

Measuring the Effect of Retail Actions on Footfall: a Weather- and Calendar-Compensated Before/After Estimator with Placebo-Guarded Significance

Indivd Analytics

October 4, 2026

Abstract

We formalize a reproducible methodology for measuring the effect of retail actions (promotions, opening-hours changes, zone/layout/entrance reconfigurations, staffing or trading-pattern interventions) on daily footfall, removing weather, calendar, and baseline-trend confounders from the comparison. The estimator is a before/after comparison of the post-intervention residual (the part of observed footfall the baseline does not explain) against a per-location counterfactual baseline (what footfall would have been expected without the action) that is fit on the pre-intervention window with weather and calendar covariates (dow, weekend, month seasonality, temperature, precipitation) and a linear trend. The load-bearing significance statement is a power bound: a minimum-detectable-effect (MDE) bound is reported so a non-significant result is honestly reported as “below detectable bar,” never as “no effect.” Two design diagnostics are reported alongside it, and neither must “agree” with the other before a number is quoted: (i) a placebo/permutation null that randomly relocates the change point across the series and re-fits the same estimator — a diagnostic of cut-placement noise whose tail carries limited inferential weight at the true change point (kept for honesty); and (ii) a positive-control recovery grid that injects a known additive effect and confirms the estimator recovers it above the detection bar, giving the smallest injected level the grid clears (here $2.0\times$ the minimum-detectable-effect bar). We prototype the method on a large zone-structure reconfiguration at a single treated store location on 2026-08-08, inferred (not log-confirmed) in the production data: the store’s distinct-zone count — the data marker of the reconfiguration — rose from 13 to 15. We show the estimated unit-level effect (-25.9 visitors/day, -4.3% of the pre mean of 599.1 visitors/day) on this inferred marker sits within the placebo null and below the detectable bar at $\alpha = 0.05$; at the achievable power of this short series it is “below the detectable bar,” not “the action did nothing.”

1 Problem and Motivation

A high-value analyst/retailer question is “what was the effect of my action?” Retail actions—a promotion, a change in opening hours, a zone/layout/entrance reconfiguration, a staffing or trading-pattern intervention—move daily footfall, but footfall is also driven by weather, day-of-week, holidays, salary days, and a smooth baseline trend. A naive before/after of raw means attributes weather and calendar swings to the action; what is needed is the *action’s* contribution after those confounders are removed.

The platform’s production forecasting system already supplies weather and calendar covariates and open-hours semantics (daily footfall is the open-hours-only sum): it trains a forward neural-network / autoregressive forecaster on weather, calendar, holiday, salary-day, and autoregressive-target features, and enforces trading-hours rules at serve time. It nevertheless keeps no retained historical counterfactual and no exposed trend/seasonality decomposition, so it cannot by itself supply the retrospective baseline this question needs. Section 2 reviews what that forecasting system can and cannot supply; a dedicated effect estimator is therefore required, and it is built *on* the same covariate layer rather than in isolation.

2 Capability assessment: the platform’s production forecasting system

Throughout, “the forecasting stack” means the platform’s production forecasting system. The assessment finds that a dedicated estimator is required; it builds *on* the existing covariate layer rather than from scratch.

- **Trend/seasonality decomposition available?** No. The stack’s covariates (periodic calendar encodings, weather, holidays, salary days, autoregressive target lags) are consumed by a forward neural-network / autoregressive forecaster; there is no exposed per-location trend or seasonality component an analyst can subtract.
- **Stored per-location counterfactual forecast?** No. The served forecast is retained for only the most recent days, so no forecast exists for a June–August episode; the platform forecast cannot serve as the retrospective baseline.
- **Inherited covariate layer.** The estimator inherits the stack’s stored hourly weather history (temperature, precipitation) and open-hours semantics.

