

Preprocessing Standards for Entrance-Line Demand Data: Capture-Window Recovery, Entrance-Zone Aggregation, and Genuine-Signal Scaling

Indivd Analytics

October 2, 2026

Abstract

We formalize and validate three preprocessing standards that govern the intake of per-location demand data on a forecasting platform. These decisions, more than any single model, shape whether the downstream forecast scale and totals can be trusted, so we propose them as preprocessing gates applied before training; their downstream forecast benefit (accuracy, calibration, and scale error) is not yet measured and is future work. For every location the raw hourly entrance-line counter is (i) *captured* within an open-hours window—the hours of the day in which the store’s entrance counters actually record traffic—so that a daily total is an *open-hours-only* daily count; (ii) *aggregated across entrance zones* so that the canonical location total is the sum over its entrance zones, and not the sum of the per-zone counters, because a visitor who crosses several zones would otherwise be counted once per zone crossed (the zone-aggregation rule); and (iii) *scaled on genuine-positive training targets*, because zero-valued rows behave as ingestion gaps (*absent zeros*) rather than genuine zero traffic—a policy hypothesis rather than an established standard (Section 4.3). We validate all three on a single retained read-only snapshot of production data. Because the intended fixed-day-of-week opening-hours schedule is empty in the production database, we recover the capture window empirically with a modal-contiguous-hour method: a window is recovered for every location that carries contiguous hourly rows, with a median span of 9 hours (a minority are genuinely 24-hour capture); the recovery’s validation is self-consistency only, and no held-out-day or external-schedule check has been run. The entrance-zone double-count is location-heterogeneous: summing per-zone counters over-counts 24.4 % of the aligned locations (Section 4.2). On the policy hypothesis that absent zeros are absence rather than traffic, the genuine-signal rescale corrects a median $1.073\times$ downward bias in the all-rows mean—a hypothesis the mechanism and worked examples support, but whose downstream benefit is untested (Section 4.3); 34 all-zero and 69 absence-dominated locations are candidates for exclusion from training. Together, the three gates make the preprocessing target well-scaled, causally scoped, and capture-consistent before any forecast is trained. Throughout, we refer to three intake rules in plain language (Section 1.1)—the zone-aggregation rule, the open-hours rule, and the labeled-series rule—and use their names as stable labels. The paper is self-contained: every domain term is defined in Section 1.1, and no system-internal knowledge or client-identifying information is required to follow the method—all worked examples are de-identified.

1 Problem and Motivation

Per-location demand forecasts drive production totals, landing pages, and per-store insights. Yet the raw signal is not a simple counter: it is an entrance-line crossing series that is (a) captured only within each store’s open-hours window, (b) aggregated across entrance zones in a way such that per-zone counters do not simply add to the location total, and (c) corrupted by absent zeros whenever an ingestion or upstream coverage gap is read as “zero traffic”. In production, each of these choices has at some point silently distorted forecast scale; three failure patterns motivate the gates below:

1. *Capture-window mismatch.* The platform treats daily counts as open-hours only, yet forecasts are produced on a 24-hour basis, and the intended fixed-day-of-week opening-hours schedule is empty in the production database. A 24-hour downscale of an open-hours-only counter therefore demands the exact window; estimating that window wrong mis-scales the total.

2. *Entrance-zone double-count.* A shopper who crosses three zones is counted in the entrance aggregate once but in the per-zone series three times. Summing the per-zone series to “recover” a location total therefore inflates the denominator, corrupting the scale of downstream models and reports (an over-scale incident traced to exactly this).
3. *Zero-absence scaling.* An under-scale incident showed that large all-zero blocks are ingestion gaps, not zero traffic. When absent zeros flow into the z-score standardization they dominate the mean and variance and pull a mean-regressing forecasting model (a model whose output level follows the average of its training target) down to a tiny, zero-heavy level (for example, a raw all-rows mean of 10.2 while its genuine open-hour average was 23.25). The preprocessing target should be the genuine-positive series, not the all-rows series.

