Dynamic Synthetic Controls vs. Panel-Aware Double Machine Learning for Geo-Level Marketing Impact Estimation

Sang Su Lee
Thumbtack, Inc.
San Francisco, CA, USA
psulee@thumbtack.com

Vineeth Loganathan
Thumbtack, Inc.
San Francisco, CA, USA
vloganathan@thumbtack.com

Vijay Raghavan
Thumbtack, Inc.
San Francisco, CA, USA
vraghavan@thumbtack.com

Abstract

Accurately quantifying geo-level marketing lift in two-sided marketplaces is challenging: the Synthetic Control Method (SCM) often exhibits high *power* yet systematically *under-estimates* effect size, while panel-style Double Machine Learning (DML) is seldom benchmarked against SCM. We build an open, fully documented simulator that mimics a *typical large-scale geo roll-out*: N_{unit} regional markets are tracked for T_{pre} weeks before launch and for a further T_{post}-week campaign window, allowing all key parameters to be varied by the user and probe both families under **five** stylised stress tests: (i) curved baseline trends, (ii) heterogeneous response lags, (iii) treated-biased shocks, (iv) a non-linear outcome link, and (v) a drifting control group trend.

Seven estimators are evaluated: three *standard* Augmented SCM (ASC) variants and four panel-DML flavours (TWFE, CRE/Mundlak, first-difference, and within-group). Across 100 replications per scenario, **ASC models consistently demonstrate severe bias and near-zero coverage** in challenging scenarios involving nonlinearities or external shocks. By contrast, **panel-DML variants dramatically reduce this bias and restore nominal 95%-CI coverage,** proving far more robust.

The results indicate that while ASC provides a simple baseline, it is *unreliable* in common, complex situations. We therefore propose a **'diagnose-first' framework** where practitioners first identify the primary business challenge (e.g., nonlinear trends, response lags) and then select the **specific DML model best suited for that scenario**, providing a more robust and reliable blueprint for analysing geo-experiments.

CCS Concepts

• **Computing methodologies** → **Causal reasoning and diagnostics**; *Machine learning approaches*; • **Information systems** → *Online advertising*; • **Applied computing** → *Electronic commerce*; *Marketing*.

Keywords

Causal Inference, Double Machine Learning, Synthetic Control, Two-Sided Marketplace, Panel Data, Marketing Analytics

ACM Reference Format:
Sang Su Lee, Vineeth Loganathan, and Vijay Raghavan. 2025. Dynamic Synthetic Controls vs. Panel-Aware Double Machine Learning for Geo-Level Marketing Impact Estimation. In *Proceedings of KDD 2025 Workshop – Causal Inference and Machine Learning in Practice.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

In two-sided marketplaces — such as ride-sharing platforms, home-sharing services, and e-commerce ecosystems — decision-makers are increasingly interested in understanding the causal impact of interventions on key business metrics. Whether it is a marketing campaign to boost user engagement or a subsidy to increase supply, the ability to reliably infer incremental gains ("lift") from such actions is crucial for optimal resource allocation. However, measuring incrementality in observational data is challenging due to confounding factors, temporal trends, and the complex dynamics inherent to marketplaces.

Causal machine learning has emerged as a promising toolset to tackle these challenges. Major platforms now integrate them into day-to-day decision making. Uber, for instance, built a spline-regularised learner that allocates marketing and incentive budgets across cities while explicitly accounting for causal lift [6]. Airbnb's data-science group reports using causal inference to infer guest-demand elasticities and to optimise marketplace outcomes at scale [2]. Together, these cases illustrate a broader trend: firms are combining traditional econometric ideas with flexible ML models to answer causal questions at industrial scale.

Despite this progress, significant methodological questions remain. Traditional econometric approaches like diff-in-diff (DiD) and Synthetic Control Methods (SCM) have been go-to solutions for causal inference on aggregate, panel-structured data. SCM, in particular, has gained popularity for evaluating interventions in a single or small set of treated units by constructing a weighted synthetic comparator [1]. SCM accounts for time-varying confounders by matching pretreatment trends, which is a major advantage over DiD.

