

Forecast Value Is a Property of the Decision: Calibrated Quantile Demand Forecasting for Urban Last-Mile Capacity and Routing

Kien Xuan Pham¹, Nhat Trinh Ngoc Minh², Linh Nguyen Thi My^{2,*}

¹*Faculty of Management Information Systems, Ho Chi Minh University of Banking, Ho Chi Minh City, Vietnam*

²*Faculty of Information Technology, Van Lang School of Technology, Van Lang University, Ho Chi Minh City, Vietnam*

*Corresponding author: linh.ntm@vlu.edu.vn

Abstract. Forecasting studies routinely read lower prediction error as operational value. For service capacity planning that inference is unsafe: the cost-minimising commitment is a demand quantile whose level is the target service level, so a median forecast silently commits an operator to a 50% target. We test this over 263 New York City zones and 39.0 million taxi trips from two services with contrasting demand geographies. Spatio-temporal graph networks, a multi-relational graph-attention model and Graph WaveNet, are trained on seasonal residuals with a monotone quantile head under a pinball objective, then calibrated by split conformalised quantile regression on a later held-out block. The same forecasts feed two decisions that consume different functionals: newsvendor capacity commitment under a 9:1 underage-to-overage penalty, which needs a calibrated tail quantile, and top-K zone selection, which needs only a ranking. Evaluation uses chronological splits, multiple seeds and Diebold–Mariano tests. The two decisions give opposite answers. Committing the calibrated 0.9 quantile instead of the median lowers capacity cost by 42–51% and raises the achieved fill rate from 94.9% to 98.9% on one service and from 86.8% to 97.4% on the other, because the median is mis-specified for an asymmetric penalty rather than inaccurate; zone selection improves by under 0.15 percentage points over a weekly historical average. The lowest-error model is not the least costly, no architecture ranks stably across the two regimes, and no hand-designed graph relation replicates. The calibration layer transfers unchanged between backbones and reproduces 40–54% cost reductions on external taxi, parcel, food-delivery and courier corpora. The appropriate unit of evaluation is the forecast–decision pair.

1. INTRODUCTION

Urban demand forecasts are routinely justified by operational narratives: pre-positioning vehicles, assigning couriers, committing capacity, or choosing areas to serve. Yet the empirical evidence

Received: Aug. 7, 2026.

2020 *Mathematics Subject Classification.* 62M20; 90B06; 62P30.

Key words and phrases. service level; probabilistic forecasting; conformal prediction; newsvendor; capacity planning; last-mile decision support; forecast value; spatio-temporal graph neural networks.

is usually a table of MAE or RMSE. This creates a gap between what a model optimises and what an operator ultimately uses. A lower average point error need not improve a decision when the decision depends on a tail quantile, a ranking, or another functional of the predictive distribution. The principle behind that statement is not new. Gneiting [23] showed that comparing point forecasts by an error measure can lead to badly misguided inferences unless the scoring function and the forecasting task are matched, and that a point forecast is well posed only relative to a scoring function fixed in advance, or equivalently to a directive naming the functional to be reported — the mean, a quantile, or another summary of the predictive distribution. What is missing in this domain is not the principle but a measurement: how large the effect is once realised operating cost is computed, and whether it is large for every decision or only for some.

This distinction is especially important in last-mile systems. Committing too little capacity can leave demand unserved, while committing too much creates idle resources. The resulting asymmetric loss is minimised by a conditional quantile, not by a conditional mean or median. By contrast, selecting a fixed number of zones before routing mostly depends on their ordering. A forecast may therefore be valuable for one decision and nearly irrelevant for another even when both decisions use the same zones, hours, and realised demand.

The forecasting literature has developed increasingly sophisticated spatial models. Taxi-demand studies combine temporal, local-spatial, and semantic views [1], while multi-graph models encode several non-Euclidean relations between regions [2]. Traffic forecasting has similarly progressed through STGCN [3], DCRNN [4], and Graph WaveNet [5]. These architectures answer an important representation question, but their predictive gains do not by themselves establish decision value. Predict-then-optimize research makes the complementary point that a prediction loss can be misaligned with a downstream objective [6]. We bring these literatures together through a deliberately simple question: *what does each decision actually consume from a forecast?*

We answer this question on two public NYC taxi services. Yellow and green taxis use the same 263-zone geography and the same evaluation period, but green taxis are restricted in where they may accept street hails in Manhattan [7]. Their spatial demand distributions are therefore sufficiently different to test whether a finding replicates across demand regimes without changing the graph support.

The paper contributes on three fronts. The first is a measurement. That a point forecast is optimal only for the functional its scoring function is matched to is Gneiting's principle [23]; what has not been established for this domain is how large the effect is in realised operating cost, or whether it is large for every decision. We evaluate identical forecast series in two last-mile decisions that consume different functionals — a 0.9 conditional quantile for service capacity commitment, a ranking for zone selection — and obtain opposite conclusions. Decision-aligned quantiles reduce asymmetric cost by 42–51% and raise the achieved fill rate by up to 10.6 percentage points, whereas learned forecasts improve zone selection by under 0.15 percentage points over a week-hour historical average. The capacity effect replicates on both demand regimes and on all

five external corpora, and the zone-selection null replicates on both regimes. The model with the lowest error is not the model with the lowest cost.

The second concerns what actually transfers between settings, and it is the decision layer rather than the architecture. A monotone quantile head with conformal calibration moves unchanged from a multi-relational GAT to Graph WaveNet, reaching nominal 80% coverage on two demand regimes and two backbones without retuning. Architectural rankings, by contrast, are unstable across regimes, and no hand-designed graph relation produces a replicable accuracy benefit.

The third is external validation of the mechanism together with an explicit account of its limits. Evidence spans a second city (Porto), real parcel order arrivals (Olist), real routing objects with measured vehicle capacity (Amazon), a same-order dispatch-to-delivery record (Meituan), a 184-day parcel history (Cainiao-LaDe), and an execution-noise bound from 19,647 observed driver routes. Section 7.9 sets out, as an auditable source matrix, why no public source combines order creation, physical capacity, and continuous route traces in one record.

Two of these findings are negative, and they are reported because they change the paper's conclusion rather than despite doing so: sophisticated forecasts add almost nothing to zone selection, and relation-level attention assigns its largest weight to the relation ablation finds least useful. The central claim is not that one network wins. It is that forecast value is a property of the forecast–decision pair.

2. RELATED WORK

2.1. Urban demand and spatio-temporal graph forecasting. Deep demand forecasting methods seek to represent temporal periodicity and spatial interaction jointly. DMVST-Net combines temporal, spatial, and semantic views for taxi demand [1]. ST-MGCN encodes several region-to-region graphs for ride-hailing demand [2], motivating our test of geographic, road-network, and origin–destination relations. In traffic forecasting, STGCN uses graph and temporal convolutions [3], DCRNN models propagation as diffusion over a directed graph [4], and Graph WaveNet learns an adaptive adjacency matrix together with dilated temporal convolutions [5]. Later work pushes the learned-adjacency idea further: MTGNN extracts relations among the series from the data itself [24], and AGCRN argues that node-specific parameter learning makes a pre-defined graph avoidable altogether [25]. Other designs model spatial and temporal dependence synchronously rather than in separate components [26], and grid-based residual convolutions forecast citywide crowd inflow and outflow without a graph at all [27]. Graph Attention Networks provide a general mechanism for learned neighbour weighting [8], although the attention they compute is static in the sense that the ranking of attention scores does not depend on the query node [28]. Our remaining graph baselines, spectral graph convolution [29] and inductive neighbourhood aggregation [30], were developed for node classification rather than for forecasting.

Most of this literature selects models using predictive metrics, as does demand forecasting more broadly: the M5 accuracy competition ranked submitted point forecasts of a large retailer's

hierarchical unit sales [31], and the choice among error measures is a studied question in its own right [32]. Two limits on what such a comparison can settle matter here. Small gaps between neural models can be produced by the random seed alone, as has been documented for sequence tagging, where comparing score distributions over repeated runs is proposed in place of single scores [33]; the wider case for improved reporting of experimental results is made in [34]. And an error measure answers a question about error, not about a decision. We instead treat the architecture as one layer in a larger decision-support pipeline. This changes the relevant comparison: a small MAE difference matters only if it changes the functional consumed by the operational decision.

2.2. Probabilistic and conformal forecasting. Quantile regression directly estimates conditional distributional functionals and is trained with the pinball loss [9], the asymmetric piecewise-linear scoring function whose optimal point forecast is the corresponding quantile [23]. Deep forecasters return predictive distributions rather than single values: DeepAR trains an autoregressive recurrent network across many related series and estimates the series' future probability distribution [35]. Attention-based architectures address the same multi-horizon setting, the Temporal Fusion Transformer using recurrent layers for local processing and self-attention for long-term dependencies [36]. Probabilistic forecasts of this kind are assessed with proper scoring rules, under which the forecaster maximises its expected score by issuing its actual predictive distribution [37]. Because neural quantiles can under-cover, conformalised quantile regression (CQR) uses held-out nonconformity scores to widen an estimated interval [10]. Under exchangeability, CQR provides finite-sample marginal coverage, inheriting the guarantee that split conformal prediction supplies for regression [38]. Hourly demand is temporally dependent, so we do not invoke that guarantee unconditionally; instead, we use a later contiguous calibration block and report realised coverage on a strictly later test block. Adaptive conformal methods retain a target coverage frequency when the data-generating distribution drifts and exchangeability fails [39]; we do not use them here. The distinction between formal exchangeable coverage and empirical time-series coverage is retained throughout.

2.3. Prediction for downstream decisions. Smart predict-then-optimize formalises the mismatch between minimising prediction error and minimising the cost induced by a prediction [6]. Decision-focused learning extends this principle by training through or around a downstream optimiser: Donti et al. learn probabilistic models end to end against the task objective of the stochastic program they feed, one of which is a classical inventory stock problem [40]. A related line removes the intermediate forecast altogether, prescribing decisions directly from historical observations and associated auxiliary quantities [41], or solving the newsvendor problem as a feature-based ordering rule rather than as a demand estimate [42]. Our goal is narrower and diagnostic: we do not differentiate through routing or capacity allocation, and we keep the forecast explicit, so that one forecast series can be read by two decisions that consume different functionals. We identify the functional each decision consumes, train that functional where needed, and evaluate realised decision cost.

For capacity commitment, the classical newsvendor result gives a particularly clean link: the optimal decision is the critical-ratio quantile. For fixed-cardinality zone selection, only the ordering of predictions determines the chosen set. The two tasks therefore expose different forms of forecast value without changing the underlying test observations.

Work on last-mile operations has developed mainly on the optimisation side of this boundary. Surveys catalogue delivery concepts together with the decision problems each one raises [43] and trace three decades of dynamic vehicle routing [44], while network-design studies size facilities and fleets for urban logistics [45] and examine the value of physical distribution flexibility in dense and uncertain urban markets [46]. Where demand is uncertain, the uncertainty is usually absorbed inside the optimiser: an a priori route is chosen to minimise expected length under a probabilistic description of demand [47], or a two-stage formulation charges a penalty for the recourse trip incurred when realised demand exceeds vehicle capacity [48]. Same-day delivery work extends the same treatment to uncertain delivery capacity, dispatching under probabilistic information about future order and courier arrivals [49]. These formulations take a demand distribution as given. Our question sits one step earlier and is deliberately smaller: which single summary of that distribution a forecasting system should be judged on, and how much the answer changes realised cost. Execution adds a third layer that neither question reaches — planned routes are not the routes drivers take, and the deviations are regular enough that machine-learning models can predict and quantify them from route structure, route attributes and driver information [50]; we use observed deviations only as a bound on execution noise.

2.4. Attention and explanation. Attention weights are often presented as indicators of input importance. Prior work shows that this interpretation is not automatically faithful [11]; manipulating the attention weights of already-trained models and observing the change in their predictions indicates that attention magnitude is at best a noisy indicator of a component's importance [51]. Subsequent work argues that validity depends on the definition and tests used [12]. We take a minimal empirical position: relation attention describes the model's internal mixture, whereas component utility must be established by intervention or ablation.

3. PROBLEM FORMULATION

3.1. Forecasting task. Let $\mathcal{Z} = \{z_1, \dots, z_N\}$ be N urban service zones and let $y_{z,t} \in \mathbb{Z}_{\geq 0}$ be observed demand in zone z during hour t . Given a history of length L , static zone attributes $\mathbf{s}_z$, and a graph $\mathcal{G}$, the model predicts horizons $\mathcal{H} = \{1, 3, 6\}$:

$$\hat{y}_{z,t+h} = f_{\theta}(\mathbf{y}_{t-L+1:t}, \mathbf{s}, \mathcal{G})_{z,h}.$$

Probabilistic models output monotone conditional quantiles $\hat{y}_{z,t+h}^{(\tau)}$ for $Q = \{0.1, 0.5, 0.9\}$.

3.2. Service capacity commitment and the target service level. The first decision is a classical service capacity problem. Before demand is observed, the operator must commit serving capacity

a_z to zone z — couriers, vehicle slots, or shift hours. Committing too little denies service to arriving demand; committing too much leaves paid capacity idle. Writing c_u for the penalty per unit of unserved demand and c_o for the penalty per unit of idle capacity,

$$C(a_z, y_z) = c_u(y_z - a_z)^+ + c_o(a_z - y_z)^+.$$

The cost-minimising commitment is a quantile of the demand distribution,

$$a_z^* = F_{Y_z}^{-1}(\tau), \quad \tau = \frac{c_u}{c_u + c_o},$$

and the quantity τ is the critical ratio, which in inventory and service operations is standard as the target cycle service level: the fraction of demand the operator intends to serve, implied by how much worse a denied customer is than an idle resource. This identity is not our observation; it is textbook [52], and setting a service target and choosing a critical ratio have long been understood as the same act expressed in two vocabularies. Fahimnia et al. [19] give a measure of how tightly the two are bound in practice: in a controlled experiment run with students and with practitioners, forecasters folded the service level into forecasts they had been instructed to keep free of it and were paid to keep accurate, and the effect traced to the service-level information itself rather than to some other anchor.

What we take to be under-used is the *evaluative* consequence rather than the identity itself. If the decision consumes the τ -quantile, then a forecasting system should be judged on whether it estimates that quantile — yet forecasting practice in this domain is still assessed almost entirely by symmetric point-error metrics. The general form of the problem is Gneiting's [23]: a point forecast is well posed only once the scoring function is fixed in advance, or the forecaster is directed to report a named functional such as a mean or a quantile, and comparing forecasts under an error measure that is not matched to the task can mislead. The sharpest form of the consequence is this: a point forecast is not service-level-neutral; it silently commits the operator to a 50% target. A conditional median is the optimal commitment only when $c_u = c_o$ — that is, only when denying service costs exactly what idling a courier costs. No last-mile service operates under that assumption. An organisation that plans capacity from a mean or median forecast has therefore chosen a service level by default rather than by decision, and it has chosen the lowest one available.