3 Method

3.1 Estimator: residual-vs-baseline before/after

Let y_t be observed daily footfall (open-hours sum) on day t , and X_t a vector of covariates: day-of-week, weekend flag, month sin/cos, temperature, precipitation, and days since the pre-window start (linear trend). A counterfactual – what footfall would have been expected without the action – is fit by ordinary least squares (OLS, the standard linear fit) on the pre-intervention window only:

$$y_t = X_t\beta + \varepsilon_t, \quad t \in \text{pre}, \quad \hat{y}_t = X_t\hat{\beta} \text{ for } t \in \text{post}. \quad (1)$$

The daily action effect is the mean post-intervention residual – the part of each day’s footfall the counterfactual does not explain:

$$\hat{\delta} = \frac{1}{n_{\text{post}}} \sum_{t \in \text{post}} (y_t - \hat{y}_t), \quad (2)$$

with standard error $\text{se} = s_{\text{pre}}/\sqrt{n_{\text{post}}}$ (how much the estimate would wobble across repeated samples; here 36 visitors/day against a pre-period baseline of 599.1 visitors/day), where s_{pre} is the pre-window residual standard deviation. The counterfactual is fit exclusively on pre-intervention data so it cannot absorb the action itself; weather/calendar seasonality and the trend are in the counterfactual, so their swings are removed from $\hat{\delta}$. This standard error assumes the daily residuals are independent; retail daily residuals are typically autocorrelated, so the bar is indicative rather than exact. The post-based sensitivity reported in Section 5 (standard error about 90 visitors/day, minimum detectable effect about 176 visitors/day) bounds the extra uncertainty that the independence assumption leaves unmodelled.

3.2 Significance: MDE bound with placebo and positive-control diagnostics

A single point estimate is not enough. The load-bearing significance statement is the minimum-detectable-effect (MDE) power bound derived below: a result under it is “below detectable bar,” never “no effect.” Two design diagnostics are reported alongside that bound at the false-alarm rate we accept, $\alpha = 0.05$ (5%); they are kept for honesty and need not “agree” before a number is quoted:

1. **Placebo / permutation null (design diagnostic).** Relocate the change point at random across the interior of the series and re-fit the identically-specified estimator, $N=2000$ times. This yields the distribution of $\hat{\delta}$ under the null (no real intervention at the placed point). The observed $\hat{\delta}$ is compared against the α -tail of this null, but because the null is inflated by trend-extrapolation bias its tail carries limited inferential weight at the true change point; it is reported as a diagnostic of cut-placement noise, not as the significance verdict.
2. **Positive-control recovery (arithmetic detection check).** Inject a known additive effect $X \cdot \text{MDE}$ onto the post window and confirm the same estimator recovers it above the α detection bar ($|\hat{\delta}| > z_{1-\alpha/2} \text{se}$ — the $z \cdot \text{se}$ threshold, about two standard errors). The smallest X the grid clears is the *empirical detection bar*; on this series the grid validates the estimator’s detection arithmetic, not counterfactual quality in general.

3.3 Minimum detectable effect

$$\text{MDE}(\alpha) = z_{1-\alpha/2} \frac{s_{\text{pre}}}{\sqrt{n_{\text{post}}}}, \quad \text{MDE}_{80\%} = (z_{1-\alpha/2} + z_{1-\beta}) \frac{s_{\text{pre}}}{\sqrt{n_{\text{post}}}}, \quad (3)$$

with $z_{1-\beta}$ chosen for 80% power ($\beta = 0.2$), i.e. an 80% chance of detecting an effect that is really there. A result below the MDE bar is reported as “below detectable bar,” never as “no effect.”

3.4 Confounder handling

Weather (temperature, precipitation) and calendar (dow, weekend, month seasonality) enter the counterfactual, so their confounded movement is subtracted from the action estimate. This removes the modelled weather, calendar, and trend components from the comparison; it does not by itself guarantee full causal separation, given the short window and the absence of a control location. The linear trend absorbs any smooth ramp. Where a valid control location exists (stable zone count, high pre-correlation ≥ 0.70 with the treated series, no concurrent intervention), a difference-in-difference leg (DiD — comparing the treated store’s change against a similar untreated store’s change) removes shared shocks; no qualifying control store exists for this episode, so DiD is not applied here.

