Journal of Technology Informatics and Engineering (JTIE)
Vol. 5 No. 1 April 2026
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 441-473
DOI: https://doi.org/10.51903/jtie.v5i1.564

Cross-Dataset Parcel Workload Priors for Last-Mile Capacity Forecasting: Integrating Package Segmentation with Delivery Operation Signals

Ruiyan Ma[1], Long Zhang*[2], Annie Bai[3]
Email: longz@smu.edu
[1]*Software Engineering, UC Irvine, CA, USA*
[2]*Transportation Systems Engineering, Southern Methodist University, TX, USA*
[3]*Data Science, Columbia University, NY, USA*
*Corresponding Author

Abstract

This paper evaluates a cross-dataset framework for parcel segmentation, next-day last-mile operational-intensity forecasting, and capacity-oriented error analysis. The operational study uses 4,514,661 delivery tasks and 6,136,147 pickup tasks from LaDe-D and LaDe-P across five cities (May–October 2022), while the visual study uses 2,197 Package Segmentation images with 7,643 annotated package instances. A segmentation model estimates package count, foreground area, and instance-area dispersion to construct an additive workload descriptor. Because the images are not paired with operational records, city-day visual variables are represented as cross-dataset distributional proxies derived from empirical-rank mapping. Segmentation baselines are evaluated independently of forecasting. The random-forest pixel classifier achieves the best segmentation performance (IoU 0.5323, Dice 0.6450), outperforming YOLO11n-seg (IoU 0.3439, Dice 0.4591). Operational intensity is represented by the first principal component of delivery orders, active couriers, areas of interest, regions, and area types, explaining 89.19% of training variance. On a 230 city-day chronological holdout, ElasticNet Operational achieves the best forecasting accuracy (MAE 2.0521). Within XGBoost models, Vision-Prior records MAE 2.4988, slightly outperforming Delivery+Pickup (MAE 2.5154) and a shuffled-prior control (MAE 2.5973). However, the 0.0166 MAE improvement is not statistically significant under a paired moving-block bootstrap (95% CI: −0.0558 to 0.0319). A separate analysis of 6,112 Amazon routes and 1,457,175 packages similarly shows only a 0.24% MAE reduction from parcel-mix features. Overall, the proposed proxy provides only limited gains, while the operational ElasticNet remains the strongest baseline. Time- and location-paired visual observations are needed to establish practical operational benefits from computer vision.

Keywords: *Computer Vision, Package Segmentation, Last-Mile Delivery, Parcel Capacity Planning, Cross-Dataset Distributional Proxy.*

I. INTRODUCTION

Last-mile parcel delivery is a capacity-planning problem before it is a routing problem. Each morning, dispatchers must decide whether available courier labor, depot staging space, and local delivery windows are sufficient for the next operating day. In high-volume e-commerce systems, this decision is difficult because parcel workload is not determined by package count alone. The same daily count can represent a compact flow of regular parcels or a more difficult mixture of irregular shapes, crowded sorting scenes, and geographically dispersed delivery points. Classical logistics planning has long emphasized that transportation capacity should be tied to the physical resources consumed by freight rather than to a single abstract demand count (Crainic & Laporte, 1997; Simchi-Levi et al., 2007). Public last-mile event data and package-level image benchmarks make it possible to examine operational histories and visual parcel composition in complementary experiments, consistent with recent transport-demand and urban-dispatch studies that connect

Received on December 2025; Revised on January, 2026; Accepted on February, 2026; Published on April, 2026

forecasting with operational planning (Xin, 2026b; Zhang et al., 2025). They do not, however, create a shared observation when the datasets lack a common parcel, location, or time key.

The operational side of this study is represented by LaDe-D and LaDe-P, a large public last-mile delivery and pickup dataset from industry (Wu et al., 2024). The visual side is represented by the Package Segmentation dataset, which provides parcel images with polygon masks for package identification, sorting, smart warehouses, and logistics applications (Ultralytics, 2024). The image benchmark and LaDe are independent. Package Segmentation contains no LaDe city or date identifier, and LaDe contains no parcel photographs. The visual descriptors used in the LaDe model are consequently not obtained by joining image records to operational events. They are generated through a fully disclosed, training-period quantile mapping and are evaluated as cross-dataset distributional proxies.

The topic is timely because recent logistics-vision research has moved from parcel detection toward high-speed recognition, localization, dimension measurement, and robotic sorting. Han et al. (2020) show that visual sorting of disorderly stacked express parcels is a central automation problem. Lu et al. (2024) develop a monocular-camera approach for high-speed parcel classification and location, while Dai et al. (2025) study visual dimension measurement for parcel boxes in high-speed sorting systems. These studies emphasize the sensing side of intelligent logistics. Adjacent transportation-perception work on uncertainty-aware late fusion, LiDAR-camera tracking, and traffic-sign recognition reinforces the principle that visual signals should be calibrated before downstream planning use (Xin, 2025, 2026a; Ye et al., 2024). The present study therefore tests the image and forecasting layers separately and reserves operational claims for comparisons that the data can support.

The central question is divided into three empirically distinguishable parts. First, can parcel masks support accurate estimation of image-level object count, foreground area, and size dispersion? Second, when the Package Segmentation training distribution is transferred to LaDe through a deterministic operational-rank mapping, do the resulting variables improve next-day forecasts beyond delivery-plus-pickup features, generic nonlinear transformations, and a shuffled-prior control? Third, in a separate dataset where package dimensions and planned service workload are paired within routes, does physical parcel mix add predictive value beyond operational controls? The second question tests a cross-dataset feature construction rather than actual city-day vision, and the third provides construct-level evidence rather than a validation of the Package-Seg-to-LaDe mapping.

Journal of Technology Informatics and Engineering (JTIE)
Vol. 5 No. 1 April 2026
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 441-473
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

The contribution is threefold. First, the paper benchmarks classical and modern package segmentation methods and converts instance masks into precisely defined image descriptors. Second, it constructs a next-day LaDe operational-intensity index, documents training-only scaling and principal-component analysis (PCA), and evaluates seasonal, linear, random-forest, and extreme-gradient-boosting models with chronological and rolling-origin designs. Third, matched generic and shuffled controls, city-level errors, paired moving-block inference, high-error-day analysis, target-construction sensitivity, and a separate Amazon route experiment delimit what can be inferred from the cross-dataset transform. The image layer is technically measurable, while its operational contribution is assessed within the restrictions imposed by unpaired datasets.

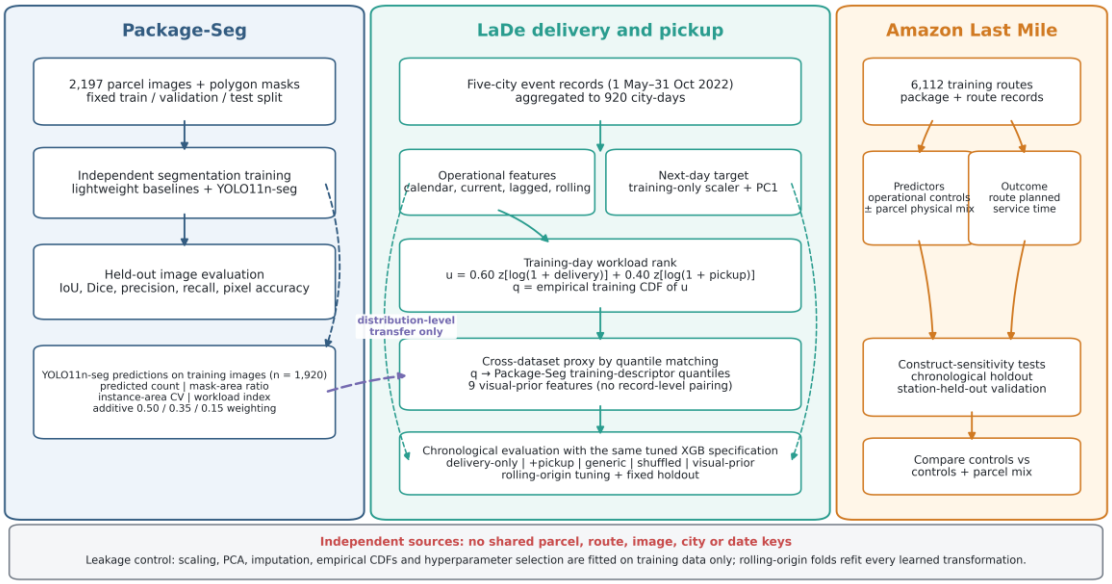


Figure 1. Study design and data boundaries. Package Segmentation supports independent segmentation evaluation, and YOLO11n-seg predictions on its training images define the descriptor distributions; LaDe operational ranks are mapped to those distributions through a training-only empirical-CDF transform; Amazon route and package records provide a separate construct-sensitivity analysis of physical parcel mix against planned service time. The datasets share no parcel, route, image, city, or date identifiers.

The absence of a shared key is treated as a design boundary rather than as an implicit data join. The Package Segmentation distribution supplies a fixed reference for a hypothesis-generating transform of LaDe operational intensity. Since every proxy is a deterministic function of the same-day delivery-pickup variables and the image training distribution, improvements could arise from nonlinear basis expansion rather than from visual information. The generic-transform and shuffled-prior controls are therefore necessary. Keeping the sensing and decision streams distinct until an explicitly defined transfer step is consistent with the discipline of calibrated late fusion

(Xin, 2025). The Amazon analysis separately pairs package dimensions with planned service workload within each route, but its dimensional features are not photographs and its route outcome is not equivalent to the LaDe city-day index.

The analyses address last-mile capacity planning at three distinct observational levels. Image segmentation is evaluated using perception and descriptor errors; LaDe models are evaluated using forecast and unitless asymmetric-error diagnostics; and Amazon is evaluated as an external construct check. Figure 1 summarizes the three branches, the points at which training-only preprocessing is fitted, and the fact that only the explicitly defined distributional proxy connects Package Segmentation to the LaDe forecasting branch.

II. LITERATURE REVIEW

Research on last-mile delivery has shown that the final leg of parcel movement is expensive, operationally constrained, and sensitive to local density. Gevaers et al. (2014) link last-mile cost to service characteristics, customer density, and delivery context, while Savelsbergh and Van Woensel (2016) describe city logistics as a domain in which operational decisions must balance service quality, congestion, and resource use. These ideas motivate a forecasting target that includes more than package count. A daily operational-intensity signal can include demand scale, observed labor deployment, spatial breadth, and operating diversity, provided that it is not interpreted as a pure measure of demand or required staffing.

City-logistics research further explains why aggregated daily capacity still matters in a field often dominated by route-level algorithms. Recent graph traffic forecasting and spatial risk-discovery work similarly treats urban operations as a spatially structured planning problem rather than as isolated local predictions (Zhang & Zhang, 2024; Zhong et al., 2023b). Last-mile systems operate under strong temporal constraints: couriers must receive tasks, load parcels, travel through local service areas, and complete customer interactions within limited windows. Even when routing algorithms are sophisticated, poor upstream planning can leave the operation with too few couriers or too much unprocessed volume at the start of the day. Daily forecasting is therefore a planning layer beneath route optimization and above strategic network design (Savelsbergh & Van Woensel, 2016).

Capacity forecasting also differs from demand forecasting because operational scale and realized supply can move together. A delivery day with the same number of packages may involve different numbers of areas of interest (AOIs), couriers, regions, and AOI types. The LaDe fields make these dimensions observable at the city-day level. The composite target used here is consequently called an operational-intensity index: it combines demand-side variables with

active-courier deployment and should not be interpreted as a direct estimate of latent demand or optimal staffing. Demand-only and active-courier targets are examined separately in sensitivity analysis.

Forecasting research supports the use of structured covariates. Hyndman and Athanasopoulos (2021) emphasize that operational forecasts should exploit trend, seasonality, and known explanatory variables. Recent transport-demand and infrastructure-workload studies make this logic more operational by pairing multi-horizon histories with risk curves, transfer settings, or uncertainty envelopes (Chen et al., 2024; He et al., 2023, 2024, 2025; Zhao et al., 2025). Machine-learning methods provide flexible alternatives when relationships are nonlinear. Random forests aggregate many decision trees and handle interactions with limited tuning (Breiman, 2001). Gradient boosting fits residual structure sequentially and is well suited to tabular prediction problems (Friedman, 2001), while XGBoost adds regularization, shrinkage, and efficient implementation for structured data (Chen & Guestrin, 2016). These properties make tree-based models appropriate for city-day logistics panels with calendar, lag, rolling, pickup, delivery, and proxy features.

