

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), and the Synthetic Difference-in-Differences (SDID). 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 $i = 1, \dots, N$ observed over periods $t = 1, \dots, T$. 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.

3 Methods

Recent work proposes estimators that accommodate latent trends. Matrix Completion models the untreated potential outcome matrix $Y(0)$ as low-rank, using singular value decomposition to impute counterfactuals for treated observations (Athey et al. 2021). Synthetic Difference-in-Differences 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). We compare these two with TWFE. The latter uses a two-way fixed-effects regression with unit-level cluster-robust standard errors. Matrix Completion uses the `fact` package with nuclear norm regularization and cross-validation for lambda selection, with bootstrap standard errors based on 200 replicates. Synthetic DiD uses the `synthdid` package with placebo-based standard errors. All estimators target the ATT for the binary outcome Y_{it}^{any} , treating it as continuous (i.e., a linear probability model) for comparability.

The three estimators are tested with spatial panel of $N = 56$ grid cells ($1^\circ \times 1^\circ$ resolution) in the Central Mediterranean 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. The latent index determining the probability of observing any deaths follows a logistic model with interactive fixed effects:

$$\text{logit}(p_{it}) = \alpha_i + \gamma_t + s \cdot \lambda'_i F_t + u_{it} + \delta \cdot D_{it},$$

where α_i are unit effects, γ_t captures seasonality and trend, $\lambda'_i F_t$ represents latent factor structure with strength controlled by parameter s , and u_{it} is AR(1) serial noise. The DGP is calibrated to match pre-period deadly-event rates (approximately 6.6%) from IOM Missing Migrants data. I vary factor strength s (where 0 means no latent factors, 0.9 represents the baseline, and 2 represents strong factors) and the treatment effect size δ on the log-odds scale.

4 Simulation Design

Two experiments are conducted. The power analysis fixes factor strength at $s = 0.9$ (moderate latent trends correlated with treatment) and varies the treatment effect $\delta \in \{0, 0.2, 0.4, 0.6, 0.8\}$, running 200 replications per value. The scenario analysis fixes $\delta = 0.6$ and varies DGP features across five conditions: baseline (realistic calibration), parallel trends world ($s = 0$), strong latent trends ($s = 2$), under-reporting ($q = 0.7$), and short-lived effect (exponential decay with half-life of 2 months). Each scenario runs 100 replications. Performance metrics include bias, RMSE, 95% CI coverage (nominal target 0.95), and rejection rate (power when $\delta > 0$, size when $\delta = 0$, at $\alpha = 0.05$).

5 Results

5.1 Power Analysis

Figure 1 and Figure 2 summarize Monte Carlo performance across treatment strengths under the calibrated DGP. The power analysis reveals a critical finding regarding the validity of the inference. At $\delta = 0$ (no true effect), both TWFE and Matrix Completion maintain proper size control, with rejection rates near the nominal 5% level (0% for TWFE, 0.5% for MC). However, SynthDiD exhibits severe size distortion, rejecting the null hypothesis 14.5% of the time when it is true. This indicates that SynthDiD's confidence intervals are systematically too narrow for this sparse binary outcome setting. All three estimators display similar positive bias that increases modestly with effect size, ranging from approximately 0.008 at $\delta = 0$ to 0.017 at $\delta = 0.8$. This upward bias is consistent across methods and reflects the challenges of estimating treatment effects with sparse binary outcomes and latent confounding.

