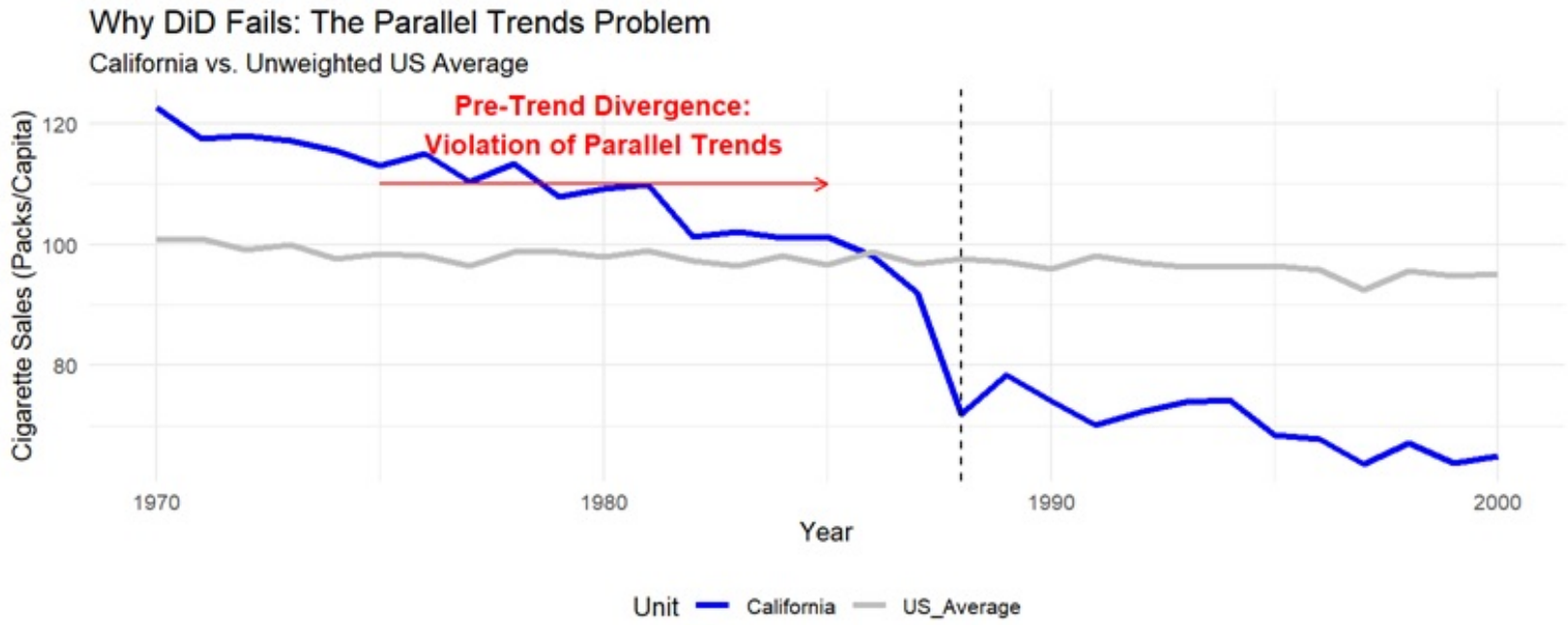
Econometrics of Policy Evaluation: Synthetic Control Method (SCM)

Cristian Huse

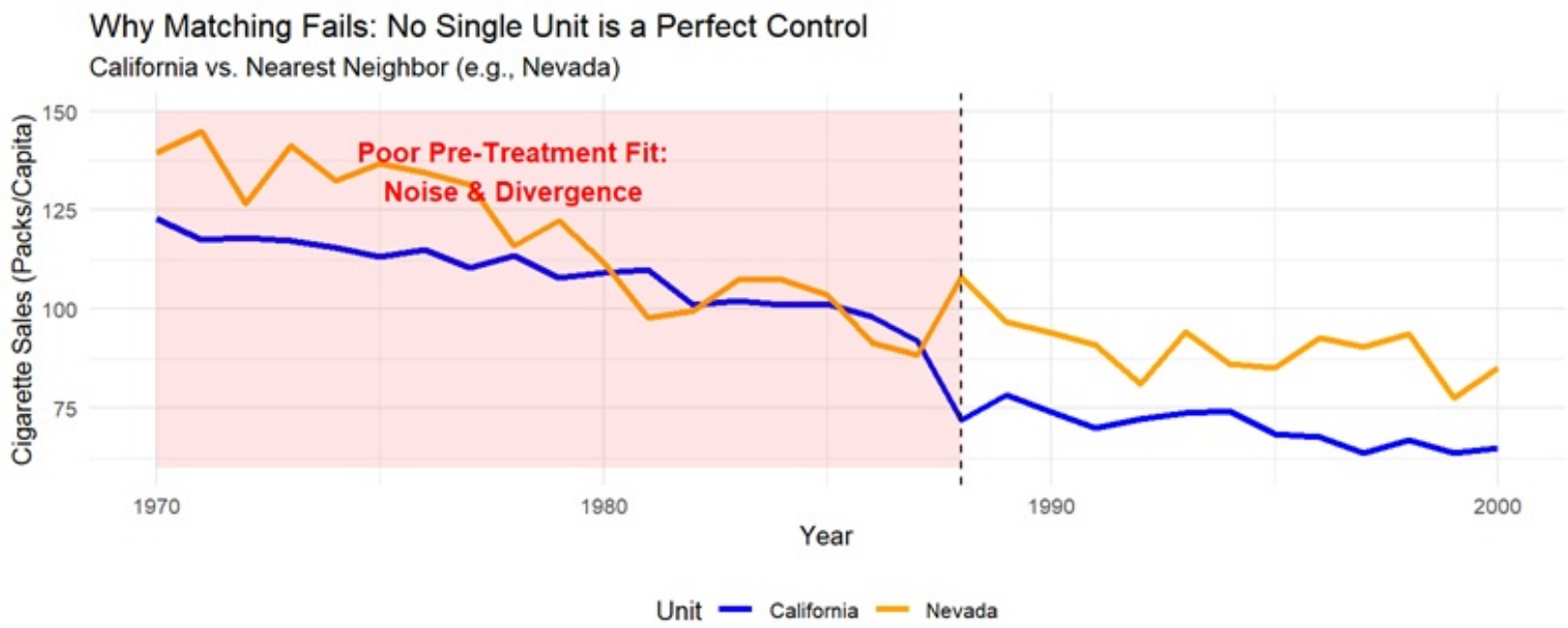
- What is our identification strategy when a treatment (a policy, an event) affects only **one** aggregate unit at one point in time?
- **Examples:**
  - A state passes a unique law (California's Prop 99);
  - A country unifies (German reunification);
  - A region is hit by a natural disaster (Hurricane Iniki on Kauai);
  - A terrorist conflict begins (Basque Country).

- **Flawed Solution 1: Simple Difference-in-Differences**
  - **Example:** Per-Capita Cigarette Sales vs. vs. Year for California vs. "Average of all other US States". California has a more health-conscious population, so sees an overall downward trend while the trend is flat for the average of other states.



  - **Conclusion:** The parallel trends assumption is clearly violated. A simple DiD would be severely biased.

- **Flawed Solution 2: The "Closest Neighbor" (Matching)**
  - **Example:** Per-Capita Cigarette Sales vs. Year for California vs. Nevada (a neighboring state).



  - **Conclusion:** The pre-trends might not match. Even if they do, we are forcing all of our causal inference to rest on the assumption that Nevada is the perfect counterfactual. This is a strong and fragile assumption.

- Now, imagine a setting where there is one treatment group and several potential "donors" (untreated individuals) which can be combined into a control group.
  - Both treated and donor set are observed for "long enough" (↗T).
- Using the pre-treatment data period, the **Synthetic Control Method (SCM)** consists of a matching algorithm that goes through all donors and assigns a weight to each of them.
  - These weights are determined in a way such that the time trend of the outcome for the treated group will be closely tracked by the time trend of the outcome for the weighted average of the control group (the **"synthetic control"** group).
  - **Example DD:** SCM amounts to taking the smooth "US Average" line and weight the states so that the curve bends and twists to match California as closely as possible (instead of equally-weighting them).
- In sum, SCM can be seen as imposing common trends on the treatment and control "group", pre-treatment.
  - Therefore, one can think of SCM as a version of DD with matching.
  - Crucially, the pre-treatment sample has to be large enough.

- SCM is *"arguably the most important innovation in the policy evaluation literature in the last 15 years"* (Athey and Imbens 2017)
- Since this is a less established method, in what follows, we combine different sources, which make complementary points and provide guidance/checklists for the empirical implementation of SCMs.
  - Please check the "Reading Guide" and references towards the end of the slide deck.

The Canonical Application: Abadie et al. (2010)

- **Paper:** Abadie, Diamond, and Hainmueller (2010). Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program. *Journal of the American Statistical Association* 105 (490): 493-505.
  - The paper studies the effects of Proposition 99, a large-scale tobacco control program that California implemented in 1988.
  - Using SCM, the paper documents how, following Proposition 99, tobacco consumption fell markedly in California relative to a comparable synthetic control region.

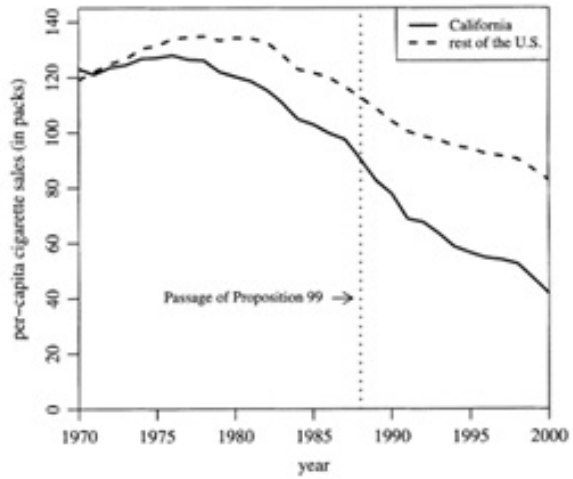


Figure 1. Trends in per-capita cigarette sales: California vs. the rest of the United States.