The computer-vision literature provides the methodological context for the image experiment. ImageNet demonstrated the value of large labeled visual benchmarks for object recognition (Russakovsky et al., 2015), and COCO extended benchmark culture to detection and segmentation tasks with object masks and rich annotations (Lin et al., 2014). Work on calibrated vision models and autonomous-driving image detection further motivates evaluating visual pipelines for robustness and uncertainty rather than only apparent accuracy (Ling et al., 2024; Zhong et al., 2023a). For dense prediction, U-Net showed that encoder-decoder segmentation can learn object masks from limited annotated images (Ronneberger et al., 2015), Mask R-CNN provided a strong instance-segmentation architecture (He et al., 2017), and the Segment Anything framework highlighted the value of general-purpose segmentation representations (Kirillov et al., 2023). These developments motivate including a current instance-segmentation baseline in addition to compact thresholding and pixel-classification methods.

Segmentation is suitable for parcel-scene measurement because masks directly support foreground-area and instance-shape calculations. Bounding boxes can provide coarser estimates of visible scale, whereas instance masks represent occupied pixels and separate annotated objects. Semantic masks and instance masks must nevertheless be distinguished. A binary pixel classifier estimates the union foreground; counts and per-object dispersion from that prediction require connected-component post-processing and may merge touching parcels. The modern instance-

segmentation baseline predicts separate instances directly (He et al., 2017; Kirillov et al., 2023; Ronneberger et al., 2015).

Real-time vision systems create a tradeoff among accuracy, speed, and interpretability. YOLO-style models are widely used because they connect detection accuracy with fast inference, while random forests and linear models remain useful for compact supervised baselines and tabular fusion (Breiman, 2001; Redmon et al., 2016). The image experiment therefore includes thresholding, a logistic pixel classifier, a random-forest pixel classifier, and a pretrained YOLO-family instance-segmentation model fine-tuned on the Package Segmentation training split. Descriptor extraction from Package Segmentation remains separate from the later operational-rank mapping used in LaDe.

Computer vision in logistics is not a hypothetical extension of general object segmentation. Han et al. (2020) study visual sorting of express parcels in complex stacked scenes, showing the importance of detection and grasp-position estimation for automated parcel handling. Lu et al. (2024) use a monocular-camera pipeline for high-speed logistics parcel classification and localization, motivated by the limits of traditional light-curtain positioning. Dai et al. (2025) focus on parcel-box dimension measurement in high-speed sorting systems. These studies demonstrate that package recognition, parcel localization, and visual measurement are established topics in intelligent logistics. The Package Segmentation dataset fits this stream because its polygon labels identify package instances and their foreground area (Ultralytics, 2024). Whether those descriptors explain a separate operational outcome is an additional empirical question.

The LaDe dataset provides the operational counterpart. Wu et al. (2024) introduced LaDe as a comprehensive public last-mile express dataset containing pickup and delivery data from multiple cities. Such data make it possible to study forecasting, estimated time of arrival, route prediction, and spatio-temporal demand. In this paper, LaDe-D and LaDe-P are used to construct a daily operating panel. Delivery events measure realized delivery activity, while pickup events provide a distinct upstream operating stream. This separation follows the broader supply-chain idea that upstream signals can help anticipate downstream capacity stress (Simchi-Levi et al., 2007), but no individual LaDe pickup is assumed to become a particular later delivery in the panel. E-commerce next-purchase studies make a parallel point at the demand-generation layer, where customer representation learning anticipates future retail actions before they become fulfillment workload (Lu & Zhou, 2023; Xin, 2026c).

The software and empirical design also follow established data-science practice. Goodfellow et al. (2016) describe representation learning as the process of converting raw signals into useful

predictive inputs. Here, the image annotations yield interpretable descriptor distributions, while the LaDe transfer yields deterministic proxy variables whose incremental value must be tested against matched controls. Pedregosa et al. (2011) emphasize consistent model evaluation in scikit-learn, and McKinney (2010) provides the data-frame foundation used for aggregation and analysis in Python. This literature supports separate, transparent evaluation of image representation, operational dynamics, and supervised forecasting without treating independent datasets as paired observations.

III. RESEARCH METHOD

A. Data sources and empirical architecture

The empirical design has three analyses. The first evaluates package segmentation and image descriptors on Package Segmentation. The second forecasts next-day city-level operational intensity from LaDe-D and LaDe-P. The third tests whether parcel dimensions add predictive value for planned route service workload in the Amazon Last Mile model-building data. The final LaDe panel contains 920 city-days across Chongqing, Hangzhou, Jilin, Shanghai, and Yantai from May 1 through October 31, 2022. Amazon is analyzed separately and is never merged with LaDe or Package Segmentation. Table 1 identifies the exact units, files, coverage, sources, terms, and empirical roles.

Table 1. Dataset sources, units, coverage, and empirical roles.

Dataset	Analysis unit and records used	Coverage / files	Official source and terms	Role and linkage
LaDe-D	4,514,661 delivery tasks	Five city files delivery_{cq,hz,jl,sh,yt}.csv; 2022-05-01 to 2022-10-31	LaDe repository and research-use terms (Wu et al., 2024)	Aggregated to city-day delivery predictors and operational-intensity target
LaDe-P	6,136,147 pickup tasks	Five city files pickup_{cq,hz,jl,sh,yt}.csv; 2022-05-01 to 2022-10-31	LaDe repository and research-use terms (Wu et al., 2024)	Aggregated to city-day upstream operational predictors
Package Segmentation	2,197 images; 7,643 annotated instances	1,920 train, 188 validation, 89 test; package-seg.zip	Package Segmentation dataset page and source terms (Ultralytics, 2024)	Independent image benchmark and fixed descriptor distribution; no shared key with LaDe
Amazon Last Mile	6,112 model-building routes; 1,457,175 packages	route_data.json and package_data.json; 17 stations and 39 dates in 2018	AWS Open Data Registry; CC BY-NC 4.0 (Merchán et al., 2022)	Separate route-level construct-validation analysis with package and workload variables paired within route

Figure 1 shows that preprocessing is fitted separately for each branch. For LaDe, the principal-component scaler, PCA loadings, imputer, ElasticNet scaler, operational-intensity distribution, descriptor quantile functions, and hyperparameters are fitted on the applicable training data. For

rolling-origin validation, each item is re-estimated inside the training portion of the fold. Package Segmentation test images and all LaDe holdout rows remain outside model selection.

B. Package annotations and image descriptors

Package Segmentation uses YOLO-style polygon labels for a single package class. Exact duplicate polygon rows within a label file are removed before rasterization. For image *i*, the package count is the number of retained polygon instances, N_i. Each polygon is rasterized at the native 640-by-640 image resolution. The foreground-area ratio A_i is the fraction of pixels covered by the union of all package masks. If at least two instances are present, instance-area dispersion CV_i is the population standard deviation of individual mask areas divided by their mean; it is set to zero for zero- or one-instance scenes. Table 2 reports the resulting split statistics.

Table 2. Package Segmentation split and descriptor statistics.

Split	Images	Package instances	Mean packages/image	Mean mask-area ratio	P90 mask-area ratio	Mean instance-area CV
Train	1,920	6,625	3.4505	0.0710	0.1388	0.4087
Validation	188	693	3.6862	0.0798	0.1497	0.4435
Test	89	325	3.6517	0.0710	0.1382	0.4023

The additive image workload descriptor is

$$W_i = 0.50z_{\text{train}}(N_i) + 0.35z_{\text{train}}(A_i) + 0.15z_{\text{train}}(CV_i),$$

where every mean and standard deviation is estimated from the relevant 1,920-image training descriptors. Table 2 uses annotated masks to describe the benchmark. For the forecasting proxy, the same count, area, dispersion, and additive formula are applied to YOLO11n-seg predictions on the Package Segmentation training images, and standardization is fitted to that predicted training distribution. Count receives the largest prespecified weight, followed by occupied foreground area and then within-image size dispersion. The weights are not learned from LaDe and are not presented as an optimal operational utility function. The separate descriptors, the combined descriptor, generic controls, and a shuffled control are all retained so that any result does not depend on the index alone.

C. *Package segmentation experiment*

Four compact baselines and one modern instance-segmentation baseline are evaluated. Otsu thresholding tests both bright- and dark-foreground polarity. The morphological variant applies elliptical opening and closing with kernels of 3, 5, 7, or 9 pixels. Polarity, kernel size, and a connected-component minimum area are selected on validation images. The selected Otsu

settings use bright polarity and a 4,096-pixel component threshold at 640 pixels; the morphological variant additionally uses a 9-pixel kernel.

The logistic and random-forest pixel classifiers use 13 features per pixel: BGR, HSV, and Lab color channels; grayscale intensity; Sobel-gradient magnitude; and normalized x and y position. Images are resized to 320 by 320. Up to 128 foreground and 128 background pixels are sampled without replacement from every training image with seed 42. The logistic model uses standardized inputs, $C = 1$, balanced class weights, and at most 300 iterations. The random forest uses 100 trees, maximum depth 20, minimum leaf size 5, square-root feature subsampling, and balanced subsample weights. Probability thresholds from 0.20 to 0.80 and component areas from 4 to 256 pixels are selected on validation Dice and count error; the selected thresholds are 0.60 for logistic regression and 0.75 for random forest, with a 256-pixel component filter. Eight-connected components convert each binary prediction to estimated object count and component-area coefficient of variation. This post-processing does not guarantee correct separation of touching parcels, so count and dispersion errors are reported explicitly.

The modern architecture comparator uses Ultralytics 8.4.101 YOLO11n-seg, initialized from the 2.84-million-parameter COCO-pretrained checkpoint. The fixed 1,920/188/89 train-validation-test split is retained, with deterministic seed 42 and CPU training. Two-stage fine-tuning comprises one 320-pixel epoch with all trainable layers, batch size 16, and AdamW with initial learning rate 0.002, final learning-rate fraction 0.01, momentum 0.937, and weight decay 0.0005, followed by one 160-pixel epoch that freezes model layers 0-9, uses batch size 32, and lowers the initial learning rate to 0.001. Hue-saturation-value, translation (0.10), scale (0.50), and horizontal-flip (0.50) augmentation are active; mosaic and copy-paste are disabled. Validation and test inference return to 320-pixel inputs with class-aware nonmaximum suppression at IoU 0.70 and at most 300 detections.

A confidence threshold of 0.30 maximizes validation mean Dice for the reported union-mask metrics, while 0.50 minimizes validation instance-count MAE for generating the 1,920 training-image descriptor distribution. Official box and mask average precision are also calculated at confidence 0.001. Operating-threshold timing uses prediction batches of four, whereas official average precision uses batches of 16 with a confidence floor of 0.001; random-forest timing is measured per image. These CPU measurements are therefore configuration-specific rather than a paired speed benchmark. Because this is a short two-stage update with one seed, it is used to broaden the architecture comparison, not to establish the best attainable deep-segmentation performance.

Segmentation metrics are computed per test image and then averaged. Intersection over union (IoU) and Dice measure foreground overlap; precision, recall, and pixel accuracy describe binary classification; count absolute error and object-area-CV absolute error evaluate the connected-component or instance outputs; foreground-area absolute error evaluates the workload-area descriptor; and inference milliseconds per image describe runtime. Nonparametric 95% intervals use 10,000 image-level bootstrap replicates. Ground-truth polygons are used for evaluation and benchmark description. The distribution transferred to the LaDe mapping is generated by the fixed YOLO11n-seg comparator on Package Segmentation training images, without Package Segmentation test labels or LaDe images.

Table 3. LaDe city-level descriptive summary, May 1-October 31, 2022.

City	Delivery tasks	Pickup tasks	Mean delivery tasks/day	Mean pickup tasks/day	Mean delivery couriers/day	Mean pickup couriers/day	City-days
Chongqing	931,351	1,172,703	5,061.69	6,373.39	372.40	454.72	184
Hangzhou	1,861,600	2,130,456	10,117.39	11,578.57	389.51	647.44	184
Jilin	31,415	261,801	170.73	1,422.83	10.49	102.61	184
Shanghai	1,483,864	1,424,406	8,064.48	7,741.34	381.81	520.84	184
Yantai	206,431	1,146,781	1,121.91	6,232.51	97.51	388.70	184

D. LaDe city-day construction

The city-day panel is constructed in the following order.

