Proceedings of the 3rd International Conference on Mechatronics and Smart Systems
DOI: 10.54254/2755-2721/135/2025.21217

Based on Bayesian Regression and Logistic Regression Model Forecast Logistics Transportation Delay

Qixuan Qiu[1,a,*]

[1]*Chang'an Dublin International College of Transportation at Chang'an University, Chang'an University, Xi'an, China*
a. qixuan.qiu@ucdconnect.ie
**corresponding author*

Abstract: In the logistics supply chain transportation, it is inevitable to encounter some unpredictable risks, these risks will greatly affect the accuracy and timeliness of logistics transportation, in order to explore the problem of logistics transport delays, this article uses Bayesian regression and logistic regression to establish a model, by analyzing the relationship between the journey risk and the logistics transport delays, Bayesian regression through Mean Squared Error(MSE) and R^2 carry out a test, logistic regression through Accuracy, Receiver Operating Characteristic Curve (ROC Curve) and Precision-Recall Curve carry out a test. This article mainly analyses some external factors in travel, including the degree of traffic congestion, the degree of weather, the degree of driver behavior, the degree of driver fatigue and other data to determine the degree of risk of logistics transportation, and then according to the risk level and the probability of delay analysis and prediction, and then can be adjusted to reduce the risk of the external factors to reduce the probability of delay, and help logistics and transport system in Southern California, USA even the global logistics industry to build a more efficient logistics supply chain framework.

Keywords: Logistics, Risk Level, Delay Probability.

1. Introduction

Due to the current competition in the global economy, the logistics supply chain industry is booming and requires logistics supply chains as a basic need, both from an individual and group perspective. In response to the global economic situation, more companies are turning to logistics standardization as a solution to complete the logistics supply chain, according to research by Bhardwaj et al., established a complete global logistics standardization framework to help standardize positioning [1]. The logistics supply chain is becoming a well-established industry system that requires a lot of interdisciplinary knowledge to respond to the changing needs and requirements of the industry. According to Baker et al., industry has decisively evolved into a complex and interdisciplinary field, demanding diverse expertise to effectively optimize the planning and execution of urban freight logistics and transport [2]. Overall, logistics is not only an important driver of economic development but also a fundamental indicator of economic growth and international trade. As the logistics supply chain industry continues to grow and develop, the logistics supply chain, from the global economy down to daily lives, has taken up a very important place in the lives [3]. The use of a Transportation Management System (TMS) is an excellent way to help logistics and transportation systems plan and

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/135/2025.21217

optimize ground transportation routes and track vehicles in real time, reducing logistics delays and risk levels to a large extent [4]. Select the best relay Internet of Devices (IoD) for data routing by using Q-Learning, which adjusts the Q-matrix of the network to identify the best relay IoD at discrete moments to get the maximum between specified pairs of device gateway total reward value, the method promotes Internet of Things (IoT) application efficiency and IoD cost control [5]. This article establishes Bayesian regression models and logistic regression models to explore the relationship between risk level and delay probability, and compares the performance of risk level weighted and unweighted, ultimately aiming to help the logistics supply chain system to better cope with future risks caused by external factors affecting the transport problem. The reason for choosing these two models in this paper is that as Ricard et al. have shown in their research, both Bayesian regression and logistic regression have a very high accuracy in exploring logistics delays, and in the research of Hudnurkar et al., it can provide a good case study as well as an idea for building a logistic regression model [6, 7].

The rest of the paper is organized as follows. Section 2 of this article analyses the data on external factors in the data set and the impact of these external data on the risk level, Section 3 comparative analysis of the gap between weighted and unweighted risk levels, and Section 4 is a conclusion for the two models.

2. Data and Method

2.1. Data Collection and Description

This data is 2024 U.S. Southern California logistics and transport data from Kaggle. When determining the logistics risk level, the first thing to determine is whether external factors can influence it and how much each external factor influences it. According to the research of Peker et al., this article can use the Multiple Criteria Decision Making (MCDM) method to help us weigh the external factors in a hierarchy [8].

