach allows estimation of the treatment effect curve which allows researcher to understand differences in treatment effect across distance whereas the indicator approach, *at best*, can only estimate an *average* effect among units experiencing effects. For example, typically one would assume treatment effects shrink over distance and evidence of this from the nonparametric approach can strengthen a causal claim. In some cases, such as a negative hyper-local shock and a postivie local shock (e.g. a local bus-stop), the treatment effect can even change sign across distances. In this case, the average effect could be near zero even though there are significant effects occuring.

Plotting estimates $\hat{\tau}_j$ can provide visual evidence for the underlying Local Parallel Trends assumption. Typically, treatment effect will stop being experienced far enough away from d_c that some estimates of $\hat{\tau}_j$ with j 'close to' L will provide informal tests for parallel trends holding. Figure 3.4 provide an example where plotting of $\hat{\tau}_j$ provide strong evidence in support of local parallel trends as it appears that after some distance, average effects are consistetly centered around zero. This is not a formal test as it could be the case that the true treatment effect curve, $\tau(\text{Dist})$ is perfectly cancelling out with the counterfactual trends curve $\lambda(\text{Dist})$ producing near zero estimates, but this is a knive's edge case.

The above proposition shows that the series estimator will consistenly estimate the treatment effect curve, $\tau(\text{Dist})$ as the number of bins L and the number of observations n both go to infinity. In finite-samples though, we will have a fixed L and hence a fixed set of treatment effect estimates $\{\tau_1, \dots, \tau_L\}$ with $\tau_L \equiv 0$ by definition. The estimates $\hat{\tau}_j$ are approximately equal to $\mathbb{E}[\tau(\text{Dist}) \mid \text{Dist} \in \mathcal{D}_j]$ or the average treatment effect within the interval $\mathcal{D}_j$.

The choice of L in finite samples is not entirely clear. Cattaneo et al. (2019a) derive the optimal choice of L which is a completely data-driven choice. The optimal L is driven by two competing terms in the mean-squared error formula. On the one hand, as L increases, the conditional expectation function is allowed to vary more across values of Dist and hence bias of the estimator decreases. However, larger values of L increase the variance of the estimator. Balancing this trade-off depends on the shape and curvature of $\tau(\text{Dist})$ which is estimated by the data. The resulting choice of L^* determines the number of bins and the rings are determined by quantiles of the data. Since the rings are themselves estimated, Cattaneo et al. (2019a) provide inference that is valid after the 'first-stage' estimation of quantiles.

This data-driven choice allows for estimation in a principled and objective way that lets the data 'speak for itself'. Further, this method removes incentives from 'pre-testing' the choice of rings to find a significant result since the estimates are determined completely by the data [10]

For a given L^*, Cattaneo et al. (2019a) show the large-sample asymptotics of the estimates $\overline{\Delta Y}_j$ and provide robust standard errors for the conditional means that account for the additional randomness due to quantile estimation. Since our estimator is a difference in means, standard errors on our estimate $\hat{\tau}_j$ are given by $\sqrt{\sigma_j^2 + \sigma_L^2}$, where σ_j is the standard error recommended by Cattaneo et al. (2019a). These standard errors are produced by the Stata/R package `binsreg`. Inference can be done by using the estimated t-stat with the standard normal distribution. There may be concerned that the standard errors need to adjust for spatial correlation. However, this is not the case under assumption (3.2) as this implies the error term is uncorrelated with distance.

Remark 3.1 (Overall Average Treatment Effect)**.** *A researcher may be tempted to 'pool' together the significant rings to estimate an overall average treatment effect among the affected. Since the estimates are quantiles in the data, a simple average of the significant $\hat{\tau}_j$'s can produce a back of the envelope treatment effect. However, infernece on this estimand is not possible, since, in repeated sampling the number of significant $\hat{\tau}_j$ can change and this additional source of 'model selection' makes inference a very difficult problem (Leeb and Pötscher, 2005). A potential solution would be to use cross-validation where half the data would determine the 'inner ring' and then the second half of the data would estimate the overall average treatment effect, though potentially requiring a large*

[10] See, for example, discussions in Andrews and Kasy (2019) and McCloskey and Michaillat (2020) about researcher degrees of freedom.

sample size.

Remark 3.2 (Covariates)**.** *A researcher may often be interested The partitioning based series-estimator can allow for covariates to be included in the model with valid inference, so the estimation can easily be done with the Stata/R package* `binsreg`*. However, including covariates changes the necessary common trends assumption to have to hold conditional on the vector of covariates X (See Sant'Anna and Zhao (2020) for modern discussion of conditonal parallel trends). In this paper, this estimation strategy would require neighborhood shocks to be common across values of X_i, i.e. shocks are experienced equally regardless of characteristics X.*

3.5 Application to Neighborhood Effects of Crime Risk

To highlight the advantages of my proposed estimator, I revisit the analysis of Linden and Rockoff (2008b). This paper analyzes the effect of a sex offender moving to a neighborhood on home prices. This paper uses the ring method with treated homes being defined as being within $1/10^{th}$ of the sex offender's home and the control units being between $1/10^{th}$ and $1/3^{rd}$ of a mile from the home. The authors make a case for the ring method by arguing that *within a neighborhood*, Local Parallel Trends holds since they are looking in such a narrow area and purchasing a home is difficult to be precisely located with concurrent hyper-local shocks. This application uses cross-sectional data which similar identification and estimation results are presented in the Appendix.

As for the choice of the treatment ring, there is little *a priori* reasons to know how far the effects of sex offender arrival will extend in the neighborhood. The authors provide graphical evidence of nonparametric estimates of the conditional mean home price at different distances in the year before and the year after the arrival of a sex offender. The published plot can be seen in Panel (b) of Figure 3.3. They 'eyeball' the point at which the two estimates are approximately equal to decide how far treatment effects extend. However, this approach is less precise than it may seem. Panels (a) and (c) show that changing the bandwidth for the kernel density estimator will produce very different guesses at how far treatment effects extend. My proposed estimator works in a data-driven way that does not require these ad-hoc decisions.

The standard rings approach is equivalent to my proposed method with two rings: $\mathcal{D}_1$ being the treated homes between 0 and 0.1 miles away and $\mathcal{D}_2$ being the control homes between 0.1 and 0.3 miles away. The average change among $\mathcal{D}_2$ estimates the counterfactual trend and the average change among $\mathcal{D}_1$ minus the estimated counterfactual trend serves as the treatment effect. Panel (a) of Figure 3.4 shows the basic results of their difference-in-differences analysis which plots estimates $\hat{\tau}_j$ for $j = 1, 2$. On average, homes between 0 and 0.1 miles decline in value by about 7.5% after the arrival of a sex offender. *As an assumption* of the rings method, homes between 0.1 and 0.3 miles away are not affected by a sex offender arrival. The choice of 0.1 miles is an untestable assumption and as seen above the evidence provided is highly dependent on the choice of bandwidth parameter. My proposed estimator does not require a specific choice for a 'treated' area.

Linden and Rockoff (2008b) only have access to a non-panel sample of home sales, so identification requires another assumption for identification, namely that the composition of homes at a given distance does not change over time. Further, since we can no longer form first-differences of the outcome variable, seperate nonparametric estimators must be estimated before and after treatment and subtracted from one another. Details of this theory are in the Appendix.

Panel (b) of Figure 3.4 applies the nonparametric approach described in Section 3.4. Two differences in results occur. First, homes in the two closest rings i.e. within a few hundred feet, are most affected by sex-offender arrival with an estimated decline of home value of around 20%. homes a bit further away but still within in Linden and Rockoff's 'treated' sample do not experience statistically significant treatment effects. As discussed above, Linden and Rockoff's estimate of $\bar{\tau}$ is attenuated towards zero because of the inclusion of homes with little to no treatment effects, leading them to understate the effect of arrival on home prices. The nonparametric approach improves on answering this question by providing a more complete picture of the treatment effect curve. The magnitude of treatment effects decrease over distance, providing additional evidence that the arrival causes a drop in home prices.[11]

[11] This is similar to estimating a dose-response function as evidence supporting a causal mechanism. The results of

The second advantage of this approach is that the produced figure provides an informal test of the local parallel trends assumption. After 0.1 miles, the estimated treatment effect curve becomes centered at zero consistently. This implies that units within each ring have the same estimated trend as the outer most ring, providing suggestive evidence that homes in this neighborhood are subject to the same trends.

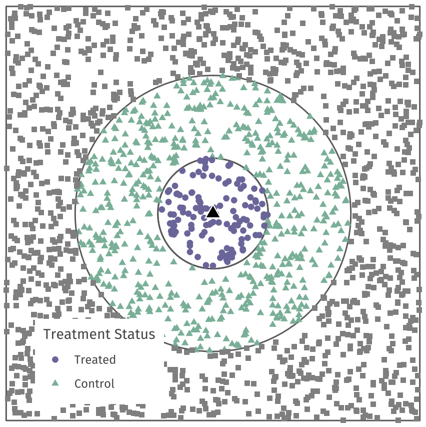
3.6 Conclusion

This article formalizes a common applied identification strategy that has a strong intuitive appeal. When treatment effects of shocks are experienced in only part of an area that would otherwise be on a common neighborhood-trend, difference-in-differences comparisons within a neighborhood can identify treatment effects. However, this paper shows that the typical *estimator* for treatment effects requires a very strong assumption and returns only an average treatment effect among affected units when this assumption holds.

This article then proposes an improved estimator that relies on nonparametric series estimators. The nonparametric estimator allows for estimation of the treatment effect at different distances from treatment, similar to a dose-response function, which can allow better understanding of *who* is experiencing effects and how this changes across ‘exposure’ to a shock. More, in some cases it can provide explanation for null results. For example, if a bus station creates negative externalities for apartments that border the station but positive externalities for apartments within walking distance, the average effect could be zero. However, nonparametric estimation would reveal the two effects seperately.

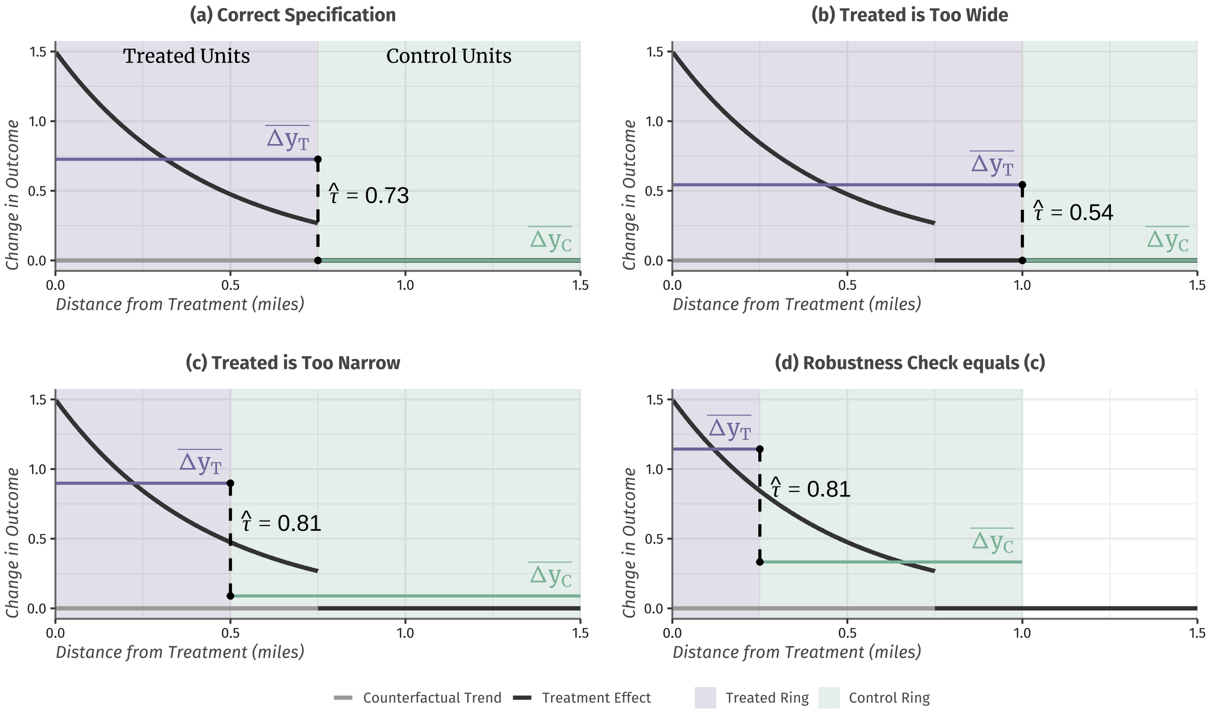
this paper are similar to the results of Callaway et al. (2021) with continuous treatment. In their framework, Local Parallel Trends is analagous to their assumption of common trends at different doses of a continuous treatment. In this setting, I am able to provide an estimator for the treatment effect curve, the average level effect in their terminology, by relying on an assumption that the treatment effect curve $\tau(\text{Dist})$ is homogeneous across units.

Figure 3.1: Rings Method



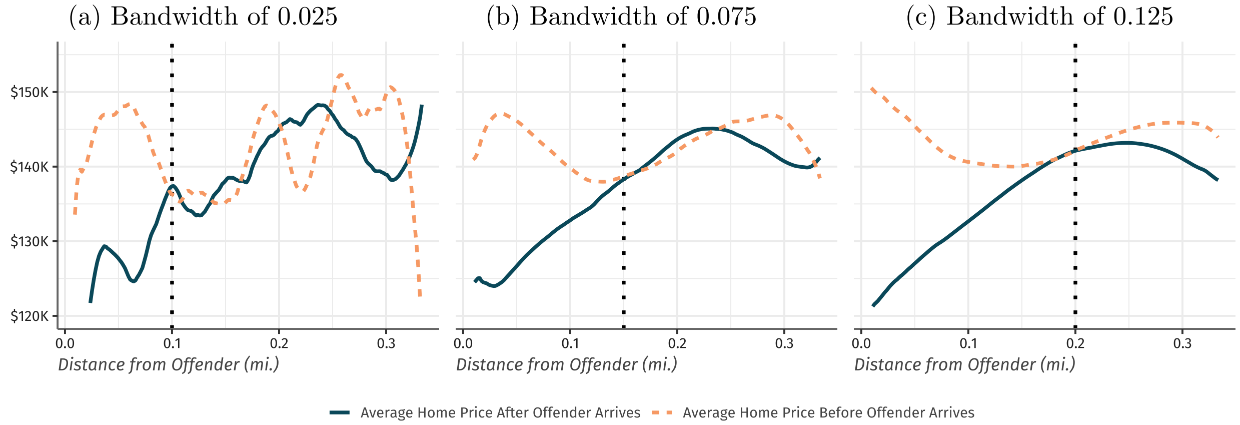
Notes. This figure illustrates the ring method. The center triangle represents the location of treatment, e.g. a foreclosed home. Units within the inner circle, marked by dots, are considered 'treated'; units between the inner and outer circles, marked in triangles, are considered control units; and then the remaining units are removed from the sample. The ring estimate compares average changes in outcomes between the inner 'treated' ring and the outer 'control' ring.