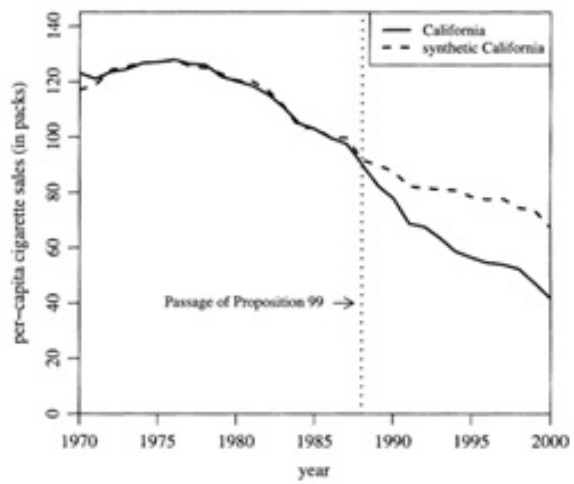


Figure 2. Trends in per-capita cigarette sales: California vs. synthetic California.

How Are the Weights Chosen?

- You want to **minimize a distance** between the treated unit and a **weighted average** of the potential donours.
- The weights are non-negative (i.e., zeroes allowed) and sum to one.
- Matching is done on **pre-treatment data** only!
- This algorithm provides a synthetic unit that **best matches** the treatment unit characteristics in the **pre-treatment period.**

What Do We Match On? (Predictors)

- **Point 1 (Most Important):** We match on pre-treatment lags of the outcome variable (e.g., cigarette sales in 1988, 1980, 1975). This is what forces the pre-treatment trend to match.
- **Point 2:** We also match on other key economic predictors of the outcome (e.g., per-capita income, retail price of cigarettes, etc.).

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.

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

- The method is transparent, see weights above.

Core Assumptions

- **Assumption 1:** We need a long pre-treatment period. SCM is not credible without it.
- We need it to
  1. find the right weights; and
  2. prove our synthetic unit is a good "doppelgänger".
- **Assumption 2 (SUTVA):** The policy in the treated unit can't affect the outcomes in the control units (no spillovers).
  - This is why Abadie et al. (2010) excluded other tobacco states from the donor pool.

The Inference Problem: Is the Gap "Real"?

- We see a gap. But we only have one treated unit. We can't run a t-test.
- How do we know this gap isn't just random noise?

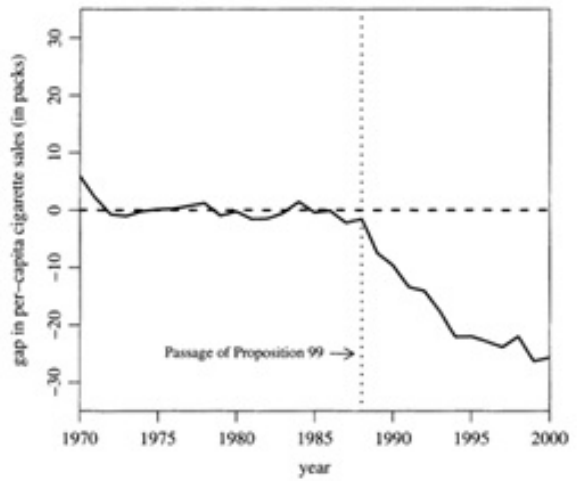


Figure 3. Per-capita cigarette sales gap between California and synthetic California.

The Solution: Placebo (Permutation) Tests

- This is the core of SCM inference:
  - **Step 1:** Run the exact same SCM analysis on every unit in our donor pool (e.g., run it on Colorado, pretending it was treated, and build a "Synthetic Colorado").
  - **Step 2:** This creates a distribution of "placebo gaps" for all the untreated states.
  - **Step 3:** We compare the "real" gap (California's) to this distribution.

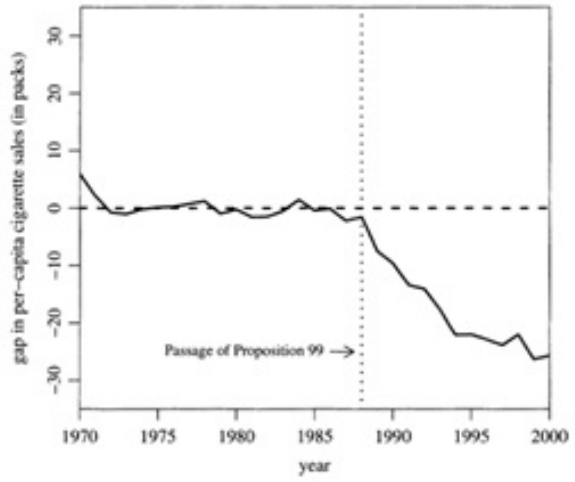


Figure 3. Per-capita cigarette sales gap between California and synthetic California.

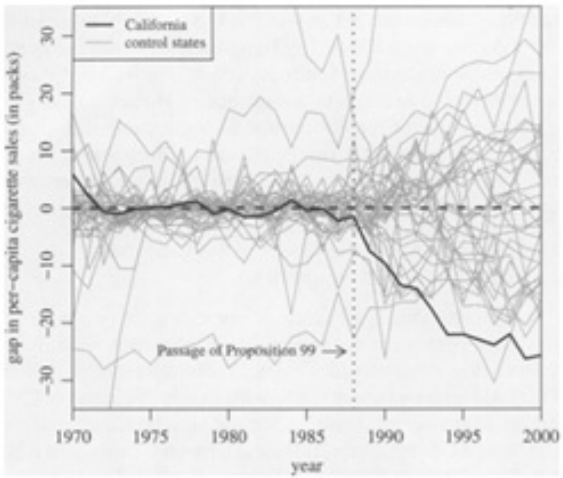


Figure 4. Per-capita cigarette sales gaps in California and placebo gaps in all 38 control states.

- **Interpretation:** This is a sort of "p-value". The effect for

- **Paper:** Bueno & Valente (2019). The effects of pricing waste generation: A synthetic control approach. *Journal of Environmental Economics and Management* 96: 2774-285.
- **Research Question:**
  - Effect of "pay-as-you-throw" waste pricing in a single municipality (Trento, IT).

Figure 2: UW, TW, RW Time Series for Trento (solid) and Synthetic Trento (dotted)

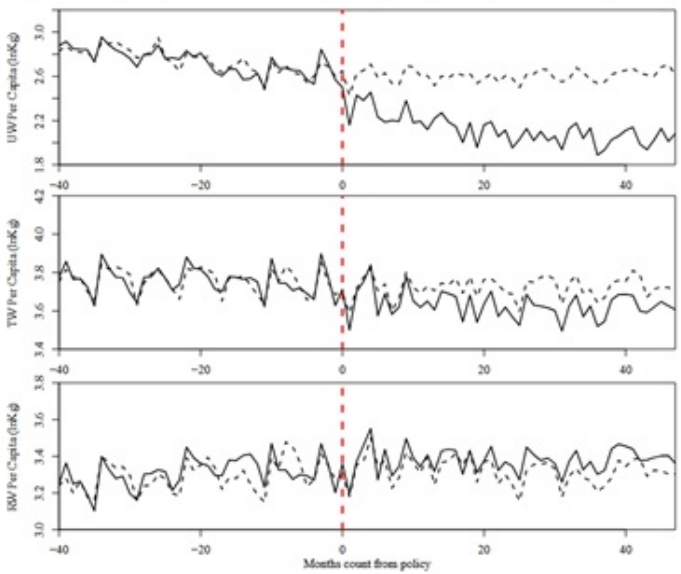
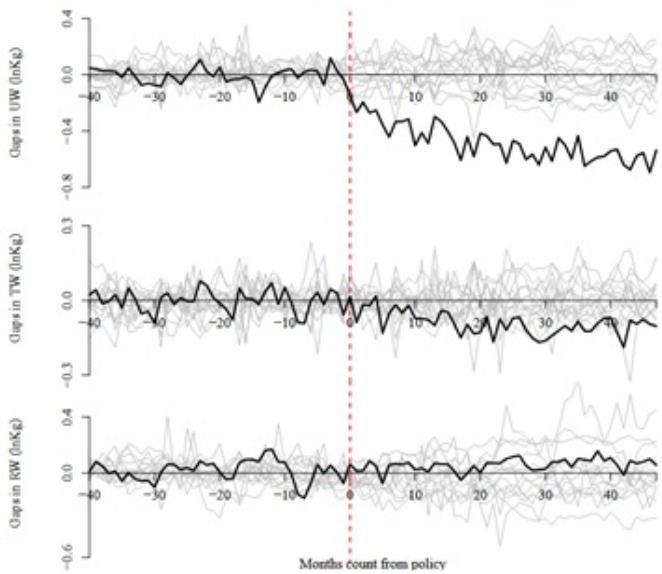


Figure 3: Placebo Tests: Trento (black line) and Control Units (grey lines)



- **Paper:** Abadie, Diamond, and Hainmueller (2015). "Comparative Politics and the Synthetic Control Method." *American Journal of Political Science* 59 (2): 495–510
- **Research Question**
  - What is the effect of the 1990 German reunification on per capita GDP in West Germany?
- **Setup**
  - Outcome: GDP per capita
  - Intervention: 1990 German reunification
  - Treated unit: West Germany
  - Donor pool: OECD countries
  - Covariates: pre-reunification values of predictors of economic growth
- **Estimation**
  - OOS algorithm

- Comparison of West Germany with equally-weighted OECD countries and SCE
  - Note discrepancy pre on Panel A (EW-OECD)