1. The five delivery CSV files and five pickup CSV files are read with a shared schema. The source ds field is a month-day code and is converted to a 2022 calendar date without a timezone transformation; timestamps are used as recorded local operation times.
2. Each delivery task contributes one row to the city-date count. Distinct courier_id, aoi_id, region_id, and aoi_type values yield active couriers, AOIs, regions, and AOI types. No task row is removed merely because its remaining fields match another row.
3. Delivery service minutes are calculated from accept_time to delivery_time; values outside 0 to 1,440 minutes are excluded only from service-time summaries. Median and 90th-percentile service time and the 90th-percentile completion minute are retained as external criteria, not as forecasting predictors or PCA components.
4. Pickup rows are aggregated analogously using task count and distinct courier_id, aoi_id, region_id, and aoi_type. Pickup cycle-time summaries are not used in the forecasting models.

Journal of Technology Informatics and Engineering (JTIE)
Vol. 5 No. 1 April 2026
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 441-473
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

5. A complete Cartesian grid of five cities and 184 dates is left-joined to the delivery and pickup summaries. Missing city-date count fields are set to zero, producing 920 city-days. The target is then shifted one day forward within city.

The raw totals and daily scales differ substantially across cities, as shown in Table 3. Pickup and delivery are treated as separate operating streams; the construction does not assume that an individual pickup task can be linked to a later delivery task.

E. *Operational-intensity target*

The target combines five delivery-side components: orders, active couriers, AOIs, regions, and AOI types. For the primary holdout, a standard scaler is fitted using rows dated no later than September 14, 2022. PCA is then fitted to those standardized training-period components. The sign of the first component is oriented so that its loadings sum positively, and every row is transformed with the fixed scaler and loading vector (1):

$$C_{c,d} = 50 + 10 \sum_{j=1}^{5} \ell_j \frac{x_{j,c,d} - \mu_{j,\text{train}}}{s_{j,\text{train}}}, \qquad y_{c,d} = C_{c,d+1}. \quad (1)$$

The 50-plus-10 scaling gives the unitless index a convenient location and spread; it does not convert the score into couriers, packages, money, or labor hours. All predictors for row (c, d) are dated at or before day *d*, whereas the target is the transformed index for day *d+1*. Current-day versions of component variables are valid one-day-ahead predictors, not same-day target values. To address persistence and construct concerns, the analysis removes the current capacity index, predicts a demand-only PCA index that excludes couriers, and predicts a separately standardized active-courier index.

Table 4. Training-period PCA loadings, date-bootstrap intervals, and explained variance.

Quantity	Main estimate	Bootstrap mean	95% interval
Delivery orders loading	0.4371	0.4371	[0.4353, 0.4388]
Delivery couriers loading	0.4598	0.4597	[0.4577, 0.4617]
Delivery AOIs loading	0.4644	0.4644	[0.4631, 0.4657]
Delivery regions loading	0.4593	0.4593	[0.4582, 0.4605]
Delivery AOI types loading	0.4133	0.4134	[0.4098, 0.4171]
PC1 explained-variance ratio	0.8919	0.8922	[0.8853, 0.8988]

Table 4 reports the primary loading vector, 1,000 date-bootstrap intervals that resample whole dates and retain the five-city cluster, and the variance explained. The intervals quantify loading stability within the observed training period. Because PCA is pooled across cities, the index captures both temporal movement and cross-city operating scale; city indicators and within-city histories are therefore included in the forecasting models, and the score is not described as a pure latent staffing requirement.

F. *Feature engineering and cross-dataset proxy construction*

Calendar and city variables comprise day of week, month, day of year, week of month, weekend, a linear date trend, and four city indicators with Chongqing as the reference. Current delivery and pickup blocks each contain five counts plus per-courier, per-AOI, and intensity-ratio variables. For each selected base variable, lag 1, lag 7, prior 7-day mean, prior 14-day mean, prior 7-day standard deviation, and the 7-over-14 mean ratio are calculated within city. Rolling windows are shifted by one day before calculation and therefore end on day *d-1*. Same-day current variables are separately available for the day-*d* forecast. This use of decision-time histories is consistent with workload and inventory studies that construct operational features before the allocation decision (He et al., 2026; Zhao et al., 2025). Table 5 enumerates every block and its count.

Table 5. Forecast feature inventory.

Feature block	Count	Variables / construction
Calendar and city	10	day of week, month, day of year, week of month, weekend, linear trend, and four city indicators
Current delivery	8	orders, couriers, AOIs, regions, AOI types, orders/courier, orders/AOI, current operational-intensity index
Delivery lag and rolling	30	lag 1, lag 7, shifted rolling 7/14 mean, shifted rolling 7 standard deviation, and 7/14 trend for orders, couriers, AOIs, regions, and index
Current pickup	8	orders, couriers, AOIs, regions, AOI types, orders/courier, orders/AOI, pickup/delivery ratio
Pickup lag and rolling	36	the same six histories for pickup orders, couriers, AOIs, regions, AOI types, and pickup/delivery ratio
Cross-dataset proxy	9	four mapped descriptor quantiles and five interactions with standardized delivery, pickup, total intensity, or flow ratio
Generic-transform control	9	rank, rank squared/cubed/square-root/logit/sine/cosine, and rank interactions with delivery and pickup
Shuffled-prior control	9	four permuted descriptor sequences and five matched interactions

The cross-dataset mapping is performed step by step. Within the applicable training window, log delivery and pickup counts are standardized, and current operating intensity show in (2).

$$I_{c,d} = 0.60 z_{\text{train}}\{\log(1 + D_{c,d})\} + 0.40 z_{\text{train}}\{\log(1 + P_{c,d})\}. \qquad (2)$$

Its empirical training cumulative probability using (3).

$$q_{c,d} = \frac{\#\{I_{c',t} \le I_{c,d} : (c', t) \in \text{training}\}}{n_{\text{training}} + 1}, \qquad (3)$$

clipped to [0.001, 0.999]. The rank is pooled over all five training cities, as indicated by c'. For each Package Segmentation descriptor $S \in \{N, A, CV, W\}$, the proxy is $V^S_{c,d} = Q^S_{\text{PackageSeg,train,pred}}(q_{c,d})$, using linear empirical quantiles of the fixed YOLO11n-seg predictions on 1,920 training images. Five interactions combine these four proxies with standardized delivery, pickup, total operating intensity, or the standardized log pickup-to-delivery

Journal of Technology Informatics and Engineering (JTIE)
Vol. 5 No. 1 April 2026
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 441-473
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

ratio. The next-day target never enters the mapping. Ties share the same empirical rank, and values beyond the training range are bounded by the clipped quantiles.

This construction does not reconstruct parcel mix for a LaDe city-day. It is a deterministic, monotone feature transformation $V = g(D, P)$ whose shape is calibrated by an independent image distribution. Two matched controls test the source of any apparent gain. The generic block supplies nine nonlinear functions and interactions of the same rank without Package Segmentation values. The shuffled block permutes each image descriptor sequence before rank indexing and retains the same number and form of interactions. A specific visual-distribution effect would require the vision-prior model to improve on delivery plus pickup and on both controls.

G. *Forecasting models, preprocessing, and validation*

Eight models are compared: Seasonal Naive, ElasticNet Operational, random forest (RF) Operational, XGB Delivery-only, XGB Delivery+Pickup, XGB Vision-Prior, XGB Generic-Transform Control, and XGB Shuffled-Prior Control. "XGB" denotes the XGBoost estimator. Training-only median imputation is used for every learned model. ElasticNet features are additionally standardized with a scaler fitted on training rows; tree inputs are not scaled. The capacity-component scaler and PCA, empirical intensity distribution, Package Segmentation quantile mapping, and imputer are refitted for each rolling-origin fold.

Hyperparameters are selected by mean absolute error (MAE) over three expanding-window validation folds. The ElasticNet grid crosses alpha {0.0001, 0.001, 0.01, 0.1} with the L1 ratio {0.1, 0.2, 0.5, 0.8}. The RF grid crosses trees {250, 500}, depth {6, 8, unrestricted}, and minimum leaf size {1, 2, 4}. The XGBoost grid crosses trees {200, 400}, depth {2, 3, 4}, and learning rate {0.03, 0.05}, with subsample and column-subsample rates fixed at 0.85 and L2 regularization fixed at 1. The XGBoost choice is selected on the delivery-plus-pickup operational block and then held fixed across all five XGBoost information variants. Table 6 gives the selected specifications.

Table 6. Forecasting model specifications after training-fold selection.

Model	Selected estimator settings	Feature set
Seasonal Naive	operational-intensity index at *t*-7	weekly persistence
ElasticNet Operational	alpha=0.1; L1 ratio=0.8; maximum iterations=20,000	92 operational variables; standardized
RF Operational	500 trees; maximum depth=8; minimum leaf=4	92 operational variables
XGB Delivery-only	200 trees; depth=2; learning rate=0.05; subsample=0.85; column subsample=0.85; L2=1	48 delivery variables
XGB Delivery+Pickup	same selected XGBoost settings	92 operational variables
XGB Vision-Prior	same selected XGBoost settings	92 operational + 9 proxy variables
XGB Generic-Transform Control	same selected XGBoost settings	92 operational + 9 generic variables
XGB Shuffled-Prior Control	same selected XGBoost settings	92 operational + 9 shuffled variables

After 14 days of history and a one-day target shift are constructed, 845 modeling rows remain from May 15 through October 30, 2022. The main training set contains 615 rows through September 14, and the final chronological holdout contains 230 rows from September 15 through October 30; their targets span the following dates through October 31. Table 7 also provides the exact rolling-origin folds used for selection and temporal stability assessment.

Table 7. Main chronological split and rolling-origin folds.

Split / fold	Training dates and rows	Validation or holdout dates and rows	Purpose
Main training	2022-05-15 to 2022-09-14; 615	-	final estimation and preprocessing fit
Main holdout	-	2022-09-15 to 2022-10-30; 230	final out-of-sample comparison
Fold 1	2022-05-15 to 2022-07-31; 390	2022-08-01 to 2022-08-15; 75	expanding-window validation
Fold 2	2022-05-15 to 2022-08-15; 465	2022-08-16 to 2022-08-31; 80	expanding-window validation
Fold 3	2022-05-15 to 2022-08-31; 545	2022-09-01 to 2022-09-14; 70	expanding-window validation

The forecasting analysis uses Python 3.12.13, NumPy 2.3.5, pandas 2.2.3, scikit-learn 1.8.0, XGBoost 3.0.2, and SciPy 1.17.0; the image analysis uses Ultralytics 8.4.101, PyTorch 2.7.1 (CPU), and OpenCV 4.11.0. The LaDe forecasting random seed is 20260718, the image seed is 42, and the Amazon route-analysis seed is 20260719. These implementations follow the data-frame and estimator interfaces described by McKinney (2010) and Pedregosa et al. (2011).

Forecast metrics are MAE, root mean squared error (RMSE), weighted absolute percentage error (WAPE), coefficient of determination (R-squared), and signed bias $\overline{\hat{y} - y}$. City-level metrics are computed over 46 holdout rows per city. For model m, the unitless asymmetric underprediction loss using eq. (4).

$$L_w(m) = \sum_i \left[w\max\left(y_i - \hat{y}_{i,m}, 0\right) + \max\left(\hat{y}_{i,m} - y_i, 0\right) \right]. \quad (4)$$

The primary weight is $w = 2$, with sensitivity at 1, 1.5, 3, and 5. This score is a preference-weighted error on the PCA index, not an observed monetary shortage cost. The mean and 90th percentile of $\max(y - \hat{y}, 0)$, the mean of $\max(\hat{y} - y, 0)$, the fraction within five index points, and the underprediction fraction are also reported; the one-sided means include all city-days, with zero assigned when the corresponding error direction does not occur. This risk-aware framing is consistent with capacity-forecasting studies that pair prediction metrics with uncertainty or demand-envelope constraints (Chen et al., 2024; He et al., 2023), while retaining the correct unitless interpretation here.

Pairwise uncertainty uses 10,000 circular moving-block bootstrap replicates with seven-date blocks. For each sampled date, all five city errors remain together; the reported contrast is MAE for XGB Vision-Prior minus comparator MAE. A negative contrast favors the proxy. The two-sided bootstrap probability is twice the smaller tail probability. The high-error analysis defines the largest decile of absolute XGB Vision-Prior errors on the final holdout.

