PHS 2000B: Econometrics Interrupted Time Series and Difference in Differences 2023

Maggie McConnell

Department of Global Health and Population

Harvard T. H. Chan School of Public Health

Plan of presentation

1. Wrap up threats to validity DiD
2. Discuss DiD versus Controlled ITS
3. DiD with multiple treatment periods
4. Synthetic control methods

Difference in Differences recap

- Diff-in-diff is a double differencing strategy: we are taking the difference of two quantities that themselves are differences.
- It is generally applied when treatment is applied at a group level and we have pre- and post-treatment outcome data for treated and control groups.
- DD can be used to estimate treatment effects even when we have:
  - multiple groups
  - multiple time points, and
  - different treatment start dates
- The control group doesn't have to be a pure control
  - Could compare *more* to *less* treated regions

Key assumptions: Counterfactual

- Key assumption in Difference-in-Differences
  - Treatment group would have followed the same trajectory as the control group in the post-intervention period if the intervention had not occurred

- This counterfactual is **not** testable
  - This counterfactual may feel more believable if
    - Control and treatment areas are qualitatively similar before the intervention
    - The composition of control and treatment groups are not changing over time in different ways
    - The control and treatment group are on a similar trajectory before the intervention

Event study option

- One way to demonstrate what is happening with pre-trends is through estimating an “event study"
- We still use basic difference in differences structure – fixed effects for each time period, fixed effects for each state
- Also estimates a time to event variable indexed on when the intervention occurred
- Allows for visualization of whether pre-trends differ in treated areas
- Allows for more flexible specification of the timing of treatment effects

Example – Impact of closing SSA offices

Who Is Screened Out? Application Costs and the Targeting of Disability Programs

Deshpande and Li (2019)

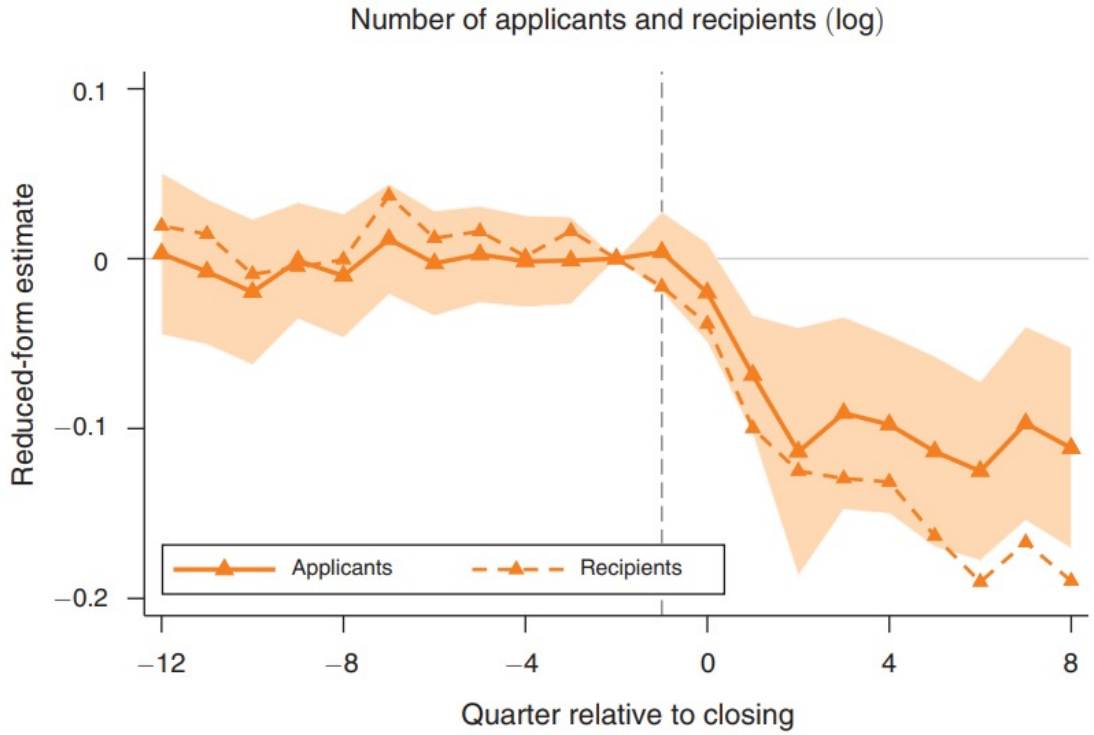


FIGURE 3. EFFECT OF CLOSINGS ON NUMBER OF DISABILITY APPLICATIONS AND ALLOWANCES

Notes: The figure plots estimates of the effect of the closing on applications (recipients) in closing zip codes in the event quarters before and after the closing. Specifically, the figure plots estimates of δ_τ coefficients from equation (4), which is a regression of the number of disability applicants (recipients) on zip code fixed effects, quarter-by-state fixed effects, a treatment indicator, event quarter indicators, and event quarter indicators interacted with the treatment indicator. The dependent variable is the log number of disability applications (solid series) or the log number of disability recipients (dashed series). The shaded region is the 95 percent confidence interval for disability applications (solid series). The sample is zip codes in which the nearest office closed after 2000 and that have an average of at least four disability applications per quarter in the year before the closing. Regressions are weighted by application or recipient volume in the year before the closing.

Valid standard errors: Clustering

- Many difference in difference studies use individual level data to understand group level policy change
  - Challenge is that outcomes are correlated within groups
- One approach is to aggregate data up to the group level and use group level averages to estimate effects like we did in ITS
  - Will substantially restrict our power
- An alternative is to cluster standard errors at the *level of treatment*, to control for correlation of individuals within the same group
- Works well with a large number of clusters

Valid standard errors: Small number of groups

- With many groups, clustering the standard errors at the group level works well to allow for intracluster correlation
- With a small number of groups, this is not a good approach
  - Can actually lead to standard errors that are *too small*
  - Could instead average the data across groups but this will limit power
- Can use bootstrap/placebo methods but more complex
  - "Wild" Bootstrap – use boottest in STATA implements Cameron et al 2008
  - Construct placebo intervention data based on random assignment of intervention to groups and construct test statistics based on the distribution of outcomes for placebo interventions
  - Can work with as few as 6 groups

