# Differnces-in-Differences with Spatial Spillover

Kyle Butts

December 3, 2020

# Spatial Spillovers

Researchers use difference-in-differences to estimate the average treatment effect on the treated:

$$\mathbb{E}\{Y_{i1}(1) - Y_{i1}(0) \mid D_i = 1\}$$

Effects of treatments at states, counties, etc. can spill over boundaries.

- e.g. large employer opening/closing in a county have positive or negative employment effects on nearby counties

# Spatial Spillovers

The difference-in-differences estimate is:

$$\hat{\mathbb{E}}\left[Y_{i1} - Y_{i0} \mid D_i = 1\right] - \hat{\mathbb{E}}\left[Y_{i1} - Y_{i0} \mid D_i = 0\right]$$

Two problems in presence of spillover effects:

- Nearby "control" units fail to estimate counterfactual trends because they are affected by treatment
- Treated units are also affected by nearby units and therefore combines "direct" effects with spillover effects

# Research Question

What are the sources of bias when researchers do not account for sptial spillover of treatment effects?

- In what contexts are these biases particularly large?
- Do current simple adjustments work?

## Preview of Results

There are two sources of bias:

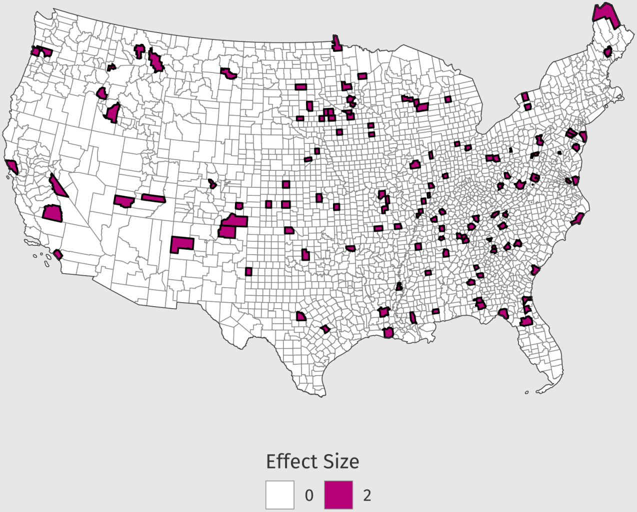
1. Treatment effects spillover onto control units
2. Treatment effects spillover onto also treated units

Solutions:

- Fixing bias 1 is relatively easy and works better when you have relatively few treated units
- Bias 2 worsens as treatment locations bunch more (spatial correlation of treatment increases)
- However, spatial correlation of treatment decreases bias 1

