Introduction to Synthetic Controls

Justin Eloriaga

Emory University

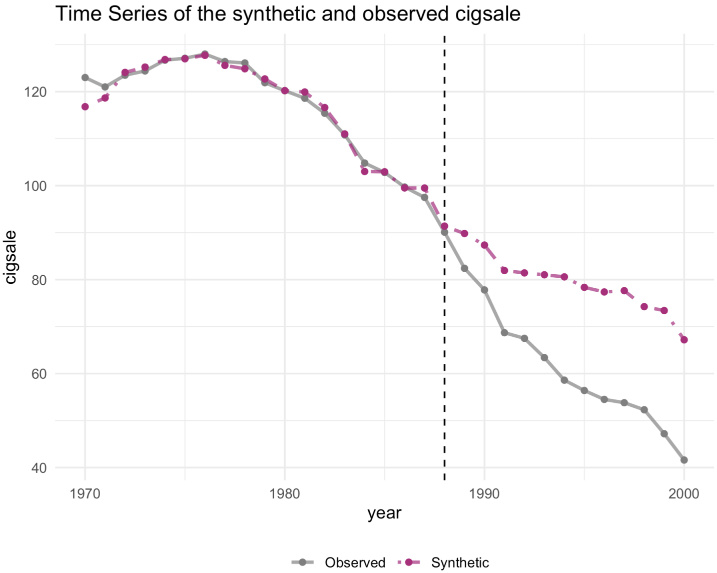
Roadmap

1. Motivation and intuition
2. Setup and notation
3. Building a synthetic control (what the R code does)
4. Results and interpretation
5. Inference: placebos and MSPE ratio
6. Assumptions, diagnostics, and good practice

Motivation

- We want the causal effect of a policy/treatment applied to one unit (e.g., California's Prop 99 on cigarette sales).
- No perfect control exists. Idea: build a "synthetic" control as a weighted average of other units.
- Choose weights so the synthetic unit closely tracks the treated unit before the intervention.
- After treatment, the gap between treated and synthetic is the estimated effect.

Idea in a picture



Setup and Notation

- Units $i = 1, \dots, J+1$: treated unit $i = 1$, donor pool $i = 2, \dots, J+1$.
- Time $t = 1, \dots, T$. Treatment starts at $T_0 + 1$.
- Outcome Y_{it} (e.g., per-capita cigarette sales).
- Synthetic weights $w = (w_2, \dots, w_{J+1})$ with $w_j \geq 0$ and $\sum_j w_j = 1$.

Weight selection (least-squares match in the pre-period):

$$\hat{w} = \arg \min_{w \geq 0,\ \mathbf{1}^\top w = 1} \sum_{t \leq T_0} \Big(Y_{1t} - \sum_{j=2}^{J+1} w_j Y_{jt} \Big)^2 .$$

Matching on Predictors (Abadie et al.)

- Often we also match on averages of predictors (covariates) over the pre-period: income, prices, demographics, etc.
- `tidysynth` builds these via `generate_predictor(...)` and outcome lags via `generate_predictor_balance()/generate_outcome()`.

Treatment effect (post-period gap):

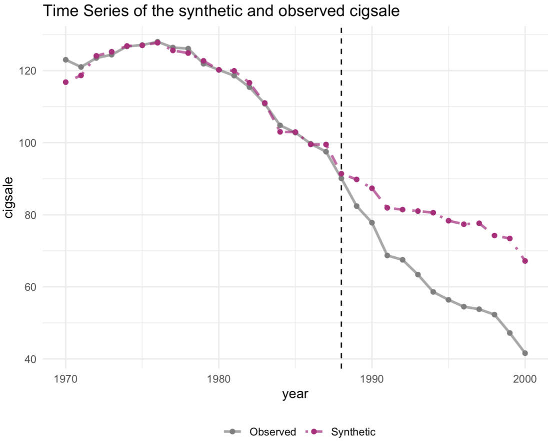
$$\hat{\tau}_t \; = \; Y_{1t} - \sum_{j=2}^{J+1} \hat{w}_j Y_{jt}, \qquad t > T_0.$$

What the R Code Does (Pipeline)

1. **Load libraries/data:** `library(tidyverse); library(tidysynth); data("smoking")`.
2. **Initialize:** `synthetic_control()` specifying treated unit (California), intervention year (1988), donor pool, and time window.
3. **Predictors:** `generate_predictor(...)` to add pre-period averages (e.g., log income, prices, youth share).
4. **Outcome lags:** `generate_outcome()` or predictor balance functions to include lagged Y for fit.
5. **Fit**: `fit_synthetic()` finds $\hat{w}$ subject to $w \geq 0$, $\sum w = 1$.
6. **Plots**: `plot_trends()`, `plot_weights()` for effects and composition.
7. **Placebos**: use `generate_placebos=TRUE` and `plot_placebos()`.
8. **MSPE ratio**: `plot_mspe_ratio()` for Fisher-style exact inference.

