

Smart Logistics Using IoT: A Data-Driven Approach to Reduce Transportation Delays

Dr. Santanu Roy¹, Shaik Mahmood Ur Rahaman², Dr. Vikas Gulati³, Dr. Virendra Kumar Verma⁴, Dr V. K. Jain⁵

¹Professor, IILM Institute for Higher Education, New Delhi, India, ORCID ID: 0000-0003-4566-1040, Email ID: rsan58@yahoo.co.uk

²Assistant Professor, University Arts & Science College (Autonomous), Kakatiya University, Subedari, Hanamakonda, India, ORCID ID: 0009-0009-4778-0407, Email ID: shaikmahmoodurrahan@gmail.com

³Associate Professor, Mechanical Engineering, Lovely Professional University, Jalandhar, India, ORCID ID: 0000-0002-9917-767, Email ID: vikasgulati1979@gmail.com

⁴Assistant Professor, Symbiosis Institute of Business Management, Symbiosis International (Deemed University), Bengaluru, India, ORCID ID: 0000-0002-1115-122, Email ID: virendra.verma@sibm.edu.in

⁵Vice Chancellor, Teerthanker Mahaveer University, Moradabad, India, Email ID: drvkjain72@gmail.com

Abstract

Urban last mile delivery is a key part of smart city logistics, but one that is hampered by delays and has a major impact on service reliability and operational efficiency - and sustainability. Despite the progress made in Internet of Things (IoT) technologies and artificial intelligence, transportation delays are still rampant, as a result of urban congestion, spatial heterogeneity and demand volatility. This study examines how big data in the logistics industry, made available through IoT, can be harnessed using machine learning algorithms to predict and minimize last mile delivery delays. Using the LaDe dataset that contains more than 4.4 million real-world delivery records across five major Chinese cities, the study combines a quantitative data-driven method combining exploratory analysis, supervised learning, hyperparameters optimization, and simulation-based intervention evaluation. Logistic Regression, Random Forest and LightGBM models were created and tested, with a random forest model fine-tuned showing the best results, (80% accuracy and 0.857 ROC-AUC). Spatial and temporal features proved to be the most influential predictors, and thus the importance of local urban context. A risk-based simulation showed that targeted interventions for higher-risk deliveries are likely to enable a reduction of overall delay rates by over 8% with city-specific improvements of more than 20% in heavily congested areas. The findings confirm the value of operationalizing predictive analytics to provide meaningful reductions in delays and enable proactive and data-driven decision making in smart urban logistics systems.

Keywords: Last-mile delivery; IoT-enabled logistics; Delivery delay prediction; Machine learning; Smart cities

1. Introduction

1.1 Background and Context

Urban logistics has become an increasingly important part of smart city development, as a result of the rapid growth of e-commerce, urban densities, and the growing expectations and demands for quick and reliable deliveries. The last-mile segment - representing the last stage of distribution from the hubs to the customers - is generally accepted as the most operationally complex, cost intensive and delay-prone segment of the logistics chain with direct implications towards customer satisfaction, escalation of operational costs and challenges for environmental sustainability objectives in the cities [1].

Simultaneously, the proliferation of such technologies as Internet of Things (IoT), artificial intelligence (AI), and massive data analytics has transformed the logistics processes into a dynamic, data-driven environment of decision-making. IoT devices (e.g., GPS trackers, embedded sensors and connected platforms) provide spatiotemporal signals that continuously change in real-time,

which could be used to monitor and make predictions across urban delivery systems [2]. Also, the direction of logistics transformation includes a growing focus on the connectivity of advanced analytics and managerial and operational planning to deliver measurable improvements in service reliability and efficiency [3].

Amid such technological innovations, transportation delays continue to be a sustained and financially valuable concern, particularly in last-mile delivery, where demand fluctuation, congestion and miniaturisation of space interact and accumulate uncertainty. The main problem is that the traditional logistics analytics may not provide real time and actionable information because of the lack of unity in the data sources and the time-consuming nature of processing. This has induced the wider usage of automated data pipelines and predictive analytics to reduce the risk of delay and enhance operational responsiveness [4].

1.2 Problem Statement

While predictive analytics has gained momentum in research and practice in logistics, there are several limitations that inhibit the use of predictive analytics techniques for delaying mitigation in a real-life last-mile delivery system. First, a lot of current research focuses on conceptual models or controlled implementations instead of illustrating the concept of integrating predictive insights into operational decision processes to enhance on-ground delivery performance. Such a gap is supported by the fact that in modern logistics, the prediction accuracy is not the only aspect, but also management strategies that convert predictions into adaptive operational interventions in the dynamic urban environment [5].

Second, the difficulty in delay prediction is fundamentally context-specific, with results depending on the structure of cities, infrastructure and heterogeneity of operations. Evidence from the transportation delay prediction problem in freight systems has shown that good predictive performance requires the identification of influential operational and contextual features and their use in supporting decision-making under uncertainty [6]. However, the research on last mile delivery delay has still not fully converged on scalable predictive frameworks validated on large, diverse, real-world datasets.

Third, there is an applied research gap in linking "data-driven logistics optimization" as a strategic concept into practical and reproducible empirical pipelines that include feature engineering, model validation and decision-oriented evaluation. There is a broad recognition of the essential role of data-driven optimization of logistics, but empirical studies showing end-to-end predictive analytics processes on a large scale are relatively scarce, particularly in the last-mile context [7].

Fourth, despite the recent advent of sophisticated learning techniques like reinforcement learning and robust optimization in IoT-based scheduling and adaptive decision making, their operational feasibility and incorporation into last-mile delivery decision support are still non-trivial, especially when the data inputs are large scale and heterogeneous [8]. As a result, there is great value in getting a rigorous baseline on the basis of large-scale IoT spatiotemporal delivery data and data interpretable machine learning, before moving on to more complex adaptive optimization pipelines.

Finally, beyond the issue of prediction, the problem of delay is intrinsically connected with the operational control. Real-time IoT architectures for logistics performance monitoring and scheduling shows that service improvement requires modular designs that link data capture, performance evaluation and intervention mechanisms [9]. However, a number of last mile delays studies are still prediction-oriented, without scaling up of the results to explicit evaluation of strategies and practical improvement situations according to operationally relevant constraints.

Accordingly, this study addresses the following problem: How can IoT-enabled data-driven modeling be used to systematically and predictably reduce transportation delays in last-mile delivery operations, accounting for the spatial and temporal heterogeneity across cities?