According to the MCDM model that Peker et al. made in their research, it can be determined that the impact of different criteria on the risk level is also different, Logistics risk sources 0.428, Logistics risk errors 0.337, Logistics risk characteristics 0.235, according to Figure 1, it can be determine which rank the different external factors in the sub-data are assigned and their weights.

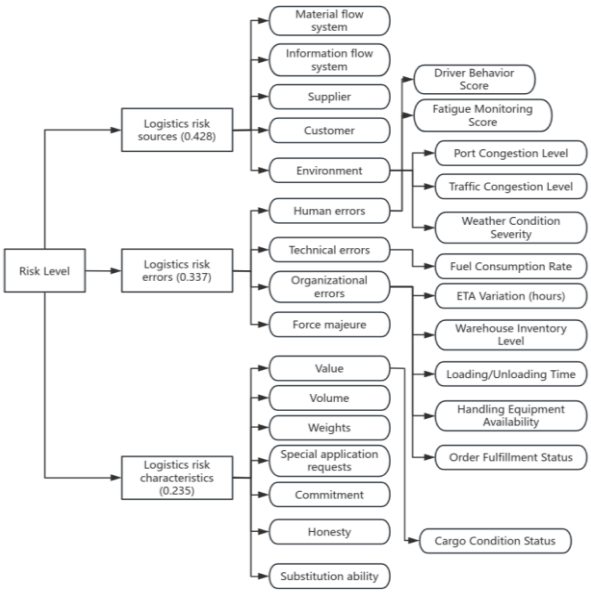

Figure 1: Flow Chart of The External Factors Weight. (Picture credit: Original).

DOI: 10.54254/2755-2721/135/2025.21217

Each external factor contributes to the risk level to a different degree. From this, it can be simply calculated the final logistics transport risk level based on the weights of these external factors, and then it can be established a transport delay model.

2.2. Data Analysis

From Figure 2, it can be concluded that Driver Behavioral Scores and Fatigue Monitoring Scores are inversely proportional, i.e., as the value of Fatigue Monitoring Scores increases, the value of Driver Behavioral Scores decreases.

As the driver proceeds with the driving task, subjective and passive fatigue will gradually increase, and the driver's horizontal gaze dispersion will increase, i.e., the vision will become more and more blurred, and the driver will not be able to concentrate as much as at the beginning of the driving task, which will in turn affect the Driver Behavioral Scores [9].

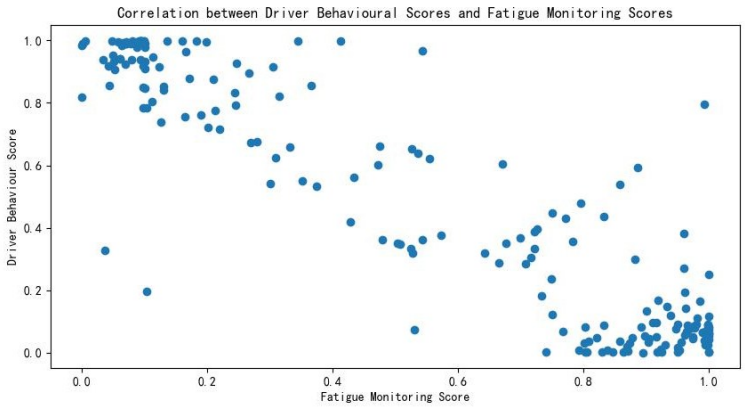

Figure 2: Scatterplot of Correlation between Driver Behaviour Scores and Fatigue Monitoring Scores. (Picture credit: Original).

From Figure 3, it can be concluded that when the fuel consumption is 5, the arrival time change value is distributed, it can be inferred that when the fuel consumption is 5 the fuel consumption in the logistics and transportation is about 5, and for other points, the greater the fuel consumption, the higher the arrival time change value so that it can be inferred that the driver in order to arrive on time, for the vehicle to increase the accelerator so that the fuel consumption of the vehicle increased. From this, it can be deduced that based on the fuel consumption of the vehicle it is possible to derive whether there will be a change in the arrival time of the vehicle transportation.