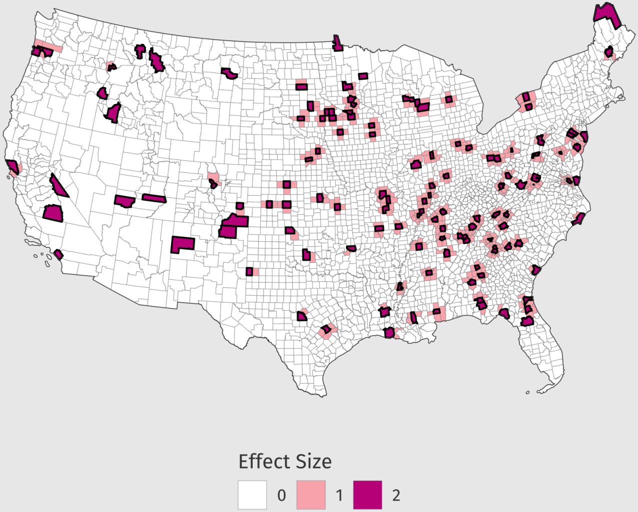
# Direct Effect of Treatment

*Pr(treatment) = 5%; Bias = 0*
*Treated units outlined in black*



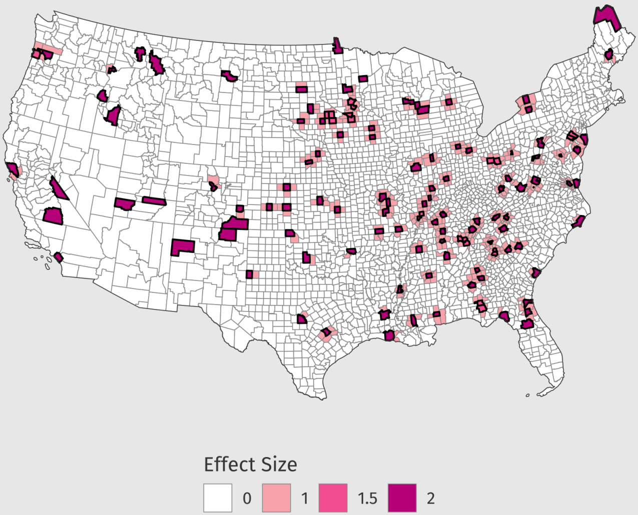
# Direct Effect + Spillover on Control

*Pr(treatment) = 5%; Bias = −0.13*
*Treated units outlined in black*



# Direct Effect + Spillover on Control and Treated

*Pr(treatment) = 0.05; Bias = −0.09 − 0.13 = −0.22*
*Treated units outlined in black*



# Outline

**Theory**

Simulations

Application: Tennessee Valley Authority

# Potential Outcomes Framework

For exposition, I will label units as counties. Assume all treatment occurs at the same time (2-periods or pre-post averages).[1]

- Let $Y_{it}(D_i, h_i(\vec{D}))$ be the potential outcome of county $i \in \{1, \dots, N\}$ at time $t$ with treatment status $D_i \in \{0, 1\}$.
- $\vec{D} \in \{0, 1\}^N$ is the vector of all units treatments.
- The function $h_i(\vec{D})$ maps the entire treatment vector into a scalar that completely determines the spillover.
- In the world with no treatment, $\vec{0}$, we say $h_i(\vec{0}) = 0$.

[1] This form adapts the framework of Vazquez-Bare (2019) for the diff-in-diff setting.

# Examples of $h_i(\vec{D})$

Examples of $h_i(\vec{D})$:

- **Treatment within $x$ miles:**
  $h_i(\vec{D}) = 1$ if there is a treated unit within $x$-miles of unit $i$ and $= 0$ otherwise
- **Average treatment among k-nearest neighbors:**
  $h_i(\vec{D}) = \sum_{j=1}^{k} D_{n(i,j)}/k$ where $n(i,j)$ represents the j-th closest neighbor to unit $i$.

# Estimand of Interest

Estimand of Interest:

$$\tau_{\text{direct}} \equiv \mathbb{E}\left[Y_{i,1}(1,0) - Y_{i,1}(0,0) \mid D_i = 1\right]$$

This is the direct effect in the absense of spillovers.

# Parallel Trends

I assume a modified version of the parallel trends assumption:

$$\mathbb{E}\left[Y_{i,1}(0,0) - Y_{i,0}(0,0) \mid D_i = 1, \vec{D} = 0\right]$$

$$= \mathbb{E}\left[Y_{i,1}(0,0) - Y_{i,0}(0,0) \mid D_i = 0, \vec{D} = 0\right],$$

- In the complete absence of treatment (not just the absence of individual $i$'s treatment), the change in potential outcomes from period 0 to 1 would not depend on treatment status

# What does Diff-in-Diff identify?

With the parallel trends assumption and random assignment of $D_i$, I decompose the diff-in-diff estimate as follows:

$$
\begin{aligned}
&\mathbb{E}_i\left[Y_{i1} - Y_{i0} \mid D_i = 1\right] - \mathbb{E}_i\left[Y_{i1} - Y_{i0} \mid D_i = 0\right] \\
&\quad = \mathbb{E}_i\left[Y_{i1}(1,0) - Y_{i1}(0,0) \mid D_i = 1\right] \\
&\qquad + \mathbb{E}_i\left[Y_{i1}(1, h_i(\vec{D})) - Y_{i1}(1,0) \mid D_i = 1\right] \\
&\qquad - \mathbb{E}_i\left[Y_{i1}(0, h_i(\vec{D})) - Y_{i1}(0,0) \mid D_i = 0\right] \\
&\quad \equiv \tau_{\text{direct}} + \tau_{\text{spillover, treated}} - \tau_{\text{spillover, control}}
\end{aligned}
$$

