Synthetic Controls for Policy Evaluation

Alberto Abadie
John F. Kennedy School of Government
Harvard University

IDB
Course on Synthetic Controls
December 3-4, 2015

Comparative Case Studies

Goal:

- Estimate effects of events or policy interventions that take place at an aggregate level (e.g., cities, states, countries).

Comparative Case Studies:

- Compare the evolution of an aggregate outcome for the unit affected by the intervention (the "treated" unit) to the evolution of the same aggregate for some control group (e.g. Card, 1990, Card and Krueger, 1994, Abadie and Gardeazabal, 2003).

Revisiting Mariel

Difference-in-differences estimate on unemployment rates
(African-American workers)

	Year		
	1979	1981	1981−1979
Miami	8.3 (1.7)	9.6 (1.8)	1.3 (2.5)
Comparison cities	10.3 (0.8)	12.6 (0.9)	2.3 (1.2)
Miami-Comparison Difference	−2.0 (1.9)	−3.0 (2.0)	−1.00 (2.8)

Adapted from Card (1990) and Angrist and Krueger (1999).
Standard errors in parentheses.

Comparative Case Studies

Advantages:

- Policy interventions often take place at an aggregate level
- Aggregate/macro data are often available

Problems:

- Selection of control group is often ambiguous
- Standard errors do not reflect uncertainty about the ability of the control group to reproduce the counterfactual of interest

The Synthetic Control Method

- A distinctive feature of comparative case studies is that the units of analysis are usually aggregate entities, like countries or regions, for which suitable single comparisons often do not exist.
- The synthetic control method is based on the observation that, when the units of analysis are a few aggregate entities, a combination of comparison units (a "synthetic control") often does a better job reproducing the characteristics of a treated unit than any single comparison unit alone.
- Motivated by this consideration, the comparison unit in the synthetic control method is selected as the weighted average of all potential comparison units that best resembles the characteristics of the treated unit(s).

The Synthetic Control Method: Advantages

- Precludes extrapolation.
- Does not require access to post-treatment outcomes in the "design" phase of the study, when synthetic controls are calculated.
- Makes explicit the contribution of each comparison unit to the counterfactual of interest.
- Allows researchers to use quantitative and qualitative techniques to analyze the similarities and differences between the units representing the case of interest and the synthetic control.
- Formalizing the way comparison units are chosen not only represents a way of systematizing comparative case studies, it also has direct implications for inference.

Synthetic Control Method: Setting

- Suppose that we observe $J+1$ units in periods $1, 2, \ldots, T$.
- Unit "one" is exposed to the intervention of interest (that is, "treated") during periods $T_0+1, \ldots, T$.
- The remaining J are an untreated reservoir of potential controls (a "donor pool").
- Let Y_{it}^N be the outcome that would be observed for unit i at time t in the absence of the intervention.
- Let Y_{it}^I be the outcome that would be observed for unit i at time t if unit i is exposed to the intervention in periods T_0+1 to T.
- We aim to estimate the effect of the intervention on the treated unit $(\alpha_{1T_0+1}, \ldots, \alpha_{1T})$, where

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

for $t > T_0$, and Y_{1t} is the outcome for unit one at time t.

Synthetic Control Method: Implementation

- Let $W = (w_2, \ldots, w_{J+1})'$ with $w_j \geq 0$ for $j = 2, \ldots, J+1$ and $w_2 + \cdots + w_{J+1} = 1$. Each value of W represents a potential synthetic control
- Let X_1 be a $(k \times 1)$ vector of pre-intervention characteristics for the treated unit. Similarly, let X_0 be a $(k \times J)$ matrix which contains the same variables for the unaffected units.
- The vector $W^* = (w_2^*, \ldots, w_{J+1}^*)'$ is chosen to minimize $\|X_1 - X_0 W\|$, subject to our weight constraints.
- Let Y_{jt} be the value of the outcome for unit j at time t. For a post-intervention period t (with $t \geq T_0$) the synthetic control estimator is:

$$\widehat{\alpha}_{1t} = Y_{1t} - \sum_{j=2}^{J+1} w_j^* Y_{jt}.$$

Synthetic Control Method: Estimation

- That is, to reproduce the counterfactual Y_{1t}^N for $t > T_0$ we find the combination of untreated units that best resembles the treated unit before the intervention in terms of the values of k relevant covariates (predictors of the outcome of interest).

- Example: 1988 California's tobacco control program:
  - Treated unit: California
  - Outcome of interest: tobacco consumption in California after 1988
  - Potential controls: other US states
  - Covariates: predictors of state-level tobacco consumption measured before 1988

Synthetic Control Method: Estimation

- We consider $\|X_1 - X_0 W\| = \sqrt{(X_1 - X_0 W)' V (X_1 - X_0 W)}$, where V is some $(k \times k)$ symmetric and positive semidefinite matrix.

- Typically, V is diagonal with main diagonal $v_1, \ldots, v_k$.

- Let X_{jm} be the value of the m-th covariates for unit j. Then, the synthetic control weights $w_2^*, \ldots, w_{J+1}^*$ minimize:

$$\sum_{m=1}^{k} v_m \left(X_{1m} - \sum_{j=2}^{J+1} w_j X_{jm} \right)^2,$$

  where v_m is a weight that reflects the relative importance that we assign to the m-th variable when we measure the discrepancy between the treated unit and the synthetic control.

Synthetic Control Method: Estimation

- To reduce interpolation biases:
  - Restrict the donor pool to units with characteristics X_j similar to X_1.
  - Include penalty terms for discrepancies in the covariates in the objective function:
$$\|X_1 - X_0 W\| + \sum_{j=2}^{J} \phi(W_j, \|X_j - X_1\|),$$
    for some positive and non-decreasing function $\phi(\cdot, \cdot)$.
- The choice of V is important:
  - W^* depends on the choice of V.
  - The synthetic control $W^*(V)$ is meant to reproduce the behavior of the outcome variable for the treated unit in the absence of the treatment.
  - Therefore, the weights, $v_1, \ldots, v_k$ should reflect the predictive value of the covariates.

Synthetic Control Method: Estimation

Choice of $v_1, \ldots, v_k$ can be based on:

- Subjective assessment of the predictive power of each of the covariates, or calibration inspecting how different values for $v_1, \ldots, v_k$ affect the discrepancies between the treated unit and the synthetic control.
- Use regression to assess the predictive power of the covariates.
- Minimize mean square prediction error (MSPE):
$$\sum_{t=1}^{T_0} \left(Y_{1t} - \sum_{j=2}^{J} w_j^*(V) Y_{jt} \right)^2 .$$
- Cross-validation:
  - Divide the pre-treatment period into an initial **training** period and a subsequent **validation** period.
  - For any given V, calculate $W^*(V)$ in the training period.
  - Minimize the MSPE of $W^*(V)$ in the validation period.

What About Unobserved Factors?

- Comparative case studies are complicated by unmeasured factors affecting the outcome variables as well as heterogeneity in the effect of observed and unobserved factors.
- However, as we will see, if the number of pre-intervention periods in the data is large, matching on pre-intervention outcomes allows to control for heterogeneous responses to multiple unobserved factors.
- Intuition: Only units that are alike in both observed and unobserved determinants of the outcome variable as well as in the effect of those determinants on the outcome variable should produce similar trajectories of the outcome variable over extended periods of time.

Synthetic Control Method: Formalization

- Suppose that Y_{it}^N is given by a factor model:

$$Y_{it}^N = \delta_t + Z_i\theta_t + \lambda_t\mu_i + \varepsilon_{it},$$

  - δ_t is an unobserved (common) time-dependent factor,
  - Z_i is a $(1 \times r)$ vector of observed covariates,
  - θ_t is a $(r \times 1)$ vector of unknown parameters,
  - λ_t is a $(1 \times F)$ vector of unknown common factors,
  - μ_i is a $(F \times 1)$ vector of unknown factor loadings,
  - ε_{it} are unobserved transitory shocks.
- Specification allows heterogeneous responses to multiple unobserved factors.
- In contrast, the Difference-in-Differences (or Fixed-Effects) model restricts λ_t to be constant.