Valid standard errors: Autocorrelation

- With autocorrelation, observations for the same group will be correlated over time
- Observations of multiple time points pre- and post-intervention will not be independent and do not add much additional information
- With a large number of groups, clustering the standard errors across time within the same group can correct for autocorrelation
- We can also estimate the autocorrelation structure, but very biased estimates with group fixed effects and small number of periods

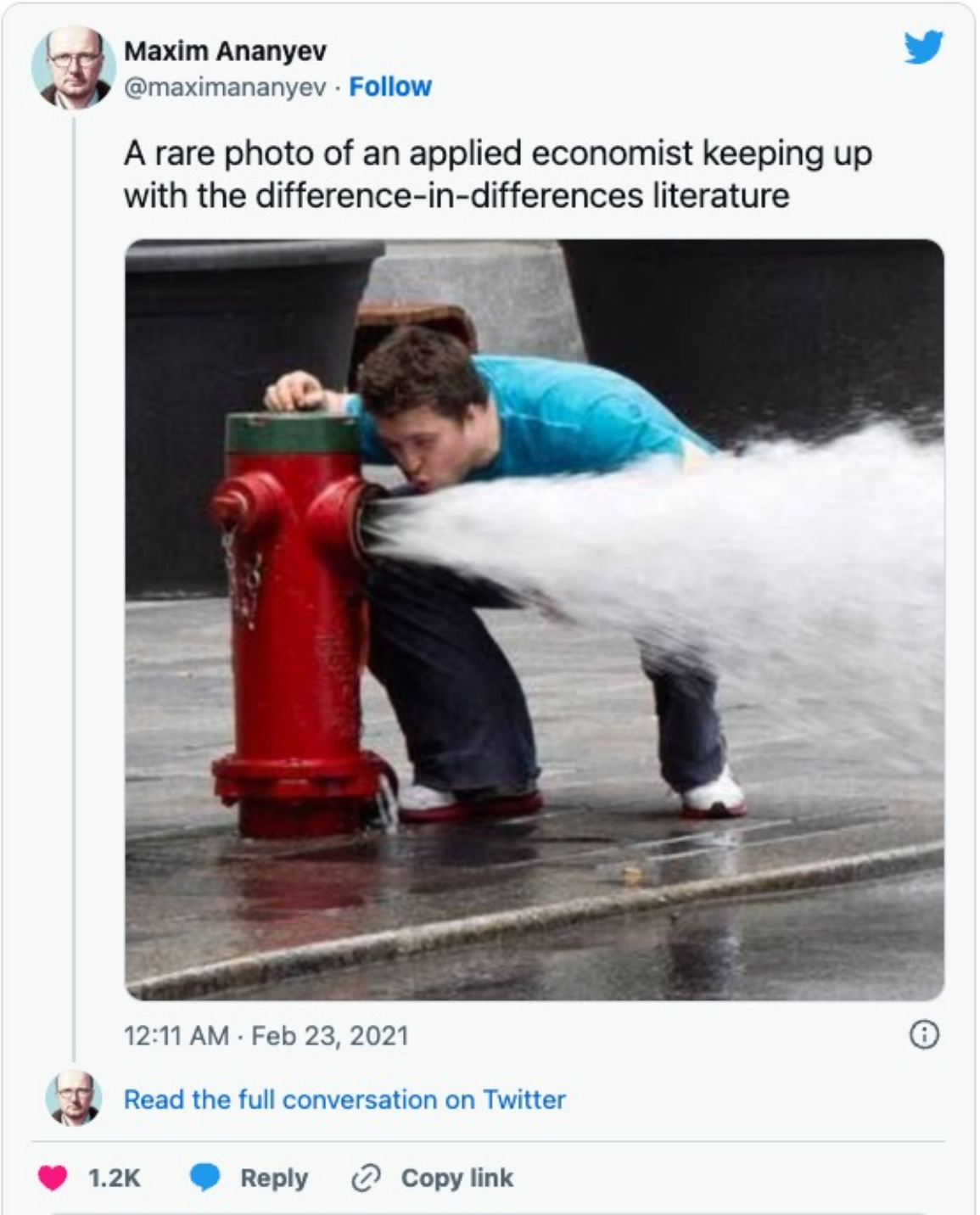
Difference-in-Differences vs. Interrupted Time Series

- *Conceptually similar*: both used to estimate the effect of group level interventions by comparing to a counterfactual; what would have happened without the intervention in the post-intervention period

- *Structural differences*
  - ITS uses aggregated data with one treated area time series whereas Difference-in-Differences usually keeps the data at the individual level and adjusts for within-group correlation
  - ITS has one treatment group (and one control group in CITS) while Difference-in-Differences works well with multiple treated and control groups
  - ITS fits a linear time trend and extrapolates counterfactual whereas Difference-in-Differences uses time fixed effects and uses control group to generate counterfactual
  - ITS may work well when change is expected to be immediate and large
  - DiD methods are more flexible – can handle situations where changes may be more subtle or phase in over time

DiD is a fast evolving method

- Today we will discuss some new developments in difference in differences
- The conversation will not be comprehensive
- The landscape is continuously evolving so can be challenging to stay up to date
- Check recent review papers for discussion of ”state of the art” methods



Recap – two way fixed effects model

- We need at least two groups – treatment and control – and two time periods – pre and post – to get a difference-in-differences estimator
- DID is a flexible methodology we can extend in many ways. Commonly we might use the following set-up:
  - We have multiple groups ($g = \{1 \dots M\}$) with some getting the treatment and some not getting it (group fixed effects)
  - We have multiple time periods ($t = \{1 \dots K\}$) both pre- and post-intervention (time fixed effects)

$$Y_{igt} = \sum_{g=1}^{M} \alpha_g + \sum_{t=1}^{K} \gamma_t + \beta\left(I_{post} * I_{g \in T}\right) + \epsilon_{igt}$$

- This model is sometimes called “two way fixed effects”

Examples of staggered rollouts

- Medicaid expansion under the Affordable Care Act
- Closures of SSA offices
- Changes to the Earned Income Tax Credit at the state level
- Treatment starts at different time in different groups

