CHEMICAL ENGINEERING TRANSACTIONS
VOL. 106, 2023
A publication of AIDIC
The Italian Association of Chemical Engineering
Online at www.cetjournal.it
Guest Editors: Jeng Shiun Lim, Nor Alafiza Yunus, Peck Loo Kiew, Hon Huin Chin

ISBN 979-12-81206-05-2; **ISSN** 2283-9216
DOI: 10.3303/CET23106039

Air Quality Modeling Using Real-time Urban Big Data

Joonsik Jo[a], Gahyun Kim[b], Juhyeon Kwak[a], Ilho Jeong[a], Donggyun Ku[c], Seungjae Lee[b,*]

[a]Department of Transportation Engineering / Department of Smart Cities, University of Seoul, Korea
[b]Department of Transportation Engineering, University of Seoul, Korea
[c]Department of Land Economy, University of Cambridge, UK
sjlee@uos.ac.kr

This study utilized real-time urban data from Seoul to assess the impact of transportation mode choice on air pollution metrics. EXtreme Gradient Boosting (XGBoost), an ensemble model, was employed for air quality analysis. Subsequently, SHapley Additive exPlanations (SHAP), one of eXplainable Artificial Intelligence (XAI) techniques, was used to understand the influence of urban factors on air quality. For spatial coverage, 50 locations with high traffic volume were selected, and the temporal coverage spanned from April 3, 2023, to April 30, 2023. Variables related to traffic, the environment, and weather were established as features, while Comprehensive Air-quality Index (CAI), PM2.5, and PM10 were determined as target variables. As a result, Root Mean Squared Error (RMSE) of the models predicting CAI, PM2.5, and PM10 were calculated as 0.57, 0.47, and 0.50. The study found that as the maximum number of pedestrians and the number of subway passengers alighting increased, the levels of CAI, PM2.5, and PM10 decreased. This indicates that the use of greener modes of transportation, such as walking and taking the subway, positively impacts air pollution reduction In addition, lower road traffic speeds were associated with higher PM2.5 levels, while increased road congestion correlated with higher PM10 levels. The observed increase in PM2.5 and PM10 levels in relation to the rise in passenger car traffic suggests that emissions from these vehicles are contributing to air pollution. Consequently, the study confirmed that traffic-related factors can influence air quality indicators, and that modifications to traffic volumes and modal splits can enhance air pollution control. This study provides a foundation for developing policies to improve air quality by quantifying and presenting various factors that impact air quality.

1. Introduction

In recent years, air pollution-related environmental and health issues have garnered significant global attention due to the detrimental effects they pose (Ku et al., 2021). High level of air pollution leads to poor air quality, causing a myriad of problems and escalating social costs (Colvile et al., 2001). The transportation sector, in particular, contributes heavily to air pollution, with this problem intensifying due to urbanization and population growth. The transportation sector is widely recognized as a major contributor to air pollutant emissions (Ku et al., 2020). Vehicles such as cars and buses emit harmful substances while burning fuel and traveling the roads, leading to the degradation of air quality (Ercan et al., 2022). Air pollutant emissions can vary widely, influenced by factors such as vehicle type, fuel used, and engine size (Van Fan et al., 2018). To reduce air pollution from the transportation sector, it is essential to accurately analyze the emissions of air pollutants by transportation and develop both technical and policy measures for improvement. Such analysis could provide valuable insights for effective air quality management and formulating efficient strategies to reduce air pollution. By gaining a deeper understanding of how different transportation modes affect air quality, more efficient measures can be derived to effectively control and reduce air pollution. Consequently, this study aims to investigate the variations in air pollutant emissions among different modes of transportation and assess their impact on air quality. The findings will aid in developing of policies and technologies that can reduce air pollution, ultimately contributing to the creation of healthier, more sustainable urban environments.

Paper Received: 12 June 2023; Revised: 3 August 2023; Accepted: 19 August 2023
Please cite this article as: Jo J., Kim G., Kwak J., Jeong I., Ku D., Lee S., 2023, Air Quality Modeling Using Real-time Urban Big Data, Chemical Engineering Transactions, 106, 229-234 DOI:10.3303/CET23106039

2. Methodology

First, real-time urban data was collected using a web crawler. This preprocessed data was then used to train air quality prediction models and to evaluate their performance. Following XAI approach, the importance of features was measured, and influencing factors were analyzed. Figure 1 illustrates the flow of this study.