Figure 3.2: Example of Problems with Ad-Hoc Ring Selection



Notes. This figure shows an example of the difficulties in estimation of treatment effects via the ring method.

Figure 3.3: Price Gradient of Distance from Offender



(a) Bandwidth of 0.025

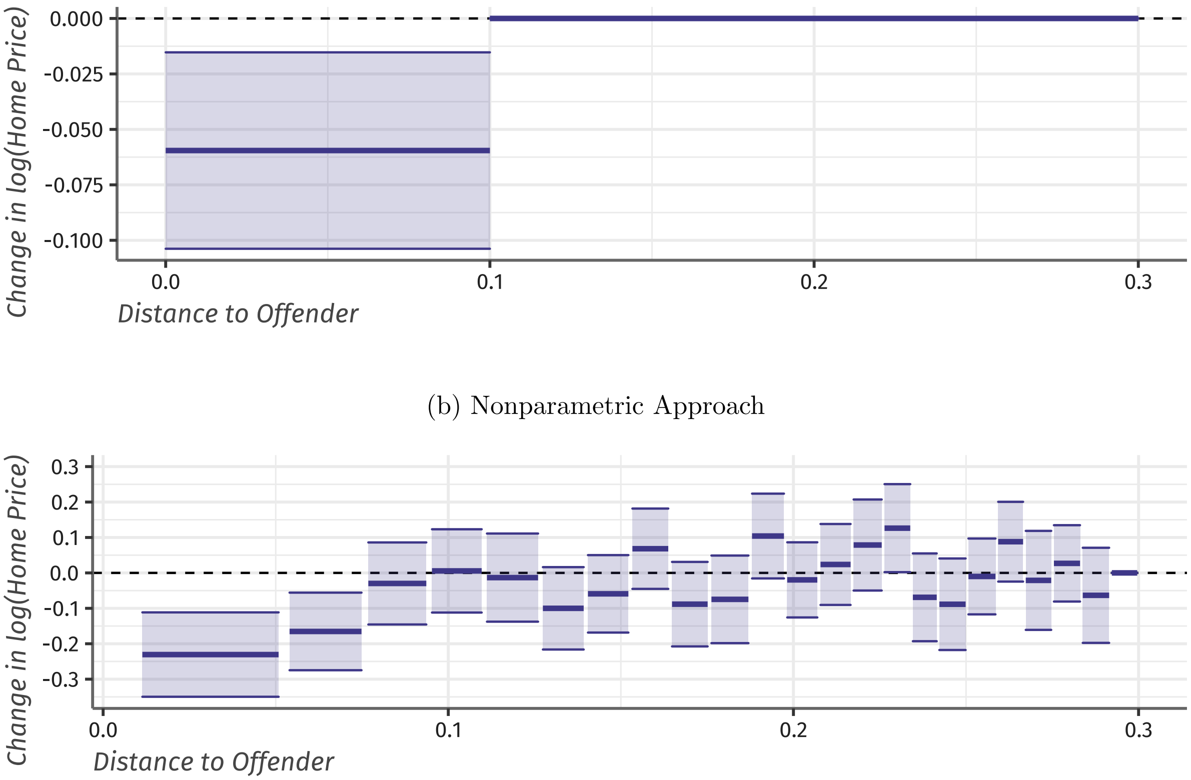
(b) Bandwidth of 0.075

(c) Bandwidth of 0.125

Notes. This figure plots estimates of home prices in the year before and the year after the arrival of a sex offender estimated using a Local Polynomial Kernel Density estimation with an Epanechnikov kernel. Panel (b) recreates Figure 2 from Linden and Rockoff (2008b) and the other panels change the bandwidth.

Figure 3.4: Effects of Offender Arrival on Home Prices (Linden and Rockoff, 2008b)

(a) Indicator Approach



(b) Nonparametric Approach

Notes. This figure plots the estimated change in home prices after the arrival of a registered sex offender as a function of distance from offender. Each line plots $\hat{\tau}_j = \overline{\Delta Y}_j - \overline{\Delta Y}_l$ with associated standard errors. Panel (a) shows an estimate from Equation 3.3 with a treatment distance of $1/10^{th}$ miles and a control distance of $1/3^{rd}$ mile. Panel (b) shows the nonparametric estimate of $\tau(\text{Dist}_i)$ proposed in Section 3.4.

Bibliography

Abadie, A. (2021). Using synthetic controls: Feasibility, data requirements, and methodological aspects. Journal of Economic Literature, 59(2):391–425.

Abadir, K. M. and Magnus, J. R. (2005). Matrix Algebra, volume 1. Cambridge University Press.

Ahn, S. C., Lee, Y. H., and Schmidt, P. (2001). Gmm estimation of linear panel data models with time-varying individual effects. Journal of Econometrics, 101(2):219–255.

Ahn, S. C., Lee, Y. H., and Schmidt, P. (2013). Panel data models with multiple time-varying individual effects. Journal of econometrics, 174(1):1–14.

Alexander, D., Currie, J., and Schnell, M. (2019). Check up before you check out: Retail clinics and emergency room use. Journal of Public Economics, 178:104050.

Andrews, I. and Kasy, M. (2019). Identification of and correction for publication bias. American Economic Review, 109(8):2766–2794.

Angelucci, M. and Di Maro, V. (2016). Programme evaluation and spillover effects. Journal of Development Effectiveness, 8(1):22–43.

Angrist, J. D. (2014). The perils of peer effects. Labour Economics, 30:98–108.

Arcidiacono, P., Ellickson, P. B., Mela, C. F., and Singleton, J. D. (2020). The competitive effects of entry: Evidence from supercenter expansion. American Economic Journal: Applied Economics, 12(3):175–206.

Arefeva, A., Davis, M. A., Ghent, A. C., and Park, M. (2021). The effect of capital gains taxes on business creation and employment: The case of opportunity zones. Available at SSRN 3645507.

Arkhangelsky, D., Athey, S., Hirshberg, D. A., Imbens, G. W., and Wager, S. (2021). Synthetic difference-in-differences. American Economic Review, 111(12):4088–4118.

Aronow, P. M. and Samii, C. (2017). Estimating average causal effects under general interference, with application to a social network experiment. Annals of Applied Statistics, 11(4):1912–1947.

Asquith, B. J., Mast, E., and Reed, D. (2021a). Local effects of large new apartment buildings in low-income areas. Review of Economics and Statistics, pages 1–46.

Asquith, B. J., Mast, E., and Reed, D. (2021b). Local effects of large new apartment buildings in low-income areas. The Review of Economics and Statistics, page 1–46.

Athey, S., Bayati, M., Doudchenko, N., Imbens, G., and Khosravi, K. (2021). Matrix completion methods for causal panel data models. Journal of the American Statistical Association, 116(536):1716–1730.

Bai, J. (2009). Panel data models with interactive fixed effects. Econometrica, 77(4):1229–1279.

Bailey, M. J. and Goodman-Bacon, A. (2015). The war on poverty's experiment in public medicine: Community health centers and the mortality of older americans. American Economic Review, 105(3):1067–1104.

Basker, E. (2005). Job creation or destruction? labor market effects of wal-mart expansion. Review of Economics and Statistics, 87(1):174–183.

Baum-Snow, N. and Ferreira, F. (2015). Causal Inference in Urban and Regional Economics, volume 5, pages 3–68. Elsevier.

Berg, T. and Streitz, D. (2019). Handling spillover effects in empirical research. Technical report, Working Paper.

Berkes, E. and Nencka, P. (2021). Knowledge access: The effects of carnegie libraries on innovation. Available at SSRN 3629299.

Billings, S. B. (2011). Estimating the value of a new transit option. Regional Science and Urban Economics, 41(6):525–536.

Borusyak, K., Jaravel, X., and Spiess, J. (2024). Revisiting event study designs: Robust and efficient estimation. Review of Economic Studies.

Brown, N., Butts, K., and Westerlund, J. (2023a). Simple difference-in-differences estimation in fixed-t panels.

Brown, N. L., Schmidt, P., and Wooldridge, J. M. (2023b). Simple alternatives to the common correlated effects model.

Busso, M., Gregory, J., and Kline, P. (2013). Assessing the incidence and efficiency of a prominent place based policy. American Economic Review, 103(2):897–947.

Butts, K. (2021a). did2s: Two-Stage Difference-in-Differences Following Gardner (2021).

Butts, K. (2021b). Difference-in-Differences with Spatial Spillovers. Working Paper.

Butts, K. (2023). Jue insight: Difference-in-differences with geocoded microdata. Journal of Urban Economics, 133:103493.

Butts, K. and Gardner, J. (2022). did2s: Two-stage difference-in-differences. R Journal, 14(3).

Callaway, B., Goodman-Bacon, A., and Sant'Anna, P. H. C. (2021). Difference-in-Differences with a Continuous Treatment. Preprint. `http://arxiv.org/abs/2107.02637`.

Callaway, B. and Karami, S. (2023). Treatment effects in interactive fixed effects models with a small number of time periods. Journal of Econometrics, 233(1):184–208.

Callaway, B. and Sant'Anna, P. H. (2021). Difference-in-differences with multiple time periods. Journal of Econometrics, 225(2):200–230.

Callaway, B. and Sant'Anna, P. H. (2020). Difference-in-differences with multiple time periods. Journal of Econometrics, page S0304407620303948.

Campbell, J. Y., Giglio, S., and Pathak, P. (2011). Forced sales and house prices. American Economic Review, 101(5):2108–2131.

Casey, M., Schiman, J. C., and Wachala, M. (2018). Local violence, academic performance, and school accountability. AEA Papers and Proceedings, 108:213–216.

Cattaneo, M. D., Crump, R. K., Farrell, M. H., and Feng, Y. (2019a). On binscatter. arXiv preprint arXiv:1902.09608.

Cattaneo, M. D., Farrell, M. H., and Feng, Y. (2019b). Large sample properties of partitioning-based series estimators. arXiv:1804.04916 [econ, math, stat]. arXiv: 1804.04916.

Chan, M. K. and Kwok, S. S. (2022). The pcdid approach: difference-in-differences when trends are potentially unparallel and stochastic. Journal of Business & Economic Statistics, 40(3):1216–1233.

Chen, J., Glaeser, E., and Wessel, D. (2023). Jue insight: The (non-) effect of opportunity zones on housing prices. Journal of Urban Economics, 133:103451.

Clarke, D. (2017). Estimating difference-in-differences in the presence of spillovers. Munich Personal RePEc Archive, page 52.

Conley, T. (1999). Gmm estimation with cross sectional dependence. Journal of Econometrics, 92(1):1–45.

Cuberes, D., Desmet, K., and Rappaport, J. (2021). Urban growth shadows. Journal of Urban Economics, page 103334.

Cui, L. and Walsh, R. (2015). Foreclosure, vacancy and crime. Journal of Urban Economics, 87:72–84.

Currie, J., Davis, L., Greenstone, M., and Walker, R. (2015). Environmental health risks and housing values: Evidence from 1,600 toxic plant openings and closings. American Economic Review, 105(2):678–709.

De Chaisemartin, C. and d'Haultfoeuille, X. (2024). Difference-in-differences estimators of intertemporal treatment effects. Review of Economics and Statistics, pages 1–45.

De Chaisemartin, C. and d'Haultfoeuille, X. (2020). Two-way fixed effects estimators with heterogeneous treatment effects. American economic review, 110(9):2964–2996.

Delgado, M. S. and Florax, R. J. (2015). Difference-in-differences techniques for spatial data: Local autocorrelation and spatial interaction. Economics Letters, 137:123–126.

Di Tella, R. and Schargrodsky, E. (2004). Do police reduce crime? estimates using the allocation of police forces after a terrorist attack. American Economic Review, 94(1):115–133.

Diamond, R. and McQuade, T. (2019). Who wants affordable housing in their backyard? an equilibrium analysis of low-income property development. Journal of Political Economy, 127(3):1063–1117.

Donaldson, D. and Hornbeck, R. (2016). Railroads and american economic growth: A “market access” approach*. The Quarterly Journal of Economics, 131(2):799–858.

Eckert, F., Fort, T. C., Schott, P. K., and Yang, N. J. (2021). Imputing missing values in the us census bureau’s county business patterns. Technical report, National Bureau of Economic Research.

Fernández-Val, I., Freeman, H., and Weidner, M. (2021). Low-rank approximations of nonseparable panel models. The Econometrics Journal, 24(2):C40–C77.

Freyaldenhoven, S., Hansen, C., Pérez, J. P., and Shapiro, J. M. (Forthcoming). Visualization, identification, and estimation in the linear panel event-study design.

Freyaldenhoven, S., Hansen, C., and Shapiro, J. M. (2019). Pre-event trends in the panel event-study design. American Economic Review, 109(9):3307–3338.

Gardner, J. (2022). Two-Stage Difference-in-Differences. Preprint. `http://arxiv.org/abs/2207.05943`.

Gerardi, K., Rosenblatt, E., Willen, P. S., and Yao, V. (2015). Foreclosure externalities: New evidence. Journal of Urban Economics, 87:42–56.

Gibbons, S. and Machin, S. (2005). Valuing rail access using transport innovations. Journal of urban Economics, 57(1):148–169.

Gobillon, L. and Magnac, T. (2016). Regional policy evaluation: Interactive fixed effects and synthetic controls. Review of Economics and Statistics, 98(3):535–551.

Goldsmith-Pinkham, P. and Imbens, G. W. (2013). Social networks and the identification of peer effects. Journal of Business & Economic Statistics, 31(3):253–264.

Goldsmith-Pinkham, P. S., Hull, P., and Kolesár, M. (2022). Contamination Bias in Linear Regressions. Number w30108 in NBER Working Paper Series. National Bureau of Economic Research.

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

Hahn, J., Liao, Z., and Ridder, G. (2018). Nonparametric two-step sieve m estimation and inference. Econometric Theory, 34(6):1281–1324.

Halloran, M. E. and Struchiner, C. J. (1995). Causal inference in infectious diseases:. Epidemiology, 6(2):142–151.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. Econometrica, 50:1029–1054.

Hu, Y., Li, S., and Wager, S. (2021). Average treatment effects in the presence of interference. arXiv:2104.03802 [econ, stat]. arXiv: 2104.03802.

Hudgens, M. G. and Halloran, M. E. (2008). Toward causal inference with interference. Journal of the American Statistical Association, 103(482):832–842.

Imbens, G., Kallus, N., and Mao, X. (2021). Controlling for unmeasured confounding in panel data using minimal bridge functions: From two-way fixed effects to factor models.