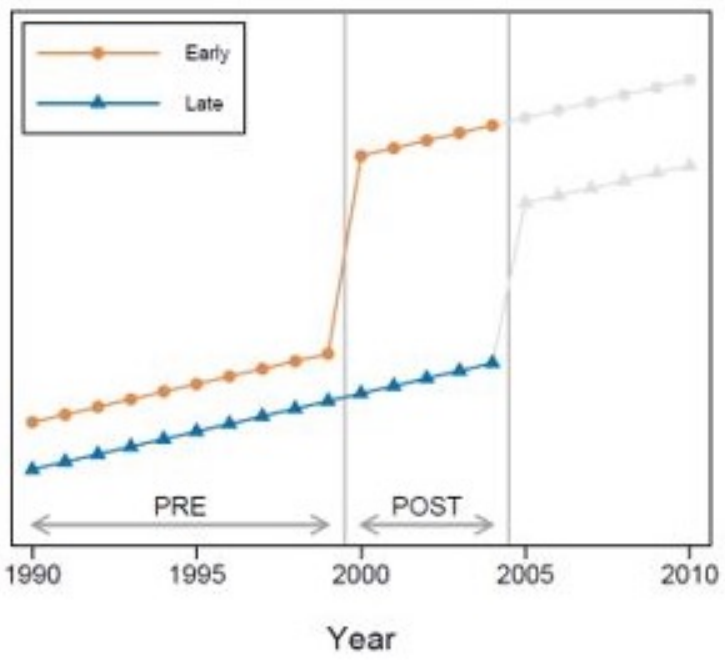
Staggered rollout DiD design

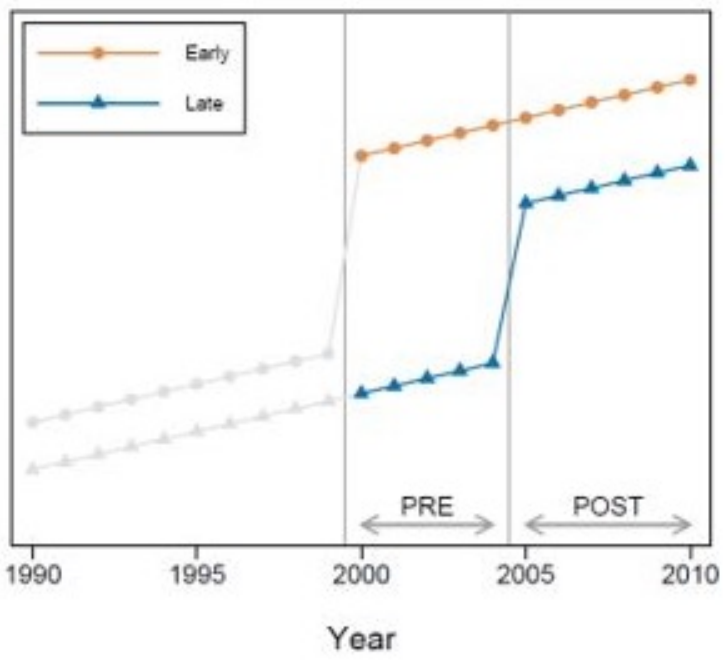
- Consider a model where different states adopted a treatment at different times and you are using two-way fixed effects (group and time fixed effects) to estimate treatment effects

- It turns out you can decompose DiD estimator into a weighted average of all possible two-by-two difference-in-differences estimators that can be constructed from your data

Example

- Imagine you are looking at the impact of transitioning to free primary schools
- Some states are early adopters and some adopt later





DiD estimator is a weight average of:

- comparisons between early adopters and later adopters over the periods when the later adopters are not yet been treated
- comparisons between early adopters and later adopters over the periods when the early adopters are treated – they become comparison group for the later adopter
- comparisons between different timing groups (e.g., early adopters or later adopters) and the never-treated

Implications

- Difference in difference designs only recover average treatment effects when treatment effects are homogeneous

- When treatment effects change over time DiD will provide biased estimates of treatment effect
  - This is because the DiD estimator uses already treated units as controls for future treated units
  - When would it be a problem to use already treated as controls for later treated?
  - If outcomes in already treated units continue to react to treatment they make bad controls

- Luckily – there are econometric strategies to address this!
  - Many of them attempt to avoid comparisons to "forbidden controls"
  - Various approaches summarized in Roth et al 2022 *What's Trending in Difference-in-Differences? A Synthesis of the Recent Econometrics Literature*

Remember – Impact of closing SSA offices

Who Is Screened Out? Application Costs and the Targeting of Disability Programs

Deshpande and Li (2019)