Double Machine Learning (DML), on the other hand, has emerged as a flexible ML-based framework to estimate treatment effects with complex confounders [7]. While DML has primarily been applied to cross-sectional data, recent advances propose adaptations for panel data. These adaptations allow DML to leverage time-invariant and

KDD 2025 Workshop – Causal Inference and Machine Learning in Practice, Toronto, ON, Canada

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

time-varying covariates while addressing unobserved heterogeneity [8, 9]. Our paper explores how such panel-aware DML methods perform relative to SCM in a two-sided marketplace context.

We simulate weekly data for 200 geos over two years, designing realistic data-generating processes that incorporate nonlinear trends, heterogeneous treatment effects, biased external shocks, and nonlinear supply-demand matching. Across five scenarios, we evaluate how SCM and several panel-DML estimators compare in estimating average treatment effects. Our results provide practical guidance for analysts selecting among causal inference tools in complex real-world settings.

2 Related Work

Causal inference in panel settings has traditionally relied on diff-in-diff (DiD) and fixed effects models. Recent work has highlighted limitations of two-way fixed effects models under treatment heterogeneity or staggered adoption (e.g., Goodman-Bacon [10]; Sun and Abraham [13]). Alternative approaches such as doubly robust DiD [12] and synthetic difference-in-differences [3] improve estimation by combining outcome modeling and propensity weighting.

Synthetic Control Methods (SCM), introduced by Abadie et al. [1], construct weighted combinations of control units to approximate treated units' counterfactual outcomes. SCM is widely used in policy evaluation and marketing analytics, particularly when a small number of treated units receive an intervention. Augmented SCM [4] further combines outcome modeling with SCM weighting for improved robustness.

Double Machine Learning (DML), proposed by Chernozhukov et al. [7], estimates treatment effects using machine learning models for nuisance function estimation and orthogonalized causal regression. Recent extensions adapt DML to panel data using fixed effects, first-differencing, or correlated random effects transformations [8, 9]. These approaches enable flexible adjustment for observed and unobserved confounding.

In industry settings, causal–ML pipelines are now a standard part of large-scale marketplace experimentation. Team [14] open-sourced GeoLift, a full geo-experiment workflow that is widely used to quantify the offline incrementality of online ad campaigns. Hermle et al. [11] (LinkedIn Ads) propose an "asymmetric-budget-split" design and accompanying estimators that deliver unobtrusive but statistically valid lift measurement at nation-wide scale. Our study complements these operational systems by benchmarking Augmented SCM versus a family of panel-aware DML estimators under five stress-test scenarios, thereby mapping recent methodological advances to practical implementation choices that practitioners must make.

3 Methodology

This section formalises the causal estimand, summarises the two estimation paradigms under comparison, and specifies the evaluation metrics.

3.1 Estimand

Let Y_{it} denote weekly gross revenue for geo i in week t ($i = 1, \dots, N$, $t = 0, \dots, T-1$). $D_{it} \in \{0, 1\}$ is an **active-treatment indicator** that equals 1 only during the 12-week exposure window of treated geos; $G_i = \max_t D_{it}$ is the **ever-treated flag**. Potential outcomes are $Y_{it}(1)$ and $Y_{it}(0)$. We target the *average treatment effect on the treated geos (ATT)* over the post-period:

$$\text{ATT} = \frac{1}{N_T(\mathcal{T}_{\text{post}})} \sum_{i:G_i=1} \sum_{t\in\mathcal{T}_{\text{post}}} \left(Y_{it}(1) - Y_{it}(0)\right) \tag{1}$$

3.2 Synthetic Control Method

Augmented Synthetic Control (ASC).. Let Y_{it} be the observed outcome for unit i at time t, $\boldsymbol{Y}_t = [Y_{1t}, \dots, Y_{Nt}]^\top$, and $D_{it} \in \{0, 1\}$ the treatment indicator. Denote by $\mathcal{T}_0$ ($\mathcal{T}_1$) the pre- (post-) treatment periods and by $\mathcal{I}_{\text{tr}}$ ($\mathcal{I}_{\text{don}}$) the index sets of treated (donor) units. ASC seeks weights $\boldsymbol{w} \in \mathbb{R}^{|\mathcal{I}_{\text{don}}|}$ that minimise

$$\min_{\boldsymbol{w}\ge 0,\ \boldsymbol{1}^\top \boldsymbol{w}=1} \sum_{t\in\mathcal{T}_0} \Big(Y_{it} - \sum_{j\in\mathcal{I}_{\text{don}}} w_j\, Y_{jt}\Big)^2 \; + \; \lambda \|\boldsymbol{w}\|_2^2 \tag{2}$$

