

Stockout-Aware Relational Demand Forecasting and Cross-Dataset Error Trade-offs

Mengyi Chen^{*1}, Jeffrey Jin²

Email: mengyichen.us@gmail.com

¹Computer Science, Lawrence Technological University, MI, USA

²Applied Analytics, Columbia University, NY, USA

*Corresponding Author

Abstract

Retail sales are a censored view of demand when products become unavailable. This study evaluates profile-weighted temporal reconstruction and relational forecasting on FreshRetailNet-50K, then examines cross-dataset error tradeoffs on M5. FreshRetailNet contains 50,000 store-product series, 4.85 million daily records, hourly sales, and stockout labels. Category-hour profiles estimate missing intraday shares, while a trailing 28-row own-series prior supplies the demand scale. Artificial contiguous-hour masking on days without labeled open-hour stockouts yields reconstruction WAPE of 8.82%–21.32%; it does not validate natural-stockout demand recovery. On 206,562 observable rows in the seven-day official test, calibrated relational forecasting achieves RMSE 0.6761, MAE 0.3621, and WAPE 29.27%, versus 30.54% for observed-sales boosting. Recovery alone does not improve WAPE, and the incremental graph contribution is small. Two earlier seven-day windows support the direction of the calibrated comparison but do not assess seasonal stability. M5 evaluation covers 30,490 series and 28 rolling one-day-ahead origins. Fixed-budget 1:1 pooling lowers relational-model RMSE from 2.1075 to 2.0716, while increasing WAPE from 71.67% to 72.21%. Alternative sampling ratios and full-target augmentation show the same metric tradeoff. Dense series receive the strongest squared-error benefit, without a WAPE advantage. The findings distinguish controlled reconstruction accuracy from observable-sales prediction and show that cross-retailer pooling does not uniformly improve forecasting.

Keywords: Stockout-Censored Demand; Temporal Reconstruction; Retail Forecasting; Relational Learning; Global Forecasting; Transfer Learning; Freshretailnet-50K; M5.

I. INTRODUCTION

Retail forecasting is often formulated as prediction from a historical sales sequence (J. Wang et al., 2024), but recorded sales and economic demand are not equivalent when a product is unavailable. If a store sells every unit on hand before the end of an operating period, the data record the realized transactions rather than the purchases customers would have made under continuous availability. Learning directly from such observations depresses estimated demand, reinforces conservative replenishment, and can create a feedback loop in which future forecasts inherit the effects of past stockouts. This problem was identified in early work on prediction from stockout-censored sales (Wecker, 1978) and is closely related to limited-dependent-variable estimation (Tobin, 1958), incomplete-observation methods (Kaplan & Meier, 1958), and censored inventory learning (Besbes & Muharremoglu, 2013; X. Ding et al., 2002; Huh & Rusmevichientong, 2009; Lu et al., 2008). The operational consequence is not merely a larger statistical residual. A downward forecast changes order quantities, service levels, assortment decisions, and the distribution of inventory across stores.

The timing and resolution of availability information determine how much latent demand can be recovered. Jain, Rudi, and Wang showed that stockout timing contains substantially more information than a single end-of-period stockout flag (Jain et al., 2015). Subsequent work separated censoring from

inventory-record inaccuracy (Mersereau, 2015), developed multiproduct data-driven policies (Shi et al., 2016), quantified uncertainty in censor-aware inventory decisions (Ban, 2020), and integrated observable features into censored-demand control (J. Ding et al., 2024). These results motivate the use of FreshRetailNet-50K's 24-hour sales and stock-status vectors rather than treating each daily record as a single binary censored observation. They also impose a clear validity boundary: the stock-status sequence identifies hours affected by availability, but it does not reveal the counterfactual purchases that would have occurred during those hours. Controlled masking can measure reconstruction error against an observable full-day total, but cannot establish accuracy for naturally unmet demand.

Retail data also have strong cross-series structure. Stores in the same market respond to common calendar effects; products in the same category share consumption rhythms; and store-product series with similar histories may contain useful signals even when their anonymized identifiers have no direct semantic mapping. Global models can share statistical strength across many series, an approach formalized through the locality-globality perspective (Montero-Manso & Hyndman, 2021) and supported by empirical studies of cross-learning (Semenoglou et al., 2021). Learned graphs offer a further mechanism for cross-series dependence. Graph WaveNet introduced adaptive adjacency when explicit edges are incomplete (Wu et al., 2019), while MTGNN learned graph structure for multivariate time series without requiring a known physical network (Wu et al., 2020). These precedents do not make anonymized retail stores geographic nodes. They instead support a relational graph whose edges are defined by observed grouping and demand-pattern similarity.

The M5 Accuracy competition provides an unusually demanding external test of transferable retail representations. Its public data contain Walmart unit sales, a calendar, events, SNAP indicators, and weekly prices organized in a large product-store hierarchy (Makridakis et al., 2022b; University of Nicosia, 2020). The competition demonstrated the strength of global machine-learning and ensemble methods (Makridakis et al., 2022), and later research showed that transfer learning can reduce the computational cost of strong M5-style models (Wellens et al., 2022). At the same time, M5 does not provide on-hand inventory, replenishment records, hourly availability, or a stockout flag. Its zeros are observed zero sales and cannot be relabeled as stockouts. M5 is therefore used here to test cross-dataset forecast transfer, not to claim external validation of stockout detection or lost-demand recovery.

Retail forecast error is not itself an inventory outcome. Prior work emphasizes the interaction among hierarchy, promotions, aggregation, and evaluation (Fildes et al., 2022) and the gap between forecasting evidence and inventory consequences (Goltsos et al., 2022). We therefore use FreshRetailNet-50K for reconstruction and M5 for forecast transfer, without converting either result into an unmeasured inventory-policy claim.

The study asks whether hourly omissions can be reconstructed under controlled pseudo-censoring, whether reconstructed histories and lagged relations improve next-day FreshRetailNet prediction, and

how joint source–target fitting changes M5 forecast error. Recovery is evaluated on artificially masked days without labeled open-hour stockouts. Fresh forecasting is assessed separately on observable sales, while M5 evaluates 853,720 bottom-level outcomes without a stockout interpretation. Figure 1 summarizes these information boundaries.

The contribution combines profile-weighted temporal reconstruction, grouped demand-pattern graphs, global boosted forecasting, and cross-dataset error-tradeoff analysis. Separate recovery ablations identify the roles of the own-series prior and the hourly profile; fixed-budget sampling ratios and store/category comparisons examine the dependence of transfer results on training composition. The analysis follows transfer learning as reuse across domains (Pan & Yang, 2010), while recognizing the distribution shift emphasized in domain-adaptation research (Ganin et al., 2016).

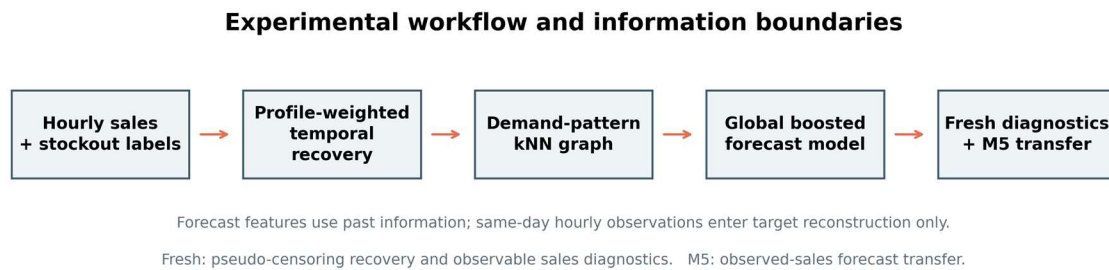


Figure 1. Experimental workflow and information boundaries.

II. RESEARCH METHOD

A. Data Audit and Experimental Units

FreshRetailNet-50K contains 4,500,000 training rows: 50,000 store–product series observed daily from March 28 through June 25, 2024. The same series appear on the next seven dates, producing 350,000 evaluation rows. The panel covers 18 cities, 898 stores, 865 products, and three category levels. Each row contains daily and 24-hour sales, 24 stock states, a 06:00–22:00 stockout count, promotions, calendar flags, and weather. The daily and hourly sales fields were globally normalized by the data provider, so absolute FreshRetailNet values are reported on the released sales scale (Dingdong-Inc, 2025; Y. Wang et al., 2025).

Full-file scans found no nulls, duplicate series–dates, missing panel dates, or static-attribute changes. Both hourly arrays had length 24. Daily sales agreed with the hourly sum within numerical tolerance, and the open-hour count equaled the sum of status positions 6–21. Status one denotes unavailability. Because 2.92% of labeled training hours and 3.08% of labeled evaluation hours still recorded positive sales, these transactions were retained in the observed daily input.

M5 contains 30,490 item–store series, 3,049 items, 10 stores, and 1,941 days in the evaluation sales matrix. Its validation matrix exactly duplicates the first 1,913 daily columns and was not concatenated.