1.3 Scope and Limitations

This research is specifically concerned with the context of last-mile delivery, using a large-scale operational data set with IoT-based spatiotemporal attributes. The scope focuses on (i) empirical description of the delay trends of the delivery, (ii) controlled machine learning-based delay prediction, (iii) controlled optimization of the models and (iv) simulated response of the delay minimization approaches based on risk-based ranking.

The research doesn't explicitly include real-time traffic feeds, weather or customer-specific behavioral features. Moreover, while the analysis supports the logic of evidence-based interventions, the analysis of their intervention is conducted within a controlled simulation framework, rather than being deployed in a live logistics platform. In addition, in the work, delay prediction and delay reduction is emphasized, not full optimization of the route, multi-stop sequencing or autonomous routing. These aspects are important in the broader aspect of smart logistics and are often placed as complementary paths that can help to improve the delivery performance, particularly if combined with real-time AI-based route adaptation and scheduling mechanisms [10].

1.4 Significance of the Study

This study offers contribution to smart logistics research in three main aspects. First, it researches an empirical foundation for understanding the last mile delivery delays by a data-driven approach based on large-scale IoT spatiotemporal delivery records. Second, it provides a repeatable machine learning pipeline, which is capable of supporting sound evaluation, optimization, and decision-oriented interpretation. Third, it connects predictive analytics with operational enhancement as it relates model outputs with quantified delay reduction opportunity through risk-based intervention logic. Taken together, these publications provide a solid evidence base on the use of smart logistics systems to minimize transportation delays and enhance the reliability of services with the help of IoT-enabled data-driven decision support.

1.5 Research Objectives

- The following research objectives are pursued by the study:
- To analyse not only the spatiotemporal patterns of last mile delivery delays using IoT-derived operational delivery data across different cities.
- To develop and test supervised machine learning models for predicting delivery delay using spatial, temporal and contextual features.
- To increase prediction model capacity by systematic hyperparameter tuning and threshold analysis in real world class imbalance.
- To quantify the operational value of predictive analytics using simulation style delay reduction assessment through risk-based intervention targeting.
- To compare delay behaviour and intervention effectiveness in different cities, focusing on spatial heterogeneity and context-specific performance.

2. Literature Review

Smart transportation and logistics systems are under rapid methodological evolution from model-based to data-based systems for forecasting and optimization. A key driving force is the development of intelligent transport systems (ITS) and the availability of high volume mobility data to allow predictive decision-making on congestion mitigation, routing and services reliability. The synthesis of the forecasting studies has indicated that transportation analytics has evolved in five major trends namely extension of classical models to reflect more real phenomena, the increased use of machine learning, use of the big data, convergence of methods towards hybrid models and the development of integrated methods that combine various paradigms to achieve better accuracy and relevancy to operations [11]. These developments offer a basis for last mile logistics research in which uncertainty is influenced by dynamic urban environments and spatiotemporal variability. Data-driven logistics also has spread beyond the thick metropolises to specialized operational areas. Evidence from maritime logistics shows how large-scale data and digital technologies (e.g., IoT, AI/ML, blockchain, and robotics) enable routes selection, risk estimation, environment impact reduction and arrival time prediction in environments characterized by high level of uncertainty and low level of infrastructure [12]. Though the geographic and operational contexts are different as compared to those of last-mile delivery, the results support a coherent principle, that is, data-driven modeling enhances the quality of decisions when it incorporates heterogeneous signals and facilitates proactive operations planning in the presence of uncertainty. This point of view is consistent with smart transportation research focused on why forecasting and optimization needs to consider complexity of the context instead than static assumptions.

Digital twin (DT) technology has come to play and become a key architecture for bridging the sensing, analytics, and operational control in the logistics system. A detailed discussion of the solutions to DT explains that logistics DTs offers real-time monitoring, prediction and simulation capability by connecting physical assets and processes to virtual replicas to support visibility and optimization in transportation and supply chain operations [13]. Notably, this literature recognizes the continued obstacles, which are data privacy, interoperability, complexity in keeping accurate digital representations, and which typically limit actual implementation. The combination of IoT gadgets with multi-source information is offered as the key to enhancing the fidelity of DT and facilitating sound predictive decision-making. In the context of smart city, wider surveys of emerging technologies give further weight to geo-information, data analytics, and machine learning increasingly becoming central to the performance of transport to support predictive and prescriptive decision systems to manage traffic and mobility services [14]. Together, DT architectures and smart transportation analytics provide motivation for research connecting spatiotemporal data to operational outcomes in terms of performance, such as delivery delay.

Real-world evidence from IoT-based operational environments suggests that tracking and dispatch systems can play an important role in reducing the outcomes of delays significantly if they link sensing to workflow decision-making. For example, empirical work in healthcare logistics shows that IoT-based real-time tracking can enhance the effectiveness of dispatch and reduce overtime delays compared with traditional methods and provides the argument that the mitigation of delays requires actionable operational intelligence and not retrospective reporting [15]. This is pertinent to last mile delivery because both fields deal with distributed agents (e.g. couriers or staff), time-sensitive tasks and performance metrics associated with quality of service. The implication of this is that the value of predictive analytics advances when combined with operational mechanisms that are capable to respond to risk signals in near real time.

Security, transparency and trust on logistics data ecosystems have now become relevant especially where high value goods, handoffs that involve more parties, or regulation demands are involved. An example of an IoT logistics system that uses blockchain is described, where it is noted that smart contracts, privacy sharing solutions, and IoT-equipped parcels can be used to facilitate traceability, identify violations, and deploy compensation logic automatically, without compromising data integrity [16]. Although these mechanisms are not themselves necessary to predict delays, they reinforce the larger smart logistics assumption: data pipelines which are used to give trustful predictions of decisions ought to be based on trustful, auditable data. In addition, the sustainability aspect is increasingly becoming part of the smart transport research. A literature-based synthesis places the emphasis on smart logistics not only on efficiency, but also a discussion on environmental performance, reduction of carbon emissions, safety, and the capability for real-time decision making and technology adoption is seen as a pathway to greener and more resilient logistics systems [17]. This sustainability framing complements predictive analytics research because reducing delay can translate to reduced consumption of fuel, congestion externalities, as well as improved service level stability.