Juodis, A. and Sarafidis, V. (2022a). An incidental parameters free inference approach for panels with common shocks. Journal of Econometrics, 229(1):19–54.

Juodis, A. and Sarafidis, V. (2022b). A linear estimator for factor-augmented fixed-t panels with endogenous regressors. Journal of Business & Economic Statistics, 40(1):1–15.

Kejriwal, M., Li, X., and Totty, E. (2021). The efficacy of ability proxies for estimating the returns to schooling: A factor model-based evaluation.

Kitchens, C. (2014). The role of publicly provided electricity in economic development: The experience of the tennessee valley authority, 1929–1955. The Journal of Economic History, 74(2):389–419.

Kline, P. (2011). Oaxaca-blinder as a reweighting estimator. American Economic Review, 101(3):532–537.

Kline, P. and Moretti, E. (2014). Local economic development, agglomeration economies, and the big push: 100 years of evidence from the tennessee valley authority. The Quarterly Journal of Economics, 129(1):275–331.

Leeb, H. and Pötscher, B. M. (2005). Model selection and inference: Facts and fiction. Econometric Theory, 21(1):21–59.

Linden, L. and Rockoff, J. E. (2008a). Estimates of the impact of crime risk on property values from megan's laws. American Economic Review, 98(3):1103–1127.

Linden, L. and Rockoff, J. E. (2008b). Estimates of the impact of crime risk on property values from megan's laws. American Economic Review, 98(3):1103–1127.

Manski, C. F. (1993). Identification of endogenous social effects: The reflection problem. The review of economic studies, 60(3):531–542.

Manson, S. M. (2020). Ipums national historical geographic information system: Version 15.0.

Marcus, M. (2021). Going beneath the surface: Petroleum pollution, regulation, and health. American Economic Journal: Applied Economics, 13(1):72–104.

McCloskey, A. and Michaillat, P. (2020). Critical values robust to p-hacking. arXiv preprint arXiv:2005.04141.

Miguel, E. and Kremer, M. (2004). Worms: Identifying impacts on education and health in the presence of treatment externalities. Econometrica, 72(1):159–217.

Mogensen, P. and Riseth, A. (2018). Optim: A mathematical optimization package for julia. Journal of Open Source Software, 3(24).

Neumark, D. and Kolko, J. (2010). Do enterprise zones create jobs? evidence from california's enterprise zone program. Journal of Urban Economics, 68(1):1–19.

Neumark, D. and Simpson, H. (2015). Place-based policies. In Handbook of regional and urban economics, volume 5, pages 1197–1287. Elsevier.

Neumark, D. and Young, T. (2019). Enterprise zones, poverty, and labor market outcomes: Resolving conflicting evidence. Regional Science and Urban Economics, 78:103462.

Neumark, D., Zhang, J., and Ciccarella, S. (2008). The effects of wal-mart on local labor markets. Journal of Urban Economics, 63(2):405–430.

Pennington, K. (2021). Does building new housing cause displacement?: the supply and demand effects of construction in san francisco.

Pesaran, M. H. (2006). Estimation and inference in large heterogeneous panels with a multifactor error structure. Econometrica, 74(4):967–1012.

Pollmann, M. (2020). Causal inference for spatial treatments. arXiv preprint arXiv:2011.00373.

Rai, B. (2023). Efficient estimation with missing data and endogeneity. Econometric Reviews, 42(2):220–239.

Rambachan, A. and Roth, J. (2023). A more credible approach to parallel trends. Review of Economic Studies, page rdad018.

Sant'Anna, P. H. and Zhao, J. (2020). Doubly robust difference-in-differences estimators. Journal of econometrics, 219(1):101–122.

Severnini, E. (2023). The power of hydroelectric dams: Historical evidence from the united states over the twentieth century. The Economic Journal, 133(649):420–459.

Sobel, M. E. (2006). What do randomized studies of housing mobility demonstrate?: Causal inference in the face of interference. Journal of the American Statistical Association, 101(476):1398–1407.

Stapp, J. (2014). The walmart effect: Labor market implications in rural and urban counties. SS-AAEA Journal of Agricultural Economics, 2014(318-2016-9525).

Sullivan, D. M. (2017). The true cost of air pollution: Evidence from the housing market. Unpublished working paper.

Sun, L. and Abraham, S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. Journal of econometrics, 225(2):175–199.

Sävje, F., Aronow, P. M., and Hudgens, M. G. (2021). Average treatment effects in the presence of unknown interference. The Annals of Statistics, 49(2).

Vazquez-Bare, G. (2023). Identification and estimation of spillover effects in randomized experiments. Journal of Econometrics, 237(1):105237.

Verbitsky-Savitz, N. and Raudenbush, S. W. (2012a). Causal inference under interference in spatial settings: A case study evaluating community policing program in chicago. Epidemiologic Methods, 1(1).

Verbitsky-Savitz, N. and Raudenbush, S. W. (2012b). Causal inference under interference in spatial settings: a case study evaluating community policing program in chicago. Epidemiologic Methods, 1(1):107–130.

Volpe, R. and Boland, M. A. (2022). The economic impacts of walmart supercenters. Annual Review of Resource Economics, 14:43–62.

Wang, Y., Samii, C., Chang, H., and Aronow, P. M. (2020). Design-based inference for spatial experiments under unknown interference. arXiv preprint arXiv:2010.13599.

Westerlund, J. (2020). A cross-section average-based principal components approach for fixed-t panels. Journal of Applied Econometrics, 35(6):776–785.

Westerlund, J., Petrova, Y., and Norkutė, M. (2019). Cce in fixed-t panels. Journal of Applied Econometrics, 34:746–761.

Windmeijer, F. (2005). A finite sample correction for the variance of linear efficient two-step gmm estimators. Journal of econometrics, 126(1):25–51.

Wooldridge, J. M. (2010). Econometric analysis of cross section and panel data. MIT press.

Wooldridge, J. M. (2021). Two-way fixed effects, the two-way mundlak regression, and difference-in-differences estimators.

Xu, R. (2023). Difference-in-differences with interference: A finite population perspective. arXiv preprint arXiv:2306.12003.

Xu, Y. (2017). Generalized synthetic control method: Causal inference with interactive fixed effects models. Political Analysis, 25(1):57–76.

Yu, C. L., Airoldi, E. M., Borgs, C., and Chayes, J. T. (2022). Estimating the total treatment effect in randomized experiments with unknown network structure. Proceedings of the National Academy of Sciences, 119(44):e2208975119.

Appendix A

Appendix to "Dynamic Treatment Effect Estimation with Interactive Fixed Effects and Short Panels"

A.1 Proofs

A.1.1 Proof of Theorem 1.1

Let $t \geq g$ for the given group g.

$$\mathbb{E}\big[y_{it} - \boldsymbol{P}(\boldsymbol{F}_t', \boldsymbol{F}_{t<g})\boldsymbol{y}_{i,t<g} \mid G_i = g\big] = \mathbb{E}[y_{it}(1) \mid G_i = g] - \mathbb{E}\big[\boldsymbol{P}(\boldsymbol{F}_t', \boldsymbol{F}_{t<g})\boldsymbol{y}_{i,t<g} \mid G_i = g\big]$$

We use the fact that

$$\begin{aligned}\mathbb{E}\big[\boldsymbol{P}(\boldsymbol{F}_t', \boldsymbol{F}_{t<g})\boldsymbol{y}_{i,t<g} \mid G_i = g\big] &= \mathbb{E}\big[\boldsymbol{F}_t'(\boldsymbol{F}_{t<g}'\boldsymbol{F}_{t<g})^{-1}\boldsymbol{F}_{t<g}'\boldsymbol{y}_{i,t<g} \mid G_i = g\big] \\ &= \mathbb{E}\big[\boldsymbol{F}_t'(\boldsymbol{F}_{t<g}'\boldsymbol{F}_{t<g})^{-1}\boldsymbol{F}_{t<g}'\big[\boldsymbol{F}_{t<g}\boldsymbol{\gamma}_i + u_{i,t<g}\big] \mid G_i = g\big] \\ &= \mathbb{E}\big[\boldsymbol{F}_t'\boldsymbol{\gamma}_i + \boldsymbol{F}_t'(\boldsymbol{F}_{t<g}'\boldsymbol{F}_{t<g})^{-1}\boldsymbol{F}_{t<g}'u_{i,t<g} \mid G_i = g\big] \\ &= \mathbb{E}[y_{it}(\infty) \mid G_i = g]\end{aligned}$$

The second equality hold by Assumption 2 and the fact that $y_{i,t<g} = y_{i,t<g}(0)$. The final equality holds by Assumption 2.

For the second part of the theorem, note that from the column span condition, there exists a $m \times p$ matrix $\boldsymbol{A}$ such that

$$\boldsymbol{F}^*\boldsymbol{A} = \boldsymbol{F} \tag{A.1}$$

$\boldsymbol{A}$ defines the linear combinations of the columns of $\boldsymbol{F}^*$ that span the columns of $\boldsymbol{F}$. Thus $\boldsymbol{F}_t^{*'}\boldsymbol{A} = \boldsymbol{F}_t'$. We then have

$$\begin{aligned}
\boldsymbol{F}_t^{*'}(\boldsymbol{F}_{t<g}^{*'}\boldsymbol{F}_{t<g}^{*'})^{-1}\boldsymbol{F}_{t<g}^{*'}\boldsymbol{F}_{t<g}\boldsymbol{\gamma}_i &= \boldsymbol{F}_t^{*'}(\boldsymbol{F}_{t<g}^{*'}\boldsymbol{F}_{t<g}^{*})^{-1}\boldsymbol{F}_{t<g}^{*'}\boldsymbol{F}_{t<g}^{*'}\boldsymbol{A}\boldsymbol{\gamma}_i \\
&= \boldsymbol{F}_t^{*'}\boldsymbol{A}\boldsymbol{\gamma}_i \\
&= \boldsymbol{F}_t^{*'}\boldsymbol{\gamma}_i
\end{aligned}$$

If $m = p$ so that $\boldsymbol{F}$ also has full column rank, we can make the stronger statement that the imputation matrices of $\boldsymbol{F}$ and $\boldsymbol{F}^*$ are equal:

$$\begin{aligned}
\boldsymbol{P}(\boldsymbol{F}_{t\geq g}, \boldsymbol{F}_{t<g}) &= \boldsymbol{F}_{t\geq g}(\boldsymbol{F}_{t<g}'\boldsymbol{F}_{t<g})^{-1}\boldsymbol{F}_{t<g}' \\
&= \boldsymbol{F}_{t\geq g}\boldsymbol{A}(\boldsymbol{A}'\boldsymbol{F}_{t<g}'\boldsymbol{F}_{t<g}\boldsymbol{A})^{-1}\boldsymbol{A}'\boldsymbol{F}_{t<g}' \\
&= \boldsymbol{F}_{t\geq g}^{*'}(\boldsymbol{F}_{t<g}^{*'}\boldsymbol{F}_{t<g}^{*})^{-1}\boldsymbol{F}_{t<g}^{*'} \\
&= \boldsymbol{P}(\boldsymbol{F}_{t\geq g}^{*}, \boldsymbol{F}_{t<g}^{*})
\end{aligned}$$

where the second equality holds because $\boldsymbol{A}$ and $(\boldsymbol{F}_{t<g}'\boldsymbol{F}_{t<g})$ are full rank.

□

A.1.2 Proof of Lemma 1.1

We first derive the averages defined in Section 2.2 in terms of the potential outcome framework:

$$\begin{aligned}
\overline{y}_{\infty,t} &= \frac{1}{N_\infty}\sum_{i=1}^{N} D_{i\infty}y_{it} = \overline{\mu}_\infty + \lambda_t + \boldsymbol{F}_t\overline{\boldsymbol{\gamma}}_\infty + \overline{u}_{t,\infty} \\
\overline{y}_{i,t\leq T_0} &= \frac{1}{T_0}\sum_{t=1}^{T_0} y_{it} = \mu_i + \overline{\lambda}_{t<T_0} + \overline{\boldsymbol{F}}_{t<T_0}\boldsymbol{\gamma}_i + \overline{u}_{i,t<T_0} \\
\overline{y}_{\infty,t<T_0} &= \frac{1}{N_\infty T_0}\sum_{i=1}^{N}\sum_{t=1}^{T_0} D_{i\infty}y_{it} = \overline{\mu}_\infty + \overline{\lambda}_{t<T_0} + \overline{\boldsymbol{F}}_{t<T_0}\overline{\boldsymbol{\gamma}}_\infty + \overline{u}_{\infty,t<T_0}
\end{aligned}$$

where $\overline{\mu}_\infty$ and $\overline{\boldsymbol{\gamma}}_\infty$ are the averages of the never-treated individuals' heterogeneity and $\overline{\boldsymbol{F}}_{t<T_0}$ and $\overline{\lambda}_{t<T_0}$ are the averages of the time effects before anyone is treated. The error averages have the same interpretation as the outcome averages.

The definition of τ_{it} is the difference between treated and untreated potential outcomes for unit i at time t, so for any (i,t), $y_{it} = d_{it}y_{it}(1) + (1-d_{it})y_{it}(\infty) = d_{it}\tau_{it} + y_{it}(\infty)$. Then

$$\begin{aligned}\tilde{y}_{it} &= d_{it}\tau_{it} + \boldsymbol{F}_t'\boldsymbol{\gamma}_i - \overline{\boldsymbol{F}}_{t<T_0}'\boldsymbol{\gamma}_i - \boldsymbol{F}_t'\overline{\boldsymbol{\gamma}}_\infty + \overline{\boldsymbol{F}}_{t<T_0}\overline{\boldsymbol{\gamma}}_\infty + u_{it} - \overline{u}_{t,\infty} - \overline{u}_{i,t<T_0} + \overline{u}_{\infty,t<T_0} \\ &= d_{it}\tau_{it} + (\boldsymbol{F}_t - \overline{\boldsymbol{F}}_{t<T_0})'(\boldsymbol{\gamma}_i - \overline{\boldsymbol{\gamma}}_\infty) + u_{it} - \overline{u}_{t,\infty} - \overline{u}_{i,t<T_0} + \overline{u}_{\infty,t<T_0}\end{aligned}$$

Taking expectation conditional on $G_i = g$ gives $\mathbb{E}[u_{it} - \overline{u}_{i,t<T_0} \mid G_i = g] = 0$ by Assumption 2 and $\mathbb{E}[\overline{u}_{\infty,t<T_0} - \overline{u}_{t,\infty} \mid G_i = g] = \mathbb{E}[\overline{u}_{\infty,t<T_0} - \overline{u}_{t,\infty}] = 0$ by random sampling and iterated expectations.

