Identification of Urban Last-Mile Courier Delivery Stops Using GPS Trajectory Data

Wei Zhang[1, *], Jing Liu[2], Hao Chen[3]

[1] School of Business, Beijing Technology and Business University, Beijing, China 100048

[2] School of Management, Capital Normal University, Beijing, China 100048

[3] School of Graduate, Chinese Academy of Fiscal Sciences, Beijing, China 100142

* Email: djhui777@163.com (Corresponding Author)

Received 21 October 2025 · Revised 5 December 2025 · Accepted 19 January 2026 · Published 30 June 2026

Abstract

Urban last-mile logistics has grown substantially in recent years, propelled by the rapid expansion of e-commerce and on-demand delivery services. Despite its operational significance, fine-grained monitoring of last-mile courier activities remains challenging: a single courier shift generates hundreds of stops, many lasting only a few minutes, and the GPS records collected by on-board fleet-tracking devices carry no explicit label indicating the purpose of each stop. This paper presents an end-to-end framework for identifying delivery stops from passively collected GPS trajectories by combining vehicle traces with electronic waybill records. Drawing on a real-world courier dataset spanning the full calendar year 2022 and covering more than 80 delivery vehicles in a major metropolitan area, the framework cleans raw trajectories, extracts stop candidates through speed-threshold segmentation and spatial merging, and constructs ground-truth labels by matching stop candidates to waybill records within calibrated temporal and spatial tolerance windows. Five interpretable features include dwell time, pre-stop speed, heading change at the stop, local stop density, and distance from the departure hub are then derived to represent each candidate. To counter the severe class imbalance between delivery and non-delivery stops, SMOTE resampling is applied exclusively to the training partition before three supervised classifiers—Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Decision Tree (DT)are trained and evaluated under a unified protocol. All three models achieve test-set identification accuracies exceeding 98.9%, confirming that the proposed interpretable feature set enables robust delivery-stop identification and supports scalable operational monitoring of urban last-mile courier services.

Keywords: Urban last-mile logistics; delivery-stop identification; GPS trajectory; stop detection; class imbalance; SMOTE; supervised learning

Identification of Urban Last-Mile Courier Delivery Stops Using GPS Trajectory Data

1. Introduction

Urban last-mile logistics—the final leg of a parcel's journey from a local distribution centre to the customer's door—has become one of the costliest, most congested, and environmentally intensive segments of contemporary supply chains (Savelsbergh & Van Woensel, 2016; Boysen et al., 2021). The sustained expansion of e-commerce has driven persistent growth in urban parcel volumes, placing mounting pressure on city road networks and raising urgent questions concerning service efficiency, environmental impacts, and street-level operations (Allen et al., 2012; Mangiaracina et al., 2019). Despite its operational importance, the fine-grained monitoring of last-mile delivery activities remains difficult: couriers may make dozens or even hundreds of stops during a shift, many of which last only a few minutes, while the GPS records generated by on-board fleet-tracking devices contain no explicit information about the purpose of each stop.

The widespread deployment of GPS tracking devices on commercial vehicles has created new opportunities for observing freight and courier operations. Compared with traditional survey- and observation-based methods, passively collected GPS data provide continuous spatiotemporal coverage, relatively low collection costs, and a close representation of actual operating conditions (Greaves & Figliozzi, 2008). Early work in this field focused largely on passenger travel behaviour and addressed methodological challenges such as positioning noise, irregular sampling, trip-end detection, and activity-episode segmentation in multi-day GPS datasets (Du & Aultman-Hall, 2007; Shen & Stopher, 2014). These studies established a methodological foundation—including stop detection, trip segmentation, and activity inference—that was subsequently extended to commercial vehicle applications.

As freight-specific GPS datasets became available, researchers began examining their potential for identifying and interpreting commercial vehicle activities. Gingerich et al. (2016) classified the purposes of stopped-truck events using entropy-based measures derived from GPS trajectories. Siripirote et al. (2020) developed statistical methods for estimating freight activity patterns from truck GPS data and demonstrated the value of distinguishing operational freight activities from other stopping events. Related research on passive GPS-based commercial vehicle monitoring has also shown that the accuracy of freight-activity analysis depends strongly on the identification of trip ends, stop purposes, and tour structures (Greaves & Figliozzi, 2008; Gingerich et al., 2016).

For urban last-mile couriers, stop identification is particularly challenging because vehicles operate in dense street networks in which delivery addresses are closely spaced and stopping durations are generally short. A single street segment may contain genuine delivery events, temporary parking, failed-delivery waiting, and other non-delivery pauses within a limited spatial and temporal range. The limited availability of curbside parking can further alter stopping behaviour, operating time, and vehicle costs (Figliozzi & Tipagornwong, 2017). The growth of on-demand and instant-delivery services has intensified these operational pressures in European and other urban contexts (Dablanc et al., 2017). Moreover, because a courier vehicle may generate a large number of stop candidates during a single shift, even a modest per-stop classification error may accumulate into substantial route- and fleet-level inaccuracies.

Three limitations motivate the present study. First, many GPS-based stop-identification procedures rely on fixed dwell-time thresholds, distance buffers, and other rule-based heuristics. Although such approaches are transparent, their performance may vary across urban environments, vehicle types, sampling frequencies, and operating models (Du & Aultman-Hall, 2007; Shen & Stopher, 2014). Second, the construction of reliable ground-truth labels often requires manual annotation or external operational records. Such processes are costly and may introduce label noise or inconsistency, which can substantially affect supervised classification performance (Frenay & Verleysen, 2014). Weak-supervision and data-programming research

has demonstrated that scalable training labels can be constructed from multiple imperfect rules or information sources, but the reliability of these labels must be explicitly assessed (Ratner et al., 2016; Zhou, 2018). Third, genuine delivery stops may constitute only a minority of all detected stop candidates, creating a class-imbalance problem that conventional classifiers may handle poorly (He & Garcia, 2009). Resampling methods such as SMOTE provide one established approach for improving minority-class learning under such conditions (Chawla et al., 2002).

To address these gaps, this paper develops an end-to-end framework for identifying delivery stops from the GPS trajectories of urban courier vehicles by integrating vehicle traces with electronic waybill records. The framework converts raw GPS tracks into labelled delivery-stop events through three structured modules: (1) trajectory preprocessing and stop-candidate generation; (2) feature engineering and dependency validation; and (3) imbalance-aware supervised classification. Five interpretable features—stop duration, pre-stop speed, heading change, local stop density, and distance from the departure hub—are designed to capture the kinematic and spatial differences between delivery and non-delivery stops in an urban environment. SMOTE is applied exclusively to the training data before the parallel training of support vector machine, k-nearest-neighbour, and decision-tree classifiers under a unified evaluation protocol. These classifiers are grounded in established statistical-learning principles (Cover & Hart, 1967; Breiman et al., 1984; Vapnik, 1995).

This paper makes three methodological contributions. First, it adapts waybill-based ground-truth construction to the urban last-mile context by calibrating spatial buffers and temporal tolerance windows to the shorter and denser stop patterns of urban courier operations. Second, it demonstrates that five compact, domain-grounded features are sufficient to achieve classification accuracy above 98.9% across three structurally different learners under severe class imbalance. This finding suggests that carefully designed operational features can support accurate and interpretable classification without requiring highly complex representation-learning architectures. Third, through a controlled cross-model comparison under identical data and evaluation conditions, the paper identifies classifier-specific biases and shared blind spots in the feature space, providing practical guidance for future feature enrichment and model selection.

At the operational level, accurate delivery-stop identification enables carriers to calculate stop-level performance indicators, including service-time distributions, on-time delivery rates, and customer-level dwell-time profiles. At the network level, identified stop clusters can support decisions concerning urban consolidation centres, micro-depots, parcel lockers, and other last-mile delivery concepts (Allen et al., 2012; Boysen et al., 2021). From a policy perspective, stop-level freight data can be integrated with congestion, parking, and environmental assessments to evaluate the local externalities of courier activities and support city-logistics planning (Crainic et al., 2009; Taniguchi et al., 2016; Lagorio et al., 2016).

The remainder of the paper is organised as follows. Section 2 presents the overall methodological framework and the three supervised classifiers. Section 3 describes the dataset and preprocessing pipeline, including stop-candidate generation and waybill-based ground-truth construction. Section 4 develops and validates the five-feature representation. Section 5 implements and compares the three classifiers under a unified evaluation protocol. Section 6 summarises the findings, limitations, and directions for future research.

2. Methodology

To identify delivery stops from urban courier GPS trajectories, this study applies three well-established supervised learning algorithms—Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Decision Tree (DT)—trained in parallel under a unified experimental scheme. The overall methodology framework, illustrated in Fig. 1, comprises three main modules: trajectory preprocessing and stop-candidate generation; feature engineering and dependency validation; and class-imbalance-aware supervised classification.