The role of big data analytics as enabling infrastructure for "smartness" across transport and service systems is widely stressed. Conceptual discussion for smart transport and healthcare areas underscores the importance of correlational and predictive analytics in supporting institutional development, governance improvement and operational optimization, while acknowledging that actual changes are likely to occur gradually because of barriers to adoption, complexity of integration and organizational constraints [18]. This is of direct relevance to the study on last-mile delivery as the effectiveness of delay prediction using machine learning based methods depends not only on the model performance, but also on integration into operational systems and decision workflows. Therefore, empirical research that combines evaluation of predictive performance and implementable intervention logic is of particular value.

Within the domain of delay prediction literature, the rail and public transportation literature offer important methodological parallels for last-mile situations due to their common features of temporal uncertainty, propagation of disruption and the high operational cost of delay. A delay prediction framework that includes dispatching commands shows the improvement in the accuracy of the model when operational decision signals and propagation mechanisms are explicitly represented, and that model interpretability (e.g. explain what factors contribute to delay risk) enhances practical decision support [19]. This points to an important gap that is often found

in last-mile studies: predictive studies can be developed, but the connection between predictions and controllable operational levers are not always systematically developed. In the field of logistics operations management, decision-support is making more and more progress towards advanced forms of AI, such as generative AI frameworks that facilitate scenario simulation, adaptive learning, and multi-criteria reasoning. A supply chain framework based on generative AI demonstrates how the hierarchies of goals, feedback-based assessment, and reasoning based on scenarios can enhance the responsiveness and sustainability metrics, supporting the future shift towards a decision system that is not single-metric optimization [20]. For purposes of last-mile delay research, these developments mean that prediction can be an input into broader optimization and policy frameworks, and not be considered an end in itself.

The real-time route optimization research also highlights how heterogeneous data streams (GPS, traffic, and weather) can be optimized in terms of their operational potential by implementing automated ETL and streaming pipelines. The practical value of end-to-end data pipelines to optimize operations is evident in a real-time optimization strategy that proves the effectiveness of the approach in integrating data, feature engineering, and streaming architectures to achieve better transport time and cost results [21]. Complementary research work in the field of applied machine learning for logistics supports that route planning and resource allocation can be improved based on predictive modeling, while also defining Logistics 4.0 as a paradigm where artificial intelligence and machine learning are used to improve visibility, risk control and system efficiency without necessarily adding cost [22]. This is quite commensurate with delay prediction research: a good delay risk model can serve as an operational signal to make changes in route, resource assignment or dispatch priorities.

At conceptual level, the convergence of smart city development and smart logistics focuses on the fact that logistics is not an isolated system of operation but an essential part of city sustainability and liveability. A smart logistics perspective supports that technology and data-driven paradigms can provide urban systems with adaptability, flexibility and competitiveness as well as economic growth and quality of life [23]. However, throughout this literature, two continuous gaps can be seen. First, many of the contributions focus on architectures, frameworks or conceptual benefits without providing large-scale empirical validation using real-world last-mile delivery event data. Second, there is still the challenge of the operationalization gap: Predictive analytics has often only been assessed on the basis of accuracy, without that accuracy being judged in relation to the potential to achieve improvement through targeted interventions or decision rules. These gaps are very useful for inspiring empirical research that (i) uses large-scale spatiotemporal data from delivery experiments, (ii) compares multiple predictive methods in the face of class imbalance and city heterogeneity, and (iii) demonstrates how predictions can result in measurable delay reduction with the help of actionable intervention logic.

In summary, the above reviewed research shows great convergence on the idea of IoT-enabled data ecosystems, predictive analytics, digital twins, and real-time decision architectures as the backbone of smart transportation and logistics. Forecasting research highlights methodological maturity and the emergence of hybrid modelling and the logistics applications highlights the importance of heterogeneous data integration in risk reduction and optimisation of operations under varying circumstances. Digital twins and smart cities technology survey lay architectural foundation for real-time monitoring and predictive control. Emphasis on adoption barriers and integration complexity. There is empirical evidence that real time tracking and secure data infrastructures can promote useful service improvements and practical feasibility of data-driven intervention. Sustainability-centred work strengthens the argument for assessing the performance of logistics through an environmental and resilience lens as an incentive to reduce delays as efficiency and sustainability. Finally, delay prediction research shows that the ability to integrate operational signals and interpretability enhance decision relevance with advanced AI approach pointing towards scenario-based and feedback-driven optimisation as the future direction for smart logistics decision support. These combined insights are warrant a last-mile research agenda which is both predictive and intervention-oriented and based on a real-world data foundation and assessed in terms of performance and practical impact.

3. Methodology

3.1 Research Design

This study uses a quantitative and data-driven research design to examine the use of Internet of Things (IoT) enabled logistics data to analyse, predict and reduce transportation delays in last mile delivery operations. The research is based on a multi-stage analytical framework which includes the following steps: (i) large-scale spatiotemporal data preprocessing (ii) exploratory delay analysis (iii) supervised machine learning based delay prediction (iv) model optimisation using hyperparameter tuning (v) simulation based evaluation of delay reduction strategies.

The design is empirical and explanatory, with the goal not only of modelling the delays of delivery, but also of converting the predictive information into actionable operational interventions, which is in line with the practical objectives of smart logistics systems. All the analyses were done by means of reproducible computational workflow as implemented in Python ensuring transparency and methodological rigour.

3.2 Data Collection Methods

3.2.1 Dataset Description

The empirical study is based on the Last-mile Delivery Dataset (LaDe), a large-scale, real-world logistics dataset that contains operational data from urban courier delivery services [24]. The dataset contains delivery transactions from five major Chinese cities (Shanghai, Hangzhou, Chongqing, Jilin and Yantai), which can be seen as representing different urban settings with regard to population density, infrastructure complexity, and geographic scale.

LaDe was created to enable rigorous empirical research in last mile logistics and spatio-temporal data mining, by overcoming limitations of previous studies based on small scale, synthetic or proprietary datasets. Those characteristics of the dataset identified in its original publication include (i) scale of coverage, i.e., the presence of approximately 10.7 million delivery records by over 21,000 couriers over a 6-month experimental period; (ii) extensive information, i.e., package-level task data, event timestamps, and mobility-related signals; and (iii) cross-city diversity, i.e., heterogeneous delivery patterns of multiple urban settings with dissimilar infrastructural and demographic characteristics.

