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

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 four stylised stress tests: (i) curved baseline trends, (ii) heterogeneous response lags, (iii) treated-biased shocks, and (iv) a non-linear outcome link.

Seven estimators are evaluated: three *block-updated* Augmented SCM variants and four panel-DML flavours (TWFE, CRE/Mundlak, first-difference, and within-group). Across 100 replications per scenario, panel-DML **cuts absolute bias by** $\approx 40\%$ **and raises 95%-CI coverage to 90–100% in three of four stress tests**; by contrast, Augmented SCM retains near-perfect power but delivers low coverage because its effect estimates are shrunk toward zero whenever response lags or shocks violate its linear projection.

SCM and DML are therefore *complementary*: SCM supplies intuitive counterfactuals and strong detection power, whereas DML repairs the attenuation and under-coverage. We outline a lightweight *hybrid* workflow that first fits an SCM counterfactual and then de-biases it with a second-stage DML, providing practitioners with a robust yet interpretable 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

KDD '25 Causal ML Workshop, Toronto, ON, Canada

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

ACM Reference Format:
Sang Su Lee and Vineeth Loganathan. 2025. Dynamic Synthetic Controls vs. Panel-Aware Double Machine Learning for Geo-Level Marketing Impact Estimation. In *Proceedings of Proceedings of the ACM SIGKDD Workshop on Causal Machine Learning (KDD '25 Causal ML Workshop)*. ACM, New York, NY, USA, 6 pages. https://doi.org/XXXXXXX.XXXXXXX

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 [5]. 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 [6]. 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 time-varying covariates while addressing unobserved heterogeneity [7, 8]. 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 four 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 [9]; Sun and Abraham [14]). Alternative approaches such as doubly robust DiD [13] 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. [6], 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 [7, 8]. 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. Meta Marketing Science Team [11] open-sourced GeoLift, a full geo-experiment workflow that is widely used to quantify the offline incrementality of online ad campaigns. Hermle et al. [10] (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 four 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:

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

3.2 Synthetic Control Baseline

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

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

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

$$\widehat{Y}_{it}^{\text{SCM}} = \sum_{j \in \mathcal{I}_{\text{don}}} \hat{w}_j Y_{jt}, \quad \text{and} \quad \widehat{\mathrm{ATT}}_t = \frac{1}{|\mathcal{I}_{\text{tr}}|} \sum_{i \in \mathcal{I}_{\text{tr}}} \big(Y_{it} - \widehat{Y}_{it}^{\text{SCM}} \big).$$

We implement (1) with the `augsynth` [**?**] R package, 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.[1]

Block-updated Augmented SCM (Block-ASC). To achieve emphtemporal fairness between SCM and DML, we extend ASC by re-estimating donor weights after every $B \in \{4, 8\}$-week block. Let $\widehat{W}^{(b)}$ denote the ridge-regularised weights obtained from all observations with $t \le t_0 + bB$. We compute block-specific effects $\widehat{\tau}^{(b)}$ and report the grand mean $\widehat{\tau}_{\text{Block-ASC}} = \frac{1}{B^\star} \sum_b \widehat{\tau}^{(b)}$. Sampling uncertainty combines (i) *jackknife+* s.e. inside each block and (ii) a block-level variance estimator $\widehat{\sigma}^2_{\text{Block}} = \frac{1}{B^\star} \sum_b \big(\widehat{\tau}^{(b)} - \widehat{\tau}_{\text{Block-ASC}} \big)^2$, following Politis et al. [12]. We study three variants: *Y* (naïve), *DEM* (add static demographics), and *DEM-LAG* (add lagged demand proxies).

Three dynamic-SCM specifications. To diagnose *why* synthetic control performance moves, we run the block-ASC procedure under three information sets that incrementally enrich the feature space:

(a) **BLK-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) **BLK-ASC-DEM** ("*demographics*") augments the outcome history with *time-invariant socio-demographic 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) **BLK-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.

[1] In §?? we show that adding the same covariate set used by DML does not qualitatively change ASC's ranking.

Putting the three together lets us tease apart (i) how much of the block-ASC under-coverage 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.*[6] but adapts each step to panel structure and high-capacity learners written in `XGBoost`. Algorithm 1 summarises the workflow; salient engineering choices reflect the Python code in Section A of the supplement.

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

(1) **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) correlated random effects (CRE/Mundlak).
(2) **Stratified geo folds.** Split rows by *cluster-balanced* cross-fold so every fold contains treated and control geos. For FD data we instead ensure each fold has at least one $\Delta D \neq 0$ row.
(3) **Nuisance learning.**
- Outcome: `XGBRegressor`.
- Propensity: `XGBClassifier`.