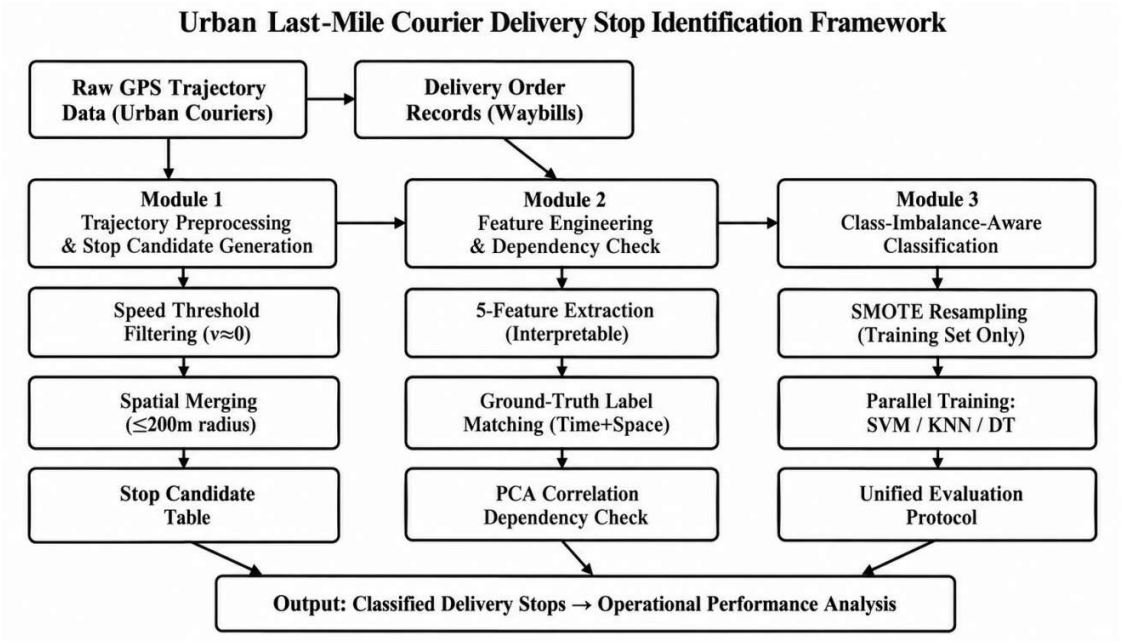


Figure 1. Framework for Identifying Urban Courier Delivery Stops

As shown in Fig. 1, the framework integrates GPS trajectory data with electronic waybill records to achieve scalable and reliable delivery-stop identification. The first module focuses on trajectory preprocessing and stop-candidate generation: noise and redundant GPS points are removed, and potential stop events are extracted based on spatiotemporal continuity and dwell characteristics. The second module constructs a feature representation for each candidate stop, emphasizing a compact, interpretable feature set grounded in urban courier operations. Feature dependencies are assessed prior to model training to prevent redundancy. The third module implements delivery-stop identification through supervised learning. Because genuine delivery stops account for only a small fraction of all detected stops, class imbalance must be explicitly addressed before training proceeds. All three classifiers are subsequently trained in parallel under a unified evaluation protocol, ensuring that observed performance differences reflect modelling assumptions rather than inconsistencies in data handling or feature construction.

2.1 Problem Formulation as Weakly Supervised Learning on a Spatio-Temporal Graph

From a methodological standpoint, the proposed framework can be reformulated as a weakly supervised learning problem on a spatiotemporal graph. Each candidate stop extracted from the GPS trajectories constitutes a graph node characterized by five interpretable behavioral features—stop duration, pre-stop speed, heading change, local stop density, and distance from the departure hub—while directed edges connect temporally adjacent stops along the same vehicle trajectory. This graph-based view provides a conceptual foundation for interpreting both the label-construction and classification stages.

Training labels are derived by matching candidate stops to waybill records within predefined spatial and temporal tolerances. While operationally practical, this procedure provides only weak supervision: genuine delivery stops falling outside the matching window receive no positive label, and certain non-delivery stops may be incorrectly labelled because of overlapping windows or address-geocoding imprecision. Under this formulation, the proposed workflow operates as a two-stage heuristic solution. The first stage reduces systematic label noise through conservative spatiotemporal matching rules and the exclusion of trajectories lacking waybill corroboration. The second stage uses interpretable stop-level features and standard supervised classifiers to learn a decision boundary that is tolerant of residual noise. Although the current implementation does not explicitly exploit the full graph structure, it is conceptually aligned with

weakly supervised learning approaches that use imperfect heuristic labels and mitigate their limitations through careful label construction and feature design (Ratner et al., 2016; Zhou, 2018; Frenay & Verleysen, 2014). The use of resampling to address class imbalance is further supported by established work on imbalanced learning and SMOTE (He & Garcia, 2009; Chawla et al., 2002).

2.2 Support Vector Machine (SVM)

The Support Vector Machine (SVM) is a supervised learning method grounded in Vapnik-Chervonenkis (VC) dimension theory and the structural risk minimization (SRM) principle (Vapnik, 1995). Its core objective is to find a maximum-margin hyperplane in feature space that separates positive (delivery) from negative (non-delivery) stop candidates. For a binary classification dataset $D = \{(x(n), y(n))\}$ where $y(n) \in \{+1, -1\}$, the separating hyperplane $\omega^T x + b = 0$ satisfies $y(n)(\omega^T x(n) + b) \geq 1$ for all $n \in \{1, \ldots, N\}$. The margin γ is maximized by solving:

$$\max^T,b \ \ 1 / \|\omega\|^2 \quad \text{subject to: } y(n)(\omega^T x(n) + b) \geq 1, \forall n \qquad (1)$$

Sample points satisfying $y(n)(\omega^T x(n) + b) = 1$ are the support vectors that define the margin boundaries. A larger margin γ implies a more stable, noise-resistant classification boundary. For non-linearly separable stop candidates, the Radial Basis Function (RBF) kernel $K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2)$ maps inputs into a higher-dimensional space where a linear separator can be found. The RBF kernel is selected on the basis of its consistently strong performance in prior GPS-based classification studies.

2.3 K-Nearest Neighbors (KNN)

KNN (Cover and Hart, 1967) classifies a test stop candidate z by majority vote among its K nearest training-set neighbors. Given training dataset $T = \{(x_i, y_i)\}$, the Euclidean distance from z to each x_i is:

$$d(z, x_i) = \|z - x_i\|_2 = \sqrt{[\Sigma_j \ (z_j - x_{ij})^2]} \qquad (2)$$

The classification decision follows plurality vote: $\hat{y} = argmax_c \ \Sigma_i \ \mathcal{A}[y_i = c, x_i \in KNN(z)]$, where KNN(z) denotes the K nearest neighbors of z. KNN makes no parametric assumptions about the feature distribution, an advantage when stop features exhibit non-Gaussian or multi-modal distributions—as commonly observed in urban courier GPS data containing delivery stops, signal pauses, break stops, and hub visits simultaneously.

2.3.1 Decision Tree (DT)

The Decision Tree (Breiman et al., 1984) recursively partitions the feature space by selecting, at each internal node, the feature and threshold that maximise the information gain IG:

$$IG(t, f, \theta) = H(t) - [n_L/n \cdot H(t_L) + n_R/n \cdot H(t_R)] \qquad (3)$$

where $H(\cdot)$ is Shannon entropy, t is the current node, t_L and t_R are left and right child nodes after splitting on feature f at threshold θ, and n_L, n_R are the respective child sample counts. Splitting continues until information gain is negligible or the stopping criterion is met. Decision Trees produce human-readable if–then rules, making them particularly valuable for operational deployment where logistics practitioners need to understand and audit classification logic without algorithmic expertise.

3. GPS Data Processing for Urban Courier Vehicles

3.1 Data Description and Preprocessing

The empirical dataset used in this study was obtained from an express parcel-courier company operating more than 80 delivery vehicles across a major metropolitan area throughout the 2022 calendar year. GPS tracking devices recorded position fixes at average intervals of approximately 30 seconds, with actual

intervals ranging from 10 to 120 seconds depending on signal quality and device firmware. The delivery network spans central urban districts and multiple suburban zones, covering road types from motorway feeder roads to narrow residential alleys. Table 1(a) and Table 1(b) present representative samples of the collected GPS and waybill data respectively.

Table 1(a). Partial GPS Trajectory Data — Urban Courier Vehicles

Vehicle ID	Record Time	Speed (km/h)	Mileage (km)	Cumul. Dist. (km)	GPS Status	Longitude	Latitude
Courier01	2022/4/12 08:31:00	0	0	0	ACC On, 3D Fix	121.48231	31.23450
Courier01	2022/4/12 08:31:30	12	0.18	0.18	ACC On, 3D Fix	121.48279	31.23482
…	…	…	…	…	…	…	…
Courier01	2022/9/7 18:22:00	0	88	88	ACC On, 3D Fix	121.51037	31.24891

**Certain details have been anonymized in compliance with data-sharing requirements.*

Table 1(b). Partial Electronic Waybill (Order) Data

Waybill ID	Vehicle ID	Hub	District	Planned Time	Earliest Arr.	Latest Arr.	Parcels
WB20220412001	Courier01	Hub-East	Jing An Dist.	2022/4/12 10:15	2022/4/12 09:15	2022/4/12 11:15	3
WB20220412002	Courier01	Hub-East	Putuo Dist.	2022/4/12 11:30	2022/4/12 10:30	2022/4/12 12:30	1
…	…	…	…	…	…	…	…
WB20220412058	Courier01	Hub-East	Changning Dist.	2022/4/12 16:45	2022/4/12 15:45	2022/4/12 17:45	2

**Certain details have been anonymized in compliance with data-sharing requirements.*

The irregular sampling intervals of the raw GPS data create both redundancy and gaps. Furthermore, some vehicles exhibit trajectory segments that cannot be matched to any waybill record, for instance, repositioning trips between hubs or personal errands during shift breaks. Because delivery and non-delivery stops cannot be reliably distinguished without waybill corroboration, these unverified segments must be excluded. Two primary preprocessing steps are applied:

Removal of redundant data. GPS data exported in daily batches produced duplicate records at batch boundaries. All duplicates (identified by matching vehicle ID and timestamp within a one-second window) were removed. On the waybill side, the same physical delivery event sometimes generates multiple waybill numbers due to split shipments or address corrections; such cases were merged by grouping records with identical vehicle ID, confirmed delivery timestamp, and geocoded destination into a single canonical delivery event.

