Panel Counterfactual Estimators for Policy Shocks under Latent Trends

A Simulation Study Motivated by the Italy–Libya MoU

Giorgio Coppola

2025-12-22

1 Introduction

Difference-in-differences (DiD) is the classic method for evaluating policy interventions in observational panel data. The "canonical" estimator for many periods/multiple groups panels is the two-way fixed effects (TWFE), which relies on the parallel trends assumption: absent treatment, treated and control units would have followed parallel trajectories. However, this assumption often fails in practice when units exhibit complex or heterogeneous trends driven by unobserved time-varying factors. Recent methodological advances propose estimators robust to violations of parallel trends by modeling low-rank latent factor structures. This project compares four panel counterfactual estimators for a spatial policy shock: TWFE, Matrix Completion (MC), Synthetic Difference-in-Differences (SDID), and the Triply Robust Panel Estimator (TROP). Using Monte Carlo simulations calibrated to the Missing Migrants Project data for the Central Mediterranean from IOM, this project assesses each method's bias, coverage, and power under varying-strength interactive fixed effects. The motivating application is the Italy–Libya Memorandum of Understanding (which came into effect in February 2017), which led to the externalization of border control from the Italian Coast Guard to Libyan Coast Guard patrols, starting in May 2017. This agreement has been, and still is, at the center of the Italian and European political and academic debate, questioning its effectiveness, as well as its ethical premises, as externalization practices are often accompanied by the criminalization of NGO's SAR operations, and therefore they are associated with increased mortality and dangerousness of the migratory routes. If this policy landscape serves as a backdrop, the research question of this project is methodological: when latent trends conflict with parallel trends (as is often the case in these kinds of settings), which estimator should practitioners use?

2 Background

Consider a balanced panel with units $i = 1, \ldots, N$ observed over periods $t = 1, \ldots, T$. The potential outcomes framework defines $Y_{it}(0)$ and $Y_{it}(1)$ as outcomes under control and treatment. A binary treatment indicator $D_{it} \in \{0, 1\}$ determines the observed outcome:

$$Y_{it} = Y_{it}(0) + D_{it} \cdot \tau_{it}, \quad \text{where } \tau_{it} = Y_{it}(1) - Y_{it}(0).$$

In a canonical 2×2 DiD design, treatment turns on for a subset of units after period T_0. TWFE estimates the average treatment effect on the treated (ATT) by regressing:

$$Y_{it} = \alpha_i + \lambda_t + \delta D_{it} + \varepsilon_{it},$$

where α_i and λ_t are unit and time fixed effects. This is unbiased for ATT if parallel trends holds: $E[Y_{it}(0)|i \in \text{treated}] - E[Y_{it}(0)|i \in \text{control}] = c$ (constant over time).

Parallel trends fails when units follow heterogeneous trends. A common generalization is the interactive fixed effects model (Bai 2009):

$$Y_{it}(0) = \alpha_i + \lambda_t + \lambda_i' F_t + \varepsilon_{it},$$

where $\lambda_i \in \mathbb{R}^K$ are unit-specific loadings and $F_t \in \mathbb{R}^K$ are common time factors. Critically, TWFE is

biased when factor loadings λ_i are correlated with treatment assignment, that is, when treated units have systematically different loadings than control units, creating differential trends.

Recent work proposes estimators that accommodate latent trends:

- Matrix Completion (MC): Models the untreated potential outcome matrix $Y(0)$ as low-rank, using singular value decomposition (SVD) to impute counterfactuals for treated observations (Athey et al. 2021).
- Synthetic Difference-in-Differences (SDID): Constructs synthetic controls using pre-period matching, combining unit weights (to match treated units' pre-period means) with time weights (to emphasize recent pre-periods), then applies a DiD-style contrast (Arkhangelsky et al. 2021).
- Triply Robust Panel Estimator (TROP): Combines unit weights, time weights, and regression adjustment. Consistent if any one of three conditions holds: unit weights balance loadings, time weights balance factors, or the regression adjustment is correctly specified (Athey et al. 2025).