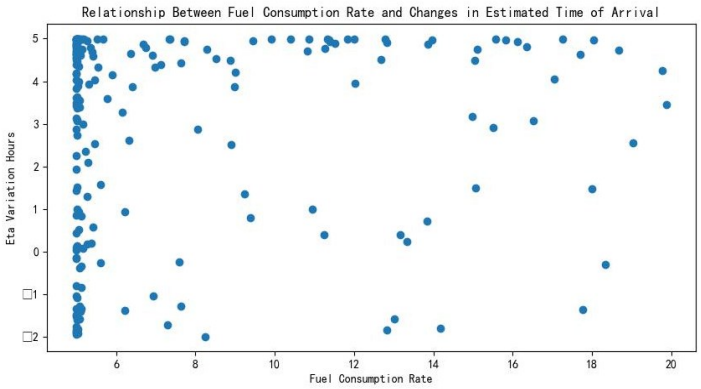

Figure 3: Scatterplot of Relationship Between Fuel Consumption Rate and Changes in Estimated Time of Arrival. (Picture credit: Original).

DOI: 10.54254/2755-2721/135/2025.21217

From Figure 4, it can be deduced that there are two peaks of traffic congestion at 0-2.5 and 7.5-10, with 0-2.5 representing smooth road traffic without congestion, and 7.5-10 representing slow road traffic with congestion, however, it can also be observed from the plot that 7.5-10 is the highest peak, which means that traffic congestion accounts for a large proportion of traffic congestion in these logistics and transport routes. Traffic jams are more important in these logistics routes.

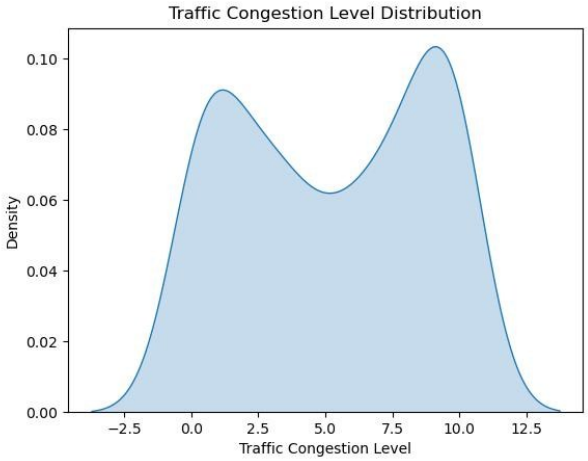

Figure 4: Kernel Density Estimation (KDM) Plot for Traffic Congestion Level. (Picture credit: Original).

From Figure 5, it can be deduced that similar to traffic congestion level, weather condition severity again has two peaks, 0-0.25 and 0.75-1. However, unlike traffic congestion level, weather condition severity is more gentle, i.e., the density of other non-peaks is greater, which can be inferred that the weather conditions are more balanced, and this data is consistent with the actual situation.

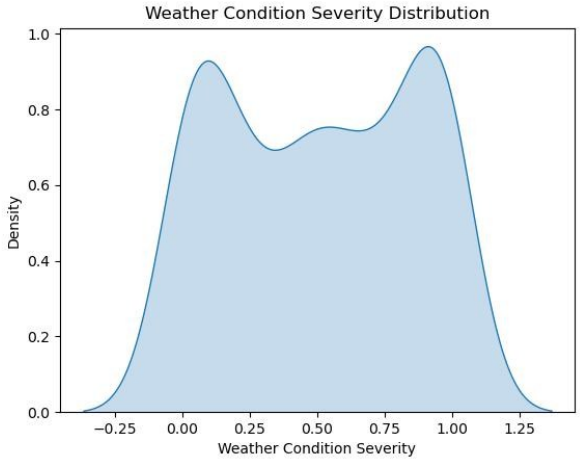

Figure 5: KDM Plot for Weather Condition Severity. (Picture credit: Original).