Throughout we study $\tau = 0.9$, corresponding to $c_u : c_o = 9 : 1$ — denied service is nine times as costly as idle capacity, a conservative reading of the reputational and re-delivery cost of a missed drop. Section 7.8 reports how conclusions move as this ratio varies. We report the achieved counterpart of the target as the fill rate, the realised share of demand served, which is the standard service-level measure in service management. The distinction that organises this paper is that the target service level is a *managerial* choice, while the ability to hit it is a *forecasting* property: it requires the conditional τ -quantile, and no amount of accuracy in a point forecast supplies that functional.

3.3. Zone selection and routing. At each hour the operator selects the K zones with highest predicted demand and constructs a depot tour over their centroids using shortest-path road distances.

Nearest-neighbour construction followed by 2-opt [53] supplies a common routing heuristic for all policies. Selection quality is the fraction of realised citywide demand contained in the chosen zones:

$$\text{coverage}_t(S_t) = \frac{\sum_{z \in S_t} y_{z,t}}{\sum_{z \in \mathcal{Z}} y_{z,t}}, \quad |S_t| = K.$$

The selected set depends on rank, not calibrated magnitude. Route length is reported separately; a demand-per-kilometre ratio is not used because it rewards short tours even when they cover little demand.

4. METHOD

4.1. Seasonal residual forecasting. Hourly urban demand is strongly periodic. For each zone and hour of week ω , a historical base $\text{HA}(z, \omega)$ is estimated from training data only. Models forecast a residual in log-transformed, standardised space, and the base is added before inversion to the original count scale. The observed residual is also supplied as an input channel at every history step. Every learned baseline receives the same seasonal base and input information.

4.2. Spatial backbones. Zone-to-zone dependence is represented as a graph and encoded by a spatial neural backbone. Because this paper’s claim concerns the decision layer rather than the network, the backbones are treated as interchangeable components and specified only to the level needed to reproduce the comparison; full construction detail, edge counts, and layer equations are in Appendix A.

We use two backbones. The first, denoted ST-MRGAT, combines a temporal convolution over each zone’s 12-hour history with graph attention applied separately over three zone relations — geographic proximity, road-network proximity, and training-period origin–destination flow — whose outputs are fused by a learned per-zone weighting. The second is Graph WaveNet [5], an established spatio-temporal architecture that learns its own adjacency instead of receiving hand-specified relations. Established baselines (GCN [29], GraphSAGE [30], GAT [8], DCRNN [4], LSTM, MLP) are trained in the same harness.

Two design choices matter for validity rather than performance. Origin–destination edges are estimated from training trips only, so no graph encodes test-period behaviour. And whether the three hand-specified relations earn their place is settled by ablation in Section 6.2, not by the architecture’s name; we report that they do not.

4.3. Transferable quantile and calibration layer. The output head emits one value for every horizon–quantile pair. Non-crossing quantiles are enforced by making the lowest quantile free and adding positive softplus increments for successive quantiles. Training minimises mean pinball loss:

$$\mathcal{L}_{\text{PB}} = \frac{1}{|Q|} \sum_{\tau \in Q} \max\{\tau e_\tau, (\tau - 1)e_\tau\}, \quad e_\tau = y - \hat{y}^{(\tau)}.$$

TABLE 1. The two New York taxi services, read over one geography of 263 zones and one period of 2,184 hours. Trip counts are raw TLC records before zone assignment and filtering, and borough shares are shares of trip activity. Green is a second demand regime inside the same city, not a second city.

Property	Yellow taxi	Green taxi
Trips read, Jan–Mar 2016	34,499,859	4,532,407
Manhattan share	89.0%	35.3%
Brooklyn share	5.3%	31.9%
Queens share	5.0%	26.1%
Neural-training seeds	5	3
Typical WMAPE	about 12%	about 25%

The same head and loss are attached to both ST-MRGAT and Graph WaveNet; the latter is denoted Graph WaveNet-Q. This isolates the transferability of the decision layer from the choice of spatial backbone.

For an intended $1 - \alpha = 0.8$ interval, CQR nonconformity scores on the calibration block are

$$E_i = \max\{\hat{y}_i^{(0.1)} - y_i, y_i - \hat{y}_i^{(0.9)}\}.$$

The empirical finite-sample corrected $1 - \alpha$ quantile of these scores widens the lower and upper endpoints. Model stopping and calibration use disjoint, chronologically ordered subsets of validation data. This is the split-conformal construction, whose finite-sample guarantee is marginal rather than per zone [38]. Section 7.8 therefore also reports a Mondrian, or group-conditional, variant [54], in which the calibration block is partitioned and a separate widening constant is fitted within each group.

5. EXPERIMENTAL DESIGN

5.1. Data and contrasting demand regimes. NYC Taxi and Limousine Commission trip records provide pickup and drop-off times and taxi-zone identifiers [7]. Drop-off counts are used as a proxy for destination-level last-mile activity, not as parcel orders.

Both datasets use 263 zones and 2,184 hourly observations. Their log zone totals have correlation 0.798, indicating related but clearly different spatial demand. Green taxis cannot accept street hails in much of lower Manhattan, producing a borough distribution unlike yellow taxis. This is a second demand regime within one city, not evidence of cross-city generalisation.

5.2. Temporal protocol. Samples are assigned by label time without shuffling:

The first 60% of validation selects the stopping epoch; the final 40% calibrates quantiles. Inputs may look backward across a block boundary, as in deployment, but no feature or statistic uses future labels. Normalisation, seasonal profiles, and OD edges are estimated from training data only.

TABLE 2. Chronological splits, shared by both taxi datasets. A window carries twelve hours of history and is assigned to a split by its label time, never shuffled; the last 40% of the validation block is reserved for conformal calibration and is never seen by early stopping.

Split	Label period	Windows
Training	1 Jan–29 Feb 2016	1,255
Validation	1–15 Mar 2016	355
Test	16–31 Mar 2016	379

Neural models use AdamW, a maximum of 150 epochs, and early stopping with patience

- (1) We compare persistence, historical average, Ridge, XGBoost, MLP, LSTM, GCN, GraphSAGE, GAT, DCRNN [4], Graph WaveNet [5], ST-MRGAT, and Graph WaveNet-Q. This set is chosen to span the model families an operator would plausibly consider — a naive rule, a seasonal average, linear and tree models, a sequential model, and graph models with fixed and with learned adjacency — rather than to survey the spatio-temporal literature. Other architecture families are not evaluated here: adaptive-graph models that infer inter-series dependence instead of receiving a predefined graph [25] [24], synchronous spatial-temporal convolution [26], grid-based residual networks for citywide crowd flow [27], and general probabilistic or attention-based sequence forecasters without a spatial component [35] [36]. Extending the backbone set would sharpen the accuracy comparison, but it would not change which functional either decision consumes, which is what this paper tests.

For the yellow dataset, each quantile backbone is retrained at five independent seeds and News vendor costs are summarised with 95% Student-t intervals over seeds. This treats initialisation variability as the replication unit rather than treating correlated hourly observations as independent replicates. Green supplies the cross-regime replication for the primary estimates, with its Graph WaveNet-Q extension reported over three seeds.

MAE, RMSE, WMAPE, and R^2 are computed after inversion to count space. These are the conventional point-error measures, and the choice among them is itself a standing question in the forecasting literature [32]; point accuracy of this kind was also the criterion of the M5 retail demand-forecasting competition [31]. We report them for comparability with prior work, but the decision evaluations that follow do not rest on them. Seed means and standard deviations describe initialisation variability. Pairwise forecast differences use the Diebold–Mariano (DM) test [13] on a series of 379 hour-level absolute-loss differences. Loss is aggregated over zones before testing to avoid treating 263 correlated zones as independent. Newey–West variance and the Harvey–Leybourne–Newbold correction [55] are used.

Because the study reports 48 DM comparisons (two datasets x three horizons x eight baselines) and eight paired ablation comparisons, raw p -values would overstate significance. We therefore

fix the test families in advance and report multiplicity adjustment for each. Families are defined by the claim: within a dataset and horizon (eight tests, the natural family for "at this horizon, does the proposed model differ from each baseline?"); within a dataset across horizons (24 tests); and across the whole study (48 tests). Ablation comparisons form a separate four-test family per dataset. Holm–Bonferroni [21] controls the family-wise error rate and is applied to all affirmative claims; Benjamini–Hochberg [22] false-discovery-rate q -values are reported alongside it. Section 7.8 gives the outcome, and full per-test values are in `outputs/multiplicity.csv`.

5.3. External-city decision validation. To test whether the decision-functional effect extends beyond NYC without making an unsupported neural-topology transfer claim, we use the public UCI Porto taxi trajectory archive [14]. We extract each trip’s first GPS coordinate, aggregate hourly pickups to a fixed 8-by-8 metropolitan grid (54 non-empty cells), and retain about 1.7 million valid trajectories. The first 70% of hours estimates a day-of-week-by-hour historical median and 0.9 quantile; the final 2,628 consecutive hours are held out. This is a deliberately model-free external test of the newsvendor mechanism. Its uncertainty interval uses 2,000 bootstrap replicates of complete calendar days, not individual zone-hours.

We also retrain Graph WaveNet-Q from scratch on this Porto representation. For this model test, the chronological split is 60% training, 10% validation, and the same final 30% test block. The Porto graph connects adjacent occupied grid cells; model parameters, adaptive adjacency, normalisation, seasonal profile, early stopping, and CQR offset are all estimated from Porto data only. We report three independent initialisations and compare them to historical median and $q_{0.9}$ policies estimated on the same 60% Porto training block.

5.4. Real parcel order-arrival-to-capacity validation. To test the demand side with actual parcel-commerce arrivals, we use the public Olist Brazilian E-Commerce dataset [16]. It contains anonymised purchase events, customer geography, order items, and delivery status. We retain 96,211 fulfilled orders whose purchase timestamps fall from 1 January 2017 through 29 August 2018, and aggregate distinct orders to 27 customer states on a complete daily calendar. The unit is a fulfilled order arrival, rather than an assumed physical package scan; the timestamp is the purchase-time arrival, never the delivery date.

We use a fixed chronological 60/10/30 train/validation/test split, leaving the final 182 days (1 March–29 August 2018) untouched until evaluation. A new Graph WaveNet-Q is initialised separately for each of three seeds. Its graph joins each state to its three nearest states using public customer-ZIP centroids; its normalisation, weekday profile, early stopping, and CQR offset use training or validation data only. Capacity is a per-state daily $q_{0.9}$ fulfilled-order commitment under the same 9:1 underage-to-overage loss. Training-only weekday median and $q_{0.9}$ policies are transparent comparators. Thus this experiment tests the full order-arrival $\rightarrow$ forecast $\rightarrow$ capacity-decision chain on real parcel-commerce observations, but not route construction.

5.5. Real parcel capacity and routing validation. We use the public Amazon Last Mile Routing Research Challenge training set [15], which contains 6,112 historical delivery routes across five US metropolitan areas. For every route it provides package-level dimensions, planned service time, delivery time windows, vehicle cubic capacity, observed stop sequence, and route-specific historically realised average travel times. We split routes chronologically: 4,362 routes departing through 12 August 2018 form the training history and 1,750 routes departing from 13–26 August form the holdout.

For capacity, we aggregate each station-day’s actual package count, parcel volume, and planned service seconds. Training-only day-of-week median and q0.9 policies are evaluated at $\tau = 0.9$ on 121 observed holdout station-days; uncertainty uses 5,000 complete-day bootstrap replicates. For routing, all 1,750 holdout routes are simulated with the released travel-time matrices. We report travel, waiting, planned service, route duration, package time-window lateness, and loaded-volume/capacity. The observed driver sequence is an operational reference, not an oracle or a proof of optimality. Nearest-neighbour travel-time routing is the common transparent comparator.

5.6. Same-order dispatch-to-delivery case study. We use the public Meituan–INFORMS operational food-delivery corpus [17] to test the data linkage unavailable in the two parcel sources. The release retains the same anonymised order ID through platform order creation, entry into dispatch, courier assignment, courier acceptance, pickup, and customer arrival. Accepted orders also join to the observed courier wave containing their order IDs. The source is released for non-commercial research under CC BY-NC 4.0 and its required acknowledgement is included below.

We retain 568,544 accepted orders whose six timestamps are positive and ordered, from 654,343 waybills, 4,901 couriers, and 23 dispatch areas. The forecasting window uses complete days: 17–20 October 2022 for training, 21 October for validation, and 22–23 October for the final test. At each 5-minute area bin, the target is the number of newly created orders. A fresh three-seed Graph WaveNet-Q uses only previous bins, a training-only time-of-day profile, and a nearest-three area graph built from pickup coordinates. It supplies a q0.9 area capacity decision; training-only time-of-day median and q0.9 profiles are the comparators. We use a one-hour moving-block bootstrap on the two test days for the median-versus-q0.9 contrast.

This is a linked order-arrival $\rightarrow$ dispatch $\rightarrow$ courier $\rightarrow$ pickup $\rightarrow$ delivery record, not a counterfactual dispatch evaluation. A courier wave and its event times reveal observed execution chronology, but the release has neither a vehicle capacity field nor continuous road-GPS traces. We therefore do not claim a solved vehicle-routing counterfactual.

5.7. Long-horizon parcel task and GPS-event validation. We additionally use the public Cainiao-LaDe industry dataset [18], whose full Shanghai delivery release covers 1,483,864 parcel tasks from 1 May through 31 October 2022. For every task it reports an anonymised courier, a dispatch region, acceptance and delivery event times, and the courier’s GPS coordinates at both events. We retain all 52 regions; every released task has non-missing event GPS and an acceptance time no later than its delivery time.

TABLE 3. Test accuracy and parameter counts on both regimes. MAE is in drop-offs per zone-hour and is averaged over the horizons $t+1$, $t+3$ and $t+6$, so it is not a next-hour figure; neural entries are seed means with the seed standard deviation, over five seeds on yellow and three on green. Diebold–Mariano tests do not separate Graph WaveNet from ST-MRGAT at any horizon on either dataset.