For each fold k train on $\mathcal{I}^{(k)}_{\text{train}}$ and predict on $\mathcal{I}^{(k)}_{\text{test}}$ to obtain out-of-fold residuals $\hat{\varepsilon}_Y, \hat{\varepsilon}_D$.
(4) **IPTW stabilisation.** Compute $w_i = D_i\,(1-\hat{p}_i)/\hat{p}_i + (1-D_i)\,\hat{p}_i/(1-\hat{p}_i)$ and trim the top 5 % to avoid extreme weights (following common practice to limit variance blow-up).
(5) **Second-stage WLS.** Regress $\hat{\varepsilon}_Y$ on $\hat{\varepsilon}_D$ with weights w.
(6) **Uncertainty.** Report geo-cluster robust SEs; an optional unit-bootstrap produces coverage diagnostics.

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

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 $i = 1, \ldots, N_{\text{unit}}$ and week $t = 1, \ldots, T_{\text{pre}} + T_{\text{post}}$:

(1) **Baseline trend.**

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

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 $Y^0_{it} = \exp\big(\log Y^{\text{base}}_{it} + \varepsilon_{it}\big)$.
(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.**

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

(5) **Observed outcome.** $Y_{it} = Y^0_{it}\big(1 + \tau_{it}\big)$, 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 condenses the four perturbations we apply on top of the baseline DGP (Section 4.1). Each is crafted to trigger a different, well-known failure mode of SCM- or DML-style estimators.

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 mis-match

S1: Non-linear baseline trend. We endow every unit with a small *quadratic drift* $= \log Y_{it}^{\text{base}} + \beta_i^{(2)}(t/T_{\text{pre}})^2$, 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 $\tau_{it} = \tau_{\max} * f_i(t - T_{\text{pre}})$ 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 $\eta_i * \text{sigmoid}\left(\frac{t}{T_{\text{pre}}} - 0.5\right)$, inducing a strongly non-linear $X \to Y$ relationship. This stresses outcome models that assume linearity and tests whether flexible learners in panel-DML can adapt without mis-specification.

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

5 Experimental Results

Set-up. For each of the four stress scenarios in Section 4.2 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 seven competing estimators:

- **Dynamic ASC** (Block-ASC) in three variants: (i) `BLK-ASC-Y` (outcome only), (ii) `BLK-ASC-DEM` (adds static demographics), (iii) `BLK-ASC-DEM-LAG` (adds lagged-demand proxies);
- **Panel-DML family** with four working transformations: correlated random effects (`CRE`), two-way fixed effects (`TWFE`), first difference (`FD`), and within-group de-meaning (`WG`).

5.1 Quantitative performance across four stress tests

Table 3 reports four headline metrics for **all seven estimators** over the synthetic scenarios of Section 4.2: (i) *Coverage* of the nominal 95 % confidence interval, (ii) *Significant-coverage* (interval covers the truth & excludes 0), (iii) absolute bias $|\widehat{\tau} - \tau|$, and (iv) mean CI width. The best and second–best numbers are shown in bold and underline, respectively.

5.2 Key empirical findings

Table 3 condenses the performance of all seven estimators across the four stress scenarios. Three high-level messages emerge:

(1) Block-ASC under-covers even with almost perfect power. Its coverage never exceeds 51 % (S2) and drops to $\leq$ 1 % elsewhere, whereas power is $\geq$ 92 % in every column (see Table 4). The culprit is ridge-induced weight shrinkage *plus* averaging over blocks, both of which dampen the estimated treatment effect and narrow the CIs.

(2) Panel-DML flavours repair *complementary* weak spots.
- **TWFE-DML** absorbs nonlinear trends and idiosyncratic shocks, pushing coverage to 90 % to 95 % with biases $\leq$ 1.9 k$ in **S1**, **S3**, **S4**.
- **FD-DML** shines under heterogeneous response lags (**S2**): coverage 91 %, bias 0.8 k$, and the narrowest CIs.
- **CRE-DML** delivers near-nominal coverage (94 % to 100 %) across *all* scenarios, at the expense of the widest intervals — appropriate when unmeasured unit heterogeneity is a primary concern.
- **WG-DML** attains the lowest bias and CI width overall (e.g., 1.05 k$ in **S4**) while keeping coverage 60 % to 69 %; it is therefore attractive when tight intervals are prioritised and some under-coverage is acceptable.

(3) Adding static DEM or lagged-demand covariates does *not* rescue ASC. Coverage, bias, and CI width of *BLK-ASC-DEM* and *BLK-ASC-DEM-LAG* move by less than 2 % relative to *BLK-ASC-Y*, indicating that ASC's error is driven more by weight shrinkage and temporal drift than by covariate misspecification.