Neither a single window constant nor a naive zone or row sum is safe. The three standards replace guesswork with an empirical, reproducible preprocessing practice that every downstream consumer (train, serve, report) inherits. We compose them as three gates applied at intake, stated in Section 2.

1.1 Terminology and scope

We use the following terms throughout; each is defined here so the paper is self-contained.

Capture window. The contiguous block of clock hours in which a location’s entrance counters consistently record rows on a typical day. A location with window 9:00–17:00 captures nine hours of traffic per day and is silent outside them. The daily sum over these hours is the *open-hours-only daily count*: totals are computed inside the window, never over the full 24 hours.

Entrance-zone aggregation. A location may expose several “entrance zones”, each with its own hourly counter (one counter per door or bank of doors). The canonical per-location total, `location_visits`, is the *sum over the entrance zones* as produced by the platform’s zone merge (the “`mergeZones: true`” setting). The per-zone counters are the individual `zone_visits` series. A consumer should *not* take the sum of the `zone_visits` as a location total: because a visitor can cross multiple zones, that sum counts some visitors more than once and over-counts the location.

Aligned window. The recent, closed multi-day interval (here the *last-14-day aligned window*) over which we compare the zone and location series day by day, matched on date. The per-location double-count factor is the median of the *daily* ratios over this window, which is robust to isolated days where one series is only partially captured.

Genuine-positive / absent zeros. A *genuine-positive* row has value strictly greater than zero. A row that is zero (or missing) because no data was ingested for that slot is an *absent zero*: it reflects absence, not zero traffic. The training target policy (gate S) drops absent zeros and keeps only the genuine-positive series.

z-score standardization. The common standardization that subtracts a series’ mean and divides by its standard deviation before training. If absent zeros enter the statistic, both the mean and the standard deviation are pulled toward a tiny, zero-heavy level, which is the under-scale failure we prevent at ingestion by using the genuine-positive series only.

The three intake rules. Three rules govern how per-location demand data is read. The *zone-aggregation rule*: a location total is the entrance-zone sum, never the sum of per-zone counters. The *open-hours rule*: daily counts cover a store’s open hours only—sum within the capture window. The *labeled-series rule*: forecast and observed series are labeled as such and never conflated. This paper validates the first two and builds the scaling guarantee the third protects.

Schema terms. `location_visits` denotes the canonical per-location hourly total (the entrance-zone sum); `zone_visits` denotes the hourly count from a single entrance zone; and the `mergeZones` setting produces the former by summing the latter. These names are introduced in ordinary words here so that no internal schema knowledge is required.

De-identified examples. Where a concrete worked example is useful, we label individual stores anonymously (“Store A”, “Store B”). The labels are arbitrary within each section and carry no information about the underlying store.

2 Method

We define `PreprocStd` as three preprocessing gates applied to the raw hourly intake before any target is built. Each gate is a deterministic, data-driven step we propose as candidate standards; the validation here is preprocessing-internal (distributions, mechanism, and worked examples), and a controlled with/without-gates comparison against an all-rows target is future work.