4 Data

All data come from the platform’s production analytics database. The treated unit is a single store location (anonymized) with a data-inferred structural change, over a 79-day design window (2026-06-20...2026-09-08) yielding 49 pre-intervention and 30 post-intervention usable open-hours daily observations. The calendar span from 2026-06-20 to 2026-09-08 is 81 days; the 79 design days exclude (i) the change day 2026-08-08 itself, dropped as a one-day washout (a buffer day held out around the intervention so the change day’s own disruption contaminates neither side), and (ii) the span’s final day 2026-09-08, dropped as an incomplete boundary day: the analysis takes only complete days via a half-open time bound, so the run-day partial at the end of the span is excluded (the design’s post days run 2026-08-09...2026-09-07). The result is 49 pre days plus 30 post days. Weather covariates come from the platform’s stored hourly weather history (temperature, precipitation; daily aggregation). Intervention: an *inferred* zone-structure change on 2026-08-08 (the distinct-zone count first reaches 15, from 13). **Caveat:** the raw action-activity log records no zone-structure event for this store/date (its only such events concern a different, unrelated store at dates not coinciding with this change), and the distinct-zone count oscillates 13/15 after the change point, so the episode is presented here as a method illustration on a data-inferred structural marker, not as a confirmed causal event.

5 Results

5.1 Estimator output on the inferred episode

Table 1 summarizes the estimator. The point estimate is -25.9 (daily, post window), i.e. -4.3% of the pre mean, with standard error 36 and $n_{\text{pre}}=49$ / $n_{\text{post}}=30$ pre/post days. Figure 1 shows observed vs counterfactual and the residual.

5.2 Placebo null and significance

The placebo/permutation null (Table 2, Fig. 2) has mean -211.1 and sd 603.7 across 2000 draws (the draws resample the finite set of interior change-point placements), far wider than the estimator’s own standard error of 36 that applies at the observed cut, and it is not centered on zero. This non-centering is expected under the placebo’s design: the null is the distribution of the same estimator under *random* change-point placement. Because the counterfactual trend is fit on the pre-window and then extrapolated over the post window, most random cuts place the boundary where that extrapolated trend runs far beyond the range it was fit on, so the residual accumulates a systematic negative drift. The null therefore measures both cut-placement noise and trend-extrapolation bias, not sampling error at the observed cut. The observed effect falls well within it (two-sided $p=0.4088$ at $\alpha=0.05$), so the unit-level effect is *not* distinguishable from zero under the placebo null. Because the null is inflated by extrapolation bias, its p -value carries limited inferential weight at the true change point; the honest-null conclusion below rests on the power/MDE bound rather than on this placebo tail.

5.3 Minimum detectable effect and power

$MDE(\alpha)=70.5$ and $MDE_{80\%}=100.8$ (daily). The observed absolute effect (25.9) is below the MDE bar, so the honest report is “no detectable effect at the achievable power,” not “no effect.” That bar is built from the *pre*-window residual scatter (sd 197.1), the only dispersion the pre-fit counterfactual supplies. The post-window residuals are markedly more scattered (rms 485.3, sd 492.9, roughly $2.5\times$ the pre-window level). That widening is driven by a few unusually large post-window days, visible as the tall residual spikes in the right-hand panel of Figure 1, rather than by a uniform rise in day-to-day scatter. If the estimator used the post-window dispersion instead, the standard error would be about 90 visitors/day and the minimum detectable effect about 176 visitors/day. Even against that more conservative bar the observed absolute effect (25.9) stays below the threshold, so the honest-null conclusion is unchanged, and strengthened, under the noisier post window.