3 Methods

A spatial panel of $N = 56$ grid cells ($1^\circ \times 1^\circ$ resolution) in the Central Mediterranean is simulated, observed monthly from April 2015 to February 2018 ($T = 35$). The treatment is defined as:

$$D_{it} = 1\{\text{distance}_i \leq 200 \text{ km from Tripoli}\} \times 1\{t \geq \text{May 2017}\}.$$

The outcome $Y_{it}^{\text{any}} = 1\{\text{deaths observed in cell } i, \text{ month } t > 0\}$ is a binary deadly-event indicator. Deaths are generated hierarchically as (1) Exposure, where migration attempts $M_{it} \sim \text{Poisson}(\mu_{it})$, where μ_{it} includes seasonality and latent trends; (2) mortality, where each attempt has per-attempt mortality probability p_{it} following a logistic model with interactive fixed effects, and (3) deaths, $\text{Deaths}_{it} \sim \text{Binomial}(M_{it}, p_{it})$. The DGP is calibrated to match pre-period deadly-event rates (~6.6%) from IOM Missing Migrants data. I vary `factor_strength` (0 = no latent factors, 0.5 = moderate, 0.9 = strong), `loading_correlation` (correlation between loadings and treatment assignment), and the treatment effect size `delta` on the log-odds scale. All estimators target the ATT for the binary outcome Y^{any}, treating it as continuous (linear probability model) for comparability. TWFE uses `feols(Y ~ D | unit_id + time_id)` with cluster-robust SEs. Matrix Completion uses the `fect` package with SVD-based imputation and nuclear norm regularization, with bootstrap SEs. SynthDID uses the `synthdid` package with jackknife SEs for valid inference. TROP uses exponential kernel weights for units and time with low-rank regression adjustment, with bootstrap SEs.

4 Simulation Design

Two experiments are conducted:

1. Power analysis: Fix `factor_strength = 0.5` and `loading_correlation = 0.5` (moderate latent trends correlated with treatment) and vary treatment effect $\delta \in \{0, 0.2, 0.4, 0.6, 0.8\}$. Run 200 replications per δ value.
2. Scenario analysis: Fix $\delta = 0.6$ and vary DGP features (parallel trends, moderate/strong latent trends, underreporting, short-lived effect). Run 100 replications per scenario.

Performance metrics: bias, RMSE, 95% CI coverage (nominal target), and rejection rate (power/size at $\alpha = 0.05$).

5 Results

5.1 One-dataset estimation example

Table 1 reports estimates on the calibrated DGP for a single simulated dataset (balanced 56×35 panel). This "one draw" is useful to see what an applied analyst would face in practice, but it can be noisy when outcomes are sparse.

Table 1: One-dataset estimates (calibrated DGP; outcome $Y_{it} = 1\text{deaths} > 0$).

method	estimate	se	true_att	bias	p_value
TWFE	0.211	0.039	0.128	0.083	1.20e-06
Matrix Completion (fect)	0.211	0.036	0.128	0.083	0.00e+00
TROP	0.195	0.047	0.128	0.066	3.93e-05
SynthDiD	0.059	0.086	0.128	-0.070	4.95e-01

In this example: TWFE and Matrix Completion (via `fect`) coincide almost exactly (same point estimate, slightly different SEs). This indicates that the `fect` MC fit is effectively behaving like a two-way fixed effects model in this design (e.g., selecting few/no latent factors), so it should not be interpreted as a meaningfully different estimator here. I could not find evidence that MC is delivering a low-rank robustness correction, but this should be better investigated. TROP is closer to the true ATT in this draw and its confidence interval covers the true effect. SynthDiD is imprecise (large SE) and does not reject in this draw; its point estimate is below the true ATT.

5.2 Power Analysis

Figure 1 and Table 2 summarize Monte Carlo performance across treatment strengths (log-odds shifts) under the calibrated DGP.