For this study, the records of delivery for each city were extracted and made available as structured CSV files. Each city dataset includes spatiotemporal attributes extracted from IoT systems, such as GPS coordinates (latitude and longitude), timestamps for accepting and delivering orders, courier identifier and area-of-interest (AOI) classification. Before analysis, schema consistency check developed ensured that all five city datasets had the same column structure, which guaranteed the integrity and comparability of data between regions. Collectively, these characteristics render LaDe to be a robust and appropriate empirical foundation for delay dynamics analysis and evaluation of machine learning based delay prediction models in large-scale urban logistics systems.

3.2.2 Temporal Data Construction

The original timestamps were registered in a truncated mode of the MM-DD HH:MM:SS mode lacking the year data. All timestamps had to be rebuilt in order to facilitate the accurate analysis in time; therefore, all timestamps were systematically constructed by appending a reference year (2022) to them and transformed into standardised date-time objects. Notably, time reconstruction was not done on date auxiliary, which avoided time leakage or misalignment. The move will be instrumental in valid delay computation and further feature engineering.

3.2.3 Delay Definition

Delivery delay was defined as the time period (minutes) between the order acceptance timestamp and delivery completion timestamp. Data quality and realism of operation was ensured by excluding deliveries with negative delays or more than 24-hour delays. A binary delay indicator was then built, in which a delivery was considered as delayed if it took more than 60 minutes, which is a threshold often used when assessing the performance of last mile logistics.

3.3 Population and Sampling

3.3.1 Target Population

The target population is urban last-mile delivery orders that are delivered by courier services in large metropolitan regions. Each observation is related to a single delivery event of a particular courier, location and time.

After preprocessing and cleaning, the final dataset contained 4,466,068 delivery records and, thus, this study is one of the largest empirical analyses of last-mile delivery delays to date.

3.3.2 Sampling Strategy

For predictive modeling, the entire dataset was divided into training data (80%) and testing data (20%) with stratified sampling, in order to maintain the natural distribution of the classes (delayed vs non-delayed delivery). This approach makes sure that in the model evaluation real-world operating conditions are considered with delayed deliveries accounting for about 72.3% of all observations.

In order to make computationally efficient hyperparameter tuning possible, while maintaining representativeness, an additional stratified subsample of 400,000 records was created, which was equally distributed between delay classes. This sampled dataset was used only for optimization and validation of the models whereas final model performance was tested using the entire test set.

3.4 Data Analysis Techniques

Exploratory Data Analysis (EDA) is a statistical technique that enables the summarization of relevant data within a dataset and the identification of patterns and trends in that data. Exploratory Data Analysis (EDA) Exploratory Data Analysis (EDA) is a statistical procedure that allows summarizing the interesting data in a dataset and detecting patterns and trends in the data.

Exploratory analysis was performed in order to characterize delay distributions and to find spatiotemporal distributions. Descriptive statistics and visualizations were used to analyze overall delay behavior, city-wise delays and trends over time between hours of the day and days of the week. These analyses provided empirical motivation for including time-based and spatial features in predictive models, and revealed large heterogeneity in the severity of delay in different cities.

3.4.2 Feature Engineering

To trade-off between predictive and interpretable accuracy, a parsimonious and yet informative feature set was built. Features included:

- Time aspects based on the timestamps of an IoT (hour of the day, day of the week),
- Spatial features according to GPS coordinates (latitude and longitude),
- Contextual features such as type of AOI,
- City identifiers, which are encoded numerically so as to reflect inter-city structural differences.

This feature design reflects realistic inputs that are available to smart logistics platforms that operate using IoT sensor data.

3.4.3 Machine Learning Models

Three supervised learning models were used for prediction of delivery delay:

1. Logistic Regression as a linear baseline model,
2. Random Forest Classifier, Non-linear interactions among spatial and temporal variables,
3. Light Gradient Boosting Machine (LightGBM), which is a state-of-the-art ensemble learning algorithm for the tabular data.

Model performance was assessed in terms of accuracy, precision, recall, F1-score, and ROC-AUC; the latter was given special emphasis because of class imbalance.

Hyperparameter optimization is a subfield of machine learning that aims to identify the optimal combination of hyperparameters to enhance the performance of a machine learning model, <[human]>3.4.4 Hyperparameter Optimization. Hyperparameter optimization is a branch of machine learning that involves finding the best combination of hyperparameters to improve the performance of machine learning model.

To improve the predictive performance and model Random Forest model were further optimized using RandomizedSearchCV with three-fold cross validation and ROC-AUC as the objective value. Hyperparameters of the model like number of trees, tree depth, minimum samples at leaf, feature

selection strategy were tuned. This process resulted in a much better model, which was better than both the untuned Random Forest and LightGBM baselines.

Decision threshold optimization was also performed to determine whether accuracy of classification could be increased beyond the default probability cutoff. Findings showed that, the default threshold offered nearly optimal behavior in real-world class distributions, which is good model calibration.

3.4.5 Delay Reduction Simulation

Beyond prediction, the study had a simulation-based intervention analysis. Predicted delay probabilities were used to identify the highest risk 20% of deliveries. An operational intervention, a hypothetical one, i.e., priority routing or early dispatch, was modeled by decreasing the delay probability of high-risk deliveries. The resulting difference in overall and city-specific delay rates were quantified, which allowed evaluation of possible delay reduction strategies.

3.5 Ethical Considerations

This research is solely based on secondary and anonymized operational data that does not contain any personally identifiable information. Courier identifiers were processed as numeric labels with no association to an actual person. No human subjects were directly involved and no behavioral tracking beyond aggregated records of operations was involved in the analysis.

All results are reported at aggregate level, so that results can't be used to assess or find couriers. The study follows principles of data minimization, fairness and transparency and the predictive models are designed to support system-level decision-making and not individual surveillance.

Overall, the methodology combines the large-scale data processing of IoT, the effective machine learning approaches, and the evaluation based on the simulations within a unified analytical system that offers the predictive performance and practical ideas to manage the delay in the smart logistics.