- (W) **Capture-window recovery.** For each location, we recover the open-hours capture window as the modal contiguous hourly set across its observed days: per local day we take the set of local hours that carry a row, keep only contiguous blocks (a contiguous run has no interior gap), and estimate the window as the modal such block. The recovery is specified as follows: a day contributes to the modal vote only if its hour set is contiguous—a day with an interior dropout (a missing interior hour) is not counted, and its hours are never bridged or imputed. The recovered window is the modal contiguous hour set across the contributing days; when two or more hour sets tie on frequency, the earliest block wins (the lexicographically smallest set of hours). We impose no minimum number of contributing days: a location is windowed if at least one of its days carries a contiguous hour set, and a location whose day sets are never contiguous is left without a recovered window. This empirical method follows a companion internal study of the same preprocessing questions [2]. The canonical daily count is then the sum over local hours within this window. A single uniform *default* window (e.g. 07:00–17:00) is never assumed, because windows vary by location and the default holds for only a minority of stores (Section 4.1).
- (Z) **Entrance-zone aggregation.** The canonical location target is `location_visits`: the entrance-zone sum (the `mergeZones: true` aggregate). A consumer should not sum raw per-zone `zone_visits` to build a location total, because a multi-zone store double-counts a visitor once per zone crossed (the zone-aggregation rule). We validate the over-count factor $r = \sum \text{zone_visits} / \text{location_visits}$ per location over an aligned window and use $r > 1.05$ to flag a double-count risk; the location total is always the entrance aggregate, not the zone sum.
- (S) **Genuine-signal scaling.** The training target vector for a location is the stream of *genuine-positive* rows: values strictly greater than zero within the capture window. Zero-valued and absent rows are dropped at intake (the zero-absence policy). The model’s z-score mean and standard deviation are therefore computed from the genuine open-hour series, not from the absent-zero-diluted all-rows histogram. The served forecast scale should match this target: an open-hours per-hour rate, not a diluted 24-hour average. We propose this gate as a policy hypothesis (the zeros are read as absence, supported by the mechanism and worked examples below) rather than as an established standard; its downstream benefit is untested (Section 6).

Each gate is paired with a measured validation metric (reported in Section 4): window match-share for (W), the over-count factor distribution for (Z), and the genuine-vs-all-rows mean ratio and zero-rate for (S).

3 Data

Evidence is a de-identified, read-only snapshot of a production analytics store taken as of 2026-09-07. Every number in Sections 4 and 5 is derived from that retained snapshot; no figure, table, or prose line contains a hand-typed value. The three gates correspond to the intake rules stated in Section 1.1. The per-location visitation series spans 2026-06-17 to 2026-09-07 across 230 locations, aligned in local time.

- *Capture window.* The fixed day-of-week opening-hours table has 0 rows and is therefore empty in the production database. We recover the window empirically per location over the full retained series: a window is recovered for 230 of the 230 locations, of which 2 are genuinely 24-hour capture and 228 are windowed. We report the span/open/close distributions and the intra-location match-share: the share of a location’s days whose captured hour set equals its modal window, which is the recovery’s internal coverage. The median span is 9 hours, the modal open and close hours are 10:00 and 18:00 local, and the *default* 07:00–17:00 is matched exactly by only 1 of the 228 windowed locations.
- *Entrance-zone aggregation.* We count locations that have both `location_visits` and `zone_visits` over the aligned last-14-day window, and report the per-location median of the daily ratio $r =$

$\sum \text{zone_visits} / \text{location_visits}$. We also report the distribution of r within the over-counting class.

- *Genuine-signal scaling.* Per location we quantify the zero rate (the absent-zero share), the all-zero count, the absence-dominated share (at least 50 % zero rows), and the scaling factor genuine-positive-mean divided by all-rows-mean.

4 Results

4.1 Capture-window recovery: the window is empirical, not a fixed constant

Figure 1 (and Table 1) characterize the recovered capture window. A modal peak sits at open 10:00 and close 18:00 local—the modal row of Table 1; the median open hour across locations is 9.0 (9:00 local, whereas the modal window opens at 10:00), from the quantile rows—with a median span of 9 hours, and the span-9 block holds 71 of the 228 windowed locations. Two stores are genuinely 24-hour, opening around the clock and therefore excluded from windowing. At the other extreme, 48 of the 230 locations yield a span-1 window (about 21 %, a single contiguous hour). A one-hour contiguous-row window plausibly reflects ingestion sparsity rather than a genuine open-hours window: it is consistent with the recovery’s minimum-support rule (a location is windowed on a single contiguous day), and the open-hours interpretation is weakest for this class. The *default* 07:00–17:00 is matched exactly by only 1 of the 228 windowed locations, so a fleet-wide 07–17 constant would mis-scale 227 of the 228 windowed locations. The intra-location match-share accompanies the recovery as a self-consistency check: a median of 80.9 % of a location’s days match its modal window, and the 90th percentile reaches 100.0 %. This match-share is high by construction for any modal estimator, so it measures internal coverage of the recovery rather than external validation. This says the preprocessing should carry a per-location window rather than a fleet-wide constant, and that the daily total is summed over that window only. A 24-hour downscale that assumes the uniform 07–17 default would therefore be structurally wrong for most locations.