Pre- and Post-Treatment Trends

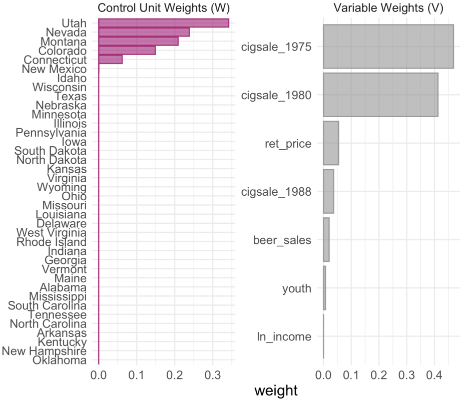
- Goal: tight pre-period fit; divergence post-period indicates an effect.
- Inspect the size and persistence of the post-period gap $\hat{\tau}_t$.



Dashed line denotes the time of the intervention.

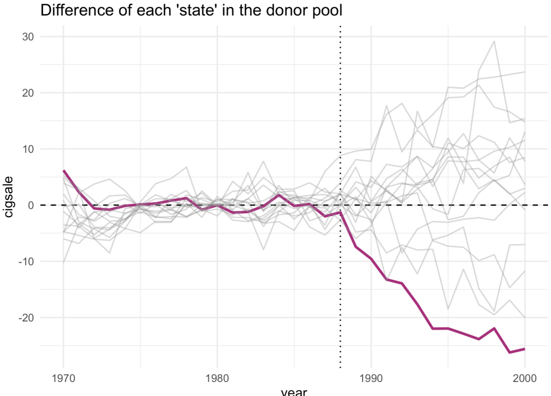
Which Donors Matter? (Weights)

- The synthetic is a convex combination of donor units.
- Heavier weights indicate donors that most resemble the treated unit pre-treatment.



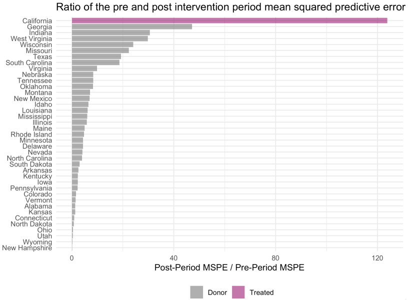
Inference via Placebos

- **Idea:** Reassign the treatment to each donor unit (one at a time) and rebuild synthetic controls to create “placebo gaps.”
- If California’s post-period gap is unusually large relative to placebos, that supports a real effect.



MSPE Ratio and Fisher-Style Exact Inference

- Mean Squared Prediction Error (MSPE):
  Pre-period: $\text{MSPE}^{\text{pre}} = \frac{1}{T_0}\sum_{t \leq T_0}(Y_{1t} - \hat{Y}^{\text{syn}}_{1t})^2$,
  Post-period: $\text{MSPE}^{\text{post}} = \frac{1}{T-T_0}\sum_{t > T_0}(Y_{1t} - \hat{Y}^{\text{syn}}_{1t})^2$.
- Ratio $R = \text{MSPE}^{\text{post}}/\text{MSPE}^{\text{pre}}$ for treated and each placebo unit.
- Rank R among all units (treated + placebos). A small (empirical) p-value indicates an unusually large post-period divergence.



Diagnostics and Good Practice

- **Pre-period fit:** RMSPE should be small; consider trimming donors with poor pre-fit.
- **Balance:** Check predictors/lag balance (tables or plots).
- **Sensitivity:** Try alternative predictor sets, leave-one-out donors, or narrower donor pools.
- **No interference:** Donor units should not be affected by the treatment.
- **Stability:** Outcomes should be driven by persistent factors that donors can span.

When to Use Synthetic Control

- Single (or few) treated units and a rich donor pool.
- Clear intervention time; long pre-period for learning weights.
- Outcomes where convex combinations are plausible counterfactuals.

When to be cautious

- Major shocks hitting donors differently than the treated unit.
- Sparse pre-period or unstable relationships.

Cheat Sheet: R Functions → Concepts

R Function	Concept
`synthetic_control()`	Choose treated unit, donor pool, time windows
`generate_predictor()`	Build pre-period predictor means/covariates
`generate_outcome()`	Add outcome lags for better fit
`fit_synthetic()`	Solve for nonnegative weights summing to one
`plot_trends()`	Pre/post trends and estimated effect (gap)
`plot_weights()`	Contribution of each donor
`plot_placebos()`	Permutation (placebo) gaps for inference
`plot_mspe_ratio()`	Fisher-style p-value via MSPE ratios

Key Takeaways

- Build a synthetic control that mimics the treated unit before treatment.
- The post-period gap is the treatment effect estimate.
- Use placebo tests and MSPE ratios to gauge significance.
- Always check pre-fit, balance, and sensitivity.

Questions?