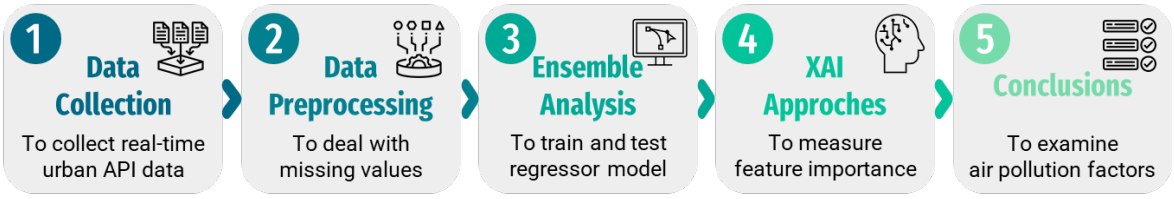

Figure 1: Framework of study

2.1 Real-time urban big data

Seoul's real-time urban data is available as an open Application Programming Interface (API), which implies that data can be collected without any access restrictions through the provided API. API facilitates the collection of real-time data as well as data that gets updated periodically (Lock et al., 2020). In this study, Python's schedule function was employed to collect data at regular intervals. The data provides real-time information on population, traffic, and environmental variables for 50 key locations in Seoul. To comprehend the impact of transportation-related air pollutant emissions, the data presented in Table 1 was collected. The data were collected in April, which is representative of the Korean climate during the year. The data spanned from April 3, 2023, to April 30, 2023.

Table 1: Seoul real-time urban data

Variable	Explanation	Property
ID	Place name	Character
Population density	Location congestion indicator based on real-time pedestrian volume (0: smooth, 1: normal, 2: less crowded, 3: crowded)	Factor
Maximum pedestrian volumes	Maximum of real-time pedestrian volume (person/5 min)	Integer
Minimum pedestrian volumes	Minimum of real-time pedestrian volume (person/5 min)	Integer
Resident population rate	Percentage of resident population (%)	Float
Traffic congestion	Status of overall road communication (0: smooth, 1: slow, 2: congestion)	Factor
Traffic speed	Average speed of overall road communication (km/h)	Integer
Time	Time of data generation	Time
Date	Date of data generation	Date
Temperature	Temperature (°C)	Float
Sensory temperature	Sensory temperature (°C)	Float
Humidity	Humidity (%)	Integer
Wind speed	Wind speed (km/h)	Float
Precipitation	Rainfall (mm)	Float
UV level	UV Index levels	Integer
PM2.5	Ultra fine dust concentration ($\mu g/m^3$)	Float
PM10	Fine dust concentration ($\mu g/m^3$)	Float
CAI	Comprehensive air-quality index	Integer
Subway passengers boarding	Number of people getting on the subway (person/h)	Integer
Subway passengers alighting	Number of people getting off the subway (person/h)	Integer

2.2 Ensemble model: XGBoost

An ensemble model is a machine learning method that combines a number of individual models to generate stronger, more reliable predictions. Each model is trained independently, makes predictions, and then their results are combined to deliver a final prediction. Ensemble models can enhance predictive performance over individual models, and generally yield more stable and robust results. XGBoost is one such ensemble technique. It is a powerful and widely-used machine learning algorithm based on the boosting algorithm (Chen and Guestrin, 2016). XGBoost is an evolution of the original gradient boosting algorithm that operates as a tree-based learning

algorithm. It takes advantage of Gradient Boosting while avoiding overfitting and increasing the generalization ability of the model. XGBoost provides a range of features and hyperparameters to fine-tune the performance and efficiency of the model. For instance, adjusting parameters such as tree depth, learning rate, number of trees, etc., allows for control over the model's complexity and generalization ability. One significant advantage of XGBoost is its ability to compute feature importance, enabling the evaluation and interpretation of the significance of variables. XGBoost can be applied to a diverse range of problems, demonstrating its strength in handling interactions between structured data and attributes. It is also efficient in terms of speed and performance, making it suitable for large datasets. The XGBoost model aggregates each individual tree model as shown in Eq(1) to take them all into account, and then generalizes them as shown in Eq(2). Ultimately, by setting the objective function as shown in Eq(3), the model's performance is enhanced via the model's regularization function to avoid overfitting.

$$Y' = a * tree_A + b * tree_B + c * tree_C + \cdots \quad (1)$$

$$Y_i' = \sum_{k=1}^{K} f_k(x_i), \qquad f_k \in F \quad (2)$$

$$obj(\theta) = \sum_{i=1}^{n} l(y_i, y_i') + \sum_{k=1}^{K} ohm(f_k) \quad (3)$$

Where y_i' is the predicted score corresponding, f_k is k th decision tree $\in$ function space F, l is loss function, and ohm is regularization function.