H. Amazon paired-route construct validation

The 2021 Amazon Last Mile Routing Research Challenge dataset contains 9,184 historical routes overall, including a public 6,112-route model-building partition with route-, stop-, and package-level information (Merchán et al., 2022). This study uses only route_data.json and package_data.json from that model-building partition. The two files contain 17 stations, 39 dates from July 19 through August 26, 2018, and 1,457,175 packages. Planned service time is present for every package, and 1,457,151 packages (99.998%) have complete positive dimensions. The 24 packages without a complete positive dimension tuple are omitted only from physical parcel summaries; all 6,112 routes remain in the analysis. Numeric median imputation and categorical most-frequent imputation are fitted within each training fold before one-hot encoding and model estimation.

The outcome is route planned service workload, defined as the sum of package-level planned-service-time seconds. The operational-control specification contains package count, drop-off-stop count, packages per stop, station indicators, sine and cosine of day of week, and date ordinal. The parcel-mix augmentation adds total, mean, and 90th-percentile package volume, volume coefficient of variation, maximum dimension, and vehicle fill ratio. Planned service time, route quality, scan outcomes, coordinates, and time-window outcomes are excluded from the predictors.

Within each station, the earliest 80% of dates form the training data and the latest 20% form the holdout; every station-date remains intact. This produces 4,604 training routes and 1,508 holdout routes across 17 stations and 14 holdout dates. Random-forest settings are selected within training data using three expanding date folds. The selected operational-control forest uses 350 trees, depth 10, minimum leaf 8, and feature fraction 0.7; the parcel-mix forest uses 350 trees, depth 10, minimum leaf 2, and feature fraction 0.7. Five-fold station-held-out validation tests geographic transfer. Paired uncertainty uses 10,000 station-date cluster-bootstrap replicates for the chronological holdout, 10,000 station-cluster replicates for station-held-out validation, and a three-date moving-block temporal sensitivity check.

IV. RESULT AND DUSCUSSION

A. Result

The annotation audit yields 7,643 package instances across 2,197 images, with an overall mean of 3.48 instances and a mean foreground-area ratio of 0.0717. The three splits have broadly similar averages, although the validation split has somewhat greater occupied area and dispersion than the training split. These statistics characterize the image benchmark only; similarity across its splits does not establish that its parcel mix represents any LaDe city-day. Figure 2 shows representative images, reference polygons, random-forest foreground estimates, and YOLO11n-seg instance predictions. The examples make both successful localization and remaining failures under clutter, contact, and occlusion visible.

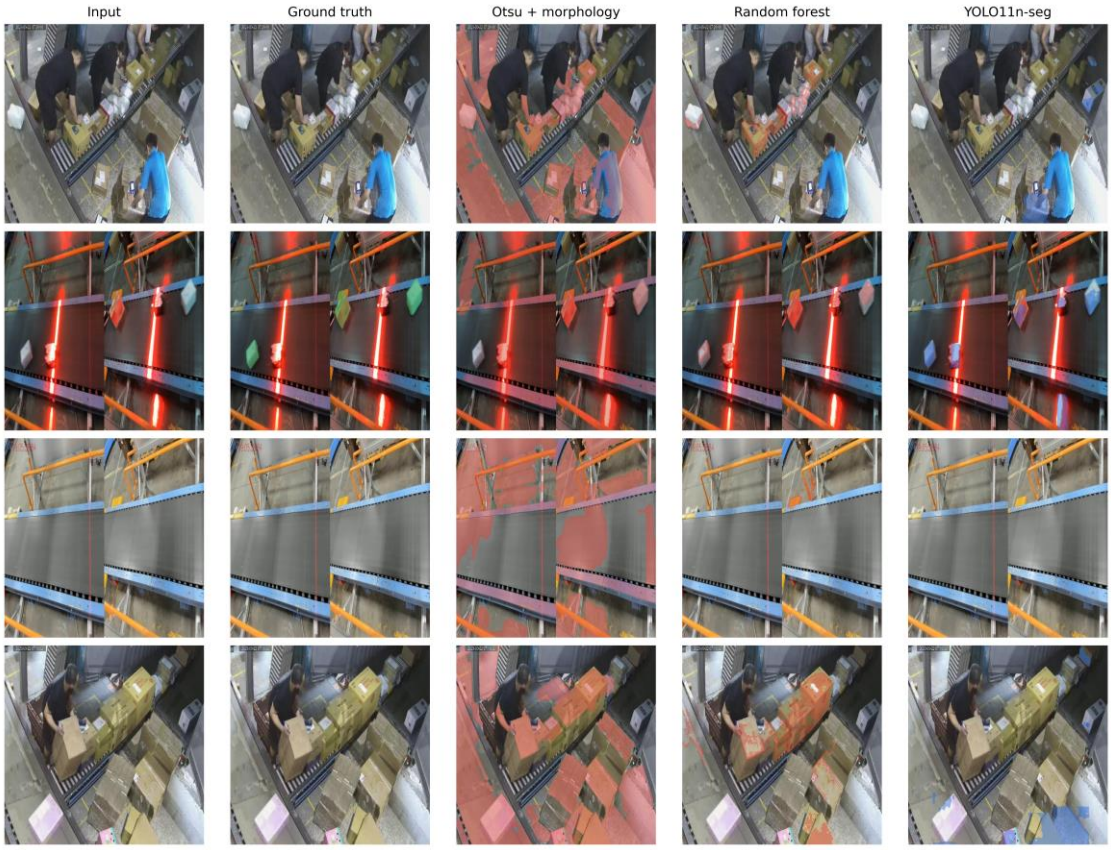
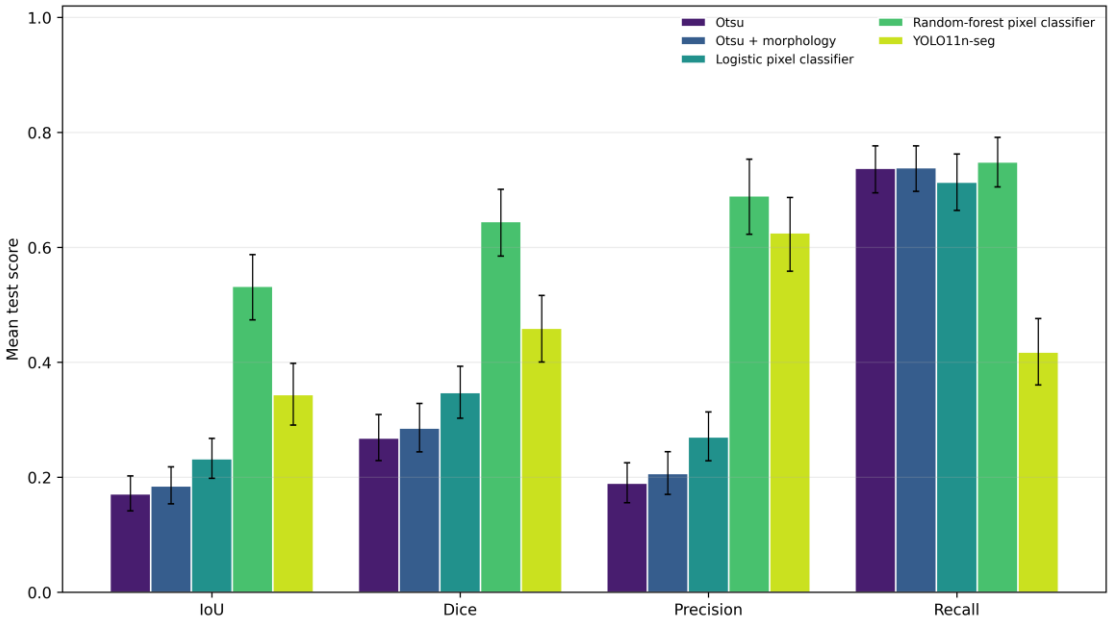


Figure 2. Qualitative results on representative Package Segmentation test images. Columns show the source image, reference masks, Otsu-plus-morphology foreground, random-forest foreground, and YOLO11n-seg instances; the rows cover difficult, typical, strong, and empty-scene cases.

Table 8 and Figure 3 compare all methods on the 89-image test split. The random-forest pixel classifier has the strongest union-mask and descriptor performance: mean IoU is 0.5323 (95% interval 0.4736-0.5874), Dice is 0.6450 (0.5846-0.7008), precision is 0.6897, recall is 0.7484,

Journal of Technology Informatics and Engineering (JTIE)
Vol. 5 No. 1 April 2026
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 441-473
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

foreground-area absolute error is 0.0216, and count absolute error is 1.4382. Its eight-connected components can nevertheless merge touching parcels, and mean CPU inference is 533.6 ms per image. Under the operating-threshold configuration described above, YOLO11n-seg requires 118.6 ms per image but has IoU 0.3439, Dice 0.4591, precision 0.6254, recall 0.4178, and count absolute error 2.6180. The paired image bootstrap gives a YOLO-minus-random-forest IoU difference of -0.1884 (95% interval -0.2682 to -0.1034) and a Dice difference of -0.1858 (-0.2647 to -0.0991). At the low confidence used for official average precision, YOLO11n-seg obtains box mAP50 of 0.7830 and box mAP50-95 of 0.4447, but mask mAP50 is 0.0669 and mask mAP50-95 is 0.0191. The modern architecture therefore broadens the comparison without strengthening the empirical segmentation claim.

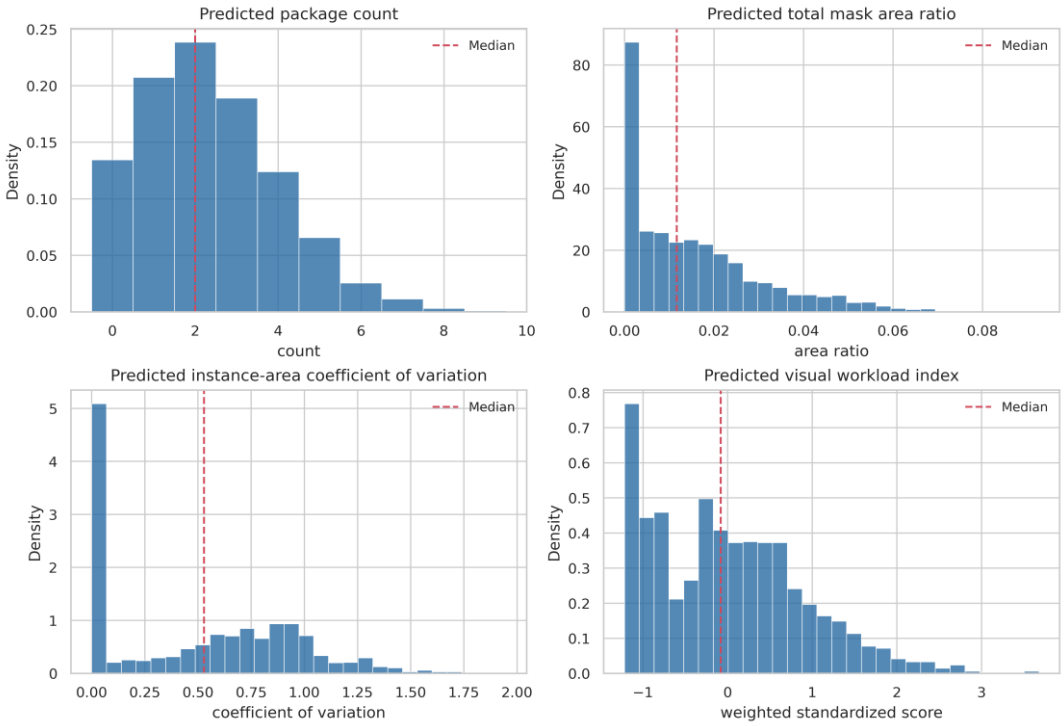


Segmentation metric comparison

Figure 3. Package Segmentation test performance with image-level 95% bootstrap intervals (89 images; 10,000 resamples).

Figure 3 shows the image-bootstrap uncertainty for the principal overlap and descriptor measures. The intervals reinforce that the random forest is stronger on this test set rather than indicating parity among the two learned approaches. Figure 4 displays the YOLO11n-seg predictions on the 1,920 Package Segmentation training images that define the cross-dataset quantile support. Predicted count has mean 2.343 and median 2; 258 images have no retained prediction. Predicted union-area ratio has mean 0.01552, and predicted instance-area CV has mean 0.4949. The resulting additive workload descriptor has standard deviation 0.9005. These are model-output distributions rather than reference annotations, and their operational meaning remains a hypothesis until examined against a paired workload outcome. The perception results establish

that foreground and instance descriptors can be estimated on Package Segmentation. They do not by themselves validate the descriptor distribution for LaDe. This distinction is important because the forecasting proxy is evaluated through controls and paired inference rather than inferred from segmentation accuracy.