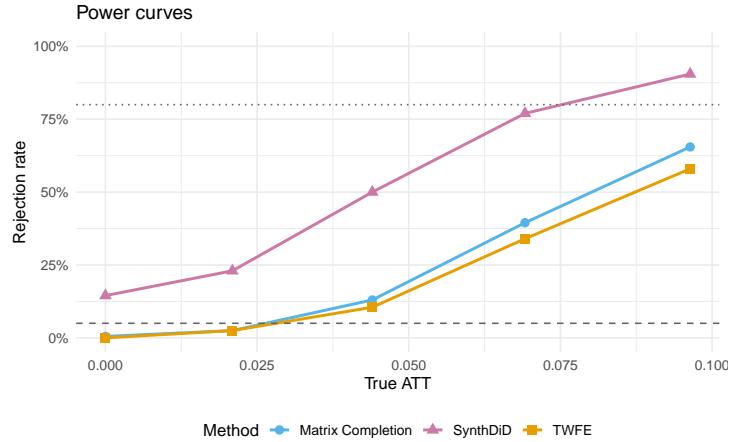


Figure 1: Rejection rates (power/size) across treatment strengths under the calibrated DGP.

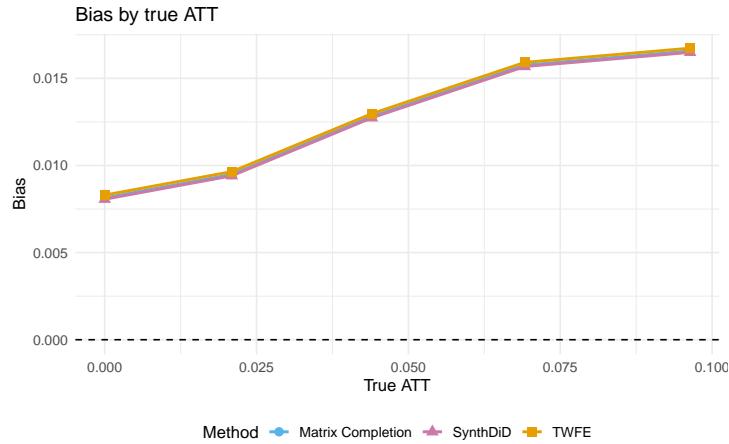


Figure 2: Bias as a function of treatment strength under the calibrated DGP.

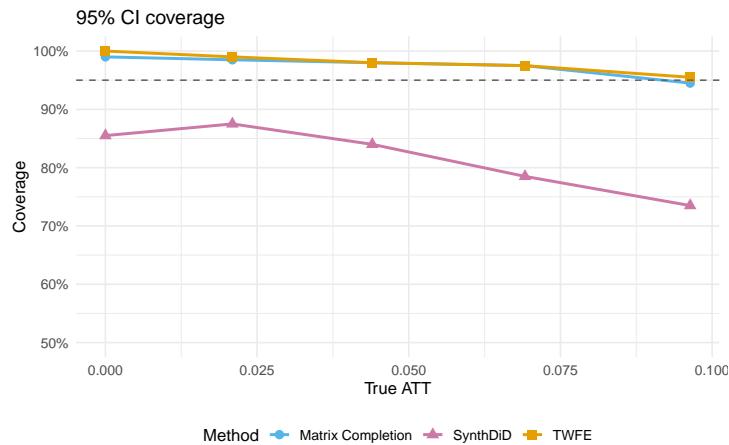


Figure 3: Coverage of nominal 95% intervals under the calibrated DGP.

The coverage results sharpen the distinction between methods. TWFE and Matrix Completion maintain coverage near or above the nominal 95% level across all effect sizes, with TWFE achieving 100% coverage at $\delta = 0$ and 95.5% at $\delta = 0.8$. Matrix Completion performs comparably, with coverage ranging from 99% to 94.5%. In contrast, SynthDiD severely undercovers throughout, starting at 85.5% when $\delta = 0$ and

declining to 73.5% at $\delta = 0.8$. This undercoverage means that SynthDiD's apparent power advantage is illusory: it rejects more often not because it detects true effects more reliably, but because its standard errors are too small. At the largest effect size ($\delta = 0.8$), the true ATT on the probability scale is approximately 0.096. Among methods with valid inference, Matrix Completion achieves 65.5% power while TWFE reaches 58%, neither attaining the conventional 80% threshold. This reflects a fundamental limitation of the design: with 56 spatial units, 35 time periods, and sparse binary outcomes, statistical power is constrained regardless of the estimation approach.