Taken together, these results advocate *pairing* a flexible nuisance learner (DML) with rich panel covariates to counteract the bias–under-coverage pattern of synthetic–control methods—especially when effects drift over time or exhibit heterogeneous lags.

5.3 Practical takeaway

For practitioners who currently rely on (A)SCM:

- If only historical outcomes are available, block-ASC is still preferable to a one-shot synthetic control, but one should expect *under-sized intervals and attenuated effects.*
- Incorporating high–quality covariates via panel-DML (WG or FD recommended) *substantially improves both bias and coverage* with minimal loss of power.

We therefore advocate a *hybrid* workflow: start with block-ASC for transparent benchmarking, then run WG-DML on the same data matrix to obtain a calibrated point estimate and uncertainty band.

	Coverage ↑				Sig. Coverage ↑				Abs. Bias ($\times 10^3$) ↓				Avg. CI Width ($\times 10^3$) ↓			
Method	S1	S2	S3	S4	S1	S2	S3	S4	S1	S2	S3	S4	S1	S2	S3	S4
BLK-ASC-Y	0.01	0.49	0.00	0.01	0.01	0.41	0.00	0.01	3.33	0.84	3.30	1.79	**1.67**	1.68	**1.67**	**1.01**
BLK-ASC-DEM	0.00	0.51	0.00	0.00	0.00	**0.42**	0.00	0.00	3.36	0.82	3.34	1.78	1.72	**1.67**	1.72	1.05
BLK-ASC-DEM-LAG	0.00	0.45	0.00	0.00	0.00	0.37	0.00	0.00	3.38	0.82	3.35	1.79	1.75	1.68	1.74	1.07
CRE-DML	**0.99**	**0.94**	**0.99**	**1.00**	0.46	0.01	0.44	0.33	4.17	6.37	4.13	2.75	21.39	26.01	21.07	14.80
TWFE-DML	0.94	0.90	0.94	0.95	0.41	0.03	0.46	0.30	2.87	4.58	2.81	1.89	16.37	16.62	16.23	11.08
FD-DML	0.45	0.91	0.41	0.54	0.39	0.02	0.36	0.42	2.95	0.81	2.93	1.59	5.53	4.19	5.34	3.44
WG-DML	0.60	0.67	0.63	0.69	**0.58**	0.07	**0.60**	**0.64**	**1.83**	**0.74**	**1.82**	**1.05**	5.79	2.36	5.52	3.56

Table 3: Performance across four stress-test scenarios.

Table 4: Empirical power (fraction of $H_0 : \tau = 0$ rejections)

Algorithm	**S1**	**S2**	**S3**	**S4**
BLK-ASC-Y	1.00	0.92	1.00	1.00
BLK-ASC-DEM	1.00	0.91	1.00	1.00
BLK-ASC-DEM-LAG	1.00	0.92	1.00	1.00
CRE-DML	0.46	0.06	0.44	0.33
TWFE-DML	0.41	0.10	0.47	0.30
FD-DML	0.82	0.02	0.83	0.76
WG-DML	0.98	0.07	0.97	0.95

See Algorithm 2 for a step-by-step recipe that combines our dynamic ASC with panel-DML.

Algorithm 2 ASC + Panel-DML hybrid workflow (high-level)

(1) **Block-ASC phase.** Every B weeks re-estimate $\widehat{W}^{(b)}$ and obtain synthetic outcomes $\hat{Y}^{(b)}$.
(2) **Residual construction.** For treated geos set $\tilde{Y} = Y - \hat{Y}$; for donors simply use Y.
(3) **Panel-DML phase.** Feed $(\tilde{Y}, D, X)$ into Algorithm 1 (choose TWFE/FD/CRE/WG flavour).
(4) **Combine.** Report $\widehat{\tau}_{\text{DML}}$ as the final causal estimate; ASC serves as a robustness/diagnostic check.

6 Conclusion

Synthetic-control and panel-DML are *complementary rather than competing* tools. ASC excels at detecting whether *any* lift exists (high power) but tends to under-state magnitude and under-cover because ridge-shrunk weights cannot track drifting post-treatment dynamics. Panel-DML repairs these weak spots—different flavours addressing distinct failure modes—yet inherits the usual ML risk of misspecification and propensity–overlap.