History was d_1-d_{1913} , containing 58,327,370 cells; the test $d_{1914}-d_{1941}$ contains 853,720 outcomes. No stockout or inventory field exists. FreshRetailNet-50K was released in May 2025, although its observations date from 2024; M5 was released in 2020 (Dingdong-Inc, 2025; University of Nicosia, 2020). Table 1 gives the experimental splits.

Table 1. Dataset composition and temporal splits

Dataset	Split	Observations	Series	Period	Availability
FreshRetailNet	Training	4,500,000	50,000	Mar 28–Jun 25, 2024	Hourly labels
FreshRetailNet	Evaluation	350,000	50,000	Jun 26–Jul 2, 2024	Hourly labels
M5	History	58,327,370	30,490	d_1-d_{1913}	Not available
M5	Test	853,720	30,490	$d_{1914}-d_{1941}$	Not available

B. Information Boundaries and Prediction Design

Reconstruction is retrospective: completed hourly sales and states determine censored training targets. Forecasting is prospective and uses demand lags, rolling summaries, lagged neighbors, day of week, and a known-schedule scenario in which target-day discount, holiday, and activity fields are available before the forecast is issued. Target-day hourly sales, stock status, and realized weather are excluded. In rolling one-day-ahead evaluation, an earlier test day becomes available before the next origin.

The primary FreshRetailNet diagnostic uses 206,562 evaluation rows with no labeled open-hour stockout. It is a conditional measurement subset selected after observation, rather than a population identifiable before forecasting. A secondary score uses all 350,000 recorded-sales rows. Neither score measures the accuracy of counterfactual demand on naturally censored days; those days support only a model-implied forecast–sales gap. In M5, every value remains observed sales. Shared features are normalized lags, rolling moments, day of week, intermittency, and relational lag aggregates; availability and Fresh-specific promotion inputs are excluded.

C. Profile Weighted Temporal Reconstruction

Let y_{it} denote daily observed scales and s_{it} hourly sales. For first-level category c , the hourly share p_{ch} is category sales in hour h divided by total 24-hour category sales, estimated using training rows without labeled open-hour stockouts. Thus, the profile estimates relative intraday timing rather than the demand level of an individual series. For a completed day, ℓ_{it} is the sum of category hourly shares over labeled stockout hours within 06:00–22:00. The temporal prior m_{it} is the same series' mean sales on days without labeled open-hour stockouts among the preceding 28 rows. If none exists, it uses the corresponding past city/category peer mean, then current observed sales if neither history is available. The reconstructed daily demand is calculated using Equation (1)

$$\hat{d}_{it} = \frac{1 - \ell_{it}}{\max(1 - \ell_{it}, 0.05)} y_{it} + \ell_{it} m_{it} \quad (1)$$

$1 - \ell_{it} \geq 0.05$, this reduces to $\hat{d}_{it} = y_{it} + \ell_{it} m_{it}$. The hourly profile supplies the missing share and the temporal prior supplies its scale. Summing direct hourly estimates $p_{ch} m_{it}$ over the When unavailable

hours gives the same addition; these are equivalent interpretations. Relative to scaling partial sales by $1/(1 - \ell_{it})$, the temporal term avoids relying entirely on the remaining sales to estimate the daily level. It can still be inaccurate when recent own-series demand is unrepresentative. If $\ell_{it} > 0.95$, the denominator floor attenuates the coefficient on observed sales to $(1 - \ell_{it})/0.05$. The rule therefore does not always preserve the full observed total. This occurred on 16,686 training rows and 1,363 evaluation rows; it is a numerical boundary of the implemented reconstruction, not evidence of accurate demand recovery under extreme censoring.

Recovery evaluation used the 206,562 evaluation rows without labeled open-hour stockouts. For each mask length of two, four, six, or eight hours, a fixed-seed generator selected an admissible contiguous open-hour interval uniformly. Its sales were removed and the full-day observed total was retained as the target. The comparators were partial sales, uniform scaling, category-profile scaling, the own-series prior alone, partial sales plus a uniform missing share times the prior, category-share temporal reconstruction, and a 50:50 temporal/peer variant. The peer variant used other same-day city/category rows without labeled stockouts and excluded the focal row. For a mask of h hours, the uniform control assigns missing share $h/16$ and scales partial sales by $1/(1 - h/16)$, assuming equal demand shares across the 16 open hours. Category shares instead retain their empirical 24-hour normalization. These experiments test artificial pseudo-censoring, not the stockout mechanism that generated natural unmet demand. Figure 2 shows the distribution of the original availability labels used to define the evaluation subset, while the complete profile-weighted temporal reconstruction procedure is detailed in Algorithm 1.

Algorithm: Training, Reconstruction, and Evaluation Procedure

Input:

Hourly transaction data
Training cutoff T
Labeled stockout information
Category hierarchy
City/category peer information

Output:

Reconstructed daily totals
Leakage-free next-day features for evaluation

- 1: Set the training cutoff T
- 2: Retain all observed hourly transactions up to the defined observation period
- 3: Using rows with $t \leq T$ and no labeled open-hour stockout:
- 4: Group observations by first-level category
- 5: Estimate one normalized 24-hour profile for each first-level category
- 6: For each completed day t :
- 7: Construct daily baseline m_{it} from eligible own-series observations over:
 $t - 28, \dots, t - 1$
- 8: If eligible own-series observations are insufficient:
- 9: Use past city/category peer information
- 10: End If
- 11: If additional information is still necessary:
- 12: Use current observed sales
- 13: End If

```

14:    Sum the corresponding profile shares over labeled unavailable hours
        6, ..., 21
        to obtain  $l_{it}$ 

15:    Apply the displayed reconstruction equation using  $m_{it}$  and  $l_{it}$ 
16: End For

17: For controlled masking:
18:    Replace the observed total with its unmasked partial total
19:    Replace the availability mask with the sampled unavailable interval
20:    Retain the complete total only as the evaluation target
21: End For

22: Construct next-day features using only completed historical information

23: During evaluation:
24:    Freeze the estimated profile

25:    After each day has closed:
26:        Update historical information
27:    End After
28: End During

```

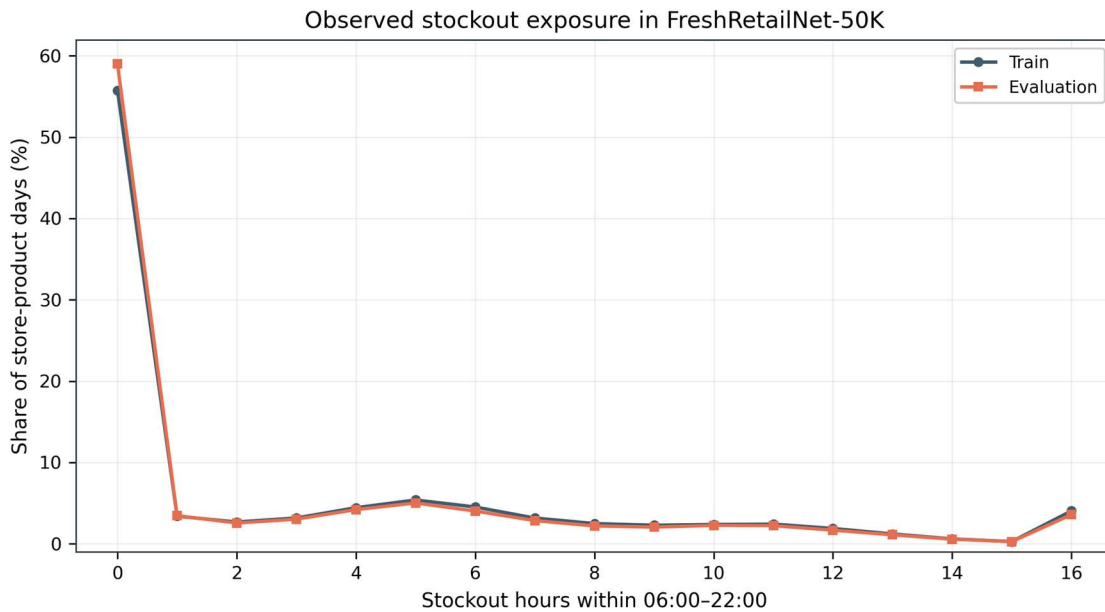


Figure 2. Distribution of labeled open-hour stockout exposure in the FreshRetailNet training and evaluation periods.

D. Learned Relational Graph

FreshRetailNet candidates share city and first-level category. Standardized demand-pattern embeddings contain log-transformed mean and standard deviation, coefficient of variation, the difference between late- and early-half means relative to the mean, lag-1 and lag-7 autocorrelation, zero frequency, seven normalized weekday means, and stockout frequency. Each node retains up to five nearest external candidates in Euclidean distance. Weights are proportional to $\exp(-\text{distance})$; groups with fewer candidates use self-padding to maintain five aggregation slots. This construction describes statistical similarity within retail groups rather than geographic proximity.