Model	Parameters	Yellow MAE	Green MAE
Graph WaveNet	147,087	7.275 +/- 0.055	1.954 +/- 0.003
Graph WaveNet-Q	147,477	7.293 +/- 0.089	1.975 +/- 0.003
ST-MRGAT	62,155	7.434 +/- 0.085	1.963 +/- 0.005
GraphSAGE	52,547	7.487 +/- 0.040	1.973 +/- 0.005
DCRNN	195,139	7.571 +/- 0.092	1.972 +/- 0.006
GCN	44,227	7.591 +/- 0.071	1.973 +/- 0.002
GAT	44,483	7.613 +/- 0.076	1.964 +/- 0.003
MLP	12,675	7.651 +/- 0.059	1.986 +/- 0.002
LSTM	53,571	7.658 +/- 0.125	1.984 +/- 0.004
XGBoost	—	7.738	1.978
Historical average	—	8.112	2.028
Persistence	—	29.177	4.555

The decision target is hourly accepted parcel-task load per region. This is an observable courier-task-slot commitment, not a proxy for physical vehicle volume: LaDe does not release parcel dimensions, a vehicle capacity field, an order-creation time, or a continuous per-task road trace. We use a chronological 60/10/30 split, with the final 55 days beginning 6 September 2022 held out. A fresh three-seed Graph WaveNet-Q uses a nearest-three region graph from delivery coordinates, training-only normalisation and week-hour profile, validation early stopping, and CQR. Training-only week-hour median and q0.9 policies are comparators; 24-hour moving-block bootstrap quantifies their held-out cost contrast.

6. RESULTS

6.1. Accuracy differences are small and rankings are unstable. Graph WaveNet has the lowest mean MAE on both datasets, but it has 2.4 times as many parameters as ST-MRGAT. DCRNN is larger still and does not win, so capacity alone is not sufficient, but an equal-budget comparison is required before attributing the gain to adaptive adjacency.

Middle rankings are not stable. Graph WaveNet-Q moves from second on yellow to seventh among neural models on green; GAT moves from seventh to third. All nine neural models occupy the narrow interval 1.954–1.986 on green, a 1.6% spread. Adding the quantile head changes Graph WaveNet’s MAE by only +0.25% on yellow and +1.06% on green.

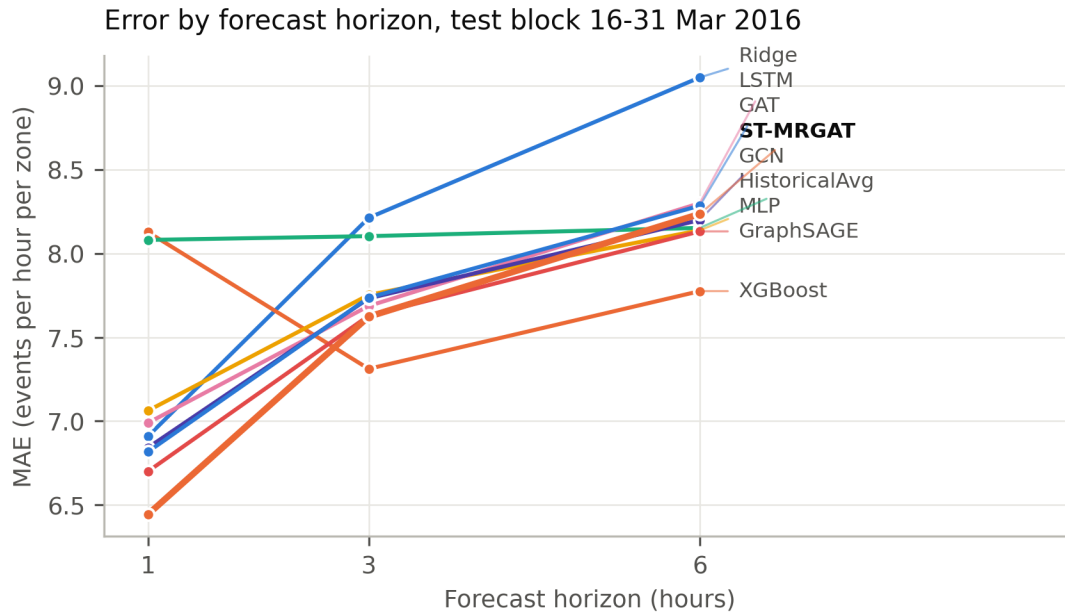


FIGURE 1. Mean absolute error by forecast horizon, yellow-taxi test block. Model separation is largest at one hour ahead and disappears by six hours.

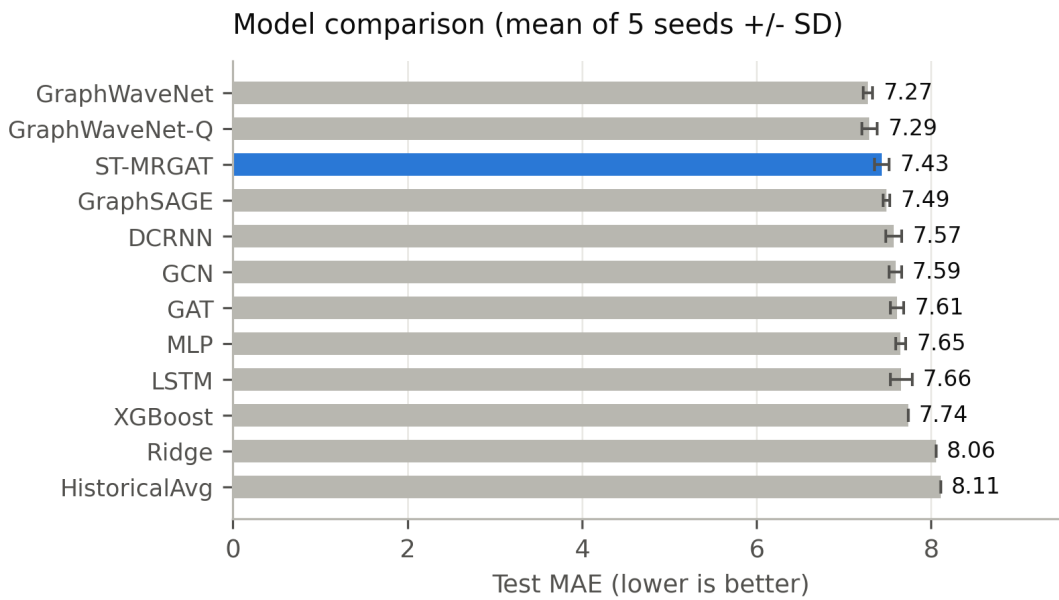


FIGURE 2. Model comparison over five seeds, yellow-taxi test block. Error bars are standard deviations across initialisations.

DM tests do not distinguish ST-MRGAT from Graph WaveNet at any horizon on either dataset. They also do not distinguish ST-MRGAT from Graph WaveNet-Q on yellow; on green, ST-MRGAT is favoured only at $t + 3$ ($p = 0.016$), not at $t + 1$ or $t + 6$. These tests support parity among the leading architectures rather than a claim that the proposed backbone is uniquely best.

TABLE 4. Relation and objective ablations, three seeds per row, which is why the full-model yellow entry is not the five-seed value reported earlier. Changes are percentages of the full model within the same dataset. The two effects that reach significance on yellow do not survive the move to green, where the reduced configurations are if anything slightly better.

ST-MRGAT configuration	Yellow MAE	Change	Green MAE	Change
Full: geo + road + flow	7.364	—	1.965	—
Without OD flow	7.371	+0.1%	1.962	-0.1%
Geographic only	7.392	+0.4%	1.962	-0.2%
Without road	7.457	+1.3%	1.961	-0.2%
Point loss instead of pinball	7.608	+3.3%	1.968	+0.15%

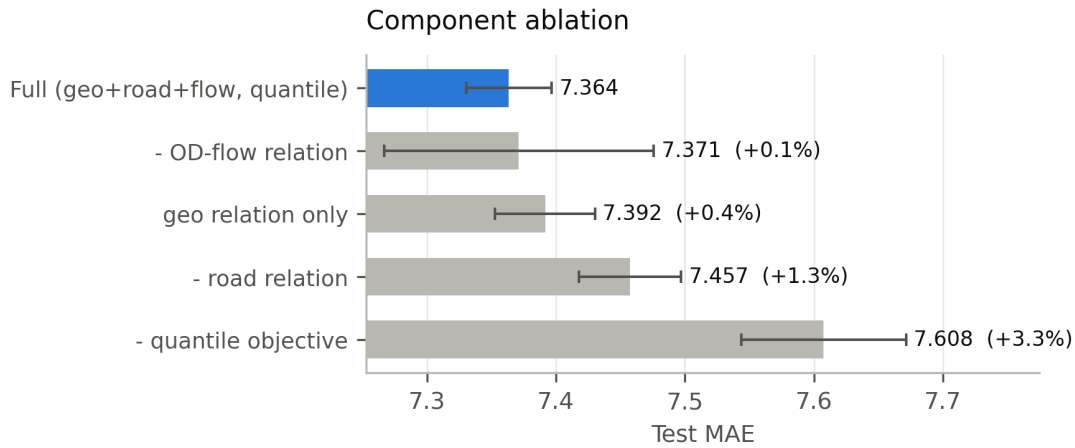


FIGURE 3. Component ablation, yellow-taxi test block. Removing the quantile objective costs eight times more than removing both auxiliary graph relations.

6.2. Hand-designed relation effects do not replicate. On yellow, removing the road relation is significant ($p = 0.005$), and replacing pinball loss is significant ($p = 0.011$). Neither effect replicates on green; the road effect reverses sign, and the loss effect shrinks to 0.15%. Removing both auxiliary relations is insignificant on yellow ($p = 0.590$) and slightly improves green performance. No hand-designed relation provides a consistent benefit across the two demand regimes.

This negative result changes the architectural interpretation. The multi-relational graph is a tested hypothesis, not an established contribution. The reproducible gain appears later, in calibration and decision alignment.

6.3. Conformal calibration transfers across backbones. For ST-MRGAT, interval width increases by 1.8% on yellow and 4.0% on green. For Graph WaveNet-Q, the increases are 1.1% and 2.1%. Thus empirical calibration is stable across both data regimes and spatial backbones. The result

TABLE 5. Empirical coverage of the nominal 80% interval before and after conformalised quantile regression, on two backbones. Entries pool the test block, all zones and all horizons, so they are marginal coverages: overshooting the target counts against it as much as undershooting. Per-zone coverage remains far from uniform.

Backbone	Yellow: raw / calibrated	Green: raw / calibrated
ST-MRGAT	74.18% / 79.76%	75.47% / 80.40%
Graph WaveNet-Q	75.91% / 80.04%	77.02% / 80.36%
Nominal coverage	80.00%	80.00%

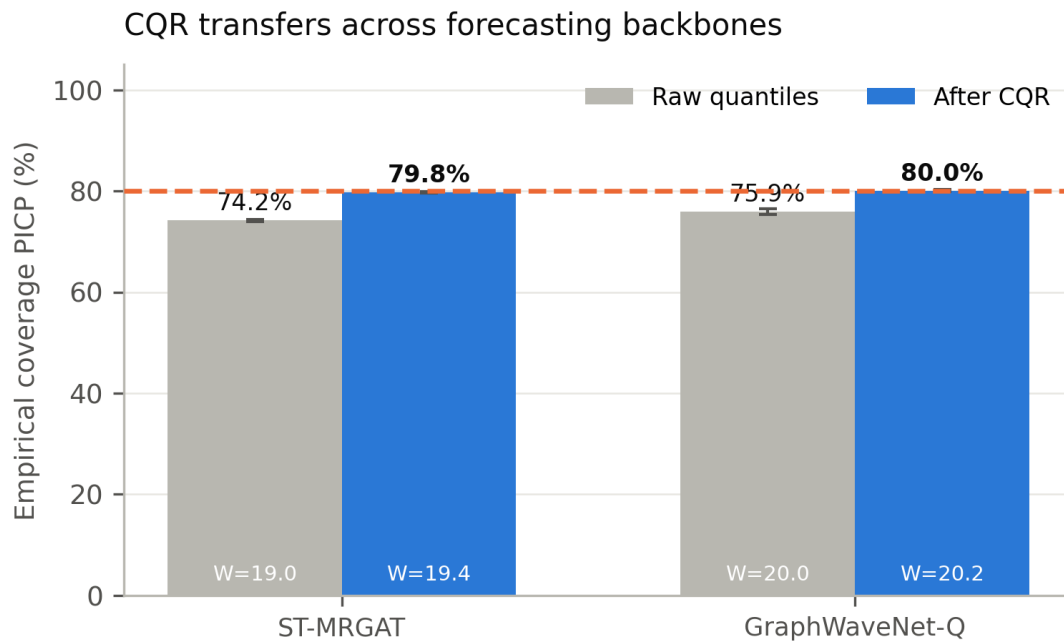


FIGURE 4. Coverage before and after CQR on yellow-taxi demand

concerns marginal coverage over the test block; it does not imply uniform conditional coverage for every zone or hour.

6.4. Decision-aligned quantiles, not lower MAE, reduce capacity cost. At $\tau = 0.9$, the realised newsvendor results are:

Read as a service–cost trade-off, the table says something simple. Every median-based policy delivers a fill rate near 95% on yellow and near 87% on green, because a median commitment targets a 50% service level and only reaches those figures incidentally. Moving the same forecast to its $q_{0.9}$ functional lifts the achieved fill rate to 99.02% and 98.03% respectively — and does so while *lowering* total cost, not paying for service with cost.

That combination is the managerially relevant point, because it is not the trade-off practitioners expect. Raising a service level normally means buying it: more capacity, higher cost. Here the

TABLE 6. Capacity cost and fill rate for one-hour-ahead commitments at a 9:1 underage-to-overage penalty. Cost is per zone-hour in demand counts and does not compare across the two columns. Intervals cover five yellow seeds; green entries are single runs, and the oracle is a perfect-foresight bound rather than a policy anyone could operate.

Policy	Yellow cost	Fill rate	Green cost	Fill rate
Oracle	0.000	100.00%	0.000	100.00%
Graph WaveNet-Q, $q_{0.9}$	1.616 (1.587–1.645)	99.02%	0.491	98.03%
ST-MRGAT, $q_{0.9}$	1.631 (1.602–1.659)	98.90%	0.498	97.43%
Graph WaveNet-Q, median	2.988 (2.547–3.428)	95.10%	0.921	88.27%
ST-MRGAT, median	3.079 (2.950–3.207)	94.92%	1.011	86.83%
Graph WaveNet, point	3.456	94.12%	1.050	86.19%
Historical average	4.959	91.33%	1.099	85.60%

service level rises and cost falls simultaneously, because the median policy was not sitting on the efficient frontier to begin with. It was mis-specified for an asymmetric penalty structure, so correcting the functional recovers service and cost at the same time. The gain is not a better forecast; it is the removal of a systematic under-commitment.

Quantitatively, across the five yellow seeds, replacing the median with $q_{0.9}$ reduces cost by 47.0% (95% CI: 45.1–48.9%) for ST-MRGAT and 45.4% (39.1–51.8%) for Graph WaveNet-Q, while raising the fill rate from 94.92% to 98.90% and from 95.10% to 99.02% respectively — gains of 4.0 and 3.9 percentage points. On green, the same substitution raises the fill rate from 86.83% to 97.43% for ST-MRGAT and from 88.27% to 98.03% for Graph WaveNet-Q, gains of 10.6 and 9.8 points; the three-seed Graph WaveNet-Q extension reduces green cost by 46.4% (43.5–49.2%). The service gain is larger on green because its demand is sparser, so a median commitment under-serves it more severely to begin with. The decision layer therefore transfers across backbones and demand regimes even though their MAE rankings do not.