From Figure 6, it can be deduced that the maximum loading/unloading time is nearly 5 hours, while the minimum loading/unloading time is less than one hour and the median value is 2 hours, which is caused by the combination of loading/unloading equipment at the loading/unloading place and the port blockage, so the driver should be prepared for the maximum loading/unloading waiting time for this situation.

DOI: 10.54254/2755-2721/135/2025.21217

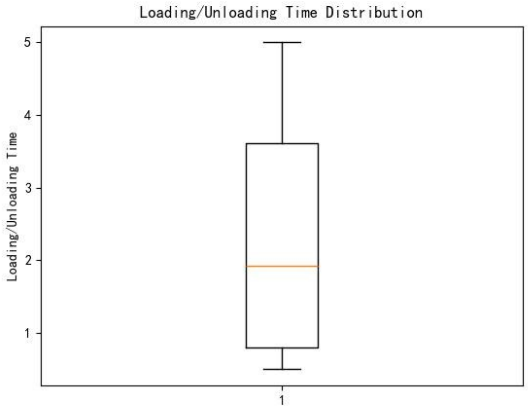

Figure 6: Box Plot for Loading/Unloading Time. (Picture credit: Original).

2.3. Model Principles

In this article, two different models are used are Bayesian regression model and the logistic regression model. The following is a description of the principles of these two models.

2.3.1. Bayesian Regression Model

Bayesian regression is a type of regression model that takes its basic form through Equation (1) and Equation (2) demonstrates that the prior distribution of the parameters is introduced through Equation (3) and combined with the likelihood function of the data in Equation (4), and combined with the idea of Bayesian statistics, the posterior distribution in Equation (5) is obtained.

$$y_i = x_i^T w + \epsilon \tag{1}$$

$$y = Xw + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma^2 I) \tag{2}$$

$$\mathrm{p(w)} = \mathcal{N}(\mathrm{w} \mid 0, \tau^2 \mathrm{I}) \tag{3}$$

$$p(y \mid X, w, \sigma^2) = \mathcal{N}(y \mid Xw, \sigma^2 I) \tag{4}$$

$$p(w \mid y, X, \sigma^2, \tau^2) = \frac{p(y \mid X, w, \sigma^2) p(w \mid \tau^2)}{p(y \mid X, \sigma^2, \tau^2)} \tag{5}$$

$$SSE = \sum_{i=1}^{n} w_i \, (y_i - \hat{y}_i)^2 \tag{6}$$

$$MSE = SSE \; / \; n = \frac{1}{n} \sum_{i=1}^{n} w_i \, (y_i - \hat{y}_i)^2 \tag{7}$$

$$SST = \sum_{i=1}^{n} w_i \, (y_i - \bar{y}_i)^2 \tag{8}$$

$$R^2 = 1 - \frac{SSE}{SST} \tag{9}$$

In Equation (1) – Equation (5), x is the input data, y is the observed data, w is the vector of weight parameters of the model, $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$ is the Gaussian noise, τ^2 denotes the variance of the prior, reflecting the a priori uncertainty about the w. $p(y|X, \sigma^2, \tau^2)$ is the normalization constant, also known as the marginal likelihood or evidence. μ is the mean value. Equation (6)- Equation (9) represents the formula for Bayesian regression model test numbers SSE and R^2.