4.2 Entrance-zone aggregation: over-counting is location-heterogeneous

Figure 2 (and Table 2) report the double-count factor $r = \sum \text{zone_visits} / \text{location_visits}$ per location over the last-14-day aligned window. For locations that hold a single entrance zone (or whose per-zone counters merge cleanly), $r \approx 1$; but for multi-zone stores r grows past 1.05, with 47 of the 193 aligned locations over-counting (24.4 %). Within that over-counting class, the median r is 2.00 $\times$, the 75th percentile is 3.27 $\times$, and the maximum reaches 17.78 $\times$; two de-identified examples over-count as well (Store A at 2.19 $\times$, Store B at 1.37 $\times$, the latter only modestly). The takeaway is that a per-location entrance aggregation is correct, and any consumer that sums raw `zone_visits` for a location total, or that cross-checks a total against zone sums, will see an over-count that grows with the number of member zones. The zone-aggregation rule encapsulates this, and r quantifies its cost.

4.3 Genuine-signal scaling: absent zeros bias the target down

Figure 3 (and Table 3) quantify the zero-absence risk. Across the 230 locations with `location_visits`, the all-rows mean understates the genuine mean by a median factor of 1.073 $\times$, with a 90th percentile of 2.36 $\times$ and a maximum of 80 $\times$. There are 69 absence-dominated locations (at least 50 % zero rows) and 34 all-zero locations that are candidates for exclusion from training; the all-zero set is a subset of the absence-dominated set by construction (a zero rate of 100 % is at least 50 %), so the two counts are not disjoint exclusion lists. The rescale gate (S) addresses this: for example, one de-identified store (Store A) has an all-rows mean of 10.2 against a genuine mean of 23.25 (a 2.27 $\times$ factor), and Store B has 19.3 against 44.47 (2.30 $\times$). The genuine/all-rows ratio distribution is computed over the 196 locations that carry at least one positive row; the 34 all-zero locations have no positive mean and are excluded from that distribution. If absent zeros flowed into the z-score standardization they would pull the mean-regressing forecasting model down to a tiny, zero-heavy level: the exact under-scale failure. We therefore offer the genuine-positive target as a policy hypothesis: on the reading that zeros are absence and not traffic, keeping genuine positives only would make the training scale equal the real open-hour traffic level rather than the zero-diluted level, but the snapshot carries no per-row ground truth to confirm that reading, and whether the choice improves downstream forecasts is untested (Section 6).

5 Validation and Quality

5.1 Validation metrics

Metric selection follows the causal scope of each gate:

- *Window match-share.* For gate (W), the share of a location’s days whose captured hour set equals its recovered modal window. A high median (80.9%, reaching 100.0%) shows the recovery is stable and self-consistent; because the metric compares each day’s captured hours against the modal window derived from those same days, it does not by itself establish that the window equals the true open-hours set. The reported match-share is a point estimate and carries no confidence interval. With the fixed opening-hours schedule empty at the snapshot anchor there is no external reference; reconciliation against the populated schedule table is future work.
- *Entrance-zone over-count factor.* For gate (Z), the per-location median of the daily ratio $\sum \text{zone_visits} / \text{location_visits}$ over the aligned window. A value above 1.05 flags a double-count risk; the over-counting share (24.4% of aligned locations, with a class maximum of 17.78 $\times$) quantifies the magnitude. The 1.05 flag is a working convention, and a threshold or window-length sensitivity sweep is not computable from the shipped de-identified snapshot (which carries no per-zone columns); performing one is future work.
- *Zero-absence scale factor.* For gate (S), the genuine-positive-mean versus all-rows-mean ratio (at or above one; median 1.073 $\times$, 90th percentile 2.36 $\times$, maximum 80 $\times$) and the zero-rate distribution (across 230 locations, 69 at 50% or more zero rows and 34 all-zero). These are mechanism and distribution evidence for the policy hypothesis that the genuine-positive series is the appropriate training scale, rather than a controlled with/without-gates forecast comparison.
- *Reproducibility.* All numbers re-derive from the retained read-only snapshot, and every figure, table, and prose value is produced programmatically rather than typed by hand. The window recovery is model-free and cheap, entrance aggregation is a direct read of the canonical column, and the scale gate is a direct threshold on positive rows: all three are interpretable and auditable in order.