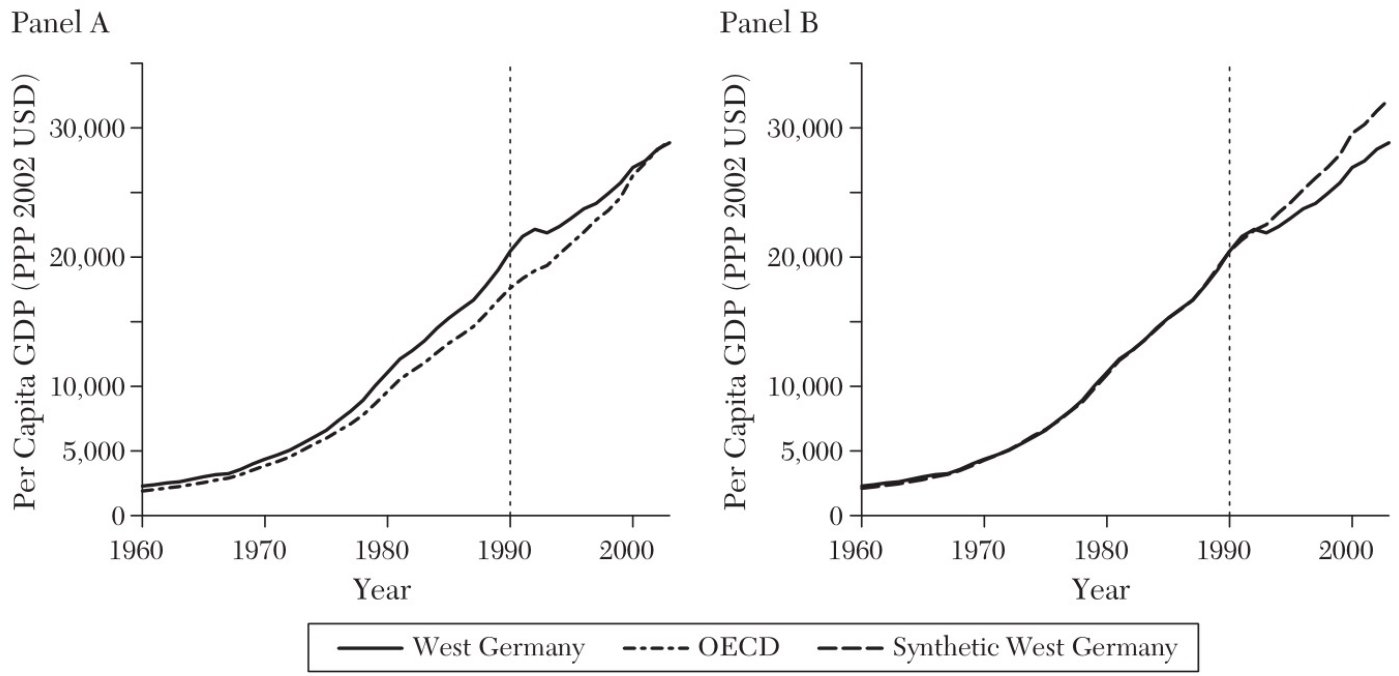


Figure 1. Synthetic Control Estimation in the German Reunification Example

Notes: Panel A compares the evolution of per capita GDP in West Germany to the evolution of per capita GDP for a simple average of OECD countries. In panel B the comparison is with a synthetic control calculated in the manner explained in subsection 3.2. See Abadie, Diamond, and Hainmueller (2015) for details.

- Predictors of different combinations of the donor pool and West Germany
  - Neither EW-OECD nor Austria are good enough whereas SWE matches pretty closely

TABLE 1
ECONOMIC GROWTH PREDICTOR MEANS BEFORE THE GERMAN REUNIFICATION

	West Germany (1)	Synthetic West Germany (2)	OECD average (3)	Austria (nearest neighbor) (4)
GDP per capita	15,808.9	15,802.2	13,669.4	14,817.0
Trade openness	56.8	56.9	59.8	74.6
Inflation rate	2.6	3.5	7.6	3.5
Industry share	34.5	34.4	33.8	35.5
Schooling	55.5	55.2	38.7	60.9
Investment rate	27.0	27.0	25.9	26.6

Note: The first column reports $\mathbf{X}_1$, the second column reports $\mathbf{X}_0\mathbf{W}^*$, the third column reports a simple average of $\mathbf{X}_j$ for the 16 OECD countries in the donor pool, and the last column reports the value of $\mathbf{X}_j$ for the nearest neighbor of West Germany in terms of predictors values. GDP per capita, inflation rate, and trade openness are averages for the 1981–90 period. Industry share (of value added) is the average for 1981–89. Schooling is the average for 1980 and 1985. Investment rate is averaged over 1980–84. See Abadie, Diamond, and Hainmueller (2015) for variable definitions and sources. The nearest neighbor in column 4 minimizes the Euclidean norm of the pairwise differences between the values of the predictors for West Germany and for each of the countries in the donor pool, after rescaling the predictors to have unit variance.

- SC weights for West Germany
  - Note sparsity: combination of some neighbours and trade partners?

TABLE 2
SYNTHETIC CONTROL WEIGHTS FOR WEST GERMANY

Australia	—
Austria	0.42
Belgium	—
Denmark	—
France	—
Greece	—
Italy	—
Japan	0.16
Netherlands	0.09
New Zealand	—
Norway	—
Portugal	—
Spain	—
Switzerland	0.11
United Kingdom	—
United States	0.22

- To be continuted in a few slides...

Variable Selection

- The choice of predictors (covariates) is key for a good performance:
  - *"The credibility of a synthetic control estimator depends on its ability to track the trajectory of the outcome variable for the treated unit for an extended pre-intervention period. Provided that a good fit for pre-intervention outcomes is attained, the researcher has some flexibility in the way pre-intervention outcomes are incorporated in* X_1 *and* X_0*"* (Abadie 2021, p. 402)
- Data-driven methods (in the spirit of the OOS algorithm above) for variable selection can be used to compare alternative sets of predictors
- **Reminder.** Only pre-intervention data should be used!

- **Example. Western Germany**
  - *"As reported in table 1, the set of predictors in X_1 and X_0 includes average per capita GDP in 1981-90, and no other pre-intervention outcome. Notice, however, that the resulting synthetic control is able to track the trajectory of per capita GDP for West Germany for the entire 1960-90 pre-intervention period. This happens because per capita GDP figures for OECD countries strongly co-move in time across countries. This co-movement of the outcome variable of interest across the different units in the data is exactly what synthetic controls are designed to exploit. It makes it possible to match the entire trajectory of GDP per capita for West Germany by fitting only the average level of GDP per capita in the 1981-90 period. Given this premise, one potential advantage from using a summary measure of prereunification GDP per capita to calculate the synthetic control for West Germany (as opposed to, say, including all ten different annual values of GDP per capita for 1981-90 as predictors) resides in a higher sparsity of the resulting synthetic control."* (Abadie 2021, p. 402)

TABLE 1
ECONOMIC GROWTH PREDICTOR MEANS BEFORE THE GERMAN REUNIFICATION

	West Germany (1)	Synthetic West Germany (2)	OECD average (3)	Austria (nearest neighbor) (4)
GDP per capita	15,808.9	15,802.2	13,669.4	14,817.0
Trade openness	56.8	56.9	59.8	74.6
Inflation rate	2.6	3.5	7.6	3.5
Industry share	34.5	34.4	33.8	35.5
Schooling	55.5	55.2	38.7	60.9
Investment rate	27.0	27.0	25.9	26.6

Note: The first column reports $\mathbf{X}_1$, the second column reports $\mathbf{X}_0\mathbf{W}^*$, the third column reports a simple average of $\mathbf{X}_j$ for the 16 OECD countries in the donor pool, and the last column reports the value of $\mathbf{X}_j$ for the nearest neighbor of West Germany in terms of predictors values. GDP per capita, inflation rate, and trade openness are averages for the 1981–90 period. Industry share (of value added) is the average for 1981–89. Schooling is the average for 1980 and 1985. Investment rate is averaged over 1980–84. See Abadie, Diamond, and Hainmueller (2015) for variable definitions and sources. The nearest neighbor in column 4 minimizes the Euclidean norm of the pairwise differences

Advantages of SCM

- The use of only pre-intervention data precludes data mining
- Sparsity helps interpretability
  - Aided by non-negativity and adding-up constraints

TABLE 2
SYNTHETIC CONTROL WEIGHTS FOR WEST GERMANY

Australia	—
Austria	0.42
Belgium	—
Denmark	—
France	—
Greece	—
Italy	—
Japan	0.16
Netherlands	0.09
New Zealand	—
Norway	—
Portugal	—
Spain	—
Switzerland	0.11
United Kingdom	—
United States	0.22

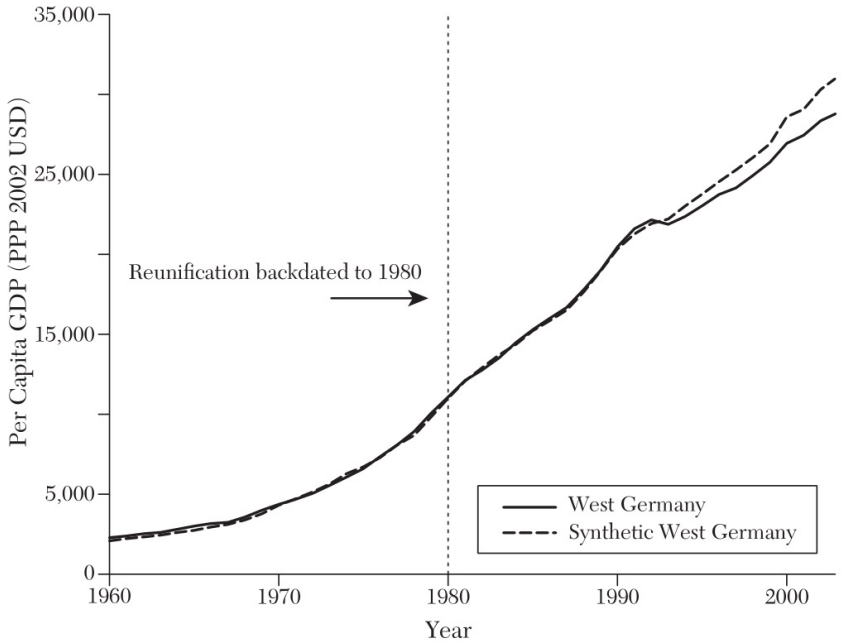
TABLE 3
REGRESSION WEIGHTS FOR WEST GERMANY

Australia	0.12
Austria	0.26
Belgium	0.00
Denmark	0.08
France	0.04
Greece	−0.09
Italy	−0.05
Japan	0.19
Netherlands	0.14
New Zealand	0.12
Norway	0.04
Portugal	−0.08
Spain	−0.01
Switzerland	0.05
United Kingdom	0.06
United States	0.13