# Signing the Bias

$$\tau_{\text{direct}} + \tau_{\text{spillover, treated}} - \tau_{\text{spillover, control}}$$

**Table 1:** Bias from Spillovers

| | + Spillover | - Spillover |
|---|---|---|
| onto Treated | + Bias | - Bias |
| onto Control | - Bias | + Bias |

# Example of Bias

**Treated Spillover:**

$\mathbb{E}_i \left[ Y_{i1}(1, h_i(\vec{D})) - Y_{i1}(1, 0) \mid D_i = 1 \right] = -0.5$ if county $i$ is within 40 miles of a treated county and $0$ otherwise.

**Control Spillover:**

$\mathbb{E}_i \left[ Y_{i1}(0, h_i(\vec{D})) - Y_{i1}(0, 0) \mid D_i = 0 \right] = 1$ if county $i$ is within 40 miles of a treated county and $0$ otherwise.

- Then, the bias term is

$$-0.5 * \% \text{ of treated within 40 miles of a treated county}$$
$$-1 * \% \text{ of control within 40 miles of a treated county}$$

# Partial Identification

Now, I turn to formalizing the above example following the literature on partial identification for treatment effects: [Armstrong and Kolesár (2020), Rambachan and Roth (2020), and Manski and Pepper (2018)]

In this framework, researchers need to specify bounds on spillovers in the context of their research question.

# Partially-identified Set

Following Rambachan and Roth (2020) and Manski and Pepper (2018), researchers can provide bounds to the maximum spillovers.

Let $\Delta$ denote the interval of potential biases:

$$\begin{aligned}\Delta &\equiv [\underline{\Delta}, \overline{\Delta}] \\ &= \left[\min \tau_{\text{spill,treated}} - \tau_{\text{spill,control}},\ \max \tau_{\text{spill,treated}} - \tau_{\text{spill,control}}\right]\end{aligned}$$

The partially identified set is formed by

$$\left[\hat{\tau} - \underline{\Delta},\ \hat{\tau} + \overline{\Delta}\right]$$

# Bias-adjusted Inference

Inference can be done following Rambachan and Roth (2020) by creating a Fixed-Length Confidence Interval:

$$\left[\hat{\tau} - \underline{\Delta} - z_{1-\alpha/2}\mathsf{SE}_{\hat{\tau}},\ \ \hat{\tau} + \overline{\Delta} + z_{1-\alpha/2}\mathsf{SE}_{\hat{\tau}}\right]$$

These standard errors should account for the spatial structure of outcomes following Conley (1999).

# Outline

Theory

**Simulations**

Application: Tennessee Valley Authority

# Monte Carlo Simulations

- Simulated the following Data Generating Process

$$
\begin{aligned}
y_{it} = &-2 + \mu_t + \mu_i + \tau_{\text{direct}} D_{it} + \tau_{\text{spillover, control}} (1 - D_{it}) \text{Near}_{it} \\
&+ \tau_{\text{spillover, treated}} D_{it} \text{Near}_{it} + \varepsilon_{it}
\end{aligned}
$$

- Observation $i$ is a U.S. county, year $t \in \{2000, \ldots, 2019\}$, treatment $D_{it}$ turns on in 2010 and is assigned randomly or through a spatial process.
- $\text{Near}_{it}$ is an indicator for having your county centroid be within 40 miles of a treated unit.
- $\tau_{\text{direct}} = 2$, $\tau_{\text{spillover, control}} = 1$ and $\tau_{\text{spillover, treat}} = -0.5$.