DOI: 10.54254/2755-2721/135/2025.21217

2.3.2. Logistic Regression Model

The logistic regression model is a model for solving classification tasks and is used for binary classification problems, according to the research of Bouhadi et al. for the probabilistic dichotomy of the data in this dataset [10]. The data with a probability of delay greater than or equal to 10% is considered as will be delayed (1), and vice versa as won't be delayed (0), and the results are calculated based on the Equation (10) – Equation (11).

$$P(y=1|x)=\frac{1}{1+e^{-(\beta_0+x^T\beta)}} \tag{10}$$

$$\ln\left(\frac{P(y=1|x)}{1-P(y=1|x)}\right)=x^T\beta \tag{11}$$

$$\text{Accuracy}=\frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \tag{12}$$

$$\text{TPR}=\frac{\text{TP}}{\text{TP+FN}} \tag{13}$$

$$\text{FPR}=\frac{\text{FP}}{\text{FP+TN}} \tag{14}$$

$$\text{Precision}=\frac{\text{TP}}{\text{TP+FP}} \tag{15}$$

$$\text{Recall}=\frac{\text{TP}}{\text{TP+FN}} \tag{16}$$

In Equation (10) – Equation (11), x is the input data, y is the observed data, $x^T\beta$ is a linear combination of the inputs, $\beta=(\beta_0,\beta_1,\ldots,\beta_d)^T$ is the parameter vector of the model, TP is true positive, TN is true negative, FP is false positive, FN is false negative. Equation (12) – Equation (16) represents the test ROC, Precision-Recall and Accuracy in the logistic regression model.

3. Results and Discussion

3.1. Experiment Configuration

The Bayesian regression model was used as an experimental configuration via Jupyter Notebook and scikit-learn with a test set of 25%, the optimizer was Stochastic Gradient Descent (SGD) with a learning rate of 0.001 and the loss function was the MSE loss function. The logistic regression model was configured as an experiment via Jupyter Notebook and scikit-learn with a test set of 25%, the optimizer was SGD, the learning rate was tuned via the C parameter, and the loss function was a log loss function.

3.2. Bayesian Regression Model

Assuming that the target prediction value is the delay probability, with risk class as the feature input, taking this into a Bayesian regression model, it can be concluded that the risk rating after weighting the external factors is linearly related to the probability of delay, and tested the model by the R^2 and MSE and concluded as follows. MSE equals 0.011351 and R^2equals to 0.869527, the model visualization is shown in Figure 7.

DOI: 10.54254/2755-2721/135/2025.21217

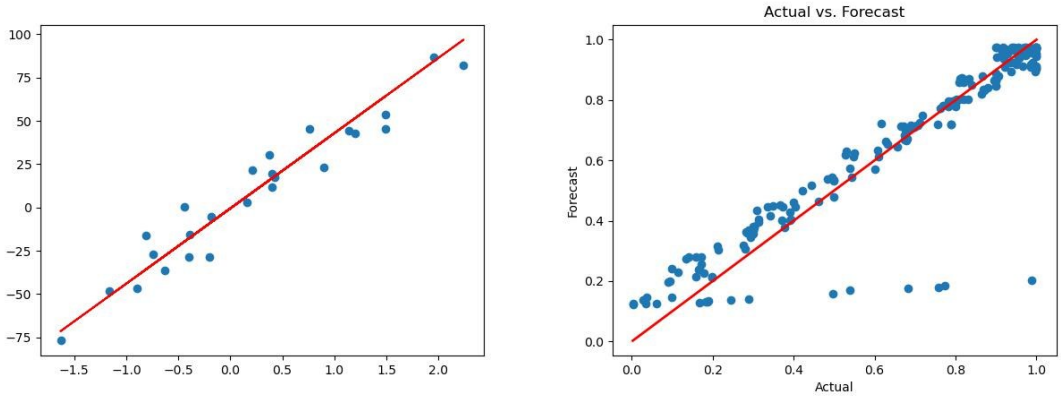

Figure 7: Delayed Prediction Bayesian Regression Model Testing. (Picture credit: Original).