4. Results

4.1 Data Presentation and Descriptive Statistics

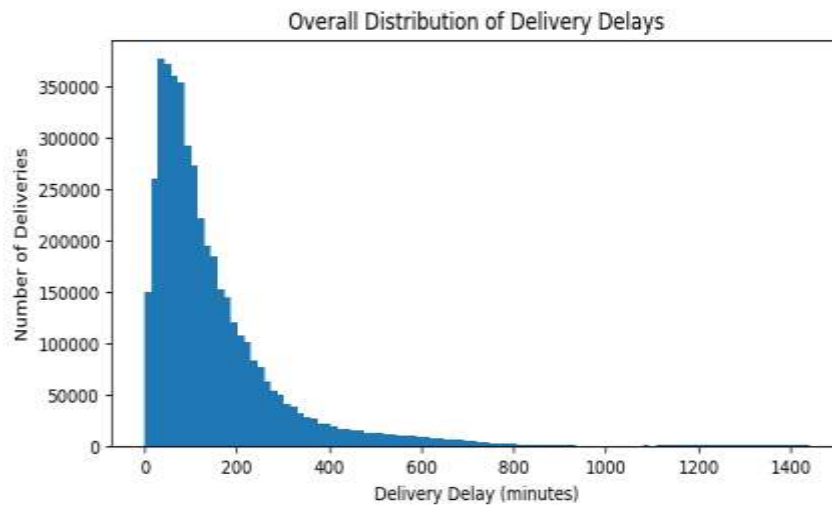
After the preprocessing and cleaning, the final analytical data set contained 4 466 068 last-mile delivery records in five cities: Shanghai, Hangzhou, Chongqing, Jilin and Yantai. Each record contained IoT derived spatiotemporal attributes, geographic coordinates, and operational context variables. Delivery delay was measured in minutes and further transformed, into a binary measure of whether or not a delivery was over a 60-minute threshold.

Table 1 includes the descriptive statistics of delays in delivery aggregated at the city level. Substantial heterogeneity is found across cities in terms of both the average and median delays.

Table 1. City-wise delivery delay statistics

City	Count	Mean Delay (min)	Median Delay (min)	Std. Dev.	Max Delay (min)
Shanghai	1,480,286	104.37	72	115.13	1440
Hangzhou	1,850,477	149.03	115	133.96	1440
Chongqing	900,117	195.45	146	185.20	1440
Jilin	31,407	203.06	175	139.21	1434
Yantai	203,781	241.91	205	167.86	1440

The overall delay distribution is strongly right skewed and with a long tail towards extreme delay values. Figure 1 shows the histogram of the delivery delays in all cities, while a big part of the deliveries are delivered in less than 200 minutes, there is a non-negligible fraction that suffer from severe delays that take more than 600 minutes.

Figure 1. Overall distribution of delivery delays (minutes)

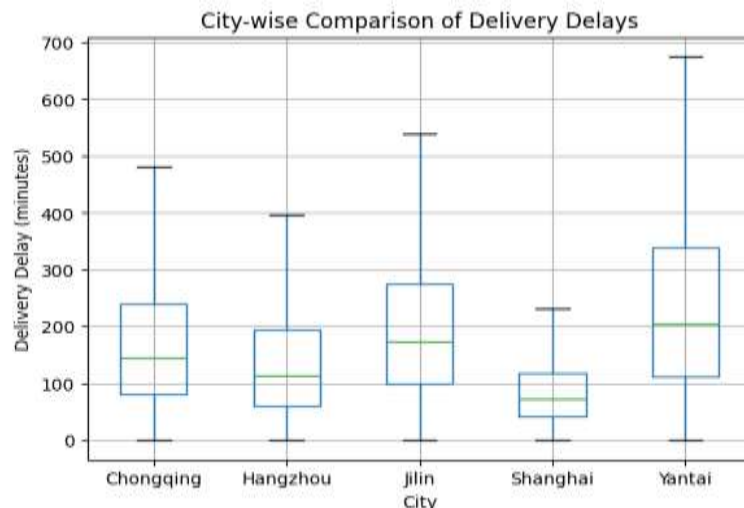
(Histogram of delivery_delay_min across all records)

Across the full dataset, 72.28% of deliveries were delayed beyond 60 minutes, underscoring the operational severity of delay-related inefficiencies in last-mile logistics.

4.2 Spatial and Temporal Delay Patterns

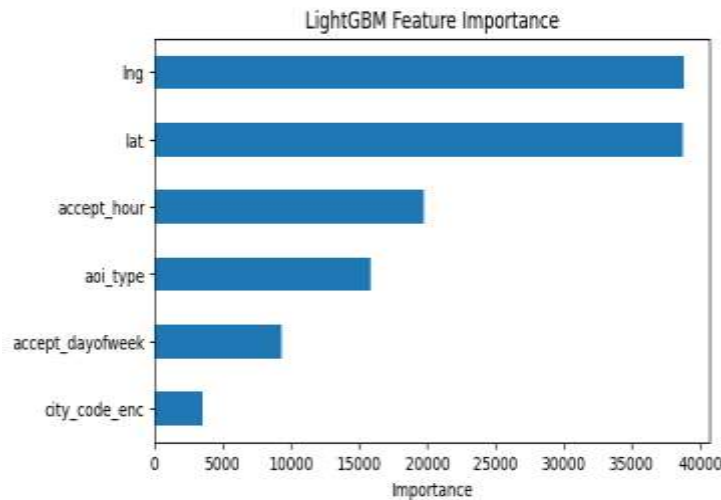
The city comparing patterns also show a systematic spatial variation in the delay behavior. Figure 2 shows boxplots of the delivery delays by city, leaving out extreme outliers for clarity's sake. Shanghai has the lowest median and interquartile range of delays in all cases and Yantai and Jilin have both higher central tendencies and greater variability.

Figure 2. City-wise comparison of delivery delays
(Boxplots of delivery_delay_min grouped by city_name)



Temporal analysis based on IoT timestamps gives more information. Figure 3 shows the feature importance scores from the LightGBM model adopted for the prediction of delivery delay. Geographic coordinates - longitude (lng) and latitude (lat) - emerge as the most influencing predictors indicating the importance of spatial factors and the role of route geography in determining the delivery delays. This suggests that distance, regional traffic patterns and urban structure have a large influence on delivery performance.

Figure 3. Feature importance ranking from the LightGBM model
(Bar charts of feature importance scores across spatial, temporal, and contextual variables)



Temporal variables also play a significant role in contributing meaning to the model. Order acceptance hour (accept_hour) exhibits considerable importance-there is an effect of time of day operation dynamics such as traffic congestion and courier availability. Similarly, the area of interest type (aoi_type) affects the delay results in terms of differences in delivery complexity in residential, commercial and mixed-use zones. In contrast, the day of the week (accept_dayofweek) and encoded city identifier (city_code_enc) indicate comparatively lower importance suggesting that variations in delivery delays are less, to say, sensitive to weekly patterns or city-level categorisation than precise location and time-based factors.