The trade-off is genuine rather than free, and the mechanism should be stated plainly: the $q_{0.9}$ policy holds substantially more idle capacity than the median policy. For ST-MRGAT on yellow, idle capacity rises from 5.67% to 16.81% of demand units while unserved demand falls from 5.01% to 1.08%; on green the exchange is steeper still, with idle capacity rising from 11.34% to 40.85%. Under a 9:1 penalty this exchange is strongly favourable, and it becomes unfavourable as the ratio approaches 1:1. Section 7.8 identifies where the crossover lies. What the operator gains is therefore not something for nothing, but the ability to *choose* a position on the service–cost frontier instead of inheriting whichever position a median forecast happens to imply.

Graph WaveNet provides the lowest MAE but not the lowest point-policy cost. Relative to the ST-MRGAT median, its yellow MAE is 2.2% lower while its asymmetric cost is 13.8% higher. It under-allocates in 5.88% of demand units versus 5.01% for ST-MRGAT; the 9:1 loss magnifies that

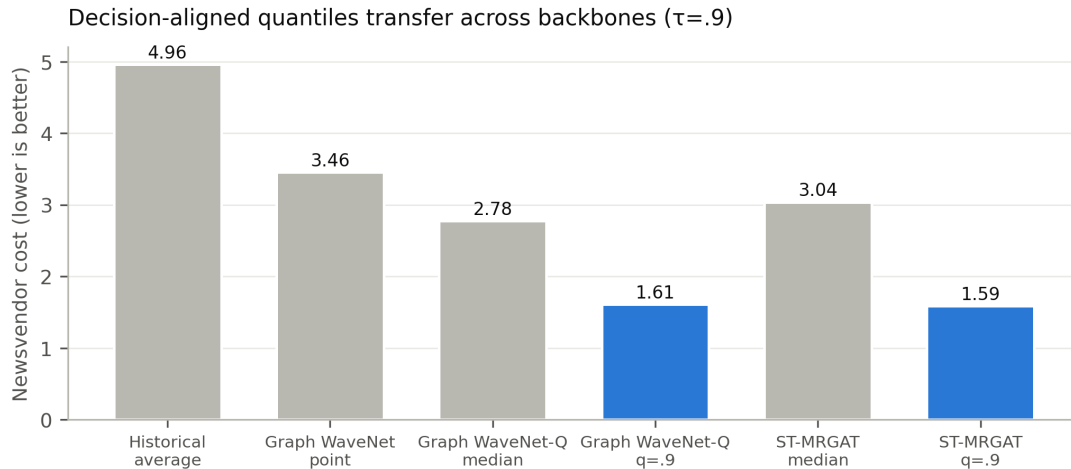


FIGURE 5. Capacity cost on yellow-taxi demand

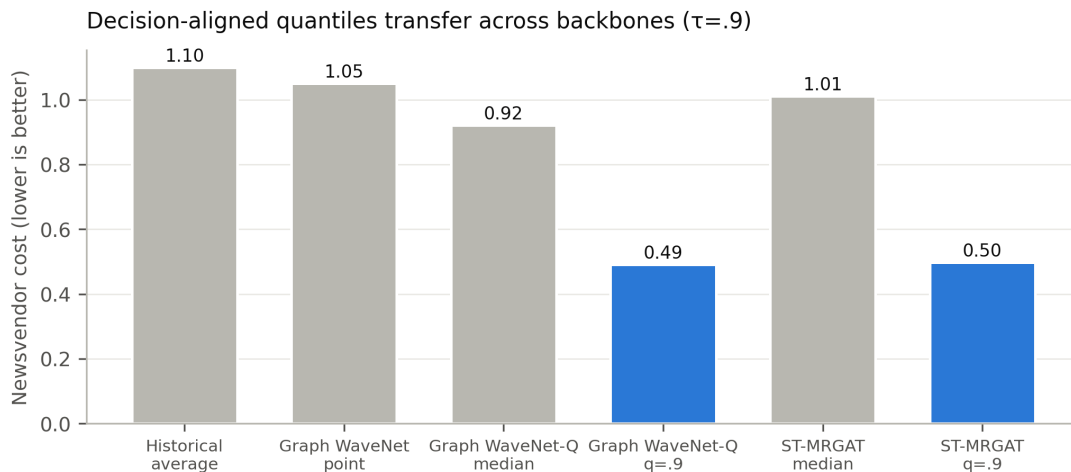


FIGURE 6. Capacity cost on green-taxi demand

difference. On green, Graph WaveNet-Q is only seventh among neural models by MAE but gives the lowest feasible decision cost. This is direct evidence that MAE ranking does not determine operational ranking, and it is the empirical form of a point already established in theory: comparing forecasts by an error measure supports an inference about the task only when the scoring function and the forecasting task are matched [23].

6.5. Sophisticated forecasts add almost no zone-selection value. Shortfall from oracle coverage, in percentage points, is:

Learned models improve coverage over the weekly historical average by at most about 0.15 points in these representative settings. Distance-only selection, however, loses 16–23 points on green. Using demand is essential; choosing among already good demand rankings is not. The stable weekly hotspot ordering explains why architecture matters little for this task.

TABLE 7. Shortfall from oracle coverage in percentage points, at three fleet sizes; zero is the hindsight optimum. Rankings come from the median $t+1$ forecast of a single run. ST-MRGAT gains about a tenth of a point on a training-only historical average, while the demand-blind distance rule gives up sixteen to twenty-three.

K zones	Yellow ST-MRGAT	Yellow hist. avg	Green ST-MRGAT	Green hist. avg	Green distance-only
12	0.40	0.54	1.19	1.31	16.09
15	0.45	0.57	1.33	1.45	18.02
20	0.44	0.59	1.39	1.47	22.78

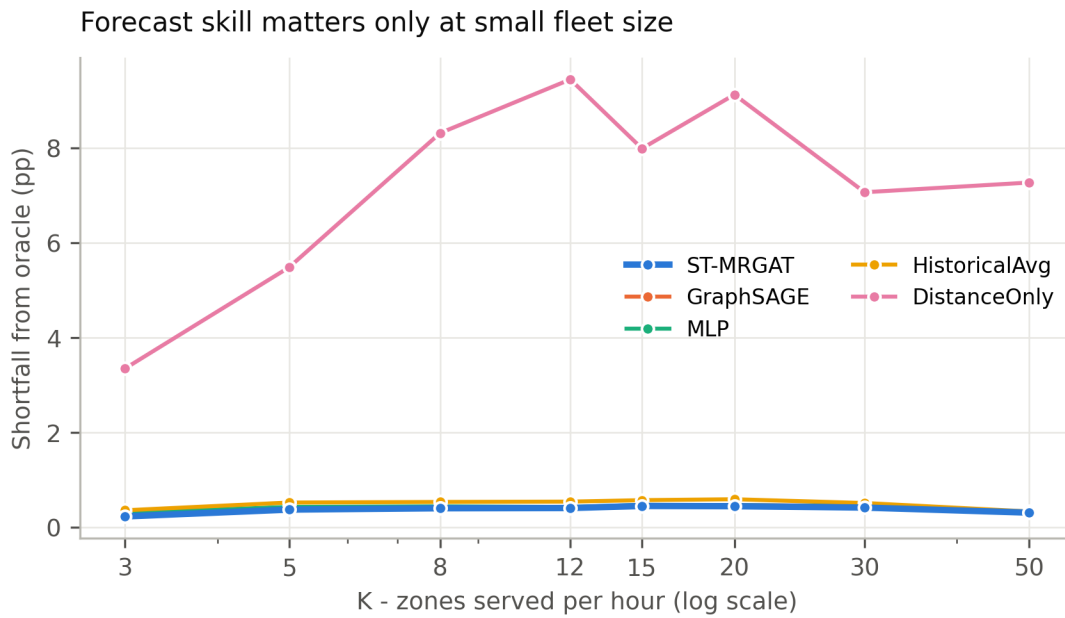


FIGURE 7. Shortfall from oracle zone-selection coverage against fleet size K . All demand-aware policies remain within a fraction of a percentage point of each other across the whole range.

6.6. Attention weights do not measure relation utility. On yellow data, mean relation-attention weights are 0.293 for geographic, 0.298 for road, and 0.409 for OD flow. Manhattan assigns 0.496 to OD flow, while Staten Island assigns 0.445 to road and 0.120 to geographic proximity. This pattern appears plausible, but the ablation result contradicts a utility interpretation: removing OD flow changes yellow MAE by only 0.1% ($p = 0.894$) and slightly improves green MAE. Attention describes internal mixing; it is not evidence that the attended relation improves forecasts.

6.7. External-city validation supports the decision mechanism. On Porto, the training-only historical median policy has q0.9 cost 0.6608, 81.36% fill rate, and 18.64% unmet demand. Replacing only its decision functional with the training-only historical q0.9 policy lowers cost to 0.3487, raises

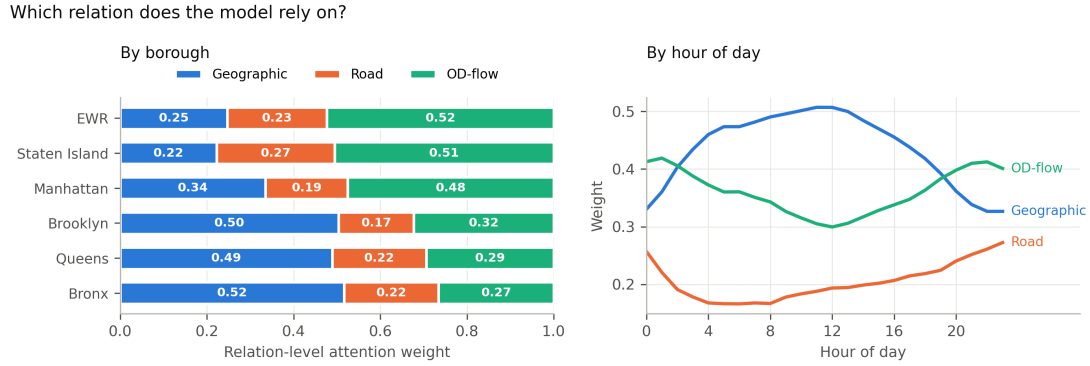


FIGURE 8. Learned relation-level attention by borough and by hour. The relation receiving the largest average weight is the one ablation finds unnecessary.

TABLE 8. Porto, with no forecaster involved. Both policies are day-of-week by hour-of-day statistics fitted on the first 70% of hours, so only the committed quantile differs. Costs are per cell-hour in pickups, and the interval is a block bootstrap over held-out calendar days.

Porto policy (training-only)	q0.9 cost	Fill rate	Unmet demand	Cost reduction vs median
Week-hour historical median	0.6608	81.36%	18.64%	—
Week-hour historical q0.9	0.3487	93.68%	6.32%	47.23% [43.31%, 50.86%]

TABLE 9. Graph WaveNet-Q retrained from scratch on Porto. Nothing carries over from New York: graph, weights, normalisation and calibration offset are all estimated on Porto data. The MAE column is not like for like, since the historical rows report the error of their own commitment while the neural row reports the error of its median forecast.

Porto model, 60/10/30 split	q0.9 cost	MAE	Fill rate	Scope
Historical median	0.6486	1.002	81.69%	Training-only week-hour policy
Historical q0.9	0.3470	1.559	93.62%	Training-only week-hour policy
Graph WaveNet-Q, fresh Porto retraining	0.2671 [0.2326, 0.3016]	0.880	96.88%	3 seeds; new graph and parameters

fill rate to 93.68%, and reduces unmet demand to 6.32%. The 47.2% cost reduction has a day-block bootstrap 95% CI of [43.3%, 50.9%] over 110 held-out calendar days.

The external neural retraining is a stricter test. With its own 60/10/30 Porto split and graph, fresh Graph WaveNet-Q achieves 0.2671 cost (three-seed 95% CI [0.2326, 0.3016]) and 0.880 MAE. The matched historical q0.9 policy has 0.3470 cost, so retraining lowers cost by 23.0% (seed-level 95% CI [13.1%, 33.0%]); the matched historical median has cost 0.6486. Mean neural fill rate is 96.88%, versus 93.62% for historical q0.9 and 81.69% for the median. All three neural runs use a new 54-cell Porto graph and fresh initialisation.

TABLE 10. Daily order-arrival capacity across 27 Brazilian states, on the final 182 days of a chronological split. Demand is a fulfilled order stamped at purchase rather than at delivery. The MAE column is not a common accuracy comparison, because each historical row reports the error of its own commitment level.

Olist fulfilled-order capacity policy	Normalised q0.9 cost	MAE
Historical median, training only	3.7965	4.380
Historical q0.9, training only	1.7759	3.121
Fresh Graph WaveNet-Q (3 seeds)	0.5453 [0.5309, 0.5598]	2.191 [2.142, 2.241]

This external result confirms that asymmetric capacity decisions benefit from the appropriate conditional functional and that the Graph WaveNet-Q decision pipeline can be retrained successfully outside NYC. It does not establish transfer for ST-MRGAT, every relation ablation, or city-scale routing.

6.8. Real parcel order arrivals validate the forecasting-to-capacity chain. The Olist experiment supplies what the route corpus does not: a public chronological stream of real parcel-commerce order arrivals. The training-only historical q0.9 policy reduces asymmetric capacity cost by 53.22% [50.75%, 56.23%] relative to its median counterpart, using a five-day moving-block bootstrap on the 182 held-out days. Fresh Graph WaveNet-Q retraining further obtains mean cost 0.5453 [0.5309, 0.5598] across three seeds, compared with 1.7759 for the historical q0.9 policy: a 69.29% reduction in mean cost. Its median forecast MAE is 2.191 [2.142, 2.241] fulfilled orders per state-day. The very large gain is consistent with the strong growth and regional heterogeneity in this e-commerce series; it is a result for this chronological corpus, not a universal architecture claim.

This is an end-to-end forecasting-to-capacity result on real parcel order arrivals. It is deliberately not presented as a routed-delivery outcome: Olist does not release matched vehicle routes, and the separate Amazon corpus does not release the corresponding arrival history. Linking them would manufacture a data relationship that is not observed.

6.9. Real parcel data validates capacity and routing objects. The Amazon data removes the taxi-proxy objection for the operational objects. On held-out station-days, the q0.9 policy reduces realised asymmetric cost by 42.37% for package count, 41.66% for loaded cubic volume, and 42.96% for planned service time. Day-block intervals remain well above zero for all three targets.

The routing comparison uses all 1,750 chronological holdout routes. Observed driver sequences have 3.028 mean travel hours and 0.084% late packages; the transparent nearest-neighbour comparator has 3.111 travel hours and 0.634% late packages. The observed operational routes therefore reduce travel time by 2.67% [2.33%, 3.10%] and late packages by 86.73% [85.04%, 88.28%] under day-block bootstrap. Mean loaded volume is 72.51% of vehicle capacity; only 0.17% of recorded routes exceed the released capacity after package-volume aggregation.