□

A.1.3 Proof of Theorem 1.1

We can appeal to standard large sample GMM theory as in Hansen (1982) due to the types of first-stage factor estimators we consider. We do not consider true "fixed effects" estimators where the number of parameters grows with the sample size. The IV and cross-sectional averages approaches are based on eliminating the factors (which are fixed in the asymptotic analysis) by reducing them to a smaller set of parameters. For example, while the CCE estimator can be implemented as a pooled regression where unit dummies are interacted with cross-sectional averages, the estimator itself takes a form similar to the within transformation in the linear fixed effects model. In fact, we prove asymptotic unbiasedness of dynamic ATT estimators using the CCE estimator in the first stage (Brown et al., 2023a)[1] .

Consider the QLD estimator of Ahn et al. (2013). They study the linear model

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{F}\boldsymbol{\gamma}_i + \boldsymbol{\epsilon}_i \tag{A.2}$$

They jointly estimate the QLD parameters $\boldsymbol{\theta}$ along with the conditional response parameters $\boldsymbol{\beta}$ using the moment conditions

$$\mathbb{E}[\boldsymbol{H}(\boldsymbol{\theta})(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}) \otimes \boldsymbol{w}_i] = \boldsymbol{0} \tag{A.3}$$

[1] We consider CCE in a separate paper because the additional modeling assumptions allow for stronger results than those considered in this paper.

They show that the estimator is well-behaved and does not suffer from asymptotic bias. As described in Windmeijer (2005), the most likely source of finite-sample bias comes from estimating the optimal weight matrix. The appendix of Ahn et al. (2013) describes a continuous updating estimator (CUE) based on their moment conditions, which may have less finite-sample bias than the optimal two-step estimator. However, we may also sacrifice efficiency in large samples if their assumed covariance structure is incorrect.

We now derive the asymptotic variance of the full estimator under a general first-step estimator of the factors. Note that $\boldsymbol{g}_{i\infty}(\boldsymbol{\theta}) \otimes \boldsymbol{g}_{ig}(\boldsymbol{\theta}, \boldsymbol{\tau}_g) = \mathbf{0}$ (from the D_{ig} terms) and $\boldsymbol{g}_{ih}(\boldsymbol{\theta}, \boldsymbol{\tau}_h) \otimes \boldsymbol{g}_{ik}(\boldsymbol{\theta}, \boldsymbol{\tau}_k) = \mathbf{0}$ almost surely uniformly over the parameter space for all $g \in \mathcal{G}$ and $h \neq k$. The covariance matrix of these moment functions, which we denote as $\boldsymbol{\Delta}$, is a block diagonal matrix.

$$\boldsymbol{\Delta} = \begin{pmatrix} \mathbb{E}[\boldsymbol{g}_{i\infty}(\boldsymbol{\theta})\boldsymbol{g}_{i\infty}(\boldsymbol{\theta})'] & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbb{E}[\boldsymbol{g}_{ig_G}(\boldsymbol{\theta}, \boldsymbol{\tau}_{g_G})\boldsymbol{g}_{ig_G}(\boldsymbol{\theta}, \boldsymbol{\tau}_{g_G})'] & \mathbf{0} & \dots & \mathbf{0} \\ \vdots & & \ddots & & \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \mathbb{E}[\boldsymbol{g}_{ig_1}(\boldsymbol{\theta}, \boldsymbol{\tau}_{g_1})\boldsymbol{g}_{ig_1}(\boldsymbol{\theta}, \boldsymbol{\tau})'] \end{pmatrix}$$

We write the individual blocks as $\boldsymbol{\Delta}_g$ for $g \in \mathcal{G} \cup \{\infty\}$. The gradient is also simple to compute because all of the moments are linear in the treatment effects. We define the overall gradient $\boldsymbol{D}$ and show it is a lower triangular matrix which we write in terms of its constituent blocks:

$$\boldsymbol{D} = \begin{pmatrix} \mathbb{E}[\nabla_{\boldsymbol{\theta}}\boldsymbol{g}_{i\infty}(\boldsymbol{\theta})] & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbb{E}[\nabla_{\boldsymbol{\theta}}\boldsymbol{g}_{ig_G}(\boldsymbol{\theta}, \boldsymbol{\tau}_{g_G})] & -\boldsymbol{I}_{T-g_G+1} & \mathbf{0} & \dots & \mathbf{0} \\ \vdots & & \ddots & & \\ \mathbb{E}[\nabla_{\boldsymbol{\theta}}\boldsymbol{g}_{ig_1}(\boldsymbol{\theta}, \boldsymbol{\tau}_{g_1})] & \mathbf{0} & \mathbf{0} & \dots & -\boldsymbol{I}_{T-g_1+1} \end{pmatrix}$$

where we write the blocks in the first column as $\boldsymbol{D}_g$ for $g \in \mathcal{G} \cup \{\infty\}$. The diagonal is made up of negative identity matrices because $\mathbb{E}\left[\frac{D_{ig_h}}{\mathbb{P}(D_{ig_h}=1)}\right] = 1$.

The overall asymptotic variance given that we use the optimal weight matrix is given by

$(\boldsymbol{D}'\boldsymbol{\Delta}^{-1}\boldsymbol{D})^{-1}$. $\boldsymbol{\Delta}$ is a block diagonal matrix so its inverse is trivial to compute. First, we have

$$\boldsymbol{\Delta}^{-1}\boldsymbol{D} = \begin{pmatrix} \boldsymbol{\Delta}_{\infty}^{-1}\boldsymbol{D}_{\infty} & \mathbf{0} & \dots & \mathbf{0} \\ \boldsymbol{\Delta}_{g_G}^{-1}\boldsymbol{D}_{g_G} & -\boldsymbol{\Delta}_{g_G}^{-1} & \dots & \mathbf{0} \\ \vdots & & \ddots & \\ \boldsymbol{\Delta}_{g_1}^{-1}\boldsymbol{D}_{g_1} & \mathbf{0} & \dots & -\boldsymbol{\Delta}_{g_1}^{-1} \end{pmatrix}$$

The transpose of the gradient matrix is

$$\boldsymbol{D}' = \begin{pmatrix} \boldsymbol{D}'_{\infty} & \boldsymbol{D}'_{g_G} & \dots & \boldsymbol{D}'_{g_1} \\ \mathbf{0} & -\boldsymbol{I}_{T-g_G+1} & \dots & \mathbf{0} \\ \vdots & & \ddots & \\ \mathbf{0} & \mathbf{0} & \dots & -\boldsymbol{I}_{T-g_1+1} \end{pmatrix}$$

so that we get

$$\boldsymbol{D}'\boldsymbol{\Delta}^{-1}\boldsymbol{D} = \begin{pmatrix} \sum_{g\in\mathcal{G}\cup\{\infty\}} \boldsymbol{D}'_g\boldsymbol{\Delta}_g^{-1}\boldsymbol{D}_g & -\boldsymbol{D}'_{g_G}\boldsymbol{\Delta}_{g_G}^{-1} & \dots & -\boldsymbol{D}'_{g_1}\boldsymbol{\Delta}_{g_G}^{-1} \\ -\boldsymbol{\Delta}_{g_G}^{-1}\boldsymbol{D}_{g_G} & \boldsymbol{\Delta}_{g_G}^{-1} & \dots & \mathbf{0} \\ \vdots & & \ddots & \\ -\boldsymbol{\Delta}_{g_1}^{-1}\boldsymbol{D}_{g_1} & \mathbf{0} & \dots & \boldsymbol{\Delta}_{g_1}^{-1} \end{pmatrix}$$

We write this matrix as

$$\begin{pmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{pmatrix}$$

where $\boldsymbol{A} = \sum_{g\in\mathcal{G}\cup\{\infty\}} \boldsymbol{D}'_g\boldsymbol{\Delta}_g^{-1}\boldsymbol{D}_g$ and $\boldsymbol{D} = \text{diag}\{\boldsymbol{\Delta}_g^{-1}\}_{g\in\mathcal{G}}$. We then apply Exercise 5.16 of Abadir and Magnus (2005) to get the final inverse. The top left corner of the inverse is $\boldsymbol{F}^{-1}$ where

$$\begin{aligned}
(\boldsymbol{F})^{-1} &= (\boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C})^{-1} \\
&= \left(\sum_{g \in \mathcal{G} \cup \{\infty\}} \boldsymbol{D}_g' \boldsymbol{\Delta}_g^{-1} \boldsymbol{D}_g - \left(\sum_{g \in \mathcal{G}} \boldsymbol{D}_g' \boldsymbol{\Delta}_g^{-1} \boldsymbol{D}_g \right) \right)^{-1} \\
&= (\boldsymbol{D}_\infty' \boldsymbol{\Delta}_\infty^{-1} \boldsymbol{D}_\infty)^{-1} \\
&= \text{Avar}(\sqrt{N}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}))
\end{aligned}$$

The rest of the first column of matrices takes the form

$$\begin{aligned}
-\boldsymbol{D}^{-1}\boldsymbol{C}\boldsymbol{F}^{-1} &= \begin{pmatrix} \boldsymbol{D}_{g_G} \\ \vdots \\ \boldsymbol{D}_{g_1} \end{pmatrix} (\boldsymbol{D}_\infty' \boldsymbol{\Delta}_\infty^{-1} \boldsymbol{D}_\infty)^{-1} \\
&= \begin{pmatrix} \boldsymbol{D}_{g_G} (\boldsymbol{D}_\infty' \boldsymbol{\Delta}_\infty^{-1} \boldsymbol{D}_\infty)^{-1} \\ \vdots \\ \boldsymbol{D}_{g_1} (\boldsymbol{D}_\infty' \boldsymbol{\Delta}_\infty^{-1} \boldsymbol{D}_\infty)^{-1} \end{pmatrix}
\end{aligned}$$

and the rest of the first row is $-\boldsymbol{F}^{-1}\boldsymbol{B}\boldsymbol{D}^{-1} = (-\boldsymbol{D}^{-1}\boldsymbol{B}'\boldsymbol{F}^{-1})' = (-\boldsymbol{D}^{-1}\boldsymbol{C}\boldsymbol{F}^{-1})'$.

Finally, the bottom-right block, which also gives the asymptotic covariance matrix of the ATT estimators, is

$$\boldsymbol{D}^{-1} + \boldsymbol{D}^{-1}\boldsymbol{C}\boldsymbol{F}^{-1}\boldsymbol{B}\boldsymbol{D}^{-1} = \boldsymbol{D}^{-1} + \begin{pmatrix} \boldsymbol{D}_{g_G} (\boldsymbol{D}_\infty' \boldsymbol{\Delta}_\infty^{-1} \boldsymbol{D}_\infty)^{-1} \boldsymbol{D}_{g_G}' & \dots & \boldsymbol{D}_{g_G} (\boldsymbol{D}_\infty' \boldsymbol{\Delta}_\infty^{-1} \boldsymbol{D}_\infty)^{-1} \boldsymbol{D}_{g_1}' \\ & \ddots & \\ \boldsymbol{D}_{g_1} (\boldsymbol{D}_\infty' \boldsymbol{\Delta}_\infty^{-1} \boldsymbol{D}_\infty)^{-1} \boldsymbol{D}_{g_G}' & \dots & \boldsymbol{D}_{g_1} (\boldsymbol{D}_\infty' \boldsymbol{\Delta}_\infty^{-1} \boldsymbol{D}_\infty)^{-1} \boldsymbol{D}_{g_1}' \end{pmatrix}$$

The g'th diagonal elements of the resulting matrix is $\boldsymbol{\Delta}_g + \boldsymbol{D}_g (\boldsymbol{D}_\infty' \boldsymbol{\Delta}_\infty^{-1} \boldsymbol{D}_\infty)^{-1} \boldsymbol{D}_g'$.

□

A.1.4 Proof of Theorem 1.2

We derive the limiting theory by multiplying $\widehat{\boldsymbol{\Delta}}_g$ by $(N_g-1)/N_g$ which produces the same limit as $N \to \infty$. We write

$$\frac{N_g-1}{N_g}\widehat{\boldsymbol{\Delta}}_g = \frac{1}{N_g}\sum_{i=1}^{N} D_{ig}\widehat{\boldsymbol{\Delta}}_{ig}\widehat{\boldsymbol{\Delta}}_{ig}' - \widehat{\boldsymbol{\tau}}_g\widehat{\boldsymbol{\tau}}_g'$$

We already know that $\widehat{\boldsymbol{\tau}}_g \operatorname{plim} \boldsymbol{\tau}_g$ by Theorem 3.1. Note that

$$\begin{aligned}\frac{1}{N_g}\sum_{i=1}^{N} D_{ig}\widehat{\boldsymbol{\Delta}}_{ig}\widehat{\boldsymbol{\Delta}}_{ig}' = &\left(\frac{1}{N_g}\sum_{i=1}^{N} D_{ig}\boldsymbol{y}_{i,t\geq g}\boldsymbol{y}_{i,t\geq g}'\right) - \left(\frac{1}{N_g}\sum_{i=1}^{N} D_{ig}\boldsymbol{y}_{i,t\geq g}\boldsymbol{y}_{i,t<g}'\right)\boldsymbol{P}(\boldsymbol{F}_{t\geq g}(\widehat{\boldsymbol{\theta}}),\boldsymbol{F}_{t<g}(\widehat{\boldsymbol{\theta}}))' \\ &- \boldsymbol{P}(\boldsymbol{F}_{t\geq g}(\widehat{\boldsymbol{\theta}}),\boldsymbol{F}_{t<g}(\widehat{\boldsymbol{\theta}}))\left(\frac{1}{N_g}\sum_{i=1}^{N} D_{ig}\boldsymbol{y}_{i,t<g}\boldsymbol{y}_{i,t\geq g}'\right) \\ &- \boldsymbol{P}(\boldsymbol{F}_{t\geq g}(\widehat{\boldsymbol{\theta}}),\boldsymbol{F}_{t<g}(\widehat{\boldsymbol{\theta}}))\left(\frac{1}{N_g}\sum_{i=1}^{N} D_{ig}\boldsymbol{y}_{i,t<g}\boldsymbol{y}_{i,t\geq g}'\right)\boldsymbol{P}(\boldsymbol{F}_{t\geq g}(\widehat{\boldsymbol{\theta}}),\boldsymbol{F}_{t<g}(\widehat{\boldsymbol{\theta}}))'\end{aligned}$$

Given $\boldsymbol{P}(\boldsymbol{F}_{t\geq g}(\widehat{\boldsymbol{\theta}}),\boldsymbol{F}_{t<g}(\widehat{\boldsymbol{\theta}}))$ is equal to its infeasible counterpart $\boldsymbol{P}(\boldsymbol{F}_{t\geq g},\boldsymbol{F}_{t<g})$ plus a term that is $O_p(N^{-1/2})$, Assumption 1 and the weak law of large numbers imply

$$\frac{1}{N_g}\sum_{i=1}^{N} D_{ig}\widehat{\boldsymbol{\Delta}}_{ig}\widehat{\boldsymbol{\Delta}}_{ig}' - \widehat{\boldsymbol{\tau}}_g\widehat{\boldsymbol{\tau}}_g' \operatorname{plim} \mathbb{E}[\boldsymbol{g}_{ig}(\boldsymbol{\theta},\boldsymbol{\tau}_g) \mid G_i = g] = \boldsymbol{\Delta}_g$$

The inverse exists with probability approaching one by Assumption 5.