where λ is a ridge penalty controlling the weight dispersion. Given the fitted weights $\hat{\boldsymbol{w}}$, the counterfactual for each treated unit i in period $t > \min \mathcal{T}_1$ is

$$\widehat{Y}_{it}^{\text{SCM}} = \sum_{j\in\mathcal{I}_{\text{don}}} \hat{w}_j\, Y_{jt}, \tag{3}$$

$$\widehat{\text{ATT}}_t = \frac{1}{|\mathcal{I}_{\text{tr}}|} \sum_{i\in\mathcal{I}_{\text{tr}}} \left(Y_{it} - \widehat{Y}_{it}^{\text{SCM}}\right) \tag{4}$$

We implement Equation (2) with the augsynth package [5], using a ridge-only prognostic function and (unless otherwise noted) **no covariates** so that all methods are compared purely on their ability to exploit panel structure.

Three ASC specifications. To diagnose the performance of synthetic controls, we evaluate three distinct specifications that incrementally enrich the feature space used to determine the donor weights:

(a) **ASC-Y** ("*outcome-only*" baseline) uses *only* the past outcomes Y_{it} when re-estimating the ridge–ASC weights. It therefore assumes that pre-intervention trends are sufficient to screen off all latent differences between treated and donor markets.

(b) **ASC-DEM** ("*demographics*") augments the outcome history with *time-invariant covariates* so that the optimisation can explicitly balance structural demand fundamentals that are fixed within geo units but heterogeneous across space. This mirrors the static-covariate term in our CRE-DML estimator.

(c) **ASC-DEM-LAG** ("*demographics + lagged demand*") further adds *lagged demand proxies*, most notably one- and two-week lagged search volume for the focal product category. These pre-treatment lags capture high-frequency fluctuations in consumer interest that are *predictive of near-future sales*, giving the synthetic control a chance to adjust for fast-moving demand shocks before the treatment starts.

Putting the three specifications together lets us tease apart (i) how much of the **ASC's estimation error** stems from ignoring stable structural differences (Y vs. DEM) and (ii) how much is due to not tracking short-run demand momentum (DEM vs. DEM-LAG). These variants appear in all subsequent tables under the abbreviations given above.

3.3 Panel-Aware Double Machine Learning

Our implementation follows the orthogonal-residual recipe of Chernozhukov *et al.*[7] but adapts each step to panel structure and high-capacity learners written in XGBoost. **Alg.** 1 summarizes the workflow.

Algorithm 1 Cross-fitted panel DML with IPTW and cluster SEs

```
Panel transformation: Convert raw $(Y_{it}, D_{it}, X_{it})$ to $(Y^{\dagger}, D^{\dagger}, X^{\dagger})$ via one of: (i) TWFE dummies, (ii) geo-demeaned (Within), (iii) first difference (FD), (iv) CRE/Mundlak.
Stratified geo folds: Split rows by cluster-balanced cross-fold so every fold contains treated and control geos. For FD data ensure at least one $\Delta D \neq 0$ per fold.
Nuisance learning:
    • Outcome model: XGBRegressor
    • Propensity model: XGBClassifier
    Train on $\mathcal{I}^{(k)}_{\text{train}}$ and predict on $\mathcal{I}^{(k)}_{\text{test}}$ to get out-of-fold residuals $\hat{\varepsilon}_Y, \hat{\varepsilon}_D$.
IPTW stabilisation: Compute
    $w_i = \frac{D_i\,(1-\hat{p}_i)}{\hat{p}_i} + \frac{(1-D_i)\,\hat{p}_i}{1-\hat{p}_i}$
    and trim the top 5 % to avoid extreme weights.
Second-stage WLS: Regress $\hat{\varepsilon}_Y$ on $\hat{\varepsilon}_D$ with weights $w$.
Uncertainty: Report geo-cluster robust SEs; optional unit-bootstrap for coverage.
```

Why four variants? We include four panel transformations because each tackles a distinct threat to identification or efficiency.

(a) **TWFE–DML (Two Way Fixed Effects; dummy absorption).** Adds $N+T-2$ dummies so the learner need not model unit or time intercepts. Best when N and T are modest and results must be comparable to a classical two–way fixed–effects regression.

