Experimental Analysis of Public Policies

Session 8

Synthetic Control, Heterogeneity, and Replication

Etienne Dagorn - Universite de Lille

Session Objectives

By the end of this session, students should be able to:

- → explain how synthetic control constructs a missing counterfactual path;
- → choose and audit a donor pool, predictors, weights, and pre-treatment fit;
- → interpret post-treatment gaps as causal only under explicit design assumptions;
- → use placebo and donor-sensitivity checks as credibility evidence, not automatic proof;
- → connect heterogeneity to ATE, ATT, LATE, group-time ATT, RDD, and SCM estimands;
- → distinguish credible pre-specified heterogeneity from post-hoc fishing;
- → audit replication, robustness, and external validity as part of causal interpretation;
- → choose methods by identifying variation rather than by estimator names.

Roadmap

Course Synthesis and Method Choice

Synthetic Control as Constructed Counterfactual

Stress-Testing Synthetic Control

Heterogeneity and External Validity

Replication and Course Close

Roadmap

Course Synthesis and Method Choice

Synthetic Control as Constructed Counterfactual

Stress-Testing Synthetic Control

Heterogeneity and External Validity

Replication and Course Close

From Local Designs to Aggregate Cases

- → Session 7 used a threshold to create a local comparison.
- → Many major policies affect one city, region, country, or institution.
- → There may be no lottery, no sharp cutoff, and few treated units.
- → The final question is how to build and challenge a credible counterfactual.

Central question

How do we synthesise, interpret, and challenge causal evidence?

Method Choice

Design	Identifying variation	Core assumption
RCT	random assignment Z	assignment independent of potential outcomes
IV	instrument-induced D	relevance, independence, exclusion, monotonicity
DiD	policy timing	parallel untreated trends
RDD	threshold rule	continuity at the cutoff
SCM	weighted donor path	donor pool reproduces untreated path

Reflex: do not begin with a favourite estimator. Begin with the comparison you can defend.

Choosing a Design

Three scenarios

1. A school lottery allocates scarce tutoring places.
2. A policy is adopted by one region in 2025, with no obvious cutoff.
3. A benefit begins exactly when income falls below a threshold.

Student output: for each scenario, name the likely method, source of identifying variation, key assumption, and local or population estimand.

Final Evidence Audit

Question	What policy effect is being claimed?
Ideal experiment	Randomly assign the policy to the target population.
Counterfactual	The treated unit or group without the policy.
Variation	Lottery, instrument, timing, threshold, or donor pool.
Assumption	The observed comparison reproduces the missing counterfactual.
Threat	Selection, invalid instrument, non-parallel trends, manipulation, poor donor fit, or limited transport.

Today: SCM builds the counterfactual; heterogeneity asks where it applies; replication asks whether we should believe it.

Roadmap

Course Synthesis and Method Choice

Synthetic Control as Constructed Counterfactual

Stress-Testing Synthetic Control

Heterogeneity and External Validity

Replication and Course Close

Why Synthetic Control?

- → A single treated unit receives a policy at a known date.
- → No untreated unit alone is a convincing comparison.
- → A weighted combination of untreated units may approximate the treated unit before treatment.
- → The post-treatment gap then becomes the object to interpret.

Counterfactual idea

The synthetic unit is credible only if the donor pool can reproduce the treated unit's untreated path.

SCM Workshop: Build the Donor Pool

Scenario: one region introduces a carbon-pricing policy in 2025.

1. Which regions are plausible untreated donors?
2. Which regions should be excluded because of spillovers or simultaneous policies?
3. Which pre-treatment predictors should the synthetic unit match?
4. What would make pre-treatment fit unconvincing?

Visible output

Write a donor-pool rule and one exclusion criterion before seeing any weights.

Donor Pool and Predictors

Choice	Question
Donor pool	Which untreated units could plausibly have received the policy?
Exclusions	Which units face spillovers, anticipation, or simultaneous reforms?
Predictors	Which pre-treatment variables forecast the outcome path?
Pre-period outcomes	Does the synthetic unit track the treated unit before treatment?

Bad-control echo: predictors should be pre-treatment or otherwise unaffected by the policy.