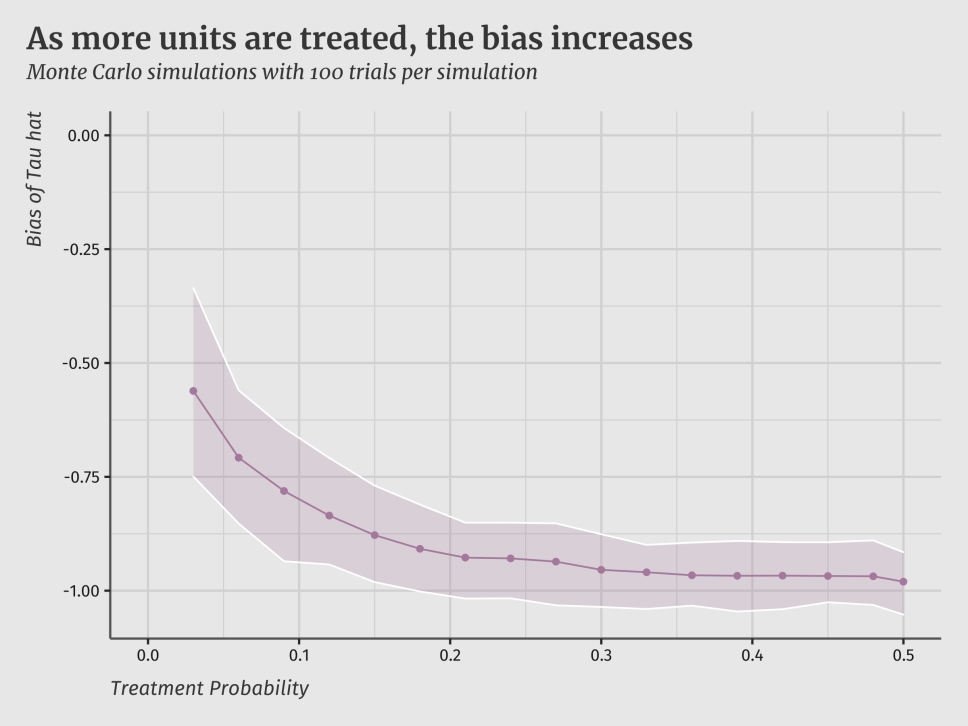
# Source of Bias 1: Control Units

Simulation 1:

- The first simulation changes the probability of treatment, $D_{it}$.
- Assume, for now, there is only spillover on the control, i.e. $\tau_{\text{spillover,treat}} = 0$.
- Estimate the following equation:

$$y_{it} = \alpha + \mu_t + \mu_i + \tau D_{it} + \epsilon_{it}$$

- The bias of our estimate is $\mathit{Bias} = \hat{\tau} - 2$

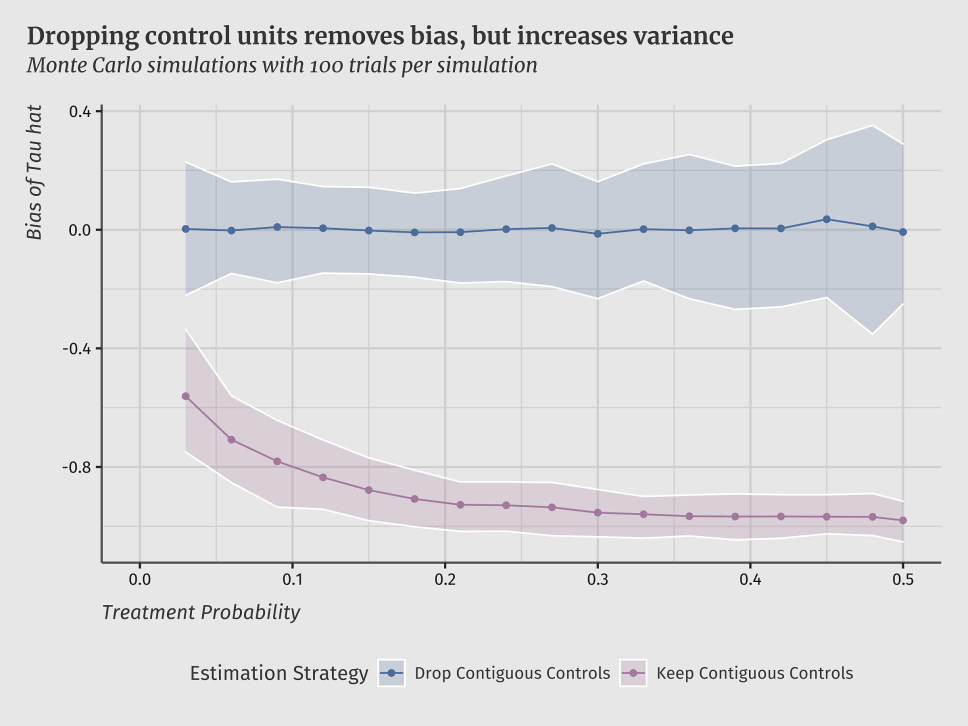


As more units are treated, the bias increases
Monte Carlo simulations with 100 trials per simulation
Bias of Tau hat
0.00
-0.25
-0.50
-0.75
-1.00
0.0
0.1
0.2
0.3
0.4
0.5
Treatment Probability

# Solution: Removing "contaminated" controls

A common solution to the problem of spillover is to restimate without neighboring control units and see what happens to the bias.

This simulation assumes the correct "contaminated" control units, i.e. only control units with $h_i(\vec{D}) \neq 0$, which is not possible for researchers.



Dropping control units removes bias, but increases variance
Monte Carlo simulations with 100 trials per simulation
Bias of Tau hat
0.4
0.0
-0.4
-0.8
0.0
0.1
0.2
0.3
0.4
0.5
Treatment Probability
Estimation Strategy
Drop Contiguous Controls
Keep Contiguous Controls

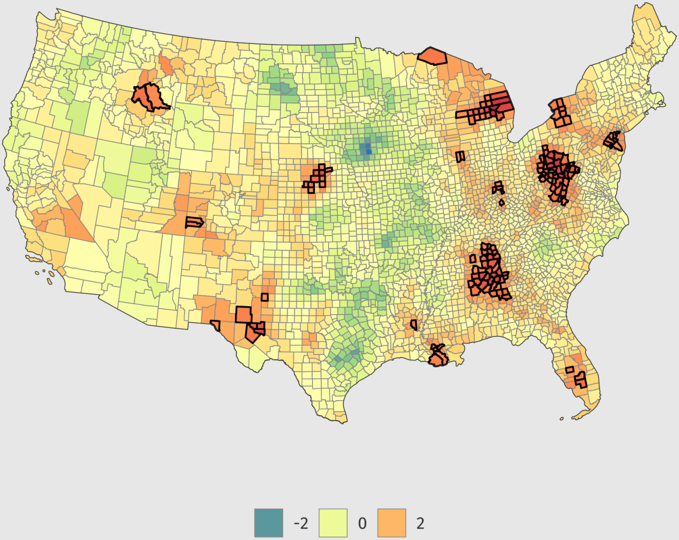
# Source of Bias 2: Treated Units

Simulation 2:

- I now add in the bias from spillover of treated units on to other treated units.
- This source of bias only occurs when treated units are located next to eachother, so the magnitude of bias depends on spatial autocorrelation of $D_{it}$.
- Using a method from geosciences, I generate correlated "fields" across the US.
- Then, for the counties in the top 10% of values (hot spots in the field), I increase the probability of treatment.
- The unconditional probability remains fixed at 10%.