Lartey et al. (2021) used the XGBoost algorithm to predict hourly traffic volume efficiently and accurately. Sun et al. (2021) compared various models for predicting highway traffic flow and used the best-performing XGBoost algorithm to predict traffic flow. In the study, XGBoost was used to create a model to predict CAI, PM2.5, and PM10 to identify the amount of air pollutants caused by transportation.

2.3 XAI: SHAP

SHAP is a form of XAI used for interpreting and explaining the predictions of machine learning models. Derived from the game theory concept of Shapley values, SHAP concentrates on explaining the influence of each input feature on the predicted outcome (Lundberg and Lee, 2017). This methodology allows for the decomposition of the model's predictions into the contributions of each feature, assessing the significance of each one. This approach enables the identification of individual feature influences and provides explanations for the model's decisions. SHAP is based on the calculation of Shapley values, which measure the degree to which participants in a cooperative game contribute to the creation of some value. This concept is applied to the model's input features to estimate the extent of their contribution to the predicted outcome. SHAP provides a consistent interpretation, regardless of the model's complexity, and presents the input-output relationships of the model in an interpretable manner. This helps in understanding the model's predictions and facilitates confident decision-making. In a SHAP model, the feature contribution is calculated by adding the difference between the prediction and the mean for the data, as shown in Eq(4). The precise Shapley value is then calculated using Monte-Carlo sampling using Eq(5). Finally, Eq(6) yields the feature importance.

$$\emptyset_s^m = \hat{f}(x_{+s}^m) - \hat{f}(x_{-s}^m) \quad (4)$$

$$\emptyset_s(x) = \frac{1}{M} \sum_{m=1}^{M} \emptyset_s^m \quad (5)$$

$$I_s = \sum_{i=1}^{n} \left| \emptyset_s^{(i)} \right| \quad (6)$$

Where $\emptyset_s$ is the contribution of the selected independent variable, $\emptyset_s(x)$ is the contribution of all variables, and I_s is the importance of the entire model.

Yang et al. (2022) developed the XGBoost model by selecting variables from past traffic accident data and used SHAP values to analyze 11 influencing factors from perspectives of the road and environment. Barredo-Arrieta et al. (2019) used XAI techniques to quantify the impact of each independent variable on the target variable for deeper insights into traffic analysis and flow prediction. This study aimed to use these models to identify which independent variables have high feature importance and how they affect CAI, PM2.5, and PM10.

3. Results

This section presents the results of the ensemble modeling and feature importance analysis conducted using SHAP to figure out the relationship between atmospheric environment and transportation.

3.1 Results of air quality prediction

Before training the model, the data was split into training and testing sets. Train-data spanned from April 3, 2023 to April 23, 2023, and test-data from April 24, 2023 to April 30, 2023. By tuning the hyperparameters of the XGBoost model, numerous air quality prediction models were trained, with the best model selected by comparing prediction results. The detailed hyperparameters adjusted include the number of boosting stages (estimators), the learning rate (learning rate), the depth of the tree (max depth), the feature sampling rate for each tree (colsample), and the observed data sampling rate for each tree (subsample). After predicting test-data using the optimal model, each model's predictive performance was evaluated using R-squared and RMSE values. The R-squared value, ranging between 0 and 1, indicates similarity to the predicted value; the models predicting CAI, PM2.5, and PM10 yielded values of 0.57, 0.47, and 0.50. RMSE, an indicator that determines the average error considering the unit of the dependent variable, resulted in values of 8.83, 5.04, and 7.22 for the models predicting CAI, PM2.5, and PM10. Detailed hyperparameter information and performance evaluation results of each model are shown in Table 2. Figure 2 is a graph comparing the actual values of test-data with the predicted values from each model, with the x-axis representing observations and the y-axis representing the dependent variable.

