Proceedings of the 2025 Winter Simulation Conference
E. Azar, A. Djanatliev, A. Harper, C. Kogler, V. Ramamohan, A. Anagnostou, and S. J. E. Taylor, eds.

ROAD TRAFFIC CONGESTION PREDICTION USING DISCRETE EVENT SIMULATION AND REGRESSION MACHINE LEARNING MODELS

Ivan Kristianto Singgih[1,2,3,4], Moses Laksono Singgih[5], and Daniel Nathaniel[1]

[1]Stud. Prog. of Ind. Eng., University of Surabaya, Surabaya, INDONESIA
[2]Indonesia Artificial Intelligence Society, Jakarta, INDONESIA
[3]Kolaborasi Riset dan Inovasi Industri Kecerdasan Artifisial (KORIKA), Jakarta, INDONESIA
[4]The Indonesian Researcher Association in South Korea (APIK), Seoul, SOUTH KOREA
[5]Dept. of Ind. and Syst. Eng., Institut Teknologi Sepuluh Nopember, Surabaya, INDONESIA

ABSTRACT

Road traffic management enters a new era with the automatic collection and analysis of big data. The traffic data could be collected continuously using various IoT sensors (light, video, etc.) and stored in the cloud. The collected data are then analyzed within a short time to make traffic control decisions, e.g., traffic redirection, traffic light duration change, and vehicle route recommendation. This study proposes (1) a traffic simulation considering a road network with several traffic lights and (2) regression machine learning models to understand the behavior of the vehicles based on the real-time characteristics of the traffic. The numerical experiment results show that (1) the best models are OrthogonalMatchingPursuitCV and the HuberRegressor, and (2) the road network behavior is affected by the condition of all intersections rather than only certain intersections or surrounding road segments.

1 INTRODUCTION

Interconnection between the physical and virtual worlds (Groth 2019) is an essential part of a digital twin system and is enabled by installing IoT sensors to record the data from the physical (real) system. The recorded data are then used to test various decisions in the virtual world (e.g., using simulation). The decisions include the traffic light duration change (Eom and Kim 2020) and dynamic vehicle rerouting (Dutta et al. 2023). After the testing, the best decision is implemented back into the physical system (Singgih 2021). Given any observed system characteristics, such continuous iterations allow for the best decisions to be made. Such a strategy reduces traffic congestion and costs (Peprah et al. 2019).

In this study, a simulation-based analysis is conducted to understand the behavior of a traffic system and the relationships between its components. The simulation is used to mimic the behavior of vehicles on a road network with several intersections. All information related to the vehicles, road utilization, etc., is then analyzed using machine learning to identify the most important features for predicting the congestion level at the whole road network. Such a research direction is recently being considered as an important issue by many researchers (Dammak et al. 2025, Bakir et al. 2024, Kafy et al. 2024). Such an understanding on the system's behavior would significantly help the decision makers to reduce the congestion by focusing on the most important features.

2 RELATED WORKS

Existing studies predict traffic volumes using statistics (de Barros et al. 2023; Min and Wynter 2011) and vehicle movement-based delay models (Lee et al. 2017; Mirchandani and Head 2001). Recent studies applied machine learning techniques for traffic prediction (Qiu et al. 2023; Weng et al. 2023; Xu et al. 2023; Zheng et al. 2023). The machine learning is selected, instead of other methods, e.g., design of experiment,

because the machine learning methods (1) work well with undesigned data and (2) does not require assumptions on the data distribution (Arboretti et al. 2022). In contrast, methods like design of experiment requires the proper understanding of the data to select the right sampling points and combinations of levels of the features (Matković et al. 2015; Arboretti et al. 2022). Despite the existence of such machine learning studies, only a few studies have used simulation for data generation (Gomes et al. 2023). It has been known that the simulation approach is necessary to model real situations with the most degree of detail, which could not be resolved solely by operations research methods that work based on many assumptions of the real systems (van Dijk and van der Sluis 2008). Our study fills such a research gap by using simulation to generate data for the prediction models. The contributions of this study are:

a. The machine learning data generation using a discrete event simulation software
b. The application of various regression machine learning models to understand the traffic behavior at a hypothetic road network

3 PROPOSED SIMULATION-MACHINE LEARNING ANALYSIS

Our study supports the implementation of the digital twin concept illustrated in Figure 1. The real-time traffic (physical world) section consists of two parts: (1) the system behavior presentation (including the vehicle arrival and movement behavior and the road network) and (2) the real system continuous running. The traffic simulation (virtual world) consists of two parts as well: (3) the real system representation (the simulation itself) and (4) the simulation run (including the data collection, system behavior prediction, and decision-making and evaluation). To enable smooth run of the digital twin concept and a good performance in the real system, the decision evaluation process must be conducted during a short time in the simulation. The complete information transfer between both worlds is explained as follows:

1. 1→3: The real system behavior is used as the basis for the simulation design.
2. 3→4: The designed simulation is then executed continuously in coordination with the real system running.
3. 2—3: The simulation is validated using the real system run. Also, the simulation model is adjusted accordingly when any new condition (system behavior) in the real system is observed.
4. 4→2: The traffic management decision tested in the simulation is implemented to optimize the real system.