Weights

$$\widehat{Y}_{1t}(0) = \sum_{j=2}^{J+1} w_j Y_{jt}, \qquad w_j \geq 0, \quad \sum_j w_j = 1.$$

- → Unit 1 is treated; units $2, \ldots, J+1$ are donors.
- → Weights choose a synthetic untreated path for the treated unit.
- → Large weights identify which donors drive the comparison.
- → Zero weights are also informative: many donors may not help match the treated path.

Interpretation: the weights describe the comparison group, not the treatment effect.

Convex Hull

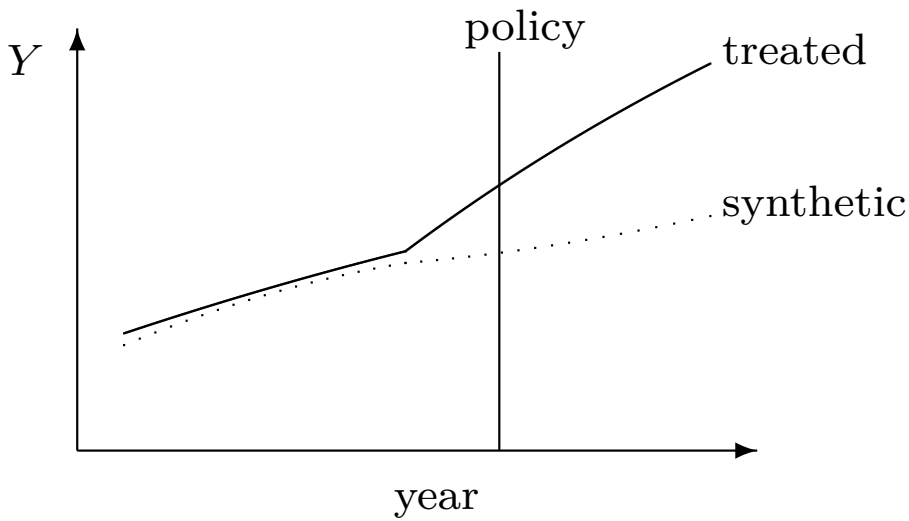
Support problem

Synthetic control cannot construct every possible treated unit. The treated unit must be approximable by weighted untreated units.

- → If the treated unit is outside the donor support, fit may be poor.
- → Extrapolation is limited by non-negative weights summing to one.
- → Poor support weakens credibility even if the post-treatment gap is large.

Connection to Session 2: this is the aggregate-unit version of common support.

Pre-Treatment Fit



- → Good pre-fit is necessary for credibility.
- → It is not proof of identification.
- → Poor pre-fit makes the post-treatment gap hard to interpret.

Treatment Gap

$$\widehat{\tau}_{1t} = Y_{1t} - \sum_{j=2}^{J+1} w_j Y_{jt}, \qquad t \geq T_0.$$

- → Y_{1t} is the treated unit's observed post-treatment outcome.
- → The weighted donor outcome estimates the missing untreated path.
- → The gap is time-specific for the treated aggregate unit.
- → It is causal only under a credible donor pool, good pre-treatment fit, no confounding post-treatment shocks, and no spillovers.

Example: Regional Carbon Policy

Object	Example choice
Treated unit	region adopting carbon pricing in 2025
Outcome	emissions per capita or industrial employment
Donors	regions without carbon pricing and without spillovers
Pre-fit	emissions path, industry mix, energy prices, population
Gap	treated minus synthetic outcome after 2025

Interpretation

Would a post-2025 energy shock unique to the treated region be an effect, a confounder, or part of the policy environment?

Roadmap

Course Synthesis and Method Choice

Synthetic Control as Constructed Counterfactual

Stress-Testing Synthetic Control

Heterogeneity and External Validity

Replication and Course Close

Placebo in Space

- → Reassign the treatment label to donor units one at a time.
- → Build a synthetic control for each placebo unit.
- → Compare the treated gap to placebo gaps.
- → Exclude or downweight placebo comparisons with very poor pre-treatment fit.

Inference logic

Placebo ranks are most meaningful when donors are plausible alternative treated units with comparable pre-fit.

Placebo in Time

- → Pretend treatment began earlier than it did.
- → A large fake gap before treatment weakens credibility.
- → This is analogous to pre-trend and placebo checks in DiD.
- → The check probes fit and timing, not all identifying assumptions.