(b) **WG–DML (Within-Group; geo demean).** Subtracts geo means before learning, $x^{w}_{it} = x_{it} - \bar{x}_i$. Algebraically identical to TWFE in linear settings yet far sparser, so boosting models avoid multicollinearity and memory blow-up when N is large.

(c) **FD–DML (First Difference).** Uses ΔY_{it} and ΔD_{it}, wiping out every time-invariant component—including those correlated with X_{it}. The price is amplified measurement noise, but bias is minimal when unit-specific trends break strict exogeneity.

(d) **CRE–DML (Correlated Random Effect; Mundlak correction).** Augments X_{it} with unit means $\bar{X}_i$ and treatment means $\bar{D}_i$, absorbing correlation between covariates and unobserved heterogeneity (μ_i). Retains level information and often shows the best bias–variance trade-off when T is moderate.

These variants let us diagnose whether flexibility (CRE), noise attenuation (TWFE/WG), or non-stationarity robustness (FD) is most valuable under the scenarios in Section 4.

3.4 Implementation Details

All learners use identical hyper-parameters across scenarios; no tuning leakage occurs because hyper-parameters are fixed *ex-ante*. Cross-fitting splits by geo (not by time) to preserve within-unit serial correlation. SCM is implemented via the `augsynth` R package with ridge-augmented option.

4 Simulation Framework and Scenarios

4.1 Full Data-Generating Process

The simulation incorporates a rich set of features to model a realistic marketplace. Each geo is described by set of **time-invariant features**. The weekly dynamics are driven by **time-variant features** representing **demand features**, **supply features**, **competitor activity**, and **conversion** funnels that link demand to the final outcome.

Table 1: Key generator parameters (defaults shown right).

Description	Symbol	Default
# geographies	N_{unit}	200
# treated geos	N_{trt}	40
Pre-period length (weeks)	T_{pre}	52
Treatment window (weeks)	T_{post}	12
Annual baseline growth	μ_{growth}	1.20
Peak proportional lift	$\tau_{\max}$	0.23
Seasonality amplitude	A_{season}	0.23
Noise s.d.	σ_{ε}	0.10
Latent intercept s.d.	σ_{η}	0.23
Weeks per season cycle	T_{season}	52

Step–by–step generation. For geo $i = 1, \dots, N_{\text{unit}}$ and week $t = 1, \dots, T_{\text{pre}} + T_{\text{post}}$:

(1) **Baseline trend.**

$$\log Y^{\text{base}}_{it} = \alpha_i + \beta_i \frac{t}{T_{\text{pre}}} + \gamma_i \sin\!\left(\frac{2\pi\, t}{T_{\text{season}}}\right),$$

with $\mathbb{E}[\beta_i] = \log(\mu_{\text{growth}})/52$ so the average unit grows by μ_{growth} in one year.

(2) **Stochastic noise.** $\varepsilon_{it} \sim \mathcal{N}(0, \sigma^2_{\varepsilon})$ and $Y^0_{it} = \exp\bigl(\log Y^{\text{base}}_{it} + \varepsilon_{it}\bigr)$.

(3) **Treatment assignment.** Randomly pick N_{trt} units to form $\mathcal{T}$ and set $D_{it} = \mathbf{1}\{i \in \mathcal{T} \wedge t > T_{\text{pre}}\}$.

(4) **Lagged-impact curve.**

$$\tau_{it} = \tau_{\max}\, \mathbf{1}\{i \in \mathcal{T}\}\Bigl[\text{sigmoid}\bigl(\tfrac{t-T_{\text{pre}}}{3}\bigr) - \text{sigmoid}\bigl(\tfrac{t-T_{\text{pre}}-T_{\text{post}}}{3}\bigr)\Bigr].$$

(5) **Observed outcome.** $Y_{it} = Y^0_{it}\bigl(1 + \tau_{it}\bigr)$, and we export the full panel $\{(Y_{it}, D_{it}, X_{it})\}$.

All constants in parentheses are easy to tweak for alternative stress tests. Here η_i introduces unobserved, unit-specific shifts in the sigmoid-linked outcome model of Scenario S4, creating nonlinear heterogeneity not captured by observed covariates.

4.2 Stress-Test Scenarios

Stress-test scenarios. Table 2 shows the five perturbations we apply on top of the baseline DGP (Section 4.1). Each is crafted to

Table 2: Stress–test scenarios. Each adds one failure mode to the base DGP.