□

A.2 The Quasi-Long-Differencing Estimator

We discuss identification and inference of the imputation estimator using the QLD estimator for the factors. We derive the results here because QLD is used in both the simulations and application.

A.2.1 Identification

We adapt the identifying assumptions from Ahn et al. (2013) to our setup, guaranteeing Assumption 1.4 holds. Part (i) of the assumption holds assuming that $\boldsymbol{F}$ is full rank, because there always exists a matrix that applies Gaussian row-reduction to a full rank matrix. For parts (ii) and (iii), we need the following matrix to be full rank:

$$\boldsymbol{I}_{T-p} \otimes \mathbb{E}\big[\boldsymbol{w}_i \boldsymbol{\gamma}_i' \mid G_i = \infty\big] \tag{A.4}$$

It implies that the instruments $\boldsymbol{w}_i$ are "strong" in the sense that they correlate with the factor loadings $\boldsymbol{\gamma}_i$. Unfortunately, this restriction is not easily testable like in the case of two-stage least squares because the variable being instrumented for is unobserved. We leave the question of testing for instrument strength in quasi-differencing for a future project. Part (iv) is an additional assumption that is routinely made in practice.

A.2.2 Asymptotic Variance

We now derive the analytical formulas for the asymptotic variance when quasi-differencing is used to estimate the factor space. Analytical standard errors can be obtained by replacing the population parameters with their estimators and expectations with the relevant sample average, e.g. expectations of the never-treated group are estimated using the average of the never-treated subsample. Conversely, one can average over the entire sample but multiply each observation by $D_{i\infty}$ and divide by N_∞/N. To get the gradient of the set of moment conditions that identify the factor space, we rewrite the moment function as

$$\begin{aligned}\boldsymbol{H}(\boldsymbol{\theta})\boldsymbol{y}_i \otimes \boldsymbol{w}_i &= \text{vec}(\boldsymbol{w}_i \boldsymbol{y}_i' \boldsymbol{H}(\boldsymbol{\theta})') \\ &= (\boldsymbol{I}_{(T-p)} \otimes \boldsymbol{w}_i \boldsymbol{y}_i') \boldsymbol{K}_{(T-p)T} \text{vec}(\boldsymbol{H}(\boldsymbol{\theta}))\end{aligned}$$

where $\boldsymbol{K}_{(T-p)T}$ is the $(T-p)T \times (T-p)T$ commutation matrix and we use the well-known relationship between vectorization and the Kronecker product[2] . Because $\text{vec}(\boldsymbol{H}(\boldsymbol{\theta})) = [\text{vec}(\boldsymbol{I}_{T-p})', \boldsymbol{\theta}']'$, the gradient of the moment function is

$$\left(\boldsymbol{I}_{(T-p)} \otimes \boldsymbol{w}_i \boldsymbol{y}_i'\right) \boldsymbol{K}_{T(T-p)} [\boldsymbol{0}'_{(T-p)^2 \times (T-p)p}, \boldsymbol{I}_{(T-p)p}]' \tag{A.5}$$

[2] See Exercise 10.18 of Abadir and Magnus (2005).

The expected gradient is obtained by taking expectations conditional on being in the never-treated group.

We now consider the gradient of the moment functions that determine the treatment effects with respect to the factor estimator for a given group treated at time g. The relevant part of the moment function for the purpose of finding the gradient is

$$\boldsymbol{F}_{t\geq g}(\boldsymbol{\theta})' \left(\boldsymbol{F}_{t<g}(\boldsymbol{\theta})'\boldsymbol{F}_{t<g}(\boldsymbol{\theta})\right)^{-1} \boldsymbol{F}_{t<g}(\boldsymbol{\theta})'\boldsymbol{y}_{i,t<g} \tag{A.6}$$

There are two leading cases to compute: $g-1 \geq T-p$ and $g-1 < T-p$. In the first case, the parameters $\boldsymbol{\theta}$ are entirely contained in the pre-treatment factor matrix. Then

$$\boldsymbol{F}_{t<g} = \begin{pmatrix} \boldsymbol{\Theta} \\ \\ \boldsymbol{E} \end{pmatrix} \tag{A.7}$$

where $\boldsymbol{E}$ is the first $(g-1)-(T-p)$ rows of $-\boldsymbol{I}_p$. Then the post-treatment factor matrix is just the lower $T-g+1$ rows of $-\boldsymbol{I}_p$ so we do not need to worry about differentiating it. In this setting,

$$\boldsymbol{F}_{t\geq g}\left(\boldsymbol{F}'_{t<g}\boldsymbol{F}_{t<g}\right)^{-1}\boldsymbol{F}'_{t<g}\boldsymbol{y}_{i,t<g} = -\left(\boldsymbol{\Theta}'\boldsymbol{\Theta} + \boldsymbol{E}'\boldsymbol{E}\right)^{-1}\begin{pmatrix}\boldsymbol{\Theta}' & \boldsymbol{E}'\end{pmatrix}\boldsymbol{y}_{i,t<g} \tag{A.8}$$

We use the notation in Chapter 13 of Abadir and Magnus (2005) to obtain the differential:

$$-\left(\boldsymbol{\Theta}'\boldsymbol{\Theta} + \boldsymbol{E}'\boldsymbol{E}\right)^{-1}\begin{pmatrix}(d\boldsymbol{\Theta})' & \boldsymbol{E}'\end{pmatrix}\boldsymbol{y}_{i,t<g} \tag{A.9}$$

$$\left(\boldsymbol{\Theta}'\boldsymbol{\Theta} + \boldsymbol{E}'\boldsymbol{E}\right)^{-1}\left((d\boldsymbol{\Theta})'\boldsymbol{\Theta}\right)\left(\boldsymbol{\Theta}'\boldsymbol{\Theta} + \boldsymbol{E}'\boldsymbol{E}\right)^{-1}\begin{pmatrix}\boldsymbol{\Theta}' & \boldsymbol{E}'\end{pmatrix}\boldsymbol{y}_{i,t<g} \tag{A.10}$$

$$\left(\boldsymbol{\Theta}'\boldsymbol{\Theta} + \boldsymbol{E}'\boldsymbol{E}\right)^{-1}\left(\boldsymbol{\Theta}'(d\boldsymbol{\Theta})\right)\left(\boldsymbol{\Theta}'\boldsymbol{\Theta} + \boldsymbol{E}'\boldsymbol{E}\right)^{-1}\begin{pmatrix}\boldsymbol{\Theta}' & \boldsymbol{E}'\end{pmatrix}\boldsymbol{y}_{i,t<g} \tag{A.11}$$

which can then be rewritten as

$$-\left(\boldsymbol{y}_{i,t<g}\otimes\left(\boldsymbol{\Theta}'\boldsymbol{\Theta} + \boldsymbol{E}'\boldsymbol{E}\right)^{-1}\right)\begin{pmatrix}\boldsymbol{K}_{(T-p)p}(d\boldsymbol{\theta})' & \boldsymbol{K}_{((g-1)-(T-p)p}\text{vec}(\boldsymbol{E})'\end{pmatrix}' \tag{A.12}$$

$$\left(\left(\boldsymbol{\Theta}\left(\boldsymbol{\Theta}'\boldsymbol{\Theta} + \boldsymbol{E}'\boldsymbol{E}\right)^{-1}\begin{pmatrix}\boldsymbol{\Theta}' & \boldsymbol{E}'\end{pmatrix}\boldsymbol{y}_{i,t<g}\right)'\otimes\left(\boldsymbol{\Theta}'\boldsymbol{\Theta} + \boldsymbol{E}'\boldsymbol{E}\right)^{-1}\right)\boldsymbol{K}_{(T-p)p}d\boldsymbol{\theta} \tag{A.13}$$

$$\left(\left(\left(\boldsymbol{\Theta}'\boldsymbol{\Theta} + \boldsymbol{E}'\boldsymbol{E}\right)^{-1}\begin{pmatrix}\boldsymbol{\Theta}' & \boldsymbol{E}'\end{pmatrix}\boldsymbol{y}_{i,t<g}\right)'\otimes\left(\boldsymbol{\Theta}'\boldsymbol{\Theta} + \boldsymbol{E}'\boldsymbol{E}\right)^{-1}\boldsymbol{\Theta}'\right)d\boldsymbol{\theta} \tag{A.14}$$

The full gradient is then

$$-\left(\boldsymbol{y}_{i,t<g}\otimes\left(\boldsymbol{\Theta}'\boldsymbol{\Theta}+\boldsymbol{E}'\boldsymbol{E}\right)^{-1}\right)\begin{pmatrix}\boldsymbol{K}'_{(T-p)p} & \boldsymbol{0}'_{((g-1)-(T-p)p\times(T-p)p}\end{pmatrix}' \tag{A.15}$$

$$\left(\left(\boldsymbol{\Theta}\left(\boldsymbol{\Theta}'\boldsymbol{\Theta}+\boldsymbol{E}'\boldsymbol{E}\right)^{-1}\begin{pmatrix}\boldsymbol{\Theta}' & \boldsymbol{E}'\end{pmatrix}\boldsymbol{y}_{i,t<g}\right)'\otimes\left(\boldsymbol{\Theta}'\boldsymbol{\Theta}+\boldsymbol{E}'\boldsymbol{E}\right)^{-1}\right)\boldsymbol{K}_{(T-p)p} \tag{A.16}$$

$$\left(\left(\left(\boldsymbol{\Theta}'\boldsymbol{\Theta}+\boldsymbol{E}'\boldsymbol{E}\right)^{-1}\begin{pmatrix}\boldsymbol{\Theta}' & \boldsymbol{E}'\end{pmatrix}\boldsymbol{y}_{i,t<g}\right)'\otimes\left(\boldsymbol{\Theta}'\boldsymbol{\Theta}+\boldsymbol{E}'\boldsymbol{E}\right)^{-1}\boldsymbol{\Theta}'\right) \tag{A.17}$$

when $g-1\geq T-p$.

The second case, when $g-1<T-p$, now has parameters in the post-treatment matrix $\boldsymbol{F}_{t\geq g}$. We redefine the parameters as $\boldsymbol{\Theta}=[\boldsymbol{\Theta}_1',\boldsymbol{\Theta}_2']'$ where $\boldsymbol{\Theta}_1$ is $(g-1)\times p$ and $\boldsymbol{\Theta}_2$ is $(T-p-g+1)\times p$. Now we write $\boldsymbol{F}_{t<g}=\boldsymbol{\Theta}_1$ and

$$\boldsymbol{F}_{t\geq g}=\begin{pmatrix}\boldsymbol{\Theta}_2\\ \\ -\boldsymbol{I}_p\end{pmatrix} \tag{A.18}$$

Because $\boldsymbol{\theta}\neq(\text{vec}(\boldsymbol{\Theta}_1)',\text{vec}(\boldsymbol{\Theta}_2)')'$, we define the matrices $\boldsymbol{E}_1=[\boldsymbol{I}_{g-1},\boldsymbol{0}_{(g-1)\times(T-p-g+1)}$ and $\boldsymbol{E}_2=[\boldsymbol{0}_{(T-p-g+1)\times(g-1)},\boldsymbol{I}_{(T-p-g+1)}]$ such that

$$\boldsymbol{\Theta}_1=\boldsymbol{E}_1\boldsymbol{\Theta} \tag{A.19}$$

$$\boldsymbol{\Theta}_2=\boldsymbol{E}_2\boldsymbol{\Theta} \tag{A.20}$$

Now we can rewrite the relevant portion of the moment function for the gradient as

$$\begin{pmatrix}\boldsymbol{E}_2\boldsymbol{\Theta}\\ \\ -\boldsymbol{I}_p\end{pmatrix}\left(\boldsymbol{\Theta}'\boldsymbol{E}_1'\boldsymbol{E}_1\boldsymbol{\Theta}\right)^{-1}\boldsymbol{\Theta}'\boldsymbol{E}_1'\boldsymbol{y}_{i,t<g} \tag{A.21}$$

We can now take the gradient with respect to the full set of parameters $\boldsymbol{\theta}$:

$$\begin{pmatrix}\boldsymbol{E}_2 d\boldsymbol{\Theta}\\ \\ \boldsymbol{0}_{p\times p}\end{pmatrix}\left(\boldsymbol{F}_{t<g}'\boldsymbol{F}_{t<g}\right)^{-1}\boldsymbol{F}_{t<g}'\boldsymbol{y}_{i,t<g} \tag{A.22}$$

$$-\boldsymbol{F}_{t\geq g}\left(\boldsymbol{F}_{t<g}'\boldsymbol{F}_{t<g}\right)^{-1}d\boldsymbol{\Theta}'\boldsymbol{E}_1'\boldsymbol{F}_{t<g}\left(\boldsymbol{F}_{t<g}'\boldsymbol{F}_{t<g}\right)^{-1}\boldsymbol{F}_{t<g}'\boldsymbol{y}_{i,t<g} \tag{A.23}$$

$$-\boldsymbol{F}_{t\geq g}\left(\boldsymbol{F}_{t<g}'\boldsymbol{F}_{t<g}\right)^{-1}\boldsymbol{F}_{t<g}'\boldsymbol{E}_1 d\boldsymbol{\Theta}\left(\boldsymbol{F}_{t<g}'\boldsymbol{F}_{t<g}\right)^{-1}\boldsymbol{F}_{t<g}'\boldsymbol{y}_{i,t<g} \tag{A.24}$$

$$+\boldsymbol{F}_{t\geq g}\left(\boldsymbol{F}_{t<g}'\boldsymbol{F}_{t<g}\right)^{-1}d\boldsymbol{\Theta}'\boldsymbol{E}_1'\boldsymbol{y}_{i,t<g} \tag{A.25}$$

where we inserted $\boldsymbol{F}_{t<g}$ and $\boldsymbol{F}_{t\geq g}$ for $\boldsymbol{E}_1\boldsymbol{\Theta}$ and $\boldsymbol{E}_2\boldsymbol{\Theta}$ respectively to preserve space, noting that these matrices are actually functions of the parameters $\boldsymbol{\theta}$ and not the true, unobserved factors. We rewrite line (A.22) so we can write the differential in terms of $\boldsymbol{\theta}$:

$$\begin{pmatrix} \boldsymbol{E}_2 d\boldsymbol{\Theta} \\ \\ \boldsymbol{0}_{p\times p} \end{pmatrix} \left(\boldsymbol{F}'_{t<g}\boldsymbol{F}_{t<g}\right)^{-1} \boldsymbol{F}'_{t<g}\boldsymbol{y}_{i,t<g} = \begin{pmatrix} \boldsymbol{E}_2 \\ \\ \boldsymbol{0}_{p\times p} \end{pmatrix} d\boldsymbol{\Theta} \left(\boldsymbol{F}'_{t<g}\boldsymbol{F}_{t<g}\right)^{-1} \boldsymbol{F}'_{t<g}\boldsymbol{y}_{i,t<g} \tag{A.26}$$

$$= \left(\left(\left(\boldsymbol{F}'_{t<g}\boldsymbol{F}_{t<g}\right)^{-1} \boldsymbol{F}'_{t<g}\boldsymbol{y}_{i,t<g}\right) \otimes \begin{pmatrix} \boldsymbol{E}_2 \\ \\ \boldsymbol{0}_{p\times p} \end{pmatrix} \right) d\boldsymbol{\theta} \tag{A.27}$$

