KDD 2025 Workshop - Causal Inference and Machine Learning in Practice

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

Sang Su (Paul) Lee, Vineeth Loganathan and Vijay Raghavan

Last updated: 2025/07/30

Today’s agenda

01 Introduction

02 Approach

03 Simulation

04 Experimentation Result

05 Conclusion

Introduction

The Challenge: Why Measuring True Lift is Difficult

- **Business Need:** In two-sided marketplaces like ours—from ride-sharing to e-commerce—understanding the **true causal impact (lift)** of marketing interventions is crucial for optimal resource allocation.
- Measuring the precise impact of our marketing efforts is challenging for two main reasons:
    - **External Noise:** Our campaigns don't happen in a vacuum. External events, such as local holidays, competitor actions, or even weather, can affect our business outcomes. It's difficult to separate the impact of our marketing from this "noise."
    - **Complex Market Dynamics:** Each geographic market has its own unique trends and a delicate balance between supply and demand. These complex, shifting dynamics make it hard to establish a stable baseline for measuring campaign performance.

Limitations of Current Methods

- **Traditional Tools:** Econometric methods like Difference-in-Differences (DiD) and Synthetic Control Methods (SCM) have been the standard solutions for panel data.
- **SCM's Popularity:** SCM has gained significant traction because it can control for time-varying confounders by creating a "synthetic twin" of the treated group based on pre-treatment trends.
- **The Unanswered Question:** While ML-based frameworks like **Double Machine Learning (DML)** have emerged, they are rarely benchmarked head-to-head against SCM in a realistic marketplace context. We know they are powerful, but how do they *really* perform relative to SCM when things get complicated?

Comparing the Tools: "Digital Twin" vs. "Flexible Learner"

To navigate these challenges, we rely on sophisticated measurement methods. This analysis compares two leading approaches:

1. **The Established Method: The "Digital Twin" (Synthetic Control Method - SCM)** This has been a standard approach. It works by creating a "synthetic twin" of the market where we ran a campaign. This twin is a weighted average of other, similar markets that were not targeted by the campaign. By comparing the performance of the real market to its twin, we can estimate the campaign's lift. Its key strength is that it's easy to understand and interpret.
2. **The Advanced Method: The "Flexible Learner" (Double Machine Learning - DML)** This is a newer, more powerful framework that uses machine learning. Its main advantage is its flexibility; it can uncover complex patterns and relationships in our data without needing rigid assumptions, making it very effective at handling the "noise" and complexity we see in the real world.

Our Goal: Finding the Most Reliable Tool

1. The Unanswered Question

- Between the "Digital Twin" (SCM) and the newer "Flexible Learner" (DML), which method is best for which situation?

2. Our Approach: A Head-to-Head Test

- We built a realistic simulator to test them side-by-side.
- We will compare their performance directly using "stress tests" that mimic real-world challenges.

3. The Expected Outcome: A Clear Recommendation

- To definitively identify the most reliable tool for measuring the marketing ROI.

Approach

A Head-to-Head Comparison of Measurement Tools

SCM (Synthetic Control Method)

How it Works: Constructs a "synthetic" twin of the treated group from a weighted average of untreated donor groups.

Key Strength: Highly interpretable. The counterfactual is a tangible combination of real-world geos.

DML (Double ML)

How it Works: Uses two ML models to separately partial out the effects of confounders on the outcome and the treatment. It then measures the relationship between the residuals.

Key Strength: Highly flexible. Excels at handling complex relationships and a large number of covariates without strong parametric assumptions.

The "Digital Twin" (Synthetic Control Method - SCM)

This method creates a 'twin' of a market from a combination of similar, untreated markets. We tested three ways to build this twin:

- **The Basic Twin (Outcome-Only):** Builds the twin using only past sales data, assuming past trends are enough to predict the future using AugSynth.
- **The Structurally-Aware Twin (+ Demographics):** Also considers stable market characteristics like population and income to ensure a fair comparison.
- **The Trend-Aware Twin (+ Demand Lags):** Adds recent, short-term data like search trends to account for fast-moving changes in consumer interest.

The "Flexible Learner" (Double Machine Learning - DML)

This method uses machine learning to flexibly model complex relationships and filter out "noise." We tested four distinct approaches for time-series panel data setting:

- **Controlling for Market & Time Effects (TWFE-DML):** A classic approach that isolates the campaign's impact from unique, stable characteristics of each market and from shocks that affect all markets at the same time (e.g., a national holiday).
- **Isolating Each Market's Unique Behavior (WG-DML):** A more efficient version of the above, making it a practical choice for large-scale tests across many markets.
- **Focusing on Weekly Growth (FD-DML):** Looks at week-over-week *changes* rather than absolute sales numbers. This is best when markets are on different growth trajectories.
- **A Flexible Hybrid (CRE-DML):** A balanced approach that controls for hidden factors that are related to the market characteristics we can already see.