5.2 Reproducibility and traceability

Every reported number is re-computable from the retained read-only snapshot, and all figures, tables, and inline numbers are produced programmatically from that same snapshot, so no hand-typed value enters the paper. The three invariance claims rest on the intake rules of Section 1.1.

6 Discussion

The capture window is best recovered per location. Because the fixed opening-hours schedule is empty, a consumer cannot know a location’s open-hours window from the scheduled hours. Our modal contiguous-hour-set recovery addresses that missing scheduled-hours source; its validation is currently self-consistency only—no held-out-day or external-schedule check has been run—accompanied by a median 80.9% intra-location stability (self-consistency of the recovery), and it is the approach the companion internal study [2] also uses. A fleet-wide constant (a single 24-hour downscale to open) would be structurally wrong: spans range from 1 to 24 hours with a median of 9, and the 07–17 default is matched by only 1 of the 228 windowed locations. The preprocessing should therefore attach a window to every location and sum daily counts within it.

The entrance-zone aggregate is the appropriate location total. Per-zone counters double-count a multi-zone shopper (the over-count class has median r of 2.00 $\times$ and a maximum of 17.78 $\times$), and 24.4% of aligned locations over-count; we propose that this entrance-zone aggregate be adopted as the ingestion contract, so that consumers need not re-sum zones.

Absent zeros most plausibly reflect absence, not zero traffic. As a matter of policy plus the observed evidence, the zero rate has a long tail (median 10.48%, with 69 locations at 50% or more zero rows), and the all-rows mean understates the genuine mean by a median factor of 1.073 $\times$ (the full distribution is in Section 4.3). We state this as a policy hypothesis rather than as a measured fact: the zeros are consistent with ingestion gaps, but the snapshot carries no per-row ground truth that would distinguish a genuine zero-traffic slot from an absent one. On that reading the training target should drop rows at or below zero and keep only genuine positives, else the z-score mean and variance are polluted and the

metric	p25	median	p75	p90
window span (h)	8.0	9.0	10.0	11.0
open hour (local)	8.2	9.0	10.0	10.0
close hour (local)	16.0	18.0	18.0	20.0
match-share (%)	68.8	80.9	98.8	100.0
store_hours rows	—	0	—	—
24h capture locations	—	2	—	—
modal open/close (local)	—	10:00 / 18:00	—	—

Table 1: Recovered capture-window characterization: per-location span, open, and close distributions, and the modal-window intra-location match-share.

mean-regressing forecasting model settles at a tiny, zero-heavy level: the under-scale failure. Measuring forecast accuracy [1] regardless of this preprocessing choice cannot repair the bias once it has been baked into the target. Dropping zero rows also changes the estimand: the target becomes positive-conditional, so a model trained on it cannot by construction predict a genuine zero-traffic hour; zero-inflated targets and imputation are out of scope here.