Table 1: Monte Carlo summary across all effect sizes (δ).

delta	method	true_att	mean_estimate	bias	rmse	coverage	rejection_rate
0.0	Matrix Completion	0.000	0.008	0.008	0.026	0.990	0.005
0.0	SynthDiD	0.000	0.008	0.008	0.036	0.855	0.145
0.0	TWFE	0.000	0.008	0.008	0.026	1.000	0.000
0.2	Matrix Completion	0.021	0.030	0.009	0.028	0.985	0.025
0.2	SynthDiD	0.021	0.030	0.009	0.035	0.875	0.230
0.2	TWFE	0.021	0.031	0.010	0.028	0.990	0.025
0.4	Matrix Completion	0.044	0.057	0.013	0.033	0.980	0.130
0.4	SynthDiD	0.044	0.057	0.013	0.040	0.840	0.500
0.4	TWFE	0.044	0.057	0.013	0.033	0.980	0.105
0.6	Matrix Completion	0.069	0.085	0.016	0.038	0.975	0.395
0.6	SynthDiD	0.069	0.085	0.016	0.046	0.785	0.770
0.6	TWFE	0.069	0.085	0.016	0.038	0.975	0.340
0.8	Matrix Completion	0.096	0.113	0.017	0.042	0.945	0.655
0.8	SynthDiD	0.096	0.113	0.016	0.050	0.735	0.905
0.8	TWFE	0.096	0.113	0.017	0.042	0.955	0.580

5.2 Scenario Analysis

The scenario analysis examines robustness to departures from baseline conditions. Table 2 summarizes performance across five scenarios. Under baseline conditions with realistic calibration, TWFE and Matrix Completion perform similarly, both achieving 99% coverage with power around 40-46%. SynthDiD shows higher apparent power (82%) but only 76% coverage, confirming that its rejection rate reflects inference failure rather than superior detection. When parallel trends hold ($s = 0$), all methods should perform well since the identifying assumption is satisfied. TWFE and MC both achieve 99% coverage. SynthDiD continues to uncover at 81%, suggesting that its inference problems are not driven by latent factor violations but rather by the sparse binary outcome structure.

Table 2: Scenario analysis (mean performance across simulations).

scenario	method	true_att	bias	rmse	coverage	power
Baseline (realistic)	Matrix Completion	0.070	0.018	0.037	0.99	0.46
Baseline (realistic)	SynthDiD	0.070	0.017	0.044	0.76	0.82
Baseline (realistic)	TWFE	0.070	0.018	0.037	0.99	0.40
Parallel trends world ($s = 0$)	Matrix Completion	0.081	0.012	0.031	0.99	0.64
Parallel trends world ($s = 0$)	SynthDiD	0.081	0.025	0.048	0.81	0.87
Parallel trends world ($s = 0$)	TWFE	0.081	0.012	0.031	0.99	0.55
Short-lived effect (decay)	Matrix Completion	0.022	0.016	0.030	0.99	0.04
Short-lived effect (decay)	SynthDiD	0.022	0.014	0.038	0.82	0.32

Table 2: Scenario analysis (mean performance across simulations).

scenario	method	true_att	bias	rmse	coverage	power
Short-lived effect (decay)	TWFE	0.022	0.016	0.030	1.00	0.02
Strong latent trends (larger s)	Matrix Completion	0.036	-0.002	0.034	0.91	0.08
Strong latent trends (larger s)	SynthDiD	0.036	-0.012	0.043	0.73	0.27
Strong latent trends (larger s)	TWFE	0.036	0.001	0.033	0.97	0.08
Under-reporting ($q < 1$)	Matrix Completion	0.070	0.007	0.028	1.00	0.28
Under-reporting ($q < 1$)	SynthDiD	0.070	-0.005	0.034	0.79	0.77
Under-reporting ($q < 1$)	TWFE	0.070	0.007	0.028	1.00	0.28