5.4 Positive-control validation

Table 3 + Fig. 3 confirm the estimator recovers an injected effect. Every grid level recovers the injected effect plus the standing observed effect (-25.9): recovered = injected + observed. This is by construction, not a validation failure. The injection adds a known effect to the same post-window observations whose baseline shortfall is already -25.9, so the linear estimator picks up both together and the recovered value tracks injected + observed exactly (the recovery bias is -25.9 at every level). The corollary is that the detection bar is reached when injected + observed clears the $z_{1-\alpha/2} \cdot se$ threshold, which is why $2.0\times$ MDE is the smallest injected level cleared. The analytic $MDE(\alpha)=70.5$ is the one-standard-error z -bar ($z_{1-\alpha/2} \cdot se$): a true effect of exactly that size is detected only about half the time, so it is a 50%-power boundary, not a guarantee. The empirical $2.0\times$ bar is larger because the recovery grid steps in coarse 0.5-MDE increments and because every recovery carries the constant offset above; both bars are reported, and the honest conclusion uses the larger, empirical one. On this series the grid therefore shows that the null result is a *power* limitation of the short dense series rather than an estimator blind to real effects: it validates the detection arithmetic on these data, not the counterfactual’s quality in general.

6 Validation & Quality

6.1 Degenerate-data guardrails

- Daily footfall is summed only over open-hours rows (hours per location).
- A day with missing weather is dropped from the design (no imputation).
- The estimate is withheld if $n_{pre} < 10$ or $n_{post} < 5$ design rows.
- The placebo null is withheld if fewer than 10 interior change-point candidates exist.
- A result below the MDE bar is reported as “below detectable bar,” never “no effect.”

6.2 Difference-in-difference applicability

We evaluated candidate control locations (Table 4). None met the joint bar (stable zone count and pre-correlation ≥ 0.70), so DiD is **not applicable** here and we report the treated-unit before/after only. The DiD leg is defined and ready when a valid control exists in a future episode.

7 Discussion and limitations

The prototype demonstrates the full estimator pipeline on a data-inferred natural-experiment marker (a structural data signal — here the distinct-zone count — that plays the role of an intervention without a logged action) and, importantly, produces a *disciplined null*: the estimated unit-level effect is below the MDE bar and inside the placebo null. The placebo null is non-centered and wide by construction (random cut placement plus trend-extrapolation drift), so the power/MDE bound, not the placebo tail, carries the honest-null conclusion. This is the correct behavior for a short (80-day) noisy retail series and is itself a meaningful result (the method refuses to overclaim). Limitations: (i) the short window limits power ($MDE_{80\%}=100.8$ daily is large relative to the 599.1 daily pre-mean baseline, i.e. an effect

on the order of the observed -4.3% swing is below the detectable bar at the achievable power); (ii) the counterfactual is a transparent OLS, not the full neural-network / autoregressive forecast—fine for interpretability but lower-power than the platform’s forecasting system would be; (iii) weather covariates come from the platform’s stored hourly weather history, so they share any of its gaps/spatial resolution; (iv) DiD requires a valid control that was not present for this episode; (v) the intervention itself is data-inferred (distinct-zone count oscillation 13/15; no recorded action-activity event for this store/date), so this prototype is a method demonstration on an inferred marker, not an estimate of a confirmed layout change; (vi) the pre/post dispersion gap: the pre-fit standard error (36) rests on the calmer pre-window, whereas the post-window scatter is $2.5\times$ larger (sd 492.9), so a post-based bar would be 176 visitors/day, still above the observed effect.

Positioning. The statistical machinery used here — OLS standard errors and the permutation-null logic — is a standard-methods application rather than a novel technique, and it is anchored in the two methodological traditions it draws on. Intervention analysis in the tradition of [1] established estimating a discrete intervention’s effect on an observed series against a modelled counterfactual, with the effect read against its uncertainty — exactly this paper’s before/after design; the placebo null follows the exact-inference logic of permutation methods [2], where the observed statistic is judged against its own randomisation distribution rather than an asymptotic approximation; the positive-control recovery grid — injecting known effects and verifying the estimator recovers them — is the simulation-based validation companion to that null. The minimum-detectable-effect and power convention follows the standard behavioural-science power-analysis framing of [3]; the difference-in-differences candidate-control evaluation follows the panel-estimation cautions of [4]. The intervention anchor, both power/inference anchors and the DiD anchor are graded *peer-reviewed*; the paper’s claims are graded as own-measured results on this series and a method demonstration, not as externally validated science.