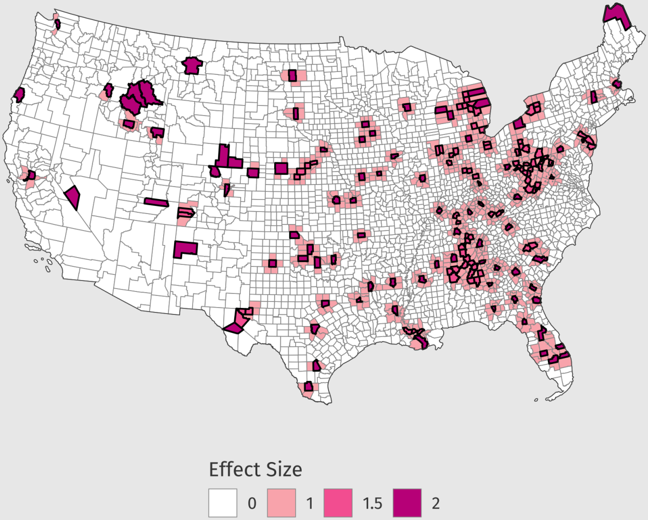
# Kriging Example

*Spatial Autocorrelation Measure = 1.4*



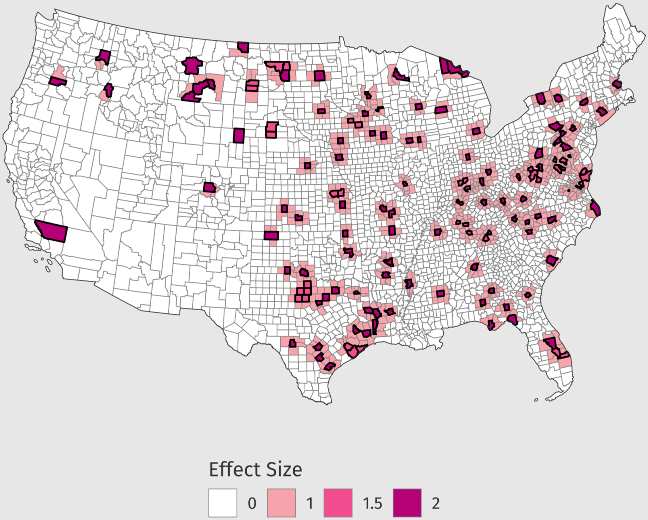
# Large Spatial Autocorrelation

*Pr(treatment) = 5%; Bias = −0.25 - 0.24 = −0.49; Spatial Autocorrelation Measure = 1.4*
*Treated units outlined in black*