TABLE 11. Cost reduction from committing the 0.9 quantile rather than the median, on three Amazon capacity targets measured over the same 121 holdout station-days. Both policies are training-only historical quantiles, so no forecasting model enters. Intervals are day-block bootstrap percentiles and are not symmetric about the estimate.

Parcel capacity target	q0.9 vs median cost reduction [95% CI]	Holdout station-days
Package count	42.37% [37.35%, 57.28%]	121
Parcel volume (cm ³)	41.66% [37.84%, 55.82%]	121
Planned service time	42.96% [38.79%, 55.97%]	121

TABLE 12. Every holdout route simulated twice over identical stop sets: once in the sequence the drivers actually took, once under greedy nearest-neighbour construction. Planned service hours are equal by construction, and most of the route-hour gap is waiting for time windows rather than driving. No forecast enters.

Holdout routing policy	Travel hours	Planned service hours	Route hours	Late packages
Observed driver sequence (reference)	3.028	4.420	7.532	0.084%
Nearest-neighbour travel time	3.111	4.420	7.847	0.634%

These are not claims that the observed driver order is globally optimal. They show that the real delivery task contains the capacity, service, time-window, and routing structures that the taxi proxy cannot expose, and that a distance-only routing rule misses material operational quality.

6.10. One corpus follows the same order from arrival to delivery. The Meituan data closes the identifier-linkage gap in a distinct delivery domain. All 568,544 retained orders have a chronologically ordered creation, dispatch, acceptance, pickup, and delivery record; every accepted (courier, order) pair matches an observed courier-wave membership. Median creation-to-delivery time is 28.50 minutes (90th percentile 48.98), and 15.01% of orders arrive after the recorded promised time. These are measured operational outcomes for the identical order records used by the forecasting experiment.

The capacity mechanism also survives this finer dispatch clock. The training-only q0.9 time-of-day policy reduces cost by 40.20% [38.83%, 41.51%] relative to its median counterpart under one-hour moving-block bootstrap. Fresh Graph WaveNet-Q retraining reaches 0.8090 [0.7485, 0.8696] mean cost over three seeds, versus 1.2381 for the historical q0.9 policy, a 34.66% lower mean cost; its median MAE is 2.856 [2.775, 2.936] orders per area per five-minute bin.

The case study validates that a real demand stream, its assignment recipient, and its observed delivery execution can coexist under one order identifier. It does not show that our q0.9 policy

TABLE 13. Meituan capacity per dispatch area in five-minute bins, on the two held-out days. The 0.9-quantile profile costs about 40% less than the median profile while scoring slightly worse on MAE, which is the divergence at issue.

Meituan arrival-capacity policy	Normalised q0.9 cost	MAE
Historical median, training only	2.0705	2.974
Historical q0.9, training only	1.2381	3.025
Fresh Graph WaveNet-Q (3 seeds)	0.8090 [0.7485, 0.8696]	2.856 [2.775, 2.936]

TABLE 14. Linkage audit of the retained Meituan orders; no model, forecast or cost enters. The complete-lifecycle figure holds by construction of the retention filter, whereas the courier-wave link is measured.

Same-order operational linkage	Observed value
Accepted orders / source waybills	568,544 / 654,343
Couriers / courier waves / dispatch areas	4,901 / 206,748 / 23
Ordered lifecycle and courier-wave link	100.00% / 100.00%
Median / 90th-percentile creation-to-delivery	28.50 / 48.98 min
Late versus recorded promise	15.01%

would improve those realised routes: doing so would require a valid counterfactual assignment-and-routing simulator or experimental policy variation, which the released observational records do not provide.

6.11. Long-horizon parcel task capacity with observed GPS events. The Cainiao-LaDe result tests the capacity mechanism on a materially longer parcel history: 1,483,864 Shanghai tasks, 52 regions, and 184 days. Every task retains GPS at both courier acceptance and delivery events, so the reported task-slot load is anchored to observed courier operations rather than taxi demand. The capacity unit remains a task slot; it is not presented as vehicle volume or parcel weight.

The historical q0.9 policy lowers cost by 54.33% [52.65%, 56.21%] relative to the training-only median. Fresh Graph WaveNet-Q retraining obtains mean cost 2.0324 [1.9684, 2.0964] over three seeds, 28.16% below the historical q0.9 cost of 2.8292; median MAE is 5.755 [5.637, 5.873] accepted tasks per region-hour.

This result strengthens long-horizon parcel generalisation and confirms that the forecast-to-capacity mechanism is useful when its operational unit is courier task slots. It does not close the physical-capacity-and-continuous-route-trace gap: the dataset's endpoint GPS events cannot be promoted to a complete vehicle trajectory.

TABLE 15. Hourly capacity on the LaDe Shanghai release, evaluated on the final 55 of 184 days. The committed unit is an accepted courier task slot, since the release records neither parcel dimensions nor any vehicle capacity limit.

LaDe Shanghai courier-task-slot policy	Normalised q0.9 cost	MAE
Historical median, training only	6.1943	7.282
Historical q0.9, training only	2.8292	7.443
Fresh Graph WaveNet-Q (3 seeds)	2.0324 [1.9684, 2.0964]	5.755 [5.637, 5.873]

7. DISCUSSION

7.1. Forecast value is decision-specific. Capacity commitment and zone selection use the same observations but consume different information. Capacity cost is asymmetric and needs a tail quantile. Training and calibrating that quantile changes cost by 42–51%. Zone selection needs only an ordering, and the ordering supplied by a weekly average is already close to oracle. Asking whether a forecast is “operationally useful” without specifying the decision functional is therefore incomplete.

7.2. The transferable contribution is above the backbone. The original architectural hypothesis does not survive replication: neither the multi-relational graph nor its individual hand-designed relations improves both demand regimes. In contrast, the quantile-and-CQR layer works on ST-MRGAT and Graph WaveNet-Q, with realised coverage near nominal and large cost reductions on both datasets. This transfer experiment is stronger evidence than an ablation on one network because it separates the decision layer from the spatial representation.

The practical design implication is modular. An operator can retain a preferred forecasting backbone and attach outputs aligned with the decision: quantiles for asymmetric capacity, rankings for fixed-cardinality selection, or other functionals for other objectives. Architectural research remains useful, but it should not be used as a proxy for decision evaluation.

7.3. Why MAE and operating cost disagree. MAE weights over- and under-prediction symmetrically. The newsvendor objective at $\tau = 0.9$ weights shortage nine times more than excess. A model can therefore reduce absolute error while concentrating too much residual mass below realised demand. Conversely, a calibrated upper quantile intentionally over-allocates more often and still reduces expected asymmetric cost. This is not a failure of MAE as a descriptive metric; it is a mismatch between MAE and the decision. The general form of the point is established: comparing point forecasts through an error measure can lead to misguided inferences unless the scoring function and the forecasting task are matched, and the remedy is either to fix the scoring function in advance or to direct the forecaster to a stated functional, such as a quantile of the predictive distribution [23]. Our capacity decision issues exactly such a directive.

TABLE 16. Seed-level view of the main capacity result. Intervals run over training seeds, not over test hours. The reduction column is a paired per-seed statistic, obtained by replacing the same trained model’s median commitment with its calibrated 0.9-quantile commitment, and so cannot be recovered from the cost column beside it. The overlapping yellow intervals separate neither backbone.

Dataset	Backbone	q0.9 cost [95% CI]	q0.9 vs median reduction [95% CI]
Yellow	ST-MRGAT	1.631 [1.602, 1.659]	47.0% [45.1, 48.9]
Yellow	Graph WaveNet-Q	1.616 [1.587, 1.645]	45.4% [39.1, 51.8]
Green	Graph WaveNet-Q	0.492 [0.490, 0.495]	46.4% [43.5, 49.2]

7.4. Implications for evaluation practice. A forecasting paper that makes operational claims should name the exact functional its decision consumes and report realised decision cost beside whatever generic error metric it uses. It should measure itself against simple operational baselines, historical averages among them, since those are what an operator already has. Where a decision consumes a distributional tail, the calibration of that tail matters more than the accuracy of the centre. Statistical tests should be run over the temporal unit at which the decision is actually taken. And any interpretation of attention weights or graph components should follow an ablation rather than precede it.

The last of these has been established elsewhere by intervention rather than by inspection: manipulating attention weights in already-trained classifiers shows that a large weight predicts a component’s influence on the prediction only noisily, and that gradient-based rankings often predict the effect of altering a component better than the weight magnitudes do [51]. A second, mechanical caution applies to the neighbour attention inside each relation rather than to the relation weights of Section 6.6: a GAT layer computes what has been termed static attention, whose ranking of neighbours is unconditioned on the querying node [28], so those weights are not a per-zone importance ordering either.

Negative results are informative in this setting. They prevent a non-replicating architectural story from obscuring a robust decision-layer result.

7.5. Completed robustness checks. The primary yellow decision result is stable over five independent training seeds. The completed green Graph WaveNet-Q replication is also reported over three seeds. Values below are mean 0.9-quantile cost with a 95% Student-t interval; reduction is the paired reduction relative to the median policy.

An earlier rolling origin (train through January, tune through 15 February, test on 16–29 February) preserves the decision effect:

The capacity-matched test shows that the primary claim is not architectural superiority; it explicitly compares low- and high-capacity pairs.

TABLE 17. The same decision effect at an earlier origin: trained through January, calibrated to mid-February, tested on the fortnight that follows. Each row contrasts two commitments drawn from one trained forecast, so the four rows are not a test between backbones.

Dataset	Backbone	q0.9 vs median reduction [95% CI]
Yellow, earlier origin	ST-MRGAT	51.4% [48.2, 54.6]
Yellow, earlier origin	Graph WaveNet-Q	49.5% [40.6, 58.3]
Green, earlier origin	ST-MRGAT	50.6% [47.1, 54.1]
Green, earlier origin	Graph WaveNet-Q	47.0% [44.3, 49.6]

TABLE 18. Test MAE at approximately matched parameter budgets, three seeds per cell. The pairing exists to remove the parameter-count confound rather than to rank architectures: the gaps run from 0.16% to 1.18%, and three of the four sit at or below the measured resolution floor.

Pair	Model A (parameters)	MAE	Model B (parameters)	MAE
Yellow: GraphWaveNet-h40_vs_ST-MRGAT-h64	GraphWaveNet (61,263)	7.342 +/- 0.048	ST-MRGAT (62,155)	7.360 +/- 0.060
Yellow: GraphWaveNet-h64_vs_ST-MRGAT-h100	GraphWaveNet (147,087)	7.273 +/- 0.019	ST-MRGAT (147,511)	7.359 +/- 0.018
Green: GraphWaveNet-h40_vs_ST-MRGAT-h64	GraphWaveNet (61,263)	1.959 +/- 0.003	ST-MRGAT (62,155)	1.962 +/- 0.005
Green: GraphWaveNet-h64_vs_ST-MRGAT-h100	GraphWaveNet (147,087)	1.954 +/- 0.004	ST-MRGAT (147,511)	1.964 +/- 0.003

These checks strengthen the decision-layer conclusion while retaining the negative architectural finding: a capacity-matched comparison is not used to claim that one backbone is universally superior.

7.6. Constrained routing extension. To test whether the routing conclusion survives a more realistic dispatch proxy, we replace fixed-cardinality TSP selection with a deterministic three-vehicle heuristic. Each vehicle has a capacity equal to 42% of total hourly demand and a 22-km maximum closed route; candidate zones are ranked by forecast score per marginal insertion cost. On the yellow test block, coverage is 45.91% for the historical average, 46.02% for ST-MRGAT, 46.03% for Graph WaveNet, and 46.06% for Graph WaveNet-Q, with total route lengths near 60.7–60.8 km. The constrained result preserves the main negative finding: forecast rankings change coverage only marginally once a route-aware policy and common operational constraints are applied. This remains a transparent heuristic rather than an exact capacitated multi-vehicle VRP. It also keeps demand uncertainty in the capacity commitment and out of the routing problem itself, unlike the stochastic vehicle routing literature, which either designs an a priori customer sequence of minimal expected total length under a probabilistic description of demand [47], or solves a two-stage problem in which planned routes are followed until vehicle capacity is reached and the resulting return to the depot is penalised [48].

TABLE 19. Realised operating points as the underage-to-overage ratio varies. Cost is scaled per 1,000 drop-offs and carries no currency. Because the penalty weights change with the critical ratio, each row prices a different objective, and the cost column must not be read downwards as a ranking.

Policy	tau	Normalized cost / 1,000 demand units	Capacity / realised demand	Fill rate
ST-MRGAT	0.60	52.70	1.044	96.50%
ST-MRGAT	0.80	40.06	1.120	98.39%
ST-MRGAT	0.90	26.54	1.157	98.92%
Graph WaveNet-Q	0.60	54.21	1.060	97.00%
Graph WaveNet-Q	0.80	41.42	1.141	98.68%
Graph WaveNet-Q	0.90	26.86	1.182	99.13%

7.7. Direct decision-focused baseline. As a controlled method comparison, we trained the same Graph WaveNet representation with a single pinball head at the operational quantile $\tau=0.9$, rather than the modular three-quantile head followed by CQR. Across three independently trained seeds, the direct head has Newsvendor cost 1.612 (95% Student-t CI [1.565, 1.658]) and MAE 11.65. The modular Graph WaveNet-Q policy, evaluated over five seeds, has cost 1.616 (95% CI [1.587, 1.645]). The estimated 0.27% reduction for direct training is not statistically distinguishable (two-sided Welch test, $p = 0.783$). Thus this is evidence of practical parity under matched inputs and backbone, not a claim that direct decision-focused training improves the modular, calibrated policy.

7.8. Economic sensitivity and calibration slices. The Newsvendor conclusion is stable across a range of critical ratios. For ST-MRGAT, increasing τ from 0.60 to 0.90 raises the committed capacity ratio from 1.044 to 1.157 and realised fill rate from 96.50% to 98.92%; Graph WaveNet-Q changes from 1.060 to 1.182 and from 97.00% to 99.13%, respectively. The corresponding normalized cost and service trade-off is shown below. Because the underage and overage weights themselves change with τ , cost values across rows are descriptive operating-point values, not a cross-ratio model ranking.

We report normalized cost per 1,000 demand units rather than monetary dollars, because taxi drop-offs are a demand proxy and no defensible parcel price is observed.

Calibration is close to nominal marginal coverage but heterogeneous conditionally. Across hours, calibrated 80% coverage ranges from 78.1% to 82.8% for ST-MRGAT and from 78.4% to 84.4% for Graph WaveNet-Q. Across zones, ranges are 54.4–100.0% and 62.0–100.0%, respectively; across low-, middle-, and high-demand regimes the ranges are 80.6–81.2% and 80.3–81.8%. This supports the marginal CQR claim while making clear that it is not a guarantee of conditional coverage for every zone or regime.