- Transparency of the counterfactual
  - Amplified if/due to sparse solution

Robustness

- Robustness is key for the credibility of results. Here we summarize two exercises
- **Placebo Test**
  - To show robustness, one could change the time of intervention slightly and show the that results are largely the same
  - Another idea is to change the timing more dramatically and show there are no results, as in Abadie (2021)



- **Robustness Tests**
  - Choice of predictors
    - e.g. remove one at a time at show the effects are largely robust
  - Choice of units in the donor pool
    - e.g., provide leave-one-out estimates and show the effects are largely robust (see below)

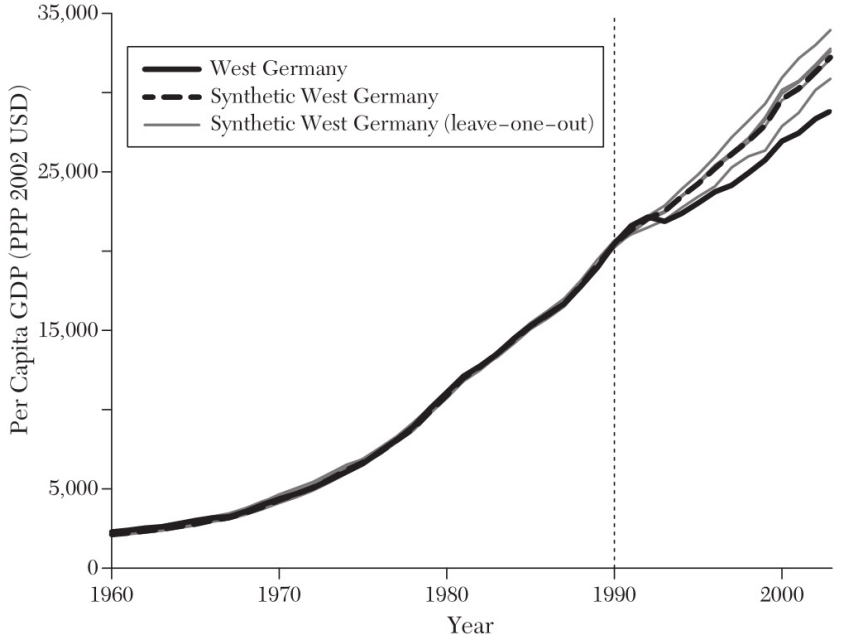


Figure 4. Leave-one-out Estimates of the Effect of the 1990 German Reunification

[We follow Abadie and Vives-i-Bastida (2021) closely]

1. Closely fitting a highly volatile series *(e.g., individual)* is likely the result of over-fitting at work, especially if it happens over a short pre-intervention period. Synthetic controls were designed for settings with **aggregate series**, where aggregation attenuates the magnitude of the noise. *(i.e., averages smooth out noisy individual series)*

2. A good control unit must closely reproduce the trajectory of the outcome variable for the treated unit over an **extended pre-intervention period** *(i.e., time series dimension is key)*. A good fit, if it is the result of a secular agreement between the treated and the synthetic control units in their responses to unobserved factors, should persist in time.

3. **A larger donor pool is not necessarily better than a smaller one** (since solutions are typically sparse)**.** Adopting a small donor pool of untreated units that are close to the treated unit in the space of the predictors helps reduce over-fitting and interpolation biases. *(i.e., restatement of classical result that parsimonious models often outperform larger ones out-of-sample)*

4. **Sparsity makes synthetic controls interpretable.** *(i.e., fewer explanatory factors)*

5. **Covariates matter** *(i.e., this is matching, so observable characteristics matter)*. A component of Z_i that is not controlled for (that is, not included in X_i) is effectively thrown into μ_j *(i.e., less observables mean more unobservables)*. Fitting Z_j is easier than fitting μ_j.

6. **Fit matters.** The bound on the bias is predicated on close fit. A deficient fit raises concerns about the validity of a synthetic control (see, however, Ferman, 2021; Ferman and Pinto, 2021, for exceptions and qualifications on this rule).

7. **Out-of-sample validation is key.** The goal of synthetic controls is to predict the trajectory of the outcome variable for the treated unit in the absence of the intervention of interest. The quality of a synthetic control can be assessed by measuring predictive power in pre-intervention periods that are left out of the sample used to calculate the synthetic control weights.

Implementation in R

- The key R package is `Synth`.
  - `dataprep()`: The function to set up the data (define treated unit, donor pool, predictors, time).
  - `synth()`: The function that runs the optimization and finds the weights.
  - `path.plot()`: Generates the main graph.
  - `gaps.plot()`: Plots the difference (the "gap") between the treated and synthetic units over time.
- Alternatives in R:
  - `gsynth, tidysynth`
- Stata:
  - `synth, synth_runner`

Take-aways

- SCM is a data-driven, transparent method for $N = 1$ case studies.
  - It imposes rigour on the choice of the control group.
  - Results are intuitive, transparent, and produce reliable estimates for a variety of data generating processes.
  - It avoids DD's parallel trends assumption by building a unit that satisfies it. More robust than picking a single "control" unit.
- Credibility rests on:
  1. A long pre-treatment period.
  2. A good pre-treatment "fit" (low RMSPE, or Root Mean Squared Prediction Error).
  3. A sensible "donor pool".
  4. Passing placebo ("spaghetti plot") tests.
- **Final Word:** When you only have one treated unit and suspect parallel trends will fail, SCM should be the first tool you reach for.

Reading Guide

- In order of complexity:
  - Abadie (2021)
  - Athey-Imbens (2017) [connection with ML, Causality in general]
  - Mitze et al (2020) [to deepen and check the current best practices]

References

- Abadie, A. (2021). Using Synthetic Controls: Feasibility, Data Requirements, and Methodological Aspects. *Journal of Economic Literature* 59(2), 391-425.
- Abadie, A., A. Diamond, and J. Hainmueller (2010). Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program. *Journal of the American Statistical Association* 105(490): 493-505.
- Abadie, Diamond, and Hainmueller (2015). Comparative Politics and the Synthetic Control Method. *American Journal of Political Science* 59 (2): 495?510
- Abadie, A. and J. Gardeazabal (2003). The Economic Costs of Conflict: A Case Study of the Basque Country. *American Economic Review* 93(1): 113-132.

References

- Abadie, A. and J. Vives-i-Bastida (2021). Synthetic Controls in Action. Working paper.
- Athey, S., and G. W. Imbens. 2017. The State of Applied Econometrics: Causality and Policy Evaluation. *Journal of Economic Perspectives* 31 (2): 3-32.
- Mitze et al (2020). Face masks considerably reduce COVID-19 cases in Germany. *Proceedings of the National Academy of Sciences* 117 (51): 32293-32301.

Overview

- **Paper:** Alberto Abadie and Javier Gardeazabal (2003). The economic costs of conflict: A case study of the Basque country. American Economic Review, 93(1): 113-132.
  - What is the effect of terrorism in the Basque Country in the 1960s on economic activity?
  - The paper compares growth in the Basque region to growth in a synthetic control group
  - Not surprisingly, the finding is that GDP in the Basque Country was negatively impacted by the conflict

Empirical Strategy

- Abadie and Gardeazabal match the Basque Country to seventeen other regions
  - The synthetic control algorithm creates weights for each of those regions in order to get common trends pre-treatment
  - In addition to the trends themselves, matching variables used were population density, education, and investment levels
  - Since the conflict started around 1970, matching is done using data up to 1969
  - The regions of Catalonia and Madrid were the strongest matches and concentrate much of the weight, which are are then applied across the entire time period

Results

- Note similarity between Basque Country and synthetic control group

TABLE 3—PRE-TERRORISM CHARACTERISTICS, 1960'S

	Basque Country (1)	Spain (2)	"Synthetic" Basque Country (3)
Real per capita GDP[a]	5,285.46	3,633.25	5,270.80
Investment ratio (percentage)[b]	24.65	21.79	21.58
Population density[c]	246.89	66.34	196.28
Sectoral shares (percentage)[d]			
Agriculture, forestry, and fishing	6.84	16.34	6.18
Energy and water	4.11	4.32	2.76
Industry	45.08	26.60	37.64
Construction and engineering	6.15	7.25	6.96
Marketable services	33.75	38.53	41.10
Nonmarketable services	4.07	6.97	5.37
Human capital (percentage)[e]			
Illiterates	3.32	11.66	7.65
Primary or without studies	85.97	80.15	82.33
High school	7.46	5.49	6.92
More than high school	3.26	2.70	3.10

Sources: Authors' computations from Matilde Mas et al. (1998) and Fundación BBV (1999).
[a] 1986 USD, average for 1960–1969.
[b] Gross Total Investment/GDP, average for 1964–1969.
[c] Persons per square kilometer, 1969.
[d] Percentages over total production, 1961–1969.
[e] Percentages over working-age population, 1964–1969.