Table 2: Hyper parameters and evaluation metrics of each model

Model	Estimators	Learning rate (%)	Max depth	Colsample (%)	Subsample(%)	R-squared (%)	RMSE
CAI	200	0.1	7	0.9	0.8	0.57	8.83
PM25	200	0.1	7	0.8	0.9	0.47	5.04
PM10	200	0.1	7	0.9	0.8	0.50	7.22

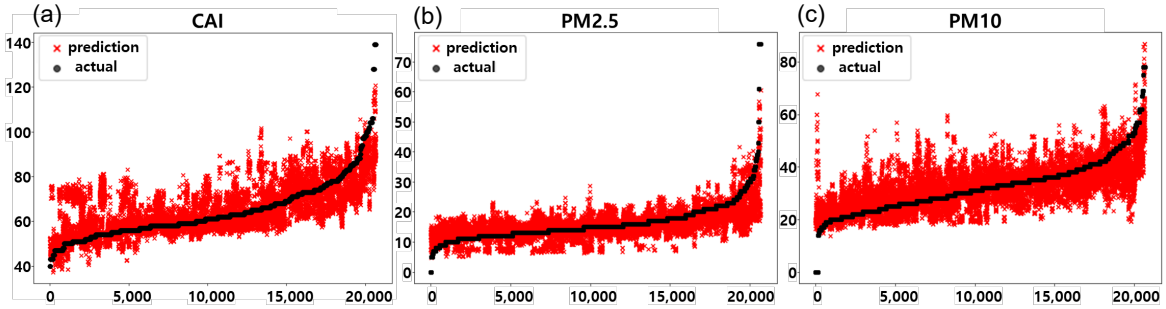

Figure 2: Prediction results of (a) CAI model (b) PM2.5 model (c) PM10 model

3.2 Influencing factor in air quality

SHAP was used to analyze how each variable influenced model predictions. Figure 3 visualizes the distribution of Shapley values for all characteristics, with the upper variables having a greater absolute impact on the prediction of each dependent variable. The top six variables with the greatest impact on CAI were humidity, hour, UV level, temperature, sensory temperature, and subway passengers alighting. For PM2.5, they were hour, humidity, temperature, UV level, subway passengers alighting, and maximum pedestrian volumes. For PM10, they were humidity, hour, temperature, UV level, maximum pedestrian volumes, and sensory temperature. The graph also indicates that a red dot representing a high feature value contributed to an increase in the dependent variable value, while a blue dot representing a low feature value contributed to a decrease. All models predicting CAI, PM2.5, and PM10 yielded high predictions for the dependent variable when maximum pedestrian volumes were low, as the feature value tended to be red when the SHAP value of maximum pedestrian volumes was less than 0. The feature values of the PM2.5 and PM10 models tended to be red when the SHAP value of subway passengers alighting was less than 0, resulting in higher predictions for the dependent variable when the number of subway passengers alighting was low. In the same way, the PM2.5 model predicted higher dependent variables when traffic speed was low, while the PM10 model predicted higher dependent variables when traffic congestion was high.