The zone-level spread above is a practical limitation, not merely a statistical footnote: an operator cannot promise a per-zone service level from an interval that is calibrated only on average.

TABLE 20. Group-conditional calibration against a single global constant, applied to the same trained model within each seed so that nothing but the calibration differs. PICP is realised marginal coverage against the nominal 80%, where closer is better rather than higher; MPIW is mean interval width in demand counts; MAD is the mean absolute deviation of per-zone coverage from 80%. Grouping evens out coverage on both datasets, but the tails move in opposite directions.

Dataset	Scheme	PICP	MPIW	Per-zone MAD	Worst zone	Zones below 70%
Yellow	Global CQR	80.37%	17.72	5.41	57.4%	6.59%
Yellow	Mondrian, zone demand	80.48%	19.11	3.79	63.9%	2.92%
Yellow	Mondrian, predicted level	80.38%	19.03	3.72	61.5%	3.93%
Green	Global CQR	80.69%	5.68	4.89	69.1%	0.38%
Green	Mondrian, zone demand	80.77%	5.99	4.22	67.0%	1.90%
Green	Mondrian, predicted level	80.77%	6.02	3.45	65.8%	1.65%

We therefore test whether *Mondrian* (group-conditional) conformal calibration, which targets conditional rather than merely marginal validity for inductive conformal predictors [54], narrows it. Instead of a single widening constant, the calibration block is partitioned into groups and a separate constant is fitted within each, so the finite-sample guarantee holds inside every group. Two partitions are compared, both defined from information available at prediction time: zone demand terciles estimated on training data only, and terciles of the model's own predicted level. Crucially, all three schemes are applied to the *same* trained model within each seed, so the comparison is paired and isolates calibration from training.

absolute deviation of per-zone coverage from the nominal 80%; lower is more uniform.

Marginal coverage is preserved in every case, and per-zone uniformity improves substantially: the mean absolute deviation falls by 29–31% on yellow and by 14–29% on green, at a cost of 4–8% in mean interval width. The improvement is also more stable across seeds than the global scheme it replaces (MAD standard deviation 0.05–0.41 against 0.09–0.13 for global).

The result comes with a trade-off that we report because it runs against the headline. Improving the *average* deviation does not uniformly improve the *worst* zones, and the direction differs by dataset. On yellow, group-conditional calibration improves both: the share of zones covered below 70% falls from 6.59% to 2.92%. On green it does not — that share rises from 0.38% to 1.65%, and the worst zone falls from 69.1% to 65.8%. Green demand is sparse, so its low-prediction stratum contains zones that are intrinsically hard to cover, and giving that stratum its own smaller constant under-widens them. A manager whose commitment is governed by the worst zone rather than the average should therefore prefer zone-demand strata, which dominate on the tail metric on yellow, and should validate the tail directly rather than trusting the uniformity summary.

A sensitivity check with five strata rather than three (yellow) leaves the conclusion intact and slightly favours zone-demand grouping (MAD 3.79 against 3.72, but 2.92% against 3.93% of zones

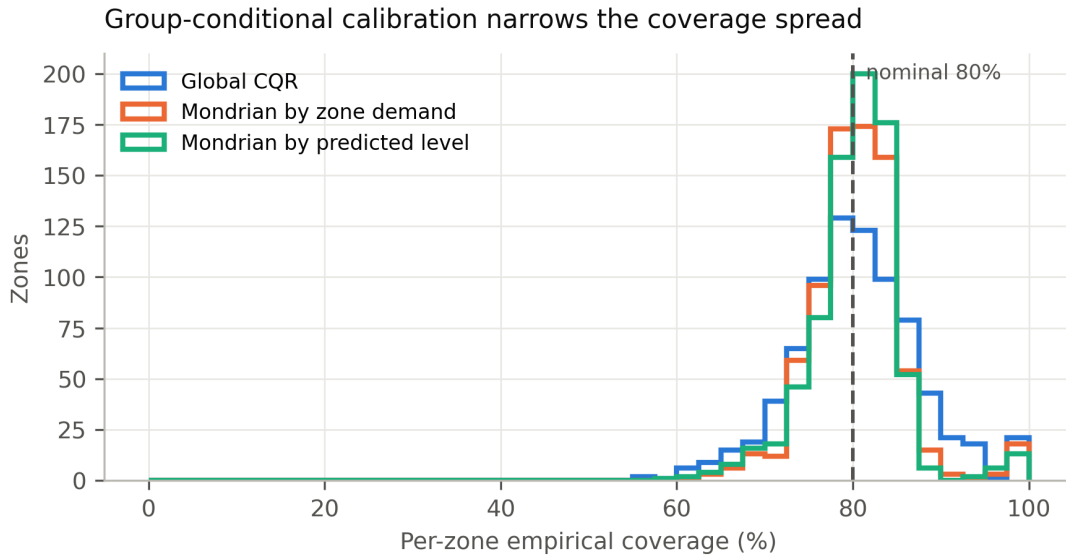


FIGURE 9. Per-zone coverage under global and group-conditional calibration. Group-conditional schemes concentrate zones nearer the nominal level.

below 70%). That comparison is not paired across bin counts, because each configuration retrains, so it is indicative rather than controlled.

For a small exact prize-collecting VRP benchmark, Held–Karp route dynamic programming and set partitioning allocate ten forecast-selected candidate zones over three capacity- and distance-constrained vehicles. Over 24 evenly spaced test hours, coverage is 31.74% for the historical average, 31.83% for ST-MRGAT, 31.82% for Graph WaveNet-Q, and 32.00% for an oracle ranking. Thus the stronger exact routing proxy confirms the main routing conclusion without claiming a city-scale exact VRP.

Every routing result above assumes that the planned stop sequence is the sequence actually driven. That assumption is testable against observed driver behaviour. Using the Konovalenko planned-versus-driven corpus [20] [50] (19,647 route pairs, 400 drivers, 249,231 stops, 32 weeks, two countries; CC BY 4.0), we measure how far execution departs from the plan. The team that released the corpus uses the same records for a different question: whether machine learning can predict, from sequential route structure, route attributes and driver information, that a driver will deviate from the planned route and by how much, treating those deviations as tacit driver knowledge rather than as execution error [50]. Only 23.3% of routes are driven exactly as planned; across all routes 44.1% of stops [95% CI 43.6, 44.5] are served out of their planned position, with a mean Kendall rank correlation of 0.741 between planned and driven order. Deviation grows with route size, from 15.8% of stops for routes of at most five stops to 65.8% for routes above twenty.

The consequence for route length is the relevant quantity. Actual and planned route lengths are almost identical on average — the actual-to-planned ratio is 0.9988 [0.9960, 1.0014] — so drivers do not systematically lengthen or shorten the tour; they reach a comparable distance by a different

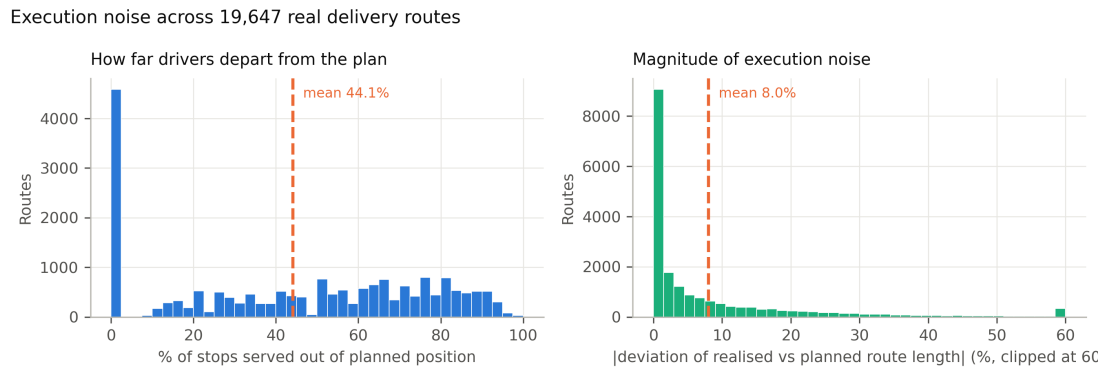


FIGURE 10. Execution noise across 19,647 observed delivery routes: the share of stops served out of their planned position, and the absolute deviation in realised route length.

order. But the *magnitude* of the per-route departure is large: the mean absolute deviation in route length is 8.56% [8.33, 8.80], with a 75th percentile of 10.2% and a 90th percentile of 25.0%. By comparison, the spread in planned route length across all forecast-based zone-selection policies in our own experiment is 5.58%, and the spread in captured demand is under one percentage point.

Execution noise is therefore about 1.5 times larger than the entire between-policy spread it would have to reveal. This does not prove the policies are identical, but it does change the interpretation of the routing null result: the differences are not merely below the resolution of our solver, they are below the resolution of real delivery execution. Two caveats bound the comparison. The corpus is a different operator, geography and period from the NYC data, so this is an order-of-magnitude argument rather than a matched control; and route length is not the same functional as captured demand, so the comparison bounds the distance channel only.

Because the study runs dozens of comparisons, some of what reaches significance will do so through the number of attempts alone. Applying the pre-specified families of Section 5.2 to the 48 DM comparisons, 23 are significant at the 5% level before adjustment, 18 survive Holm correction within dataset-and-horizon families, and 15 survive Holm correction applied across the entire study — the most conservative treatment available. Benjamini–Hochberg retains 21.

No conclusion drawn in this paper changes. The comparisons that lose significance are uniformly the weaker, longer-horizon ones — for example ST-MRGAT against MLP and LSTM at $t + 3$ on yellow, and against MLP at $t + 6$ on green — none of which carries a claim. The findings the argument rests on are unaffected. Parity between ST-MRGAT and Graph WaveNet was never significant and remains so, which is the direction that matters, since the paper’s claim is the absence of architectural superiority rather than its presence. Both ablation effects that the paper does assert on yellow survive Holm correction (removing the road relation, adjusted $p = 0.019$; removing the quantile objective, adjusted $p = 0.033$), and every green ablation comparison remains non-significant, which is what “the effects do not replicate” requires.

TABLE 21. Numbers of comparisons significant at the 5% level, before and after adjustment. The first three rows repartition one set of 48 Diebold–Mariano comparisons into families of increasing width, which is why the unadjusted count never moves; the last row is a separate set of eight ablation comparisons, corrected as two per-dataset families of four. Entries are counts of tests, not p-values, and they record no direction.

Family	Tests	Raw $p < 0.05$	Holm	Benjamini–Hochberg
Dataset x horizon	48 (6 families of 8)	23	18	20
Dataset	48 (2 families of 24)	23	15	21
Whole study	48 (1 family)	23	15	21
Ablation, per dataset	8 (2 families of 4)	2	2	2

TABLE 22. Four retrainings of one configuration with the seed held fixed, leaving nondeterministic reduction on the GPU as the only source of variation. The range as a share of the mean is this study’s resolution floor, and there is one floor per dataset and per quantity rather than a single figure for the study. Seed variability is separately larger.

Dataset	Quantity	Mean	SD	Range	Range as % of mean
Yellow	Test MAE	7.3375	0.0302	0.0609	0.83%
Yellow	Capacity cost at $q_{0.9}$	1.5958	0.0126	0.0259	1.62%
Green	Test MAE	1.9634	0.0058	0.0138	0.70%
Green	Capacity cost at $q_{0.9}$	0.4975	0.0006	0.0013	0.27%

Adjustment therefore strengthens rather than weakens the paper: the surviving significant comparisons are exactly those against the weaker baselines, while the claims of parity and non-replication are, by construction, not vulnerable to multiplicity in the first place.

Every comparison in a GPU-trained study is bounded below by run-to-run nondeterminism: reduction kernels are not deterministic, so retraining an identical configuration with an identical seed does not reproduce an identical model. Papers rarely quantify this, which makes it impossible to judge whether a reported difference is larger than the apparatus that measured it. We therefore retrained one configuration (ST-MRGAT, seed 0, loss-based stopping) four times and measured the spread:

These figures are the study’s resolution floor, and we use them to grade our own claims rather than only our baselines’:

Three consequences follow, and two of them cut against earlier readings of our own results. First, the effects the paper asserts are real: the quantile objective and the architecture gap sit well clear of the floor. Second, the road-relation effect that survived Holm correction is only 1.5x the

TABLE 23. Effects reported elsewhere in the paper, each set beside the resolution floor for its dataset. All five are MAE effects, so no cost result is graded here. The floor is a measurement limit rather than a significance threshold, and the two criteria disagree: removing the road relation survives Holm correction yet clears the floor by only half again.

Effect	Size	Floor	Verdict
Removing the quantile objective (MAE)	3.30%	0.83%	4.0x floor — resolved
Graph WaveNet vs ST-MRGAT (MAE)	2.20%	0.83%	2.7x floor — resolved
Spread of all nine neural models, green (MAE)	1.60%	0.70%	2.3x floor — barely resolved
Removing the road relation (MAE)	1.26%	0.83%	1.5x floor — marginal
Geo-only vs full multi-relational (MAE)	0.38%	0.83%	below floor — unresolvable

TABLE 24. Two early-stopping criteria compared with everything else held fixed. A positive change means the cost-selected branch was the more expensive one. No cell is significant, the sign alternates across both backbone and dataset, and the magnitudes sit close to the cost floors, so the experiment is inconclusive at its own scale rather than evidence that the criterion does not matter.

Backbone	Dataset	Cost-selected	Loss-selected	Change	p
ST-MRGAT	Yellow	1.6826	1.6153	+4.16%	0.102
ST-MRGAT	Green	0.4978	0.4992	-0.29%	0.279
Graph WaveNet-Q	Yellow	1.6082	1.6281	-1.22%	0.222
Graph WaveNet-Q	Green	0.4935	0.4921	+0.27%	0.144

floor, so it should be read as marginal rather than established — a caution we add against our own significance test. Third, and most consequentially, the entire nine-model spread on green demand is under 2x the floor. The instability of architectural rankings reported in Section 6.1 is therefore not a curiosity of that dataset; on green, model identity is barely a measurable quantity at all. That strengthens the paper’s central claim while narrowing what any architectural comparison in this setting can legitimately assert.

The seed-to-seed standard deviation of ST-MRGAT on yellow (0.085) is roughly 2.8 times the nondeterminism standard deviation (0.030), which confirms that initialisation contributes real variance beyond kernel noise and that reporting over seeds is the right unit of replication rather than a formality.

If forecasts should be *evaluated* by the decision functional, a natural question is whether they should also be *selected* by it. We tested this directly: two branches identical in every respect except the early-stopping metric on the validation-stopping block — pinball loss in one, realised $q_{0.9}$ capacity cost in the other — over three seeds, two backbones, and both datasets.