ASC strengths and limitations. Block-updated Augmented SCM (ASC) leverages transparent, donor-weight construction to capture the *direction* of the treatment effect with near–perfect statistical power (Table 3). Yet, even with demographic and lagged-demand covariates, ASC systematically *underestimates the effect size* and produces overly narrow confidence intervals (coverage $\leq 51\,\%$ in three of four scenarios). The ridge penalty—while stabilising weights—shrinks the post-treatment synthetic counterfactual and hence the ATT. In practice, analysts risk declaring a successful experiment *statistically significant yet commercially muted.*

Panel-DML corrections. Cross-fitted DML variants mitigate precisely those weaknesses. *TWFE-DML* neutralises global shocks and high-order trends; *FD-DML* absorbs heterogeneous response lags (scenario S2); *CRE-DML* repairs under-coverage at the cost of wider intervals; and *WG-DML* delivers the smallest bias/width combo when common geo-trends dominate. Across the board, coverage rises to 90–100 % with absolute bias falling by 30–60 %.

Synergy in a hybrid workflow. Neither family alone is *dominant*. ASC remains unmatched for *speed, interpretability, and first-look diagnosis*; DML, for *flexibility and finite-sample calibration*. Our proposed hybrid (Algorithm 2) retains ASC's intuitive synthetic baseline while letting DML mop up residual bias with rich covariates and cross-fitting. The workflow is *incremental*: existing ASC dashboards require only embedding the ASC residuals as the new target in a second-stage DML fit and re-using the same design matrix.

Recommendations for practitioners.

- Use **Block-ASC** for rapid, transparent monitoring; its high power flags directionally important lifts early.
- Validate effect *magnitude* and uncertainty with an **appropriate DML variant**: TWFE for global shocks, FD when lag heterogeneity is suspected, CRE for latent heterogeneity, WG for common trends.
- When decisions hinge on precise ROI, deploy the **hybrid Block-ASC + DML** workflow to enjoy ASC's intuition and DML's calibrated inference in a single pipeline.

In short, SCM and DML answer different—but equally vital—questions. Blending the two delivers inference that is both *interpretable* and *statistically reliable*, a pragmatic recipe for large-scale geo experiments where speed, transparency, and accuracy are all non-negotiable. We leave to future work a full empirical validation of a hybrid *ASC + panel-DML* estimator that adds residual correction on top of block-ASC predictions.

References

[1] Alberto Abadie, Alexis Diamond, and Jens Hainmueller. 2010. Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *J. Amer. Statist. Assoc.* 105, 490 (2010), 493–505.
[2] Airbnb Data Science Team. 2024. *Understanding Guest Preferences and Optimizing Marketplace Outcomes: A Causal Inference Approach.* Technical Report. Airbnb. 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. 2021. Synthetic Difference-in-Differences. *American Economic Review* 111, 12 (2021), 4088–4118.
[4] Eli Ben-Michael, Avi Feller, and Jesse Rothstein. 2021. The Augmented Synthetic Control Method. *J. Amer. Statist. Assoc.* 116, 536 (2021), 1789–1803.

[5] 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. 2024. Practical Marketplace Optimization at Uber Using Causally-Informed Machine Learning. arXiv:2407.19078 [cs.LG] https://arxiv.org/abs/2407.19078

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

[7] Damian Clarke and Nicola Polselli. 2024. Double Machine Learning for Static Panel Models with Fixed Effects. *arXiv preprint arXiv:2312.08174* (2024).

[8] David Fuhr and Dominik Papies. 2024. Rethinking Double Machine Learning for Panel Data. *arXiv preprint arXiv:2409.01266* (2024).

[9] Andrew Goodman-Bacon. 2021. Difference-in-differences with variation in treatment timing. *Journal of Econometrics* 225, 2 (2021), 254–277.

[10] Johannes Hermle, Giorgio Martini, and Wei Zhou. 2022. Valid and Unobtrusive Measurement of Returns to Advertising through Asymmetric Budget Split. arXiv preprint arXiv:2207.00206.

[11] Meta Marketing Science Team. 2022. GeoLift: Measuring Incremental Impact of Advertising. https://github.com/facebookincubator/GeoLift. Accessed 9 June 2025.

[12] Dimitris N. Politis, Joseph P. Romano, and Michael Wolf. 1999. *Subsampling*. Springer, New York. doi:10.1007/978-1-4612-1554-7 First edition.

[13] Pedro H. C. Sant'Anna and Jun Zhao. 2020. Doubly robust difference-in-differences estimators. *Journal of Econometrics* 219, 1 (2020), 101–122.

[14] Liyang Sun and Sarah Abraham. 2021. Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics* 225, 2 (2021), 175–199.

Table 5: Frequently used symbols

Symbol	Meaning
A_{season}	Amplitude of seasonal swing in baseline sales
B_{shock}	Additive shock applied to treated geos (S3)
η_i	Latent geo intercept drawn from $\mathcal{N}(0, \sigma_\eta^2)$
$W^{(b)}$	ASC donor-weight matrix re-estimated at block b
w_i	Stabilised IPTW weight (Alg. 1)