Removal of trajectories lacking waybill corroboration. Trajectory segments recorded outside scheduled shift windows, or in locations inconsistent with the reported cumulative mileage, were flagged as operationally unexplained and excluded. After cleaning, 83 of the 87 vehicles in the raw dataset retained complete, corroborated trajectories suitable for analysis.

3.2 Identification of Stops in GPS Data

Raw GPS datasets consist of discrete position points and contain no explicit stop-event records. To identify stops, consecutive GPS points with instantaneous speed below 2 km/h were first aggregated into contiguous stationary intervals. A threshold of 2 km/h rather than strict zero is adopted to accommodate GPS speed noise at rest, which regularly produces non-zero readings of 1–3 km/h even for a stationary vehicle. Stop candidates with total duration below one minute were discarded as likely signal artefacts or momentary idling at traffic signals. Adjacent stop fragments within 200 meters were then merged spatially, since GPS noise and minor vehicle repositioning within a delivery block—such as moving the van to the other side of a building entrance—can fragment a single stop event into multiple discrete entries. The resulting stop-candidate data are presented in Table 2.

Table 2. Processed Stop-Candidate Records (Representative Sample)

StopID	Vehicle ID	StartTime	EndTime	Duration (min)	TimeDelta (min)	MeanLat	MeanLon	Distance (m)
0	Courier01	2022/4/12 08:31:00	2022/4/12 08:35:22	4.4	0	31.23450	121.48231	0
1	Courier01	2022/4/12 08:48:10	2022/4/12 08:51:05	2.9	12.8	31.23694	121.48512	412.7
…	…	…	…	…	…	…	…	…
287	Courier01	2022/4/12 17:09:14	2022/4/12 17:18:42	9.5	18.3	31.24877	121.51021	9847.6

**StopID is sequenced per vehicle per day. TimeDelta is elapsed time since the previous stop. Certain details anonymized.*

4. Feature Extraction for Urban Courier Delivery Stops

4.1 Characteristics of Urban Courier Delivery Stops

To train machine learning classifiers for delivery-stop identification, it is essential to select features that capture genuine behavioral differences between delivery and non-delivery stops. This section derives five features from close observation of urban courier GPS trajectories, each grounded in an explicit operational rationale drawn from the daily workflow of urban couriers.

Feature 1—Stop Duration (F1). Delivery stops require the courier to retrieve parcels from the vehicle, transport them to the recipient, and obtain a signature or scan confirmation. Even for a single small parcel, this physical process typically takes three to eight minutes. Traffic-signal stops last under two minutes, while double-parking pauses for loading or unloading at commercial premises can overlap with this range. Stop duration is therefore the most direct proxy for delivery activity and consistently emerges as the primary split feature in decision-tree classifiers trained on urban courier data.

The simulated route map in Fig. 2 illustrates the spatial distribution of delivery stops, non-delivery stops, and the departure hub for a representative urban courier shift. Delivery stops (red stars) are distributed across residential and commercial blocks throughout the service territory, while non-delivery stops (orange squares) tend to cluster near major intersections and service facilities.

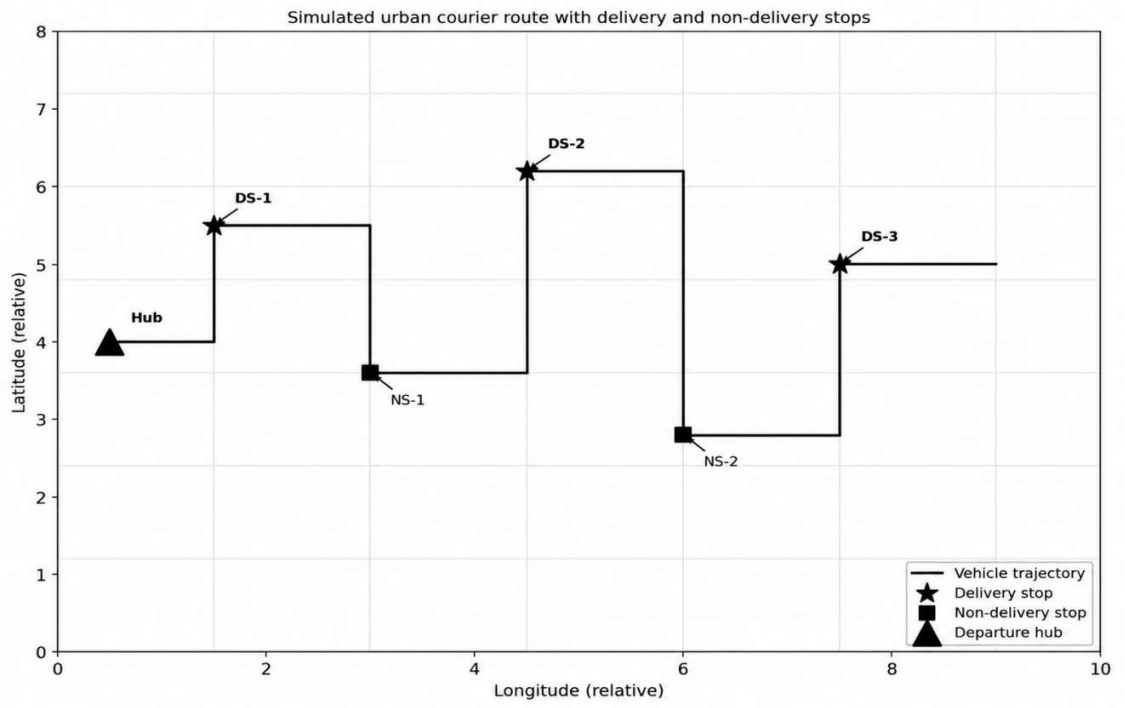


Figure 2. Simulated Urban Courier Route and Stop Locations

Feature 2—Pre-stop Speed (F2). Delivery addresses in urban environments are typically located on secondary streets, residential lanes, or inside compound driveways where vehicle speeds are low. In contrast, non-delivery stops such as waits at traffic signals often occur on major arterial roads where the vehicle has been travelling at 30 to 50 km/h. The mean GPS-recorded speed during the 60 seconds immediately preceding each stop provides a robust proxy for road type and access context, serving as an indicator of whether the stop is situated on a low-speed access road consistent with a delivery address.

Feature 3—Heading Change (F3). After completing a delivery, a courier typically retraces the approach path to return to the main road network before proceeding to the next stop. This creates an acute heading change at the stop location—the departure bearing is close to the reverse of the arrival bearing. In contrast, stops at traffic signals or rest areas usually involve minimal heading change because the vehicle continues in the same general direction. Fig. 3 illustrates this contrast.

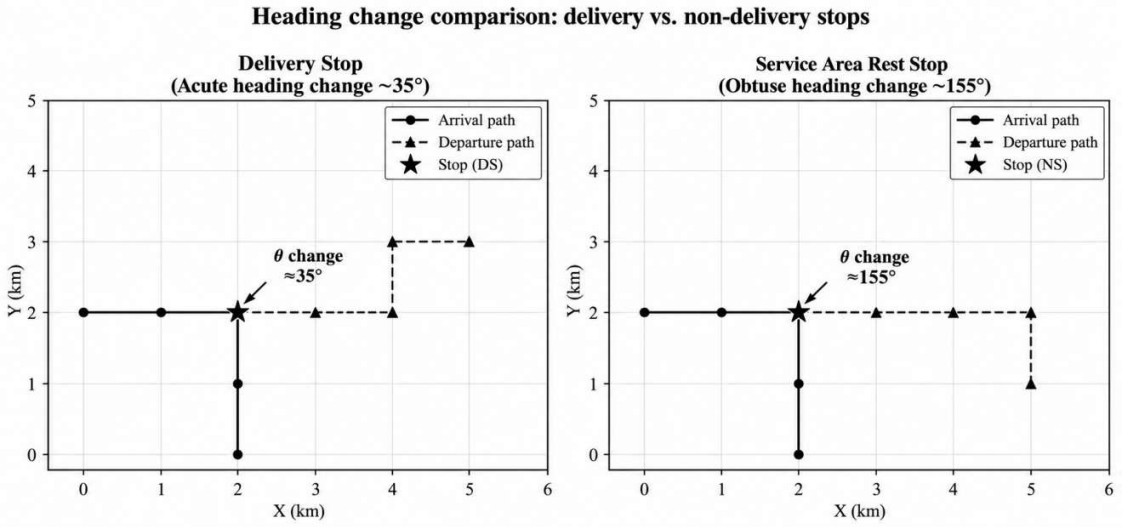


Figure 3. Heading Changes at Delivery and Non-Delivery Stops

To compute the heading change accurately, WGS84 latitude-longitude coordinates must first be converted to a three-dimensional Cartesian reference frame using the WGS84 ellipsoid (semi-major axis a = 6,378,137.0 m; semi-minor axis b = 6,356,752.3145 m). Fig. 4 illustrates how the Z and X/Y coordinates are derived from the latitude and longitude circles respectively.