8 References

References

- [1] Box, G. E. P., and Tiao, G. C. Intervention analysis with applications to economic and environmental problems. *Journal of the American Statistical Association*, 70(349):70–79, 1975. doi:10.1080/01621459.1975.10480264. Peer-reviewed. Finds intervention analysis: estimating a discrete intervention’s effect on a time series against a modelled counterfactual, read against its uncertainty — the design this paper applies to an inferred layout action.
- [2] Ernst, M. D. Permutation methods: a basis for exact inference. *Statistical Science*, 19(4), 2004. doi:10.1214/088342304000000396. Peer-reviewed. Establishes judging an observed statistic against its randomisation (permutation) distribution as exact inference — the logic of this paper’s placebo null.
- [3] Cohen, J. *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Lawrence Erlbaum Associates, 1988. doi:10.4324/9780203771587. Standard research monograph (book; not peer-reviewed in the journal sense). The minimum-detectable-effect and power convention this paper’s estimator tables use.
- [4] Bertrand, M., Duflo, E., and Mullainathan, S. How much should we trust differences-in-differences estimates? *The Quarterly Journal of Economics*, 119(1):249–275, 2004. doi:10.1162/003355304772839588. Peer-reviewed. Establishes the serial-correlation and inference cautions behind the candidate-control evaluation this paper reports for its DiD layer.
- [5] Indivd Analytics (2026) *Forecast-Quality Methodology*. Benchmarking forecast quality on short, sparse retail series, including degenerate-data guardrails for small samples. Internal sibling paper; source-and-method anchoring only, not methodological positioning.
- [6] Indivd Analytics (2026) *Data Catalog*. Platform documentation of the observed visitor metric, open-hours capture, and data-quality rules. Internal source-and-method documentation; anchors the measurement layer.
- [7] Indivd Analytics (2026) *Entrance-Line Visitation Patterns in Retail*. Characterising entrance-line visitation patterns (open-hours window, peak-hour concentration, weekend/weekday and weekly shape) across the retail network. Internal sibling paper; source-and-method anchoring only.

Figures and Tables

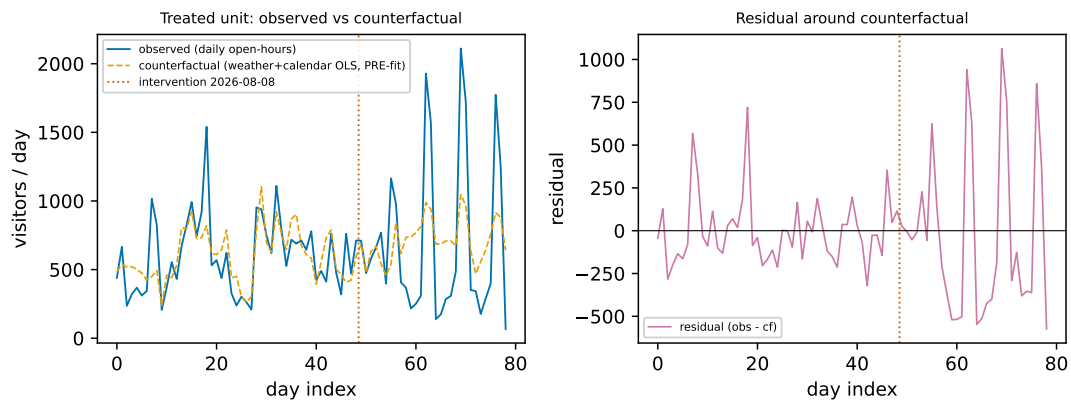


Figure 1: Observed daily open-hours footfall vs weather/calendar counterfactual, and the residual, for the treated unit. Vertical dashed line = intervention change point (2026-08-08).