The strong latent trends scenario ($s = 2$) tests robustness to severe violations of parallel trends. Here, Matrix Completion achieves 91% coverage, a meaningful improvement over earlier implementations that used a narrower regularization grid. TWFE maintains 97% coverage, performing well despite the identifying assumption being violated. SynthDiD coverage drops to 73%. The true ATT in this scenario is smaller (0.036) because strong latent factors absorb more of the treatment-correlated variation, leaving less signal attributable to the policy. Under measurement error from under-reporting ($q = 0.7$, meaning 30% of events go unrecorded), Matrix Completion and TWFE both achieve perfect or near-perfect coverage (100% and 100%, respectively) with modest bias around 0.007. SynthDiD undercovers at 79%. The short-lived effect scenario, in which treatment impact decays exponentially with a half-life of 2 months, poses a challenge for all methods. The true ATT shrinks to 0.022 because the effect dissipates before the post-period ends. Power collapses to 2-4% for TWFE and MC, as the diluted signal cannot be reliably distinguished from noise. SynthDiD shows 32% power but with only 82% coverage, meaning approximately half of its rejections may be false positives.

6 Discussion

The simulation results yield three principal findings. First, SynthDiD produces invalid inference for sparse binary panel data in this setting. Across all scenarios, its confidence intervals are too narrow, leading to severe undercoverage (73-86%) and inflated rejection rates. SynthDiD constructs unit weights by matching pre-period trajectories, but with a 6.6% event rate, most cell-months have zero deaths, making pre-period means noisy and discrete rather than smooth. The placebo standard error method assumes that the distribution of placebo effects (computed by treating each control unit as if it were treated) approximates the true sampling distribution of the estimator. It might be that with sparse binary data, this assumption breaks down: placebo effects might concentrate near zero with occasional spikes, producing SE estimates that are systematically too small. Notably, SynthDiD undercovers even when parallel trends hold (81% coverage in the $s=0$ scenario), confirming that the problem is not about identifying assumptions but about inference calibration for rare-event data. The method was developed for continuous, well-behaved outcomes, and the implicit regularity conditions underlying its asymptotic theory appear to fail when outcomes are sparse and binary. Bootstrap standard errors, which directly estimate the sampling distribution through resampling, may restore valid coverage, though this remains to be tested. Second, Matrix Completion and TWFE perform comparably in this design. Point estimates are nearly identical across methods, and both maintain valid coverage. Matrix Completion shows modest advantages in specific scenarios: slightly higher power under baseline conditions and better robustness to under-reporting. However, these differences are small relative to sampling variability. The finding that MC tracks TWFE so closely suggests that the nuclear norm regularization is not selecting substantial low-rank structure beyond what two-way fixed effects already capture. Third, the study is fundamentally underpowered. Even at the largest effect size examined, power reaches only 58-66% for methods with valid inference. This limitation arises from the combination of a few spatial units ($N=56$), a moderate time span ($T=35$), and sparse binary outcomes (6.6% pre-period event rate). No estimation method can overcome insufficient statistical information.

7 Future Directions

Several extensions merit investigation. The SynthDiD inference failure warrants deeper examination. Alternative standard error methods, such as conformal inference or different bootstrap schemes, may restore valid coverage for binary outcomes. Additionally, the `synthdid` package was designed for continuous outcomes, and developing variants with appropriate link functions could improve performance. The similarity between MC and TWFE in this setting raises questions about when matrix completion provides meaningful robustness gains. Designs with stronger factor structure, more units, or continuous outcomes may show greater differentiation. Cross-validation diagnostics for rank selection could help identify when low-rank imputation is actively improving estimation. Finally, extending the analysis to heterogeneous treatment effects would increase realism. The current simulations assume constant δ across treated units, but effects may plausibly vary with distance from Libya or time since policy implementation. It would also be valuable to check whether the results change with more statistical information. However, these are outside of the scope of this project.