The FreshRetailNet graph contains 50,000 nodes and 250,000 directed aggregation slots, comprising 249,320 external neighbor slots and 680 self-padding slots; 39 groups are singletons. Each neighbor

series is divided by its own initial scale before weighted aggregation. Forecast features are neighbor demand at lags 1 and 7 and the neighbor seven-day mean. The observed-history graph ablation constructs both its topology and lag inputs from recorded sales.

M5 candidates share state and category. Its target-specific graph uses the same demand-pattern statistics, omitting unavailable stockout information. Graph identities and weights are fitted within each dataset; model features have a shared definition. Zero-shot transfer reuses source-model parameters, whereas pooled fitting estimates a new model from source and target examples. The graph construction and feature timing procedures are detailed in Algorithm 2.

Algorithm 2. Graph Construction and Feature Timing Procedure
Input: - Training histories - Representation cutoff T - Dataset indicator (FreshRetailNet or M5) - Completed sales or reconstructed histories Output: - Fixed neighbor graph - Time-valid graph-based and focal-series features <pre>1: At representation cutoff T: 2: For each series i: 3: Summarize its training history 4: End For 5: Standardize embedding columns across all series 6: For each focal series i: 7: If dataset = FreshRetailNet: 8: Form candidate set using series with the same city and category 9: Else if dataset = M5: 10: Form candidate set using series with the same state and category 11: End If 12: Compute distances between focal series i and all external candidates 13: Select up to five nearest external neighbors 14: If fewer than five external neighbors are available: 15: Pad the remaining neighbor slots with focal series i 16: Assign distance = 0 to each padded slot 17: End If 18: For each of the five neighbor slots j: 19: Compute raw weight: 20: $w_{ij} = \exp(-\text{distance}_{ij})$ 21: End For 22: Normalize $\{w_{ij}\}$ across the five neighbor slots 23: End For 24: For each one-day-ahead forecast origin t: 25: For each focal series i: 26: For each neighbor j in the five fixed neighbor slots: 27: Take neighbor j's completed history available before origin t 28: Divide this history by neighbor j's fixed initial scale 29: Extract normalized value at $t - 1$ 30: Extract normalized value at $t - 7$ 31: Compute normalized mean over $t - 7, \dots, t - 1$ 32: End For 33: Aggregate the neighbor values using the normalized neighbor weights 34: Combine the three neighbor-based values with:</pre>

```

- focal-series normalized lags
- rolling summaries
- intermittency features
- calendar features

34:      Exclude:
      - target-day sales
      - target-day availability
35:      End For

36:      Keep graph topology fixed throughout the evaluation window

37:      After the current day is completed:
38:          Update only completed sales histories
            or reconstructed histories
39:      Use the updated histories for the next one-day-ahead origin

40: End For

```

E. Forecast Models and Transfer Protocol

Baselines were seasonal naive, a seven-day Fresh mean, a 28-day M5 mean, and Croston-SBA with alpha 0.10. Croston’s method (Croston, 1972) and its Syntetos-Boylan correction (Syntetos & Boylan, 2005) provide the reported intermittent benchmark; TSB (Teunter et al., 2011) and probabilistic occurrence-size models (Snyder et al., 2012) motivate the sparsity analysis.

Histogram boosting was selected because global trees perform well in retail forecasting (Makridakis et al., 2022a; Wellens et al., 2024), and histogram-based split construction scales efficiently (Ke et al., 2017). We used scikit-learn (Pedregosa et al., 2011). The 12 common features are normalized lags 1, 7, 14, and 28; rolling means and standard deviations over 7 and 28 days; expanding zero frequency; day-of-week sine and cosine; and expanding standard deviation. Relational variants add the three neighbor features. Fresh-only models may additionally use discount, holiday, and activity under the known-schedule scenario. The full stockout-aware relational forecaster (SARF) combines reconstructed histories, graph features, scheduled variables, and scale calibration. The source transfer model excludes the three scheduled variables and receives no Fresh scale-calibration multiplier.

Each series is scaled by its observed mean over an initial history ending before the first fitted origin: the first 28 FreshRetailNet days and d_1-d_{1185} for M5, with a minimum scale of 0.1. Normalized targets are clipped to $[0, 30]$ and features to $[-30, 30]$. Predictions are constrained to be nonnegative and converted back to the dataset’s released scale. Every histogram booster uses squared-error loss, learning rate 0.06, 55 iterations, 31 leaves, minimum leaf size 40, L2 regularization 0.25, and disabled early stopping. Random procedures use seed 20260803; masking uses this seed plus mask length. Ridge regression uses alpha 2.0.

Fresh fitting uses all 3,100,000 valid series–origin examples from days 29–90. A separate model fitted through June 18 predicts June 19–25; 198,987 rows without labeled open-hour stockouts yield a scale-calibration multiplier of 0.9451444, computed as total observed sales divided by total predicted sales. It is then frozen for the official evaluation. Profiles and graph relations use the complete official training period through June 25, so the calibration interval adjusts scale and is not an isolated validation set for representation learning or model selection.

To examine sensitivity to the short source horizon, two additional seven-day windows begin on June 5 and June 12. Each model, category profile, graph, and embedding standardization is rebuilt using only dates preceding that window. Its multiplier is estimated from the final seven pre-window days under the same scale-calibration design. These checks remain close in time and do not test seasonal drift.

M5 fitting uses all 30,490 series at 52 fortnightly origins $d_{1186}, d_{1200}, \dots, d_{1900}$, giving $N = 1,585,480$ examples. The augmentation protocol retains all N target examples and adds N sampled Fresh examples. A separate fixed-budget sensitivity experiment keeps N total rows while varying source:target ratios over 0:1, 1:4, 1:2, 1:1, and 2:1, using seeded sampling without replacement. Both nonrelational and relational boosters are fitted for every ratio. Increasing the source share under this budget replaces target rows; augmentation instead doubles the training budget. The two protocols therefore answer different questions about source reuse.

Each M5 test date is a rolling one-day-ahead origin. Earlier test outcomes enter lagged features, seasonal-naive forecasts, and Croston updates; the 28-day mean remains fixed at its pretest value. The representations use complete official pretest histories and are frozen throughout the test. Forecasts are evaluated at the bottom level without reconciliation; hierarchical combination (Hyndman et al., 2011) and MinT (Wickramasuriya et al., 2019) remain possible extensions. Table 2 summarizes the model settings.

Table 2. Model Settings and Training Budgets

Component	Setting	Scope
Recovery	Category-hour missing share with 28-row temporal prior; denominator floor 0.05	Fresh completed days
Graph	Five slots per node; city/category or state/category groups	Training histories
Common features	12 lag, rolling, intermittency and calendar features	Both datasets
Booster	Learning rate 0.06; 55 iterations; 31 leaves; min leaf 40; L2 0.25	All HGB models
Augmented pooling	1,585,480 source + 1,585,480 target examples	Augmented 1:1 fitting
Fixed-budget pooling	1,585,480 total; source:target 0:1 to 2:1	Ratio sensitivity
Scale calibration	0.9451444 from 198,987 eligible rows	Fresh official test
Seed	20260803; masking uses seed + hours	Sampling and bootstrap

F. Metrics and Statistical Analysis

MAE and RMSE are measured on the released sales scale for FreshRetailNet and in item units for M5. WAPE is the sum of absolute errors divided by the sum of absolute observed sales; relative bias uses signed errors in the numerator. Ordinary MAPE is excluded because zero sales make it undefined (Hyndman & Koehler, 2006). Mean RMSSE first divides each series' test mean squared error by its mean squared first difference across the entire official training history, including leading zeros, then averages the square roots over eligible series. It is an equal-series bottom-level measure, not the official hierarchical, revenue-weighted M5 WRMSSE.

For calibrated SARF versus observed-sales HGB in FreshRetailNet, and fixed-budget and augmented pooled versus target-only relational HGB in M5, we calculate the difference in each series' MAE. A 1,000-replicate paired series bootstrap supplies a 95% interval, alongside the proportion of series with lower MAE and a two-sided Wilcoxon signed-rank diagnostic (Demšar, 2006). Shared dates, stores, products, and graph neighbors remain dependent, so these intervals describe conditional series resampling rather than cluster-robust population uncertainty. Store, category, and zero-sales-prevalence analyses use the same fixed predictions.

III. RESULTS AND DISCUSSION

A. Data characteristics and censoring severity

The FreshRetailNet training panel had mean daily sales of 0.9986, median 0.7, maximum 44.9, and a 4.46% zero-sales rate. Evaluation mean sales were 1.1931, with median 0.8, maximum 49.9, and a 4.16% zero rate. Labeled open-hour stockouts affected 44.27% of training rows and 40.98% of evaluation rows. The mean number of affected hours was 3.180 and 2.897, respectively. All 16 open hours were labeled unavailable on 4.04% of training days. Availability-related censoring is therefore a common feature of this panel.