We put this expression with the others to get the final gradient:

$$= \left(\left(\boldsymbol{F}'_{t<g}\boldsymbol{F}_{t<g}\right)^{-1} \boldsymbol{F}'_{t<g}\boldsymbol{y}_{i,t<g}\right) \otimes \begin{pmatrix} \boldsymbol{E}_2 \\ \\ \boldsymbol{0}_{p\times p} \end{pmatrix} \tag{A.28}$$

$$- \left(\boldsymbol{E}'_1\boldsymbol{F}_{t<g}\left(\boldsymbol{F}'_{t<g}\boldsymbol{F}_{t<g}\right)^{-1} \boldsymbol{F}'_{t<g}\boldsymbol{y}_{i,t<g}\right)' \otimes \left(\boldsymbol{F}_{t\geq g}\left(\boldsymbol{F}'_{t<g}\boldsymbol{F}_{t<g}\right)^{-1}\right) \boldsymbol{K}_{(T-p)p} \tag{A.29}$$

$$- \left(\left(\boldsymbol{F}'_{t<g}\boldsymbol{F}_{t<g}\right)^{-1} \boldsymbol{F}'_{t<g}\boldsymbol{y}_{i,t<g}\right)' \otimes \left(\boldsymbol{F}_{t\geq g}\left(\boldsymbol{F}'_{t<g}\boldsymbol{F}_{t<g}\right)^{-1} \boldsymbol{F}'_{t<g}\boldsymbol{E}_1\right) \tag{A.30}$$

$$+ \left(\boldsymbol{y}'_{i,t<g}\boldsymbol{E}_1\right) \otimes \left(\boldsymbol{F}_{t\geq g}\left(\boldsymbol{F}'_{t<g}\boldsymbol{F}_{t<g}\right)^{-1}\right) \boldsymbol{K}_{(T-p)p} \tag{A.31}$$

A.3 Inference of Aggregate Treatment Effects

As in Callaway and Sant'Anna (2021), we can form aggregates of our group-time average treatment effects. For example, event-study type coefficients would average over the τ_{gt} where $t - g = e$ for some relative event-time e with weights proportional to group membership. Consider a general aggregate estimand δ which we define as a weighted average of $ATT(g,t)$:

$$\delta = \sum_{g\in\mathcal{G}} \sum_{t>T_0} w(g,t)\tau_{gt} \tag{A.32}$$

where the weights $w(g,t)$ are non-negative and sum to one. Table 1 of Callaway and Sant'Anna (2021) and the surrounding discussion describes various treatment effect aggregates and discuss explicit forms for the weights.

Our plug-in estimate for δ is given by $\hat{\delta} = \sum_{g \in \mathcal{G}} \sum_{t > T_0} \hat{w}(g,t)\hat{\tau}_{gt}$. Inference on this term follows directly from Corollary 2 in Callaway and Sant'Anna (2021) if we have the influence function for our τ_{gt} estimates. Rewriting our moment equations in an asymptotically linear form, we have:

$$\sqrt{N}\Big((\widehat{\boldsymbol{\theta}}', \widehat{\boldsymbol{\tau}}')' - (\boldsymbol{\theta}', \boldsymbol{\tau}')'\Big) = -\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}(\boldsymbol{D}'\boldsymbol{\Delta}^{-1}\boldsymbol{D})^{-1}\boldsymbol{D}'\boldsymbol{\Delta}^{-1}\boldsymbol{g}_i(\boldsymbol{\theta}, \boldsymbol{\tau})\right) + o_p(1). \tag{A.33}$$

This form comes from the fact that the weight matrix is positive definite with probability approaching one[3] . The first term on the right-hand side is the influence function and hence inference on aggregate quantities follows directly. This result allows for use of the multiplier bootstrap to estimate standard errors in a computationally efficient manner.

A.4 Inference in Two-Way Fixed Effect Model

We derive the asymptotic distribution of our imputation estimator based off of the two-way error model in equation (1). First, we note that this estimator can be written in terms of the imputation matrix from Section 2. In particular, let $\mathbf{1}_t$ be a $T \times 1$ vector of ones up the t'th spot, with all zeros after. Define $\overline{\boldsymbol{y}}_{\infty} = (\overline{y}_{\infty,1}, ..., \overline{y}_{\infty,T})'$ be the full vector of never-treated cross-sectional averages. Then our imputation transformation can be written as

$$\tilde{\boldsymbol{y}}_i = [\boldsymbol{I}_T - \boldsymbol{P}(\mathbf{1}_T, \mathbf{1}_{T_0})]\,(\boldsymbol{y}_i - \overline{\boldsymbol{y}}_{\infty}) \tag{A.34}$$

where the t^{th} component of the above T-vector is

$$d_{it}\tau_{it} + \tilde{u}_{it}, \tag{A.35}$$

with $\tilde{u}_{it}$ is defined as the same transformation as $\tilde{y}_{it}$.

The imputation step of our estimator is a just-identified system of equations. As such, we do not need to worry about weighting in implementation and inference comes from standard theory of M-estimators. In fact, we have the following closed-form solution for the estimator of a group-time average treatment effect:

$$\widehat{\tau}_{gt} = \frac{1}{N_g}\sum_{i} D_{ig}\tilde{y}_{it}, \tag{A.36}$$

[3] This is a well-known expansion for analyzing the asymptotic properties of GMM estimators. See Chapter 14 of Wooldridge (2010) for example.

where $N_g = \sum_i D_{ig}$ is the number of units in group g.

The following theorem characterizes estimation under the two-way error model:

Theorem A.1. *Assume untreated potential outcomes take the form of the two-way error model given in equation (1). Suppose Assumptions 1 and 3 hold, as well as Assumption 2 with $\boldsymbol{\gamma}_i = 0$. Then for all (g, t) with $g > t$, $\widehat{\boldsymbol{\tau}}_{gt}$ is conditionally unbiased for $\mathbb{E}[\tau_{it} \mid D_{ig} = 1]$, has the linear form*

$$\sqrt{N_g}\big(\widehat{\tau}_{gt} - \tau_{gt}\big) = \frac{1}{\sqrt{N_g}} \sum_{i=1}^{N} D_{ig}(\tau_{it} - \tau_{gt} + u_{it} - \overline{u}_{i,t<T_0} - \overline{u}_{\infty,t} + \overline{u}_{\infty,t<T_0}) \tag{A.37}$$

and

$$\sqrt{N_1}(\widehat{\tau}_{gt} - \tau_{gt}) \xrightarrow{d} N(0, V_1 + V_0) \tag{A.38}$$

as $N \to \infty$, where V_1 and V_0 are given below and $\tau_{gt} = \mathbb{E}[y_{it}(g) - y_{it}(\infty) \mid D_{ig} = 1]$ is the group-time average treatment effect (on the treated). ■

Theorem (A.1) demonstrates the simplicity of our imputation procedure under the two-way error model. While the general factor structure requires more care, estimation and inference will yield a similar result.

A.4.1 Proof of Theorem Theorem A.1

The transformed post-treatment observations are

$$\tilde{y}_{it} = \tau_{it} + u_{it} - \overline{u}_{\infty,t} - \overline{u}_{i,t<T_0} + \overline{u}_{\infty,t<T_0} \tag{A.39}$$

To show unbiasedness, take expectation conditional on $D_{ig} = 1$. This expected value is

$$\mathbb{E}[\tau_{it} + u_{it} - \overline{u}_{i,t<T_0} - \overline{u}_{\infty,t} + \overline{u}_{\infty,t<T_0} \mid D_{ig} = 1] = \mathbb{E}[\tau_{it} \mid D_{ig} = 1] \tag{A.40}$$

by Assumption 2 and 3.

For consistency, note that averaging over the sample with $D_{ig} = 1$, subtracting τ_{gt}, and multiplying $\sqrt{N_g}$ gives

$$\sqrt{N_g}\big(\widehat{\tau}_{gt} - \tau_{gt}\big) = \frac{1}{\sqrt{N_g}} \sum_{i=1}^{N} D_{ig}(\tau_{it} - \tau_{gt} + u_{it} - \overline{u}_{i,t<T_0}) + \frac{1}{\sqrt{N_g}} \sum_{i=1}^{N} D_{ig}(-\overline{u}_{\infty,t} + \overline{u}_{\infty,t<T_0}) \tag{A.41}$$

which is two normalized sums of uncorrelated iid sequences that have mean zero (by iterated expectations) and finite fourth moments.

Rewriting the second term in terms of the original averages $\frac{1}{N_\infty}\sum_{i=1}^{N} -u_{i,t} + \overline{u}_{i,t<T_0}$ gives:

$$\sqrt{N_g}\big(\widehat{\tau}_{gt} - \tau_{gt}\big) = \frac{1}{\sqrt{N_g}}\sum_{i=1}^{N} D_{ig}(\tau_{it} - \tau_{gt} + u_{it} - \overline{u}_{i,t<T_0}) \\ + \sqrt{\frac{N_g}{N_\infty}}\bigg(\frac{1}{\sqrt{N_\infty}}\sum_{i=1}^{N} D_{i\infty}(-u_{i,t} + \overline{u}_{i,t<T_0})\bigg)$$

Since these terms are mean zero and uncorrelated, we find the variance of each term separately.

The first term has asymptotic variance

$$V_1 = \mathbb{E}\Big[\Big(\tau_{it} - \tau_{gt} + u_{it} - \overline{u}_{i,t<T_0}\Big)\Big(\tau_{it} - \tau_{gt} + u_{it} - \overline{u}_{i,t<T_0}\Big)' \mid D_{ig} = 1\Big] \tag{A.42}$$

and the second term has asymptotic variance

$$V_0 = \frac{\mathbb{P}(D_{ig} = 1)}{\mathbb{P}(D_{i\infty} = 1)}\,\mathbb{E}\Big[\Big(\overline{u}_{i,t<T_0} - u_{i,t}\Big)\Big(\overline{u}_{i,t<T_0} - u_{i,t}\Big)' \mid D_{i\infty} = 1\Big] \tag{A.43}$$

The result follows from the independence of the two sums.

A.5 Including Covariates

We now discuss the inclusion of covariates in the untreated potential outcome mean model. Allowing for covariates further weakens our parallel trends assumption by allowing selection to hold on unobserved heterogeneity as well as observed characteristics. Identifying the effects of covariates requires some kind of time and unit variation because we manually remove the level fixed effects.

A common inclusion in the treatment effects literature is time-constant variables with time-varying slopes. Suppose $\boldsymbol{x}_i$ is $1 \times K$ vector of time-constant covariates. We could write the mean model of the untreated outcomes as

$$\mathbb{E}[y_{it}(\infty) \mid x_i, \mu_i, \boldsymbol{\gamma}_i, D_i] = \boldsymbol{x}_i\boldsymbol{\beta}_t + \mu_i + \lambda_t + \boldsymbol{F}_t'\boldsymbol{\gamma}_i \tag{A.44}$$

which allows observable covariates to have trending partial effects; covariates with constant slopes are captured by the unit effect. After removing the additive fixed effects, $\boldsymbol{x}_i\boldsymbol{\beta}_t$ will take the same form as the residuals of factor structure. Estimating $\boldsymbol{\theta}$ can be done jointly with the time-varying coefficients by applying the QLD transformation to the vector of $\tilde{y}_{it} - \tilde{x}_i\tilde{\beta}_t$. We cannot identify the underlying partial effects because of the time-demeaning, but we can include them for the sake of strengthening the parallel trends assumption.

Time-constant covariates (or time-varying covariates fixed at their pre-treatment value) are often employed because there is little worry that they are affected by treatment. However, we could also include time- and individual-varying covariates of the form $\boldsymbol{x}_{it}$ that are allowed to have identifiable constant slopes if we assume their distribution is unaffected by treatment status. Let $\boldsymbol{x}_{it}$ be a $1 \times K$ vector of covariates that vary over i and t. We can jointly estimate a $K \times 1$ vector of parameters $\boldsymbol{\beta}$ along with $\boldsymbol{\theta}$ using the moments

$$\mathbb{E}\Big[\boldsymbol{H}(\boldsymbol{\theta})'(\tilde{\boldsymbol{y}}_i - \tilde{\boldsymbol{X}}_i\boldsymbol{\beta}) \otimes \boldsymbol{w}_i \mid G_i = \infty\Big] = \mathbf{0} \tag{A.45}$$

where $\tilde{\boldsymbol{X}}_i$ is the $T \times K$ matrix of stacked covariates after our double-demeaning procedure.

We could also allow slopes to vary across groups and estimate them via the group-specific pooled regression $D_{ig}y_{it}$ on $D_{ig}\boldsymbol{x}_{it}$ with unit-specific slopes on $D_{ig}\tilde{\boldsymbol{F}}(\widehat{\boldsymbol{\theta}})_t$ for $t = 1, ..., g-1$. Then we include the covariates and their respective slopes into the moment conditions

$$\mathbb{E}\Big[(\tilde{\boldsymbol{y}}_{i,t\geq g} - \tilde{\boldsymbol{X}}_{i,t\geq g}\boldsymbol{\beta}_g) - \boldsymbol{P}(\tilde{\boldsymbol{F}}_{t\geq g}, \tilde{\boldsymbol{F}}_{t<g})(\tilde{\boldsymbol{y}}_{i,t<g} - \tilde{\boldsymbol{X}}_{i,t<g}\boldsymbol{\beta}_g) - \boldsymbol{\tau}_g \mid G_i = g\Big] = \mathbf{0} \tag{A.46}$$

We note that the above expression requires treatment to not affect the evolution of the covariates, a strong assumption in practice. Chan and Kwok (2022) make a similar assumption for their principal components difference-in-differences estimator. We study this assumption in the context of the common correlated effects model in Brown et al. (2023a).