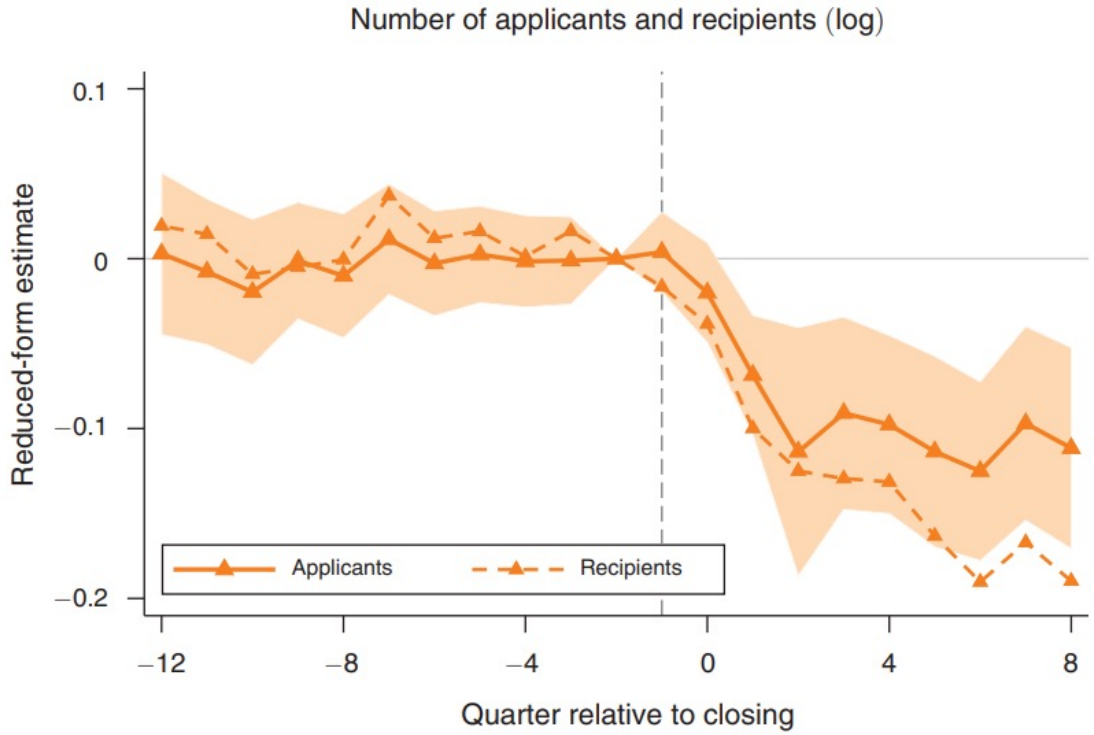


FIGURE 3. EFFECT OF CLOSINGS ON NUMBER OF DISABILITY APPLICATIONS AND ALLOWANCES

Notes: The figure plots estimates of the effect of the closing on applications (recipients) in closing zip codes in the event quarters before and after the closing. Specifically, the figure plots estimates of δ_τ coefficients from equation (4), which is a regression of the number of disability applicants (recipients) on zip code fixed effects, quarter-by-state fixed effects, a treatment indicator, event quarter indicators, and event quarter indicators interacted with the treatment indicator. The dependent variable is the log number of disability applications (solid series) or the log number of disability recipients (dashed series). The shaded region is the 95 percent confidence interval for disability applications (solid series). The sample is zip codes in which the nearest office closed after 2000 and that have an average of at least four disability applications per quarter in the year before the closing. Regressions are weighted by application or recipient volume in the year before the closing.

How do you handle this if you are estimating difference in differences?

Economics strategy:
Estimate using different estimators – demonstrate robustness to model specification

Braghieri et al 2022, Social Media and Mental Health

Uses the rollout of Facebook across US colleges as the "treatment"

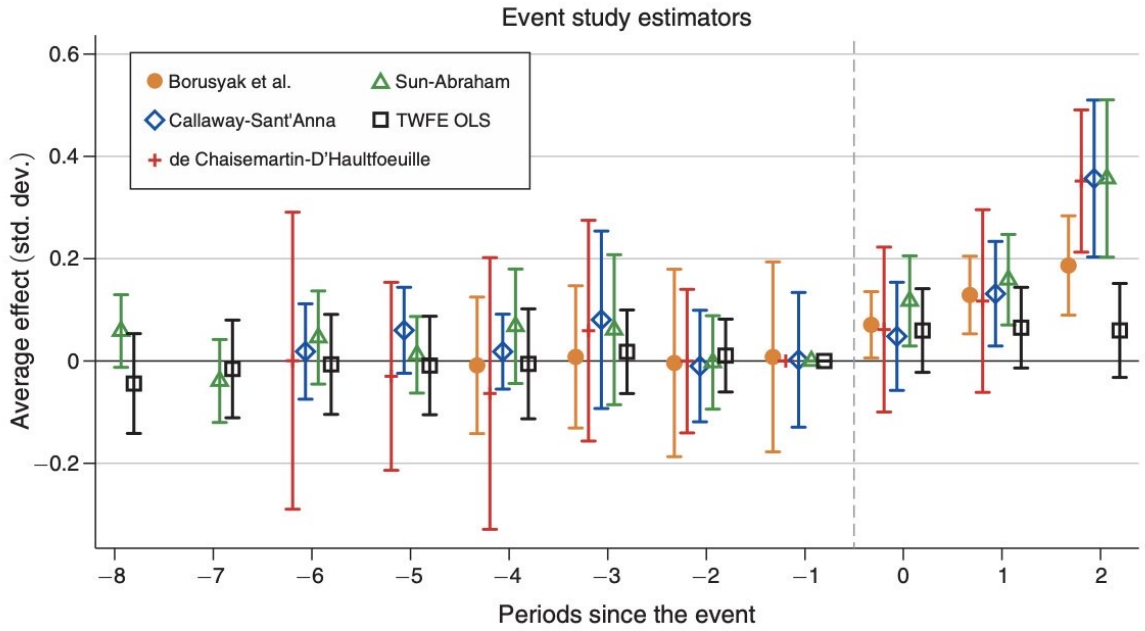


FIGURE 2. EFFECTS OF FACEBOOK ON THE INDEX OF POOR MENTAL HEALTH BASED ON DISTANCE TO/FROM FACEBOOK INTRODUCTION

Notes: This figure overlays the event-study plots constructed using five different estimators: a dynamic version of the TWFE model, equation (2), estimated using OLS (in black with square markers); Sun and Abraham (2021) (in green with triangle markers); Callaway and Sant'Anna (2021) (in blue with diamond markers); De Chaisemartin and d'Haultfoeuille (2020) (in red with cross markers); and Borusyak, Jaravel, and Spiess (2021) (in orange with circle markers). The outcome variable is our overall index of poor mental health. The time variable is the survey wave and the treatment group variable is given by the semester in which the college attended by the student was granted Facebook access. The figure displays only two postperiods because the estimation of additional post periods would require employing already treated units as controls for newly treated units. In the presence of heterogeneous dynamic treatment effects, such comparisons would bias the estimation and, therefore, they are shut down by all the newly introduced robust estimators. As a result, the maximum number of postperiods that can be estimated robustly is two. For the Borusyak, Jaravel, and Spiess (2021) estimator, we estimate four preperiods since estimating more preperiods dramatically increases the standard errors in the preperiod (Borusyak, Jaravel, and Spiess 2021, p. 24). Similarly, for the estimator by De Chaisemartin and d'Haultfoeuille (2020), the maximum number of preperiods that can be estimated in our panel is only five. In order to estimate the standard errors for the $t + 2$ estimate, the De Chaisemartin and d'Haultfoeuille (2020) estimator includes controls for age and age squared. For appropriate estimation of the coefficients on $t = -8$ and $t = -7$ using the Sun and Abraham (2021) estimator, we include data from additional preperiods, even though, in those preperiods, we do not observe all four Facebook expansion groups (Sun and Abraham 2021, p. 13). For a detailed description of the outcome and treatment variables, see online Appendix Table A.31. The bars represent 95 percent confidence intervals. Standard errors are clustered at the college level.