Predicted visual descriptor distributions

Figure 4. Distributions of YOLO11n-seg descriptors predicted on the 1,920 Package Segmentation training images. The workload index combines train-standardized predicted count, union-mask area ratio, and instance-area coefficient of variation with weights 0.50, 0.35, and 0.15.

Table 8. Package Segmentation test performance; overlap and descriptor entries are means with 95% image-bootstrap intervals, and inference is mean CPU time.

Model	IoU	Dice	Count absolute error	Area-ratio absolute error	Area-CV absolute error	CPU ms/image
Otsu	0.1711 [0.1414, 0.2024]	0.2683 [0.2289, 0.3091]	2.5730 [2.1685, 2.9775]	0.2423 [0.2160, 0.2694]	0.5195 [0.4446, 0.5971]	5.65
Otsu + morphology	0.1848 [0.1535, 0.2179]	0.2857 [0.2441, 0.3284]	2.1685 [1.8315, 2.5056]	0.2189 [0.1937, 0.2452]	0.4849 [0.4149, 0.5587]	8.25
Logistic pixel classifier	0.2321 [0.1980, 0.2676]	0.3475 [0.3025, 0.3929]	4.7865 [4.2584, 5.3258]	0.1070 [0.0921, 0.1246]	0.4247 [0.3750, 0.4775]	76.27
Random-forest pixel classifier	0.5323 [0.4736, 0.5874]	0.6450 [0.5846, 0.7008]	1.4382 [1.1573, 1.7303]	0.0216 [0.0159, 0.0291]	0.2692 [0.2040, 0.3440]	533.57
YOLO11n-seg	0.3439 [0.2909, 0.3983]	0.4591 [0.4004, 0.5163]	2.6180 [2.1685, 3.1011]	0.0370 [0.0295, 0.0450]	0.5034 [0.4262, 0.5865]	118.55

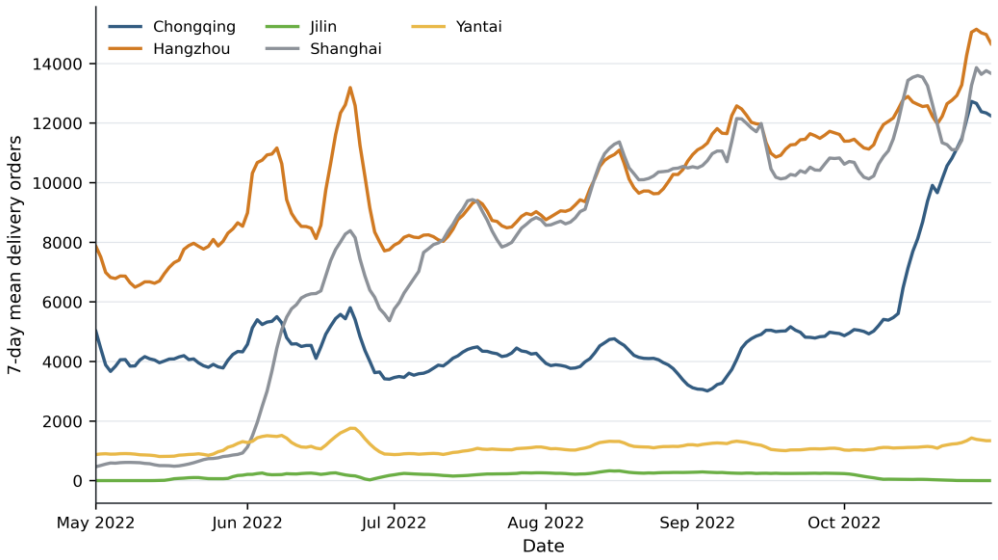
Journal of Technology Informatics and Engineering (JTIE)
Vol. 5 No. 1 April 2026
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 441-473
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

1. *LaDe operating scale and capacity-index behavior*

The operational panel displays large cross-city and temporal differences. Figure 5 plots seven-day mean delivery volume by city. Hangzhou and Shanghai form high-volume regimes, Chongqing is also substantial, Yantai is lower on delivery but high on pickup, and Jilin remains the smallest delivery city. Table 3 quantifies these differences and motivates city indicators together with histories calculated within city.

Table 9. Chronological holdout forecasting performance.

Model	MAE	RMSE	WAPE	R-squared	Bias
ElasticNet Operational	2.0521	2.9699	0.0361	0.9876	-0.7889
XGB Vision-Prior	2.4988	3.6654	0.0439	0.9811	-1.1485
XGB Delivery+Pickup	2.5154	3.6884	0.0442	0.9808	-1.0036
XGB Delivery-only	2.5522	3.8484	0.0448	0.9791	-1.2236
XGB Shuffled-Prior Control	2.5973	3.7537	0.0456	0.9801	-1.1404
XGB Generic-Transform Control	2.6499	3.9103	0.0466	0.9784	-1.3065
RF Operational	2.8878	4.2634	0.0507	0.9744	-1.5367
Seasonal Naive	3.3929	4.8376	0.0596	0.9670	-0.9848

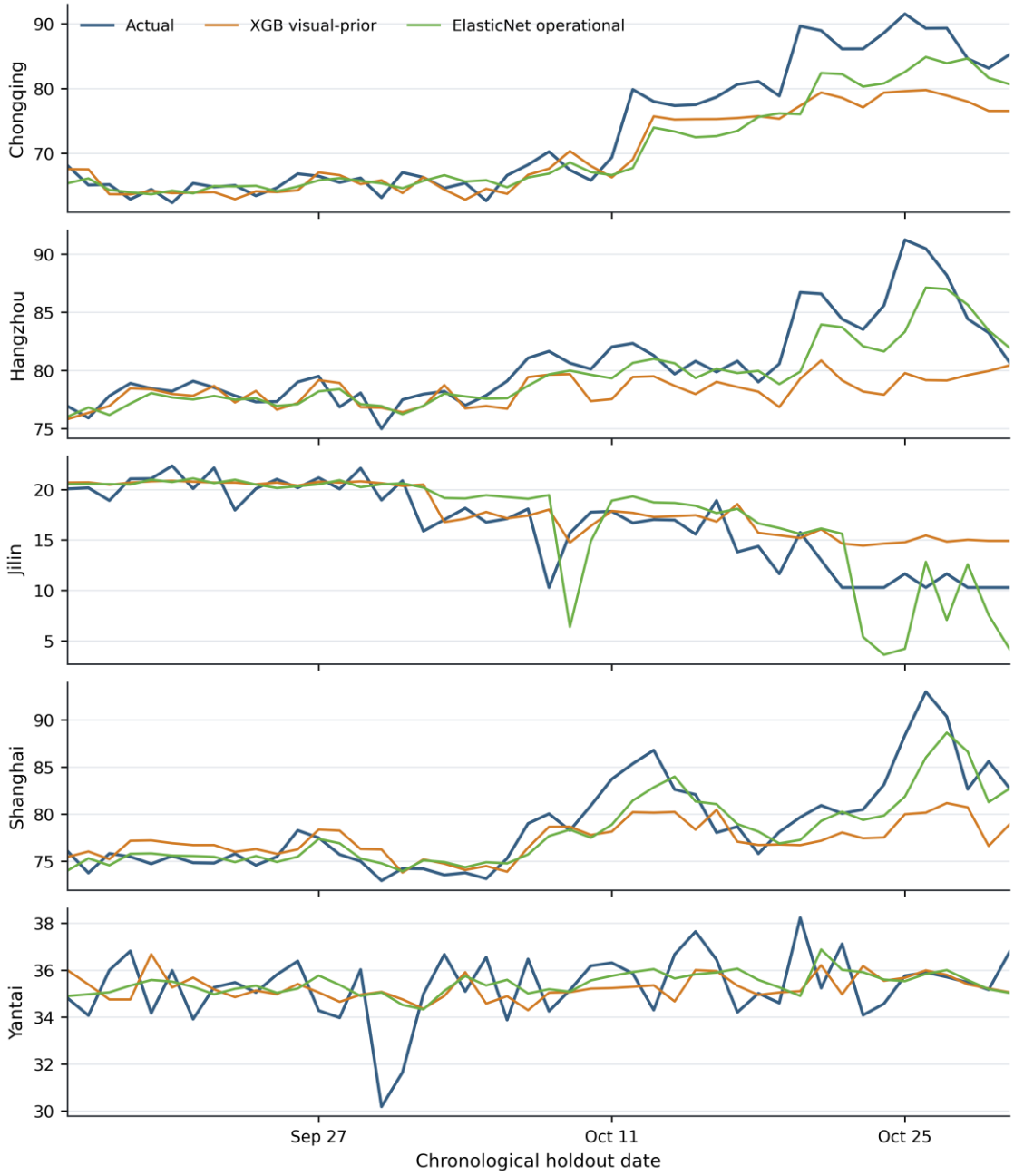


Seven-day delivery trends

Figure 5. Seven-day moving means of delivery orders across the five LaDe cities from May through October 2022.

The pickup stream is not a scaled copy of delivery. Across all cities, pickup tasks exceed delivery tasks, but their ratio differs markedly: Yantai records 1.15 million pickups and 0.21 million deliveries, whereas Shanghai records 1.42 million pickups and 1.48 million deliveries. Pickup covariates can therefore contribute an upstream activity signal, although the data do not identify whether any pickup becomes a specific later delivery. The primary PCA explains 89.19% of standardized training-period variance and has positive loadings from 0.4133 to 0.4644. Table 4 gives the estimates and their narrow date-bootstrap intervals; the same intervals are visualized in Figure 10. Stability of the loading vector supports reproducible construction within this sample,

but high variance explained is not sufficient to label the score a pure capacity requirement: the pooled index also reflects persistent city scale and realized courier deployment.



Holdout trajectories by city

Figure 6. Actual next-day operational-intensity index, ElasticNet forecast, and XGB Vision-Prior forecast for each city in the chronological holdout.

2. *Main holdout forecasting results*

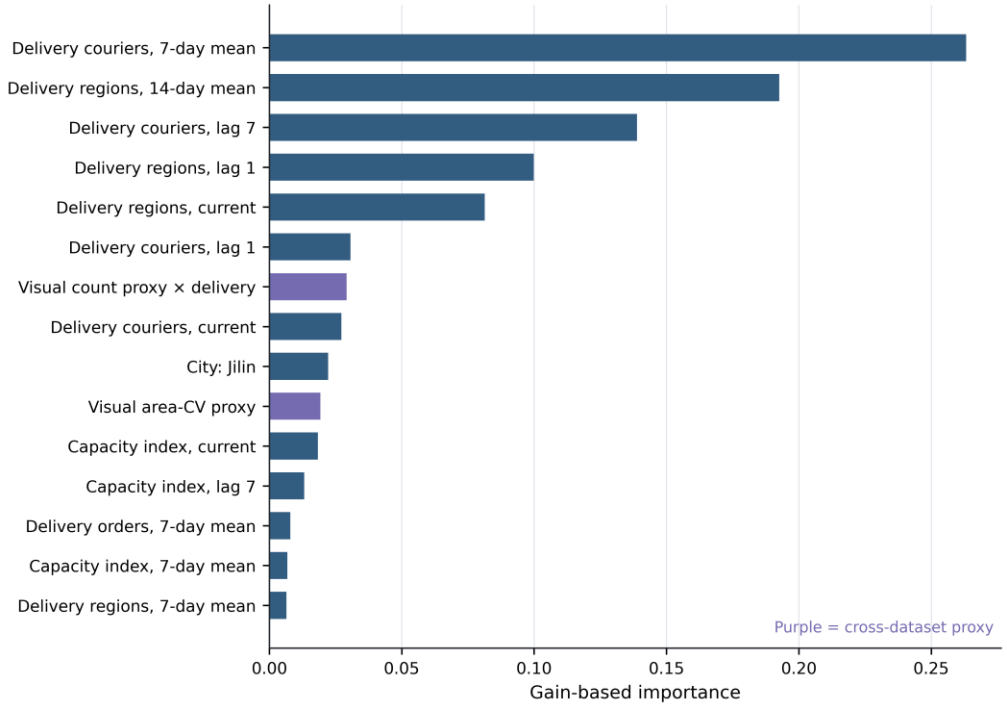
Table 9 reports the 230-row chronological holdout. ElasticNet Operational is the strongest overall model, with MAE 2.0521, RMSE 2.9699, WAPE 0.0361, and R-squared 0.9876. XGB Vision-

Journal of Technology Informatics and Engineering (JTIE)
Vol. 5 No. 1 April 2026
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 441-473
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

Prior has the lowest MAE within the XGBoost family (2.4988), followed by Delivery+Pickup (2.5154), Delivery-only (2.5522), the shuffled-prior control (2.5973), and the generic-transform control (2.6499). The proxy reduces MAE by 0.0166, or 0.66%, relative to Delivery+Pickup on this holdout, but its MAE remains 0.4467 above ElasticNet. The result is therefore an incremental within-family comparison rather than evidence of general superiority.