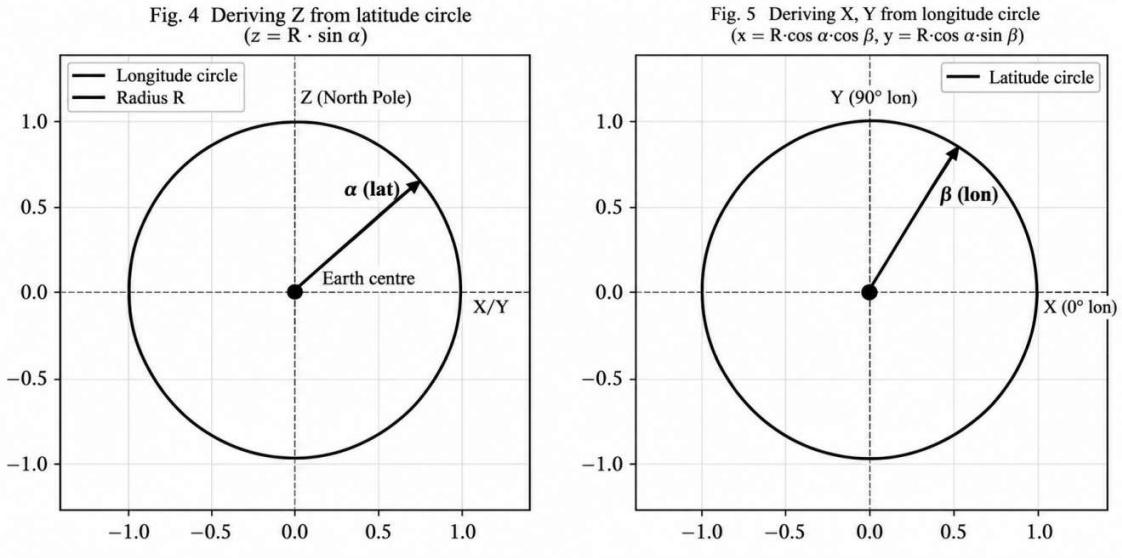


Figure 4. Coordinate Components Derived from Latitude and Longitude

The coordinate transformation equations are:

$$z = b \cdot \sin(\alpha) \quad (4)$$

$$x = a \cdot \cos(\alpha) \cdot \cos(\beta) \quad (5)$$

$$y = a \cdot \cos(\alpha) \cdot \sin(\beta) \quad (6)$$

$$x_corr = (a^2 \cos(\alpha) \cos(\beta)) / \sqrt{(a^2 \cos^2(\alpha) + b^2 \sin^2(\alpha))} \quad (7)$$

where α is WGS84 geodetic latitude and β is longitude. The heading change at each stop is computed as the angle between the unit arrival vector and the unit departure vector in Cartesian space using the vector dot product. Equation (7) corrects for the ellipsoidal shape of the Earth, which would otherwise introduce position errors of up to 21 km at mid-latitudes if spherical formulas were used.

The urban courier dataset also contains many stop records at departure hubs, transfer sorting centers, and service stations. These locations share kinematic characteristics with genuine delivery stops—long dwell times, low pre-stop speeds, and acute heading changes—making them difficult to distinguish using Features F1–F3 alone. The critical observation is that hubs and sorting centers are visited by many vehicles on every shift, producing dense spatial clusters of stop events throughout the dataset, whereas individual delivery addresses are visited only rarely.

Feature 4—Local Stop Density (F4). The count of all stop candidates within a 1 km radius of a given stop captures this spatial density contrast. Hub locations typically exhibit densities of 30 to 80 near stops within the 1 km radius, whereas individual delivery stops show densities below 5. Fig. 5 illustrates the clear spatial distinction.

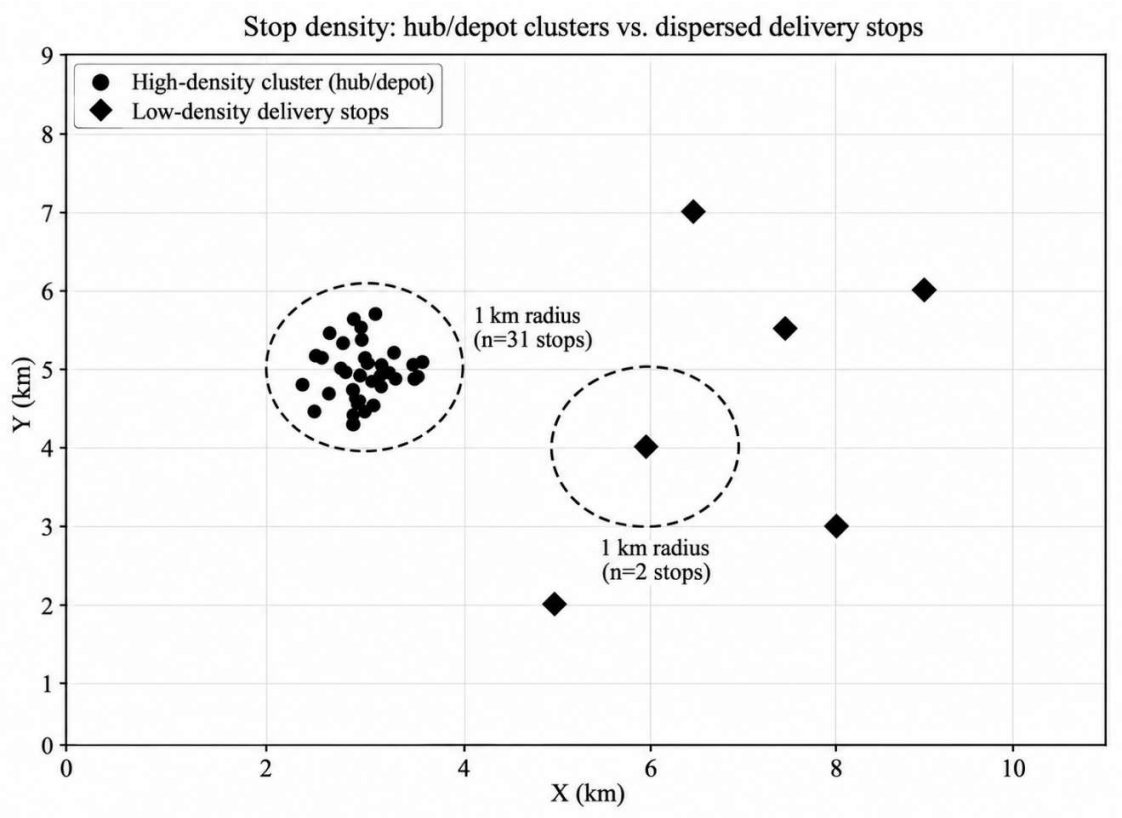


Figure 5. Local Stop Density Near Hubs and Delivery Locations

Inter-stop distances are calculated using the Haversine formula:

$$d = 2r \cdot \arcsin(\sqrt{[\sin^2(\Delta\varphi/2) + \cos(\varphi_1) \cdot \cos(\varphi_2) \cdot \sin^2(\Delta\lambda/2)]}) \quad (8)$$

where φ_1, φ_2 are stop latitudes, $\Delta\varphi$ and $\Delta\lambda$ are latitude and longitude differences, and r = 6,371 km is the mean Earth radius.

Feature 5—Distance from Departure Hub (F5). Urban couriers operate within a bounded service territory radiating from a local distribution hub. Stops very close to the hub (within 0.5 to 2 km) are likely to

be loading, sorting, or administrative activities at the hub premises rather than customer deliveries. Genuine delivery stops are distributed more broadly across the service territory, typically more than 2 km from the hub. The straight-line Euclidean distance from each stop centroid to the nearest hub geocoordinate provides a simple but effective spatial-context feature that complements the local density signal of F4.

In summary, the five selected features are: (1) stop duration, (2) pre-stop speed, (3) heading change at the stop, (4) count of stop candidates within a 1 km radius (local stop density), and (5) straight-line distance from the departure hub. All five features are hand-crafted and explicitly grounded in domain knowledge of urban courier operations. Critically, they require no external GIS layers or map-matching infrastructure, making the framework deployable with minimal additional data collection costs. Their interpretability also allows logistics practitioners to validate and adjust the feature set in operational deployment—an advantage that is far more difficult to realize with latent embeddings from black-box models.

4.2 Correlation Analysis of Features

Adding irrelevant or highly correlated features to a machine learning model typically degrades generalization performance and inflates computational cost without improving accuracy. To verify that the five selected features are mutually non-redundant, their Pearson correlation matrix is computed and presented in Table 3.

Table 3. Pearson Correlation Matrix of the Five Stop-Level Features

	Duration (min)	**Pre-stop Speed (km/h)**	**Heading Change (°)**	**Stop Density (1 km)**	**Distance-Hub (km)**
Duration (min)	1.000	−0.097	−0.112	0.083	−0.074
Pre-stop Speed (km/h)	−0.097	1.000	0.287	−0.274	0.131
Heading Change (°)	−0.112	0.287	1.000	−0.198	0.158
Stop Density (1 km)	0.083	−0.274	−0.198	1.000	−0.241
Distance-Hub (km)	−0.074	0.131	0.158	−0.241	1.000

As shown in Table 3, all pairwise correlations are well below the conventional multicollinearity threshold of 0.5. The largest observed coefficient is 0.287 between pre-stop speed and heading change, reflecting the mild tendency for approaches from faster arterial roads to also involve less heading deviation—a plausible operational relationship that nonetheless leaves both features carrying largely non-redundant information. Consequently, no dimensionality reduction is required, and all five features are retained for subsequent model training.