A.6 Testing Mean Equality of Factor Loadings

We develop this test in the context of the QLD estimation of Ahn et al. (2013). Specifically, we need $\mathbb{E}[\boldsymbol{\gamma}_i] = \mathbb{E}[\gamma_i \mid G_i = g]$ for all $g \in \mathcal{G}$. Our imputation approach allows us to identify these terms up to a rotation. To see how, let $\boldsymbol{A}^*$ be the rotation that imposes the Ahn et al. (2013)

normalization. Then

$$\begin{aligned}\boldsymbol{P}(\boldsymbol{I}_p, \boldsymbol{F}(\boldsymbol{\theta})_{t<g})\,\mathbb{E}[\boldsymbol{y}_{i,t<g} \mid G_i = g] &= \left(\boldsymbol{F}(\boldsymbol{\theta})'_{t<g}\boldsymbol{F}(\boldsymbol{\theta})_{t<g}\right)^{-1}\boldsymbol{F}(\boldsymbol{\theta})'_{t<g}\boldsymbol{F}_{t<g}\,\mathbb{E}[\boldsymbol{\gamma}_i \mid G_i = g]\\ &= \left(\boldsymbol{F}(\boldsymbol{\theta})'_{t<g}\boldsymbol{F}(\boldsymbol{\theta})_{t<g}\right)^{-1}\boldsymbol{F}(\boldsymbol{\theta})'_{t<g}\boldsymbol{F}(\boldsymbol{\theta})_{t<g}(\boldsymbol{A}^*)^{-1}\,\mathbb{E}[\boldsymbol{\gamma}_i \mid G_i = g]\\ &= (\boldsymbol{A}^*)^{-1}\,\mathbb{E}[\boldsymbol{\gamma}_i \mid G_i = g]\end{aligned}$$

where $\boldsymbol{F}(\boldsymbol{\theta}) = \boldsymbol{F}\boldsymbol{A}^*$.

It is irrelevant that the means of the factor loadings are only known up to a nonsingular transformation, because $\boldsymbol{A}^*$ is the same for each $g \in \mathcal{G}$ by virtue of the common factors. We note that

$$\mathbb{E}[\boldsymbol{\gamma}_i \mid G_i = g] - \mathbb{E}[\boldsymbol{\gamma}_i] = \mathbf{0} \iff (\boldsymbol{A}^*)^{-1}(\mathbb{E}[\boldsymbol{\gamma}_i \mid G_i = g] - \mathbb{E}[\boldsymbol{\gamma}_i]) = \mathbf{0} \tag{A.47}$$

The results above show how we can identify $(\boldsymbol{A}^*)^{-1}\,\mathbb{E}[\boldsymbol{\gamma}_i \mid G_i = g]$ by imputing the pre-treatment observations onto an identify matrix.

Collect the moments

$$\begin{gathered}\mathbb{E}\left[\frac{D_{i\infty}}{\mathbb{P}(D_{i\infty} = 1)}\boldsymbol{H}(\boldsymbol{\theta})\tilde{\boldsymbol{y}}_i \otimes \boldsymbol{w}_i\right] = \mathbf{0}\\ \mathbb{E}\left[\frac{D_{i\infty}}{\mathbb{P}(D_{i\infty} = 1)}\left(\boldsymbol{P}(\boldsymbol{I}_p, \boldsymbol{F}(\boldsymbol{\theta}))\boldsymbol{y}_i - \boldsymbol{\gamma}^*\right)\right] = \mathbf{0}\\ \mathbb{E}\left[\frac{D_{ig_G}}{\mathbb{P}(D_{ig_G} = 1)}\left(\boldsymbol{P}(\boldsymbol{I}_p, \boldsymbol{F}(\boldsymbol{\theta})_{t<g_G})\boldsymbol{y}_{i,t<g_G} - \boldsymbol{\gamma}^*_{gG}\right)\right] = \mathbf{0}\\ \vdots\\ \mathbb{E}\left[\frac{D_{ig_1}}{\mathbb{P}(D_{ig_1} = 1)}\left(\boldsymbol{P}(\boldsymbol{I}_p, \boldsymbol{F}(\boldsymbol{\theta})_{t<g_1})\boldsymbol{y}_{i,t<g_1} - \boldsymbol{\gamma}^*_{gG}\right)\right] = \mathbf{0}\end{gathered}$$

The parameters $(\boldsymbol{\gamma}^*, \boldsymbol{\gamma}^*_{gG}, ..., \boldsymbol{\gamma}^*_{g1})$ represent the rotated means of the factor loadings. $\boldsymbol{\gamma}$ is the unconditional mean $(\boldsymbol{A}^*)^{-1}\,\mathbb{E}[\boldsymbol{\gamma}_i]$ and $\boldsymbol{\gamma}_g$ is the conditional mean $(\boldsymbol{A}^*)^{-1}\,\mathbb{E}[\boldsymbol{\gamma}_i \mid G_i = g]$ for $g \in \mathcal{G}$. We include estimation of the factors for convenience, so that one does not need to directly calculate the effect of first-stage estimation on the asymptotic variances of conditional means.

Joint GMM estimation of the above parameters, including $\boldsymbol{\theta}$, then allows one to test combinations of the rotated means. Specifically, we have the following result:

Theorem A.2. *If* $\mathbb{E}[\boldsymbol{\gamma}_i \mid G_i = g] = \mathbb{E}[\boldsymbol{\gamma}_i]$ *for all* $g \in \mathcal{G}$*, then*

$$\boldsymbol{\gamma}^* = \boldsymbol{\gamma}^*_{g_G} = ... = \boldsymbol{\gamma}^*_{g_1} \tag{A.48}$$

■

Appendix B

Appendix to "Difference-in-Differences with Spatial Spillovers"

B.1 Proofs

B.1.1 Proof of Proposition 2.1

$$
\begin{aligned}
&\underbrace{\mathbb{E}[Y_{i1} - Y_{i0} \mid D_i = 1] - \mathbb{E}[Y_{i1} - Y_{i0} \mid D_i = 0]}_{\text{Difference-in-Differences}} \\
&= \mathbb{E}[Y_{i1}(1, h_i(\boldsymbol{D})) - Y_{i0}(0, \mathbf{0}) \mid D_i = 1] - \mathbb{E}[Y_{i1}(0, h_i(\boldsymbol{D})) - Y_{i0}(0, \mathbf{0}) \mid D_i = 0] \\
&= \mathbb{E}[Y_{i1}(1, h_i(\boldsymbol{D})) - Y_{i0}(0, \mathbf{0}) \mid D_i = 1] - \mathbb{E}[Y_{i1}(0, \mathbf{0}) - Y_{i0}(0, \mathbf{0}) \mid D_i = 0] \\
&\qquad - \mathbb{E}[Y_{i1}(0, h_i(\boldsymbol{D})) - Y_{i1}(0, \mathbf{0}) \mid D_i = 0] \\
&= \mathbb{E}[Y_{i1}(1, h_i(\boldsymbol{D})) - Y_{i0}(0, \mathbf{0}) \mid D_i = 1] - \mathbb{E}[Y_{i1}(0, \mathbf{0}) - Y_{i0}(0, \mathbf{0}) \mid D_i = 1] \\
&\qquad - \mathbb{E}[Y_{i1}(0, h_i(\boldsymbol{D})) - Y_{i1}(0, \mathbf{0}) \mid D_i = 0] \\
&= \mathbb{E}[Y_{i1}(1, h_i(\boldsymbol{D})) - Y_{i1}(0, \mathbf{0}) \mid D_i = 1] - \mathbb{E}[Y_{i1}(0, h_i(\boldsymbol{D})) - Y_{i1}(0, \mathbf{0}) \mid D_i = 0] \\
&\equiv \tau_{\text{total}} - \tau_{\text{spill}}(0),
\end{aligned}
$$

where the first equality comes from the potential outcomes and the No Anticipation assumption, the second from adding and subtracting terms, the third equality follows from the Parallel Counterfactual Trends assumption, and the final equality follows from the definitions of total and spillover effects.

B.1.2 Proof of Proposition 2.2

$$
\begin{aligned}
&\mathbb{E}[Y_{i1} - Y_{i0} \;\mid\; D_i = 1] - \mathbb{E}[Y_{i1} - Y_{i0} \;\mid\; D_i = 0, S_i = 0] \\
&\qquad = \mathbb{E}[Y_{i1}(1, h_i(\boldsymbol{D})) - Y_{i0}(0, \boldsymbol{0}) \;\mid\; D_i = 1] - \mathbb{E}[Y_{i1}(0, \boldsymbol{0}) - Y_{i0}(0, \boldsymbol{0}) \;\mid\; D_i = 0, S_i = 0] \\
&\qquad = \mathbb{E}[Y_{i1}(1, h_i(\boldsymbol{D})) - Y_{i0}(0, \boldsymbol{0}) \;\mid\; D_i = 1] - \mathbb{E}[Y_{i1}(0, \boldsymbol{0}) - Y_{i0}(0, \boldsymbol{0}) \;\mid\; D_i = 1] \\
&\qquad = \mathbb{E}[Y_{i1}(1, h_i(\boldsymbol{D})) - Y_{i0}(0, \boldsymbol{0}) - Y_{i1}(0, \boldsymbol{0}) + Y_{i0}(0, \boldsymbol{0}) \;\mid\; D_i = 1] \\
&\qquad = \mathbb{E}[Y_{i1}(1, h_i(\boldsymbol{D})) - Y_{i1}(0, \boldsymbol{0}) \;\mid\; D_i = 1] \\
&\qquad \equiv \tau_{\text{total}},
\end{aligned}
$$

where the first equality comes from the fact that $S_i = 0$ implies $h_i(\boldsymbol{D}) = 0$, the second equality from parallel trends assumption (2.4), and the last by definition of the total effect.

B.1.3 Identification of Imputation Estimator

This section will give identification arguments for the total effect and spillover effects. First, note under Parallel Counterfactual Trends (Staggered), we have

$$
\begin{aligned}
\mathbb{E}[Y_{it}(d_{it}, h_i(\boldsymbol{D}_t)) - \mu_i - \lambda_t \mid d_{it} = 1] &= \mathbb{E}[\tau_{it}(d_{it}, h_i(\boldsymbol{D}_t)) + \varepsilon_{it} \mid d_{it} = 1] \\
&= \mathbb{E}[\tau_{it}(d_{it}, h_i(\boldsymbol{D}_t)) \mid d_{it} = 1] \\
&= \tau_{total}.
\end{aligned}
$$

In the first stage, we estimate μ_i and λ_t using observations with d_{it} and $S_{it} = 0$. Unbiasedness/consistency follows from Spillovers Are Local and No Anticipation (Staggered) and the arguments of the appendix of Gardner (2022).

A similar argument follows for spillover effects. Again, assuming Parallel Counterfactual Trends (Staggered), we have

$$
\begin{aligned}
\mathbb{E}[Y_{it}(d_{it}, h_i(\boldsymbol{D}_t)) - \mu_i - \lambda_t \mid d_{it} = 0, S_{it} = 1] &= \mathbb{E}[\tau_{it}(d_{it}, h_i(\boldsymbol{D}_t)) + \varepsilon_{it} \mid d_{it} = 0, S_{it} = 0] \\
&= \mathbb{E}[\tau_{it}(d_{it}, h_i(\boldsymbol{D}_t)) \mid d_{it} = 0, S_{it} = 1].
\end{aligned}
$$

Appendix C

Appendix to “Difference-in-Differences with Geocoded Microdata”

C.1 Proofs

C.1.1 Proof of Proposition 3.1

Proof. Note using our model (3.2), we have

$$\begin{aligned}
s\,\mathbb{E}\left[\hat{\beta}_1\right] &= \mathbb{E}[\Delta Y_{it} \mid \mathcal{D}_t] - \mathbb{E}[\Delta Y_{it} \mid \mathcal{D}_c] \\
&= \mathbb{E}[\tau(\text{Dist}_i) + \lambda(\text{Dist}_i) + \Delta\varepsilon_i \mid \mathcal{D}_t] - \mathbb{E}[\tau(\text{Dist}_i) + \lambda(\text{Dist}_i) + \Delta\varepsilon_i \mid \mathcal{D}_c] \\
&= \mathbb{E}[\tau(\text{Dist}_i) \mid \mathcal{D}_t] - \mathbb{E}[\tau(\text{Dist}_i) \mid \mathcal{D}_c] + \mathbb{E}[\lambda(\text{Dist}_i) \mid \mathcal{D}_t] - \mathbb{E}[\lambda(\text{Dist}_i) \mid \mathcal{D}_c] \\
&\quad + \mathbb{E}[\Delta\varepsilon_i \mid \mathcal{D}_t] - \mathbb{E}[\Delta\varepsilon_i \mid \mathcal{D}_c].
\end{aligned}$$

By construction, $\Delta\varepsilon_i$ is uncorrelated with distance, so the final two terms in the sum is zero giving us result (i). Result (ii) comes from the fact that within d_c, $\lambda(\text{Dist}_i) = \lambda$. Result (iii) comes from the fact that if d_t is the correct cutoff $\mathbb{E}[\tau(\text{Dist}_i) \mid \mathcal{D}_c] = 0$. □

C.1.2 Proof of Proposition 3.2

Proof. Note that $L \to \infty$ implies $d_t \leq F_n^{-1}(\frac{L-1}{L})$ by assumption (3.5). This implies $\overline{\Delta Y}_L \to^p \lambda$ as $n \to \infty$ by assumption (3.5).

From assumption (3.2) and from our model (3.2), we have

$$
\begin{aligned}
\hat{\tau}_j &= \overline{\Delta Y}_j - \overline{\Delta Y}_L \\
&\to^p \mathbb{E}[\tau(\text{Dist}) \mid \text{Dist} \in \mathcal{D}_j] + \lambda - \lambda \\
&= \mathbb{E}[\tau(\text{Dist}) \mid \text{Dist} \in \mathcal{D}_j]
\end{aligned}
$$

As $L \to \infty$ and $n \to \infty$, we have that $\mathcal{D}_j$ approaches a set containing a singular point, say d_j. Therefore

$$\hat{\tau}_j \to^p \mathbb{E}[\tau(\text{Dist}) \mid \text{Dist} = d_j]$$

The sum of $\hat{\tau}_j$ therefore approach the conditional expectation function of $\tau(\text{Dist})$ pointwisely. See Cattaneo et al. (2019b) for proof of uniform convergence and underlying smoothness conditions for nonparametric consistency. □

C.2 Monte Carlo Simulations

In the following section, I describe a set of Monte Carlo simulations that compare the standard rings method with the proposed nonparametric method. I generate a number of units on the unit circle for two periods using the following data-generating process proposed in Diamond and McQuade (2019):

$$p_{it} = 1 + \tau(\text{Dist}_i)\, 1_{t=1} + \beta_{\text{Lat}}\text{Lat}_i + \lambda\, 1_{t=1} + \beta_{\text{Lon}}\text{Lon}_i + \varepsilon_{it}, \tag{C.1}$$

where $(\text{Lat}_i, \text{Lon}_i)$ is units' x and y coordinates on the unit circle, $\beta_{\text{Lat}} \sim N(0, 0.036)$ and $\beta_{\text{Lon}} \sim N(0, 0.036)$ determine how the price levels evolve over the unit circle, $\lambda \sim N(0, 0.025)$ is the constant common trend, $\varepsilon_{it} \sim N(0, 0.036)$ are idiosyncratic errors (uncorrelated with distance).