Table 10. Holdout underprediction and asymmetric-error diagnostics at weight two.

Model	Mean one-sided underprediction	P90 one-sided underprediction	Mean one-sided excess	Within +/-5	Underprediction fraction	Unitless asymmetric loss
ElasticNet Operational	1.4205	4.4413	0.6316	0.9087	0.5826	798.70
XGB Vision-Prior	1.8237	5.6228	0.6752	0.8609	0.6130	994.18
XGB Delivery+Pickup	1.7595	5.6299	0.7559	0.8565	0.6130	983.23
XGB Delivery-only	1.8879	5.8456	0.6643	0.8565	0.6043	1,021.22
XGB Shuffled-Prior Control	1.8689	5.8293	0.7285	0.8522	0.6043	1,027.23
XGB Generic-Transform Control	1.9782	6.2624	0.6717	0.8565	0.6304	1,064.48
RF Operational	2.2123	6.6419	0.6755	0.8217	0.6391	1,173.01
Seasonal Naive	2.1889	8.4361	1.2040	0.7696	0.5304	1,283.81

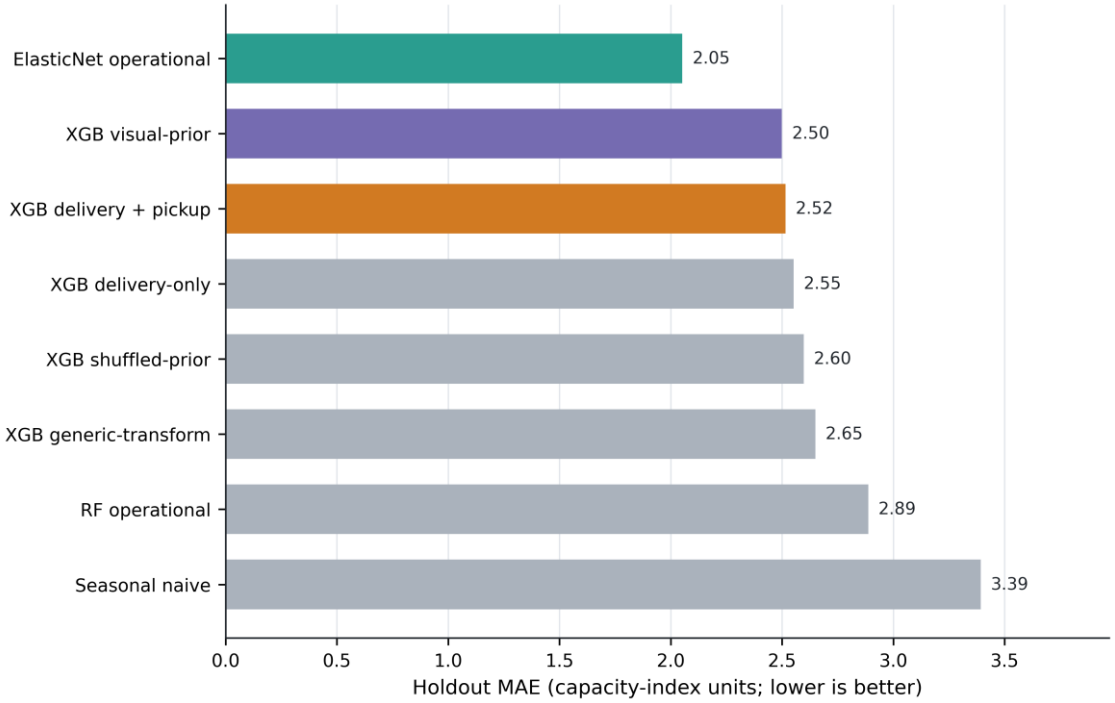


XGBoost gain-based feature importance

Figure 7. Top 15 gain-based feature importances from XGB Vision-Prior. Long feature names are retained so that operational and proxy variables can be distinguished.

The underprediction diagnostics in Table 10 preserve the same overall interpretation. ElasticNet has the smallest mean one-sided underprediction (1.4205), highest within-five hit rate (0.9087),

and lowest primary asymmetric loss (798.70). XGB Vision-Prior has mean one-sided underprediction 1.8237 and asymmetric loss 994.18, compared with 1.7595 and 983.23 for Delivery+Pickup. Its underprediction fraction is 0.6130, or 141 of 230 city-days. Although the proxy has lower MAE than Delivery+Pickup, the weight-two loss is slightly higher because its underprediction magnitude averaged over all days is larger. These quantities are errors in index points, not observed late deliveries, overloaded couriers, or economic costs. When the underprediction weight is varied from 1 to 5, ElasticNet remains the lowest-loss model.



Holdout MAE comparison

Figure 8. Chronological-holdout mean absolute error across forecasting models, including generic-transform and shuffled-prior controls. Lower values indicate better next-day operational-intensity forecasts.

Figure 6 shows actual and predicted trajectories separately for all five cities. ElasticNet tracks most high-volume movements more closely than XGB Vision-Prior, while both models have difficulty with sharp late-October increases. Presenting separate panels avoids hiding city-specific errors through a cross-city date average.

Table 11. Rolling-origin summary for XGBoost information blocks.

Model	Mean MAE	MAE SD	Mean RMSE	RMSE SD	Mean WAPE	Mean R-squared
XGB Shuffled-Prior Control	1.7133	0.2581	2.4348	0.5038	0.0313	0.9883
XGB Delivery+Pickup	1.7268	0.2883	2.4700	0.5717	0.0316	0.9879
XGB Delivery-only	1.7353	0.2439	2.4962	0.5588	0.0318	0.9877
XGB Generic-Transform Control	1.7514	0.2842	2.4874	0.5541	0.0320	0.9877
XGB Vision-Prior	1.7746	0.2265	2.5195	0.4862	0.0325	0.9875

Journal of Technology Informatics and Engineering (JTIE)
Vol. 5 No. 1 April 2026
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 441-473
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

Figure 7 reports gain-based importance for the XGB Vision-Prior model. The most important variables are operational: seven-day mean delivery couriers (0.2632), 14-day mean delivery regions (0.1926), and seven-day-lag delivery couriers (0.1389). A proxy interaction, predicted count prior by standardized delivery, ranks seventh at 0.0292; the predicted area-CV prior ranks tenth at 0.0193. Importance measures tree-split gain allocation rather than a causal or stable marginal effect. The concentration in operational histories is consistent with the small within-family holdout difference.

Table 12. Paired seven-day moving-block bootstrap contrasts on holdout MAE.

Comparator	Vision minus comparator MAE	95% interval	Two-sided bootstrap probability	Interpretation
XGB Delivery+Pickup	-0.0166	[-0.0558, 0.0319]	0.4492	small, statistically uncertain proxy advantage
XGB Generic-Transform Control	-0.1511	[-0.2726, -0.0448]	0.0008	proxy lower on the final holdout
XGB Shuffled-Prior Control	-0.0985	[-0.1389, -0.0669]	<0.0002	proxy lower on the final holdout
ElasticNet Operational	0.4467	[-0.0401, 1.1094]	0.0996	observed MAE favors ElasticNet

Figure 8 compares holdout MAE for all models and controls. XGB Vision-Prior has a small advantage over the other XGBoost specifications on this holdout, whereas ElasticNet remains clearly better overall. Showing the generic and shuffled controls in the same panel prevents the within-family difference from being interpreted as a general model advantage.

3. *Rolling-origin and statistical comparisons*

The three-fold rolling-origin summary in Table 11 is less favorable to the proxy than the single final holdout. The shuffled-prior control has the lowest mean MAE (1.7133), followed by Delivery+Pickup (1.7268), Delivery-only (1.7353), and the generic-transform control (1.7514). XGB Vision-Prior has the largest mean MAE (1.7746) and RMSE (2.5195) among the five XGBoost variants. The fold means therefore do not show a temporally stable proxy advantage.

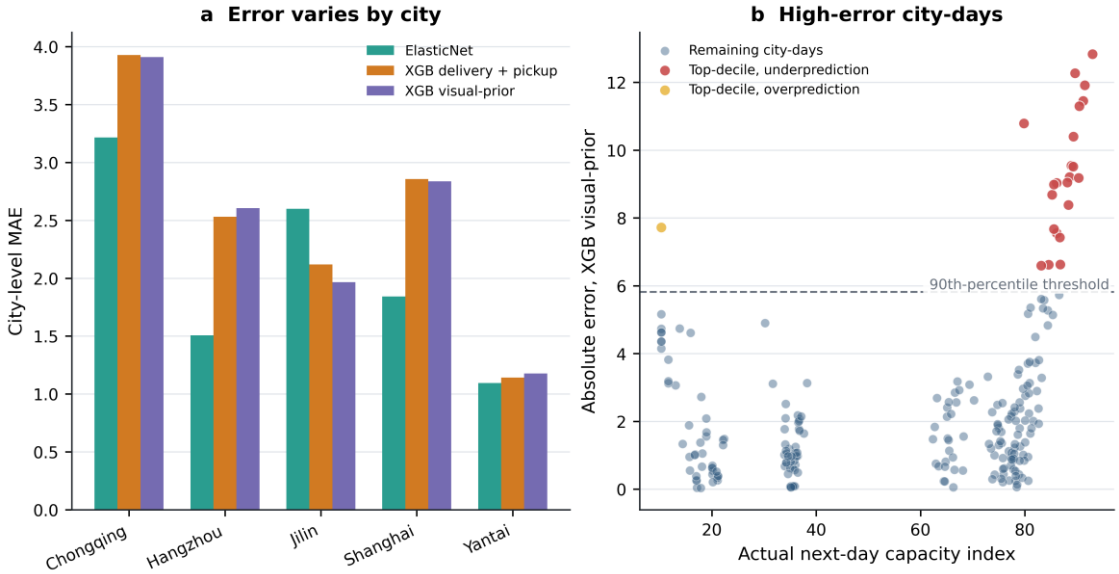
Table 13. City-level holdout errors for the principal model comparison.

City	Model	MAE	RMSE	WAPE	R-squared	Bias
Chongqing	ElasticNet Operational	3.2157	4.4157	0.0439	0.7873	-2.5032
Chongqing	XGB Delivery+Pickup	3.9281	5.2992	0.0536	0.6937	-3.1302
Chongqing	XGB Vision-Prior	3.9105	5.2650	0.0534	0.6976	-3.1828
Hangzhou	ElasticNet Operational	1.5069	2.1235	0.0187	0.6624	-1.0599
Hangzhou	XGB Delivery+Pickup	2.5316	3.7407	0.0314	-0.0475	-2.1956
Hangzhou	XGB Vision-Prior	2.6047	3.8490	0.0323	-0.1090	-2.3475
Jilin	ElasticNet Operational	2.5999	3.4535	0.1569	0.2316	0.2785
Jilin	XGB Delivery+Pickup	2.1187	2.9371	0.1279	0.4442	1.6045
Jilin	XGB Vision-Prior	1.9652	2.6871	0.1186	0.5348	1.3976
Shanghai	ElasticNet Operational	1.8432	2.4760	0.0234	0.7325	-0.7752
Shanghai	XGB Delivery+Pickup	2.8566	3.8864	0.0362	0.3410	-1.3771
Shanghai	XGB Vision-Prior	2.8375	3.8850	0.0360	0.3415	-1.5577
Yantai	ElasticNet Operational	1.0948	1.4269	0.0310	-0.0257	0.1152
Yantai	XGB Delivery+Pickup	1.1419	1.4896	0.0323	-0.1179	0.0801
Yantai	XGB Vision-Prior	1.1763	1.5254	0.0333	-0.1722	-0.0521

Paired moving-block results in Table 12 quantify the uncertainty in the main holdout differences. Relative to Delivery+Pickup, XGB Vision-Prior has 0.0166 lower MAE, with a 95% interval from -0.0558 to 0.0319 and two-sided bootstrap probability 0.4492. It is lower than the selected generic-transform control by 0.1511 MAE and lower than the shuffled-prior control by 0.0985; both intervals exclude zero. The proxy-versus-ElasticNet interval includes zero under the date-block design, although the observed difference favors ElasticNet by 0.4467. The final holdout thus contains a detectable distinction from the two matched transformations but not from the operational XGBoost specification. Together with ElasticNet's lower aggregate error and the unfavorable rolling-origin result, this supports only a limited within-family increment.