Synthetic control -- Intuition

- Intuition – we are trying to find a control group that is similar to treated group in pre-treatment period
  - Similar covariates that affect the outcome
  - Similar values of the outcome
  - Helps with the plausibility of the counterfactual
    - The treatment group would have followed the path of the control group if it had not been treated in the post-intervention period
- We may try to find this as a naturally occurring phenomena
  - Sometimes this isn't possible – especially with just one treatment group and few control options
- If researchers can "chose" their preferred control group – many options for cherry picking a control group that gives a desired outcome
- The idea of synthetic control is to "create" this control group using a data driven and transparent process

Synthetic control – How it works

- To create a *synthetic control* we use weighted average of possible control groups as counterfactual
  - Weights tells us which control units (“donors”) contribute the most to the counterfactual
  - We get to see weights – decide if they are intuitive
- Choose weights to match outcome and covariates in the treated group during the pre-interventional period
  - We can perfectly replicate pre-trends in the treated group
- Counterfactual – treated group would have continued in the same way as the synthetic control in the absence of treatment

Synthetic control

- Weights chosen so that the weighted average of the untreated control units (or donors) most perfectly matches the pre-treatment trends in the treated unit
- Weights (usually) constrained to be non-negative and sum to 1.
- Negative weights can give better fit – but difficult to interpret and imply extrapolation → treated group predicted outcome predicted to move up when negative weight group moves down

Diff-in-diff – Synthetic control

- Let $\mathbf{X}_1$ be a $(k \times 1)$ vector of pre-intervention characteristics for the treated unit
- Let $\mathbf{X}_0$ be a $(k \times J)$ matrix of same characteristics for the potential control units
  - Variables in X matrices could include the outcomes we want to analyze!
- Then we choose the vector of weights to minimize differences in the $\mathbf{X}$ between the treated unit and the synthetic control
- Choose $\mathbf{W}^* = \left(w_2^*, \ldots w_{J+1}^*\right)$ to minimize some general measure of "distance" between treatment and synthetic control covariates

$$\|\mathbf{X}_1 - \mathbf{X}_0\mathbf{W}\|$$

  - Usually by minimizing mean squared prediction error

- Then estimate the causal effect of the treatment in time t as

$$\hat{\alpha}_{1t} = Y_{1t} - \sum_{i=2}^{J+1} w_i^* Y_{it}$$

Overfitting and Validation

- Model may be over fit – it performs very well in the pretreatment period due to chance
  - Particularly true if there are a lot of potential control groups and pretreatment time period is short
  - May bias results
- Some strategies to address
  - Only fit model in training period subsample of the full pre-intervention period
  - Validate the fitted model by using it to predict outcomes in the data held back from fitting.
  - Use weights that preform well in both fit in training period and out of sample validation
  - Similar approach to machine learning trying to use big data

Hypothesis testing

- How likely is it we see the gap between treated group and synthetic control due to chance ?
- Placebo approach – time dimension and space dimension
  - Try setting different random placebo “intervention” groups and times in control units
  - Estimate the “effect” of these fake intervention by creating synthetic controls form them from the other control groups.
  - Under assumption of no effect of the placebo “interventions” we get a distribution of observed “effects” under the null of no effect
- Reject the null of no effect if estimated effect size for the treatment is outside the 95% confidence interval generated by the placebo effect distribution

Assumptions for Synthetic control methods

- Availability of donor groups to construct controls. Similar to treatment group – driven by similar shocks
  - But did **not** get the treatment
  - Did not get large idiosyncratic shocks in the post-intervention period that did not affect the treatment group
- Synthetic control matches actual outcome data closely in the pre-intervention period – testable
- Size of the effect is large relative to noise
- No anticipation of treatment
  - How would it affect selection of synthetic controls if the treatment group outcomes started to change before the policy change went into effect?
- No spillovers of treatment effect between groups

Example

- Research question: What was the effect of California's Proposition 99 on cigarette sales?
  - Increased cigarette tax by 25 cents/pack
  - Tax revenues earmarked for anti-smoking education
  - Funded anti-smoking media campaigns
  - Clean indoor-air ordinances
  - Passed in 1988
- Policy passed in only one state
  - Are there any states that make good "controls" for California?

- Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program (Abadie et al 2010)

Synthetic Control Example

- *Step 1*: Throw out 11 other states that passed similar tobacco control legislation (contaminates control)
- *Step 2*: Decide which covariates predict post-period outcomes and should be used for determining synth control weights (minimize MSPE)
- Step 3: Look at the weights assigned by the MSPE minimization
  - Are they "reasonable"? What might you consider reasonable?
  - If not, might reconsider whether you're using the right variables to construct weights

Table 1. Cigarette sales predictor means

Variables	California		Average of 38 control states
	Real	Synthetic	
Ln(GDP per capita)	10.08	9.86	9.86
Percent aged 15–24	17.40	17.40	17.29
Retail price	89.42	89.41	87.27
Beer consumption per capita	24.28	24.20	23.75
Cigarette sales per capita 1988	90.10	91.62	114.20
Cigarette sales per capita 1980	120.20	120.43	136.58
Cigarette sales per capita 1975	127.10	126.99	132.81

NOTE: All variables except lagged cigarette sales are averaged for the 1980–1988 period (beer consumption is averaged 1984–1988). GDP per capita is measured in 1997 dollars, retail prices are measured in cents, beer consumption is measured in gallons, and cigarette sales are measured in packs.