TWFE: Two-Way Fixed Effect
WG: Within Group
FD: First Difference
CRE: Correlated Random Effect

Simulation

Simulation Setting

To find the most reliable tool, we didn't just compare the 7 models in theory. We built a realistic simulator to see how each model performs when faced with tough, real-world marketing challenges.

We pitted our 7 competing models against 5 challenging "stress-test" scenarios:

7 Competing Models

3 SCM Variants
The Basic Twin (Outcome-Only) (ASC-Y)
The Basic Twin + Demographic (ASC-DEM)
The Basic Twin + Demographics + Demand Lags (ASC-DEM-LAG)

4 DML Variants
Controlling for Market & Time Effects (TWFE-DML)
Isolating Each Market's Unique Behavior (WG-DML)
Focusing on Weekly Growth (FD-DML)
A Flexible Hybrid (CRE-DML)

The 5 Real-World “Stress Tests”

S1: Nonlinear Baseline Trend
S2: Geo-Specific Response Lags
S3: Treated-Only Shock
S4: Nonlinear Outcome Link
S5: Control Group Anomaly

Simulation Setting

To test our models, we created a realistic, virtual environment that mimics our business. The key parameters of this simulated world are as follows:

Parameter	Description	Value
Experiment Scale	Total number of geographic markets	210
	Number of markets receiving the treatment	40
	Number of times the simulation was run for each scenario	100
Timeline	Total duration of data for each market	104 weeks (2 years)
	Pre-treatment data period	52 weeks (1 year)
	Week the treatment begins	Week 53
	Duration of the marketing campaign	12 weeks
Treatment Effect	The true "lift" from the campaign	15% to 35% (randomly assigned)

Simulation Features

The following features were included in the simulation to model the real-world complexities of our markets.

Time-Invariant Features (Fixed characteristics of each market)
Geo based static features to reflect the demographics

Time-Variant Features (Metrics that change weekly)
Demand Metrics, Supply Metrics:, Conversion Metrics (The ratio of demand to conversion), and Competitor Metrics

Key Outcomes & Identifiers
These are the main variables we measure, along with identifiers.
Core Outcomes: Revenue
Identifiers: Geo name, week id, and flags indicating if a Geo is in the test group and if the treatment is active in a given week.

Simulation Scenario Definitions and Real-World Marketing Analogs

S1: Nonlinear Baseline Trend
The adoption of smart home security systems (like Ring or Simplisafe) in a new neighborhood often shows accelerating early adoption curves. Linear models underestimate lift in such fast-growth markets.

S2: Geo-Specific Response Lags
In marketing campaigns, some geos react later due to postal delays or local behavior patterns (e.g., weekly booking cycles). Lift may not appear immediately after launch.

S3: Treated-Only Shock
During a test, treated regions might experience a local holiday, weather anomaly, or unrelated PR boost, artificially inflating post-treatment outcomes.

S4: Nonlinear Outcome Link
In advertising, more search may not linearly translate into revenue due to click-through-rate saturation or conversion plateaus.

S5: Control Group Anomaly
An external event, like a competitor leaving, causes sales to grow in the control group alone. This drift creates a false baseline, leading to a severe underestimation of the campaign's real impact.

Experimentation Result

Experimentation Results

Our experiment shows there is no single "best" model. The right choice depends on the specific business challenge you face.

Below is a summary of which model performed best in each of the five real-world stress-test scenarios.

Scenario	Key Challenge	Recommended Model	Why It Won
S1	Nonlinear Growth	WG-DML	Most accurately predicted the accelerating growth curve.
S2	Delayed Responses	FD-DML	Best at capturing time-shifted effects by focusing on weekly changes.
S3	Isolating Shocks	WG-DML	Best at filtering out external noise to isolate the true campaign effect.
S4	Complex Outcomes	WG-DML	Most accurately modeled the complex, nonlinear link between spend and revenue.
S5	Unreliable Control Group	CRE-DML	The only model that could reliably handle a biased or drifting control group.

A Guide to Our Key Metrics

1. Absolute Bias (Accuracy)

- What it answers: "On average, how far off is the model's estimate from the true value?"
- What's good: Lower is better. A value of 0 means perfect accuracy.