The key component of my simulations is the treatment effect curve, $\tau(\text{Dist}_i)$, which I vary across simulations:

$$\tau_1(\text{Dist}) = 0.15 * 1_{\text{Dist}<0.4}$$

$$\tau_2(\text{Dist}) = \left(0.5 * (0.8 - \text{Dist})^2\right) * 1_{\text{Dist}<0.8}$$

$$\tau_3(\text{Dist}) = \left(-0.15 + 1.2875 * \text{Dist} - 1.375 * \text{Dist}^2\right) * 1_{\text{Dist}<0.8}$$

$$\tau_4(\text{Dist}) = \left(0.5 * (0.8 - \text{Dist})^2\right) * 1_{\text{Dist}<0.25}$$

First, to compare the rings method and the nonparametric under favorable conditions for the ring method, I use τ_1 which assumes the treatment effect curve is flat for a fixed distance which aligns with the ring method. Second, I use the treatment effect curve proposed by Diamond and McQuade (2019) which is a positive effect that declines to 0 by Dist = 0.8. This treatment effect curve is difficult to estimate since it evolves smoothly over space. However, the estimate of a ring might be close to the average treatment effect since the curve is positive everywhere. Third, I use τ_3 which starts with a negative effect which becomes positive around Dist = 0.3 and then becomes 0 by Dist = 0.8. This treatment effect curve is the most difficult for the rings method since the average treatment effect is near zero even though there are significant positive and negative effects. Last, I modify τ_2 to cuttoff at Dist = 0.25, so that only units *really* close to treatment are affected by treatment. This is potentially difficult for the rings estimtator which will average over many non-affected units.

For each treatment effect curve, I generate data according to (C.1) and estimate the treatment effect curve in three ways. I use that standard rings method with two rings $\{(0, 0.6], (0.6, 1]\}$, with three rings $\{(0, 0.3], (0.3, 0.6], (0.5, 1]\}$, and the nonparametric ring method proposed in section 3.4. Note that these simulations are performed with the correct maximum treatment distance satisfied, which in practice is not known. In cases where the true treatment effect distance is not known, this would introduce additional bias in the treatment effect curve estimate.

I compare each estimator on their ability to estimate the true treatment effect curve. To do so, for each point in the data, I compute the predicted treatment effect $\hat{\tau}(\text{Dist}_i)$ and compare it to the true treatment effect $\tau(\text{Dist}_i)$. I'll label this quantity, $\upsilon_i \equiv \hat{\tau}(\text{Dist}_i) - \tau(\text{Dist}_i)$, as the prediction error. To summarize the prediction error, I calculate the mean-squared prediction error, $\hat{E}(\upsilon_i^2)$, and the average absolute prediction error, $\hat{E}(|\upsilon_i|)$, where the average is taken over the sample of observations with positive treatment effects. These numbers are a combination (i) how well the predicted treatment effect curve is approximated (bias of estimator) and (ii) the estimator's noise in repeated sampling (variance of estimator). Additonally, since it is often desirable to accurately estimate the largest treatment effects, I report the bias in predicted treatment effect at 0.4 miles following Diamond and McQuade (2019). For each metric, I divide by the nonparametric estimator's value, so that values > 1 perform worse than the nonparametric method and values < 1 perform better.

The results are presented in Table C.1. For τ_1, the 2 ring and 3 ring method performed better than the nonparametric method across metrics and across sample sizes. This is due to (i) the large rings accurately approximate the treatment effect curve and (ii) the sampling variation of the nonparametric estimator. For τ_2 and τ_3, the rings estimators perform better at small sample sizes, but by 5000 observations, the nonparametric method performs better in all 3 metrics which makes sense given that nonparametric estimators perform better with larger amounts of data. Last, for τ_4, the nonparametric estimator performs better than the rings method even at small sample sizes. This is because the rings method is able to better tease out the large treatment effect *very close* to treatment than the rings method. Overall, the results of Table C.1 show that the rings method performs better in most settings with at least a few thousand observations except for in the most-favorable case where the treatment effect curve is approximately flat.

C.3 Cross-Sectional Data

In the case of cross sectional data, we have individuals i that appear in the data in period $t(i) \in \{0, 1\}$. However, since we no longer are able to observe units in both periods, we are not able to take first differences. Our model therefore will have a λ term for both periods. Therefore, λ includes the average of μ_i, covariates, and period shocks at a given distance.

Table C.1: Monte Carlo Simulations

	$\hat{E}(\upsilon_i^2)$		$\hat{E}(\lvert\upsilon_i\rvert)$		Bias at 0.4 miles	
n	2 Rings	3 Rings	2 Rings	3 Rings	2 Rings	3 Rings
τ_1						
250	0.33	0.46	0.58	0.73	0.59	0.87
500	0.26	0.36	0.52	0.66	0.54	0.79
1000	0.21	0.29	0.47	0.58	0.49	0.69
5000	0.12	0.16	0.36	0.45	0.38	0.55
τ_2						
250	0.81	0.56	1.13	0.81	1.68	0.91
500	1.05	0.55	1.32	0.84	2.06	1.06
1000	1.47	0.61	1.59	0.93	2.54	1.29
5000	3.81	1.15	2.48	1.32	4.32	2.22
τ_3						
250	0.62	0.72	0.85	0.89	1.06	0.98
500	0.72	0.78	0.93	0.94	1.21	1.09
1000	0.91	0.95	1.06	1.04	1.45	1.31
5000	2.18	2.17	1.61	1.54	2.49	2.24
τ_4						
250	1.57	0.93	1.34	1.00	1.37	1.01
500	2.17	1.16	1.55	1.14	1.60	1.18
1000	3.63	1.85	1.95	1.44	2.06	1.52
5000	6.96	3.37	3.64	2.67	5.76	4.24

Notes. The table shows the results from 2,000 Monte Carlo simulations. For each treatment effect curve τ_i, I generate data following (C.1) and estimate the treatment effect curve using the standard ring methods with 2 or 3 rings and the nonparametric method. Columns (2)-(3) report the mean-squared prediction error, (4)-(5) report the mean absolute prediction error, and (6)-(7) report the prediction error at 0.4 miles. All results are normalized by the nonparametric value, so that values > 1 performed worse than the nonparametric method and vice-versa.

$$Y_i = \tau(\text{Dist}_i)\, 1_{t(i)=1} + \lambda_{t(i)}(\text{Dist}_i) + \nu_{it}. \tag{C.2}$$

The parallel trends assumption must be modified now in the case of cross sections:

Assumption C.1 (Local Parallel Trends (RC)). *For a distance $\bar{d}$, we say that 'local parallel trends' hold if for all positive $d, d' \leq \bar{d}$, then $\lambda_1(d) - \lambda_0(d) = \lambda_1(d') - \lambda_0(d')$.*

Assumption C.2 (Average Parallel Trends (RC)). *For a pair of distances d_t and d_c, we say that 'average parallel trends' hold if*

$$\mathbb{E}[\lambda_1(d) \mid 0 \leq d \leq d_t] - \mathbb{E}[\lambda_0(d) \mid 0 \leq d \leq d_t] = \mathbb{E}[\lambda_1(d) \mid d_t < d \leq d_c] - \mathbb{E}[\lambda_0(d) \mid d_t < d \leq d_c].$$

The parallel trends assumption is a bit more complicated now and is theoretically more strict. Local Parallel Trends (RC) still require changes in outcomes over time for a given unit i must be constant across distance (or on average in the case of Average Parallel Trends (RC)). However since the composition of units can change over time, this also requires that the average of individual fixed effects must be constant across time. This is well understood in the hedonic pricing literature that the composition of homes being sold can not change over time for identification (e.g. Linden and Rockoff (2008b)).

For completeness, I rewrite the other necessary assumption

Assumption C.3 (Correct d_t). *A distance d_t satisfies this assumption if (i) for all $d \leq d_t$, $\tau(d) > 0$ and for all $d > d_t$, $\tau(d) = 0$ and (ii) $F(d_c) - F(d_t) > 0$.*

The ring estimate in the case of cross-sections is given by:

$$\Delta Y_i = \beta_0 + \beta_1\, 1_{i \in \mathcal{D}_c}\, 1_{t(i)=1} + \beta_2\, 1_{i \in \mathcal{D}_t}\, 1_{t(i)=0} + \beta_4\, 1_{i \in \mathcal{D}_t}\, 1_{t(i)=1} + u_{it}. \tag{C.3}$$

Proposition C.1 (Decomposition of Ring Estimate (RC)). *Given that units follow model (C.2),* *(i) The estimate of β_4 in (3.3) has the following expectation:*

$$\begin{aligned}
\mathbb{E}\left[\hat{\beta}_4\right] &= \mathbb{E}[\Delta Y_{it} \mid \mathcal{D}_t] - \mathbb{E}[\Delta Y_{it} \mid \mathcal{D}_c] \\
&= \underbrace{\mathbb{E}[\tau(\mathit{Dist}) \mid \mathcal{D}_t] - \mathbb{E}[\tau(\mathit{Dist}) \mid \mathcal{D}_c]}_{\text{Difference in Treatment Effect}} \\
&\quad + \left(\mathbb{E}[\lambda_1(\mathit{Dist}) \mid \mathcal{D}_t, t(i) = 1] - \mathbb{E}[\lambda_0(\mathit{Dist}) \mid \mathcal{D}_t, t(i) = 0]\right) \\
&\quad - \left(\mathbb{E}[\lambda_1(\mathit{Dist}) \mid \mathcal{D}_c, t(i) = 1] - \mathbb{E}[\lambda_0(\mathit{Dist}) \mid \mathcal{D}_c, t(i) = 0]\right)
\end{aligned}$$

(ii) If d_c satisfies Local Parallel Trends (RC) or, more weakly, if d_t and d_c satisfy Average Parallel Trends (RC), then

$$\mathbb{E}\left[\hat{\beta}_4\right] = \underbrace{\mathbb{E}[\tau(\mathit{Dist}) \mid \mathcal{D}_t] - \mathbb{E}[\tau(\mathit{Dist}) \mid \mathcal{D}_c]}_{\text{Difference in Treatment Effect}}.$$

(iii) If d_c satisfies Local Parallel Trends (RC) and d_t satisfies Assumption 3.4, then

$$\mathbb{E}\left[\hat{\beta}_4\right] = \bar{\tau}.$$

Proof. With some algebraic manipulation, we can rewrite our difference-in-differences estimator as

$$\begin{aligned}
&\left(\mathbb{E}[Y_i \mid \mathcal{D}_t, t(i) = 1] - \mathbb{E}[Y_i \mid \mathcal{D}_t, t(i) = 0]\right) - \left(\mathbb{E}[Y_i \mid \mathcal{D}_c, t(i) = 1] - \mathbb{E}[Y_i \mid \mathcal{D}_c, t(i) = 0]\right)) \\
&\quad = \left(\mathbb{E}[\tau(\text{Dist}_i) + \lambda_1(\text{Dist}_i) + \nu_{i1} \mid \mathcal{D}_t, t(i) = 1] - \mathbb{E}[\lambda_0(\text{Dist}_i) + \nu_{i0} \mid \mathcal{D}_t, t(i) = 0]\right) \\
&\quad\quad - \left(\mathbb{E}[\tau(\text{Dist}_i) + \lambda_1(\text{Dist}_i) + \nu_{i1} \mid \mathcal{D}_c, t(i) = 1] - \mathbb{E}[\lambda_0(\text{Dist}_i) + \nu_{i0} \mid \mathcal{D}_c, t(i) = 0]\right) \\
&\quad = \mathbb{E}[\tau(\text{Dist}_i) \mid \mathcal{D}_t, t(i) = 1] - \mathbb{E}[\tau(\text{Dist}_i) \mid \mathcal{D}_c, t(i) = 0] + \\
&\quad\quad \left(\mathbb{E}[\lambda_1(\text{Dist}_i) \mid \mathcal{D}_t, t(i) = 1] - \mathbb{E}[\lambda_0(\text{Dist}_i) \mid \mathcal{D}_t, t(i) = 0]\right) \\
&\quad\quad - \left(\mathbb{E}[\lambda_1(\text{Dist}_i) \mid \mathcal{D}_c, t(i) = 1] - \mathbb{E}[\lambda_0(\text{Dist}_i) \mid \mathcal{D}_c, t(i) = 0]\right),
\end{aligned}$$

where the terms consisting of ν_{it} cancel out as they are uncorrelated with distance. Propositions (ii) and (iii) follow the same arguments as in the panel case. □

Part (i) of this theorem shows that under no parallel trends assumption, the 'Difference in Trends' term becomes the change in λ_t for the treated ring minus the change in for the control ring. As discussed above, this change in lambdas can be do to period specific shocks or changes in the composition of units observed in each period.

C.3.1 Nonparametric Estimation

Since we can no longer perform a single nonparmaetric regression on first differences in the context of cross-sections, our nonparametric estimator must be adjusted. The modified procedure will fit a nonparametric estimate of $\mathbb{E}[Y_i \mid \text{Dist}_i, t]$ seperately for $t = 0$ and $t = 1$ with a restriction that the bin intervals $\{\mathcal{D}_1, \dots, \mathcal{D}_L\}$ must be the same in both samples.[1] Then, for each distance bin we calculate an estimate of $\bar{Y}_{j,t}$ which corresponds to the sample average of observations in period t in bin $\mathcal{D}_j$.

Then estimates of τ_j can be formed as

$$\hat{\tau}_j = \left[\bar{Y}_{j,1} - \bar{Y}_{j,0}\right] - \left[\bar{Y}_{L,1} - \bar{Y}_{L,0}\right],$$

where, as before, the change in trends in the last ring serve as an estimate for the counterfactual trend. Under Local Parallel Trends (RC), estimates of $\hat{tau}_j$ are consistent for $\mathbb{E}[\tau(\text{Dist}_i) \mid i \in \mathcal{D}_j]$ and the treatment effect curve converges uniformly to the treatment effect curve $\tau(d)$.

Standard errors are formed similarly as before, but is the difference of four means so they can be formed as $\sqrt{\sigma^2_{j,1} + \sigma^2_{j,0} + \sigma^2_{L,1} + \sigma^2_{L,0}}$. These individual estimates and standard errors can be produced by the Stata/R package `binsreg`.

[1] The number of intervals are decided based on a different IMSE condition described in Cattaneo et al. (2019a) and the quantiles are calculated using the distribution of distances in both periods.