Synthetic Control Group Designs: Key Concepts, Recent Extensions, and An Application

Workshop In Methods
Alex Hollingsworth
Coady Wing
Today’s Talk Is Going To Be Great

What is the synthetic control method good for? And how does it work?

Guide to key bits of notation, central concepts, and confusing bits. Examples to make things concrete

An extension to the method that we have been working on.

Practice code for you to take home with you.

---

Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California’s Tobacco Control Program

Alberto ABACHE, Alesis DIAMOND, and Jens HAINMUeller

Buildings as an idea in Abache and Gaudiano (2012), this article investigates the application of synthetic control methods to comparative case studies. We discuss the advantages of these methods and apply them to study the effects of Proposition 99, a large-scale tobacco-control program that California implemented in 1988. We demonstrate that, following Proposition 99, tobacco consumption fell sensibly in California relative to a comparable synthetic control region. We estimate that by the year 2000 annual spending per capita on tobacco controls was about 26% lower than what it would have been in the absence of Proposition 99. Using new empirical methods proposed in this article, we demonstrate the significance of our estimates. Given that many policy interventions involve public policy remains in the social sciences and a question of interest is to estimate the aggregate outcome of interest like in Card 1990 and Card and Koeng 1994.

For the causal impact of interventions, the data are available for all of the aggregate outcomes. In particular, we are interested in the impact of the aggregate outcomes on interest, on information, on the data on sampled population. We can sometimes be used to estimate the aggregate outcomes. Interestingly, interest is in the social sciences such as policy problems that affect the empirical outcome. First, comparative case studies are likely to provide a higher standard of robustness about how comparability criteria are chosen. Researchers often select comparability groups on the basis of subjective assessment of the comparability of the units. Second, comparative case studies typically employ data on a sample of disaggregated units and inferential techniques that measure only uncertainty about the aggregate values of the data in the population. Uncertainty about the values of aggregate parameters can be estimated for aggregate data that are available. However, the availability of aggregate data does not imply that the effect of the event or intervention of interest can be estimated without error. Even if aggregate data are employed, there remains uncertainty about the ability of the control group to reproduce the conditional outcomes on the affected units would have experienced the absence of the intervention or event of interest. This type of uncertainty is not reflected by the standard errors computed with traditional inferential techniques for comparative case studies.

This article addresses current methodological shortcomings of case study analyses. We advocate the use of data-driven procedures to construct suitable comparison groups. As in Abache,
A lot of people are on the synthetic control train!

Maybe you should get on the train too?
What was the economic cost of the 1990 German Reunification?
After WWII Germany was split into two large blocks:
  West Germany + West Berlin was part of NATO
  East Germany was part of the Warsaw Pact

Then the Berlin Wall fell

In 1990, East and West Germany were re-unified.
East and West Germany had very different economies...

In the late 1980s, GDP per capita was about three times higher in West Germany than in East Germany.

Huge income disparities after being separated for 45 years.
How did reunification affect West Germany’s Economy?

What would West Germany GDP have looked like if it had not merged with East Germany?
Before 1990
West Germany had GDP Per Capita that was a bit higher than the OECD average.

The gap was actually growing a bit.

After 1990
GDP gap widens and then converges.

Is the OECD average a good “counterfactual”
Some notation...

There are \( j = 1 \ldots J + 1 \) units.

Each unit observed in periods \( t = 1 \ldots T \)

Unit \( j = 1 \) is treated
  Exposed to control condition from \( t = 1 \ldots T_0 - 1 \)
  Exposed to treatment condition from \( t = T_0 \ldots T \)

The other \( J \) units are untreated for all periods.
  “Donor pool”

German Reunification Example

The units are OECD countries.

Each country’s GDP per capita is observed in each country 1960 to 2003.

West Germany is unit 1. We don’t know it’s GDP after 1990 because it becomes “Germany”.
Treatment Effects

\( Y_{jt}^{(1)} \) is the treated potential outcome for unit \( j \) in period \( t \).
West Germany's PC GDP under re-unification.

\( Y_{jt}^{(0)} \) is the untreated potential outcome for unit \( j \) in period \( t \).
West Germany's PC GDP in the absence of re-unification.

\[ \alpha_{1t} = Y_{jt}^{(1)} - Y_{jt}^{(0)} \] is the treatment effect for unit \( j \) in period \( t \).

The goal is to compute the treatment effect for unit \( j \) in each period after \( T_0 \).

How much lower or higher was West Germany PC GDP because of re-unification?
Build a synthetic control group...