3.3. Logistic Regression Model

For the logistic regression model, this research treats the probability of delay below 0.15 as if it will not be delayed, denoted as 1, and conversely greater than or equal to 0.15 as if it will be delayed, denoted as 0. Assuming that the target predicted value is the probability of delay, the risk level is taken as the feature input into the logistic regression model and the model is validated by examining its Accuracy, the ROC Curve, Confusion Matrix, and the Precision-Recall Curve. This model compares the weighted and unweighted risk levels, yielding the following result. The Accuracy value weighted is 0.9250, and unweighted is 0.8370, the model visualization is shown in Figure 8, Figure 9 and Figure 10.

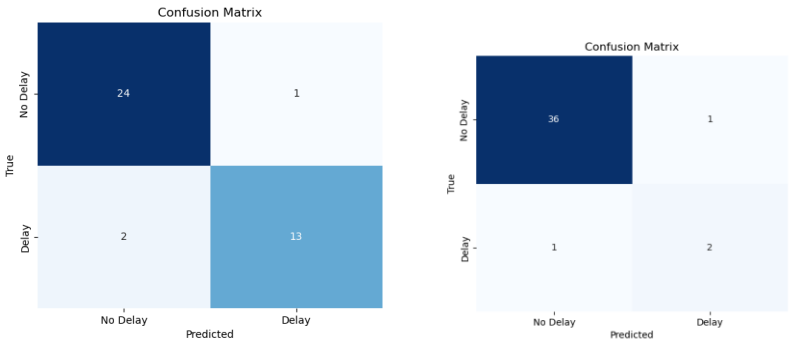

Figure 8: Weighted Confusion Matrix. (Picture credit: Original).

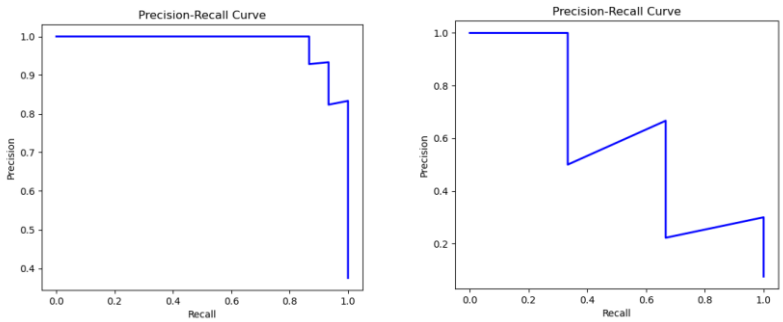

Figure 9: Weighted Precision-Recall Curve. (Picture credit: Original).

DOI: 10.54254/2755-2721/135/2025.21217

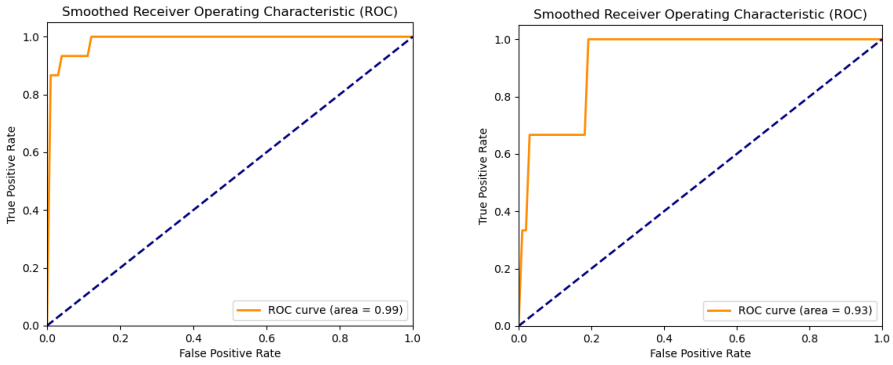

Figure 10: Weighted ROC Curve. (Picture credit: Original).