Synthetic Controls: Time-Varying Variables

- Recall the factor model:

$$Y_{it}^N = \delta_t + Z_i\theta_t + \lambda_t\mu_i + \varepsilon_{it},$$

- The vector Z_i may contain pre- and post-intervention values of time-varying variables, as long as they are not affected by the intervention.
- For example, if $T = 2$, $T_0 = 1$, $Z_i = (Z_{i1}, Z_{i2})$,

$$\theta_1 = \begin{pmatrix} \beta \\ 0 \end{pmatrix} \quad \text{and} \quad \theta_2 = \begin{pmatrix} 0 \\ \beta \end{pmatrix},$$

then $Z_i\theta_t = Z_{it}\beta$.

Synthetic Control Method: Formalization

- Let $W = (w_2, \ldots, w_{J+1})'$ with $w_j \geq 0$ for $j = 2, \ldots, J+1$ and $w_2 + \cdots + w_{J+1} = 1$. Each value of W represents a potential synthetic control.
- Suppose that we can choose W^* such that:

$$\sum_{j=2}^{J+1} w_j^* Z_j = Z_1, \ \sum_{j=2}^{J+1} w_j^* Y_{j1} = Y_{11}, \ \cdots \ , \ \sum_{j=2}^{J+1} w_j^* Y_{jT_0} = Y_{1T_0}.$$

- Then (if T_0 is large relative to the scale of ε_{it}), an approximately unbiased estimator of α_{1t} is:

$$\widehat{\alpha}_{1t} = Y_{1t} - \sum_{j=2}^{J+1} w_j^* Y_{jt}$$

for $t \in \{T_0 + 1, \ldots, T\}$.

Synthetic Control Method: Formalization

Suppose that $E|\varepsilon_{jt}|^p < \infty$ for some $p > 2$. Then,

$$|E[\widehat{\alpha}_{1t} - \alpha_{1t}]| < C(p)^{1/p} \left(\frac{\bar{\lambda}^2 F}{\underline{\xi}} \right) J^{1/p} \max \left\{ \frac{\bar{m}_p^{1/p}}{T_0^{1-1/p}}, \frac{\bar{\sigma}}{T_0^{1/2}} \right\},$$

where

$$\sigma_{jt}^2 = E|\varepsilon_{jt}|^2, \quad \sigma_j^2 = \frac{1}{T_0} \sum_{t=1}^{T_0} \sigma_{jt}^2, \quad \bar{\sigma}^2 = \max_{j=2,\ldots,J+1} \sigma_j^2,$$

$$m_{pjt} = E|\varepsilon_{jt}|^p, \quad m_{pj} = \frac{1}{T_0} \sum_{t=1}^{T_0} m_{pjt}, \quad \bar{m}_p = \max_{j=2,\ldots,J+1} m_{pj},$$

for p even, $|\lambda_{tf}| \leq \bar{\lambda}$ for all $t = 1, \ldots, T$ and $f = 1, \ldots, F$, and

$$\underline{\xi} \leq \xi(M) = \text{smallest eigenvalue of } \frac{1}{M} \sum_{t=T_0-M+1}^{T_0} \lambda_t' \lambda_t.$$

$\Rightarrow$ Bias is small when the number of pretreatment periods, T_0, is large relative to the scale of the transitory shocks, ε_{jt}.

Synthetic Control Method: Formalization

- Synthetic controls can provide unbiased estimates in contexts other than the factor model considered so far.
- Consider, the following autoregressive model with time-varying coefficients:

$$\begin{aligned} Y_{it+1}^N &= \alpha_t \, Y_{it}^N + \beta_{t+1} \, Z_{it+1} + u_{it+1}, \\ Z_{it+1} &= \gamma_t \, Y_{it}^N + \Pi_t \, Z_{it} + v_{it+1}, \end{aligned}$$

  where u_{it+1} and v_{it+1} have mean zero given $\{Y_{js}, Z_{js}\}_{1 \leq j \leq N, s \leq t}$.
- Suppose that we can choose $\{w_j^*\}_{2 \leq j \leq N}$ such that:

$$\sum_{j=2}^{J+1} w_j^* \, Y_{jT_0} = Y_{1T_0}, \quad \text{and} \quad \sum_{j=2}^{J+1} w_j^* \, Z_{jT_0} = Z_{1T_0}.$$

- Then, the synthetic control estimator is unbiased.

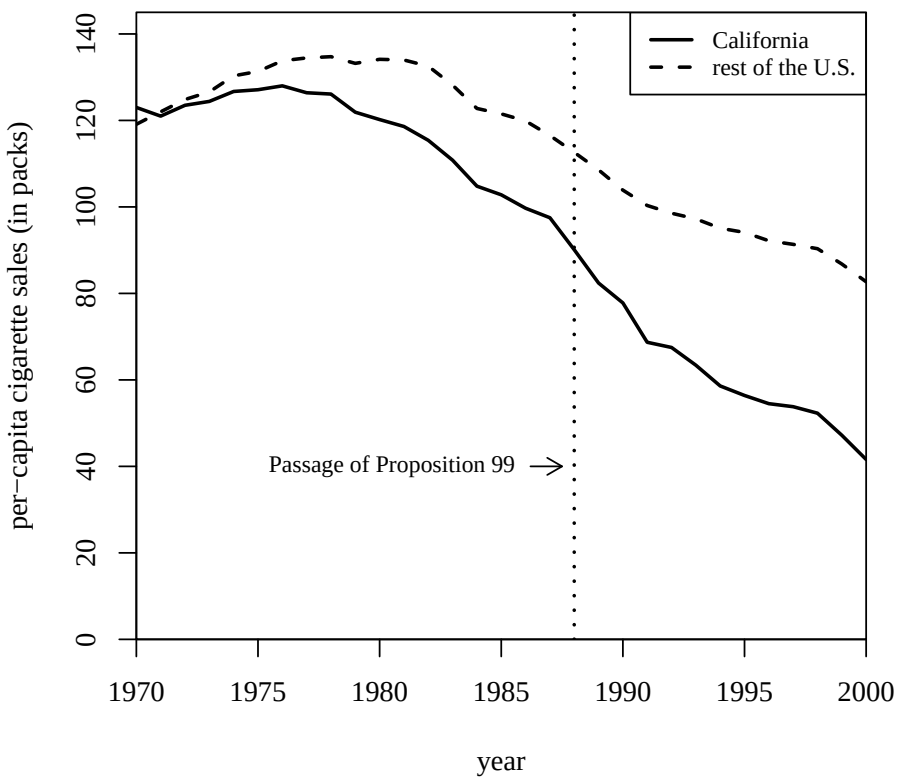
Application: California's Proposition 99

In 1988, California first passed comprehensive tobacco control legislation:

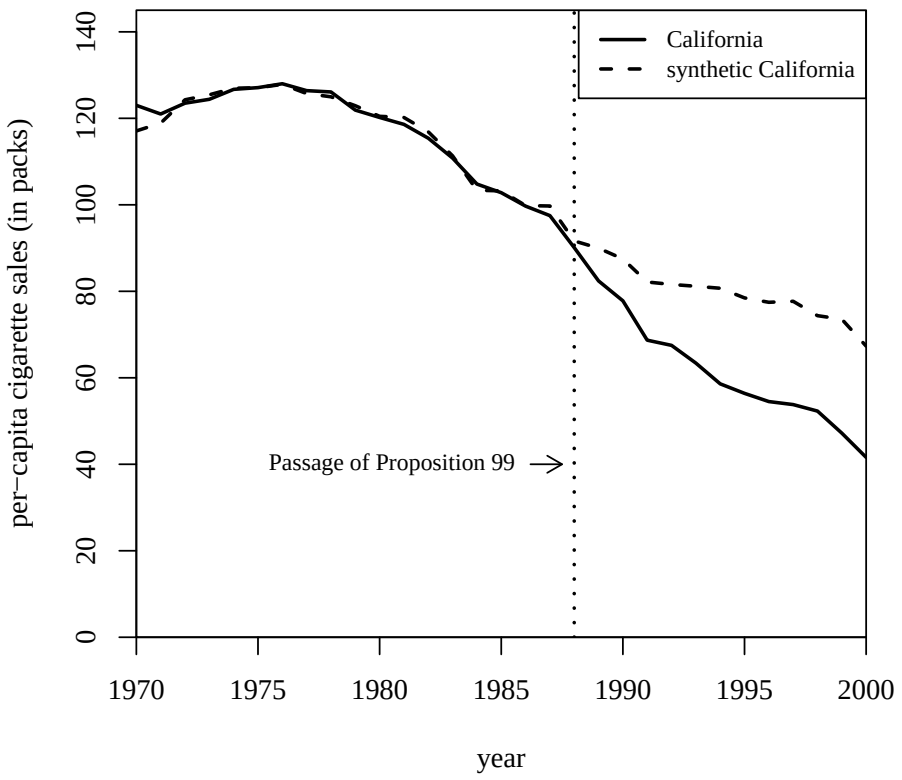
- increased cigarette tax by 25 cents/pack
- earmarked tax revenues to health and anti-smoking budgets
- funded anti-smoking media campaigns
- spurred clean-air ordinances throughout the state
- produced more than $100 million per year in anti-tobacco projects

Other states that subsequently passed control programs are excluded from donor pool of controls (AK, AZ, FL, HA, MA, MD, MI, NJ, NY, OR, WA, DC).

Cigarette Consumption: CA and the Rest of the U.S.



Cigarette Consumption: CA and synthetic CA

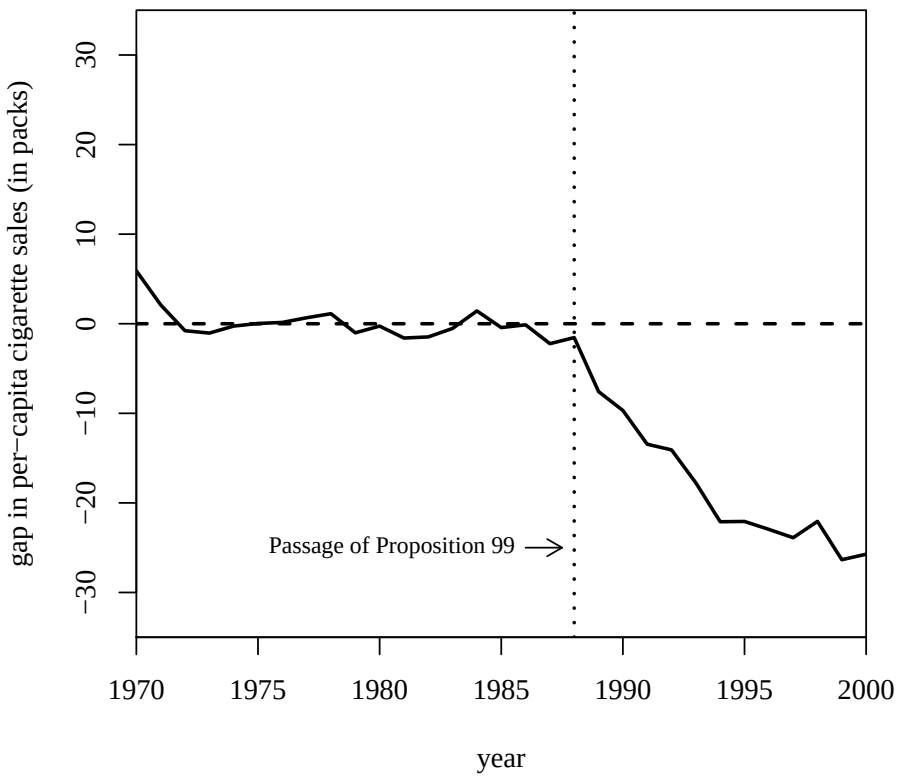


Predictor Means: Actual vs. Synthetic California

Variables	California Real	California Synthetic	Average of 38 control states
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).

Smoking Gap Between CA and synthetic CA



Inference

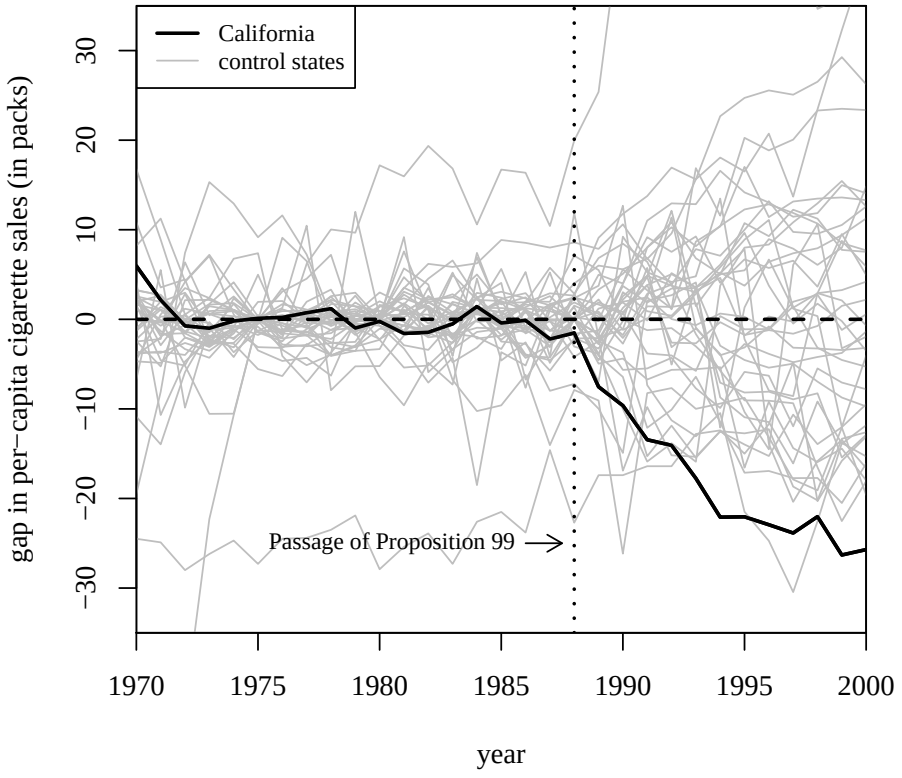
- To evaluate the significance of our results, we conduct a series of placebo experiments where we reassign the tobacco control program to states other than California.
- We proceed as follows:
  - Iteratively apply the synthetic method to each state in the donor pool and obtain a distribution of placebo effects.
  - Compare the gap (RMSPE) for California to the distribution of the placebo gaps. For example, the post-Prop. 99 RMSPE for California is:

$$\text{RMSPE} = \left(\frac{1}{T - T_0} \sum_{t=T_0+1}^{T} \left(Y_{1t} - \sum_{j=2}^{J+1} w_j^* Y_{jt} \right)^2 \right)^{1/2}.$$

  - Question is whether the effect estimated by the synthetic control for California is large relative to the effect estimated for a state chosen at random.