Training mean sales were 0.07293 in status-zero open hours and 0.00492 in status-one hours, a ratio of 14.83. The evaluation means were 0.08526 and 0.00567. Positive sales in some status-one hours could reflect within-hour timing, delayed status recording, or partial availability; the anonymized data cannot distinguish these explanations. They motivate retaining the actual daily transactions in the reconstruction input rather than forcing labeled hours to zero. M5 differs markedly. Its 30,490 series contain 66,927,173 item units across d_1 – d_{1941} , with 67.998% zero cells and a maximum daily value of 763. Zeros comprise 68.196% of training history. This much greater intermittency creates substantial distribution shift from FreshRetailNet, whose explicit availability labels also have no target-domain equivalent.

B. Controlled Pseudo Censoring Recovery

In Table 3, uncorrected partial-sales WAPE rises from 12.32% with two hidden hours to 51.16% with eight. Category-profile scaling gives 9.01%, 15.20%, 21.29%, and 27.73% across the four mask lengths. Category-share temporal reconstruction reduces these errors to 8.82%, 13.97%, 18.06%, and 21.32%. Its relative WAPE reduction against partial sales ranges from 28.39% to 58.33%. The two additional controls separate the temporal prior from the hourly profile. The prior alone has WAPE 32.28% because it discards the remaining same-day observations. Adding a uniform missing-share estimate to the partial total lowers WAPE to 9.58%, 14.77%, 18.52%, and 21.92%. Replacing the uniform share by category-hour shares reduces WAPE by 0.46–0.80 percentage points. The strongest recovery results therefore combine contemporaneous partial sales, an own-series demand scale, and an estimate of the missing intraday share.

Table 3. Recovery Under Controlled Contiguous Hour Masking

Method	2 h WAPE	4 h WAPE	6 h WAPE	8 h WAPE	8 h bias
Partial sales	12.32%	25.78%	38.91%	51.16%	-51.16%
Uniform scaling	10.01%	16.28%	21.71%	28.55%	-2.33%
Category profile scaling	9.01%	15.20%	21.29%	27.73%	0.68%
Own-series prior only	32.28%	32.28%	32.28%	32.28%	-7.67%
Uniform share + prior	9.58%	14.77%	18.52%	21.92%	-5.00%
Category share + prior	8.82%	13.97%	18.06%	21.32%	-3.58%
Category share + prior + peers	9.80%	16.35%	21.88%	26.65%	-1.59%

The peer-augmented variant has higher WAPE than category-share temporal reconstruction at every mask length. A same-day group mean can differ from the focal series' own demand scale; this is a plausible explanation rather than a separately identified cause. Figure 3 compares all seven methods. None of these masks triggers the denominator floor, and none uses the known full-day total as a prior fallback. These results establish recovery accuracy only for the specified artificial omission process. Noted that for Table 3 every cell uses 206,562 real days without labeled open-hour stockouts. The prior-only control predicts the full-day total; the two share-plus-prior controls retain the unmasked partial total. These are pseudo-censoring errors, not natural-stockout demand errors.

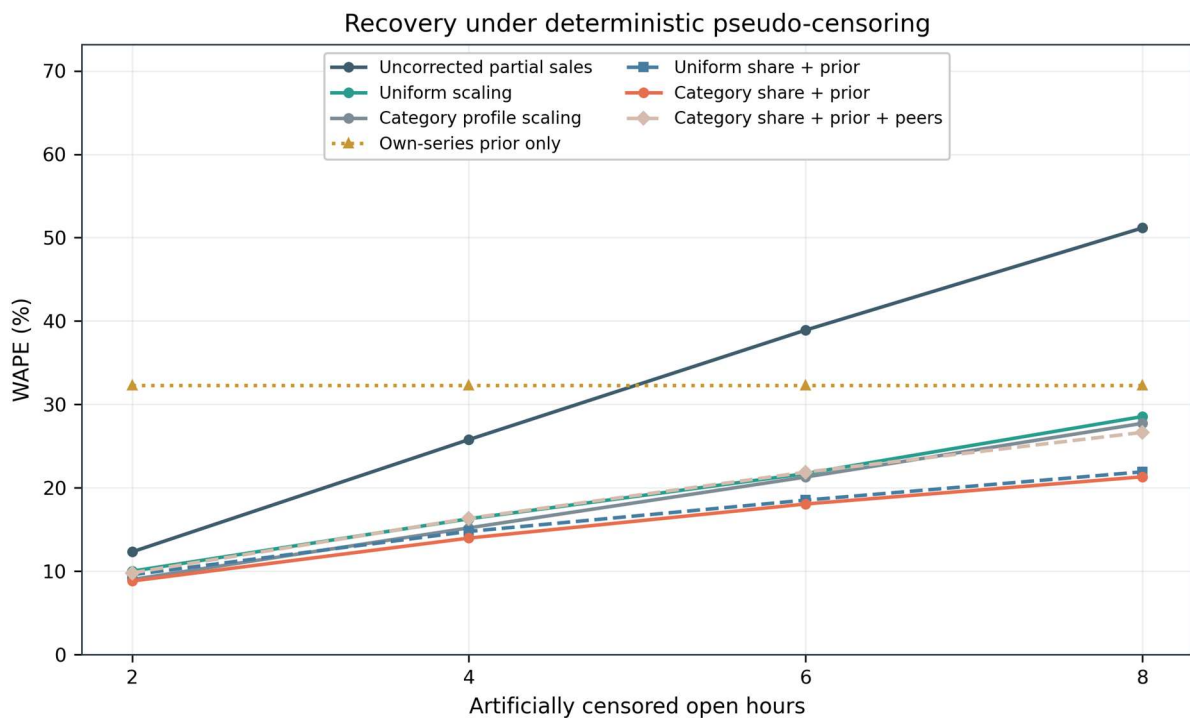


Figure 3. WAPE under fixed-seed contiguous-hour masking of observed sales. The own-prior and uniform-share controls isolate the two reconstruction components.

C. FreshRetailNet forecast comparison

Table 4 reports the primary comparison on 206,562 rows without labeled open-hour stockouts. Seasonal naive achieves WAPE 38.91%; the rolling seven-day mean and Croston-SBA achieve 31.17% and

32.50%. Ridge regression reaches 30.43%, close to observed-sales HGB at 30.54%. These scores concern observable sales within the selected subset.

Table 4. FreshRetailNet forecasting on rows without labeled open hour stockouts

Model	RMSE	MAE	WAPE	Bias	Mean RMSSE
Seasonal naive	0.7979	0.4813	38.91%	-0.77%	0.9861
Seven-day mean	0.6408	0.3856	31.17%	-1.50%	0.7949
Croston-SBA	0.7261	0.4021	32.50%	-10.42%	0.8023
Ridge (observed sales)	0.6911	0.3765	30.43%	-1.87%	0.7652
HGB (observed sales)	0.7617	0.3778	30.54%	-2.11%	0.7529
HGB + recovery	0.6914	0.3816	30.84%	8.33%	0.7788
HGB + recovery + exogenous	0.6740	0.3736	30.20%	7.08%	0.7603
HGB + graph + exogenous, observed history	0.7181	0.3698	29.89%	-2.22%	0.7431
SARF (uncalibrated)	0.6693	0.3735	30.19%	8.43%	0.7649
SARF with scale calibration	0.6761	0.3621	29.27%	2.48%	0.7366

Recovery alone reduces RMSE from 0.7617 to 0.6914, but increases WAPE from 30.54% to 30.84% and introduces 8.33% positive bias. Adding scheduled exogenous variables lowers WAPE to 30.20%. The observed-history graph model with the same exogenous variables reaches 29.89%; this comparison includes both graph and exogenous inputs relative to the plain observed-sales model. Within the matched recovery-plus-exogenous pair, adding graph features changes WAPE from 30.1994% to 30.1928%, only 0.0066 percentage points. The graph's incremental absolute-error contribution is small in this run.

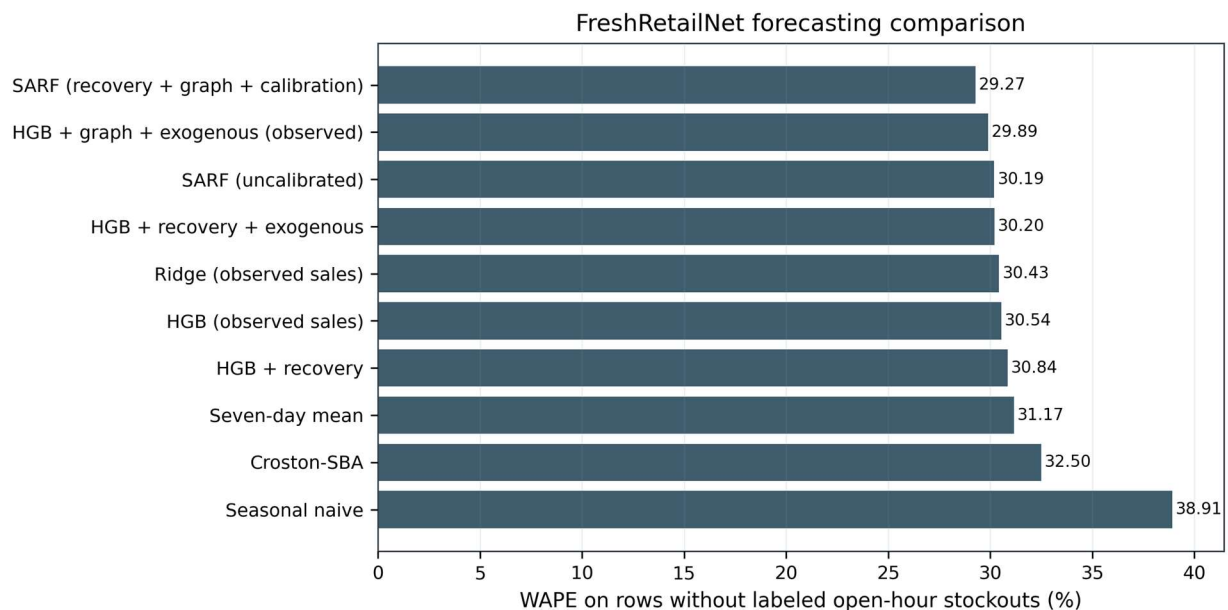


Figure 4. Primary FreshRetailNet WAPE across forecasting methods. Absolute errors use the released sales scale.