Results

- The values of the outcome are very similar in the pre-treatment period for the treated and synthetic control group

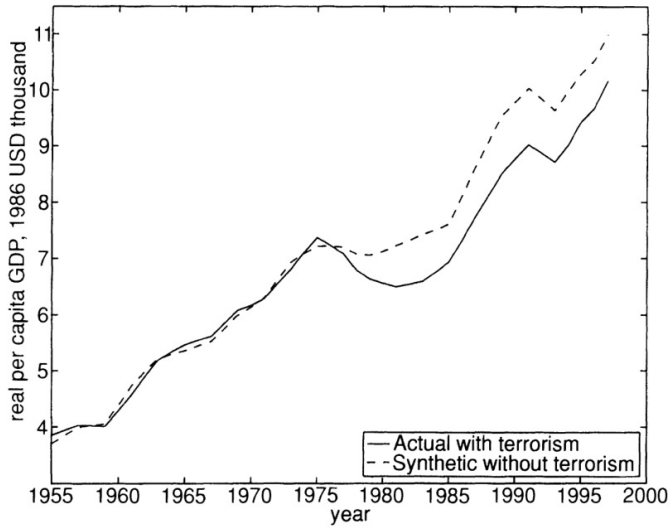


FIGURE 1. PER CAPITA GDP FOR THE BASQUE COUNTRY

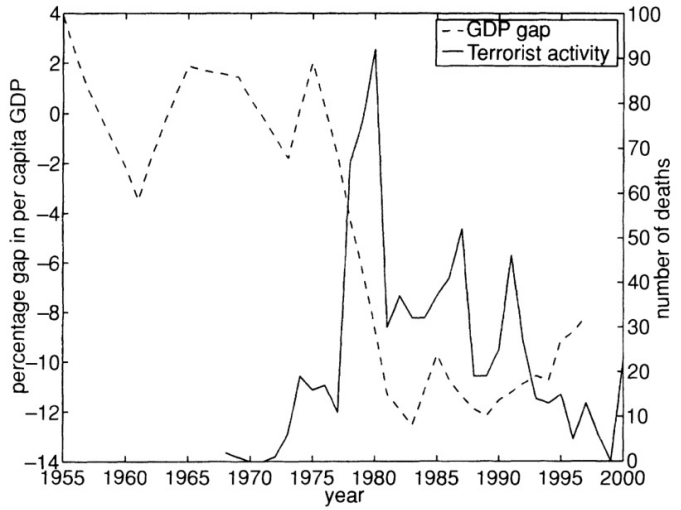


FIGURE 2. TERRORIST ACTIVITY AND ESTIMATED GAP

- They continue to trend very closely together for a few years after treatment starts, at least until the conflict intensified
- Thereafter, the Basque Country's GDP dips, and more strongly so than for the synthetic control group
- In sum, the conflict takes its toll `abc`

Overview

- **Paper:** Mitze et al (2020). Face masks considerably reduce COVID-19 cases in Germany. Proceedings of the National Academy of Sciences 117 (51): 32293-32301.
- Face mask wearing in public transport and shops became mandatory at different points in time across German states/regions

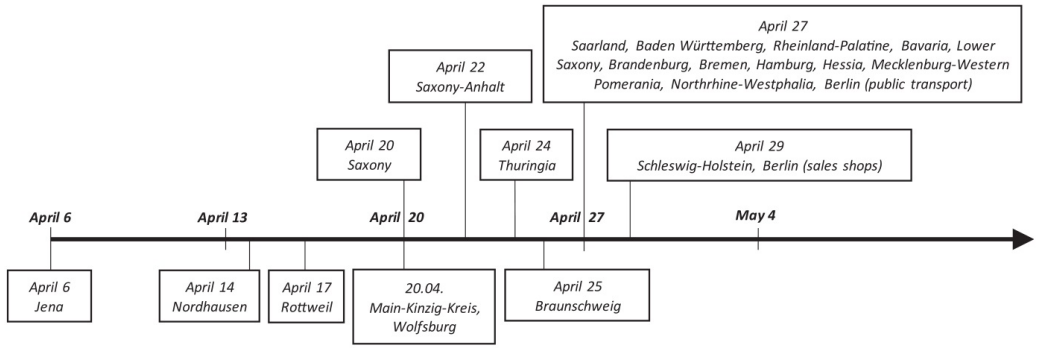


Fig. 5. The timing of mandatory face mask wearing in German federal states (*Top*) and individual regions (*Bottom*). The figure shows the regional variation in the introduction of face masks in public transport and shops over time. While text boxes above the timeline on the horizontal axis indicate the timing when the wearing of face masks became compulsory in the respective federal states (NUTS1 level), text boxes below the timeline identify individual NUTS3 regions that have anteceded the general introduction of face masks at the federal state level. The first NUTS3 region that introduced mandatory face masks in Germany was Jena on 6 April. By 29 April face masks had become mandatory in all German regions. (See *SI Appendix*, section A.1 for more background.)

- This papers essentially uses SCM to compare a treated unit (e.g., Jena) with untreated units in an "optimal" way
  - Results point to a 15-75% reduction in cases within 20 days of the introduction of mask mandates

- **Data**
  - 401 municipal district (Landkreise)
  - Variables considered include prior COVID- 19 cases, the demographic composition, and the local health care system of municipalities
- **Assumptions**
  - Exogeneity of introduction (timing)
  - Time window is 20 days (see timing for context)

- **Timing**
  - 6th April 2020 (major measure, overall 1st-10th April 2020): Masks mandatory in Jena
  - Six further regions made masks compulsory before the introduction at the federal state level
  - 20th-29th April 2020: Face masks became mandatory in all federal states

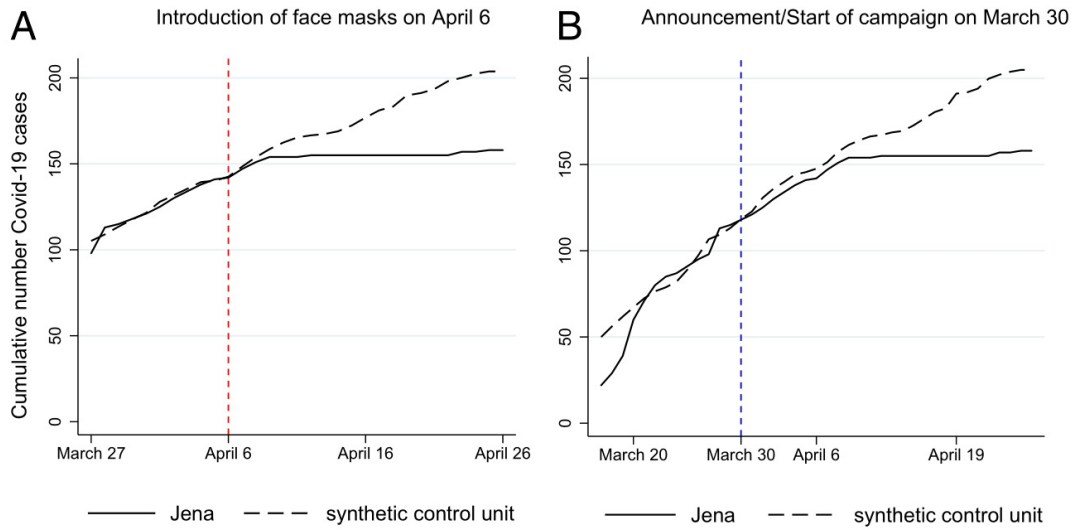


Fig. 1. Treatment effects of the mandatory introduction of face masks (6 April) and of its announcement (30 March) in Jena. The figure shows the development of the cumulative number of COVID-19 cases in Jena (treated region, black solid line) compared to a synthetic control group (gray dashed line) over time. (Details on the construction of the synthetic control group and SCM estimation are given in *Method and Data*.) Both panels distinguish between a pretreatment and treatment period. In *A*, the treatment period starts on 6 April when face masks became mandatory in public transport and shops. The start of the treatment period is indicated by the dashed vertical line (red). In *B*, the treatment period is set to begin on 30 March as starting date of the local campaign in Jena to wear face masks in public. The start of the treatment period is indicated by the dashed vertical line (blue). The panels show that the conclusion is independent of the starting dates: Face masks strongly reduced the number of COVID-19 cases in Jena.

- **Empirical Strategy**
  - SCM applied to Jena as baseline
  - Robustness by looking at all regions that introduced masks by 22 April ($\sim$ 8% of all German regions)
- **Results**
  - Larger for Jena
  - Smaller for robustness set

Table S13: Summary of treatment effects of face mask introduction in Germany

Difference between treated region(s) and synthetic control group(s)	**Single Treatment (Jena)**	**Multiple treatments (all districts)**	**Multiple treatments (larger cities)**
Absolute change in cumulative number of Covid-19 cases over 20 days	-46.9	-7.0	-28.4
Percentage change in cumulative number of Covid-19 cases over 20 days	-22.9%	-2.6%	-8.9%
Percentage change in newly registered Covid-19 cases over 20 days	-75.6%	-15.7%	-51.2%
Difference in daily growth rates of Covid-19 cases in percentage points	-1.28%	-0.13%	-0.46%
Reduction in daily growth rates of Covid-19 cases (in percent)	70.6%	14.0%	47.3%

- **Robustness Checks**
  - With the results in the pocket, it's time to perform a battery of robustness checks...
    - **Sensitivity analysis:** length of the pre-intervention period (Cross-Validation Tests), composition of the control pool (Changing the Donor Pool)
    - **SCM assumptions:** unobserved macro effects shared by many regions (Placebo-in-Space Tests), test for anticipation effects potentially caused by other public health measures (Placebo-in-Time Tests), Diff-in-diff
- Bottom line: Alternative explanations are ruled out

- **Cross-validation tests** (Lag of predictors)
  - Recall that need to find
    - Left: largely robust results to changes in lags
    - Right: no results after a placebo-in-time (lack of anticipation)