ID	Added complexity	Failure target
S1	Quadratic baseline trend: $\tau(t) = 1 + \alpha_1 t + \alpha_2 t^2$	ASC extrapolation bias
S2	Geo-specific response lags/decays	Dynamic mis-specification bias
S3	Shock $+B_{\text{shock}}$ only in treated units	Hidden confounding
S4	Sigmoid outcome link with η_i intercepts	Non-linear model mismatch
S5	Control group trend drift: $+\alpha_{drift}(t - T_{pre})$	Parallel trends violation

trigger a different, well-known failure mode of SCM- or DML-style estimators.

S1: Non-linear baseline trend. We endow every unit with a small *quadratic drift* = $\log Y_{it}^{\text{base}} \ + \ \beta_i^{(2)}\big(t/T_{\text{pre}}\big)^2$, where $\mathbb{E}[\beta_i^{(2)}] < 0$. Because Ridge-regularised ASC relies on (approximately) linear projection of donor trends, curvature causes systematic *under-extrapolation* and thus downward-biased ATT.

S2: Heterogeneous response lags. Treatment effects now follow geo-specific impact curves $\tau_{it} = \tau_{\max} \ * f_i\big(t - T_{\text{pre}}\big)$ with randomly drawn onset, peak and decay parameters. Static weights become mis-aligned with the moving effect window, leading to *dynamic-misspecification* bias. First-difference or de-meaned DML variants are expected to fare better.

S3: Shock larger in treated. We inject an exogenous post-period shock $\delta_{\text{shock}} \sim \mathcal{N}(0, \sigma^2_{\text{shock}})$ *only* for units in $\mathcal{T}$. ASC must separate the genuine treatment signal from this hidden confounder; DML can mitigate bias if predictive covariates proxy the shock mechanism.

S4: Non-linear outcome link. The linear growth term in the revenue equation is replaced by $\eta_i \ * \text{sigmoid}\big(\frac{t}{T_{\text{pre}}} - 0.5\big)$, inducing a strongly non-linear $X \rightarrow Y$ relationship. This stresses outcome models that assume linearity and tests whether flexible learners in panel-DML can adapt without mis-specification.

S5: Control group drift. We add a scenario where the baseline trend for control units is modified during the post-treatment period. Specifically, the log-baseline outcome for control units is augmented by an additional term, $\mathbb{I}(i \notin \mathcal{I}_{tr} \wedge t \geq T_{\text{pre}}) \cdot \alpha_{\text{drift}} \cdot (t - T_{\text{pre}})$, where $\alpha_{\text{drift}} > 0$. This directly violates the parallel trends assumption, creating a situation where the control group is no longer a valid baseline and is likely to cause severe underestimation of the true treatment effect.

Together these stress tests probe (i) trend-extrapolation, (ii) timing heterogeneity, (iii) hidden confounding, (iv) functional-form robustness, and (v) robustness to baseline drift—dimensions where SCM and DML are known to exhibit complementary strengths and weaknesses.

5 Experimental Results

Set-up. For each of the five stress scenarios, we generate $R = 100$ independent panels, each containing $N_{\text{unit}} = 200$ geo units of which $N_{\text{trt}} = 40$ are randomly assigned to treatment after a $T_{\text{pre}} = 52$-week pre-period, followed by a $T_{\text{post}} = 12$-week intervention window. We then estimate the average treatment effect (ATT) on every replicate with the competing estimators.

To evaluate performance, we focus on four key metrics: (i) **Absolute Bias**, the average difference between the estimated and true effect (accuracy); (ii) **Coverage**, the frequency with which the 95% confidence interval contains the true effect (reliability); (iii) **Power**, the frequency with which a true effect is correctly detected as statistically significant (sensitivity); and (iv) **Avg. CI Width**, the average width of the confidence interval (precision).

5.1 Performance in Scenario 1: Nonlinear Baseline Trend

This scenario tests model performance against a quadratic baseline trend, a common feature of markets with accelerating growth. The results, shown in Table 3, indicate that the linear assumptions of ASC models lead to severe under-estimation and high bias. In contrast, the flexible DML models adapt to the trend, with WG-DML providing the most accurate estimate (lowest bias) while maintaining high power.

