The Reitan Store Testing

A Data-Driven Framework for Store Testing at Reitan

Zahra Shamlou - Feb 2026

To move from intuition-based decisions to evidence-driven growth, Reitan utilizes the Synthetic Control Method (SCM) for store testing. Unlike traditional before-after comparisons that are often biased by weather or seasonal trends, this framework isolates the True Lift of any intervention by creating a control twin for any test. By focusing on rigorous preparation, strategic operational blackouts, and the elimination of statistical noise, we ensure that every rollout, from new aromas to layout changes, is a proven contributor to our bottom line.

1. The Mission: Why We Test?

In the world of convenience retail, intuition is our starting point, but evidence is our decision-maker. This framework ensures that every change we make — whether it's a new aroma, a shelf layout change, an assortment change, or a color update — actually adds measurable value to the bottom line without compromising customer satisfaction or overall business performance. At its core, this framework is designed to **isolate the real effect** of a change by filtering out other noises such as seasonal trends, weather fluctuations, or general market shifts.

2. Method Description: How We Isolate the Real Effect

To extract the real effect, we use a statistical method that creates a **Counterfactual**. This is a model of what would have happened in the store if we had done nothing. By subtracting this Expected performance from the Actual performance, we are left with the True Lift, the pure impact of our intervention.

Because no two Pressbyrån or 7-Eleven stores, and no two periods of time are identical, we cannot simply compare one store to another or use historical records or before-after analysis. Traditional comparisons are often unfair and biased, they fail to extract the pure intervention effect and isolate it from other unobserved factors, trends, or noises. To address these issues, we primarily use the **Synthetic Control Method (SCM)**.

2.1. What is Synthetic Control Method (SCM)

The Synthetic Control Method is a data-driven approach to estimate the causal effects of interventions when a randomized experiment is not feasible.[1] The core idea is to construct a **synthetic version** of the treated unit as a **weighted combination** of untreated stores and use them as the control unit.

This synthetic unit mimics the treated stores' pre-treatment behavior, so differences after treatment can be argued as the result of the intervention. Essentially, it is a what-if scenario, a way to estimate what would have happened to the treated stores if the intervention didn't exist.

2.2. Why This Works Well for Store Testing at Reitan

- **Captures Trends:** The synthetic control unit captures unobserved trends and patterns in the treated unit before the intervention, which helps account for other factors that could influence outcomes.
- **High Reliability with Small Samples:** SCM is specifically designed to provide high-confidence results from small size samples and a few number of tested stores. This makes the testing process highly cost-effective.
- **Transparency:** The weights show exactly which control stores contribute most to the synthetic version, making the results easy to verify and explain to stakeholders. This contrasts with machine learning methods, such as Random Forests[2], where the contribution of individual inputs is often unclear or difficult to interpret.
- **Light Assumptions:** It doesn't rely on strong assumptions about the relationship between the treatment and outcome, unlike some other methods, which makes this method highly practical and useful for real-world retail.

3. Step-by-Step Execution Guide

To ensure our results are as reliable and actionable as possible, the testing framework is structured into three main phases: **Preparation, Execution, and Evaluation.**

---

[1] The Synthetic Control Method differs from A/B testing, which is used when a randomized experiment is possible. A/B testing is common in digital marketing, SaaS, e-commerce, online platforms, or highly controlled environments where randomization can be implemented.

[2] Random Forests is a machine learning method that can predict outcomes by combining multiple decision trees. While powerful for prediction, it is less transparent than Synthetic Control, as the contribution of individual stores or features is difficult to interpret.

3.1 Preparation

The goal of preparation is to ensure the test is set up to give a clear and reliable answer, making the process scalable and consistent across the company. Good preparation saves time during analysis and ensures we are measuring what truly matters, while making the insights usable for future similar situations.

Economic Viability

It's nice to perform a pre-test cost-benefit analysis before running the test. This confirms that the expected True Lift significantly outweighs the operational investment required to execute the test.