4. City-level and high-error analysis

Table 13 reports the key three-model comparison by city, and Figure 9, panel a, visualizes city MAE. ElasticNet is best in Chongqing, Hangzhou, Shanghai, and Yantai, while XGB Vision-Prior is best among the three in Jilin. Relative to Delivery+Pickup, the proxy improves MAE in Chongqing by 0.0176, Jilin by 0.1534, and Shanghai by 0.0192; it worsens Hangzhou by 0.0731 and Yantai by 0.0344. The pooled within-XGBoost improvement is therefore heterogeneous and is driven most visibly by Jilin rather than by a consistent five-city gain.



City-level and high-error diagnostics

Figure 9. City-level and high-error diagnostics on the chronological holdout. Panel a reports MAE by city for the three focal models; panel b identifies the top decile of absolute XGB Vision-Prior errors and distinguishes underprediction from overprediction.

The highest-error decile for XGB Vision-Prior contains 23 city-days, shown in Figure 9, panel b. All occur in October; 12 are in Chongqing, five in Hangzhou, five in Shanghai, and one in Jilin. Twenty-two are underpredictions. The largest errors occur for Shanghai on October 26 (actual

Journal of Technology Informatics and Engineering (JTIE)
Vol. 5 No. 1 April 2026
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 441-473
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

92.99, predicted 80.16, absolute error 12.84), Chongqing on October 20 (89.63 versus 77.36, error 12.27), Chongqing on October 25 (91.51 versus 79.60, error 11.91), and Hangzhou on October 25 (91.22 versus 79.77, error 11.45). The dominant failure mode is a late-period upward shift at high index values, not an isolated parcel-composition pattern.

Table 14. Capacity-target and current-index sensitivity using XGB Delivery+Pickup.

Target / predictor specification	MAE	RMSE	WAPE	R-squared	Unitless asymmetric loss
Composite operational-intensity index	2.5154	3.6884	0.0442	0.9808	983.23
Composite index; current index excluded	2.4309	3.5647	0.0427	0.9821	945.29
Composite index; all current delivery variables excluded	3.0754	4.3439	0.0540	0.9734	1,179.96
Demand-only index	2.0176	2.7784	0.0365	0.9854	730.73
Active-courier index	1.2902	2.0278	0.0236	0.9790	544.29

5. Capacity-index and feature-importance sensitivity

Table 14 addresses whether the forecasting task is made artificially easy by including current variables that also define the next-day target. With the same XGB Delivery+Pickup design, removing the current composite index improves MAE from 2.5154 to 2.4309. Removing all eight current delivery variables, including every PCA component and the two contemporaneous delivery ratios, raises MAE to 3.0754, while R-squared remains 0.9734 using lagged delivery histories, calendar terms, and pickup variables. The difference confirms that one-day persistence is useful, but the predictor row never contains the next-day values used to form its target. A demand-only target has MAE 2.0176 and R-squared 0.9854; a standardized active-courier target has MAE 1.2902 and R-squared 0.9790. Since every target is standardized to a comparable 50-plus-10 scale, the composite is only one possible operational construct rather than a uniquely validated capacity quantity.

Table 15. Capacity-index correlations with criteria excluded from PCA.

Criterion	Spearman rho	*p* value	Valid city-days
Delivery orders per courier	0.6706	2.76×10^{-110}	836
Median delivery service minutes	-0.7395	1.56×10^{-145}	836
P90 delivery service minutes	-0.6021	1.34×10^{-83}	836
P90 completion minute	0.1972	8.95×10^{-9}	836

Table 15 relates the composite index to LaDe criteria excluded from PCA. The positive association with delivery orders per courier (Spearman rho 0.671) and later 90th-percentile completion minute (0.197) is consistent with operating intensity. Median and 90th-percentile service durations are negatively related to the index (-0.740 and -0.602). The negative duration relationships may reflect cross-city scale efficiency, task mix, or aggregation effects; they rule out interpreting the PCA score as a simple monotone measure of per-task difficulty or required

labor time. Figure 10 displays the loadings from Table 4 and their date-bootstrap intervals. Their signs and relative magnitudes are stable within the training window, although this stability does not resolve the broader interpretation of the combined demand-and-courier score.

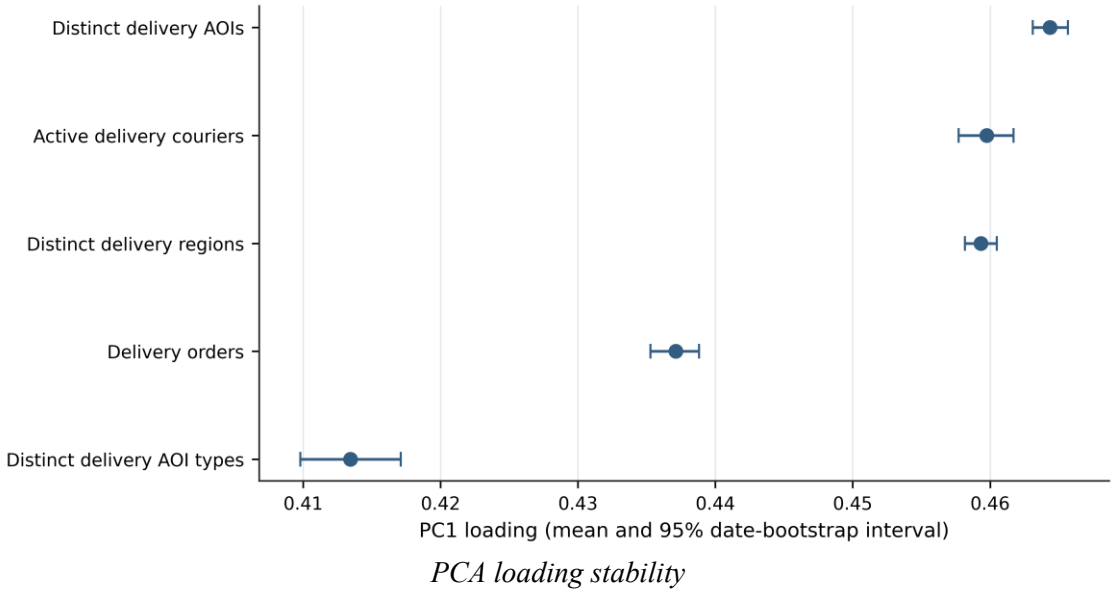


Figure 10. Training-period PC1 loadings with 95% date-bootstrap intervals for the five components of the operational-intensity index.

Table 16 gives the numerical values underlying Figure 7. Two proxy variables appear among the top 15, but their gain importance is far below the leading courier and region histories. The table is descriptive and does not supply uncertainty or causal attribution.

Table 16. Top 15 gain-based feature importances from XGB Vision-Prior.

Rank	Feature	Gain-based importance
1	delivery_couriers_roll7_mean	0.2632
2	delivery_regions_roll14_mean	0.1926
3	delivery_couriers_lag7	0.1389
4	delivery_regions_lag1	0.0999
5	delivery_regions	0.0814
6	delivery_couriers_lag1	0.0307
7	visual_count_x_delivery	0.0292
8	delivery_couriers	0.0272
9	city_Jilin	0.0222
10	visual_area_cv_prior	0.0193
11	capacity_index	0.0183
12	capacity_index_lag7	0.0132
13	delivery_orders_roll7_mean	0.0079
14	capacity_index_roll7_mean	0.0069
15	delivery_regions_roll7_mean	0.0065

6. Amazon external construct-validation results

Table 17 reports the paired-route analysis. In the station-specific chronological holdout, adding parcel physical-mix variables reduces MAE from 2,003.32 to 1,998.56 seconds, a 4.77-second or 0.238% reduction; RMSE falls by 6.56 seconds and WAPE by 0.00030. The station-date cluster-

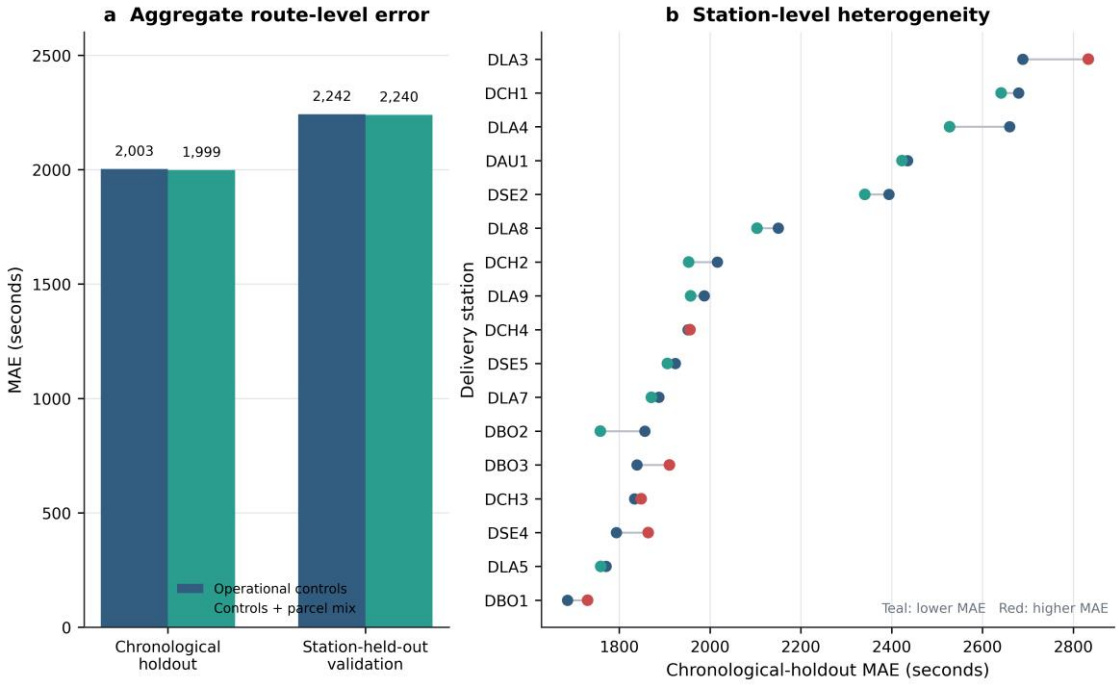
Journal of Technology Informatics and Engineering (JTIE)
Vol. 5 No. 1 April 2026
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 441-473
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

bootstrap 95% interval for the augmented-minus-control MAE difference is -28.74 to 17.70 seconds, and the three-date moving-block interval is -21.99 to 11.89 seconds. In station-held-out validation, MAE decreases by 2.64 seconds (0.118%) while RMSE increases by 1.00 second; the station-cluster MAE interval is -24.36 to 18.09 seconds. None of these intervals excludes zero.

Table 17. Amazon paired-route external construct-validation performance.

Evaluation	Specification	Routes	MAE (s)	RMSE (s)	WAPE	R-squared	Bias (s)
Station-specific chronological holdout	Operational controls	1,508	2,003.32	2,535.35	0.12735	0.29816	283.52
Station-specific chronological holdout	Controls + parcel physical mix	1,508	1,998.56	2,528.79	0.12704	0.30179	322.27
Five-fold station-held-out	Operational controls	6,112	2,242.43	2,809.71	0.13783	0.19613	34.61
Five-fold station-held-out	Controls + parcel physical mix	6,112	2,239.79	2,810.71	0.13767	0.19556	50.49

Figure 11, panel a, makes the sub-0.3% pooled change visible in both evaluations. Panel b shows that station effects are heterogeneous: 11 of 17 stations improve, and six worsen in the chronological holdout, with MAE changes from -132.35 seconds at DLA4 to +144.25 seconds at DLA3. This heterogeneity is consistent with the wide paired interval.



Amazon construct validation

Figure 11. Independent Amazon Last Mile construct-sensitivity analysis. Panel a compares route-level MAE from operational controls with and without parcel physical-mix features under chronological and station-held-out evaluation; panel b shows heterogeneous station-level changes on the chronological holdout.

The highest-error decile of the augmented chronological model begins at 4,210.39 seconds (70.17 minutes) and contains 151 routes. Of these, 28.5% are underpredictions and 71.5% are overpredictions. High-error routes are concentrated at DLA7 (25 routes), DLA9 (18), and DBO3 (17), and on August 12 (24) and August 15 (23). Their mean observed workload is lower than that of the remaining holdout routes, indicating that unexpectedly low planned-service routes are a major source of overprediction.