$Y_{st}^0$ is the untreated potential outcome for the synthetic unit in period $t$.

$$Y_{st}^0 = \sum_{j=2}^{J+1} \theta_j \times Y_j^0$$

The synthetic time series is a weighted average of the $J$ time series in the donor pool.

Example: Counterfactual West German PC GDP is "some" weighted average of PC GDP in the other OECD countries in the donor pool.
Terms that come up when you read about synthetic control weights

In the pre-treatment period, the treatment unit must be “inside the convex hull” of the donor pool.
   Interpolation and extrapolation.
   Non-negative weights that sum to 1.

How good is pre-treatment fit?
   RMSE
   Our idea: cohen’s d

Trimming rules
   Trim the donor pool
   Trim the treatment units in studies with multiple treatment units.
The synthetic control weights...

Don’t worry about how they obtain the weights just yet.

But notice that some of the donor pool countries get zero weight.

The synthetic control is actually a fairly small set of countries.

Austria, Japan, Netherlands, Switzerland, and the United States
Synthetic West Germany looks a lot like West Germany in the pre-period...

This is a balancing table.

You would see the same thing in a propensity score matching study, or in a randomized experiment.

Remember that the weights are chosen to minimize the imbalance in the pre-treatment time series.
Reunification did lower West Germany GDP!

**Before 1990**
Close match between synthetic control and real West Germany.

**After 1990**
Synthetic West Germany grew faster than real West Germany.

The gap implies a negative treatment effect of reunification.
Plot the difference between the two lines...

Over the 1990 to 2003 period, West German PC GDP was reduced by about $1600 per year, on average.

That’s about 8% of the 1990 level.

In 2003, synthetic West Germany’s PC GDP was about 12% higher than real West Germany.
Timing Based Placebo Test

Suppose we do the whole thing over again but pretend that re-unification happened in 1975.

If the synthetic control “discovers” an effect in 1975...then the whole thing is a bit unimpressive.
Country Based Placebo Tests

Pretend each of the control countries is the treated unit.

Use the same method to construct a synthetic comparison group for Norway, Greece, Italy,...

In each case, we pretend that “something happened” in 1990.
Summary Statistic to Gauge Performance

Compute the gap between the lines in each period.

Square the gap. (It’s a residual.)

Compute the average of the squared residuals over the pre-period. Take the square root.

Repeat for post period.

Compute the ratio: $\frac{RMSE_{post}}{RMSE_{pre}}$

Did the gap between the lines get bigger in the post period?

\[
RMSE_{pre} = \sqrt{\left(\frac{1}{T} \sum_{t=1960}^{1989} (Y_{jt} - Y_{st})^2\right)}
\]

\[
RMSE_{post} = \sqrt{\left(\frac{1}{T} \sum_{t=1960}^{1989} (Y_{jt} - Y_{st})^2\right)}
\]
Placebo Test Results

West Germany’s gap was much worse in the post period. (Because there is a big treatment effect.)

In the placebo countries, the fit is always a bit worse.

But West Germany is a big outlier.

This makes you think that it’s not just statistical noise. Germany
How do they choose the weights?
Nitty Gritty: how do they make the synthetic West Germany?

**Two Kinds of Weights:**
- Country weights
- Importance weights

**Pre-unification Attributes:**
- PC GDP
- Inflation Rate
- Industry Share of Value Added
- Investment Rate
- Schooling Level
- Measure of Trade Openness

Use **country weights** to make the synthetic time series

\[ Y_{st}^0 = \sum_{j=2}^{J+1} \theta_j \times Y_{jt}^0 \]

Choose country weights to minimize differences in pre-unification attributes.

But what if you are well matched on inflation rates and badly matched on trade openness?

Solution: minimize an importance weighted average of differences on pre-unification attributes.
Technical Version

\( X_1 \) is a \((k \times 1)\) vector of pre-treatment statistics in West Germany. PC GDP, Inflation Rate, Industry Share of Value Added, Investment Rate, Schooling Level, Measure of Trade Openness.

\( X_0 \) is a \((k \times J)\) matrix of the same pre-treatment statistics from the donor pool.

\( \theta \) is a \(J \times 1\) vector of country weights.

\( X_0 \theta \) is the \(k \times 1\) vector of country weighted pre-unification statistics.
Weighted sum of differences

Choose country weights to minimize:

$$\min_w ||X_1 - X_0 \theta||$$

Where: $$||X_1 - X_0 \theta|| = \sqrt{(X_1 - X_0 \theta)^T V (X_1 - X_0 \theta)}$$

V is a diagonal matrix of importance weights.