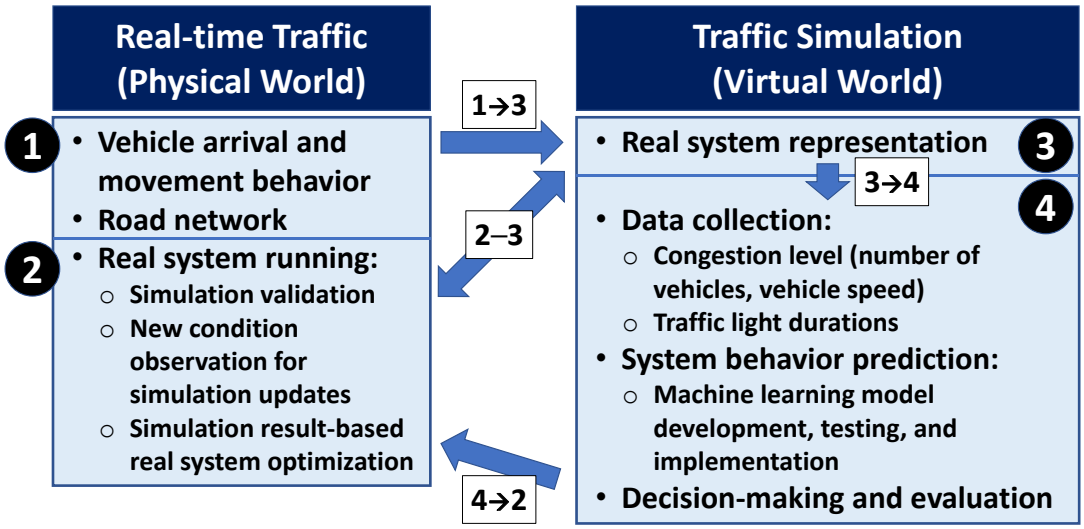

Figure 1: The digital twin concept for the traffic management.

Our study implements a partial function of the digital twin concept, which is within the virtual world part of the system, in which the simulation is run, the data are collected, and the data relationships are defined using machine learning.

4 DISCRETE EVENT SIMULATION DESIGN

The discrete event simulation is designed using the Simulation of Urban MObility open-source software (version 1.10.0), which is executed with the Python programming language (SUMO 2024; Figure 2). As shown in Figure 2, a hypothetical road network is considered, consisting of four intersections with a traffic light at each intersection. Each road segment is bidirected with one or two lanes for each direction. An example of the road lanes at an intersection is shown in Figure 3.

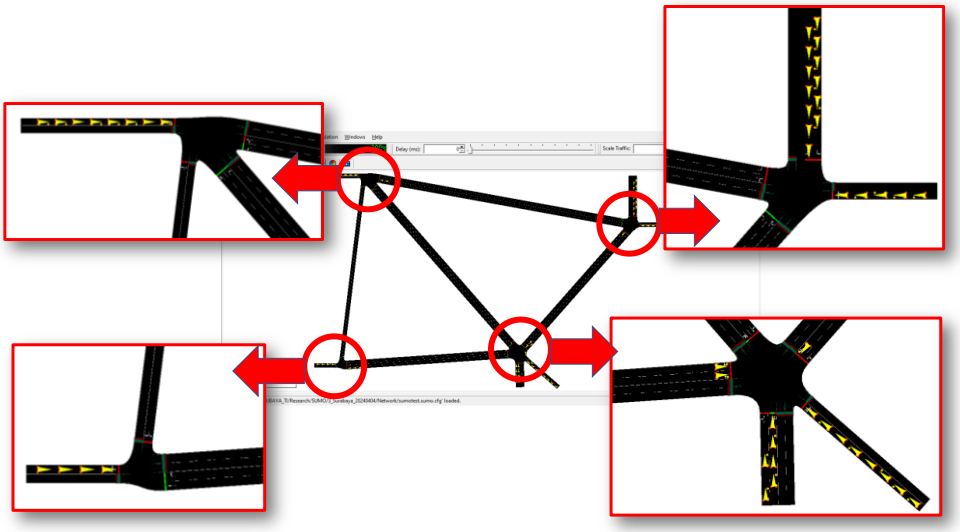

Figure 2: The designed discrete event simulation.