Student output: explain what a large pre-2025 placebo gap would imply for the carbon-policy example.

Donor Sensitivity

- → Drop one influential donor and re-estimate.
- → Drop donors exposed to possible spillovers.
- → Vary predictor sets and pre-treatment windows.
- → Check whether the qualitative conclusion depends on one donor or one modelling choice.

Interpretation

Robustness is convincing when it addresses a named threat, not when it is a long menu of unchecked alternatives.

Inference in SCM

- → SCM usually has few treated units, so large-sample formulas are limited.
- → Placebo and permutation logic are common.
- → Ranks and pseudo-p-values should be interpreted cautiously.
- → A donor with poor pre-fit is not a good placebo benchmark.

Reporting habit: show the treated gap, placebo gaps, pre-fit quality, and donor sensitivity together.

SCM vs DiD

	DiD	**SCM**
Comparison	untreated group trend	weighted donor path
Key assumption	parallel untreated trends	synthetic path approximates $Y_{1t}(0)$
Strength	simple and transparent	visible construction for one treated unit
Threat	different trends	poor donor support or post shocks

Connection: both ask whether untreated units can reveal the treated unit's missing untreated path.

Roadmap

Course Synthesis and Method Choice

Synthetic Control as Constructed Counterfactual

Stress-Testing Synthetic Control

Heterogeneity and External Validity

Replication and Course Close

Heterogeneous Treatment Effects

- → "Does it work?" is rarely the final policy question.
- → Effects can differ by baseline need, institution, timing, compliance behaviour, or exposure duration.
- → Heterogeneity matters for targeting and scale-up.
- → The credibility of a heterogeneity claim inherits the credibility of the underlying design.

Policy question

For whom, where, and under what implementation conditions should the policy be expanded?

Local Estimands Across the Course

Design	Estimand	Locality
Session 1	ATE, ATT, ATU	population or treated status
Session 4	LATE	compliers moved by Z
Session 6	$ATT(g,t)$	cohort and calendar time
Session 7	$\tau_{SRD}(c)$, $\tau_{FRD}(c)$	cutoff and local compliers
Session 8	$\widehat{\tau}_{1t}$	treated aggregate unit and time

External-validity reflex: local credibility is valuable, but it narrows where the estimate travels.

Credible Heterogeneity

- → Pre-specify one or two theoretically motivated dimensions.
- → Prefer baseline need, pre-treatment predicted exposure, randomised dosage, or eligibility intensity.
- → Show subgroup sample sizes and uncertainty.
- → Explain why the design remains credible within the subgroup contrast.

Avoid

Conditioning on realised post-treatment exposure can change the estimand or introduce selection bias.

Interactions and Subgroups

$$Y_i = \alpha + \tau D_i + \theta(D_i \times HighNeed_i) + \gamma HighNeed_i + u_i.$$

- → τ: treatment contrast for lower-need units.
- → θ: difference in the treatment contrast for high-need units.
- → $\tau + \theta$: treatment contrast for high-need units.
- → These are causal effects only if the treatment contrast is identified within the relevant groups.

Reading habit: ask whether subgroup analysis was planned before seeing the results.

Multiple Testing

Fishing problem

If many subgroups are searched after the fact, some large differences will appear by chance.

→ Limit the number of primary heterogeneity dimensions.

→ Report all tested dimensions, not only the most exciting one.

→ Treat exploratory heterogeneity as hypothesis-generating.

→ Compare subgroup estimates to policy relevance, not only stars.

Roadmap

Course Synthesis and Method Choice

Synthetic Control as Constructed Counterfactual

Stress-Testing Synthetic Control

Heterogeneity and External Validity

Replication and Course Close

Replication

- → Replication asks whether the data, code, specification, and interpretation support the claim.
- → It is not only rerunning a regression.
- → A replication audit checks design choices and assumptions as well as arithmetic.
- → Robustness should be tied to the most credible threat.

Final question

What evidence would make you trust the result less?