4.3 Predictive Model Performance

4.3.1 Baseline Models

In order to set predictive baselines, logistic regression and untuned random forest classifier were fitted on the entire data set based on an 80/20 stratified train-test split.

Logistic regression had a 73 percent accuracy with ROC-AUC of 0.731, which has a limited ability to detect the nonlinearity of the relationship between spatiotemporal delivery data. The untuned random forest showed a significant improvement in terms of performance as it achieved 77% accuracy and ROC-AUC of 0.807, confirming the value of nonlinear ensemble methods for delay prediction.

4.3.2 LightGBM Performance

A gradient boosting method based on LightGBM was then tested. The LightGBM model attained ROC-AUC of 0.813 and accuracy of 77% which is almost the same as the accuracy of the untuned random forest. The analysis of feature importance showed that geographic coordinates (latitude and longitude) and order acceptance hour were the strongest predictors, which underlines the predominance of spatial signals of the IoT and demand patterns in time.

4.3.3 Tuned Random Forest Model

The best predictive performance has been achieved with hyperparameter optimization of the random forest classifier. Using the ROC-AUC as optimization objective, the performance of the tuned model attained a ROC-AUC of 0.857 on the whole test dataset, which is a significant improvement from all baseline models.

The tuned random forest showed an overall accuracy of 80.03% at the default classification threshold (0.5), and strong recall on delayed deliveries (0.89). Threshold optimization gave only marginal improvements in accuracy indicating that the model was well-calibrated for real world class imbalance conditions.

Table 2. Comparison of predictive model performance

Model	Accuracy	ROC-AUC
Logistic Regression	0.73	0.731
Random Forest	0.77	0.807
LightGBM	0.77	0.813
Tuned Random Forest	0.80	0.857

These results show that model optimization provides meaningful improvements in performance, especially in discriminative ability, which is important in operational decision-making.

4.4 Feature Importance and IoT Insights

Feature importance analysis thus always pointed to the dominance of spatial and temporal IoT features. Latitude and longitude were the most important predictors in all ensemble models, which indicated the influence of urban geography, road networks and spatial congestion. Order acceptance hour ranked next in importance which confirmed that temporal demand cycles significantly influence delivery outcomes.

City-level encoding and day-of-week effects, in their turn, had a relatively small level of importance, which implies that finer spatial-temporal features are highly predictive compared to more coarse ones of a categorical nature.

4.5 Delay Reduction Simulation and Operational Impact

Beyond prediction, the study assessed the usefulness of predictive information through a simulation-based analysis of the reduction of delays. Predicted delay probabilities from the tuned random forest were used to identify the highest risk top 20% deliveries. Although hypothetical intervention which is represented by proactive operational activities like priority routing or resource reallocation was simulated by lowering the probability of delay of these high-risk deliveries by 30%.

At the aggregate level, this intervention achieved a reduction in the overall delay rate from 72.28% to 66.42%, which is a relative improvement of 8.10%.

City-specific results, shown in Table 3, show that there is a great deal of variation in the amount of delay reduction that can be achieved.

Table 3. City-specific delay reduction results

City	Baseline Delay Rate	Post-Intervention Delay Rate	Improvement (%)
Yantai	0.903	0.703	22.17
Jilin	0.880	0.706	19.78
Hangzhou	0.750	0.674	10.07
Chongqing	0.836	0.772	7.63
Shanghai	0.593	0.580	2.09

The baseline delay rates of cities are larger, showing that relative improvements are larger in cities with higher delay rates, meaning that predictive interventions work especially well in more operationally constrained or congested settings.

4.6 Summary of Key Findings

Overall, the results demonstrate that:

- Last-mile delivery delays are common, with more than 70% of deliveries subject to them.
- There is large heterogeneity in space and time between cities and time periods.
- Ensemble machine learning models are significantly better than linear baselines.
- Hyperparameter tuning yields a robust predictive model with 80% accuracy and 0.857 ROC-AUC.

- Predictive insights can be translated into meaningful delays reduction with improvements of more than 20% in cities with high delays. These results confirm the efficacy of a smart logistics framework that incorporates IoT data, machine learning and operational decision support for transportation delays mitigation.

5. Discussion

5.1 Interpretation of Key Results

The empirical results of this study are strong evidence that big data in logistics, comprising IoT-enabled logistics data, can be effectively used to both forecast and minimize last-mile delivery delays in complex urban settings. The descriptive analysis showed that over 70% of deliveries across the five cities took more than 60 minutes confirming that delay is not an exceptional event, but rather a structural urban last-mile logistics characteristic. This high prevalence makes predictive, as opposed to reactive, operational strategies necessary.

One of the most salient results is that of the overwhelming control of spatial features, especially latitude and longitude, on delay prediction performance. In both Random Forest and LightGBM models, geographical coordinates were the most important predictors. This suggests that delivery delays are strongly rooted in the urban spatial structure such as road network density, land use distribution, and local congestion situation. These findings support the belief that the performance in the last mile is fundamentally influenced by micro-geographic constraints rather than just organizational or temporal constraints, which is also the case in smart transportation analytics research [14], [11].

Temporal features, particularly the hour of acceptance, were also highly influential, underscoring the influence of time of day operation dynamics such as peak traffic congestion, courier workload distribution and service demand surges. In contrast, day of week effects were less important, indicating that daily operational rhythms are more important than weekly rhythms in last mile delivery settings. This is consistent with the results of studies of delay prediction for freight and transportation systems using data, where short-term temporal dynamics frequently dominate predictive performance [6], [19].

5.2 Comparison with Existing Literature

The predictive performance achieved in this study compares favorably and in some cases surpasses results of previous research on prediction of logistics and transportation delays. The Tuned Random Forest model was able to reach ROC-AUC of 0.857 and approximately 80% accuracy on a full scale test set of over 890,000 observations. This level of performance is significant considering the extreme class imbalance that exist in the last-mile delivery data from the real world, and the lack of external inputs, in the form of real-time traffic or weather feeds.

Previous works have shown good prediction effects utilizing advanced ensemble or boosting methods in structured transportation fields like rail systems [6], [19]. However, these studies tend to be based on more narrow operational contexts or in more controlled infrastructure environments. In contrast, last mile delivery is associated with greater uncertainty, spatial heterogeneity and decentralised execution. The results of this study, therefore, build on previous research by showing that interpretable ensemble models, when trained on large-scale IoT delivery data, can be used to achieve robust predictive accuracy even in highly complex urban logistics environments. Furthermore, in times when recent work has focused more on advanced AI paradigms, such as reinforcement learning and optimization using digital twin [8], [5], the current work demonstrates that good baseline improvements can be made in using conventional supervised learning, as long as the data scale, feature engineering, and model optimization is addressed rigorously. This provides support for arguments in the data driven logistics literature that there are often practical gains to be made through systematic application of established methods rather than early or unnecessary dependence on highly complex architectures [7], [3].