Table 3: Performance in Scenario 1: Nonlinear Baseline Trend

Model	Abs. Bias	Coverage	Power	Avg. CI Width
ASC-Y	5020.24	0.01	0.19	4415.34
ASC-DEM	5037.64	0.01	0.15	4547.66
ASC-DEM-LAG	5046.62	0.01	0.15	4631.11
CRE-DML	4166.24	0.99	0.46	21385.11
TWFE-DML	2870.10	0.94	0.41	16366.87
FD-DML	2950.08	0.45	0.82	5527.35
WG-DML	1832.97	0.60	0.98	5785.75

5.2 Performance in Scenario 2: Geo-Specific Response Lags

This scenario introduces heterogeneous treatment lags, a challenge that resulted in critically low power for nearly all models, as shown in Table 4. Although WG-DML offered the lowest bias and highest precision among DMLs, its 7% power makes it unreliable for effect detection. In contrast, the structurally superior FD-DML model delivered the highest reliability (91% coverage). Despite its own low power (2%), its high reliability makes it the most trustworthy choice for this specific problem.

5.3 Performance in Scenario 3: Treated-Only Shock

Here, we test robustness to an external shock that positively impacts only the treated units. The results in Table 5 show that ASC models are unable to distinguish the shock from the treatment effect, leading to severe bias. DML models are more effective at

Table 4: Performance in Scenario 2: Geo-Specific Response Lags

Model	Abs. Bias	Coverage	Power	Avg. CI Width
ASC-Y	269.80	1.00	0.01	4444.94
ASC-DEM	228.27	1.00	0.00	4423.62
ASC-DEM-LAG	224.97	1.00	0.00	4434.81
CRE-DML	6372.29	0.94	0.06	426005.15
TWFE-DML	4584.69	0.90	0.10	416622.86
FD-DML	814.64	0.91	0.02	24187.34
WG-DML	744.03	0.67	0.07	12357.36

isolating the true lift, with WG-DML providing the best balance of low bias, high power, and precision.

Table 5: Performance in Scenario 3: Treated-Only Shock

Model	Abs. Bias	Coverage	Power	Avg. CI Width
ASC-Y	4973.75	0.01	0.19	4403.81
ASC-DEM	5000.56	0.01	0.15	4545.83
ASC-DEM-LAG	5010.39	0.01	0.14	4589.85
CRE-DML	4130.50	0.99	0.44	621071.17
TWFE-DML	2811.69	0.94	0.47	716234.08
FD-DML	2928.82	0.41	0.83	5340.88
WG-DML	1822.77	0.63	0.97	55522.08

5.4 Performance in Scenario 4: Nonlinear Outcome Link

This scenario tests a complex, nonlinear relationship between inputs and outcomes, such as ad saturation. As seen in Table 6, the DML models' flexibility is a clear advantage. WG-DML is again the top performer, delivering the lowest bias by a significant margin while maintaining high power.

Table 6: Performance in Scenario 4: Nonlinear Outcome Link

Model	Abs. Bias	Coverage	Power	Avg. CI Width
ASC-Y	2722.90	0.03	0.17	2678.95
ASC-DEM	2713.44	0.02	0.11	2784.08
ASC-DEM-LAG	2719.71	0.03	0.11	2816.92
CRE-DML	2749.44	1.00	0.33	514798.05
TWFE-DML	1894.94	0.95	0.30	211080.88
FD-DML	1593.13	0.54	0.76	93441.59
WG-DML	1046.38	0.69	0.95	83559.34

5.5 Performance in Scenario 5: Control Group Anomaly

This final scenario tests a critical failure of the parallel trends assumption, where the control group trend drifts. As shown in Table 7, nearly all models failed, producing highly biased and unreliable estimates. The CRE-DML model was the sole exception. Despite having the lowest precision (the widest confidence interval), it was the only model to deliver both high reliability (98% coverage) and the highest accuracy (lowest bias), proving it to be the most trustworthy choice for this severe challenge.