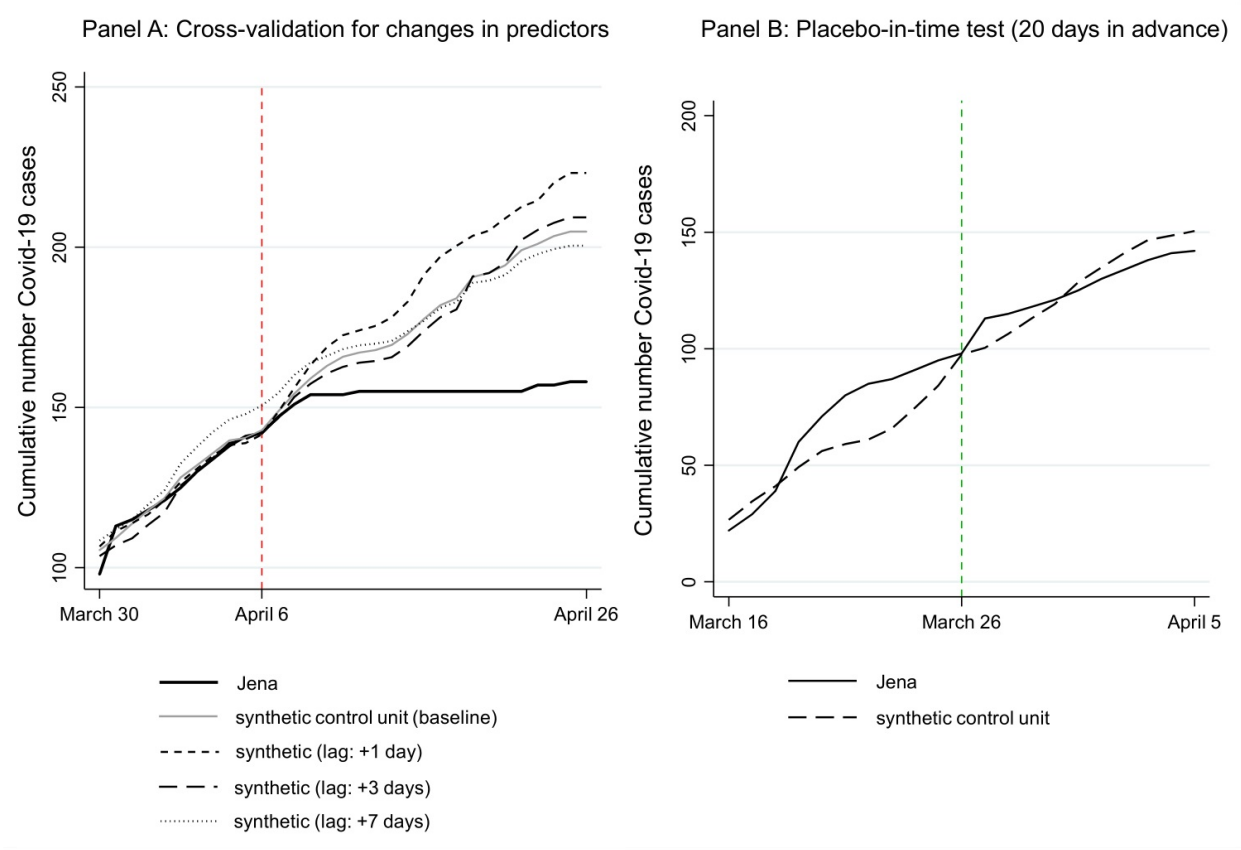


Figure S11: Cross-validation test for changes in time-varying predictor variables and placebo-in-time test

- Mission accomplished

- **Changes in donor pool**
  - Recall that need to find
    - Left: largely robust results when dropping a particular demographic
    - Right: largely robust results when dropping a part of the donor pool

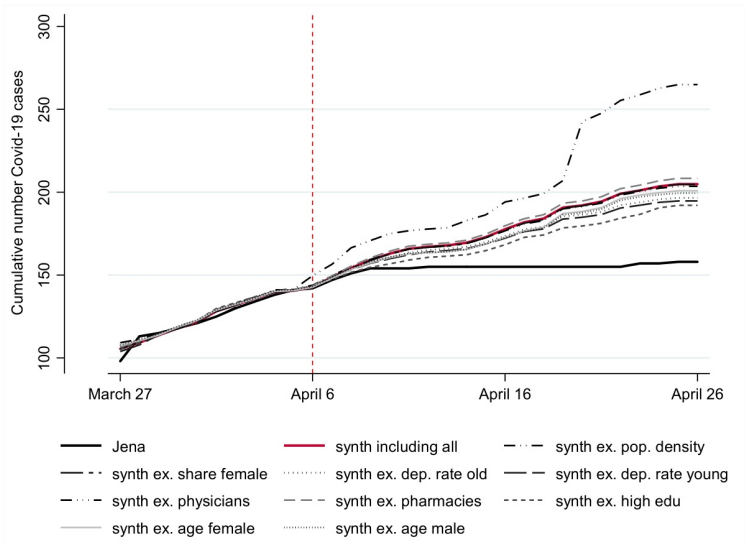


Figure S12: Cross-validation test for changes in the set of time-constant predictors

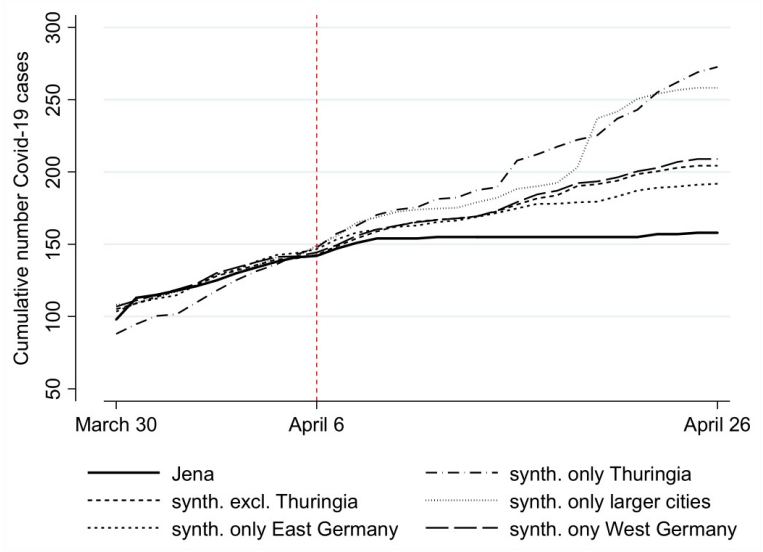


Figure S13: Treatment effects for changes in donor pool used to construct synthetic Jena

- Mission accomplished

- **Placebo-in-space tests**
  - Check whether other cities that did not introduce face masks on 6 April have experienced a similar decline in the number of registered COVID-19 cases
  - If this had been the case, the TE might have been driven by (latent) factors other than by face masks

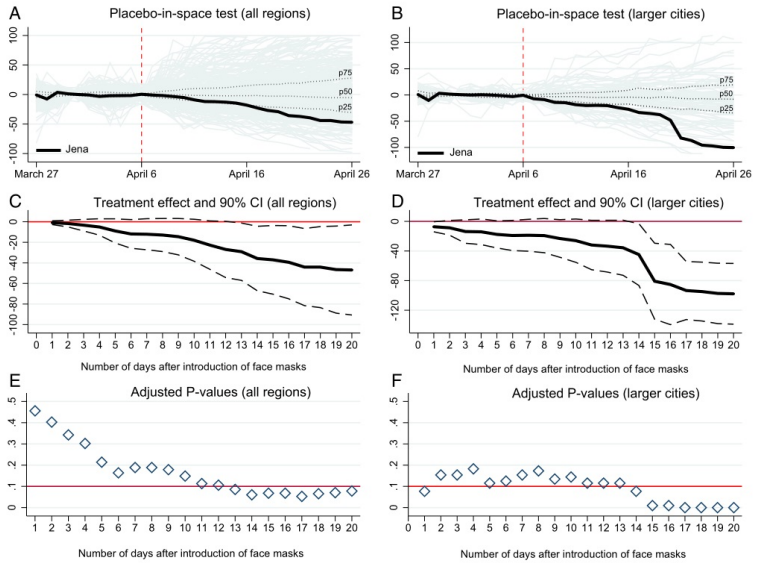


Fig. 2. Comprehensive placebo-in-space tests for the effect of face masks on COVID-19 cases. The figure compares the difference in the cumulative number of COVID-19 cases between Jena and its synthetic control group with the differences between all other regions in the donor pool and their respective controls over time. Differences have been calculated for the treatment period after 6 April when Jena de facto introduced mandatory face masks. For all other regions, the reported differences in the treatment period constitute pseudo-treatment effects since these regions had not introduced face masks back then. In *A*, the black solid line measures the treatment effect for Jena. The solid gray lines measure the pseudo-treatment effects for all other German regions in the donor pool (see *Method and Data* for details on the specification of the donor pool). Plots are shown for donor regions with an acceptable pretreatment root-mean-square prediction error (RMSPE), that is, less than 10 times the RMSPE of Jena; dotted lines indicate the median (p50), 25th percentile (p25), and 75th percentile (p75) of pseudo-treatment effects. In *B*, the donor pool is reduced to comprise only larger cities (*kreisfreie Städte*). *C* and *D* also plot the treatment effect for Jena and 90% confidence intervals (gray dashed lines) for the full sample of regions and the subsample of larger cities, respectively. Confidence intervals are constructed on the basis of pseudo *P* values as shown in *E* and *F* for the first 20 d after the introduction of face masks in Jena. These *P* values are adjusted for the pretreatment match quality (see *Method and Data* for details). The red horizontal line in *E* and *F* indicates a threshold *P* value of 0.1.