One traffic cycle for any traffic light is 180 seconds, and each traffic light operation is divided into three to five phases with equally distributed initial durations. The duration of each yellow light is 5 seconds, and the remaining duration is distributed to the green lights. An example of traffic light phases is shown in Figure 4. A complete cycle of the traffic light at the intersection starts from phases 1, 2, 3, and up to phase 8. For each light phase, its duration is set arbitrarily into 31 s, 40 s, and 55 s for the green (the same green light duration for all phases in an intersection) and 5 s for yellow light, respectively.

The generated traffic light durations and proportion fit the values presented in Han et al. (2024) and Lin et al. (2023), who defined that the yellow light duration is significantly smaller than the green light duration, and green light duration is more than 30 seconds. The simulation is run with various vehicle arrival rates (to allow considering various traffic conditions) as follows: 0.1, 0.2, 1, 2, 4, and 10 vehicles per second. During each simulation run, 99 data rows are collected. Each row consists of complete information about input and output features listed in Table 1. The pseudocode of the Python code is illustrated in the following steps:

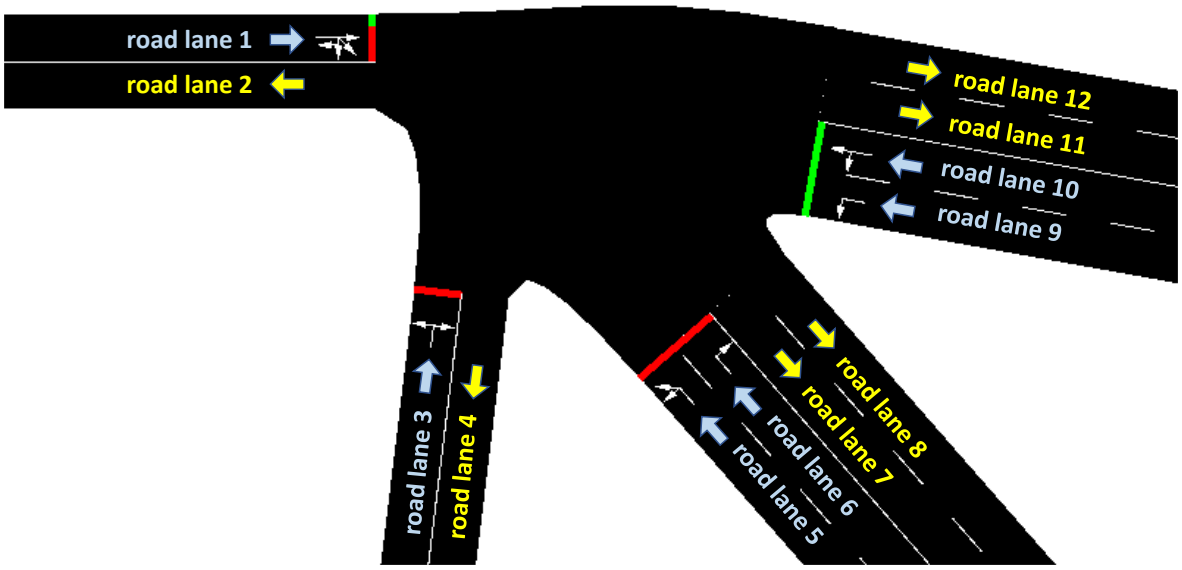

Figure 3: Example of road lanes at an intersection.