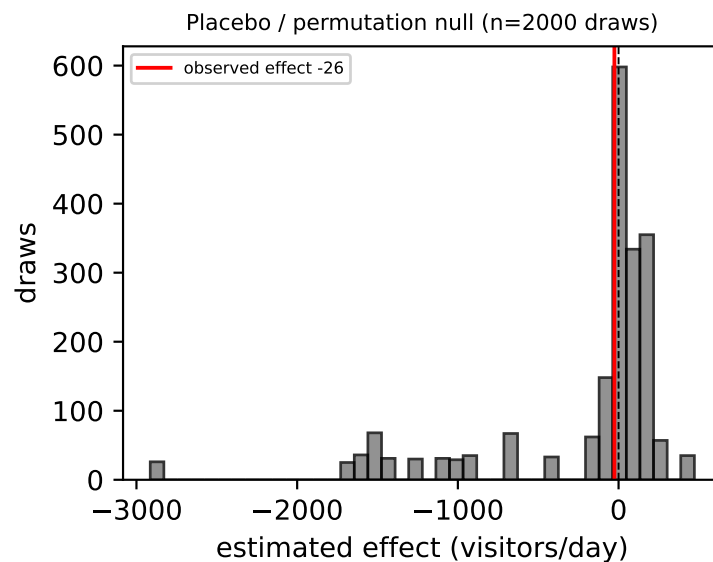


Figure 2: Placebo/permutation null: distribution of the estimated effect across random change-point shifts. Red line = observed effect.

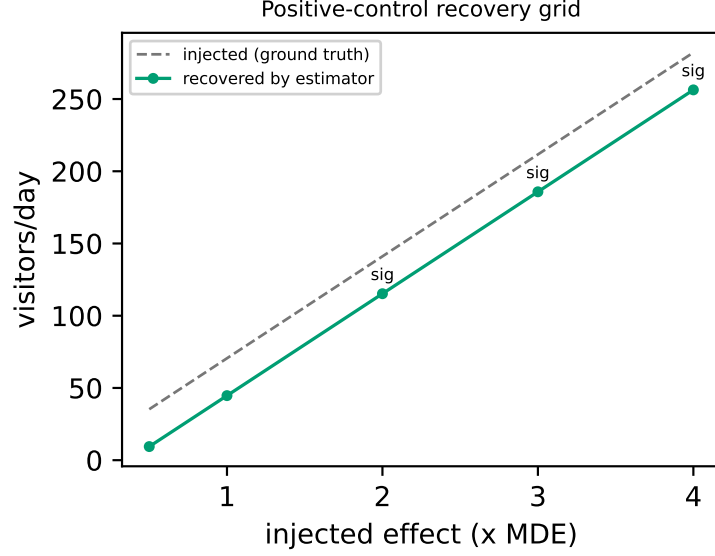


Figure 3: Positive-control recovery grid: injected vs recovered effect at MDE multiples.

Estimator input / output	value
Treated unit	one store location (anonymized)
Intervention (change date)	2026-08-08
Zone count before $\rightarrow$ after	13 $\rightarrow$ 15
Pre window	2026-06-20 - 2026-08-08
Post window	2026-08-08 - 2026-09-08
Pre days (n_{pre})	49
Post days (n_{post})	30
Pre mean (visitors/day)	599.1
Estimated effect (visitors/day)	-25.9 (-4.32%)
Std error of mean effect	36
Min detectable effect ($\alpha=0.05$)	70.5
Min detectable effect (80% power)	100.8
Placebo null p (two-sided tail)	0.4

Table 1: Estimator summary for the treated unit.

Statistic	value
Null draws	2000
Null mean (visitors/day)	-211.1
Null std (visitors/day)	603.7
Two-sided p ($ \text{tail} \geq \text{observed} $)	0.4

Table 2: Placebo/permutation null statistics.

Injected effect (x MDE)	injected/day	recovered/day	detected
0.5	35.3	9.4	no
1.0	70.5	44.7	no
2.0	141.1	115.2	yes
3.0	211.6	185.8	yes
4.0	282.2	256.3	yes

Table 3: Positive-control recovery grid.

Candidate control	pre-corr	n(common)	stable zones
Candidate 1	0.6	49	no
Candidate 2	0.3	49	yes
Candidate 3	0.4	49	yes
Candidate 4	-0	49	yes
Candidate 5	0.6	49	no
Candidate 6	0.1	49	no
Candidate 7	0.6	49	yes
Candidate 8	0.3	49	yes

Table 4: Difference-in-difference candidate controls evaluation.