  - Mission accomplished
  - **Note.** TEs only become statistically significant roughly 2 weeks after the introduction of face masks

- **Placebo-in-time tests**
  - Recall that need to find no significant treatment effects for Jena prior to the introduction of face masks on 6 April or its announcement on 30 March

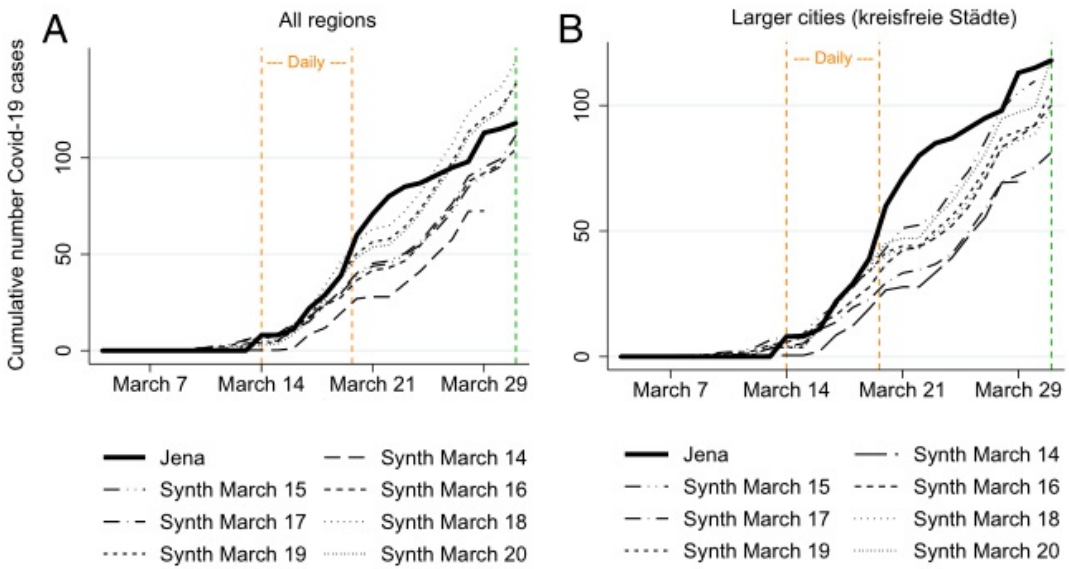


Fig. 3. Placebo-in-time tests for (pseudo) treatment effects in the period 14 to 20 March. The figure shows the empirical development of the cumulative number of COVID-19 cases in Jena (treated region, black solid line) and the estimated development in different synthetic control groups over time. The key difference between these synthetic control groups lies in the starting point of the (pseudo) treatment period. The starting point varies on a daily basis between 14 and 20 March. In *A*, the full sample of German regions is used as the donor pool for the construction of synthetic control groups (see *Method and Data* for details on the specification of the donor pool). Vertical dashed lines (orange) indicate the time corridor in which the respective (pseudo) treatment periods start. In *B*, the donor pool is reduced to comprise only larger cities (*kreisfreie Städte*). Again, vertical dashed lines (orange) indicate the time corridor in which the respective (pseudo) treatment periods start.

  - Mission accomplished for right, questionable for left graph

- **Difference-in-difference tests**
  - Estimate incremental difference-in-differences (IDiD) models – M regressions of the form

$$Y_{it} = \beta \times \Delta Y_{it-1} + \gamma \times base_{it} + \delta_m \times add_{it}^m + D_{dow} + \mu_i + \Psi_{k(t)} + e_{it}$$

  - Outcome: log of cumulative number of Covid-19 in municipal m at day t
  - $base_{it}$ is the the baseline treatment dummy
  - add_{it}^m is the additional treatment dummy from day m onwards
  - μ_i are municipality fixed-effects, D_{dow} are day-of-week dummies, $\Psi_{k(t)}$ are calendar weeks fixed-effects, e_{it} is an i.i.d. error term
  - sample period is 14th March-6th May
  - m varies between 15th March and 25th April
- See SI for estimation details, e.g., first-step probit

- **Difference-in-difference estimates**
  - Panel A: total treatment effects $(\gamma + \delta_m)$
  - Panel B: add-on effects (δ_m)
  - Panel C: expected timing of public health measure considering a total delay of 19 days for incubation period and reporting lag

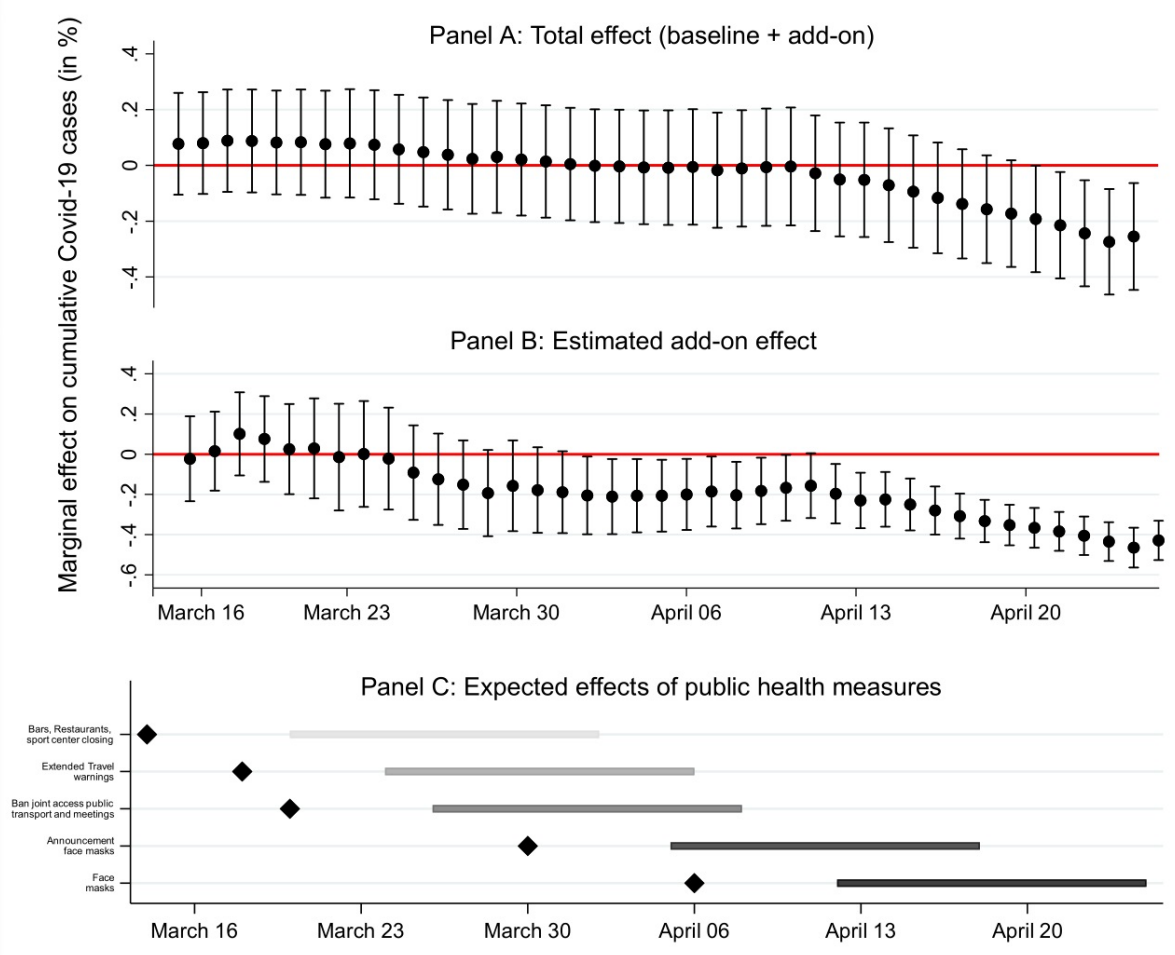


Figure S16: Estimated effects from incremental difference-in-difference (IDiD) model

- **Difference-in-difference estimates**
  - TE estimates become significant approximately one-two weeks after the introduction of face masks on 6th April
  - Given the 19-day delay, findings are consistent with an effect on the announcement of the policy on 30th March (30th March, see Panel C)
  - The magnitude of the effect is a reduction of 20% in cumulative cases (consistent with SCM)

Results for Other Regions

- Paper examines all regions which introduced face masks on or before 22 April

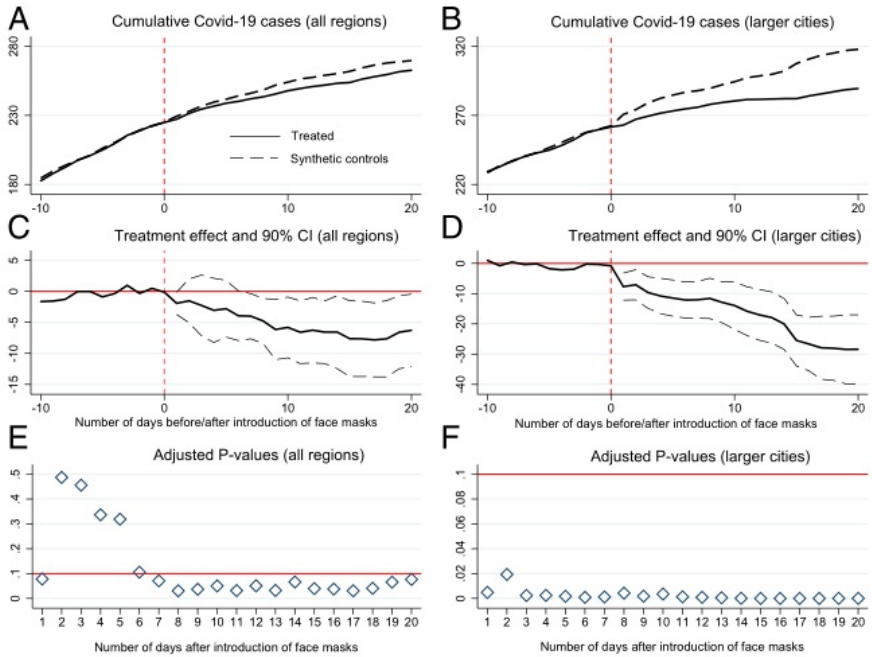


Fig. 4. Average treatment effects for the introduction of face masks with multiple treated units. The figure shows the average development of the cumulative number of COVID-19 cases for treated units (defined as all regions that introduced face masks on or before 22 April), for their synthetic control groups and estimated treatment effects. *A* plots the average number of cumulative COVID-19 cases in treated regions (black solid line) and synthetic controls (black dashed line) for the full sample of German regions in the donor pool (see *Method and Data* for details on the specification of the donor pool). The horizontal axis plots the number of days before and after the start of the treatment (mandatory introduction of face masks). The vertical dashed line (red color) indicates the start of the treatment period, which is allowed to vary by treated regions. In *B*, the number of treated and the donor pool for control regions is limited to larger cities (*kreisfreie Städte*). *C* and *D* plot the estimated average treatment effects, that is, average reduction in the cumulative number of COVID-19 cases (black solid lines) over time joint with 90% confidence intervals (gray dashed lines) for the two samples. Confidence intervals are constructed on the basis of pseudo *P* values as shown in *E* and *F* for the first 20 d after the start of the treatment. These (one-sided) *P* values are adjusted for the pretreatment match quality (see *Method and Data* for details). Inference has been conducted on the basis of a randomly drawn sample of 1 million placebo averages. The red horizontal line in *E* and *F* indicates a threshold *P* value of 0.1.