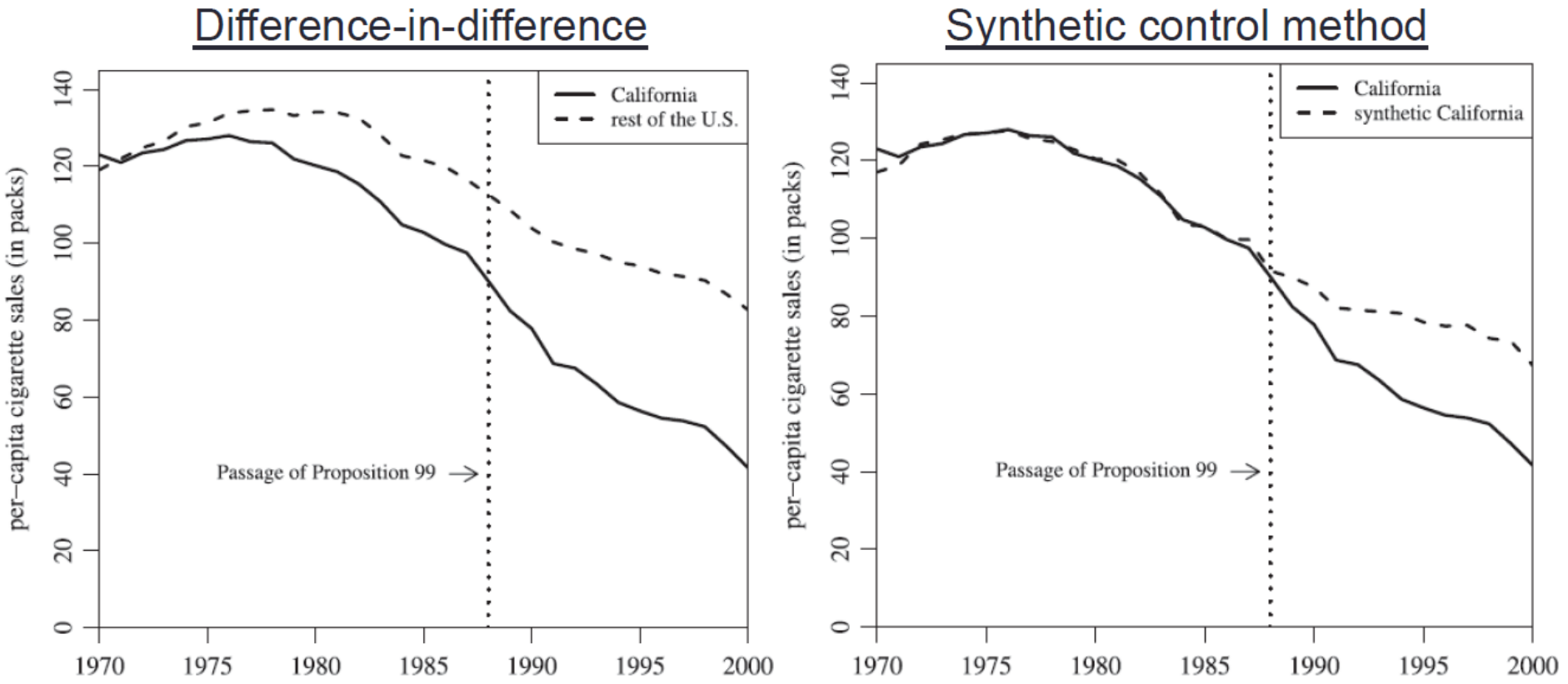
Synthetic Control Example

Table 2. State weights in the synthetic California

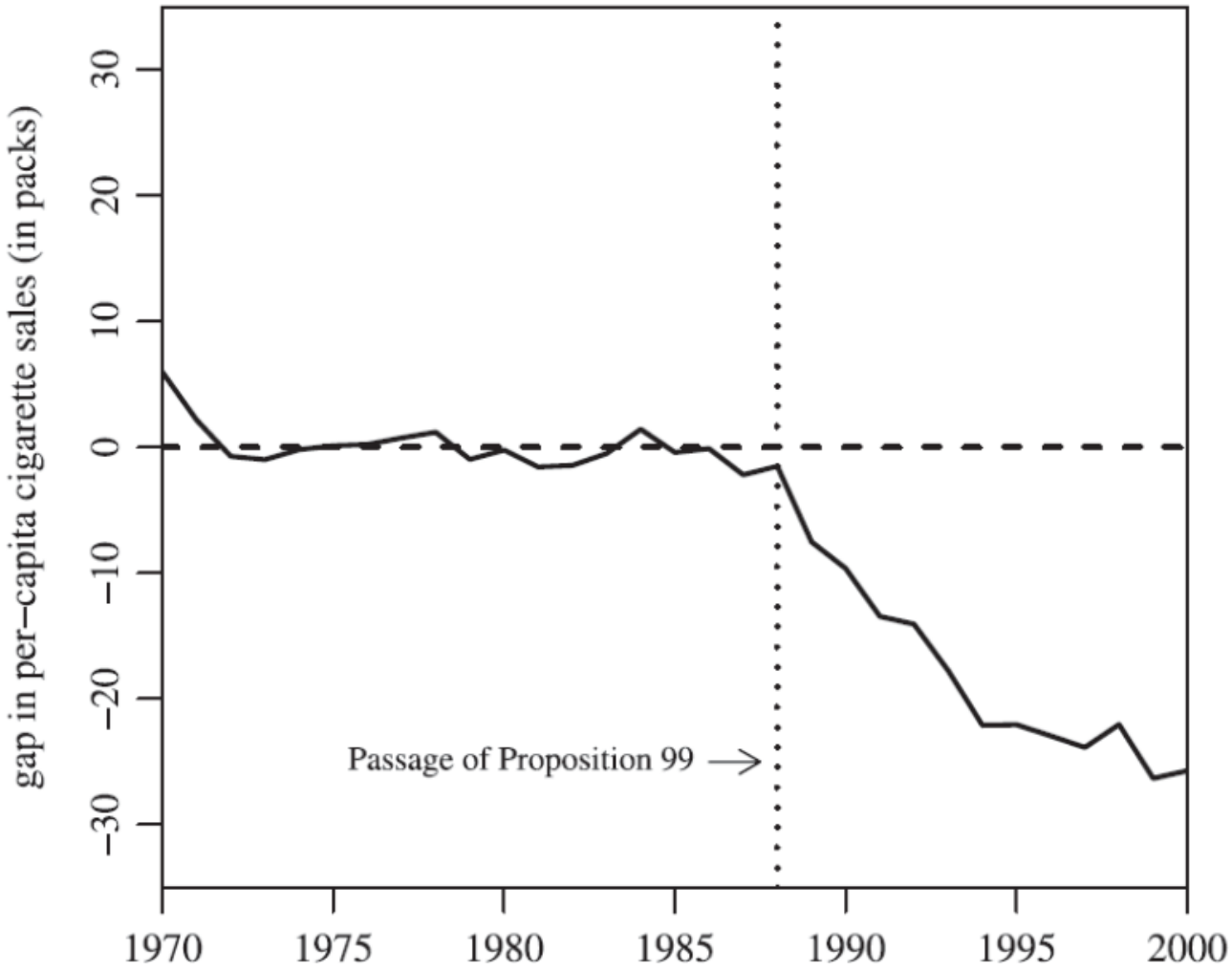
State	Weight	State	Weight
Alabama	0	Montana	0.199
Alaska	–	Nebraska	0
Arizona	–	Nevada	0.234
Arkansas	0	New Hampshire	0
Colorado	0.164	New Jersey	–
Connecticut	0.069	New Mexico	0
Delaware	0	New York	–
District of Columbia	–	North Carolina	0
Florida	–	North Dakota	0
Georgia	0	Ohio	0
Hawaii	–	Oklahoma	0
Idaho	0	Oregon	–
Illinois	0	Pennsylvania	0
Indiana	0	Rhode Island	0
Iowa	0	South Carolina	0
Kansas	0	South Dakota	0
Kentucky	0	Tennessee	0
Louisiana	0	Texas	0
Maine	0	Utah	0.334
Maryland	–	Vermont	0
Massachusetts	–	Virginia	0
Michigan	–	Washington	–
Minnesota	0	West Virginia	0
Mississippi	0	Wisconsin	0
Missouri	0	Wyoming	0