Limitations and future work. The analysis rests on a single snapshot of a production store (as of 2026-09-07, a retained span of roughly 12 weeks); it carries no seasonality or holiday controls, so the reported factors describe this window rather than a stable year-round quantity. A separate study of the same preprocessing questions ([2]) observes the same empty fixed opening-hours schedule; once a business-hours reconciliation populates it, the empirical recovery should be reconciled against the official fixed-day-of-week windows and the open-hours daily-total definition re-verified. The per-location double-count factor is a median over an aligned window in which some locations carry as few as one aligned day (minimum 1), and the $r > 1.05$ flag is a working convention rather than a calibrated threshold, so the 24.4% over-count share is sensitive to both that threshold and the aligned-window length, and no sensitivity analysis is included. The shipped de-identified artifacts carry no per-zone columns—the joined panel is a single fleet-level daily total—so a threshold or window-length sweep is not computable from the snapshot; performing one would require the raw per-zone series, which is future work. Fleet-wide the ratio has a median of 1.0 and a 90th percentile of 2.0, so the over-counting share describes the tail rather than the typical location. A zone-membership-weighted model of the double-count, and a per-(zone, location) row-identity check, are future work. The zero-absence reading is likewise not sharpenable from the shipped artifacts: telling a genuine zero-traffic slot from an ingestion gap would need the raw per-slot counter series, which the snapshot does not carry, so a zero-block adjacency or context audit is future work. The scale gate’s motivating incident and its worked example are the same de-identified store (Store A); it is a single illustrative case, not independent corroboration. The three gates here cover intake; serve-time quality (whether the served forecast is degenerate, scaled, or corrupted) is out of scope for this paper and belongs to a separate study of forecast quality. In short, the evidence supports window recovery, entrance-aggregate-only totals, and a genuine-positive target, offered here as policy hypotheses rather than as binding policy. No controlled downstream comparison (forecast accuracy, calibration, or total-scale error against an all-rows-trained baseline) exists in this study; the case that the gates improve the downstream forecast rests on the mechanism and the two worked examples, and measuring that benefit is future work.

References

- [1] R. J. Hyndman and A. B. Koehler. Another look at measures of forecast accuracy. *International Journal of Forecasting*, vol. 22, no. 4, pp. 679–688, 2006.
- [2] Indivd Analytics research team. *Retail demand data preprocessing: capture windows, entrance aggregation, and genuine signal scaling*. Working note, Indivd Analytics, 2026.

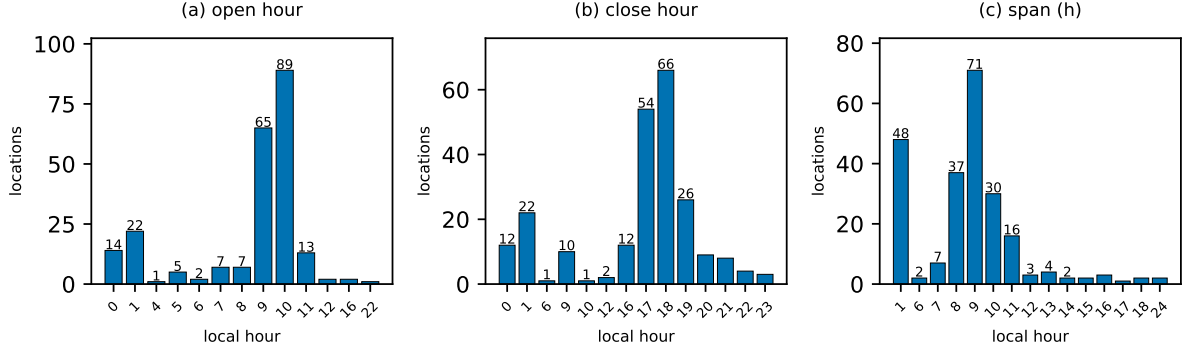


Figure 1: Capture-window recovery: empirical recovery from capture data (no schedule table available), so the window is recovered from the observed hourly rows rather than from any fixed opening-hours schedule. (a) local open-hour distribution; (b) local close-hour distribution; (c) span distribution in hours; per-location quantiles and the modal hours are in Table 1.

metric	p25	median	p75	p90
all locations $r = \text{zone}/\text{loc}$	1.00	1.00	1.00	2.00
over-counting class ($r > 1.05$)	2.00	2.00	3.27	8.97
n locs with both (aligned 14d)	193	—	—	—
over-counting locs	47 (24.4%)	—	—	—
example store A	2.19x	—	—	—
example store B	1.37x	—	—	—

Table 2: Entrance-zone double-count factor r per location (median over the last-14-day aligned window), with the over-counting share, the over-counting class distribution, and de-identified worked examples.