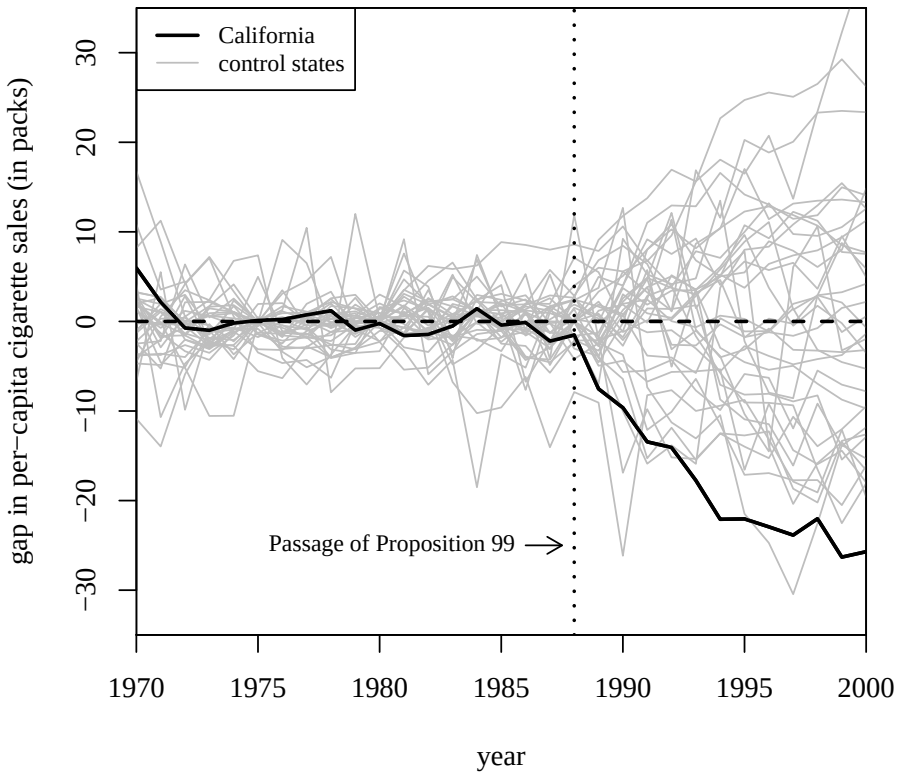
Smoking Gap for CA and 38 control states

(All States in Donor Pool)



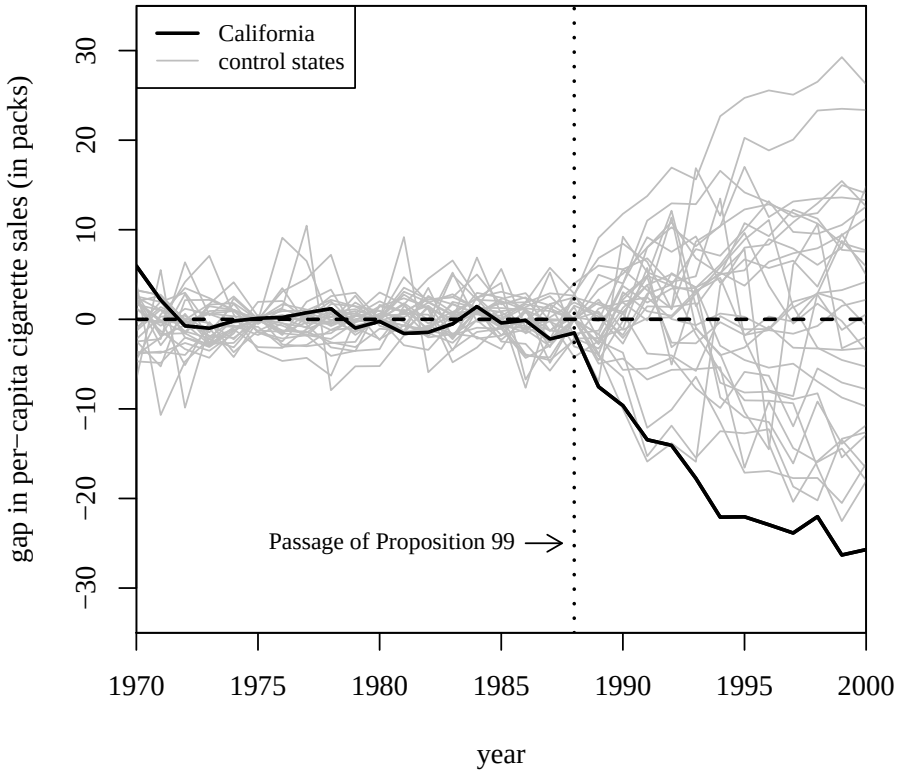
Smoking Gap for CA and 34 control states

(Pre-Prop. 99 MSPE $\leq$ 20 Times Pre-Prop. 99 MSPE for CA)



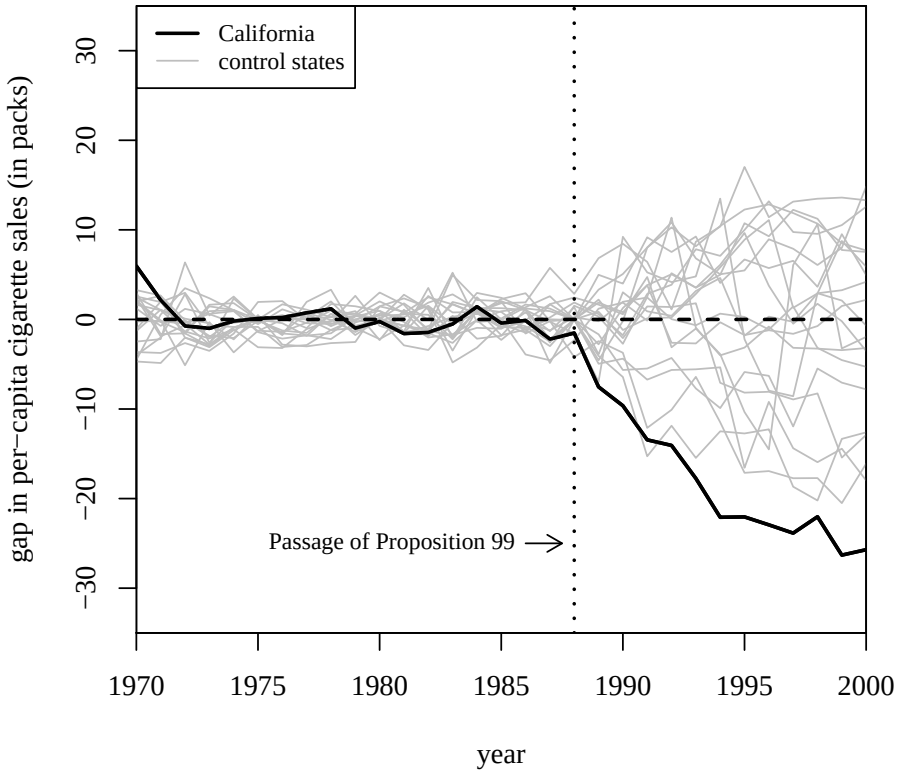
Smoking Gap for CA and 29 control states

(Pre-Prop. 99 MSPE $\leq$ 5 Times Pre-Prop. 99 MSPE for CA)