Synthetic control example

- Synthetic control is that it uses high-tech methods to get “right” counterfactual but **you can still make nice visuals**



Synthetic control example – event study



ITS and Synthetic Control example

- You can also use synthetic control methods within ITS – situations where we might have just one treatment group

- Example: Impact of tobacco taxes in South Africa
- South Africa increased tobacco taxes considerably after 1994.
- Advertising and smoking bans were implemented in 2001.
- What was the effect on tobacco consumption?

- Counterfactual – synthetic control from similar countries
- Chelwa, Grieve, Corné van Walbeek, and Evan Blecher. "Evaluating South Africa's tobacco control policy using a synthetic control method." *Tobacco Control* 26.5 (2017): 509-517.

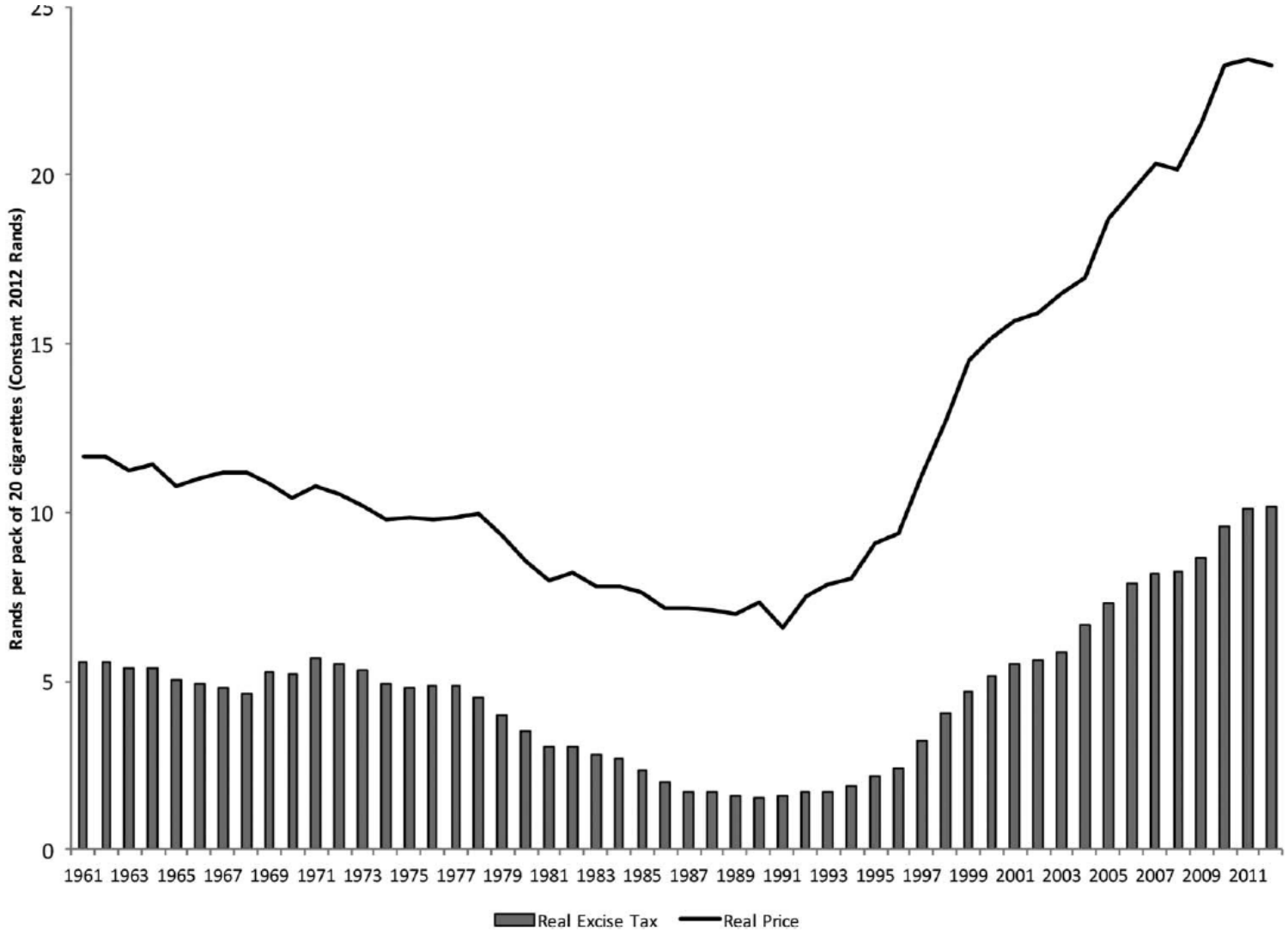


Figure 1 Real excise tax per pack of cigarettes and real price per pack of cigarettes, South Africa 1960–2012.