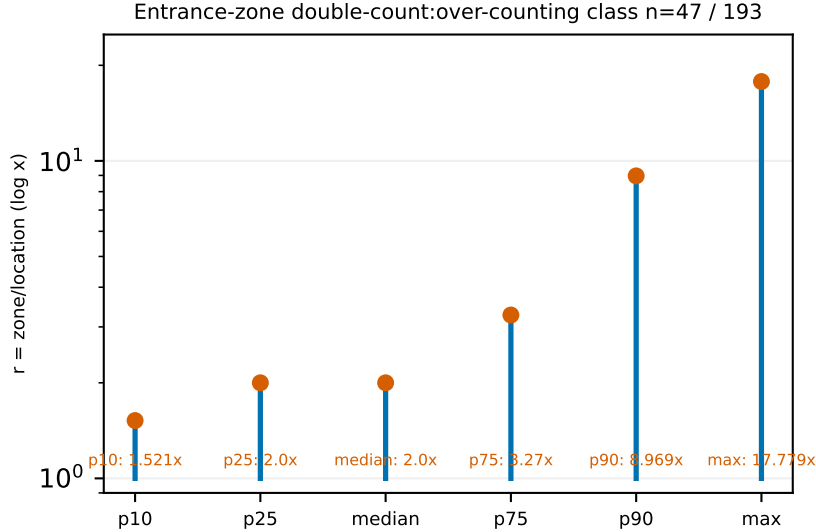


Figure 2: Entrance-zone double-count factor $r = \sum \text{zone_visits} / \text{location_visits}$ per location (median over the last-14-day aligned window, log axis). Most locations sit near $r \approx 1$ (a single or merged entrance zone), but multi-zone stores over-count whenever zones are summed; the over-counting share, class range, and worked examples are in Table 2. The entrance aggregate should be used exclusively.

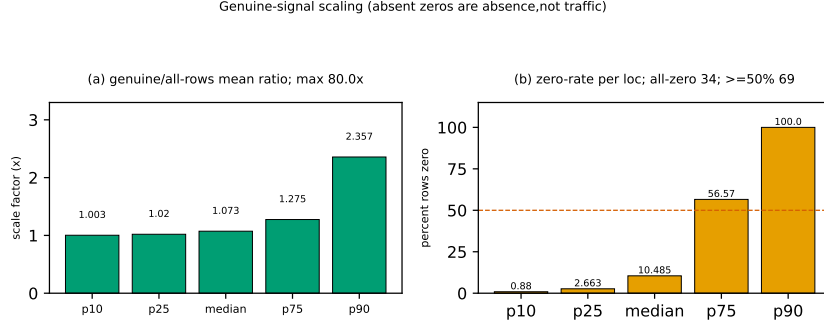


Figure 3: Zero-absence scaling. (a) Distribution of the genuine-positive-mean over all-rows-mean ratio (linear axis; at or above one; median $1.073\times$; full quantiles in Table 3). (b) Zero-rate distribution across locations (at least 50 % zero rows: 69; all-zero: 34). On the policy hypothesis that most zeros are absence rather than traffic, the genuine-positive target and its z-score scale are the appropriate choice, and the all-rows mean understates the real open-hour level.

metric	p25	median	p75	p90
genuine/all-rows mean ratio	1.020	1.073	1.275	2.357
zero rate (%)	2.66	10.48	56.57	100.00
n locations	230	—	—	—
all-zero locations	34	—	—	—
absence-dominated ($\geq 50\%$ zero)	69	—	—	—
example store A: all/pos mean	10.2 / 23.25 (2.27x)			
example store B: all/pos mean	19.3 / 44.47 (2.30x)			

Table 3: Genuine-signal scaling: zero-absence risk per location (genuine-positive mean divided by all-rows mean, zero rate, and the absence and zero classes), with de-identified worked examples.