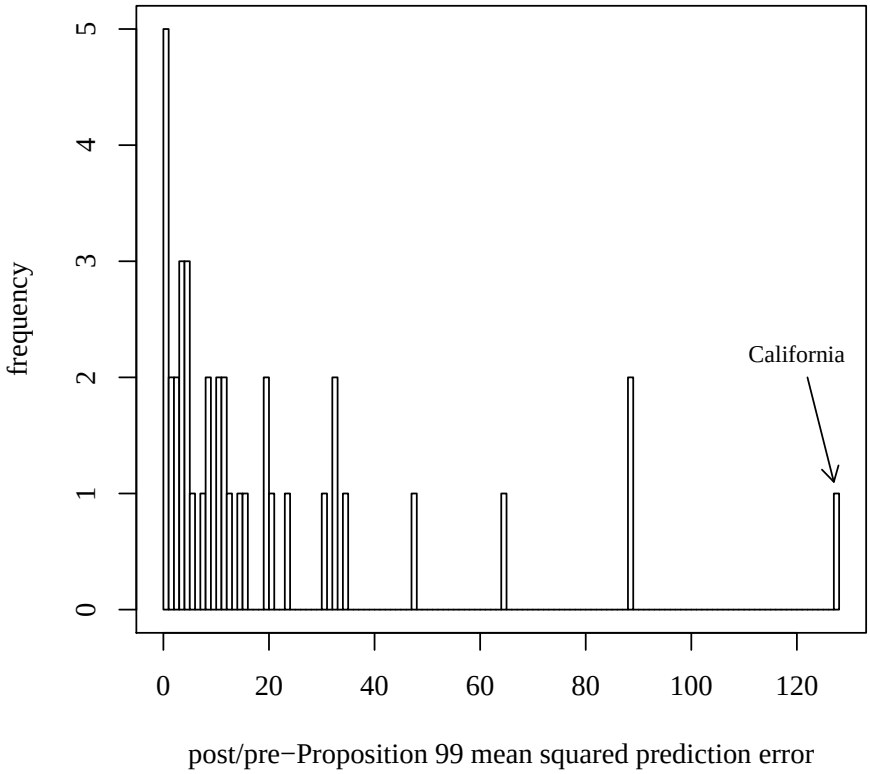
Smoking Gap for CA and 19 control states

(Pre-Prop. 99 MSPE $\leq$ 2 Times Pre-Prop. 99 MSPE for CA)



Ratio Post-Prop. 99 MSPE to Pre-Prop. 99 MSPE

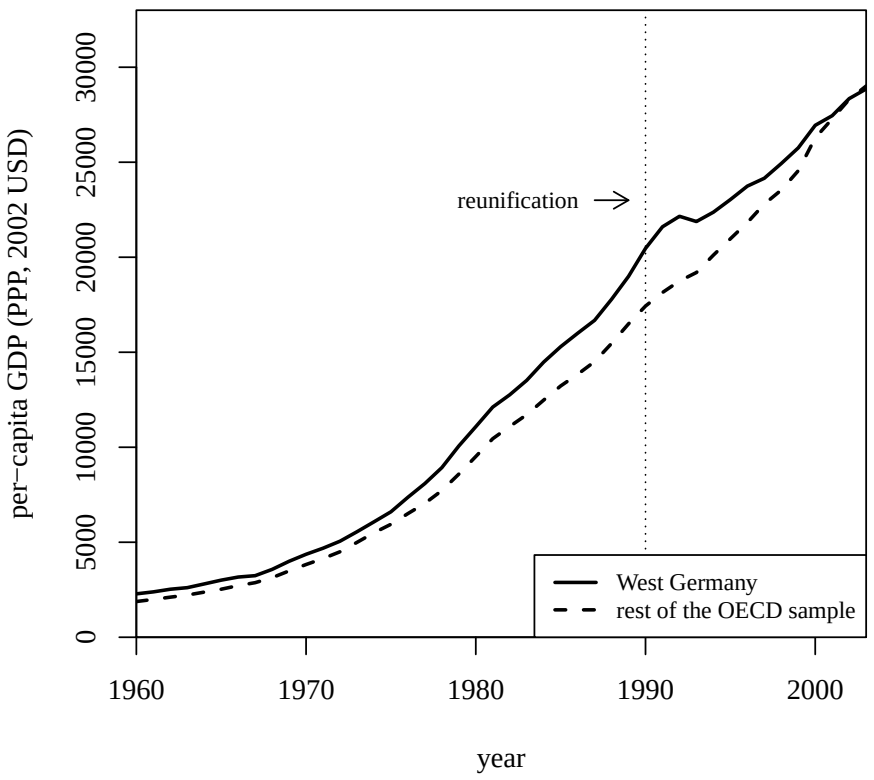
(All 38 States in Donor Pool)



Application: The 1990 German Reunification

- Cross-country regressions are often criticized because they put side-by-side countries of very different characteristics.
- "What do Thailand, the Dominican Republic, Zimbabwe, Greece, and Bolivia have in common that merits their being put in the same regression analysis? Answer: For most purposes, nothing at all." (Harberger, 1987).
- The synthetic control method provides a data-driven procedure to select a comparison unit.
- Application: the economic impact of the 1990 German reunification in West Germany.
- Donor pool is restricted to 16 OECD countries.

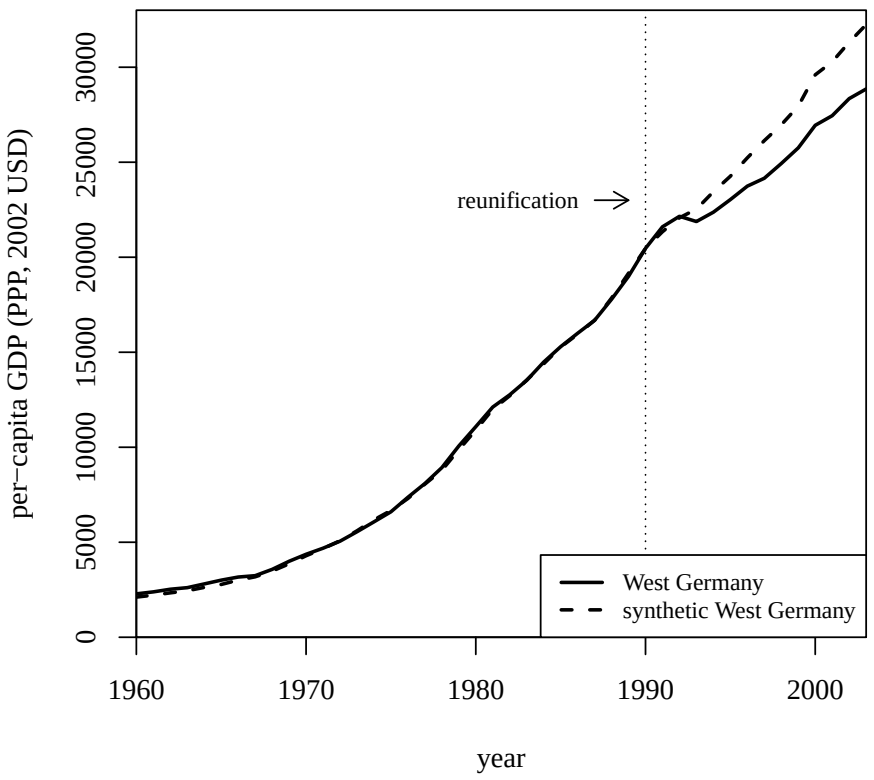
West Germany and the OECD sample



Covariate Averages before 1990

	West Germany	Synthetic West Germany	OECD Sample
GDP per-capita	15808.9	15800.9	8021.1
Trade openness	56.8	56.9	31.9
Inflation rate	2.6	3.5	7.4
Industry share	34.5	34.4	34.2
Schooling	55.5	55.2	44.1
Investment rate	27.0	27.0	25.9