The frozen scale-calibration multiplier lowers full-model WAPE from 30.19% to 29.27% and bias from 8.43% to 2.48%, while increasing RMSE from 0.6693 to 0.6761. Calibrated SARF has MAE 0.3621 and mean RMSSE 0.7366. Its WAPE is 4.16% below observed-sales HGB and 24.77% below seasonal naive. Figure 4 shows the complete comparison. The combined gain should not be assigned to recovery or graph features alone, because the ablations reveal different effects on scale, squared error, and absolute error.

Calibrated SARF has lower WAPE than observed-sales HGB on all 7 official evaluation dates, ranging from 27.58% to 30.31% (Table 5 and Figure 5). Each origin incorporates observations from completed previous days, so this is a rolling next-day test rather than a recursive seven-step forecast. Agreement across one week does not establish longer-term temporal stability.

Table 5. FreshRetailNet performance by rolling evaluation date

Date	SARF WAPE	HGB WAPE	Naive WAPE	SARF RMSE	HGB RMSE
2024-06-26	29.96%	31.04%	38.08%	0.6202	0.7036
2024-06-27	29.31%	30.88%	38.56%	0.6005	0.6932
2024-06-28	29.38%	30.08%	41.37%	0.5338	0.5567
2024-06-29	27.58%	28.75%	38.49%	0.5808	0.6301
2024-06-30	28.88%	30.30%	36.60%	0.9113	1.0036
2024-07-01	30.16%	30.82%	38.90%	0.6736	0.7749
2024-07-02	30.31%	32.72%	40.67%	0.8189	0.9699

The mean per-series MAE difference for calibrated SARF minus observed-sales HGB is -0.01940, with a 95% conditional series-bootstrap interval of [-0.02122, -0.01771]. SARF has lower MAE on 53.31% of series. The Wilcoxon diagnostic gives $p < .001$. These summaries indicate a favorable average change with heterogeneous series-level effects, subject to dependence not captured by independent series resampling.

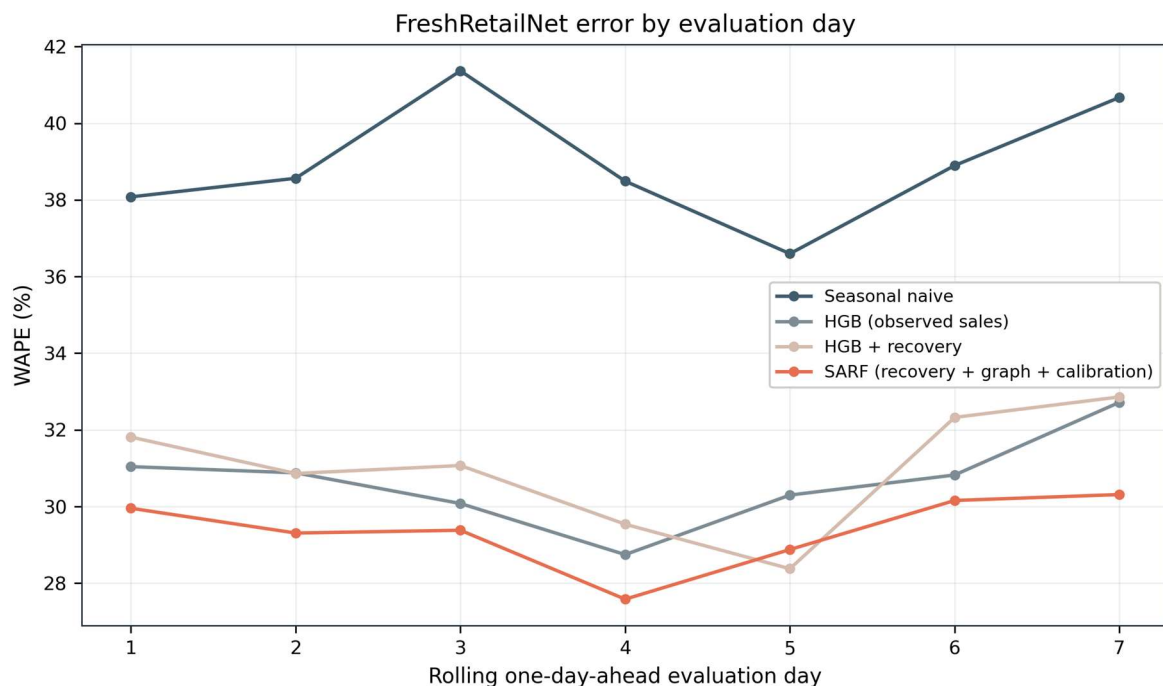


Figure 5. FreshRetailNet WAPE by rolling one-day-ahead evaluation date.

The additional forward windows provide a short-range temporal check (Table 6). Calibrated SARF improves WAPE in both June windows, but the uncalibrated model is worse than observed-sales HGB on WAPE in both. The scale multipliers also differ across windows. Calibration therefore matters to the observed absolute-error gain, and its stability should be assessed over a longer history before operational

use. Noted that for Table 6, all windows span seven days. Each uses only rows without labeled open-hour stockouts for scoring. The calibration interval precedes the scored window but shares its pre-window representation-fitting period; it is a scale-adjustment interval, not a fully isolated validation set.

Table 6. Forward window WAPE with representations fitted before each window

Window	Eligible rows	Observed HGB	Uncalibrated SARF	Calibrated SARF	Multiplier
2024-06-05 to 2024-06-11	206,229	34.00%	35.39%	32.35%	0.86004
2024-06-12 to 2024-06-18	196,407	32.34%	33.70%	31.21%	0.87414
2024-06-26 to 2024-07-02	206,562	30.54%	30.19%	29.27%	0.94514

On all 350,000 official evaluation rows, including naturally censored days, calibrated SARF has WAPE 33.86% versus 33.71% for observed-sales HGB. This comparison uses recorded sales as its target and therefore penalizes some upward demand adjustments. It does not identify which model better predicts latent demand under natural stockouts.

D. Relational Robustness and Natural Stockout Implications

City-level WAPE improves in 17 of 18 cities. Relative improvement has mean 4.40%, median 3.47%, and an interquartile range 2.35%–5.62% (Table 7). The range extends from -2.26% to 16.47%. This variation describes the combined calibrated model rather than the isolated contribution of the graph. Figure 6 compares city-level improvement with training stockout prevalence. Cities with similar exposure can have different changes in error. This descriptive comparison does not identify whether censoring, assortment, promotion mix, or cross-series dependence caused the differences.

Table 7. Distribution of city level relative WAPE improvement

Statistic	Value
Cities with improvement	17 / 18
Mean	4.40%
Median	3.47%
25th percentile	2.35%
75th percentile	5.62%
Minimum	-2.26%
Maximum	16.47%

Table 8 reports model-implied forecast–sales gaps by natural stockout exposure. For 13–16 affected hours, recorded sales average 0.1943 and the forecast averages 1.2184, giving a 527.01% gap. The large percentage partly reflects the very small observed-sales denominator. It is neither a measured loss nor validated recovered demand.

Table 8. Model implied forecast sales gaps on natural stockout bands

Stockout hours	Rows	Observed mean	Forecast mean	Implied gap
0	206,562	1.2371	1.2678	2.48%
1-4	46,283	1.5184	1.4244	-6.18%
5-8	49,253	1.2557	1.2954	3.16%
9-12	28,673	0.9128	1.1995	31.41%
13-16	19,229	0.1943	1.2184	527.01%

Natural counterfactual demand remains unobserved. These forecast–sales gaps describe model behavior as availability declines; they cannot determine recovery accuracy. To test that interpretation directly, we

would need inventory positions, replenishment timing, substitutions, or recorded unmet orders. For Table 8, values are on the FreshRetailNet released sales scale. The percentage is $(\text{aggregate forecast} / \text{aggregate recorded sales} - 1) \times 100$. Natural unmet demand is not observed; these gaps are not validated recovery estimates. Figure 7 summarizes the fitted graph relations. Mean history correlation is 0.4024 for FreshRetailNet and 0.1051 for M5.