Replication Audit: Debrief Grid

Audit item	**Question**
Sample	Who is included or excluded?
Variables	Do Z, D, Y, X, and timing match the claim?
Specification	Is the estimator aligned with the design?
Inference	Are standard errors or placebo checks design-aware?
Robustness	Does each check address a named threat?
Interpretation	Is the estimand local or population-wide?

Specification Curve

- → A specification curve displays results across defensible choices.
- → It can reveal whether a conclusion depends on one arbitrary specification.
- → It should be based on choices justified before looking at the answer.
- → It is not a licence to search until a preferred estimate appears.

Connection to robustness: a curve is most informative when each specification reflects a real design uncertainty.

Robustness Checklist

1. What is the main identifying assumption?
2. Which diagnostic most directly probes it?
3. Which alternative specification addresses the main threat?
4. Does the estimand change under the alternative?
5. Does the conclusion survive in magnitude, sign, and policy meaning?

Method Map: Assignment and Timing

Method	Variation	Assumption	Estimand
RCT	assignment Z	randomisation	SATE, ATE, ITT
IV	instrument Z	exclusion, monotonicity	LATE
DiD	timing	parallel trends	treated-period ATT

Shared question

Does the comparison reconstruct the right missing counterfactual?

Method Map: Local and Constructed

Method	Variation	Assumption	Estimand
Staggered DiD	cohort timing	clean comparisons	$ATT(g,t)$
RDD	cutoff rule	continuity	cutoff effect
SCM	donor weights	credible synthetic path	treated-unit gap

Main threats

Already-treated controls, cutoff manipulation, poor pre-treatment fit, spillovers, and contemporaneous shocks change the interpretation of the design.

Final Exam Skills

- → Translate a policy question into an estimand.
- → Identify the missing counterfactual.
- → Name the source of identifying variation.
- → State the assumption in plain language and notation when useful.
- → Interpret the estimate for the correct population.
- → Separate identification, estimation, inference, and external validity.

Course Synthesis

The course reflex

1. Where does the identifying variation come from?
2. What assumption makes the comparison causal?
3. What estimand or population does the resulting estimate describe?

→ Good empirical work defends a comparison.

→ The method name is secondary to the credibility of the counterfactual.

→ Local evidence can be powerful when interpreted honestly.

Common Mistakes

Watch for these

- → Treating pre-fit, balance, or pre-trends as proof.
- → Calling a local estimate an ATE without justification.
- → Interpreting subgroup coefficients causally without subgroup identification.
- → Treating placebo ranks as generic large-sample p-values.
- → Reporting robustness checks that do not address a specific threat.
- → Starting with a method name before defining the counterfactual.

Practice Bridge

- → Theory sheet: `exercises/theory/session_08_theory.md`.
- → QCM: `assessments/qcm/session_08_qcm.md`.
- → Final exam: `assessments/exams/final_exam_template.md`.
- → Mini code task: inspect donor weights, pre-fit, treatment gaps, and placebo gaps.

Student deliverable: a one-page evidence audit: counterfactual, assumption, diagnostic, estimate, inference, external-validity claim, and remaining doubt.

Key Takeaways

- → Synthetic control constructs the missing untreated path with weighted donors.
- → SCM evidence depends on donor validity, pre-fit, placebo logic, and sensitivity.
- → Heterogeneity changes interpretation and external validity.
- → Replication is a causal-design audit, not only a coding exercise.
- → The best method is the one whose comparison and assumptions you can defend.

Closing line: first defend the comparison; then interpret the estimate.

Appendix: Extensions and References

- → Abadie, Diamond, and Hainmueller (2010), "Synthetic Control Methods for Comparative Case Studies".
- → Abadie (2021), "Using Synthetic Controls: Feasibility, Data Requirements, and Methodological Aspects".
- → Brodersen et al. (2015), "Inferring Causal Impact Using Bayesian Structural Time-Series Models".
- → Simonsohn, Simmons, and Nelson (2020), "Specification Curve Analysis".
- → Steegen et al. (2016), "Increasing Transparency Through a Multiverse Analysis".
- → Athey et al. (2021), "Matrix Completion Methods for Causal Panel Data Models".
- → Arkhangelsky et al. (2021), "Synthetic Difference-in-Differences".

Reading reflex

For any extension, ask what counterfactual it constructs and what assumption makes that construction credible.