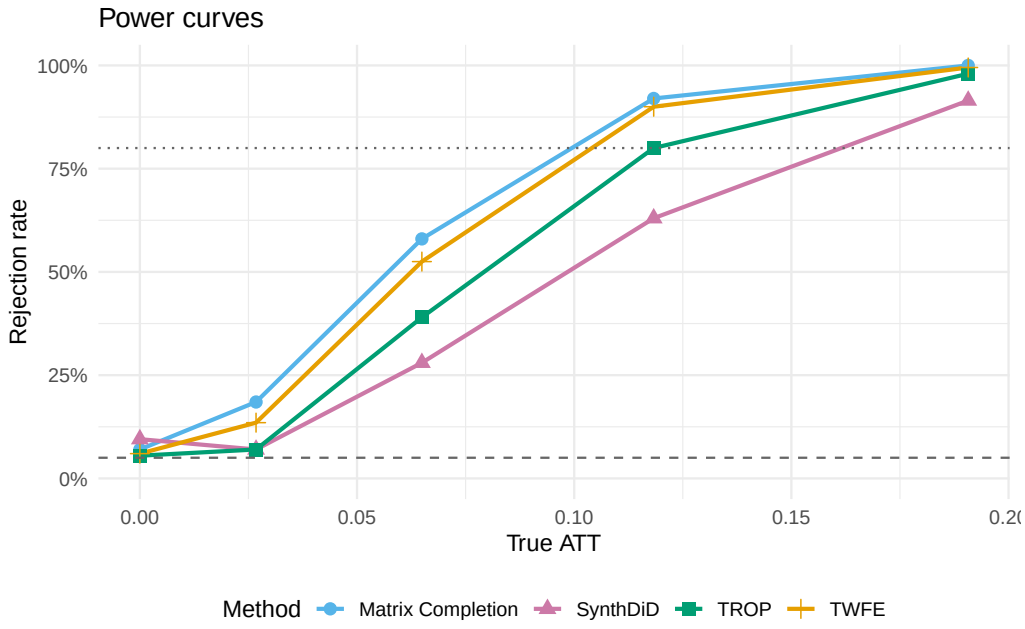

Figure 1: Rejection rates (power/size) across treatment strengths under the calibrated DGP.

Key takeaways from the calibrated power study include that size control is imperfect for some methods. When $\delta = 0$ (no effect), TWFE is close to nominal size, whereas Matrix Completion (via `fect`) and SynthDiD reject too often, indicating inflated Type I error. TROP is closest to nominal size and exhibits the best coverage at $\delta = 0$. Matrix Completion (via `fect`) does not improve point-estimation accuracy relative to TWFE in this design. Across values of δ, MC and TWFE display essentially the same average bias and RMSE; differences mainly arise in inference (standard errors, coverage, and power), rather than in point estimates. Overall, the results reflect a bias–variance tradeoff. TROP behaves more conservatively, with higher coverage and lower rejection rates, while SynthDiD is typically less powerful and tends to attenuate estimated effects in this binary-outcome setting.