The Amazon analysis removes the cross-dataset pairing problem for physical dimensions and planned workload within this auxiliary study, but it does not validate the Package-Seg-to-LaDe mapping. Amazon contains dimensions but no package images, and its outcome is planned service time rather than realized courier hours, overtime, delay, or economic loss. Its very small and statistically uncertain incremental result supports only a cautious construct-level conclusion: parcel physical mix can be represented as an operational covariate, but the public data do not show a material forecasting gain after basic route controls.

V. CONCLUSION AND RECOMMENDATION

This study separates three empirical questions that cannot be answered from one public source. Package Segmentation supports a direct perception benchmark; LaDe-D and LaDe-P support a five-city next-day forecast; and Amazon supplies a separate paired route-package analysis. The connection from Package Segmentation to LaDe is a documented training-period distributional transform of operational intensity, not a join of images to city-day events.

The image benchmark contains learnable foreground structure, but the model comparison is mixed. The random-forest pixel classifier reaches IoU 0.5323 and Dice 0.6450, whereas the short YOLO11n-seg adaptation reaches 0.3439 and 0.4591. YOLO11n-seg supplies instance predictions for the transfer distribution, but neither architecture result validates workload relevance in LaDe. The LaDe target is highly forecastable and its PC1 loadings are stable within the training period. Removing all current delivery variables increases error, as expected when useful one-day persistence is withheld, yet the lag-and-pickup model still attains R-squared 0.9734. The demand-only and courier sensitivities, together with negative service-duration correlations, show that the composite is an operational-intensity index rather than a direct measure of labor time, latent demand, or optimal staffing.

The cross-dataset proxy has limited incremental value. On the final holdout, XGB Vision-Prior reduces MAE from 2.5154 to 2.4988 relative to XGB Delivery+Pickup and is better than the generic and shuffled controls. Its 0.0166 contrast with Delivery+Pickup is not statistically distinguishable from zero, however, and the proxy improves three of five cities rather than all

five. It also trails Delivery+Pickup and the shuffled control in the rolling-origin average. ElasticNet Operational remains best overall at MAE 2.0521. The Amazon augmentation changes chronological MAE by only 0.24%, with an interval that includes zero. Taken together, these results support a small holdout increment within one nonlinear model family, not general superiority of image-derived priors for capacity forecasting.

Computer-vision logistics systems should be evaluated with perception metrics and downstream outcomes that share an observation key. A segmentation model with accurate masks is not automatically valuable for capacity planning, and a forecasting model with low MAE can still produce consequential asymmetric errors. This decision-oriented perspective is consistent with uncertainty-aware perception and urban transportation forecasting, where sensing or forecasting is judged by downstream consequences as well as raw model scores (Xin, 2025, 2026b; Ye et al., 2024; Zhang et al., 2025). Here, the asymmetric score is unitless rather than an observed business cost, so economic value cannot be inferred without staffing, overtime, service-failure, or related outcome data.

For panels similar to LaDe, the regularized operational model is the appropriate forecasting baseline. A direct operational test of vision would timestamp depot or loading-area images, link each observation to a depot, route, or city-day, and record realized handling time and labor outcomes. Such a design would permit camera-derived count, occupied area, physical dimensions, and size dispersion to be tested beyond the event stream and across city, depot, and period shifts. Until paired observations are available, Package Segmentation, LaDe, and Amazon remain complementary sources whose findings should not be treated as one jointly observed pipeline.

Data Availability

LaDe-D and LaDe-P are available from the official LaDe repository at https://github.com/wenhaomin/LaDe. Package Segmentation, including its fixed training, validation, and test splits, is available at https://docs.ultralytics.com/datasets/segment/package-seg/. The Amazon Last Mile model-building data and CC BY-NC 4.0 license are documented at https://registry.opendata.aws/amazon-last-mile-challenges/. The two Amazon files used here can be obtained directly as route_data.json from https://amazon-last-mile-challenges.s3.us-west-2.amazonaws.com/almrrc2021/almrrc2021-data-training/model_build_inputs/route_data.json and package_data.json from https://amazon-last-mile-challenges.s3.us-west-2.amazonaws.com/almrrc2021/almrrc2021-data-training/model_build_inputs/package_data.json; the license text is available at https://amazon-last-mile-challenges.s3.us-west-2.amazonaws.com/almrrc2021/License.txt.

REFERENCES

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32. https://doi.org/10.1023/A:1010933404324

Chen, S., He, S., & Sun, E. (2024). Risk-bounded GPU resource oversubscription via conformal demand envelopes in production AI clusters. JACS, 4(5), 119-134. https://doi.org/10.69987/JACS.2024.40509

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794). https://doi.org/10.1145/2939672.2939785

Crainic, T. G., & Laporte, G. (1997). Planning models for freight transportation. European Journal of Operational Research, 97(3), 409-438. https://doi.org/10.1016/S0377-2217(96)00298-6

Dai, N., Hu, X., Xu, K., Yuan, Y., & Lu, Z. (2025). Research on dimension measurement algorithm for parcel boxes in high-speed sorting system. Scientific Reports. https://doi.org/10.1038/s41598-025-07730-y

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. The Annals of Statistics, 29(5), 1189-1232. https://doi.org/10.1214/aos/1013203451

Gevaers, R., Van de Voorde, E., & Vanelslander, T. (2014). Cost modelling and simulation of last-mile characteristics in an innovative B2C supply chain environment with implications on urban areas and cities. Procedia - Social and Behavioral Sciences, 125, 398-411. https://doi.org/10.1016/j.sbspro.2014.01.1483

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press.

Han, S., Liu, X., Han, X., Wang, G., & Wu, S. (2020). Visual sorting of express parcels based on multi-task deep learning. Sensors, 20(23), 6785. https://doi.org/10.3390/s20236785

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (pp. 2961-2969). https://doi.org/10.1109/ICCV.2017.322

He, S., Chang, X., & Sun, E. (2024). Cross-cloud transfer learning for AI training capacity forecasting under workload and topology distribution shift. JACS, 4(1), 100-120. https://doi.org/10.69987/JACS.2024.40108

He, S., Li, C., & Rao, H. (2025). Few-shot cold-start workload forecasting for new AI inference tenants with time-series foundation models. Journal of Technology Informatics and Engineering, 4(1), 306-324. https://doi.org/10.51903/jtie.v4i1.546

He, S., Nie, J., & Li, C. (2026). Power-aware inventory planning for AI infrastructure using job-level forecasting and LLM workload explanations. Journal of Technology Informatics and Engineering, 5(1), 341-359. https://doi.org/10.51903/jtie.v5i1.548

He, S., Tu, H., & Liu, I. (2023). Safe PD capacity forecasting with time-series foundation models and calibrated uncertainty for heterogeneous GPU clusters. JACS, 3(4), 48-66. https://doi.org/10.69987/JACS.2023.30404

Hyndman, R. J., & Athanasopoulos, G. (2021). Forecasting: Principles and practice (3rd ed.). OTexts.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W. Y., Dollár, P., & Girshick, R. (2023). Segment anything. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 4015-4026).

Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. In D. Fleet, T. Pajdla, B. Schiele, & T. Tuytelaars (Eds.), Computer Vision - ECCV 2014 (pp. 740-755). Springer. https://doi.org/10.1007/978-3-319-10602-1_48

Ling, Z., Xin, Q., Lin, Y., Su, G., & Shui, Z. (2024). Optimization of autonomous driving image detection based on RFAConv and triplet attention. In Proceedings of the 2nd International Conference on Software Engineering and Machine Learning (SEML 2024).

Lu, S., & Zhou, D. (2023). LLM-augmented customer representation learning for next-purchase prediction in online retail. JACS, 3(3), 50-64. https://doi.org/10.69987/JACS.2023.30305

Lu, Z., Dai, N., Hu, X., Xu, K., & Yuan, Y. (2024). Research on high-speed classification and location algorithm for logistics parcels based on a monocular camera. Scientific Reports, 14, 15901. https://doi.org/10.1038/s41598-024-66941-x

McKinney, W. (2010). Data structures for statistical computing in Python. In Proceedings of the 9th Python in Science Conference (pp. 56-61). https://doi.org/10.25080/Majora-92bf1922-00a

Merchán, D., Arora, J., Pachon, J., Konduri, K., Winkenbach, M., Parks, S., & Noszek, J. (2022). 2021 Amazon Last Mile Routing Research Challenge: Data Set. Transportation Science, 58(1), 8-11. https://doi.org/10.1287/trsc.2022.1173

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., VanderPlas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 779-788). https://doi.org/10.1109/CVPR.2016.91

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In N. Navab, J. Hornegger, W. M. Wells, & A. Frangi (Eds.), Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015 (pp. 234-241). Springer. https://doi.org/10.1007/978-3-319-24574-4_28

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). ImageNet large scale visual recognition challenge. International Journal of Computer Vision, 115, 211-252. https://doi.org/10.1007/s11263-015-0816-y

Savelsbergh, M., & Van Woensel, T. (2016). 50th anniversary invited article: City logistics: Challenges and opportunities. Transportation Science, 50(2), 579-590. https://doi.org/10.1287/trsc.2016.0675

Simchi-Levi, D., Kaminsky, P., & Simchi-Levi, E. (2007). Designing and managing the supply chain: Concepts, strategies and case studies (3rd ed.). McGraw-Hill.

Ultralytics. (2024). Package Segmentation Dataset. https://docs.ultralytics.com/datasets/segment/package-seg/

Wu, L., Wen, H., Hu, H., Mao, X., Xia, Y., Shan, E., Zheng, J., Lou, J., Liang, Y., Yang, L., Zimmermann, R., Lin, Y., & Wan, H. (2024). LaDe: The first comprehensive last-mile express dataset from industry. In Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. https://doi.org/10.1145/3637528.3671548

Xin, Q. (2025). Uncertainty-aware late fusion for 3D perception (confidence calibration + fusion rule learning). Journal of Technology Informatics and Engineering, 4(1), 215-238. https://doi.org/10.51903/jtie.v4i1.485

Xin, Q. (2026a). LiDAR-camera object-level fusion for multi-target tracking using JPDA and EKF: A reproducible empirical study on a PandaSet-parameterised five-sequence dataset. Journal of Technology Informatics and Engineering, 5(1), 54-76. https://doi.org/10.51903/jtie.v5i1.486

Xin, Q. (2026b). Probabilistic bike-sharing demand forecasting under changing weather and seasonal regimes with transformer-based models. Transport Findings. https://doi.org/10.32866/001c.157499

Xin, Q. (2026c). Self-supervised customer representation learning for segmentation and next-purchase prediction on UCI online retail. Journal of Information and Technology, 14(1). https://doi.org/10.32664/j-intech.v14i01.2229

Ye, T., Ma, R., & Luo, S. (2024). Vision-language traffic sign recognition for self-driving: Robust classification, uncertainty calibration, and LLM safety explanations on a GTSRB-compatible benchmark. JACS, 4(10), 84-102. https://doi.org/10.69987/JACS.2024.41007

Zhang, L., Ma, R., & Greg, P. (2025). Digital-twin dispatching for urban mobility via spatio-temporal transformers and offline reinforcement learning. Journal of Technology Informatics and Engineering, 4(2), 337-363. https://doi.org/10.51903/jtie.v4i2.501

Zhang, L., & Zhang, E. (2024). LLM-explained graph traffic forecasting for DOT corridor operations: Full empirical evaluation on METR-LA and PEMS-BAY with crash evidence. JACS, 4(5), 135-145. https://doi.org/10.69987/JACS.2024.40510

Zhao, S., Bai, J., & Roberson, D. (2025). Multi-horizon GPU demand forecasting with workload semantics and operational risk curves: An empirical study on Alibaba Clusterdata GPU trace. Journal of Technology Informatics and Engineering, 4(3), 544-571. https://doi.org/10.51903/jtie.v4i3.498

Zhong, Z. S., Ma, R., & Zhao, H. (2023a). Human-uncertainty distillation for calibrated vision models on CIFAR-10H. JACS, 3(2), 77-89. https://doi.org/10.69987/JACS.2023.30206

Zhong, Z. S., Zhang, L., & Ma, D. (2023b). Spatial RAG for urban crash hotspot discovery and safety countermeasure recommendation. JACS, 3(12), 45-61. https://doi.org/10.69987/JACS.2023.31206