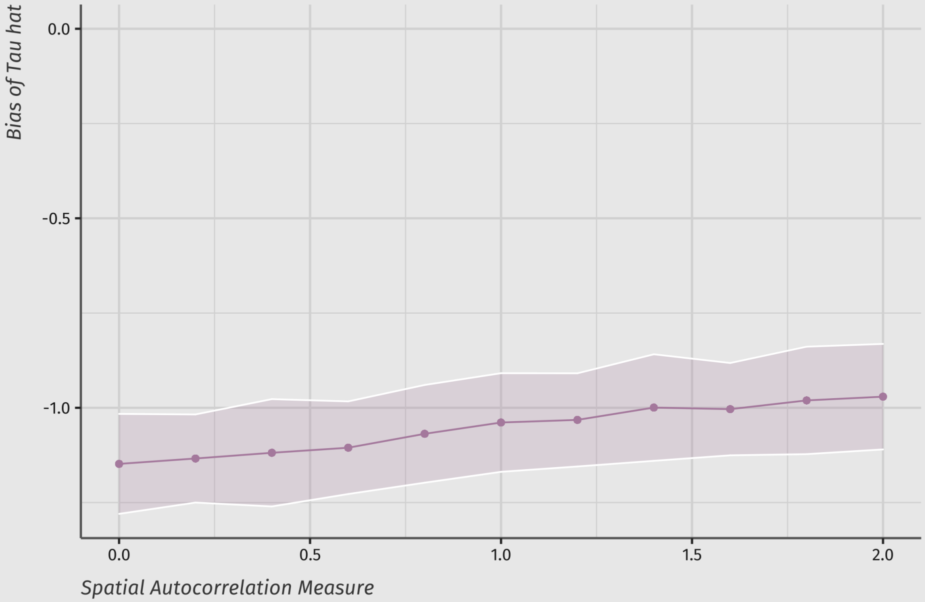
## Small Spatial Autocorrelation

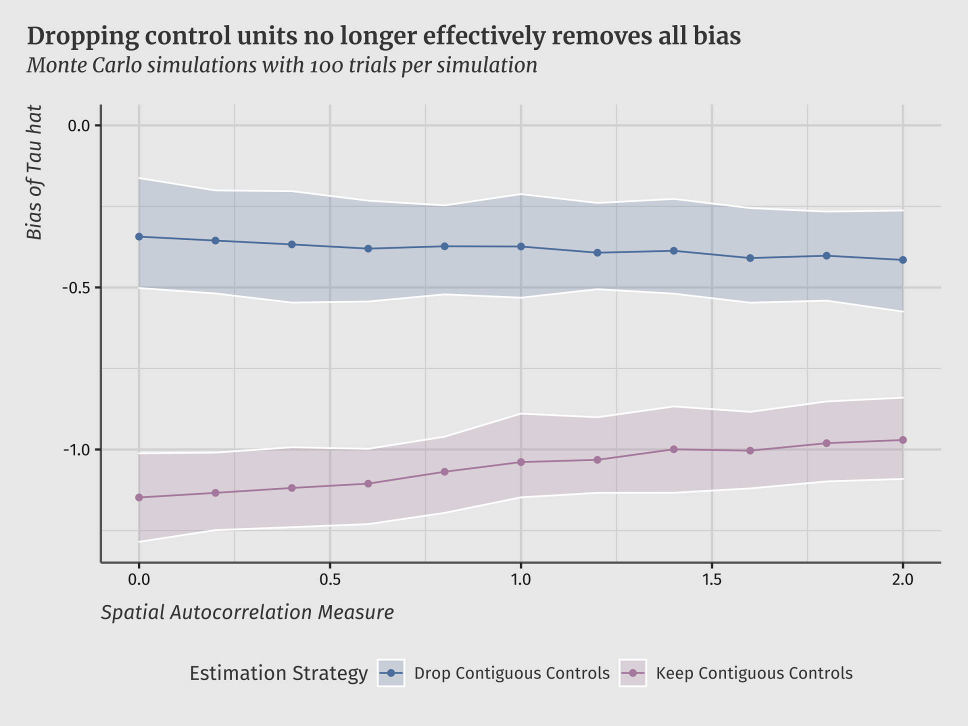
*Pr(treatment) = 5%; Bias = −0.15 - 0.24 = −0.39; Spatial Autocorrelation Measure = 0.4*
*Treated units outlined in black*



# Bias as spatial correlation changes

*Monte Carlo simulations with 100 trials for each simulation*





Dropping control units no longer effectively removes all bias
Monte Carlo simulations with 100 trials per simulation
Bias of Tau hat
0.0
-0.5
-1.0
0.0
0.5
1.0
1.5
2.0
Spatial Autocorrelation Measure
Estimation Strategy
Drop Contiguous Controls
Keep Contiguous Controls

# Spillovers as estimand of interest

Until now, we assumed our estimand of interest is $\tau_{\text{direct}}$.

However, the two other spillover effects are of interest as well:

- $\tau_{\text{spillover, control}}$: Do the benefits of a treated county come at a cost to neighbor counties?
- $\tau_{\text{spillover, treated}}$: Does the estimated effect change based on others treatment? (This is what you should consider if you are a policy maker)

To estimate the spillover effects, we have to parameterize $h_i(\vec{D})$ function and the potential outcomes function $Y_i(D_i, h_i(\vec{D}))$.

# Outline

Theory

Simulations

**Application: Tennessee Valley Authority**

# Tennessee Valley Authority

Kline and Moretti (2014) look at the long-run impacts of the Tennessee Valley Authority (TVA).

- The TVA was a large-scale federal investment started in 1934 that focused on improving manufacturing economy.
- The program focused on electrification through dams and transportation canals in order to improve the manufacturing economy
- Hundreds of dollars spent anually per person

## Identification

Kline and Moretti (2014) run the county-level difference-in-differences specification:

$$y_{c,2000} - y_{c,1940} = \alpha + \text{TVA}_c \tau + X_{c,1940}\beta + (\varepsilon_{c,2000} - \varepsilon_{c,1940}) \quad (1)$$

- $y$ are outcomes for agricultural employment, manufacturing employment, and median family income.
- $\text{TVA}_c$ is the treatment variable
- $X_{c,1940}$ allow for different long-term trends based on covariates in 1940.