Synthetic Weights

- Donor pool: other low and middle-income countries
  - Pattern of increasing affordability of cigarettes over the period (price/income per capita)
- Donor pool excludes Thailand (and South Africa) and any countries with missing data
- Variables: Per capita cigarette consumption (the outcome variable), real price of cigarettes, per capita real GDP, per capita alcohol consumption and the proportion of adults in the total population.
- 24 control countries

Table 1 Donor pool and synthetic weights

Country	Weight
Argentina	0.276
Brazil	0.476
Chile	0.146
China	0
Colombia	0
Costa Rica	0
Ecuador	0
Egypt	0
India	0
Indonesia	0
Cote d'Ivoire	0
Jordan	0
Malaysia	0
Morocco	0
Pakistan	0
Panama	0
Peru	0
Philippines	0
Romania	0.007
Senegal	0
Sri Lanka	0
Tunisia	0.094
Uruguay	0
Vietnam	0

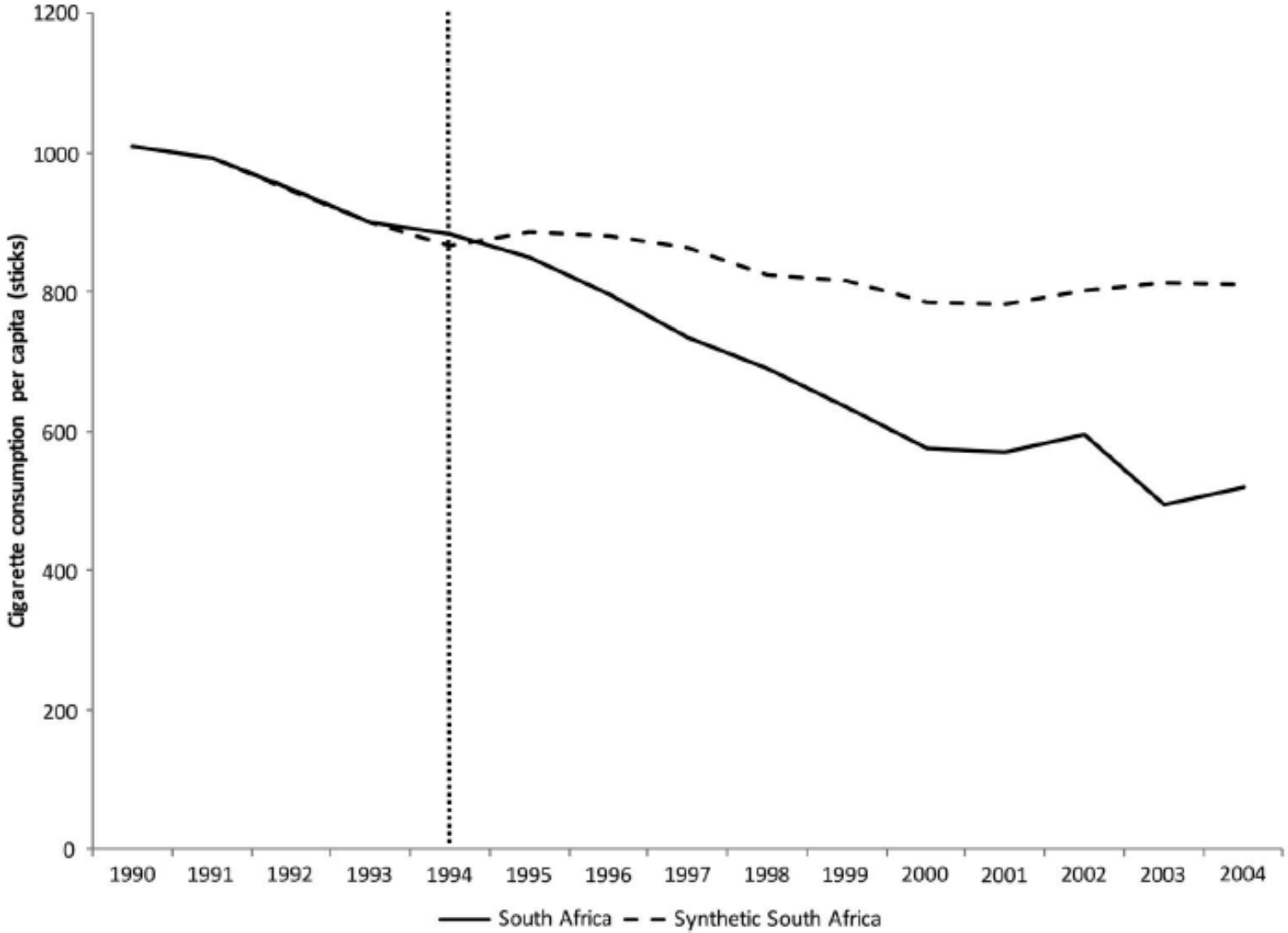


Figure 2 Cigarette consumption per capita, South Africa versus synthetic South Africa.

Summary – Synthetic Control

- In settings with few treated units and few control units, synthetic control can be a powerful way to improve on diff-in-diff
  - Weights can be estimated *before* an intervention even begins; allows for pre-specification of counterfactual
- If used to complement diff-in-diff or ITS: Be transparent! Show regular diff-in-diff / ITS and synth control estimates
  - Weights can be sensitive to specification choices – good to show different options
- Requires a lot of high-quality pre-treatment data to get a good fit
- Inference requires a variety of methods: placebo methods, randomization inference

Additional Slides

Formal difference in differences assumptions

Laura Hatfield has created a terrific website that provides a lot of formal detail surrounding difference in differences for health policy:

https://diff.healthpolicydatascience.org/

Recommended readings

- Roth, Jonathan, Pedro H. C. Sant'Anna, Alyssa, Bilinski and John Poe. 2022. *What's Trending in Difference-in-Differences? A Synthesis of the Recent Econometrics Literature.* Journal of Econometrics (conditionally accepted), available at **https://psantanna.com/files/RSBP_DiD_Review.pdf**

Recommended readings

- Abadie, Alberto. "Using synthetic controls: Feasibility, data requirements, and methodological aspects." Journal of Economic Literature 59.2 (2021): 391-425.