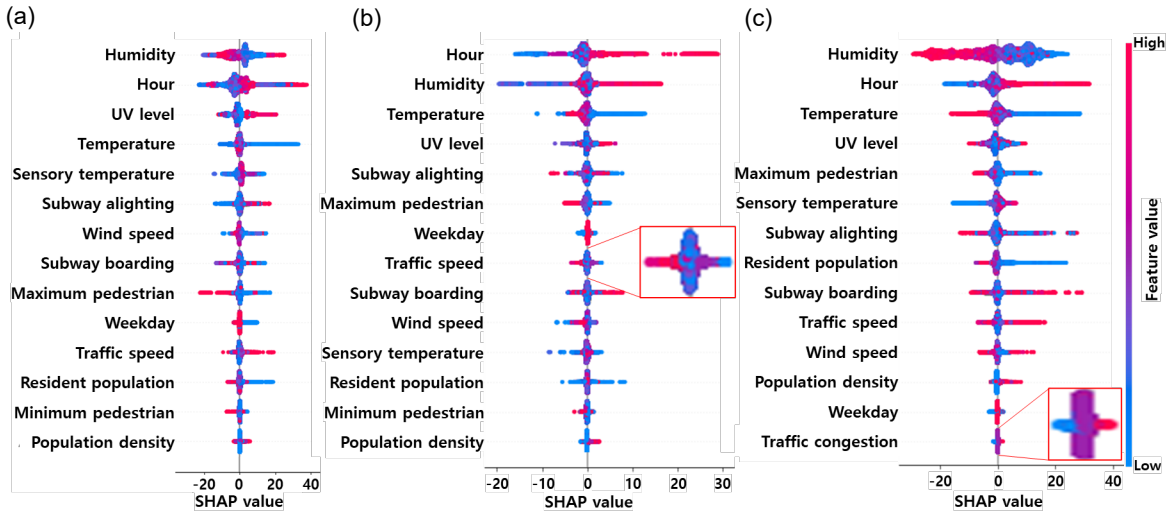

Figure 3: Feature importance of (a) CAI model (b) PM2.5 model (c) PM10 model

To specifically examine how transportation variables affect the dependent variables, the independent feature importance of each variable was confirmed in Figure 4. The SHAP value of CAI is lower than the dotted red line only when the maximum pedestrian volumes are more than 100,000. Conversely, the value is lower than the dotted red line only when the subway passengers alighting are less than 1,000. In other words, the air quality was mainly predicted as low when there were numerous pedestrians and alighting subway passengers, which aligns with the results of the overall interactive feature importance analysis discussed above. In addition, when checking the feature importance of population density in the PM2.5 and PM10 models, the dotted red line indicating the median of the SHAP value when the population density was 0 was significantly lower than the median of other population densities. It means that PM2.5 and PM10 were predicted as low when the population density was 0, meaning that the area was not congested.

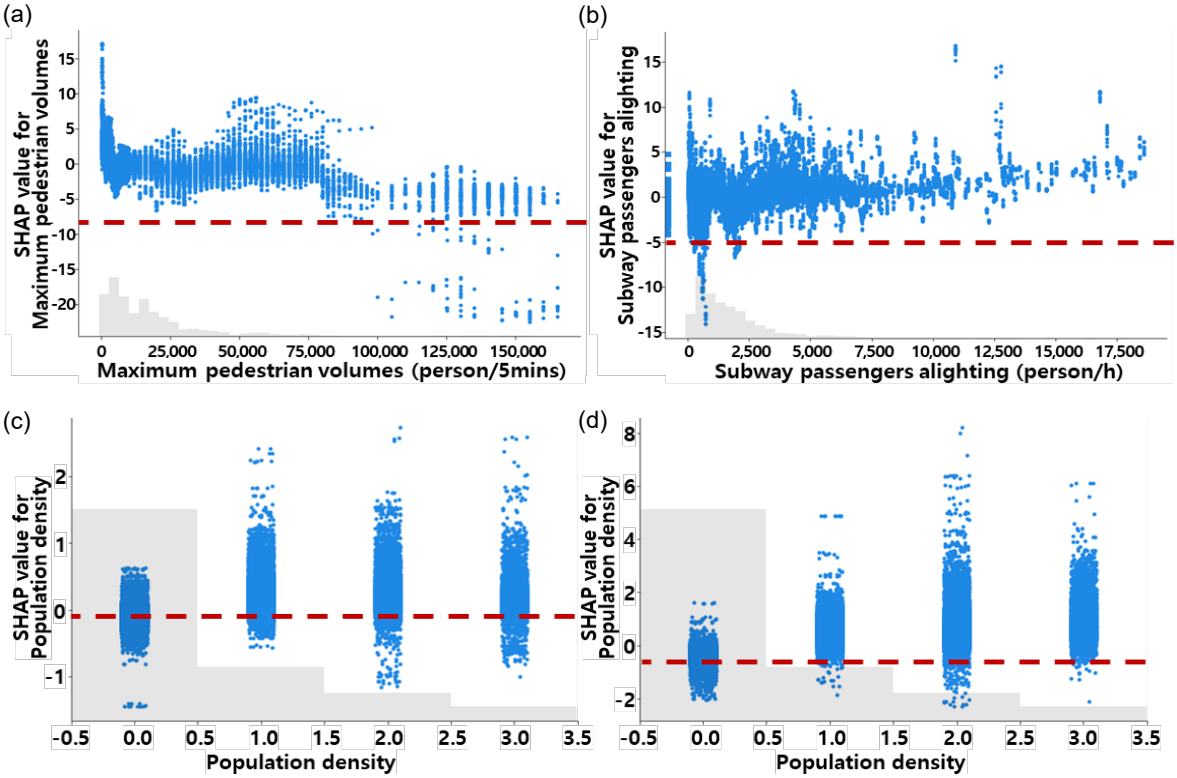

Figure 4: SHAP value for (a) Maximum pedestrian volumes of CAI model (b) Subway passengers alighting of CAI model (c) Population density of PM2.5 model (d) Population density of PM10 model