They trim the sample using a logit regression to predict treatment using $X_{c,1940}$ and keep control units in the top 75% of predicted probability.

## Spillovers in the TVA Context

In our context, there is reason to believe spillovers can occur to nearby counties

- Agricultural employees might be drawn to hire wages for new manufacturing jobs in Tennessee Valley
- Manufacturing jobs that would have been created in the control units in the absence of treatment might move to the Tennessee Valley
- Since manufacutring jobs pay higher than agricultural jobs, the above two spillover effects can cause changes in median family income

# Specification including spillovers

$$\Delta y_c = \alpha + \text{TVA}_i \tau + \sum_{d \in \text{Dist}} \text{Between}(d) \delta_d + X_{i,1940} \beta + \Delta \varepsilon_c \qquad (2)$$

- Between($d$) is a set of indicators for being in the following distance bins (in miles) from the Tennessee Valley Authority:

$$d \in \{(0, 50], (50, 100], (100, 150]\}$$

# Effective Sample and Spillover Variables

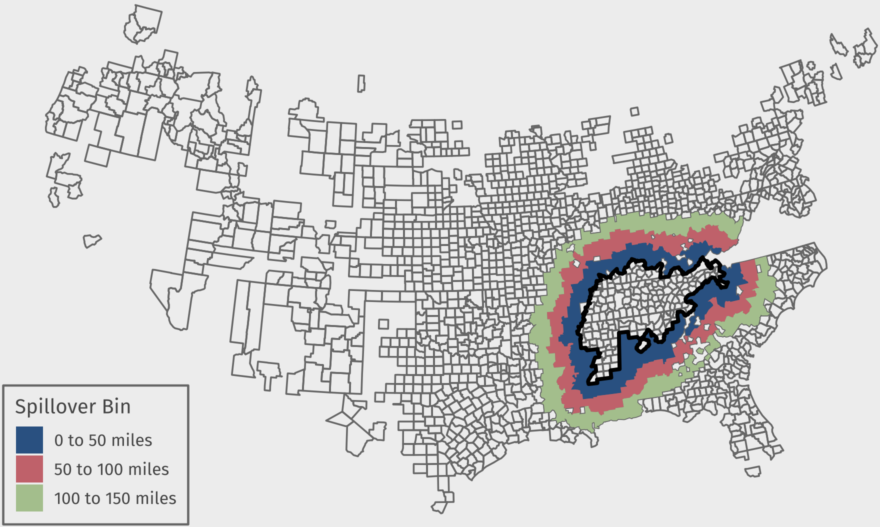


**Table 2:** Effects of Tennessee Valley Authority on Decadel Growth

| | **Diff-in-Diff** | **Diff-in-Diff with Spillovers** | | | |
|---|---|---|---|---|---|
| *Dependent Var.* | TVA | TVA | TVA between 0-50 mi. | TVA between 50-100 mi. | TVA between 100-150 mi. |
| Agricultural employment | $-0.0514^{***}$ | $-0.0678^{***}$ | $-0.0310^{**}$ | $-0.0112$ | $-0.0252^{***}$ |
| | (0.0087) | (0.0102) | (0.0123) | (0.0094) | (0.0084) |
| Manufacturing employment | $0.0560^{***}$ | $0.0461^{**}$ | $-0.0104$ | $-0.0128$ | $-0.0248^{*}$ |
| | (0.0187) | (0.0210) | (0.0205) | (0.0257) | (0.0147) |
| Median family income | $0.0191^{***}$ | $0.0196^{**}$ | 0.0056 | $-0.0062$ | $-0.0011$ |
| | (0.0067) | (0.0077) | (0.0079) | (0.0060) | (0.0032) |

$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.