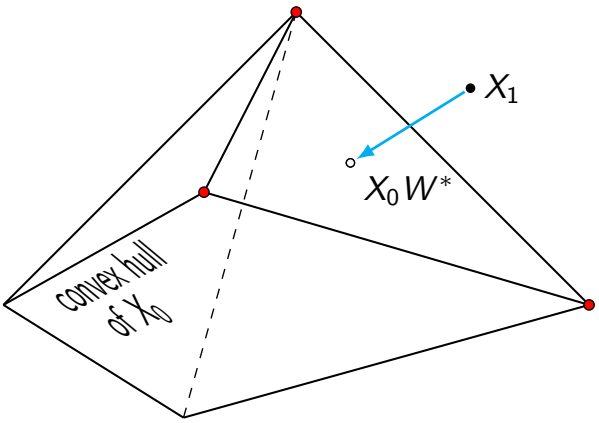
West Germany and synthetic West Germany



Country Weights in the Synthetic West Germany

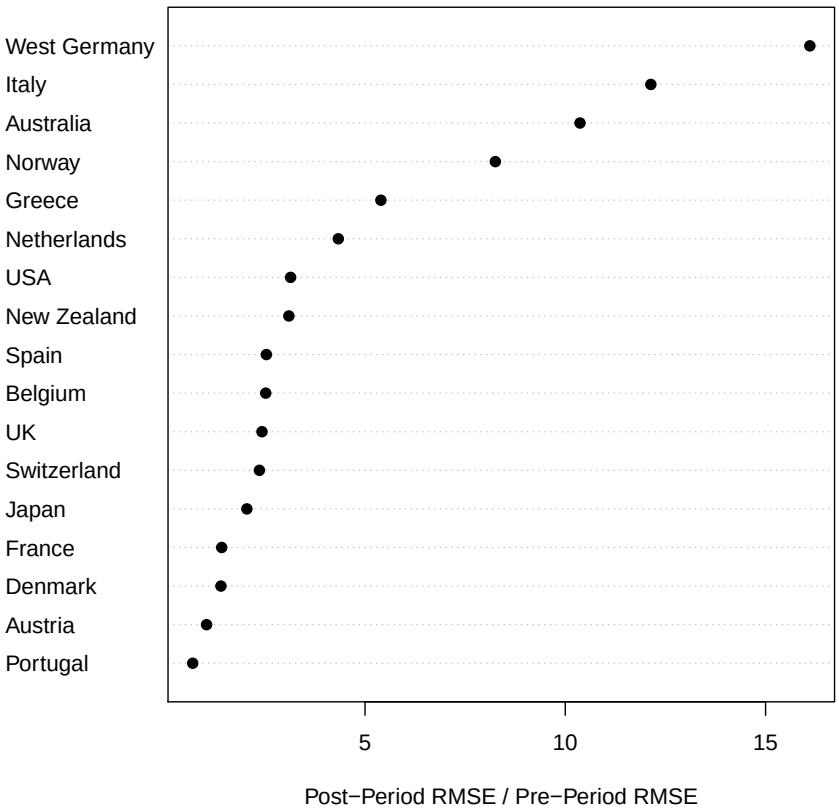
Country	Weight	Country	Weight
Australia	0	Netherlands	0.10
Austria	0.42	New Zealand	0
Belgium	0	Norway	0
Denmark	0	Portugal	0
France	0	Spain	0
Greece	0	Switzerland	0.11
Italy	0	United Kingdom	0
Japan	0.16	United States	0.22

Synthetic Control Method: Sparsity

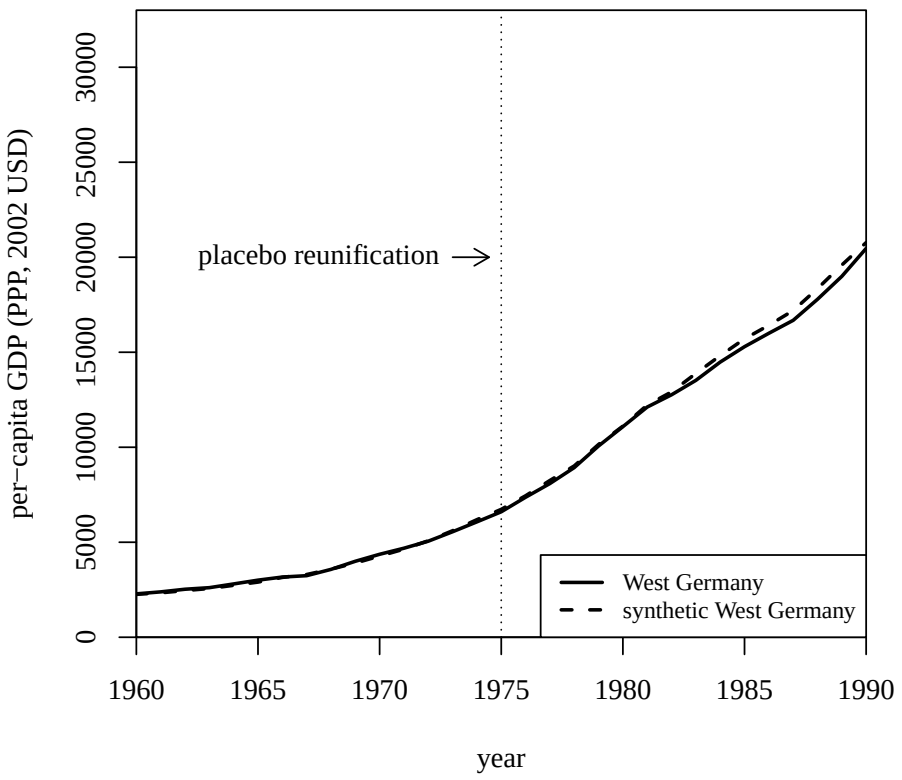


Synthetic controls are typically sparse because they are obtained by projecting X_1 on the convex hull of X_0.

Post-Reunification to Pre-Reunification RMSPE Ratio



Placebo Reunification: 1975



Comparison to Regression

- Constructing a synthetic comparison as a linear combination of the untreated units with coefficients that sum to one may appear unusual.
- However, we will see that a regression-based approach does exactly this, albeit in an implicit way.
- In contrast to the synthetic control method, the regression approach does not restrict the coefficients of the linear combination that define the comparison unit to be in between zero and one, therefore allowing extrapolation outside the support of the data.

Comparison to Regression

- Let T_1 be the number of post-intervention periods and:
  X_1: $(k \times 1)$-matrix of covariates for the treated unit
  X_0: $(k \times J)$-matrix of covariates for control units
  Y_0: $(T_1 \times J)$-matrix of post-intervention outcomes for control units

- Let
$$\widehat{B} = (X_0 X_0')^{-1} X_0 Y_0'$$
be the $(k \times T_1)$ matrix of regression coefficients of Y_0 on X_0.

- That is, each column of $\widehat{B}$ contains the regression coefficients of Y_0 on X_0 for a post-intervention period.

- A regression-based counterfactual of the outcome for the treated unit in absence of the treatment is given by the $(T_1 \times 1)$ vector $\widehat{B}' X_1$.

Comparison to Regression

- Notice that
$$\widehat{B}' X_1 = Y_0 W^{\text{reg}},$$
where
$$W^{\text{reg}} = X_0'(X_0 X_0')^{-1} X_1.$$

- As a result the regression-based estimate of the counterfactual of interest is a linear combination of post-treatment outcomes for the untreated units, with weights W^{reg}.

- Let ι be a $(J \times 1)$ vector of ones. The sum of the regression weights is $\iota' W^{\text{reg}}$. It can be proven that
$$\iota' W^{\text{reg}} = 1.$$

Regression Weights Sum to One

- To prove this claim, assume that, as usual, the regression includes an intercept, so the first row of X_0 is a vector of ones.
- Notice that $(X_0 X_0')^{-1} X_0 \, \iota$ is the vector of coefficients of the regression of ι on X_0.
- Therefore $(X_0 X_0')^{-1} X_0 \, \iota$ is the $(k \times 1)$ vector $(1, 0, \cdots, 0)'$.
- This implies that

$$\begin{aligned} \iota' W^{\text{reg}} &= \iota' X_0' (X_0 X_0')^{-1} X_1 \\ &= (1, 0, \cdots, 0) \, X_1 \end{aligned}$$

- Then, because the first element of X_1 is equal to one, we obtain $\iota' W^{\text{reg}} = 1$.