2. Coverage (Reliability)

- What it answers: "How often does the model's 95% confidence interval actually contain the true value?"
- What's good: Closer to 95% is better. This shows the model is "honest" about its own uncertainty.

3. Power (Sensitivity)

- What it answers: "When there is a real effect, how often does our model successfully detect it?"
- What's good: Higher is better. High power means we won't miss the impact of a successful campaign.

4. Avg. CI (Confidence Interval) Width (Precision)

- What it answers: "How precise is the estimate? Is the range of uncertainty narrow or wide?"
- What's good: Narrower is more useful, but only if the Reliability (Coverage) is also high. A precise but wrong answer is not helpful.

Scenario 1: Nonlinear Baseline Trend

In a high-growth environment, SCM models fail while flexible DML models win.

Key Finding:
In a market with a nonlinear trend (i.e., accelerating growth), traditional ASC models significantly underestimated the true effect. This resulted in very high bias and unreliable results (low coverage).

In contrast,

DML models successfully adapted to this growth curve thanks to their inherent flexibility, providing more accurate estimates.

Model	Abs. Bias (Accuracy)	Coverage (Reliability)	Power (Sensitivity)	Avg. CI Width (Precision)
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	4166.24	0.99	0.46	21385.11
TWFE	2870.1	0.94	0.41	16366.87
FD	2950.08	0.45	0.82	5527.35
WG	1832.97	0.6	0.98	5785.75

Scenario 2: Geo-Specific Response Lags

When customer response times vary, specialized models are required to avoid misinterpreting the results.

Key Finding:

This scenario tested a common marketing challenge where customers in different regions react to a campaign with different time lags. This proved challenging for most models, and **nearly all exhibited critically low power**.

The results reveal a key trade-off between the top DML contenders:

- **WG-DML** delivered the highest accuracy (lowest bias) and precision, but its very low power (7%) makes it unreliable for consistently detecting a real effect.
- In contrast, the **FD-DML** model is structurally the most appropriate for this problem. While its power is also extremely low (2%), it successfully handles the varied lags and delivers the highest reliability (91% coverage), making it the most trustworthy choice.

Model	Abs. Bias (Accuracy)	Coverage (Reliability)	Power (Sensitivity)	Avg. CI Width (Precision)
ASC-Y	269.8	1	0.01	4444.94
ASC-DEM	228.27	1	0	4423.62
ASC-DEM-LAG	224.97	1	0	4434.81
CRE	6372.29	0.94	0.06	426005.15
TWFE	4584.69	0.9	0.1	416622.86
FD	814.64	0.91	0.02	24187.34
WG	744.03	0.67	0.07	12357.36

Scenario 3: Treated-Only Shock

When an external shock hits only the test group, flexible DML models successfully separate the true campaign lift from the noise.

Key Finding:
This scenario tested a situation where an external event (like a local holiday or PR boost) positively impacted **only the markets where the campaign was active.**

The ASC models failed this test, as they could not distinguish the external shock from the campaign's true effect. This resulted in very high and misleading bias.

In contrast, the DML models were much more effective.

WG-DML is able to filter out the shock and provide a more accurate and reliable estimate of the campaign's actual contribution.

Model	Abs. Bias (Accuracy)	Coverage (Reliability)	Power (Sensitivity)	Avg. CI Width (Precision)
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	4130.5	0.99	0.44	621071.17
TWFE	2811.69	0.94	0.47	716234.08
FD	2928.82	0.41	0.83	5340.88
WG	1822.77	0.63	0.97	55522.08

Scenario 4: Nonlinear Outcome Link

When the link between marketing and revenue isn't a straight line, DML models show their strength.

Key Finding:
This scenario tested a case where marketing efforts don't produce results in a simple, linear way (e.g., due to ad saturation or conversion plateaus). As in other scenarios, the ASC models could not adapt to this complexity, resulting in high bias and low reliability.

The DML models, however, were built for this challenge. Their flexible machine-learning core allowed them to model the complex outcome link successfully. The **WG-DML** model was the clear winner, delivering the lowest bias by a significant margin while maintaining very high power.

Model	Abs. Bias (Accuracy)	Coverage (Reliability)	Power (Sensitivity)	Avg. CI Width (Precision)
ASC-Y	2722.9	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	2749.44	1	0.33	514798.05
TWFE	1894.94	0.95	0.3	211080.88
FD	1593.13	0.54	0.76	93441.59
WG	1046.38	0.69	0.95	83559.34

Scenario 5: Control Group Anomaly

When the control group is unreliable, only CRE-DML provided a trustworthy result.