Table 7: Performance in Scenario 5: Control Group Anomaly

Model	Abs. Bias	Coverage	Power	Avg. CI Width
ASC-Y	12936.77	0.00	0.00	4562.70
ASC-DEM	13307.00	0.00	0.00	4519.29
ASC-DEM-LAG	13442.66	0.00	0.00	4603.92
CRE-DML	978.85	0.98	0.40	221408.58
TWFE-DML	2760.93	0.90	0.42	114984.97
FD-DML	2952.90	0.42	0.78	85453.30
WG-DML	3043.73	0.34	0.76	5931.74

5.6 Summary of Key Findings

The experimental results highlight the comparative strengths of DML and ASC models when faced with common, complex marketing dynamics. In our simulated environment, DML-based estimators consistently proved more robust than the standard ASC models, which showed significant bias and low reliability in scenarios with nonlinear trends, external shocks, or baseline drift.

However, these findings should be interpreted in the context of two important limitations of the simulation. First, our simulation does not include an expert-driven "geo pre-selection" step (the process of manually selecting a more comparable set of control group markets before analysis), which in practice can significantly improve the performance of ASC by creating a more comparable control group. Second, the simulation does not explicitly model strong unobserved confounders, though scenarios like heterogeneous lags (S2) and external shocks (S3, S5) implicitly test the models' robustness to such challenges.

Considering these points, our findings do not suggest abandoning SCM, but rather adopting a more sophisticated, context-aware approach. The clear outperformance of specific DML models in each stress test points to a powerful practical takeaway: a **"diagnose-first" strategy**. Analysts should first identify the most likely real-world challenge (e.g., nonlinear growth, potential for shocks) and then select the specific DML model best equipped to handle it. This approach, potentially used in parallel with a carefully curated SCM analysis, provides a robust and trustworthy framework for modern marketing analytics.

5.7 Practical takeaway

Our findings, which balance the simulation's results with its real-world limitations, lead to a set of actionable recommendations for practitioners. Instead of advocating for a wholesale replacement of one method with another, we recommend a more strategic, context-aware approach to geo-level impact analysis.

- **Acknowledge Simulation Limitations in Practice:** Recognize that the raw performance of SCM in a simulation may not reflect its true performance in a business setting where expert "geo pre-selection" is applied. A carefully curated SCM can serve as a powerful and intuitive baseline.

- **Adopt a "Diagnose-First" DML Strategy:** The primary takeaway is to move from a "one-size-fits-all" model to a diagnostic approach. Before analysis, identify the most likely business challenge your campaign faces by referencing the five scenarios:
  - For *nonlinear trends or external shocks*, start with **WG-DML**.
  - For *suspected response lags*, use **FD-DML**.
  - If you have concerns about *unreliable control group trends*, **CRE-DML** is the most robust choice.
- **Use Both Models for Robust Validation:** For high-stakes analyses, the most robust workflow is to run both a carefully curated SCM and the appropriate DML model in parallel. If their results align, confidence in the measured impact is high. If they diverge, it serves as a critical signal for a deeper investigation into potential unobserved confounders or model misspecifications.
- **Invest in Feature Engineering:** The success of DML models is highly dependent on a rich feature set. Continued investment in identifying and collecting data on potential business drivers is crucial for improving the accuracy and reliability of these advanced methods.

6 Conclusion

Our simulation provides a clear verdict on the comparative performance of Augmented Synthetic Controls and a suite of panel-aware Double Machine Learning models under several challenging, real-world scenarios. The results consistently demonstrated that while traditional ASC models provide an intuitive baseline, they are fragile. In the face of common complexities such as nonlinear trends, external shocks, and unreliable control groups, ASC models produced results with severe bias and near-zero coverage.

In contrast, DML models proved far more robust, consistently delivering more accurate and reliable estimates. Our key finding, however, is that no single DML model is universally superior. The optimal model is context-dependent: **WG-DML** excelled at handling nonlinearities and shocks, **FD-DML** was uniquely suited for heterogeneous response lags, and **CRE-DML** was the only effective tool against a drifting control group.

These findings should be contextualized by the realities of applied work. In practice, the performance of SCM is often enhanced by expert-driven **geo pre-selection**, and the success of DML is contingent upon rich **feature engineering** to make potential confounders observable. A purely model-vs-model comparison does not capture the critical role of this human-in-the-loop expertise.

Therefore, we conclude not with a declaration of a single "winner," but with a recommendation for a more strategic workflow. We advocate for a **"diagnose-first" approach**, where practitioners first identify the primary challenge of their specific geo-experiment and then select the DML model best equipped to handle it. This framework, which combines the analytical power of the right DML model with the crucial context provided by expert-driven data preparation, offers a more robust, reliable, and practical blueprint for modern marketing analytics. Future work should focus on empirically validating this hybrid, diagnostic workflow in a live production environment.