Synthetic *vs.* Regression Weights

Country	Synthetic Control Weight	Regression Weight	Country	Synthetic Control Weight	Regression Weight
Australia	0	0.12	Netherlands	0.10	0.14
Austria	0.42	0.26	New Zealand	0	0.12
Belgium	0	0	Norway	0	0.04
Denmark	0	0.08	Portugal	0	-0.08
France	0	0.04	Spain	0	-0.01
Greece	0	-0.09	Switzerland	0.11	0.05
Italy	0	-0.05	UK	0	0.06
Japan	0.16	0.19	USA	0.22	0.13

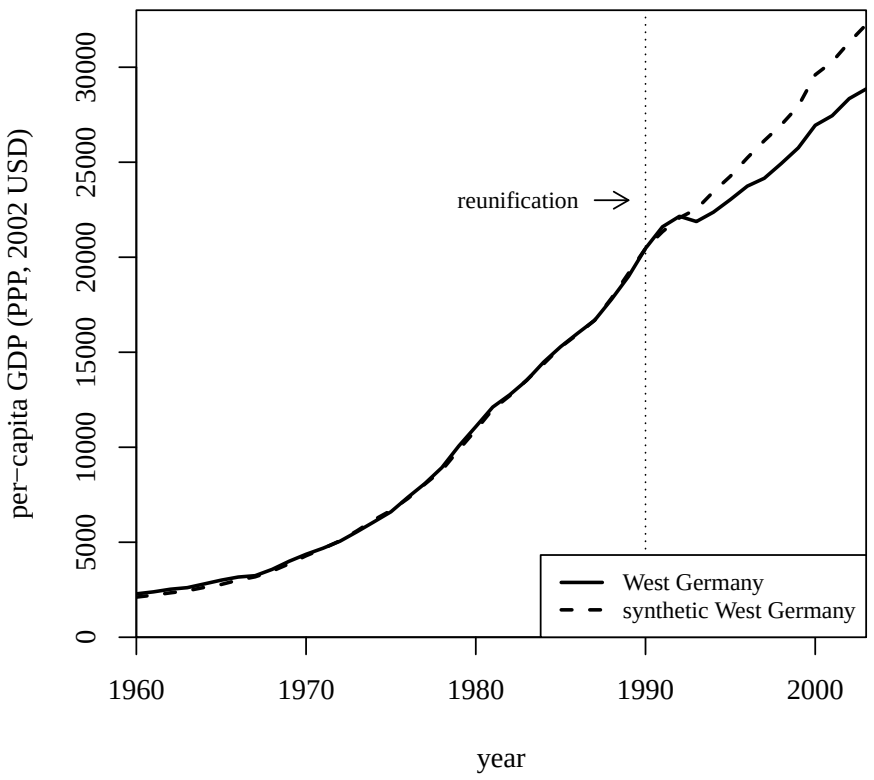
Sparse Synthetic Controls

	West Germany	Synthetic West Germany *Number of countries in synthetic unit:*					OECD Sample
		5	4	3	2	1	
GDP per-capita	15808.9	15800.9	15800.5	15486.4	15576.1	14817.0	8021.1
Trade openness	56.8	56.9	55.9	52.5	61.5	74.6	31.9
Inflation rate	2.6	3.5	3.6	3.6	3.8	3.5	7.4
Industry share	34.5	34.4	34.6	34.8	34.3	35.5	34.2
Schooling	55.5	55.2	57.6	57.7	60.7	60.9	44.1
Investment rate	27.0	27.0	27.2	26.8	25.6	26.6	25.9

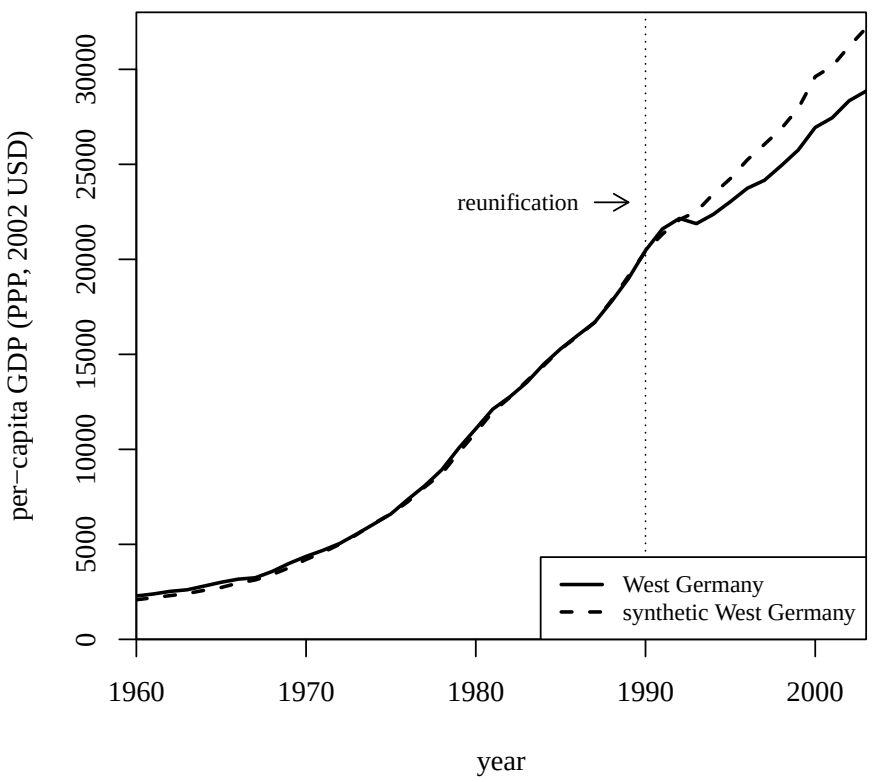
Sparse Synthetic Controls

Synthetic Combination:	Country Weights				
Five Control Countries	Austria 0.42	USA 0.22	Japan 0.16	Switzerland 0.11	Netherlands 0.10
Four Control Countries	Austria 0.56	USA 0.22	Japan 0.13	Switzerland 0.10	
Three Control Countries	Austria 0.59	USA 0.26	Japan 0.15		
Two Control Countries	Austria 0.76	USA 0.24			
One Control Country	Austria 1				