Key Finding:
This scenario tested a critical and dangerous situation where the control group's sales trend changed on its own, making it a "false baseline" for comparison.

The results revealed a clear hierarchy of performance. The ASC models, along with FD-DML and WG-DML, failed by producing highly unreliable estimates with low coverage. While TWFE-DML was reliable (90% coverage), its accuracy was poor, with a bias nearly three times higher than the winner's.

The **CRE-DML** model was the sole standout. 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), making it the most trustworthy choice for this challenging scenario.

Model	Abs. Bias (Accuracy)	Coverage (Reliability)	Power (Sensitivity)	Avg. CI Width (Precision)
ASC-Y	12936.77	0	0	4562.7
ASC-DEM	13307	0	0	4519.29
ASC-DEM-LAG	13442.66	0	0	4603.92
CRE	978.85	0.98	0.4	221408.58
TWFE	2760.93	0.9	0.42	114984.97
FD	2952.9	0.42	0.78	85453.3
WG	3043.73	0.34	0.76	5931.74

Conclusion & Next Steps

From Experimental Results to Real-World Strategy

Our simulation results were clear: in a raw, uncontrolled environment, **DML models consistently and decisively outperformed traditional SCM models**. This confirms that DML is a more powerful and flexible technology.

However, our goal isn't just to find the best model in a simulation, but the best process for our company. This requires incorporating two key pieces of expert knowledge that were not part of the experiment:

SCM is not basic: Industry uses a rigorous geo-unit pre-selection process that makes the SCM baseline much stronger than the one tested.

DML is not magic: Its success depends on the expert feature selection to make sure all critical business drivers are visible to the model.

Because both models' real-world success depends on our expert processes, the best strategy is not to replace one with the other, but to build a workflow that leverages both.

Final Recommendation: A Practical Guide to Model Selection

Our experiments show that the best model depends entirely on the specific marketing challenge. Instead of a one-size-fits-all approach, we recommend using this guide to select the most appropriate and reliable tool for your analysis.

When to Use SCM (Curated with Geo-Selection)

SCM is a good choice in stable, predictable situations where you can confidently identify a very similar control group. Use SCM when:

- The market is mature with stable, linear growth trends (The opposite of S1 & S4).
- You expect customer reactions to be quick and consistent across regions (The opposite of S2).
- There are no major, market-specific shocks expected during the campaign (The opposite of S3 & S5).

In these cases, a well-curated SCM provides a transparent and intuitive baseline.

When to Use DML

DML is the necessary choice for complex, dynamic, and uncertain situations. You must use a DML model when:

- The market is new or experiencing accelerating, nonlinear growth (S1 & S4) -> Use WG-DML.
- You expect customer responses to be delayed or varied by region (S2) -> Use FD-DML.
- A shock event impacts only the test group (S3) -> Use WG-DML.
- You suspect the control group may not be a perfect parallel to the test group (S5) -> Use CRE-DML.

In these common real-world cases, DML is essential for an accurate and reliable measurement.

Next Step: Experiment Velocity Improvement using DML/SCM

Which model requires less post-test data?

This approach answers the question: "If we need to detect a 5% lift, which model will allow us to end the campaign and get a reliable answer the fastest?"

1. Fix the Effect Size: We set the pre-test period to 52 weeks and fix our target MDE.

2. Iterate on the Campaign Duration: We run the simulation multiple times, progressively shortening the treatment_window (e.g., 12 weeks, then 10, then 8, etc.).

3. Measure Power for Each Model: For each duration, we calculate the power of each model.

4. Determine the Winner: The model that achieves our target power (e.g., 80% power) with the shortest campaign duration is the winner. This model provides the fastest path to a conclusive result.

Next Step: Placebo Test Simulation

How can we be sure our model isn't finding effects that aren't there?

This approach answers the question: "If there were no true marketing effect, how often would our model incorrectly report a significant lift?"

1. **Run the Simulation with Zero Effect:** We set the true treatment effect size to zero and run the full simulation.

2. **Analyze with Our Recommended Model:** We analyze the resulting data using our recommended DML model for that scenario.

3. **Measure the False Positive Rate:** We count how often the model incorrectly finds a statistically significant effect.

4. **Check for Reliability:** A reliable model should have a false positive rate at or below the expected level (e.g., around 5% for a 95% confidence level). This test builds confidence that the effects we find are real.

Thank you!