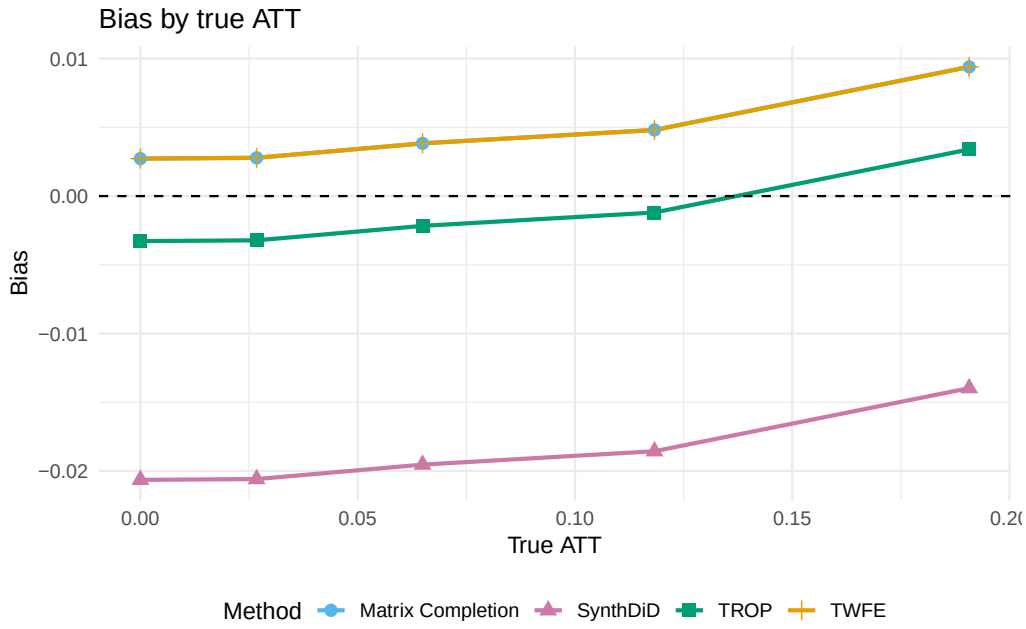

Figure 2: Bias as a function of treatment strength under the calibrated DGP.

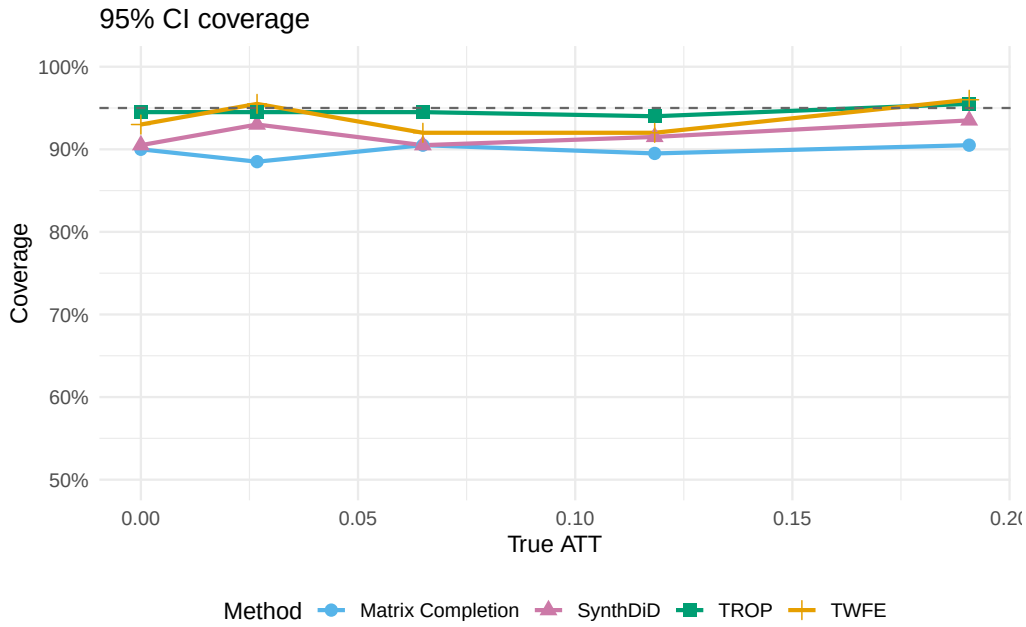

Figure 3: Coverage of nominal 95% intervals under the calibrated DGP.