5.3 Implications for Smart Logistics and Urban Operations

A key aspect of the contribution of this work will be to show that predictive analytics can be operationalized to achieve measurable delay reduction, as opposed to being an isolated forecasting exercise. The analysis of the intervention based on the simulation model indicated that saving the

top 20% highest-risk deliveries could save about 8% of the total delay rates; the improvement is more than 20% in cities such as Yantai and Jilin. There are significant practical implications of these results.

First, they validate the principle of risk-based prioritization, which, with the help of machine learning predictions, can dramatically improve the reliability of service without having to completely rethink the entire system. Such interventions could be in the form of priority dispatching, dynamic reallocation of resources or early rerouting, all of which are compatible with already existing IoT-enabled logistics platforms [9], [10].

Second, the city-specific variation in intervention effectiveness identified suggests the need for localized strategies of logistics. Cities that had higher baseline delay rates had a much higher amount of potential improvement, which could suggest that predictive interventions have higher marginal benefit in highly constrained or congested environments. This finding is consistent with the smart city logistics research that stresses the inductility of the one-size-fits-all optimization strategies in heterogeneous urban environments [2], [23]. Third, from a sustainability point of view there are implications of delay reduction beyond service performance. Reduced delays can result in reduced fuel consumption, failed delivery attempts and congestion externalities. These outcomes support more general smart city goals with respect to environmental sustainability and resource efficiency [1], [17].

5.4 Methodological and Practical Limitations

Despite its contributions, this study has a number of limitations that should be taken into account. First, while the dataset is extremely large and diverse, it does not contain actual real-time traffic conditions or weather information or customer-level attributes, all of which may have further effects on delivery performance. Incorporation of such features may potentially increase predictive accuracy and effectiveness of intervention.

Second, the analysis on delay reduction is simulation based, i.e., hypothetical operational intervention, not direct deployment in a live logistics system. While simulation has been established in logistics research as a decision support evaluation tool [5], the implementation of simulation in a real-life context could reveal further constraints related to system integration, workforce behaviour, and organizational policy.

Third, the predictive models work on the classification of delay, in binary format as opposed to continuous duration forecasting. Although this option fits the operational service-level thresholds, the future research might consider the use of hybrid solutions, which integrate classification and regression to enable even more detailed decision-making.

Finally, while ensemble models were able to provide high-level performance, interpretability is an outstanding challenge to operational adoption. While feature importance gives valuable insight, it would be great for the explanation to be further integrated into the explainability techniques (for example, SHAP-based explanation) to strengthen trust and usability in the real world [19].

5.5 Directions for Future Research

The results of this study create several exciting areas for future study. First of all, if real time traffic, weather and event-based data streams could be integrated into the process, the predictive robustness could be improved and near-real time intervention could be delivered. Second, the customized models of the framework in cities or the adaptation of the framework to specific city contexts may allow for further enhancements in the performance of the heterogeneous urban environment.

Third, future research could incorporate predictive models in the context of digital twin or reinforcement learning, enabling a dynamic feedback between prediction, intervention, and system response, as called for in recent research on smart logistics [8], [13]. Finally, the longitudinal evaluation of deployed interventions would allow to assess the adaptation of behavior, resilience of the system, and long-term sustainability effects.

In summary, this discussion has shown that the results of this study are not only robust statistically but are theoretically based as well as operationally meaningful. By bridging massive scale IoT data, machine learning, and decision-oriented evaluation, this research improves the empirical basis of predictive and proactive smart logistics systems that can reduce delays in last-mile delivery in complex urban settings.

6. Conclusion

This study aimed to understand how leveraging large-scale and IoT-enabled logistics data and the use of machine learning can be used to predict and reduce last-mile delivery delays in urban environments. Using an extensive collection of real-time data of millions of delivery occurrences in multiple cities, the research used clarified that delivery delays are a structural and spatially embedded phenomenon and not isolated operational failures. The empirical results indicated that machine learning models based on ensemble approach were able to fulfil the robust predictive performance issues under severe class imbalance, especially tuned Random Forest and gradient boosting machine learning models, of which have strong discriminative power and the stable generalization at scale. Spatial features and short-term temporal factors became the most influential factors in delivery delay risk, making localised urban context a key factor in last mile logistics performance.

Beyond prediction, the study provided evidence that predictive analytics can be translated into the operational value. Simulation-based analysis of intervention suggested that risk-targeted interventions can produce significant improvements in overall delay rates, with the greatest benefits conferred in high congestion cities. These findings verify that predictive models, if included in a framework of decision support workflow, can be used to support proactive rather than only retrospective performance assessment logistics management.

From the viewpoint of practical implementation, the presented findings demonstrate the feasibility of being able to implement data-driven delay management methods with existing IoT infrastructure and traditional machine learning methods. Logistics operators can use delay risk scores to prioritize resources, make dispatch decisions and maximize the reliability of service without having to redesign the entire system radically. For urban planners and policymakers, the results give new legitimacy to data-centric logistics strategies as part of the advancement towards efficiency, sustainability and the quality of service in smart-cities.

This study suggests that logistics organizations pursue city-specific and risk-based operational policies backed by ongoing data collection and updating of the models. This research should investigate combining real-time traffic and real-time environmental information, acceptable frameworks for adaptive and feedback based learning, and on-field pilot tests with real-world applications to understand long-term behavioral and system-level effects. By promoting predictive, proactive and context-aware logistics management, further improvements of the resilience and sustainability of urban delivery systems can be expected in future work.