5. Implementation of Urban Courier Delivery Stop Identification

5.1 Addressing Class Imbalance in the Stop Dataset

A fundamental challenge in this dataset is the severe imbalance between delivery and non-delivery stops. Of the 28,456 stop candidates extracted from the GPS trajectories of the 83 usable vehicles, only 832 are confirmed as delivery stops through waybill match a positive-class rate of 2.92%, corresponding to approximately 33.2 non-delivery stops for every genuine delivery stop. Standard classification algorithms trained on such highly imbalanced data tend to predict the majority class for almost all samples, achieving deceptively high overall accuracy while failing to identify most of the rare delivery stops.

This study adopts the Synthetic Minority Over-sampling Technique (SMOTE) proposed by Chawla et al. (2002), which generates synthetic minority-class samples by interpolating in feature space between existing minority-class samples and their K nearest minority-class neighbors. Unlike random over-sampling,

which merely duplicates existing positive samples, SMOTE creates new, plausible delivery-stop instances that enrich the decision boundary in the minority-class region. Critically, SMOTE is applied only to the training partition; the test set retains the original imbalanced class distribution to ensure that reported metrics faithfully reflect real-world deployment conditions. The SMOTE implementation steps are illustrated in Fig. 6.

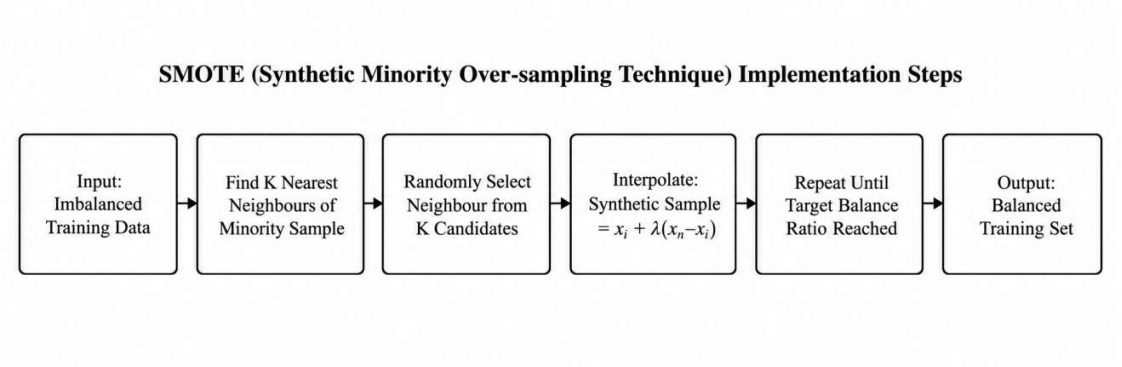


Figure 6. SMOTE Implementation Procedure

After SMOTE augmentation, the training set achieves a balanced 1:1 delivery-to-non-delivery ratio. The test set contains 5,691 stop candidates (20% of the total dataset) at the original 2.92% positive-class rate. All evaluation metrics reported in Sections 5.2 through 5.4 are computed on this original-distribution test set, with training-set metrics included for reference.

5.2 Stop Classification Using Support Vector Machines

5.2.1 Selection of an Appropriate Kernel Function

Four commonly used kernel functions were evaluated on a 30% random subsample of the SMOTE-augmented training data to select the most suitable kernel: linear, polynomial, RBF, and sigmoid. The comparison results, including identification accuracy and computation time, are shown in Fig. 7.

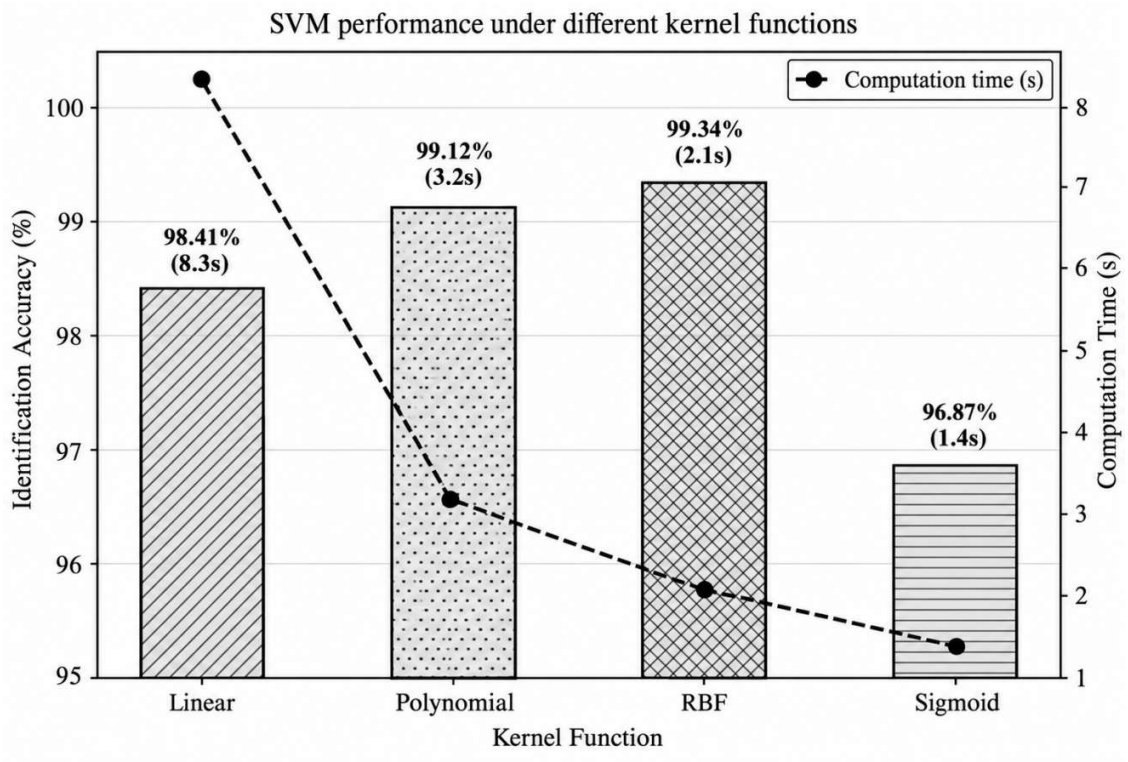


Figure 7. SVM Accuracy and Computation Time across Kernel Functions

As shown in Fig. 7, the linear kernel achieves acceptable accuracy (98.41%) but is the slowest at 8.3 seconds on the subsample, as the large number of support vectors required makes the quadratic program computationally expensive. The sigmoid kernel is fastest but shows the lowest accuracy (96.87%), reflecting its known instability for non-Gaussian feature distributions. Both the polynomial kernel (99.12%, 3.2 s) and the RBF kernel (99.34%, 2.1 s) perform well. Given that the RBF kernel achieves the highest accuracy with the second-shortest computation time, and is consistent with prior GPS-based freight-stop classification studies (Gingerich et al., 2016), it is selected for all subsequent SVM training.

5.2.2 Hyperparameter Tuning for the SVM Model

SVM performance depends primarily on two hyperparameters: the penalty coefficient C and the RBF kernel bandwidth γ. A large C forces the model to classify all training samples correctly (risking overfitting), while a small C permits more training misclassifications (risking underfitting). A large γ restricts each support vector to a highly localised influence (overfitting risk), while a small γ spreads that influence broadly (underfitting risk). To identify optimal values, a nested 5-fold cross-validation strategy is applied with a grid search over $C \in \{0.01, 0.1, 0.5, 1, 2, 5, 10\}$ and $\gamma \in \{0.001, 0.01, 0.1, 1, 5, 10\}$. The cross-validation accuracy heatmap is shown in Fig. 8 and the three-dimensional surface in Fig. 9.