How important is it that the synthetic control and the treated unit “match” on each of the covariates.
Synthetic Control Using Lasso

An extension to regular Synthetic Control
Synthetic Control Using Lasso

\( \hat{\alpha}_{1t} = Y_{jt}^{(1)} - Y_{st}^{(0)} \) is the treatment effect for unit \( j \) in period \( t \).

How do we determine the synthetic control \( Y_{st}^{(0)} \)?

What kind of “things” can be in the “donor pool”?
Adopt a regression framework

\[ Y_{st}^0 \] is the untreated potential outcome for the synthetic unit in period \( t \).

\[
Y_{st}^0 = \sum_{j=2}^{J+1} Y_{jt}^0 \times \theta_j
\]

Equivalent to write it like:

\[
Y_{st}^0 = Y_{jt}^0 \theta
\]

The synthetic control looks a lot like the predicted value from a regression of the treated unit on the vector of donor pool units.
Could simply choose the synthetic control weights using OLS regressions...

\[ W_{GDP_t} = USGDP_t \theta_1 + UKGDP_t \theta_2 + \cdots + AustriaGDP_t \theta_J + \epsilon_t \]

This looks so easy. The synthetic control is the predicted value from the regression.

Coefficients are a mix of importance weights and country weights...But so what.

Relaxes some constraints: regular synth weights have to be non-negative, sum to 1, etc.
Problem with OLS approach

Very easy to overfit the time series and then there will be poor out of sample performance.

OLS can’t accommodate situations where the donor pool has more candidates than there are time periods to analyze.

Solution: LASSO
Choose weights with lasso

Choose weights to satisfy:

\[
\arg\min_\beta \left\{ \frac{1}{2N} \sum_{t=1}^{T_0} \left( Y_{1t} - \sum_{s=2}^{S} \beta_s Y_{st} \right)^2 + \lambda \left( \sum_{s=2}^{S} |\beta_s| \right) \right\}
\]

The first term is just regular OLS.
Second term is a penalty for complexity.
How does this work?

\[
\arg\min_\beta \left\{ \frac{1}{2N} \sum_{t=1}^{T_0} \left( Y_{1t} - \sum_{s=2}^{S} \beta_s Y_{st} \right)^2 + \lambda \left( \sum_{s=2}^{S} |\beta_s| \right) \right\}
\]

\(\lambda\) is a parameter that controls the penalty.

When \(\lambda = 0\) you have OLS.
When \(\lambda > 0\) you shrink the coefficients towards zero and sometimes you set some coefficients to zero. (Sparsity)

Choose \(\lambda\) using cross-validation

Split the pre-period into training and evaluation sets
Why not just do simple regression?

Set $\lambda = 0$ to obtain OLS Weights

Over-fitting to the pretest data may lead to poor out of sample forecasts.

Every candidate control gets “some” weight. It might be nice/interpretable to form the synthetic control a smaller number of states.

Regression doesn’t seem to let you “match” many pretest outcomes at once.

• That is, you may have more combinations of “donor units” and regressors than observations
Advantages From Lasso Weights

Useful when more regressors than observations

Regularization helps avoid overfitting, which aids extrapolation.

Allows for non-outcome of interest products to contribute to the prediction

Allows for negative weights

Interpretable- It’s still a regression, which is easy to understand.

It’s easy to extend to multivariate case using seemingly unrelated regression with Lasso.

Removes the researcher degree of freedom in model selection for both specification of interest and placebo goods that form the basis of statistical inference
Why would you have so many controls?

Donor pool could include:

- Time series of the dependent variable in other territories.
  Example: Annual Per capita GDP in a bunch of other countries.

- Time series of other variables in other territories.
  Examples:
  - Annual unemployment rate in a bunch of other countries.
  - Annual defense budget in a bunch of other countries.
  - Annual measures of the value of a stock index.
  - Etc., Etc., Etc.

Once you start adding things up, it’s easy for $J \gg T$. 
Application to Recreational Marijuana Laws in Colorado
What happens to Beer Sales when you legalize marijuana?

Use retail scanner data to measure the quantity of products sold every month in Colorado.

Scanner data are organized by UPC codes. There are so many types of alcohol and tobacco products!

And what is the best donor pool of comparison time series for each possible “type” of product?
What could possibly be a good synthetic control for this crazy time series?

- Light beer in other states?
- Wine in other states?
- Milk in other states?
- Shampoo, razor blades, diapers in other states?

The list is very large and the chance of overfitting is very high.

Important problem for a lot of newly available data sets.
Things we want to learn from the marijuana alcohol study...

What is the effect of the policy on substitution to other products?

Can we implement a synthetic control strategy in a very high dimensional setting and still make sense of the results?

Can we find effective ways to avoid researcher degrees of freedom, over fitting, etc.