No cell is significant, the sign alternates across both backbone and dataset, and the magnitudes sit at or near the dataset-specific cost resolution floors established above. On yellow, whose floor is 1.62%, the Graph WaveNet-Q effect (1.22%) is below it and the ST-MRGAT effect (4.16%) is 2.6 times it; on green, whose floor is a much tighter 0.27%, both effects (0.29% and 0.27%) sit essentially *on* the floor. We therefore report this as inconclusive at the scale of our experiment, not as evidence that the criterion does not matter — a distinction the resolution floor makes it possible to state honestly.

The one regularity worth recording is mechanical rather than statistical: cost-selected runs stop systematically earlier (61–120 epochs against 78–150). The realised-cost metric depends only on the upper tail of the predictive distribution over 213 validation windows, whereas pinball loss averages over three quantiles and three horizons, so the former is the noisier stopping signal and triggers patience sooner.

This suggests a distinction we did not anticipate. A loss function acts on every gradient step and shapes the entire optimisation trajectory; a stopping rule only selects a point on a trajectory that has already been shaped. If the objective has already aligned training with the decision functional, the choice of point on that trajectory appears to be second-order. For practitioners the implication is narrow but useful: align the training objective with the decision, and do not expect a decision-aligned stopping rule to add to it.

7.9. Limitations. First, taxi drop-offs remain a proxy for the primary hourly forecasting task, and the external validations are drawn from separate sources rather than one joint record. Three properties would be needed to observe the full order-to-execution chain in a single dataset: (i) order or parcel creation timestamps, (ii) a measured physical vehicle capacity constraint, and (iii) a continuous GPS trace of the driven route rather than an ordered stop list. We searched the public last-mile and mobility data landscape for a source providing all three simultaneously and did not find one. The audit is reported here so that the negative claim is checkable rather than asserted:

The union of these sources covers all three properties, but no single source links them, so the joint effect cannot be measured. Concretely, this paper can show that capacity-aware quantile allocation changes which zones are served and at what committed capacity, but it cannot show whether that decision changes the physical path actually driven, because no source records the driven path continuously alongside creation times and a capacity limit. The paper therefore makes no claim about a common, long-horizon parcel order-to-route deployment with measured vehicle capacity.

One caveat on the audit itself: a negative search result is evidence of absence only within the sources examined. Proprietary carrier data plainly contains all three properties, and any future public release of such a record would supersede this limitation.

Second, fresh Graph WaveNet-Q retraining supports the external neural result in Porto, but the experiment does not repeat every NYC architecture, multi-relation ablation, calibration-slice

TABLE 25. Public sources audited against the three properties needed to follow an order through to its execution within one record. Cells take only yes, no and partial. The union of these sources covers all three properties; nothing links them, which is the gap. The finding is an absence among the sources examined, not a claim about proprietary carrier data.

Source	(i) order creation	(ii) vehicle capacity	(iii) continuous GPS route	Binding limitation
Amazon LMRRC	no	yes (vehicle cm ³ , package dimensions)	no	Packages arrive as a route manifest; execution is an ordered stop sequence with a travel-time matrix
LaDe	yes (task accept/pickup/deliver events)	no	partial	GPS is recorded at task events, not as a continuous trace; couriers have no recorded volume limit
Meituan	yes (same-order creation, dispatch, delivery)	no	no	Food delivery, eight-day window
Olist	yes (order purchase/approval timestamps)	no	no	Order records only; no vehicle or route object
Konovalenko et al. (Mendeley 10.17632/kkwgfvmtxn.1, CC BY 4.0)	no	no	no	Verified by inspection: 19,647 planned/driven route pairs, 400 drivers, 249,231 stops. Fields are stop indices, arrival times, time windows and Haversine inter-stop distances — an ordered stop list, not a road-level trace. Used separately in §7.8 to quantify execution noise
NYC TLC, Porto taxi	no	no	no	Trip endpoints only; no parcel or vehicle-capacity semantics

analysis, or routing task. It therefore supports cross-city retraining of one decision pipeline, not universal architecture transfer.

Third, the yellow capacity results include five-seed decision intervals and the green Graph WaveNet-Q replication includes three seeds. The rolling-origin check also uses three seeds per

backbone; uncertainty is reported over initialisations, not by treating correlated hourly observations as independent replicates.

Fourth, the capacity-matched pairs show that the primary architectural comparison is sensitive to model capacity; they are therefore reported as a robustness check, not as evidence of universal backbone superiority.

Fifth, the Amazon routing validation is route sequencing for already assigned single-vehicle routes. It includes vehicle capacity, service, time windows, and travel times, but does not solve fleet assignment, dynamic re-optimisation, or monetary labour and fuel costs.

Finally, standard split-conformal coverage guarantees assume exchangeability, under which they are finite-sample but only marginal [38]. Chronological calibration reduces leakage but does not make hourly observations exchangeable. The reported 80% values are empirical test-block coverage, not a proof of distribution-free validity under arbitrary temporal dependence. A deployed version of this layer would be better served by an online variant that continuously re-estimates the widening parameter as the distribution moves, which attains the intended coverage frequency over long intervals without assuming exchangeability [39].

8. MANAGERIAL IMPLICATIONS

The findings translate into guidance for service organisations that buy, build, or govern demand forecasts. We state them as decisions a manager actually faces.

8.1. Set the service level first, then ask what the forecast must estimate. The most consequential result of this study is also the least technical. A capacity plan built on a mean or median forecast is a 50% service-level policy, whether or not anyone chose that. Because the cost-minimising commitment is the τ -quantile with $\tau = c_u / (c_u + c_o)$, the target service level and the forecast functional are the same decision seen from two sides. The managerial sequence should therefore be inverted relative to common practice: first agree what fraction of demand the service intends to serve and what a denied customer costs relative to an idle courier, and only then specify what the forecasting system must produce. That ordering is the managerial counterpart of the prescriptive-analytics programme, which uses data to prescribe the decision itself rather than only to predict the quantity that enters it [41], and of feature-based newsvendor work that estimates the stocking decision directly and bypasses the intermediate demand question, in a field that has traditionally kept estimation apart from decision optimisation [42]. On our data, making that sequence explicit moved the achieved fill rate from roughly 95% to 99% on one service and from 87% to 98% on the other, at lower total cost.

One boundary is worth stating. The critical-ratio argument assumes the operator can commit a capacity level and then hold it. Where capacity itself arrives through the day — as in crowd-shipping, where in-store customers supplement company drivers and delivery capacity is itself uncertain, with both online orders and willing couriers arriving through the day [49] — the target

service level remains a managerial choice, but committing to it is a harder operational problem, and the value of flexibility in physical distribution becomes a question in its own right [46].

8.2. Do not procure or reward forecasts on accuracy alone. Forecast vendors, internal analytics teams, and model-selection pipelines are routinely evaluated on error metrics. Our results show that this criterion can select the wrong system. That the criterion a model is trained or selected on can diverge from the criterion by which its output is finally judged is a documented failure mode in stochastic optimisation, where models trained end-to-end against the task objective outperformed ones trained only to predict well in the applications studied [40]. The architecture with the lowest mean absolute error on both datasets did not deliver the lowest asymmetric operating cost on either: on yellow-taxi demand its error was 2.2% lower than a competing backbone while its capacity cost was 13.8% higher, because it under-committed slightly more often and the 9:1 penalty magnified that bias. A service-level agreement or vendor scorecard written in terms of MAE or MAPE can thus be satisfied by a system that is worse for the business. The recommendation is concrete: contract and evaluate on realised decision cost and achieved fill rate, and treat accuracy as a diagnostic rather than an objective.

8.3. Spend on the decision layer before spending on the model. For managers allocating a limited analytics budget, the relative sizes matter. Across nine neural architectures the entire span of accuracy was 1.6% on one service and 5.3% on the other, and the ranking of the middle of the field reversed between the two. Changing the decision functional on a *fixed* model moved cost by 42–51%. The same quantile-and-calibration layer produced those gains on two different backbones and in an external city. A programme that replaces its forecasting architecture while continuing to commit capacity at the median is therefore optimising the component with the smaller and less reliable effect. Upgrading the decision layer is also cheaper: it required no additional data and, on one backbone, no measurable accuracy cost.

8.4. Ask whether the decision needs a magnitude or only a ranking. Not every decision rewards forecast investment, and the distinction is diagnosable in advance. Committing capacity consumes a calibrated magnitude and was highly sensitive to forecast quality. Choosing which zones to serve consumes only a ranking, and there a simple week-hour historical average came within 0.15 percentage points of every learned model, across fleet sizes from three to fifty zones — while ignoring demand altogether lost up to 23 points. The operational reading is that urban demand hotspots are stable enough to rank without machine learning, so forecasting effort directed at zone selection is largely wasted, whereas the same effort directed at capacity sizing is not. Managers can apply this test to their own decisions by asking which functional of the demand distribution the decision consumes.

8.5. Report calibration, and expect it to hold only on average. The calibration layer reached its nominal 80% coverage on both services and both backbones without retuning, which is what makes it deployable rather than a per-case fit. But coverage was materially heterogeneous across zones

(54% to 100%). A manager who guarantees a service level per zone on the basis of a marginally calibrated interval will over-promise in some zones. Marginal calibration supports portfolio-level commitments; per-zone commitments require per-zone validation.

8.6. Treat plan adherence as part of the system, not an implementation detail. Finally, the routing analysis assumed planned sequences are executed, and real driver data shows they are not: only 23.3% of 19,647 observed routes followed the plan exactly, and the resulting variation in route length exceeded the entire spread between competing forecast policies. Investment in more precise route planning is therefore likely to be absorbed by execution variance unless adherence is managed alongside it. This is an argument for measuring adherence, not for abandoning planning — the same data shows that ignoring demand entirely is far more costly than any planning imprecision. The assumption runs deep in the routing literature as well: formulations built for uncertain demand still commit to an a priori sequence and absorb the realisation through a defined recourse, the planned order otherwise being followed as designed [47] [48]. A separate literature on dynamic vehicle routing takes up the case where the problem's inputs change during execution [44].

9. CONCLUSION

Lower forecast error is not synonymous with higher operational value. Across two NYC demand regimes, specialised spatio-temporal architectures differ only modestly in MAE, their middle rankings change, and hand-designed graph-relation effects do not replicate. The stable result lies elsewhere: a monotone quantile head with conformal calibration transfers across two backbones and reduces asymmetric capacity cost by 42–51%. The same learned forecasts provide almost no benefit over a historical average when the decision only needs a zone ranking.

An external Porto test strengthens this conclusion: a training-only historical policy achieves a 47.2% reduction in asymmetric cost when it supplies the 0.9 quantile rather than the median, and a fresh Porto Graph WaveNet-Q retraining reduces cost by a further 23.0% relative to that historical q0.9 policy. This does not establish universal architecture transfer, but it shows that the central forecast–decision alignment result and one neural decision pipeline are not confined to NYC.

The two parcel studies close distinct proxy gaps without pretending that their records are linked. Olist provides an actual order-arrival-to-capacity chain: training-only q0.9 capacity reduces cost by 53.2% versus a median policy, and a fresh three-seed Graph WaveNet-Q reduces mean cost by 69.3% relative to the historical q0.9 comparator. Amazon provides actual capacity, service, time-window, and route-sequencing objects: q0.9 policies reduce capacity cost by 42% across count, volume, and planned service time, while observed delivery sequences materially outperform a distance-only rule on both travel time and late packages. Meituan then supplies a same-order operational food-delivery record: 568,544 orders link from creation through courier assignment, pickup, a delivery wave, and final arrival, while the matched q0.9 forecasting policy reduces capacity cost. This closes the identifier-linkage gap, though its eight-day span and absence of vehicle capacity/GPS

TABLE 26. The three zone relations and how each is built. Counts are directed non-zero entries of the symmetrised adjacency, so they run to roughly twice the undirected edge count; origin-destination edges use training-period trips only.

Relation	Construction	Directed edge entries
Geographic	symmetrised 8-NN on centroid haversine distance	2,506
Road	symmetrised 8-NN on OSM shortest-path distance	2,584
OD flow	strongest training-period origin-destination flows	3,864

route traces mean that a long-horizon common parcel order-to-route deployment with measured vehicle capacity remains future work. LaDe independently adds 184 days of parcel task load and GPS event evidence, with q0.9 policy gains that replicate the decision mechanism, but correctly remains a courier-task-slot study rather than a physical-capacity route experiment.

The appropriate unit of evaluation is therefore the forecast-decision pair. Models should be judged by whether they estimate the functional required by the downstream decision and by the realised cost of using it. For the setting studied here, the defensible contribution is not a universally superior graph architecture, but a calibrated and transferable decision layer whose value is measured where forecasts are actually consumed.

APPENDIX A. SPATIAL BACKBONE SPECIFICATION

This appendix contains the architectural detail deliberately kept out of Section 4, where the paper’s argument concerns the decision layer rather than the network. None of the results in Sections 6–8 depend on these choices; Section 6.2 reports that the multi-relational construction described here does not improve accuracy.

A.1. Zone relations. Three undirected relations are constructed over the 263 zones:

Road distance is computed over a network of 55,405 nodes and 139,864 segments, with each zone centroid anchored to its nearest street node. Origin-destination edges use training-period trips only, so no relation encodes test-period behaviour. The relations are structurally distinct: geographic and road relations have Jaccard similarity 0.725, while geographic and flow relations have similarity 0.136 — two zones separated by a river are close geographically but distant by road.

A.2. ST-MRGAT. A dilated causal temporal convolution encodes each zone’s 12-hour history into $\mathbf{h}_z$. Two graph-attention layers [8] are then applied separately to each relation, and their outputs are fused by node-specific relation attention:

$$\mathbf{h}'_z = \sum_{r \in \mathcal{R}} w_{z,r} \mathbf{h}_z^{(r)}, \quad w_{z,r} = \text{softmax}_r(\mathbf{v}^\top \mathbf{h}_z^{(r)}).$$

A skip connection carries the temporal embedding to the output head. Graph attention uses four heads with self-loops; the hidden width is 64 unless a capacity-matched comparison states otherwise.

Section 6.6 examines the fused weights $w_{z,r}$ and finds that the relation receiving the highest average attention is the one ablation shows to be least useful. The weights are reported there as a caution against reading attention as evidence of component value, not as an interpretability claim.

A.3. Graph WaveNet. Graph WaveNet [5] is used as published in structure: gated dilated causal temporal convolutions with residual and skip connections, and graph convolution over forward and backward transition matrices together with a learned adaptive adjacency $\text{softmax}(\text{ReLU}(\mathbf{E}_1 \mathbf{E}_2^\top))$. Because its adjacency is learned rather than specified, it serves as a control for whether the hand-designed relations in A.1 are the limiting factor. Our implementation replaces the multi-step decoder with a direct multi-horizon head so that all compared models share one training and evaluation harness.