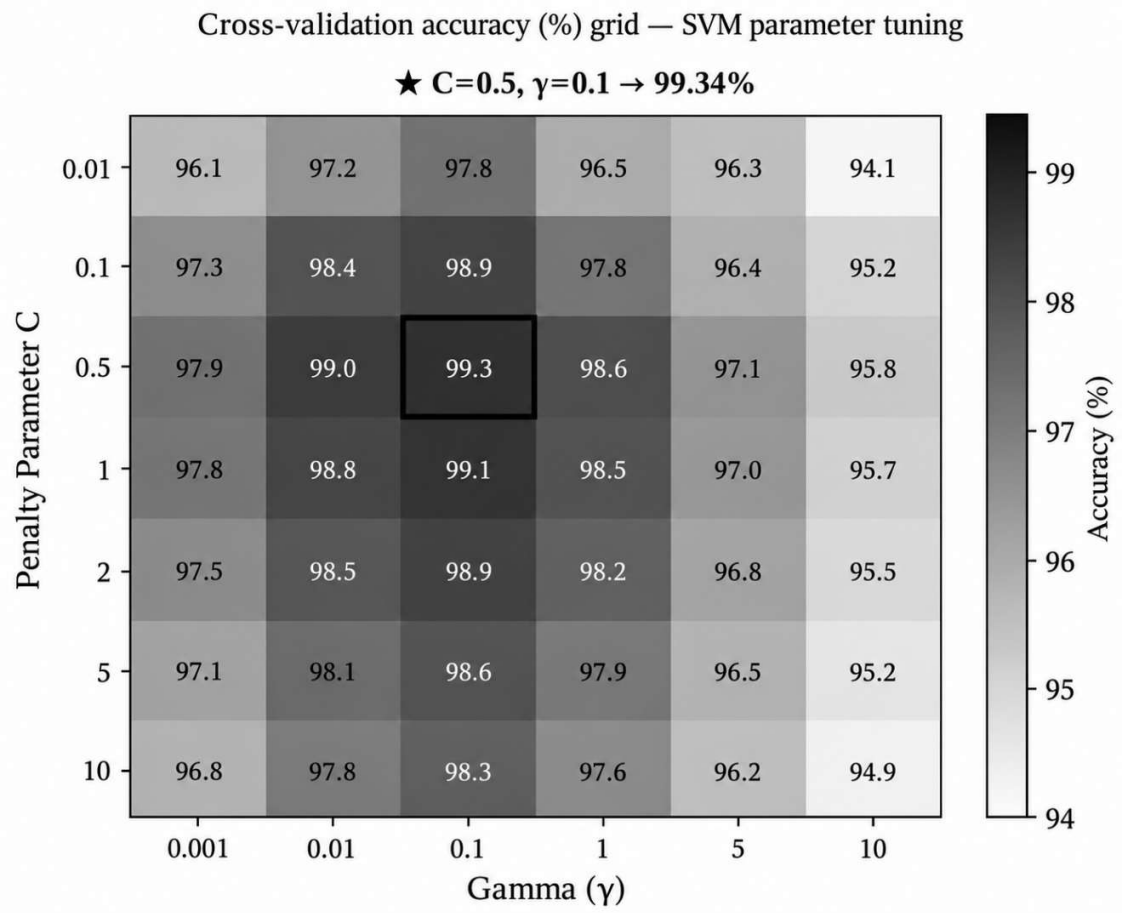


Figure 8. Cross-Validation Accuracy for SVM Hyperparameter Selection

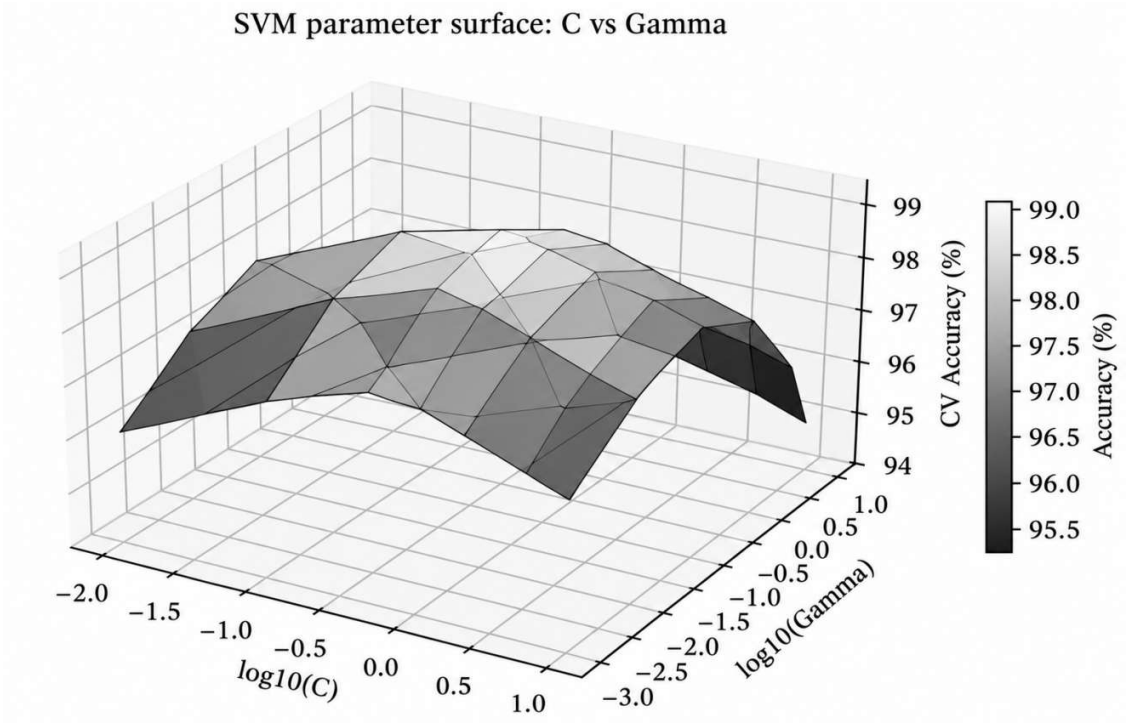


Figure 9. SVM Accuracy Surface across C and γ

The maximum cross-validation accuracy of 99.34% is achieved at C = 0.5 and γ = 0.1. These values are adopted as the final SVM hyperparameters for full-dataset training and evaluation.

5.2.3 Classification Results Based on SVM

The SMOTE-augmented SVM with RBF kernel (C = 0.5, γ = 0.1) was implemented in Python 3.10 using scikit-learn. Classification results for both training and test sets are presented in Table 4.

Table 4. Classification Results Based on SVM

	Training Set	**Test Set**
Identification Accuracy	0.9998	0.9932
False Negatives (FN)	0	38
False Negative Rate (FNR)	0.0000	0.2280
False Positives (FP)	4	2
False Positive Rate (FPR)	0.0001	0.0001

**FN: delivery stop incorrectly classified as non-delivery. FP: non-delivery stop incorrectly classified as delivery.*

The SVM achieves 99.98% training-set accuracy and 99.32% test-set accuracy. While overall classification performance is highly satisfactory, the test-set false negative rate of 22.80% corresponding to 38 missed delivery stops out of approximately 167 in the test sets the primary limitation. This conservative behavior is consistent with SVM's margin-maximization objective: even after SMOTE augmentation, the maximum-margin hyperplane tends to be positioned conservatively toward the majority class, trading reduced false alarms for a higher rate of missed deliveries. The false positive rate is negligibly low (0.01%), meaning non-delivery stops are almost never misclassified as delivery events.

5.3 Stop Classification Using K-Nearest Neighbors

5.3.1 Hyperparameter Tuning for the KNN Model

The key hyperparameter for KNN is the number of neighbors K. An excessively small K increases model sensitivity to noise in the SMOTE-augmented feature space, while an excessively large K over-smoothes the decision boundary and reduces recall. A 5-fold cross-validation grid search over K ∈ {1, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21} was applied to identify the optimal value. Results are shown on Fig. 10.

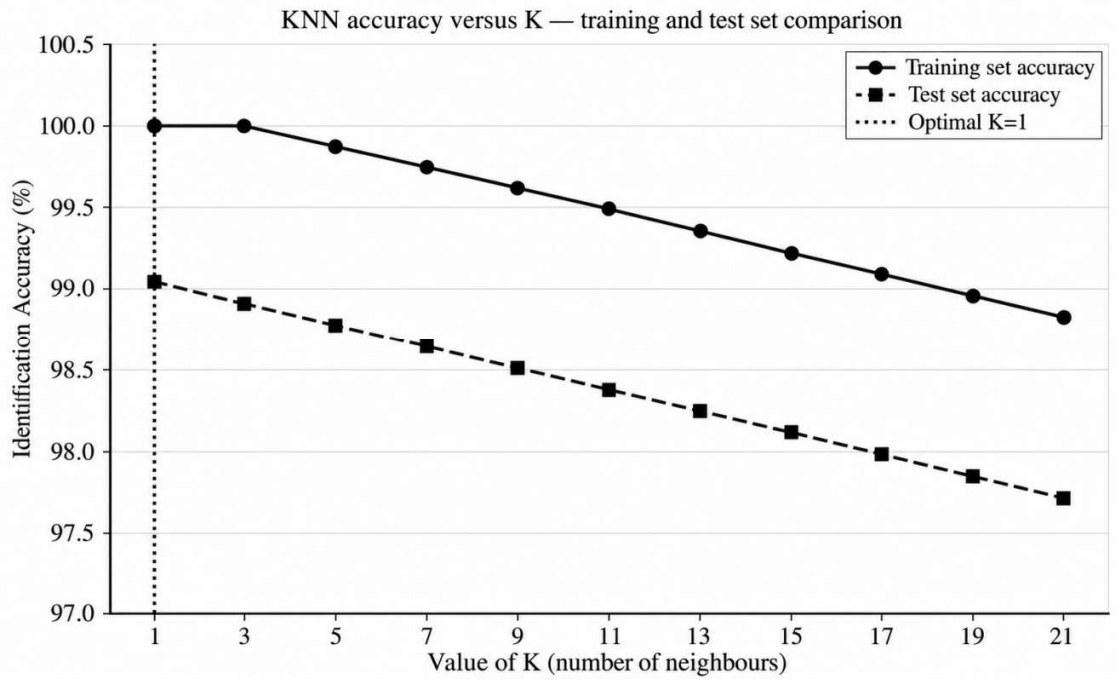


Figure 10. KNN Training and Test Accuracy across K Values

Both training and test accuracy decrease monotonically as K increases. Test-set accuracy peaks at K = 1 (99.06%), after which it drops steadily. At K = 1, each test sample is classified by its single nearest neighbor in the SMOTE-augmented feature space, placing the decision boundary at the midpoint between adjacent positive and negative training samples—tighter than SVM's global maximum-margin hyperplane. This explains why KNN achieves a substantially lower false negative rate than SVM at the cost of slightly more false positives. K = 1 is adopted for the final model.