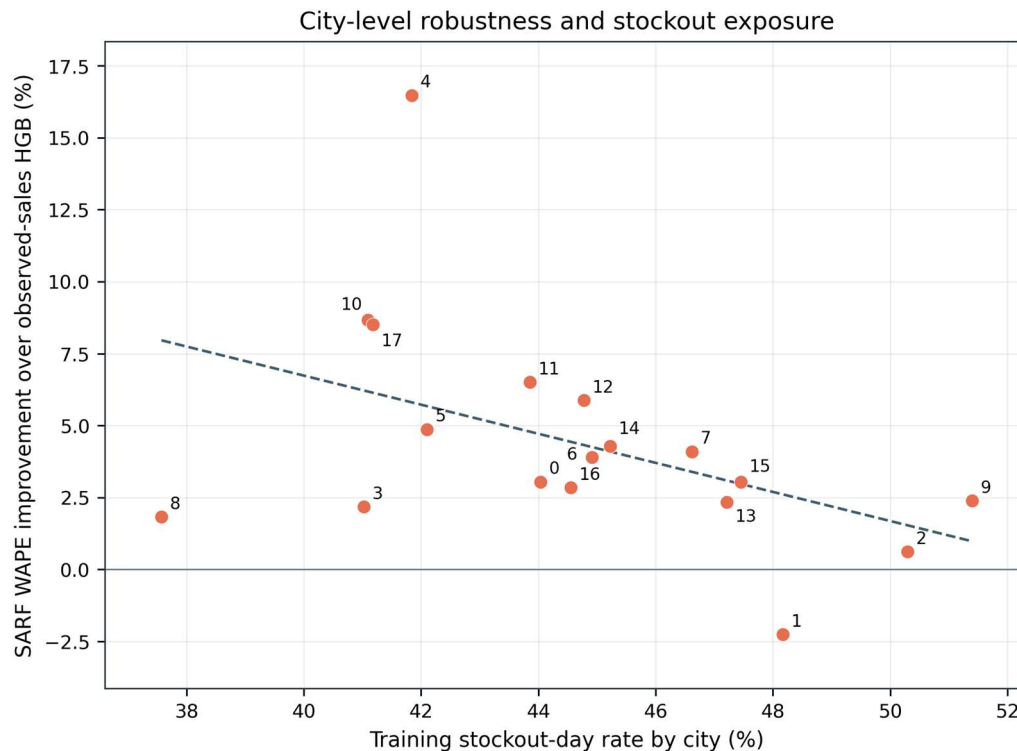


Figure 6. City-level association between training stockout prevalence and calibrated SARF WAPE improvement over observed-sales HGB.

The corresponding positive-correlation shares are 96.25% and 84.16%. Fresh statistics include its self-padding slots, while all 152,450 M5 slots are external neighbors. The weaker M5 correlations are consistent with limited relational gains, but do not establish why a particular model succeeds or fails.

E. M5 External Transfer

Table 9 shows that transfer changes the balance between squared and absolute error. Seasonal naive has WAPE 86.27%; the fixed 28-day mean and Croston-SBA reach 73.86% and 73.30%. Their competitiveness is consistent with M5's high zero-sales prevalence and the usefulness of a stable local demand level. Zero-shot source reuse gives RMSE 2.1536 and WAPE 76.74%, below the target-trained boosters in absolute-error performance. M5-only nonrelational HGB has RMSE 2.1035 and WAPE 71.60%; adding the target graph changes these to 2.1075 and 71.67%. Thus, the graph does not improve aggregate target-only WAPE or RMSE. Statistical similarity between sparse sales histories may add little beyond own-series lags, and can introduce redundant or noisy predictors; the present comparisons do not distinguish these explanations.

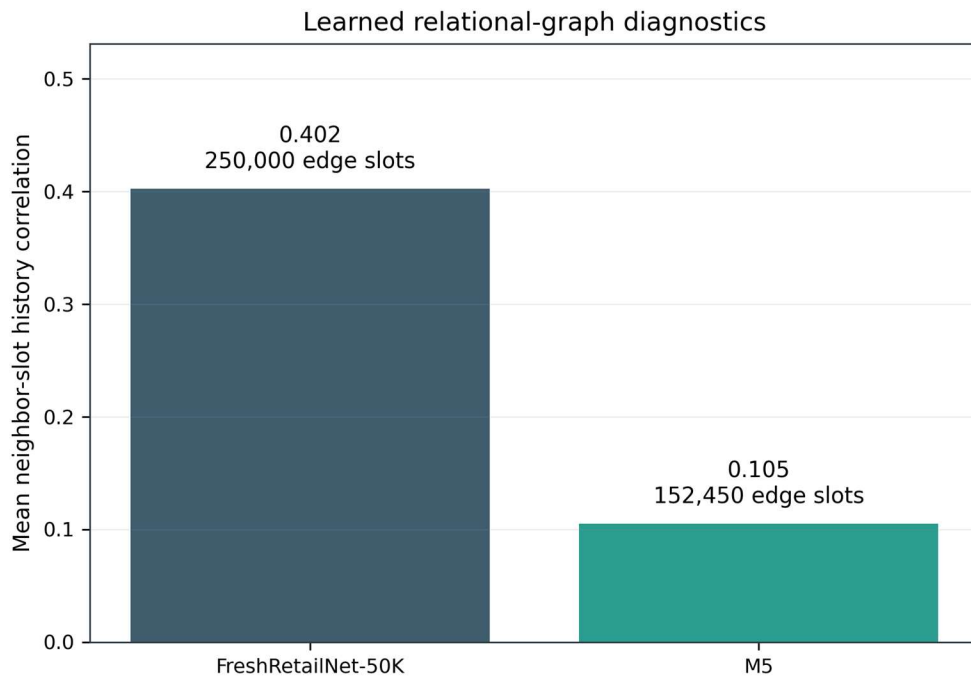


Figure 7. Mean training-history correlation across directed neighbor slots. FreshRetailNet counts include self-padding; M5 counts are external neighbors.

With all target examples retained, 1:1 source augmentation yields relational-model RMSE 2.0706, MAE 1.0403, and WAPE 72.10%. Relative to target-only relational HGB, RMSE improves by 1.75%, while WAPE increases by 0.43 percentage points. Under a fixed total budget, 1:1 pooling yields RMSE 2.0716 and WAPE 72.21%, a 1.70% RMSE improvement and a 0.54-point WAPE increase. Figure 8 displays WAPE for the main fixed-budget comparison. The nonrelational M5-only model retains the lowest WAPE. Noted that for Table 9, each model is evaluated on 853,720 observed sales outcomes. Augmented 1:1 models use 3,170,960 training rows; fixed-budget 1:1 models use 1,585,480. M5 zeros are not labeled as stockouts.

Table 9. M5 bottom level forecasts over 28 rolling one day ahead origins

Model	RMSE	MAE	WAPE	Bias	Mean RMSSE
Seasonal naive	2.6815	1.2447	86.27%	-1.49%	1.2170
Fixed 28-day mean	2.2430	1.0657	73.86%	-3.91%	0.9240
Croston-SBA	2.1683	1.0576	73.30%	-5.67%	0.9207
Fresh zero-shot relational	2.1536	1.1072	76.74%	7.91%	0.9400
M5-only HGB	2.1035	1.0330	71.60%	-6.15%	0.9054
M5-only relational HGB	2.1075	1.0341	71.67%	-5.83%	0.9051
Augmented 1:1 nonrelational	2.0650	1.0463	72.52%	2.33%	0.9096
Augmented 1:1 relational	2.0706	1.0403	72.10%	0.16%	0.9043
Fixed-budget 1:1 nonrelational	2.0639	1.0429	72.28%	1.57%	0.9057
Fixed-budget 1:1 relational	2.0716	1.0418	72.21%	0.31%	0.9045

The fixed-budget 1:1 relational comparison has a mean per-series MAE difference of 0.00772, with a 95% conditional bootstrap interval [0.00652, 0.00892]; 33.88% of series improve. The augmented model gives 0.00619 [0.00507, 0.00731], with 34.50% improving. Both intervals favor target-only

learning for average absolute error. Lower pooled RMSE therefore does not imply broad series-level improvement.