- TEs turn significant after roughly 1 week for the overall sample and suggest anticipation effects of face masks in urban areas

- All in all, paper documents that the introduction of face masks reduced the number of newly registered COVID-19 cases over the next 20 days by 75% relative to the synthetic control group, a sizeable effect
- The introduction of mandatory face masks and – before it – the associated signal to the local population to take the risk of person-to-person transmissions seriously apparently helped considerably in reducing the spread of COVID-19
- There is evidence of heterogeneous TEs when looking at different cities, yet the results remain sizeable and significant

- Suppose we observe $j = 1, ..., J+1$ aggregate units, such as states or countries, for $t = 1, ..., T$ periods.
- The first unit ($j = 1$) is exposed to a policy **intervention**, or some other **event** or **treatment** of interest, at time $t = T_0 + 1$, with $T_0 + 1 \leq T$. The remaining J units are not exposed to the intervention of interest.
- We aim to estimate the effect of the treatment on some outcome of interest during the post-treatment periods, $T_0 + 1, ..., T$.
  - Let Y_{jt}^N be the potential outcome observed for unit $j \in \{1, ..., J+1\}$ and time $t = \{1, ..., T\}$ in the absence of the intervention ($N = No$).
  - Let Y_{1t}^I be the potential outcome observed for the treated unit at time $t = T_0 + 1, ..., T$ under the intervention ($I = Intervention$).
- For each unit and time period, Y_{jt} is the observed outcome. Therefore, observed outcomes for untreated units, $j = 2, ..., J+1$ are equal to Y_{jt}^N.

- For the treated unit, the observed outcome is equal to Y_{1t}^N for $t = 1, ..., T_0$, and equal to Y_{1t}^I for $t = T_0 + 1, ..., T$
- The object of interest is the treatment effect on the treated unit,

$$\tau_{1t} = Y_{1t}^I - Y_{1t}^N$$

for $t = T_0 + 1, ..., T$. Since $Y_{1t}^I = Y_{1t}$ for post-treatment periods,

$$\tau_{1t} = Y_{1t} - Y_{1t}^N$$

for $t = T_0 + 1, ..., T$

  - That is, because Y_{1t}^I is observed in the post-treatment periods, estimating τ_{1t} for $t = T_0 + 1, ..., T$ boils down to estimating Y_{1t}^N

- **Challenge:** Estimate the **counterfactual outcome** $Y_{1t}^N, t > T_0$, i.e., how the outcome of interest would have evolved for the affected unit in the absence of the intervention

- A synthetic control estimator (SCE) of Y_{1t}^N is a weighted average of the outcomes for the "donor pool" of J untreated units,

$$\hat{Y}_{1t}^N = \sum_{j=2}^{J+1} W_j Y_{jt}$$

  where $W_2, ..., W_{J+1}$ are non-negative and sum to one

- A SCE of τ_{1t} is equal to the difference between the outcome values for the treated units and the outcomes values for the synthetic control

$$\hat{\tau}_{1t} = Y_{1t} - \sum_{j=2}^{J+1} W_j Y_{jt}$$

- The weights represent the contribution of each untreated observation to the estimate of the counterfactual of interest, $\hat{Y}_{1t}^N$

- **Some Comments**
  - SCM is based on the idea that, when the units of observation are a small number of aggregate entities, a combination of unaffected units often provides a more appropriate comparison than any single unaffected unit alone
  - The weights can be chosen via data-driven selectors
  - The solution to this type of constrained optimization problem problem is typically **sparse**
- **What is a sparse solution?**
  - A solution where only a few units in the donor pool obtain non-zero weigths
  - This poses analytical and numerical challenges yet helps with interpretability/transparency (details below)

- How to optimally choose the weights $\boldsymbol{W}^* = (W_2^*, ..., W_{J+1}^*)'$?

Notation

- Let $\boldsymbol{X_j} = (X_{1j}, ..., X_{kj})', j = 1, ..., J+1$ be a $(k \times 1)$-vector of pre-intervention values of predictors of Y_{jt}^N , with $t = T_0 + 1, ..., T$ (i.e., matching on observables! choice of X is key!)
- Let $\boldsymbol{X}_0$ be the $(k \times J)$-matrix that concatenates $\boldsymbol{X}_2, ..., \boldsymbol{X}_{J+1}$

General formulation

- Obtain the weights which minimize the distance

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

where $||.||$ is a norm (a measure of distance between $\boldsymbol{X}_1 = \textit{treated}$ and $\boldsymbol{X}_0 \boldsymbol{W}$ in this case, where $\boldsymbol{X}_0 = \textit{untreated}$)

Examples

- The Euclidean norm is given by $||x - y|| = \left(\sum_i [x_i - y_i]^2\right)^{1/2}$
- The L_1-norm is given by $||x - y|| = (\sum_i |x_i - y_i|)$
- If a single unit, m, in the donor pool is used as a comparison, then $w_m = 1, w_j = 0, j \neq m$ and

$$\hat{\tau}_{1t} = Y_{1t} - Y_{mt}$$

  - For NN estimators, m is the index value that minimizes $||\boldsymbol{X}_1 - \boldsymbol{X}_j||$ over j for some norm $||.||$

Least-squares Approach (Abadie & Gardeazabal 2003, Abadie et al 2010)

- The weights $\boldsymbol{W}$ can be obtained by minimizing the distance (conditioned on v_h's)

$$||\boldsymbol{X}_1 - \boldsymbol{X}_0\boldsymbol{W}|| = \left(\sum_{h=1}^{k} v_h \left(X_{h1} - W_2 X_{h2} - ... - W_{J+1} X_{hJ+1}\right)^2\right)^{1/2}$$

  s.t. weights being non-negative and sum to one

- The v_h's are non-negative constants
  - e.g., inverse of variance of the h-th variable (in this case, used to scale variables of different magnitudes all to have unit variance)

- **Intuition:** (constrained) Weighted Least Squares problem on characteristics (X, instead of outcomes, Y) of individual units

Out-of-sample Validation (Abadie et al 2015)

- The aim is to choose $\boldsymbol{V} = (v_1, ..., v_k)$ such that $\boldsymbol{W}(\boldsymbol{V}) = (w_2(\boldsymbol{V}), ..., w_{J+1}(\boldsymbol{V}))'$ by minimizing the MSPE (Mean Squared Predicted Error) for a subset of pre-intervention periods $t = 1, 2, ..., T_0$

$$\sum_{t=1}^{T_0} (Y_{1t} - w_2(V) Y_{2t} - ... - w_{J+1}(V) Y_{J+1t})^2$$

- **Note**
  - Abadie (2021, p. 397) describes it briefly
  - Such types of algorithms will be covered in detail later on

OOS Validation: Heuristic Algorithm (simplified)

1. Divide the pre-intervention periods into an initial **training** period and a subsequent **validation** period

(Assume T_0 even and the training and validation periods are $t = 1, ..., t_0$ and $t = t_0 + 1, ..., T_0$, $t_0 = T_0/2$)

2. For every value $\boldsymbol{V}$,let $\{\tilde{w}_k(\boldsymbol{V})\}_{j=2}^{J+1}$ be obtained computing the **training sample** (by LS) and compute the corresponding MSPE in the **validation sample**

$$\sum_{t=t_0}^{T_0} (Y_{1t} - \tilde{w}_2(\boldsymbol{V}) Y_{2t} - ... - \tilde{w}_{J+1}(\boldsymbol{V}) Y_{J+1t})^2$$

3. Do Step 2 for a number of candidates $\boldsymbol{V} \in \mathcal{V}$. Call the best of them $\boldsymbol{V}^*$ and $\boldsymbol{W}^* = \boldsymbol{W}(\boldsymbol{V}^*)$

e.g. set $\boldsymbol{V}$ equal to the inverse of the variance, of the std. deviation, the IQ range etc (This set is called $\mathcal{V}$ above)

4. Use the best weights $\boldsymbol{W}^* = \boldsymbol{W}(\boldsymbol{V}^*)$ in the post-intervention period

- Here I briefly summarize the results of detailed studies reported in Abadie (2021) and Abadie and Vives-i-Bastida (2021)
- The starting point is a linear factor model assumed to generate Y_{jt}^N

$$Y_{jt}^N = \delta_t + \boldsymbol{\theta}_t \boldsymbol{Z}_j + \lambda_t \boldsymbol{\mu}_j + \varepsilon_{it}$$

  where δ_t is a time trend, $\boldsymbol{Z}_j$ and $\boldsymbol{\mu}_j$ are vectors of **observed** and **unobserved** predictors of Y_{jt}^N, respectively, with coefficients $\boldsymbol{\theta}_t$ and $\boldsymbol{\lambda}_t$, and ε_{jt} is zero mean individual transitory shock
- In the time-series literature in econometrics, $\boldsymbol{\theta}_t$ and $\boldsymbol{\lambda}_t$ are referred to as common factors, and $\boldsymbol{Z}_j$ and $\boldsymbol{\mu}_j$ as factor loadings. The term δ_t is a common factor with constant loadings across units, while $\boldsymbol{\lambda}_t$ represents a set of common factors with varying loadings across units
- A difference-in-differences/fixed effects panel model can be obtained by setting $\boldsymbol{\lambda}_t = \boldsymbol{\lambda}$ (time invariant)

Summary of findings

- The SCE performs better in the post-intervention period...
  - The lower the $Var(\varepsilon_{jt})$
  - The larger the pre-intervention period T_0
  - The lower the number of unobserved factors
  - The lower the number of donors in the donor pool
    - (the sparser the solution, the lower the risk of overfitting)
- See both papers and references therein for details
- See also Sections 5-6 of Abadie (2021) for empirical guidance