5.3.2 Classification Results Based on KNN

KNN achieves 100% training-set accuracy (trivially, since K = 1 maps each sample to itself) and 99.06% test-set accuracy. The false negative rate falls to 6.59% (11 missed delivery stops)—a substantial improvement over SVM—while the false positive rate rises to 0.88% (48 false alarms). Since both error rates remain comfortably below 10%, the overall KNN performance is highly satisfactory (Table 5). The lower FNR reflects KNN's instance-based nature: at K = 1, any test stops spatially close in feature space to a genuine or synthetic delivery stop in the training set will be classified positively, making the classifier inherently more sensitive to the minority class than SVM's global margin approach.

Table 5. Classification Results Based on KNN

	Training Set	Test Set
Identification Accuracy	1.0000	0.9906
False Negatives (FN)	0	11
False Negative Rate (FNR)	0.0000	0.0659
False Positives (FP)	0	48
False Positive Rate (FPR)	0.0000	0.0088

**FN: delivery stop incorrectly classified as non-delivery. FP: non-delivery stop incorrectly classified as delivery.*

5.4 Stop Classification Using Decision Trees

Decision Trees serve as a highly intuitive and transparent supervised learning method. The tree structure learned from the SMOTE-augmented training dataset is illustrated in Fig. 11. The root node splits on stop duration (F1) at a threshold of 8 minutes, consistent with its role as the primary discriminator identified in Section 4.1. Subsequent splits involve pre-stop speed, heading change, stop density, and hub distance, demonstrating that the classifier has internalized the same operational reasoning that motivated the feature design.

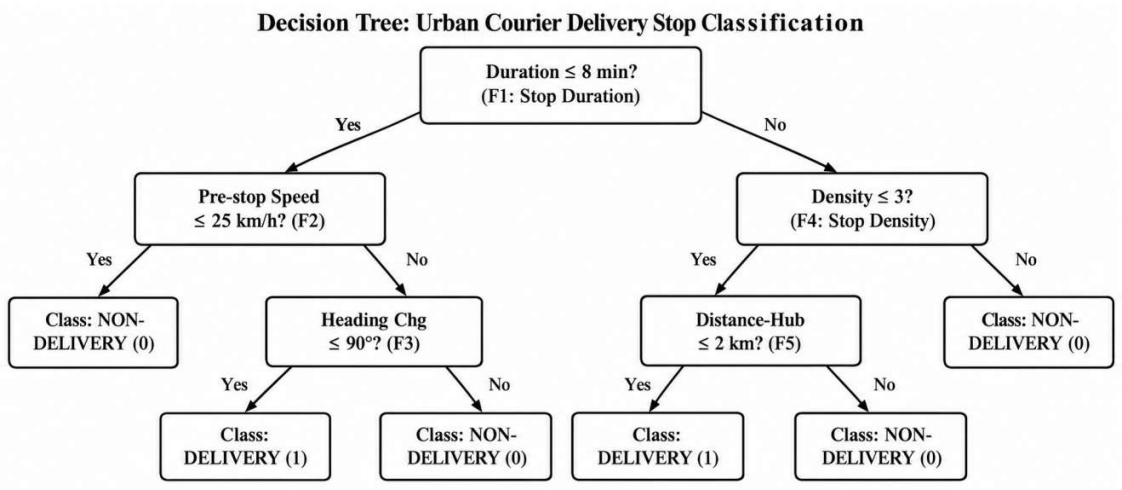


Figure 11. Decision Tree for Delivery Stop Classification

The DT achieves 100% training-set accuracy and 98.93% test-set accuracy (Table 6). The false negative rate of 8.98% (15 missed delivery stops) is marginally higher than KNN but dramatically lower than SVM, while the false positive rate of 0.84% is comparable to KNN. All three models achieve highly satisfactory training outcomes, but with meaningfully different errors and trade-offs.

Table 6. Classification Results Based on Decision Tree

	Training Set	Test Set
Identification Accuracy	1.0000	0.9893
False Negatives (FN)	0	15
False Negative Rate (FNR)	0.0000	0.0898
False Positives (FP)	0	46
False Positive Rate (FPR)	0.0000	0.0084

**FN: delivery stop incorrectly classified as non-delivery. FP: non-delivery stop incorrectly classified as delivery.*

Table 7 presents a cross-model comparison of test-set error patterns, including an estimate of the delivery stops missed by all three classifiers simultaneously, derived using set-intersection logic on the error indices.

Table 7. Three-Model Test-Set Error Comparison and Estimated Shared Misclassifications

Metric	SVM	KNN	Decision Tree (DT)
Test-set accuracy	0.9932	0.9906	0.9893
False Negatives (FN)	38	11	15
False Negative Rate (FNR)	0.2280	0.0659	0.0898
False Positives (FP)	2	48	46
False Positive Rate (FPR)	0.0001	0.0088	0.0084
Est. FN shared by all three	≈6–9 samples (≈4–5% of test-set delivery stops)		

**FN: delivery stop misclassified as non-delivery. FP: non-delivery misclassified as delivery. Test-set count (≈167) back-calculated from each model's FNR; dataset split 80/20. Shared-FN lower bound estimated by set-intersection logic.*

Comparing error profiles across the three classifiers reveals systematic differences reflecting both individual model biases and the shared limitations of the current feature set. SVM exhibits the highest false negative rate (FNR = 22.80%), consistent with its margin-maximization objective that remains conservative toward the minority class even after SMOTE augmentation. In contrast, KNN (FNR = 6.59%; FPR = 0.88%) and DT (FNR = 8.98%; FPR = 0.84%) achieve substantially lower missed-detection rates at the cost of slightly more false alarms, because their instance-based and partition-based local rules are more aggressive in claiming the positive class near the decision boundary.

The near-identical false positive counts of KNN (48) and DT (46) point to a shared feature-space blind spot rather than a model-specific artefact: non-delivery stops with high local density and moderate dwell times (4 to 8 minutes) occupy a region of feature space that all three classifiers find difficult to separate cleanly from genuine delivery events. Fig. 12 illustrates three distinct hard-case false-negative patterns as radar profiles of their feature values.

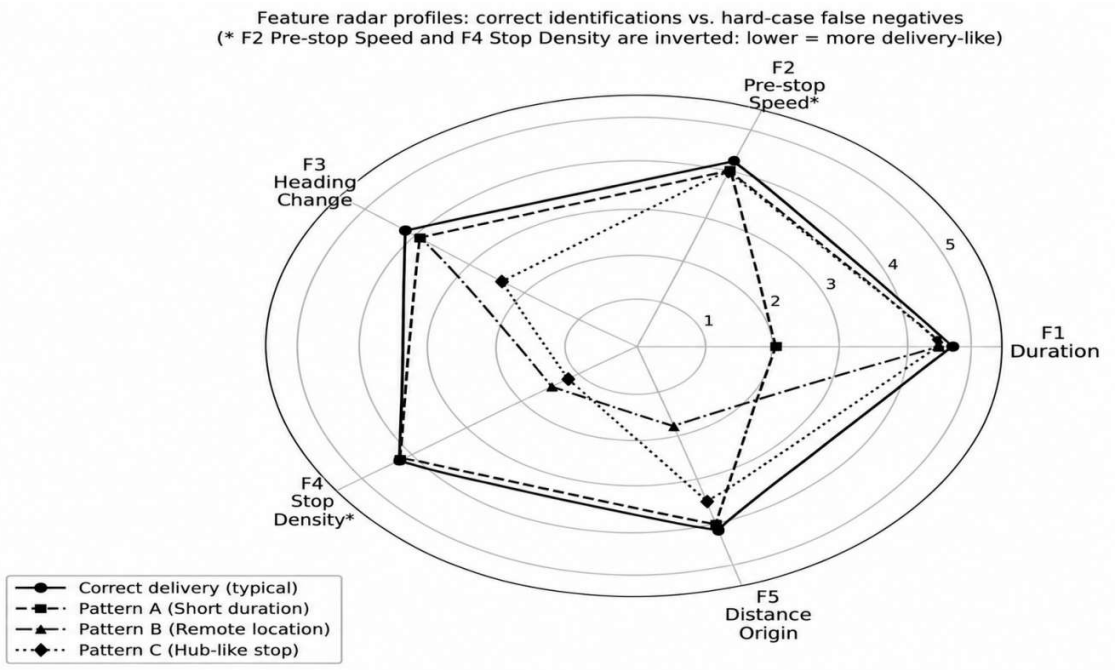


Figure 12. Feature Profiles of Typical Delivery Stops and False-Negative Cases

The three hard-case false-negative patterns are characterized in detail in Table 8.