References

[1] Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. *Journal of the American Statistical Association*, 105(490):493–505, 2010. doi: 10.1080/01621459.2010.507311.

[2] Airbnb Data Science Team. Understanding guest preferences and optimizing marketplace outcomes: A causal inference approach. Technical report, Airbnb, dec 2024. https://airbnb.tech/wp-content/uploads/sites/19/2024/12/Understanding-Guest-Preferences-and-Optimizing-.pdf.

[3] Dmitry Arkhangelsky, Susan Athey, David A. Hirshberg, Guido W. Imbens, and Stefan Wager. Synthetic difference-in-differences. *American Economic Review*, 111(12):4088–4118, 2021. doi: 10.1257/aer.20181347.

[4] Eli Ben-Michael, Avi Feller, and Jesse Rothstein. The augmented synthetic control method. *Journal of the American Statistical Association*, 116(536):1789–1803, 2021. doi: 10.1080/01621459.2021.1934492.

[5] Eli Ben-Michael, Avi Feller, and Jesse Rothstein. *augsynth: The Augmented Synthetic Control Method*, 2021. URL https://CRAN.R-project.org/package=augsynth. R package version 0.2.0.

[6] Bobby Chen, Siyu Chen, Jason Dowlatabadi, Yu Xuan Hong, Vinayak Iyer, Uday Mantripragada, Rishabh Narang, Apoorv Pandey, Zijun Qin, Abrar Sheikh, Hongtao Sun, Jiaqi Sun, Matthew Walker, Kaichen Wei, Chen Xu, Jingnan Yang, Allen T. Zhang, and Guoqing Zhang. Practical marketplace optimization at uber using causally-informed machine learning. arXiv preprint: https://arxiv.org/abs/2407.19078, 2024.

[7] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018. doi: 10.1093/ectj/utx024.

[8] Damian Clarke and Nicola Polselli. Double machine learning for static panel models with fixed effects. *arXiv preprint arXiv:2312.08174*, 2024. doi: 10.48550/arXiv.2312.08174.

[9] David Fuhr and Dominik Papies. Rethinking double machine learning for panel data. *arXiv preprint arXiv:2409.01266*, 2024. doi: 10.48550/arXiv.2409.01266.

[10] Andrew Goodman-Bacon. Difference-in-differences with variation in treatment timing. *Journal of Econometrics*, 225(2):254–277, 2021. doi: 10.1016/j.jeconom.2020.12.001.

[11] Johannes Hermle, Giorgio Martini, and Wei Zhou. Valid and unobtrusive measurement of returns to advertising through asymmetric budget split. arXiv preprint: https://arxiv.org/abs/2207.00206, 2022.

[12] Pedro H. C. Sant'Anna and Jun Zhao. Doubly robust difference-in-differences estimators. *Journal of Econometrics*, 219(1):101–122, 2020. doi: 10.1016/j.jeconom.2019.08.006.

[13] Liyang Sun and Sarah Abraham. Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*, 225(2):175–199, 2021. doi: 10.1016/j.jeconom.2021.03.005.

[14] Meta Marketing Science Team. Geolift: Measuring incremental impact of advertising. GitHub: https://github.com/facebookincubator/GeoLift, 2022. Accessed: 2025-06-09.

Table 8: Frequently used symbols

Symbol	Meaning
Core Variables & Estimand	
Y_{it}	Outcome variable (weekly gross revenue) for unit i at time t
D_{it}	Treatment indicator for unit i at time t
ATT	Average Treatment effect on the Treated, the target estimand
Simulation Parameters	
N_{unit}	Total number of geo units in the simulation
T_{pre}	Length of the pre-treatment period (in weeks)
T_{post}	Length of the post-treatment (campaign) window
A_{season}	Amplitude of seasonal swing in baseline sales
Scenario-Specific Parameters	
B_{shock}	Additive shock applied to treated geos (S3)
η_i	Latent geo intercept for the non-linear term in Scenario 4 (S4)
α_{drift}	Coefficient for the time-trend drift in Scenario 5 (S5)
Model-Specific Parameters	
$\mathcal{I}_{tr}$	The index set of treated (treatment group) units
w_i	Stabilised IPTW weight (Alg. 1)