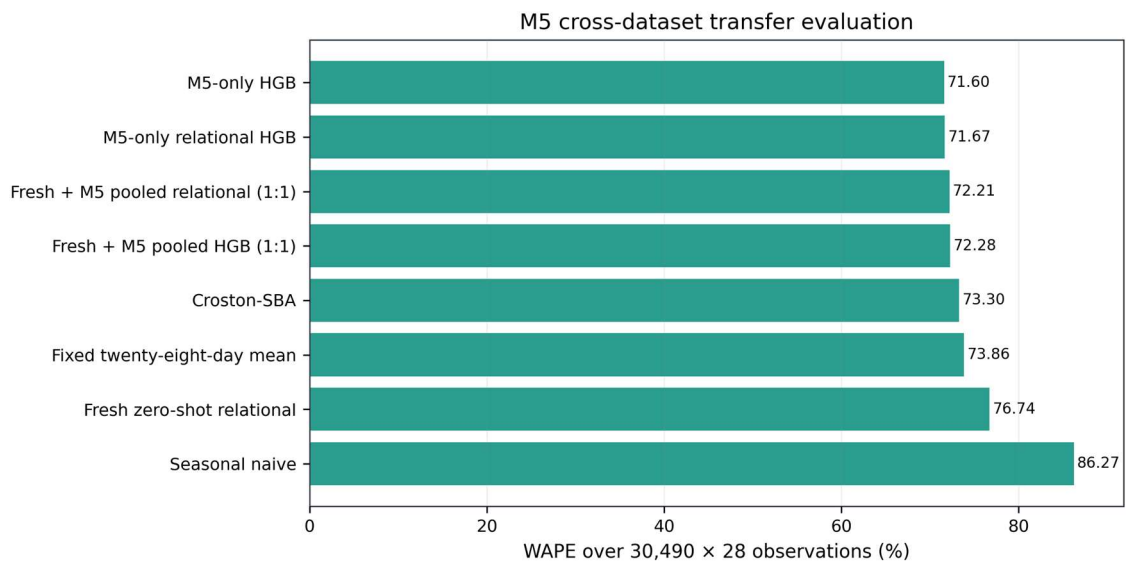


Figure 8. M5 WAPE across 853,720 outcomes. Pooled bars use the fixed-budget 1:1 protocol.

Table 10 varies source:target composition while keeping the total training budget fixed. Every nonzero source ratio lowers RMSE relative to the corresponding target-only feature set, while every ratio increases WAPE. Within the relational family, the smallest pooled WAPE occurs at 1:4 and the smallest RMSE at 2:1. The relative ordering of graph and nonrelational variants also changes with the ratio. These results support an error tradeoff over the tested range, rather than a uniquely favorable 1:1 balance or a general advantage from adding relational features. Noted that for the table 10, all models use 55 boosting iterations. Source and target rows are sampled without replacement using fixed seeds. Increasing source share removes target examples under this budget; it is not equivalent to holding all target examples fixed and adding source data.

Table 10. Source target ratio sensitivity at a fixed training budget

Ratio	Source rows	Target rows	No graph RMSE	No graph WAPE	Graph RMSE	Graph WAPE
0:1	0	1,585,480	2.1035	71.60%	2.1075	71.67%
1:4	317,096	1,268,384	2.0658	71.98%	2.0774	71.92%
1:2	528,493	1,056,987	2.0647	72.19%	2.0734	72.02%
1:1	792,740	792,740	2.0639	72.28%	2.0716	72.21%
2:1	1,056,987	528,493	2.0639	72.58%	2.0638	72.10%

Table 11 localizes the fixed-budget 1:1 comparison by category and store. RMSE improves in all three categories and all ten stores, but its category reduction is largest for FOODS and nearly zero for HOBBIES. WAPE worsens in every category and improves only at CA_2 among stores. The augmented model has the same qualitative WAPE pattern. Pooling therefore affects particular domains and error magnitudes differently; the aggregate RMSE gain is not a domain-wide absolute-error gain. Table 12 shows that the dense group has pooled WAPE 48.26% versus 48.05% for target-only relational HGB.

Pooled WAPE is also higher in the intermediate, sparse, and very sparse groups. Croston-SBA reaches 127.65% in the very sparse group, below pooled WAPE of 131.93%. The squared-error pattern differs: pooled RMSE improves by 2.95% among dense series, compared with 0.012% among very sparse series.

Table 11. M5 category and store comparisons for fixed budget 1 to 1 relational pooling

Domain	Series	Target RMSE	Pooled RMSE	Target WAPE	Pooled WAPE	MAE wins
FOODS	14,370	2.6206	2.5636	64.51%	64.91%	37.06%
HOBBIES	5,650	1.6809	1.6808	102.60%	103.93%	27.63%
HOUSEHOLD	10,470	1.4085	1.3974	80.06%	80.67%	32.87%
CA_1	3,049	2.1966	2.1489	72.29%	72.54%	36.93%
CA_2	3,049	2.0877	2.0333	73.07%	72.81%	44.77%
CA_3	3,049	2.5848	2.5644	65.46%	66.37%	34.54%
CA_4	3,049	1.4222	1.4200	86.14%	87.34%	27.58%
TX_1	3,049	1.7191	1.6739	75.16%	75.66%	30.60%
TX_2	3,049	1.9064	1.8812	72.27%	73.31%	32.80%
TX_3	3,049	2.0233	1.9872	71.60%	72.10%	30.14%
WI_1	3,049	1.7012	1.6684	73.43%	73.64%	37.29%
WI_2	3,049	2.9340	2.8933	65.64%	66.22%	30.14%
WI_3	3,049	2.0837	2.0315	71.74%	72.31%	33.98%

One interpretation is that FreshRetailNet contributes useful mappings for the larger positive-demand fluctuations present in dense target series, while its relatively low zero frequency provides less information about occurrence dynamics in very sparse series. This is an empirical interpretation of error patterns, not a demonstrated transfer mechanism. No sparsity group shows a pooled WAPE advantage in this comparison, and a routing rule between pooled and intermittent-demand models would require separate development and evaluation.

Table 12. M5 WAPE by training zero sales prevalence

Zero-sales segment	Series	Pooled graph	M5 graph	Croston-SBA	Fresh zero-shot
Dense (<30%)	2,247	48.26%	48.05%	49.27%	51.86%
Intermediate (30–70%)	11,414	71.41%	70.95%	73.20%	75.51%
Sparse (70–90%)	11,545	102.56%	101.64%	102.96%	109.01%
Very sparse (≥90%)	5,284	131.93%	129.70%	127.65%	140.89%

F. Validity and Practical Interpretation

Four limitations bound the conclusions. First, FreshRetailNet provides availability labels but no counterfactual demand during natural stockouts; zero labeled open-hour exposure also does not certify availability outside 06:00–22:00. Artificial masking measures omitted-sales reconstruction under a controlled mechanism. Second, all FreshRetailNet entities occur in both splits, so the study evaluates temporal generalization rather than cold starts. Third, the official source test lasts seven days and the two added windows remain within June. They do not assess seasonal drift, changing promotion regimes, or calibration stability over longer periods. Fourth, M5 evaluates rolling daily bottom-level sales forecasts without stockout labels, inventory costs, or service outcomes.

The graph is a statistical relation among demand histories within observed retail groups. It does not encode travel distance, inventory-transfer feasibility, or communication between autonomous agents. Its small incremental source WAPE gain and unfavorable target-only M5 comparison limit claims that relational features are generally beneficial. In practice, hourly sales and availability records can be processed after each day closes, followed by target reconstruction and next-day forecasting using lagged histories. Scheduled promotions must be available at issuance for the corresponding features to be valid. Profiles, graph relations, and the global scale multiplier require monitoring when demand patterns change (Zhang, 2023). Inventory optimization could consume these forecasts, but its cost and service effects require a separate simulation or field study, consistent with the forecasting–inventory distinction in prior work (Goltsos et al., 2022).

IV. CONCLUSION

Profile-weighted temporal reconstruction reduced WAPE under artificial contiguous-hour masking of real FreshRetailNet sales, reaching 8.82%–21.32% across two- to eight-hour masks. The own-prior and uniform-share controls show that remaining same-day sales, a temporal demand scale, and intraday allocation each matter in this diagnostic. The experiment does not establish the accuracy of latent demand recovered during natural stockouts. On the seven-day official subset of 206,562 rows without labeled open-hour stockouts, calibrated SARF achieved RMSE 0.6761, MAE 0.3621, and WAPE 29.27%, compared with 30.54% WAPE for observed-sales HGB. Recovery alone did not improve WAPE, the incremental graph effect was small, and scale calibration contributed substantially to the final absolute-error reduction. Two earlier seven-day windows showed the same calibrated comparison direction, while remaining too short to support seasonal stability claims.

M5 evaluation covered 30,490 series at 28 rolling daily origins. Fixed-budget 1:1 relational pooling lowered RMSE from 2.1075 to 2.0716, but increased WAPE from 71.67% to 72.21%. Full-target augmentation produced the same tradeoff, and the ratio analysis showed no universal advantage for 1:1 pooling. Dense series had the strongest squared-error benefit, while all four sparsity groups had higher pooled WAPE. Category and store results further showed that benefits depend on the error measure and domain. These findings support evaluating stockout reconstruction, observable-sales forecasting, and cross-dataset transfer as distinct tasks. The available evidence favors a calibrated forecasting combination within the observed source windows and shows metric-dependent transfer effects on M5. Natural unmet demand and downstream inventory outcomes remain separate validation questions.