A.4. Training configuration. All neural models use AdamW, a maximum of 150 epochs, early stopping with patience 25 on the first 60% of the validation block, and identical inputs: the seasonal residual channels of Section 4.1 plus static zone attributes. Metrics are computed after inversion to count space. Every learned model — including Ridge and XGBoost — receives the same training-only seasonal base and predicts the same residual, so no model gains an information advantage.

Acknowledgements: This research was supported by data provided by Meituan.

Funding: This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Data Availability: Every dataset analysed here is publicly available. The primary demand data are the New York City Taxi and Limousine Commission trip records [7]. The external corpora are the Porto taxi trajectory challenge set [14], the Amazon Last Mile Routing Research Challenge data [15], the Olist Brazilian e-commerce release [16], the Meituan on-demand food delivery release [17], the Cainiao-LaDe last-mile express dataset [18], and the Konovalenko planned-versus-driven route corpus [20]. No proprietary or restricted data were used, and no new data were generated.

Conflicts of Interest: The authors declare that there are no conflicts of interest regarding the publication of this paper.

REFERENCES

- [1] H. Yao, F. Wu, J. Ke, X. Tang, Y. Jia, et al., Deep Multi-View Spatial-Temporal Network for Taxi Demand Prediction, in: Proceedings of the AAAI Conference on Artificial Intelligence, Association for the Advancement of Artificial Intelligence (AAAI), 2018. <https://doi.org/10.1609/aaai.v32i1.11836>.

- [2] X. Geng, Y. Li, L. Wang, L. Zhang, Q. Yang, et al., Spatiotemporal Multi-Graph Convolution Network for Ride-Hailing Demand Forecasting, in: Proceedings of the AAAI Conference on Artificial Intelligence, Association for the Advancement of Artificial Intelligence (AAAI), 2019, pp. 3656–3663. <https://doi.org/10.1609/aaai.v33i01.33013656>.
- [3] B. Yu, H. Yin, Z. Zhu, Spatio-Temporal Graph Convolutional Networks: A Deep Learning Framework for Traffic Forecasting, in: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, International Joint Conferences on Artificial Intelligence Organization, California, 2018, pp. 3634–3640. <https://doi.org/10.24963/ijcai.2018/505>.
- [4] Y. Li, R. Yu, C. Shahabi, Y. Liu, Diffusion Convolutional Recurrent Neural Network: Data-Driven Traffic Forecasting, in: International Conference on Learning Representations, 2018. <https://arxiv.org/abs/1707.01926>.
- [5] Z. Wu, S. Pan, G. Long, J. Jiang, C. Zhang, Graph WaveNet for Deep Spatial-Temporal Graph Modeling, in: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, International Joint Conferences on Artificial Intelligence Organization, California, 2019, pp. 1907–1913. <https://doi.org/10.24963/ijcai.2019/264>.
- [6] A.N. Elmachtoub, P. Grigas, Smart “Predict, Then Optimize”, *Manag. Sci.* 68 (2022), 9–26. <https://doi.org/10.1287/mnsc.2020.3922>.
- [7] New York City Taxi and Limousine Commission, TLC Trip Record Data and User Guide. <https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page>.
- [8] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, et al., Graph Attention Networks, in: International Conference on Learning Representations, 2018. <https://arxiv.org/abs/1710.10903>.
- [9] R. Koenker, G. Bassett, Regression Quantiles, *Econometrica* 46 (1978), 33–50. <https://doi.org/10.2307/1913643>.
- [10] Y. Romano, E. Patterson, E. Candès, Conformalized Quantile Regression, in: Advances in Neural Information Processing Systems, Curran Associates, 2019, pp. 3538–3548.
- [11] S. Jain, B.C. Wallace, Attention Is Not Explanation, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Stroudsburg, PA, USA, 2019, pp. 3543–3556. <https://doi.org/10.18653/v1/N19-1357>.
- [12] S. Wiegrefe, Y. Pinter, Attention Is Not Not Explanation, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Stroudsburg, PA, USA, 2019, pp. 11–20. <https://doi.org/10.18653/v1/D19-1002>.
- [13] F.X. Diebold, R.S. Mariano, Comparing Predictive Accuracy, *J. Bus. Econ. Stat.* 13 (1995), 253–263. <https://doi.org/10.1080/07350015.1995.10524599>.
- [14] L. Moreira-Matias, J. Gama, M. Ferreira, J. Mendes-Moreira, L. Damas, Taxi Service Trajectory - Prediction Challenge, ECML PKDD 2015, UCI Machine Learning Repository, 2015. <https://doi.org/10.24432/C55W25>.
- [15] D. Merchán, J. Arora, J. Pachon, K. Konduri, M. Winkenbach, et al., 2021 Amazon Last Mile Routing Research Challenge: Data Set, *Transp. Sci.* 58 (2024), 8–11. <https://doi.org/10.1287/trsc.2022.1173>.
- [16] Olist, Brazilian E-Commerce Public Dataset by Olist, Kaggle, 2018. <https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce>.
- [17] Meituan and INFORMS Transportation Science and Logistics Society, Meituan Operational-Level On-Demand Food Delivery Data for the 2025 INFORMS TSL Data-Driven Research Challenge, 2025. <https://github.com/meituan/Meituan-INFORMS-TSL-Research-Challenge>.
- [18] L. Wu, H. Wen, H. Hu, X. Mao, Y. Xia, et al., LaDe: The First Comprehensive Last-mile Express Dataset from Industry, in: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, ACM, New York, NY, USA, 2024, pp. 5991–6002. <https://doi.org/10.1145/3637528.3671548>.

- [19] B. Fahimnia, M. Arvan, T. Tan, E. Siemsen, A Hidden Anchor: The Influence of Service Levels on Demand Forecasts, *J. Oper. Manag.* 69 (2023), 856–871. <https://doi.org/10.1002/joom.1229>.
- [20] A. Konovalenko, L.M. Hvattum, K.A.H. Iversen, Last-Mile Delivery Route Deviations Dataset: Planned vs. Actual Routes, Mendeley Data, 2024. <https://doi.org/10.17632/kkwgfvmtxn.1>.
- [21] S. Holm, A Simple Sequentially Rejective Multiple Test Procedure, *Scand. J. Stat.* 6 (1979), 65–70. <https://www.jstor.org/stable/4615733>.
- [22] Y. Benjamini, Y. Hochberg, Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing, *J. R. Stat. Soc. Ser. B Stat. Methodol.* 57 (1995), 289–300. <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>.
- [23] T. Gneiting, Making and Evaluating Point Forecasts, *J. Am. Stat. Assoc.* 106 (2011), 746–762. <https://doi.org/10.1198/jasa.2011.r10138>.
- [24] Z. Wu, S. Pan, G. Long, J. Jiang, X. Chang, et al., Connecting the Dots: Multivariate Time Series Forecasting with Graph Neural Networks, in: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, New York, NY, USA, 2020, pp. 753–763. <https://doi.org/10.1145/3394486.3403118>.
- [25] L. Bai, L. Yao, C. Li, X. Wang, C. Wang, Adaptive Graph Convolutional Recurrent Network for Traffic Forecasting, in: *Advances in Neural Information Processing Systems*, Curran Associates, 2020, pp. 17804–17815.
- [26] C. Song, Y. Lin, S. Guo, H. Wan, Spatial-Temporal Synchronous Graph Convolutional Networks: A New Framework for Spatial-Temporal Network Data Forecasting, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, Association for the Advancement of Artificial Intelligence (AAAI), 2020, pp. 914–921. <https://doi.org/10.1609/aaai.v34i01.5438>.
- [27] J. Zhang, Y. Zheng, D. Qi, Deep Spatio-Temporal Residual Networks for Citywide Crowd Flows Prediction, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, Association for the Advancement of Artificial Intelligence (AAAI), 2017. <https://doi.org/10.1609/aaai.v31i1.10735>.
- [28] S. Brody, U. Alon, E. Yahav, How Attentive Are Graph Attention Networks?, in: *International Conference on Learning Representations*, 2022. <https://arxiv.org/abs/2105.14491>.
- [29] T.N. Kipf, M. Welling, Semi-Supervised Classification with Graph Convolutional Networks, in: *International Conference on Learning Representations*, 2017. <https://arxiv.org/abs/1609.02907>.
- [30] W.L. Hamilton, R. Ying, J. Leskovec, Inductive Representation Learning on Large Graphs, in: *Advances in Neural Information Processing Systems*, Curran Associates, 2017, pp. 1024–1034. <https://proceedings.neurips.cc/paper/2017/hash/5dd9db5e033da9c6fb5ba83c7a7e9bea9-Abstract.html>.
- [31] S. Makridakis, E. Spiliotis, V. Assimakopoulos, M5 Accuracy Competition: Results, Findings, and Conclusions, *Int. J. Forecast.* 38 (2022), 1346–1364. <https://doi.org/10.1016/j.ijforecast.2021.11.013>.
- [32] R.J. Hyndman, A.B. Koehler, Another Look at Measures of Forecast Accuracy, *Int. J. Forecast.* 22 (2006), 679–688. <https://doi.org/10.1016/j.ijforecast.2006.03.001>.
- [33] N. Reimers, I. Gurevych, Reporting Score Distributions Makes a Difference: Performance Study of LSTM-networks for Sequence Tagging, in: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Stroudsburg, PA, USA, 2017, pp. 338–348. <https://doi.org/10.18653/v1/d17-1035>.
- [34] J. Dodge, S. Gururangan, D. Card, R. Schwartz, N.A. Smith, Show Your Work: Improved Reporting of Experimental Results, in: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Association for Computational Linguistics, Stroudsburg, PA, USA, 2019, pp. 2185–2194. <https://doi.org/10.18653/v1/d19-1224>.
- [35] D. Salinas, V. Flunkert, J. Gasthaus, T. Januschowski, DeepAR: Probabilistic Forecasting with Autoregressive Recurrent Networks, *Int. J. Forecast.* 36 (2020), 1181–1191. <https://doi.org/10.1016/j.ijforecast.2019.07.001>.
- [36] B. Lim, S.Ö. Arık, N. Loeff, T. Pfister, Temporal Fusion Transformers for Interpretable Multi-Horizon Time Series Forecasting, *Int. J. Forecast.* 37 (2021), 1748–1764. <https://doi.org/10.1016/j.ijforecast.2021.03.012>.

- [37] T. Gneiting, A.E. Raftery, Strictly Proper Scoring Rules, Prediction, and Estimation, *J. Am. Stat. Assoc.* 102 (2007), 359–378. <https://doi.org/10.1198/016214506000001437>.
- [38] J. Lei, M. G'Sell, A. Rinaldo, R.J. Tibshirani, L. Wasserman, Distribution-Free Predictive Inference for Regression, *J. Am. Stat. Assoc.* 113 (2018), 1094–1111. <https://doi.org/10.1080/01621459.2017.1307116>.
- [39] I. Gibbs, E. Candès, Adaptive Conformal Inference Under Distribution Shift, in: *Advances in Neural Information Processing Systems*, Curran Associates, 2021, pp. 1660–1672.
- [40] P.L. Donti, B. Amos, J.Z. Kolter, Task-Based End-to-End Model Learning in Stochastic Optimization, in: *Advances in Neural Information Processing Systems*, Curran Associates, 2017, pp. 5484–5494.
- [41] D. Bertsimas, N. Kallus, From Predictive to Prescriptive Analytics, *Manag. Sci.* 66 (2020), 1025–1044. <https://doi.org/10.1287/mnsc.2018.3253>.
- [42] G.-Y. Ban, C. Rudin, The Big Data Newsvendor: Practical Insights from Machine Learning, *Oper. Res.* 67 (2019), 90–108. <https://doi.org/10.1287/opre.2018.1757>.
- [43] N. Boysen, S. Fedtke, S. Schwerdfeger, Last-Mile Delivery Concepts: A Survey from an Operational Research Perspective, *OR Spectr.* 43 (2021), 1–58. <https://doi.org/10.1007/s00291-020-00607-8>.
- [44] H.N. Psaraftis, M. Wen, C.A. Kontovas, Dynamic Vehicle Routing Problems: Three Decades and Counting, *Networks* 67 (2016), 3–31. <https://doi.org/10.1002/net.21628>.
- [45] M. Winkenbach, P.R. Kleindorfer, S. Spinler, Enabling Urban Logistics Services at La Poste through Multi-Echelon Location-Routing, *Transp. Sci.* 50 (2016), 520–540. <https://doi.org/10.1287/trsc.2015.0624>.
- [46] A. Snoeck, M. Winkenbach, The Value of Physical Distribution Flexibility in Serving Dense and Uncertain Urban Markets, *Transp. Res. Part A Policy Pract.* 136 (2020), 151–177. <https://doi.org/10.1016/j.tra.2020.02.011>.
- [47] D.J. Bertsimas, A Vehicle Routing Problem with Stochastic Demand, *Oper. Res.* 40 (1992), 574–585. <https://doi.org/10.1287/opre.40.3.574>.
- [48] M. Gendreau, G. Laporte, R. Séguin, An Exact Algorithm for the Vehicle Routing Problem with Stochastic Demands and Customers, *Transp. Sci.* 29 (1995), 143–155. <https://doi.org/10.1287/trsc.29.2.143>.
- [49] I. Dayarian, M. Savelsbergh, Crowdsipping and Same-Day Delivery: Employing In-Store Customers to Deliver Online Orders, *Prod. Oper. Manag.* 29 (2020), 2153–2174. <https://doi.org/10.1111/poms.13219>.
- [50] A. Konovalenko, L.M. Hvattum, K.A.H. Iversen, Predicting Last-Mile Delivery Route Deviations Using Machine Learning, *Expert Syst. Appl.* 298 (2026), 129921. <https://doi.org/10.1016/j.eswa.2025.129921>.
- [51] S. Serrano, N.A. Smith, Is Attention Interpretable?, in: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Stroudsburg, PA, USA, 2019, pp. 2931–2951. <https://doi.org/10.18653/v1/p19-1282>.
- [52] E.A. Silver, D.F. Pyke, D.J. Thomas, *Inventory and Production Management in Supply Chains*, CRC Press, 2016. <https://doi.org/10.1201/9781315374406>.
- [53] G.A. Croes, A Method for Solving Traveling-Salesman Problems, *Oper. Res.* 6 (1958), 791–812. <https://doi.org/10.1287/opre.6.6.791>.
- [54] V. Vovk, Conditional Validity of Inductive Conformal Predictors, *Mach. Learn.* 92 (2013), 349–376. <https://doi.org/10.1007/s10994-013-5355-6>.
- [55] D. Harvey, S. Leybourne, P. Newbold, Testing the Equality of Prediction Mean Squared Errors, *Int. J. Forecast.* 13 (1997), 281–291. [https://doi.org/10.1016/s0169-2070\(96\)00719-4](https://doi.org/10.1016/s0169-2070(96)00719-4).