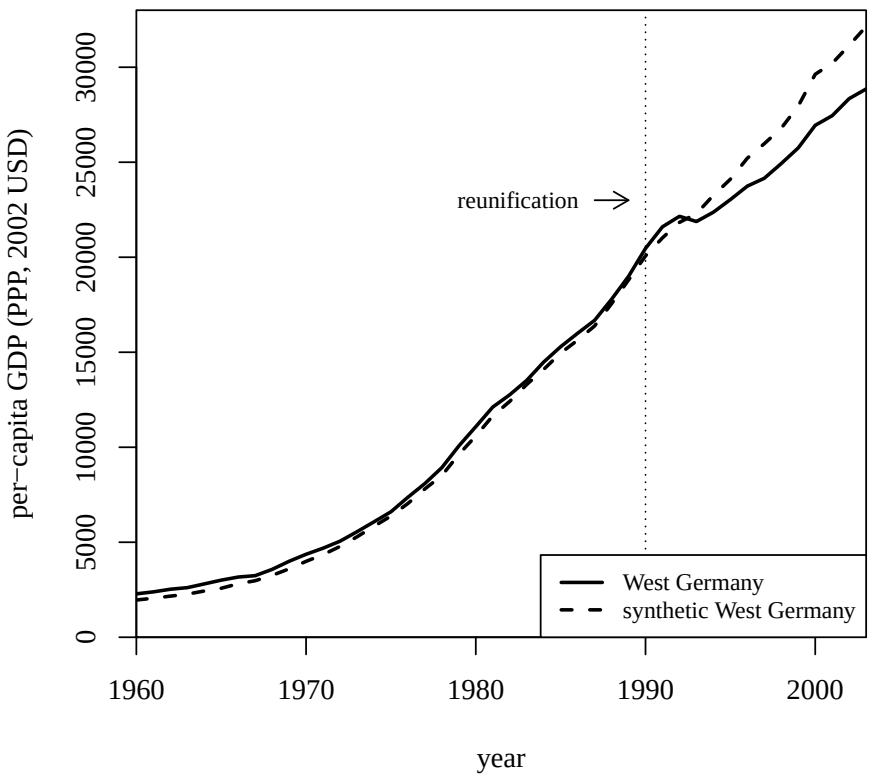
Synthetic Control with 5 Countries



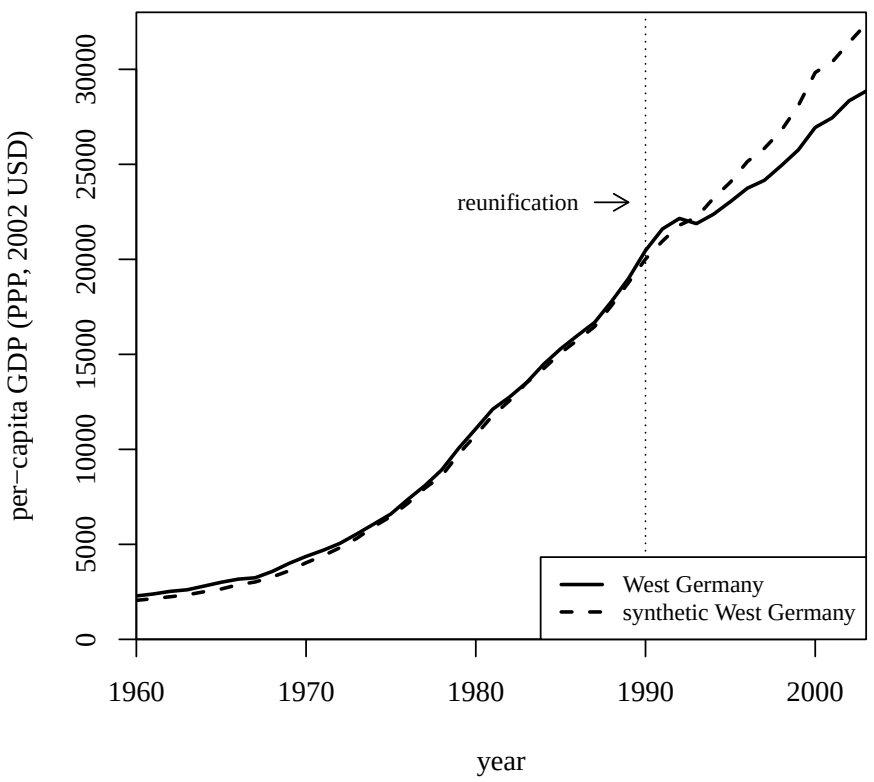
Synthetic Control with 4 Countries



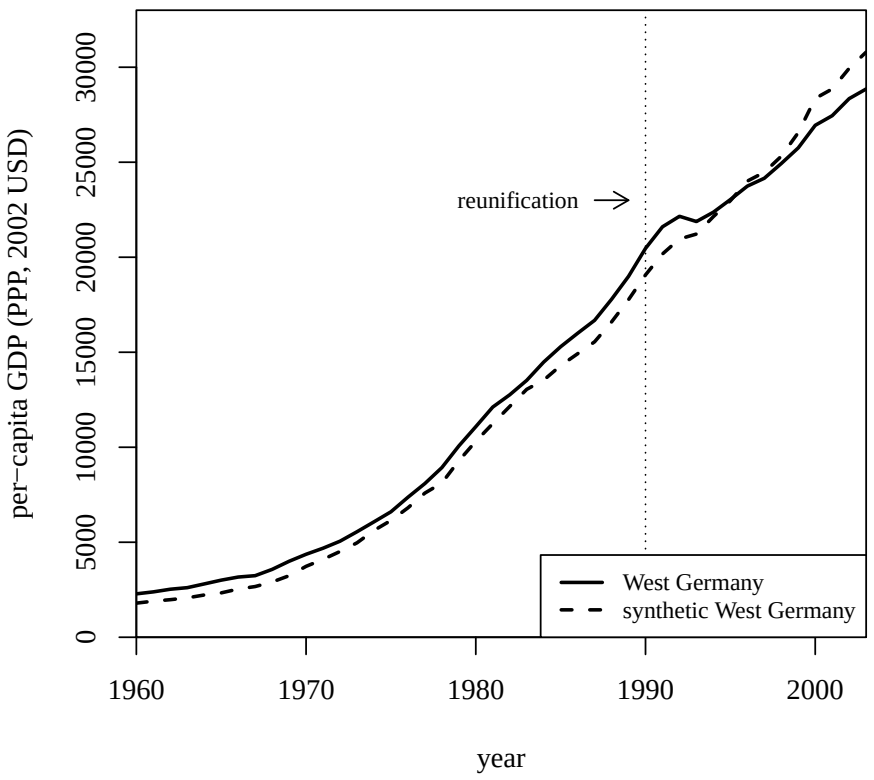
Synthetic Control with 3 Countries



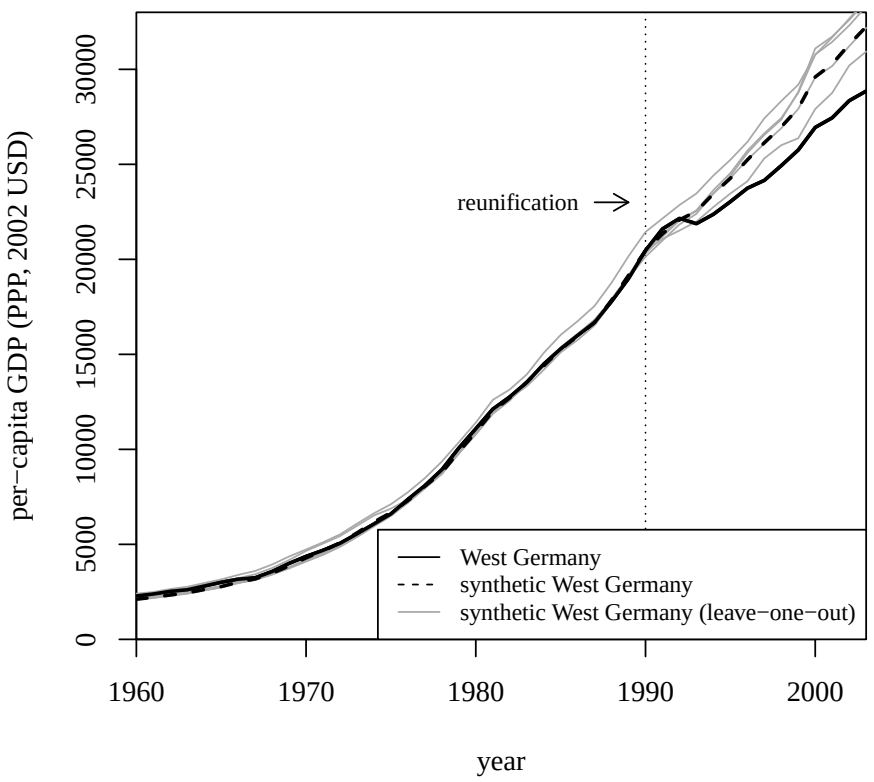
Synthetic Control with 2 Countries



Synthetic Control with 1 Country



Robustness: Leave-One-Out Distribution



Resources

- The material in this presentation comes from:
  - Abadie, A. and J. Gardeazabal. 2003. "The Economic Costs of Conflict: A Case Study of the Basque Country." **American Economic Review,** 93(1): 112-132.
  - 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, A., A. Diamond, and J. Hainmueller. 2015. "Comparative Politics and the Synthetic Control Method". **American Journal of Political Science,** 59(2): 495-510.
- Code: *Synth* (Stata, R, and Matlab). Available at:
  `http://web.stanford.edu/~jhain/synthpage.html`

Thank you!

`http://www.hks.harvard.edu/fs/aabadie/`