4. Conclusions

In this study, air quality was predicted by using a variety of variables created by preprocessing urban data collected in real-time. Subsequently, the SHAP method was used to understand how each variable impacts air quality. The analysis revealed that as the maximum number of pedestrians and the number of subway passengers alighting increased, meanwhile CAI, PM2.5, and PM10 levels decreased. This implies that use of green transportation, such as walking and subway, positively impacts air pollution reduction. Additionally, PM2.5 was higher when the road speed was lower, and PM10 was higher when the road congestion level was higher. This suggests that as the traffic volume of passenger cars increases, PM2.5 and PM10 increase, indicating that emissions from passenger cars contribute to air pollution. Consequently, this study implies that mode choice of transportation can significantly affect air pollution levels. Based on these findings, they can provide a foundation for developing policies aimed at improving air quality by quantifying and presenting factors that influence it. For instance, it would be possible to analyze the effect on air pollution before and after implementing Transportation Demand Management (TDM) project, such as congestion pricing or low emission zones. Consequently, this could help in moving towards a sustainable city by influencing changes in traffic volumes and modal split. In further studies, a more accurate and robust air quality prediction model could be established by collecting more diverse control variables. In addition, a feedback cycle could be set up to understand mutual effects by not only analyzing the impact of transportation use on the atmospheric environment but also studying how weather conditions influence passengers' transportation choice behavior.

Acknowledgments

This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2022R1A6A3A01087328). This work was also supported by the Korea Ministry of Land, Infrastructure, and Transport (MOLIT) as an Innovative Talent Education Program for Smart City.

References

Barredo-Arrieta A., Laña I., Del Ser J., 2019, What Lies Beneath: A Note on the Explainability of Black-box Machine Learning Models for Road Traffic Forecasting, 2019 IEEE Intelligent Transportation Systems Conference (ITSC), 27th-30th October, Aukland, New Zealand, 2232–2237.

Chen T., Guestrin C., 2016, XGBoost: A Scalable Tree Boosting System, Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, 13th-17th August, San Francisco, United States of America, 785-794.

Colvile R.N., Hutchinson E.J., Mindell J.S., Warren R.F., 2001, The transport sector as a source of air pollution, Atmospheric Environment, 35(9), 1537–1565.

Ercan T., Onat N.C., Keya N., Tatari O., Eluru N., Kucukvar M., 2022, Autonomous electric vehicles can reduce carbon emissions and air pollution in cities, Transportation Research Part D: Transport and Environment, 112, 103472.

Ku D., Bencekri M., Kim J., Lee S., Lee S., 2020, Review of European Low Emission Zone Policy, Chemical Engineering Transactions, 78, 241-246.

Ku D., Kwak J., Na S., Lee S., Lee S., 2021, Impact Assessment on Cycle Super Highway Schemes, Chemical Engineering Transactions, 83, 181-186.

Lartey B., Homaifar A., Girma A., Karimoddini A., Opoku D., 2021, XGBoost: A tree-based approach for traffic volume prediction, 2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 17th-20th October, Melbourne, Australia, 1280–1286.

Lock O., Bednarz T., Pettit C., 2020, The visual analytics of big, open public transport data–a framework and pipeline for monitoring system performance in Greater Sydney, Big Earth Data, 5(1), 134-159.

Lundberg S.M., Lee S.I., 2017, A Unified Approach to Interpreting Model Predictions, Advances in Neural Information Processing Systems, 30.

Sun B., Sun T., Jiao P., 2021, Spatio-Temporal Segmented Traffic Flow Prediction with ANPRS Data Based on Improved XGBoost, Journal of Advanced Transportation, 1-24.

Fan, Y. V., Klemes J.J., Perry S., Lee C.T., 2018, An Emissions Analysis for Environmentally Sustainable Freight Transportation Modes: Distance and Capacity, Chemical Engineering Transactions, 70, 505-510.

Yang Y., Wang K., Yuan Z., Liu D., 2022, Predicting Freeway Traffic Crash Severity Using XGBoost-Bayesian Network Model with Consideration of Features Interaction, Journal of Advanced Transportation, 2022.