Design the Hypothesis

Define Intervention: Start by being specific. What exactly are we changing, and where? Is it an assortment change with a focus on specific cold drinks, or a shelf layout change with more available chocolate options?

Define Context: Where and when is this happening? e.g., Stockholm's stores located in transportation hubs during morning hours, or stores located specifically in airports or hospitals during lunch time.

Define the Metrics:

- Primary Outcome: What is your main outcome you wanna see the results on? e.g., Sales, Average Transaction Value (ATV), Items per Basket, Growth, or Net Profit.
- **Guardrail Metrics:** What must we keep an eye on to ensure no negative side effects? e.g., Service speed, customer trust, or customer satisfaction.
- **Behavioral Channel:** This is the channel where we expect the customer to act differently through. e.g., increasing conversion rate, or attracting new customers.

Define the Expected Effect and Affected Product: Specify the expected direction and size of the impact on the affected product or category (e.g., +10% category sales).

Hypothesis: If we [**Intervention**] in [**Context**], then [**Behavioral Channel**], leading to an [**Expected Effect**] in [**Primary Outcome**], without affecting [**Guardrail**].

Hypothesis Example: If we use cinnamon bun scents in transportation-hub Pressbyrån stores during the afternoon rush, then we will see a higher conversion rate of people entering the store,

leading to a 10% increase in pastry sales, without making the store smell too intense or hurting our customers' trust in our freshly baked quality.

Sample Size & Statistical Power

Choosing the number of test stores **N** is a strategic balance. We must weigh the cost of the experiment against the reliability of the results. We don't simply pick a number of stores, we calculate the optimal required number of test stores by analyzing historical natural variation within our stores. This ensures our results are statistically significant rather than coincidental. To do this, we balance two critical thresholds:

Significance Level (α) : Confidence that the effect we found is real (Avoiding False Positives)

Statistical Power (1 - β) : Ensuring the test is sensitive enough to detect the intervention's impact amidst daily retail noise. (Reduces the risk of a False Negative)

Selecting the number of test stores is about finding a balance between execution costs and reliable results. Fewer number of tested stores is cheaper and easy to manage, but it's risky, local noises can easily hide the real results, causing us to miss out on a great idea(high risk of False Negatives). Conversely, a large sample size is much better at filtering out that noise and proving a lift is real, but it is expensive, harder for staff to execute perfectly, and prevents us from running other tests in those stores. Our goal is to find the **optimal sample size**, using just enough stores to be confident in the results without wasting time or money.

3.2 Execution

Once the setup is complete, the focus turns to maintaining a stable environment in the stores. In a busy retail world, the main challenge is to keep the noise of daily operations from drowning out the results of our intervention. To isolate the real impact, we treat the test store as a clean room. This involves a strategic blackout of other major changes. We aim to avoid overlapping interventions, such as price shifts, secondary marketing campaigns, or major shelf reorganizations, during the test period. If too many variables move at once, we lose the ability to tell which change actually drove the results.

The Blackout Strategy

To ensure the true Lift isn't actually just a hidden marketing push, we implement a Strategic Blackout. Avoid overlapping the test with other interventions, marketing campaigns, or local competitor activity. For example, if we are testing a New Coffee Aroma, we must freeze price changes, assortment changes, or major staffing adjustments in those stores during the test window.

Operational Alignment

The most sophisticated statistical model will fail if the human element, as the last mile of execution, isn't aligned. In a busy store, an intervention that is too complex or time-consuming will suffer from low compliance, skewing the data. We prioritize making every test easy to implement. By keeping the tasks low-effort and the communication transparent, we ensure the intervention is executed consistently, preventing operational fatigue from biasing the results.

3.3 Evaluation

The final phase is analysing the results and turning the data into clear business insights using the Synthetic Control Method.

Applying SCM Model

We use our **R Code-Base** model to create the **Synthetic Control Unit** as the counter factual and compute the test effect through comparing tested stores performance against the **control** unit.