DATA AVAILABILITY STATEMENT

FreshRetailNet-50K is available from Dingdong-Inc on Hugging Face (Dingdong-Inc, 2025). M5 sales, calendar, and price data are available through the M5 Forecasting—Accuracy competition (University of Nicosia, 2020). The accompanying supplementary analysis package contains the experiment scripts,

input-file checksums, fitted models, prediction outputs, summary tables, and figure-generation code. The two benchmarks retain their distinct roles: FreshRetailNet supplies hourly availability labels, whereas M5 supplies observed sales only.

AI USAGE AND GENERATIVE AI DECLARATION

ChatGPT (OpenAI) assisted with analytical code, visualizations, and manuscript preparation.

REFERENCES

- Ban, G.-Y. (2020). Confidence intervals for data-driven inventory policies with demand censoring. *Operations Research*, 68(2), 309–326. <https://doi.org/10.1287/opre.2019.1883>
- Besbes, O., & Muharremoglu, A. (2013). On implications of demand censoring in the newsvendor problem. *Management Science*, 59(6), 1407–1424. <https://doi.org/10.1287/mnsc.1120.1654>
- Croston, J. D. (1972). Forecasting and stock control for intermittent demands. *Operational Research Quarterly*, 23(3), 289–303. <https://doi.org/10.1057/jors.1972.50>
- Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7(1), 1–30. <https://jmlr.org/papers/v7/demsar06a.html>
- Ding, J., Huh, W. T., & Rong, Y. (2024). Feature-based inventory control with censored demand. *Manufacturing & Service Operations Management*, 26(3), 1157–1172. <https://doi.org/10.1287/msom.2021.0135>
- Ding, X., Puterman, M. L., & Bisi, A. (2002). The censored newsvendor and the optimal acquisition of information. *Operations Research*, 50(3), 517–527. <https://doi.org/10.1287/opre.50.3.517.7752>
- Dingdong-Inc. (2025). FreshRetailNet-50K (Version 1.0) [Data set]. Hugging Face. <https://huggingface.co/datasets/Dingdong-Inc/FreshRetailNet-50K>
- Fildes, R., Ma, S., & Kolassa, S. (2022). Retail forecasting: Research and practice. *International Journal of Forecasting*, 38(4), 1283–1318. <https://doi.org/10.1016/j.ijforecast.2019.06.004>
- Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., & Lempitsky, V. (2016). Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59), 1–35. <https://jmlr.org/papers/v17/15-239.html>
- Goltsos, T. E., Syntetos, A. A., Glock, C. H., & Ioannou, G. (2022). Inventory–forecasting: Mind the gap. *European Journal of Operational Research*, 299(2), 397–419. <https://doi.org/10.1016/j.ejor.2021.07.040>
- Huh, W. T., & Rusmevichientong, P. (2009). A nonparametric asymptotic analysis of inventory planning with censored demand. *Mathematics of Operations Research*, 34(1), 103–123. <https://doi.org/10.1287/moor.1080.0355>
- Hyndman, R. J., Ahmed, R. A., Athanasopoulos, G., & Shang, H. L. (2011). Optimal combination forecasts for hierarchical time series. *Computational Statistics & Data Analysis*, 55(9), 2579–2589. <https://doi.org/10.1016/j.csda.2011.03.006>
- Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4), 679–688. <https://doi.org/10.1016/j.ijforecast.2006.03.001>
- Jain, A., Rudi, N., & Wang, T. (2015). Demand estimation and ordering under censoring: Stock-out timing is (almost) all you need. *Operations Research*, 63(1), 134–150. <https://doi.org/10.1287/opre.2014.1326>
- Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282), 457–481. <https://doi.org/10.1080/01621459.1958.10501452>

- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In I. Guyon, U. von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 30, pp. 3146–3154). Curran Associates. <https://papers.nips.cc/paper/6907-lightgbm-a-highly-efficient-gradient-boosting-decision-tree>
- Lu, X., Song, J.-S., & Zhu, K. (2008). Analysis of perishable-inventory systems with censored demand data. *Operations Research*, 56(4), 1034–1038. <https://doi.org/10.1287/opre.1080.0553>
- Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2022a). M5 accuracy competition: Results, findings, and conclusions. *International Journal of Forecasting*, 38(4), 1346–1364. <https://doi.org/10.1016/j.ijforecast.2021.11.013>
- Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2022b). The M5 competition: Background, organization, and implementation. *International Journal of Forecasting*, 38(4), 1325–1336. <https://doi.org/10.1016/j.ijforecast.2021.07.007>
- Mersereau, A. J. (2015). Demand estimation from censored observations with inventory record inaccuracy. *Manufacturing & Service Operations Management*, 17(3), 335–349. <https://doi.org/10.1287/msom.2015.0520>
- Montero-Manso, P., & Hyndman, R. J. (2021). Principles and algorithms for forecasting groups of time series: Locality and globality. *International Journal of Forecasting*, 37(4), 1632–1653. <https://doi.org/10.1016/j.ijforecast.2021.03.004>
- Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10), 1345–1359. <https://doi.org/10.1109/TKDE.2009.191>
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830. <https://jmlr.org/papers/v12/pedregosa11a.html>
- Semenoglou, A.-A., Spiliotis, E., Makridakis, S., & Assimakopoulos, V. (2021). Investigating the accuracy of cross-learning time series forecasting methods. *International Journal of Forecasting*, 37(3), 1072–1084. <https://doi.org/10.1016/j.ijforecast.2020.11.009>
- Shi, C., Chen, W., & Duenyas, I. (2016). Technical Note—Nonparametric data-driven algorithms for multiproduct inventory systems with censored demand. *Operations Research*, 64(2), 362–370. <https://doi.org/10.1287/opre.2015.1474>
- Snyder, R. D., Ord, J. K., & Beaumont, A. N. (2012). Forecasting the intermittent demand for slow-moving inventories: A modelling approach. *International Journal of Forecasting*, 28(2), 485–496. <https://doi.org/10.1016/j.ijforecast.2011.03.009>
- Syntetos, A. A., & Boylan, J. E. (2005). The accuracy of intermittent demand estimates. *International Journal of Forecasting*, 21(2), 303–314. <https://doi.org/10.1016/j.ijforecast.2004.10.001>
- Teunter, R. H., Syntetos, A. A., & Babai, M. Z. (2011). Intermittent demand: Linking forecasting to inventory obsolescence. *European Journal of Operational Research*, 214(3), 606–615. <https://doi.org/10.1016/j.ejor.2011.05.018>
- Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica*, 26(1), 24–36. <https://doi.org/10.2307/1907382>
- University of Nicosia. (2020). M5 forecasting—Accuracy [Data set]. Kaggle. <https://www.kaggle.com/competitions/m5-forecasting-accuracy>
- Wang, J., Xin, Q., Liu, Y., Wang, J., & Yang, T. (2024). Predicting enterprise marketing decision making with intelligent data-driven approaches. *Journal of Industrial Engineering and Applied Science*, 2(3), 12–19. <https://doi.org/10.5281/zenodo.11357252>

- Wang, Y., Gu, J., Long, L., Li, X., Shen, L., Fu, Z., Zhou, X., & Jiang, X. (2025). FreshRetailNet-50K: A stockout-annotated censored demand dataset for latent demand recovery and forecasting in fresh retail (Version 1) [Preprint]. arXiv. <https://arxiv.org/abs/2505.16319v1>
- Wecker, W. E. (1978). Predicting demand from sales data in the presence of stockouts. *Management Science*, 24(10), 1043–1054. <https://doi.org/10.1287/mnsc.24.10.1043>
- Wellens, A. P., Boute, R. N., & Udenio, M. (2024). Simplifying tree-based methods for retail sales forecasting with explanatory variables. *European Journal of Operational Research*, 314(2), 523–539. <https://doi.org/10.1016/j.ejor.2023.10.039>
- Wellens, A. P., Udenio, M., & Boute, R. N. (2022). Transfer learning for hierarchical forecasting: Reducing computational efforts of M5 winning methods. *International Journal of Forecasting*, 38(4), 1482–1491. <https://doi.org/10.1016/j.ijforecast.2021.09.011>
- Wickramasuriya, S. L., Athanasopoulos, G., & Hyndman, R. J. (2019). Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization. *Journal of the American Statistical Association*, 114(526), 804–819. <https://doi.org/10.1080/01621459.2018.1448825>
- Wu, Z., Pan, S., Long, G., Jiang, J., Chang, X., & Zhang, C. (2020). Connecting the dots: Multivariate time series forecasting with graph neural networks. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 753–763). Association for Computing Machinery. <https://doi.org/10.1145/3394486.3403118>
- Wu, Z., Pan, S., Long, G., Jiang, J., & Zhang, C. (2019). Graph WaveNet for deep spatial-temporal graph modeling. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence* (pp. 1907–1913). International Joint Conferences on Artificial Intelligence Organization. <https://doi.org/10.24963/ijcai.2019/264>
- Zhang, H. (2023). DriftGuard: Multi-signal drift early warning and safe re-training/rollback for CTR/CVR models. *Journal of Advanced Computing Systems*, 3(7), 24–40. <https://doi.org/10.69987/JACS.2023.30703>