Table 8. Detailed Feature Characterization of the Three Hard-Case False-Negative Patterns

Feature	Typical delivery (correctly classified)	Pattern A: Short-duration	Pattern B: Dense-area	Pattern C: Hub-district
F1 Stop Duration	Long (>5 min)	Short (2–4 min); overlaps with traffic signal waits [Key ambiguity]	Normal to long	Long
F2 Pre-stop Speed	Low; secondary road or residential lane	Low; still indicative	Low; normal	Low
F3 Heading Change	Acute; sharp directional reversal on departure	Acute; still indicative	Acute; normal	Moderate; variable in commercial districts
F4 Stop Density	Low (address rarely revisited)	Low; normal	High (dense commercial area) [Key ambiguity]	High; hub vicinity mimics depot signature [Key ambiguity]
F5 Distance-Hub	Far (>3 km from hub)	Far; normal	Moderate to far	Near hub district [Key ambiguity]
Suggested enrichment	—	Intra-shift temporal position: does stop fall within expected delivery time window?	Land-use type: residential vs. commercial vs. industrial block	First-vs-repeated visit indicator to distinguish first arrival at high-density locations

**Pattern A: short dwell overlaps with signal stops. Pattern B: dense commercial blocks resemble hub/depot areas. Pattern C: deliveries near hub premises sharing F4 and F5 characteristics with hub activities.*

6. Conclusions

This paper has presented an end-to-end framework for identifying delivery stops from passively collected GPS trajectories of urban last-mile courier vehicles, addressing a problem that is both operationally important and methodologically non-trivial due to the extreme scarcity of delivery events relative to the full stop population. Using 83 courier vehicles operating across a metropolitan area throughout 2022, the study extracted 28,456 stop candidates, constructed ground-truth delivery labels by matching candidates to electronic waybill records through calibrated spatiotemporal constraints, and trained three supervised classifiers—SVM, KNN, and DT—under a unified evaluation protocol that applied SMOTE resampling exclusively to the training partition.

The principal findings are as follows. First, five interpretable kinematic and spatial features—stop duration, pre-stop speed, heading change, local stop density, and distance from hub—are sufficient to achieve test-set classification accuracies of 99.32% (SVM), 99.06% (KNN), and 98.93% (DT), demonstrating that compact, domain-grounded features can match or exceed the practical utility of black-box representations when sample sizes are limited and misclassification costs are asymmetric. Second, the three classifiers exhibit systematically different error profiles reflecting their inductive biases: SVM is precision-oriented (very low FPR; high FNR), while KNN and DT are recall-oriented (low FNR; moderate FPR). This trade-off carries direct operational implications—SVM is preferable when false alarms are costly, while KNN or DT are preferable when missed deliveries carry the higher cost. Third, cross-model error analysis identifies three hard-case false-negative patterns—short-duration deliveries, deliveries in dense commercial areas, and deliveries near hub locations—each with specific feature ambiguities and concrete suggestions for feature enrichment.

The study has several limitations that point to productive directions for future research. First, the dataset covers a single carrier operating in one metropolitan area during one calendar year; generalizability to different urban morphologies, carrier operating models, or parcel types should be validated with additional datasets from other cities (Taniguchi et al., 2016; Dablanc et al., 2017). Second, the five kinematic features do not capture contextual signals such as time of day, land-use type, or driver confirmation logs that may further sharpen classification in the hard-case zones identified in Table 8 (Zhou, 2018). Third, while SMOTE is effective for moderate imbalance, the 33:1 ratio in this dataset remains challenging; future work could compare SMOTE with cost-sensitive learning, deep generative over-sampling, or ensemble approaches specifically designed for severe imbalance (Chawla et al., 2002; He and Garcia, 2009). Fourth, extending the framework to emerging delivery modes—cargo bikes, autonomous robots, and crowd-sourced couriers—would require redefining features appropriate to their distinct kinematics (Lagorio et al., 2016; Boysen et al., 2021). Fifth, future work could explicitly model the sequential structure of stops along a delivery route using graph neural networks or recurrent architectures, which may resolve the hard-case ambiguities that persist across all three current classifiers (Ratner et al., 2016; Zhou, 2018).

Acknowledgement

The authors gratefully acknowledge the express parcel-courier company for providing the GPS trajectory and electronic waybill data used in this study, and thank the anonymous reviewers for their constructive comments.

Author Contributions

Conceptualization, W.Z.; Methodology, W.Z.; Formal Analysis, W.Z.; Investigation, W.Z. and J.L. and H.C.; Data Curation, W.Z. and J.L. and H.C.; Software, W.Z. and J.L. and H.C.; Validation, W.Z. and J.L. and H.C.; Resources, J.L. and H.C; Visualization, W.Z. and J.L. and H.C.; Writing – Original Draft, W.Z.; Writing – Review & Editing, W.Z. and J.L. and H.C.; Supervision, W.Z.; Project Administration, W.Z. We have read and approved the final manuscript and take responsibility for the work.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Ethical Approval and Consent

This study analyzed previously collected courier-vehicle GPS trajectories and electronic waybill records. No new human participant recruitment or experimental intervention was undertaken. Selected operational details were anonymized in the reported examples.

Data Availability Statement

Representative data records and aggregate analytical results are presented in Tables 1–8. The data preprocessing, feature extraction, and evaluation procedures are described in the manuscript.

Declaration of Generative AI Use

Generative AI was used to assist with manuscript editing, formatting, and the preparation of disclosure statements. The authors take full responsibility for the accuracy, integrity, and originality of the manuscript.

References

Allen, J., Browne, M., Woodburn, A., & Leonardi, J. (2012). The role of urban consolidation centres in sustainable freight transport. Transport Reviews, 32(4), 473–490. doi: 10.1080/01441647.2012.688074.

Boysen, N., Fedtke, S., & Schwerdfeger, S. (2021). Last-mile delivery concepts: A survey from an operational research perspective. OR Spectrum, 43, 1–58. doi: 10.1007/s00291-020-00607-8.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and regression trees. Wadsworth & Brooks/Cole.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357. doi: 10.1613/jair.953.

Cover, T. M., & Hart, P. E. (1967). Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13(1), 21–27. doi: 10.1109/TIT.1967.1053964.

Crainic, T. G., Ricciardi, N., & Storchi, G. (2009). Models for evaluating and planning city logistics systems. Transportation Science, 43(4), 432–454. doi: 10.1287/trsc.1090.0279.

Dablanc, L., Morganti, E., Arvidsson, N., Woxenius, J., Browne, M., & Saidi, N. (2017). The rise of on-demand 'instant deliveries' in European cities. Supply Chain Forum: An International Journal, 18(4), 203–217. doi: 10.1080/16258312.2017.1375375.

Du, J., & Aultman-Hall, L. (2007). Increasing the accuracy of trip rate information from passive multi-day GPS travel datasets: Automatic trip end identification issues. Transportation Research Part A: Policy and Practice, 41(3), 220–232. doi: 10.1016/j.tra.2006.05.001.

Figliozzi, M. A., & Tipagornwong, C. (2017). Impact of last mile parking availability on commercial vehicle costs and operations. Supply Chain Forum: An International Journal, 18(2), 60–68. doi: 10.1080/16258312.2017.1333386.

Frenay, B., & Verleysen, M. (2014). Classification in the presence of label noise: A survey. IEEE Transactions on Neural Networks and Learning Systems, 25(5), 845–869. doi: 10.1109/TNNLS.2013.2292894.

Gingerich, K., Maoh, H., & Anderson, W. (2016). Classifying the purpose of stopped truck events: An application of entropy to GPS data. Transportation Research Part C: Emerging Technologies, 64, 17–27. doi: 10.1016/j.trc.2016.01.002.

Greaves, S. P., & Figliozzi, M. A. (2008). Collecting commercial vehicle tour data with passive Global Positioning System technology: Issues and potential applications. Transportation Research Record, 2049(1), 158–166. doi: 10.3141/2049-19.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263–1284. doi: 10.1109/TKDE.2008.239.

Lagorio, A., Pinto, R., & Golini, R. (2016). Research in urban logistics: A systematic literature review. International Journal of Physical Distribution & Logistics Management, 46(10), 908–931. doi: 10.1108/IJPDLM-01-2016-0008.

Mangiaracina, R., Perego, A., Seghezzi, A., & Tumino, A. (2019). Innovative solutions to increase last-mile delivery efficiency in B2C e-commerce: A literature review. International Journal of Physical Distribution & Logistics Management, 49(9), 901–920. doi: 10.1108/IJPDLM-02-2019-0048.

Ratner, A. J., De Sa, C. M., Wu, S., Selsam, D., & Ré, C. (2016). Data programming: Creating large training sets, quickly. In Advances in Neural Information Processing Systems 29 (pp. 3567–3575).

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

Shen, L., & Stopher, P. R. (2014). Review of GPS travel survey and GPS data-processing methods. Transport Reviews, 34(3), 316–334. doi: 10.1080/01441647.2014.903530.

Siripirote, T., Sumalee, A., & Ho, H. W. (2020). Statistical estimation of freight activity analytics from Global Positioning System data of trucks. Transportation Research Part E: Logistics and Transportation Review, 140, 101986. doi: 10.1016/j.tre.2020.101986.

Taniguchi, E., Thompson, R. G., & Yamada, T. (2016). New opportunities and challenges for city logistics. Transportation Research Procedia, 12, 5–13. doi: 10.1016/j.trpro.2016.02.004.

Vapnik, V. N. (1995). The nature of statistical learning theory. Springer. doi: 10.1007/978-1-4757-2440-0.

Zhou, Z.-H. (2018). A brief introduction to weakly supervised learning. National Science Review, 5(1), 44–53. doi: 10.1093/nsr/nwx106.