Extracting True Lift

By comparing the **Realized Outcomes** to the **Expected Performance**, we isolate the pure impact of our intervention from all other background noise. You can transform the results to actionable insights through different visualizations for example %Att over pre treatment period, validation period, and post treatment is a good way of showing the results as well as the accuracy of the model.

Actionable Insights & Visualization

To transform raw results into clear business insights, we can utilize various visualizations.

Visualization

To transform test results into clear business insights, you can use different visualizations. Below are the most common plots, which provide a transparent look at the results as well as the confidence intervals, overfitting risk, and estimation errors.

The Raw Path (Actual vs. Counterfactual) This plot tracks the performance of the test stores alongside their Synthetic Twin. By observing how closely these lines follow each other during the training period, we can verify that the model has established a solid baseline before the intervention begins.

The Att plot (Average Treatment Effect on the Treated)**:** The Att plot or Gap plot centers the results at zero to isolate the intervention effect. By removing the background performance, it

becomes easier to calculate the exact percentage of True Lift and see how it evolves over the course of the test.

Confidence Intervals and Estimation Errors To ensure a result is meaningful rather than a statistical fluke, we use error distributions. If the shaded confidence intervals remain clear of the zero line, it suggests the change is a direct result of the test rather than random noise.

Validation and Overfitting Checks: We use a Validation Period to ensure the model hasn't overfit to historical noise. If the model can accurately "predict" the store's behavior during a slice of time where no change occurred, we can have much higher confidence in the post-treatment results.

Before we look at the post-treatment results, we hold back a portion of the pre-treatment data (e.g., the 4 weeks immediately preceding the test) as a **Validation Window**. We train the model on the historical data and predict this window. If the model can't accurately track the store when nothing happened, we don't trust it to measure the lift when something does happen. A good way of visualizing the results is using the **%Att plot** (Average Treatment Effect on the Treated Units) over three distinct time periods:

1. **Pre-treatment - Training Period:** Showing the control unit's ability to mimic the store historically during the training period.
2. **Pre-treatment - Validation Period:** Proving the model accuracy to follow test stores before the test.
3. **Post-treatment Period:** Showing the expected performance (by control stores trend) vs. realized outcomes (by tested stores trend) of our intervention on tested stores over time.

This three-period visualization is also a handy tool to judge the results while avoiding over-fitting[3] risk. Over-fitting is a situation when the model is too perfect in the Training Period but fails during the Validation Period, so we can't trust the results of the post-treatment. If a plot shows perfect mimicking during the training period but fails to follow the trend during the validation period, it indicates that the model has over-fitted to training noise. In these cases, the model has memorized historical noise rather than the store's true patterns, and we cannot trust the results. A reliable model must maintain its accuracy through the validation phase to ensure the post-treatment lift is reliable.

---

[3] Over-fitting occurs when a statistical model is so "tuned" to historical noise and random fluctuations that it loses its ability to predict the future. In store testing, an over-fitted model might track a store's past perfectly but fail to provide a reliable counterfactual during the actual test period. Think of over-fitting like memorizing the answers to a specific past exam instead of learning the subject. You will get a perfect score on the old test (Training Period), but you will fail when a new, slightly different test begins (Post-treatment).

The Learning Bank

To maintain a high quality consistent store testing framework across the organization, the setup details of every intervention is documented using the **Experiment Setup & Design Template**. This ensures that the knowledge gained from one test is structured properly and remains easily accessible for future decision-making.

To complete the cycle, all findings and results are reported and stored using the **Experiment Results Template**. By using these twin templates, we ensure that every test—from its initial hypothesis to its final outcome—is recorded in a consistent format within our Learning Bank.

We archive every result in a consistent format. Successes provide a confirmation for informed rollouts, while failures provide valuable data that prevents us from repeating the same experiments in the future.

For more information You can also take a look at the full mind map for the Guidance Checklist for Initiating New Tests. It covers the entire process from defining the mission and hypothesis, to execution, evaluation, and extracting actionable insights using the Synthetic Control Method.