In the Bayesian regression model, the effect of weighted and unweighted risk levels on predicting the probability of transport delay can be clearly contrasted, due to the fact that the present data has certain noise and outliers, for unweighted due to the relatively poor generalization ability, it is easy to overfitting, and the weighting can be used to improve the robustness and generalization ability of the model by assigning different weights to the data, which can be seen by the R^2 and the MSE as well. It can be judged that there is basically no linear relationship between the risk level and delay probability in the model, while after weighting $R^2$is 0.8695 and MSE is 0.0114. It can be deduced that the linear relationship is very strong, and 86.95% of the changes in the dependent variable can be predicted by the model, therefore the model after weighting, meets the prediction requirements and can further help this research to predict the delay probability.

In the logistic regression model, the difference between the weighted and unweighted models can be clearly contrasted by the accuracy, ROC curve, confusion matrix and precision-recall curve, and it can be seen from both the distribution of the confusion matrix and the ROC or precision-recall curve that the generalization ability and robustness of the model after weighting is stronger than that of the unweighted model. Model after weighting than the unweighted model.

4. Conclusion

Ultimately, the aim of this article is to make the analysis of the logistics data and the relationship between the logistics data more concise and easy to understand through the visualization of these data in Jupyter Notebook. In both Bayesian and linear regression, weighting the risk levels is better than unweighting them, and for Bayesian regression, it is straightforward to predict the exact value of the probability of delay, whereas for linear regression it is intuitive to predict whether or not there will be a delay, and both models play a crucial role. Nevertheless, the test ROC and Precision-Recall Curve in the linear regression equation cannot be fully curved due to the small sample size, which will be solved in the subsequent research. These two models can help to improve the logistics and transport system in Southern California, USA, and to some extent, the global logistics industry to improve the transport environment and reduce the probability of transport delays.

References

[1] *Bhardwaj, D., et al. (2024). Streamlining Global Logistics and Supply Chain Operations: A Process Standardization Framework. Dynamics in Logistics. LDIC 2024. 20-28.*

[2] *Baker, D., Briant, S., Hajirasouli, A., Yigitcanlar, T.,et al. (2023). Urban freight logistics and land use planning education: Trends and gaps through the lens of literature. Transportation Research Interdisciplinary Perspectives, 17, 100731.*

[3] *Pehlivan, P., Aslan, A. I., et al. (2024). Determination of logistics performance of G20 countries using quantitative decision-making techniques. Sustainability, 16(5), 1852.*
[4] *Kwon, N., Kim, J., et al. (2018). T-Reduce: Route-Aware Mobile Trajectory Data Reduction in Transportation Management Systems. IEEE Access, 6, 38934-38946.*
[5] *Sharma, N., Thota, V. S. P., Tankala, Y., Tripathi, S., & Pandey, O. J. (2024). OptRISQL: Toward Performance Improvement of Time-Varying IoT Networks Using Q-Learning. IEEE Transactions on Network and Service Management, 21(3), 3008-3020.*
[6] *Ricard, L., Desaulniers, G., Lodi, A., & Rousseau, L.-M. (2022). Predicting the probability distribution of bus travel time to measure the reliability of public transport services. Transportation Research Part C: Emerging Technologies, 138, 103619.*
[7] *Hudnurkar, M., Renji, K. M., Ambekar, S., Sahu, G., & Joseph, K. M. (2024). Predicting Delays for Truck Delivery Logistics: An Application of AI and ML. 2024 Ninth International Conference on Science Technology Engineering and Mathematics (ICONSTEM), 1-10.*
[8] *Peker, I., Korucuk, S., & Baki, B. (2019). Firm selection based on logistics risk factors: A multiple criteria decision making approach. International Journal of Operations Research and Information Systems, 10(3), 31-43.*
[9] *He, J., Li, Z., Ma, Y., Sun, L., & Ma, K.-H. (2023). Physiological and Behavioral Changes of Passive Fatigue on Drivers during On-Road Driving. Applied Sciences, 13(2), 1200.*
[10] *Bouhadi, O. E., Azmani, M., Azmani, A., & Mouna Atik, e. f. (2022). Using a fuzzy-bayesian approach for predictive analysis of delivery delay risk. International Journal of Advanced Computer Science and Applications, 13(7),316-326.*