Table 2: Monte Carlo summary at a moderate effect size ($\delta = 0.8$).

method	true_att	mean_estimate	bias	rmse	coverage	rejection_rate
Matrix Completion	0.065	0.069	0.004	0.034	0.905	0.580
TWFE	0.065	0.069	0.004	0.034	0.920	0.525
TROP	0.065	0.063	-0.002	0.037	0.945	0.390
SynthDiD	0.065	0.045	-0.020	0.046	0.905	0.280

5.3 Scenario Analysis

The scenario analysis varies the strength of latent trends, measurement error, and whether the treatment effect is transient (May–June only). Table 3 summarizes performance.

Table 3: Scenario analysis (mean performance across simulations).

scenario	method	true_att	bias	rmse	coverage	power
Moderate latent trends	Matrix Completion	0.165	0.007	0.054	0.86	0.94
Moderate latent trends	SynthDiD	0.165	0.013	0.065	0.86	0.86
Moderate latent trends	TROP	0.165	0.003	0.067	0.89	0.76
Moderate latent trends	TWFE	0.165	0.007	0.054	0.91	0.92
Parallel trends (no latent factors)	Matrix Completion	0.172	-0.004	0.042	0.94	0.96
Parallel trends (no latent factors)	SynthDiD	0.172	0.005	0.060	0.91	0.89
Parallel trends (no latent factors)	TROP	0.172	0.000	0.053	0.98	0.87
Parallel trends (no latent factors)	TWFE	0.172	-0.004	0.042	0.95	0.95
Short-lived (May–June only)	Matrix Completion	0.124	-0.091	0.107	0.45	0.18
Short-lived (May–June only)	SynthDiD	0.124	0.000	0.082	0.92	0.42
Short-lived (May–June only)	TROP	0.124	-0.005	0.085	0.92	0.38
Short-lived (May–June only)	TWFE	0.124	-0.003	0.075	0.94	0.50
Strong latent + underreporting	Matrix Completion	0.150	0.022	0.074	0.78	0.82
Strong latent + underreporting	SynthDiD	0.150	0.034	0.084	0.81	0.76
Strong latent + underreporting	TROP	0.150	0.028	0.077	0.82	0.76
Strong latent + underreporting	TWFE	0.150	0.022	0.074	0.82	0.81
Strong latent trends	Matrix Completion	0.150	0.004	0.067	0.78	0.81
Strong latent trends	SynthDiD	0.150	0.014	0.069	0.82	0.78
Strong latent trends	TROP	0.150	0.005	0.070	0.84	0.76
Strong latent trends	TWFE	0.150	0.004	0.067	0.83	0.78

The main qualitative patterns are: in persistent-effect scenarios, TWFE and Matrix Completion behave similarly in terms of bias and RMSE, but Matrix Completion tends to undercover more and can appear "more powerful" largely because it reports smaller standard errors. In the short-lived effect scenario, Matrix Completion can break down sharply, with large bias and severe under-coverage, which is consistent with a mismatch between the estimator's implicit treatment structure and a transient policy shock. In that scenario, TROP and SynthDiD remain much more stable, albeit with lower power. These simulations are designed to be application-motivated rather than "best case" for any method: sparse binary outcomes, interactive fixed effects with loadings correlated with treatment, and a limited number of treated units. The most important empirical takeaway from the current implementation is that Matrix Completion via `fect` is not acting like a distinct low-rank imputation estimator in this design. In both the one-dataset illustration and the Monte Carlo summaries, its point estimates track TWFE extremely closely. When this happens, MC should be interpreted as "TWFE with different uncertainty quantification," not as a robustness improvement.

6 Future Directions

Several extensions merit investigation:

1. SDID for discrete outcomes: Develop variants of SDID with link functions (logit, Poisson) appropriate for binary or count data, or test performance with continuous transformations like Y_{ihs}.
2. Heterogeneous treatment effects: Current simulations assume constant δ across treated units. Realistic settings often exhibit effect heterogeneity by distance or time.

3. Optimal rank selection: Investigate data-driven methods for choosing the number of latent factors in MC and TROP (e.g., cross-validation, information criteria).
4. Inference under interference: Spatial settings may exhibit spillover effects between adjacent cells, violating SUTVA.