REFERENCES

1. B. M. Mohsen, "AI-driven optimization of urban logistics in smart cities: Integrating autonomous vehicles and IoT for efficient delivery systems," *Sustainability*, vol. 16, no. 24, Art. no. 11265, 2024. <https://doi.org/10.3390/su162411265>
2. H. Fatorachian, H. Kazemi, and K. Pawar, "Enhancing smart city logistics through IoT-enabled predictive analytics: A digital twin and cybernetic feedback approach," *Smart Cities*, vol. 8, no. 2, Art. no. 56, 2025. <https://doi.org/10.3390/smartcities8020056>
3. H. Tran-Dang, J. W. Kim, J. M. Lee, and D. S. Kim, "Shaping the future of logistics: Data-driven technology approaches and strategic management," *IETE Technical Review*, vol. 42, no. 1, pp. 44–79, 2025. <https://doi.org/10.1080/02564602.2024.2445513>
4. [4] S. Yerra, "Optimizing supply chain efficiency using AI-driven predictive analytics in logistics," *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.*, vol. 11, no. 2, pp. 1212–1220, 2025. <https://doi.org/10.32628/CSEIT25112475>
5. J. Li and J. Wang, "Digital twin-driven management strategies for logistics transportation systems," *Scientific Reports*, vol. 15, no. 1, Art. no. 12186, 2025. <https://doi.org/10.1038/s41598-025-96641-z>
6. J. Pineda-Jaramillo, F. Bigi, T. Bosi, F. Viti, and A. D'Ariano, "Short-term arrival delay time prediction in freight rail operations using data-driven models," *IEEE Access*, vol. 11, pp. 46966–46978, 2023. <https://doi.org/10.1109/ACCESS.2023.3275022>
7. J. Kanyepe, "Data-driven logistics optimization: Powering the digital supply chain," in *Impacts of Technology on Operations Management: Adoption, Adaptation, and Optimization*, Hershey, PA, USA: IGI Global, 2025, pp. 63–96. <https://doi.org/10.4018/979-8-3693-6205-1.ch003>

8. Y. Lu, "A multimodal deep reinforcement learning approach for IoT-driven adaptive scheduling and robustness optimization in global logistics networks," *Scientific Reports*, vol. 15, no. 1, Art. no. 25195, 2025. <https://doi.org/10.1038/s41598-025-10512-1>
9. Â. F. Brochado, E. M. Rocha, and D. Costa, "A modular IoT-based architecture for logistics service performance assessment and real-time scheduling towards a synchromodal transport system," *Sustainability*, vol. 16, no. 2, Art. no. 742, 2024. <https://doi.org/10.3390/su16020742>
10. K. Hussain, "Revolutionizing route optimization systems with artificial intelligence for a smarter, sustainable logistics ecosystem," *Int. J. Comput. Sci. Mobile Comput.*, vol. 14, no. 2, 2025. <https://doi.org/10.47760/ijcsmc.2025.v14i02.008>
11. S. Mrad and R. Mrahi, "An overview of model-driven and data-driven forecasting methods for smart transportation," in *Data Analytics and Computational Intelligence*, Cham, Switzerland: Springer, 2023, pp. 159–183. https://doi.org/10.1007/978-3-031-38325-0_8
12. K. Paardenkooper, "The role of data-driven logistics in Arctic shipping," in *Arctic Maritime Logistics*, Cham, Switzerland: Springer, 2022, pp. 173–191. https://doi.org/10.1007/978-3-030-92291-7_10
13. T. K. Vashishth et al., "Digital twin solutions for smart logistics and transportation," in *Digital Twins for Smart Cities and Villages*, Oxford, U.K.: Elsevier, 2025, pp. 353–376. <https://doi.org/10.1016/B978-0-443-28884-5.00016-6>
14. K. L.-M. Ang et al., "Emerging technologies for smart cities' transportation: Geo-information, data analytics and machine learning approaches," *ISPRS Int. J. Geo-Inf.*, vol. 11, no. 2, Art. no. 85, 2022. <https://doi.org/10.3390/ijgi11020085>
15. S. W. Huang et al., "Enhancing hospital efficiency and patient care: Real-time tracking and data-driven dispatch in patient transport," *Sensors*, vol. 24, no. 12, Art. no. 4020, 2024. <https://doi.org/10.3390/s24124020>
16. M. Balfaqih et al., "A blockchain-enabled IoT logistics system for efficient tracking and management of high-price shipments," *Sustainability*, vol. 15, no. 18, Art. no. 13971, 2023. <https://doi.org/10.3390/su151813971>
17. M. Shamsuddoha, M. A. Kashem, and T. Nasir, "Smart transportation logistics: Achieving supply chain efficiency with green initiatives," in *Data Analytics for Supply Chain Networks*, Cham, Switzerland: Springer, 2023, pp. 243–258. https://doi.org/10.1007/978-3-031-29823-3_10
18. S. Pourroostaei Ardakani and A. Cheshmehzangi, "Big data analytics and the future of smart transport and healthcare systems," in *Big Data Analytics for Smart Transport and Healthcare Systems*, Singapore: Springer, 2023, pp. 175–184. https://doi.org/10.1007/978-981-99-6620-2_10
19. T. Gao, J. Chen, and H. Xu, "Data-driven train delay prediction incorporating dispatching commands: An XGBoost-metaheuristic framework," *IET Intell. Transp. Syst.*, vol. 18, no. 10, pp. 1777–1796, 2024. <https://doi.org/10.1049/itr2.12461>
20. K. Y. Lin, "Generative artificial intelligence-driven sustainable supply chain management: A UNISONE framework for smart logistics and predictive analytics under Industry 5.0," *Int. J. Logistics Res. Appl.*, pp. 1–32, 2025. <https://doi.org/10.1080/13675567.2025.2540855>
21. J. Wang, "Real-time optimization algorithm for intelligent logistics transportation routes based on big data analysis," in *Proc. Int. Conf. Machine Intelligence and Digital Applications*, 2024, pp. 193–199. <https://doi.org/10.1145/3662739.3672306>
22. A. Reyana and S. Kautish, "Machine learning techniques for route optimization and logistics management," in *Computational Intelligence Techniques for Sustainable Supply Chain Management*, Oxford, U.K.: Academic Press, 2024, pp. 197–224. <https://doi.org/10.1016/B978-0-443-18464-2.00010-8>
23. M. Azari et al., "From city logistics to a merger of smart cities and smart logistics," in *Proc. 14th Int. Colloq. Logistics Supply Chain Management (LOGISTIQUA)*, 2022, pp. 1–5. <https://doi.org/10.1109/LOGISTIQUA55056.2022.9938020>
24. L. Wu et al., "LaDe: The first comprehensive last-mile express dataset from industry," in *Proc. 30th ACM SIGKDD Conf. Knowledge Discovery and Data Mining*, 2024, pp. 5991–6002. <https://doi.org/10.1145/3637528.3671548>