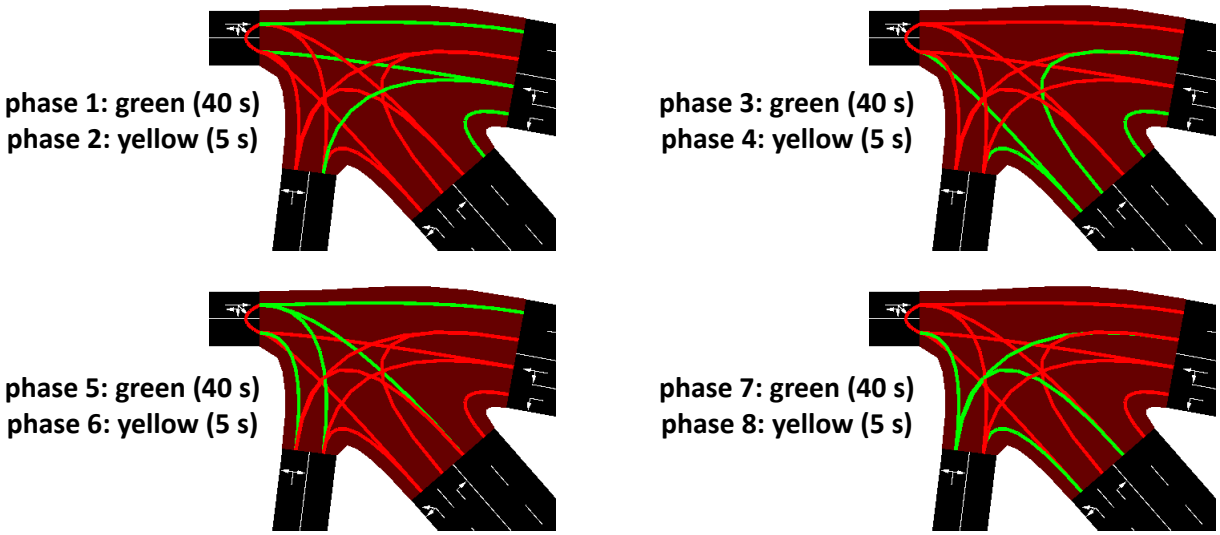

Figure 4: Example of traffic light phases at an intersection (green lines indicate the lanes enabled during each traffic light phase, while the red lines represent the inactive lanes during the phase).

Part 1: Simulation of Urban Mobility

1. **input** config_file, network_file, vehicle_arrivals_and_route_file, traffic_light_information
2. **input** simulation_length, green_light_duration, yellow_light_duration, traffic_cycle_time
3. **while** simulation_time < simulation_length **do**
4. **increase** simulation_time and **simulate** vehicle movements
5. **calculate** and **store** metrics into a database: number_of_vehicles_passing_intersections, average_vehicle_speed_per_lane, number_of_vehicles_per_lane,
6. **end while**
7. **copy** database into Excel

Part 2: Machine Learning

8. **input** regression machine learning models, input features and prediction target
9. **perform** data normalization
10. **divide** data into training and testing data
11. **train** regression machine learning models
12. **evaluate** regression machine learning models using testing data
13. **observe** and **analyze** feature importance

5 REGRESSION MACHINE LEARNING PREDICTION MODELS

The data collected during the simulation is presented in Table 1. The input and output data for the machine learning models are listed. Various input data are observed, including the vehicle arrival rates, the number of vehicles on each road segment, the average vehicle speed on each road segment, and the phase duration of each traffic light. After each complete traffic light cycle, several output data are observed: the number of vehicles passing each intersection and the number of vehicles passing all intersections.

Table 1: Data collected during the simulation.

Input/Output data	Feature	Description	Number of features
input	(1) arrival_rate_per_second	Vehicle arrival rate between each pair of origin and destination nodes (number of vehicles per second)	1
input	(2) number_of_vehicles_ <road_segment_ID>	Number of vehicles on each road segment (accumulation of all lanes on the segment)	22
input	(3) average_speed_ <road_segment_ID>	The average speed of vehicles on each road segment (average of all lanes on the segment) (km/h)	22
input	(4) light_duration_ <intersection ID>_ <phase ID>	The phase duration at each road intersection (seconds)	32
output	(5) number_of_passing_ vehicles_<intersection ID>	Number of vehicles passing each intersection during a complete traffic light cycle	4
output	(5) number_of_passing_ vehicles_all_intersections	Number of vehicles passing all intersections during a complete traffic light cycle	1

Each output data is predicted using the following regression machine learning models (scikit-learn, 2024; with the references on traffic prediction studies that utilized the models within the same variant classes): Random Forest Regression (Luitel et al. 2025), Linear Regression (Luitel et al. 2025), RidgeCV (Lin et al. 2022), ElasticNetCV, LarsCV, LassoCV (Luitel et al. 2025), LassoLarsCV (Luitel et al. 2025), OrthogonalMatchingPursuitCV, ARDRegression, BayesianRidge (Lin et al. 2022), HuberRegressor, RANSACRegressor, TheilSenRegressor, Poisson-Regressor (Bonela and Kadali 2023; Goyani et al. 2024), TweedieRegressor (Goyani et al. 2024), and PassiveAggressiveRegressor. Some models that have no references but exist in the Python library are tested to observe the match possibility between the studied problem and the model characteristics.

6 NUMERICAL EXPERIMENTS

Correlations between the input and output data are observed to allow predicting the output properly (Figure 5). It is shown that: (a) when there are more vehicles, the average speed at the road segment is reduced, (b) when there are more vehicles on the road segment, more vehicles pass the intersection, and (c) when the average speed of vehicles is higher on the road segments, less vehicles pass the intersections (the reason is that less vehicles enter the intersection).

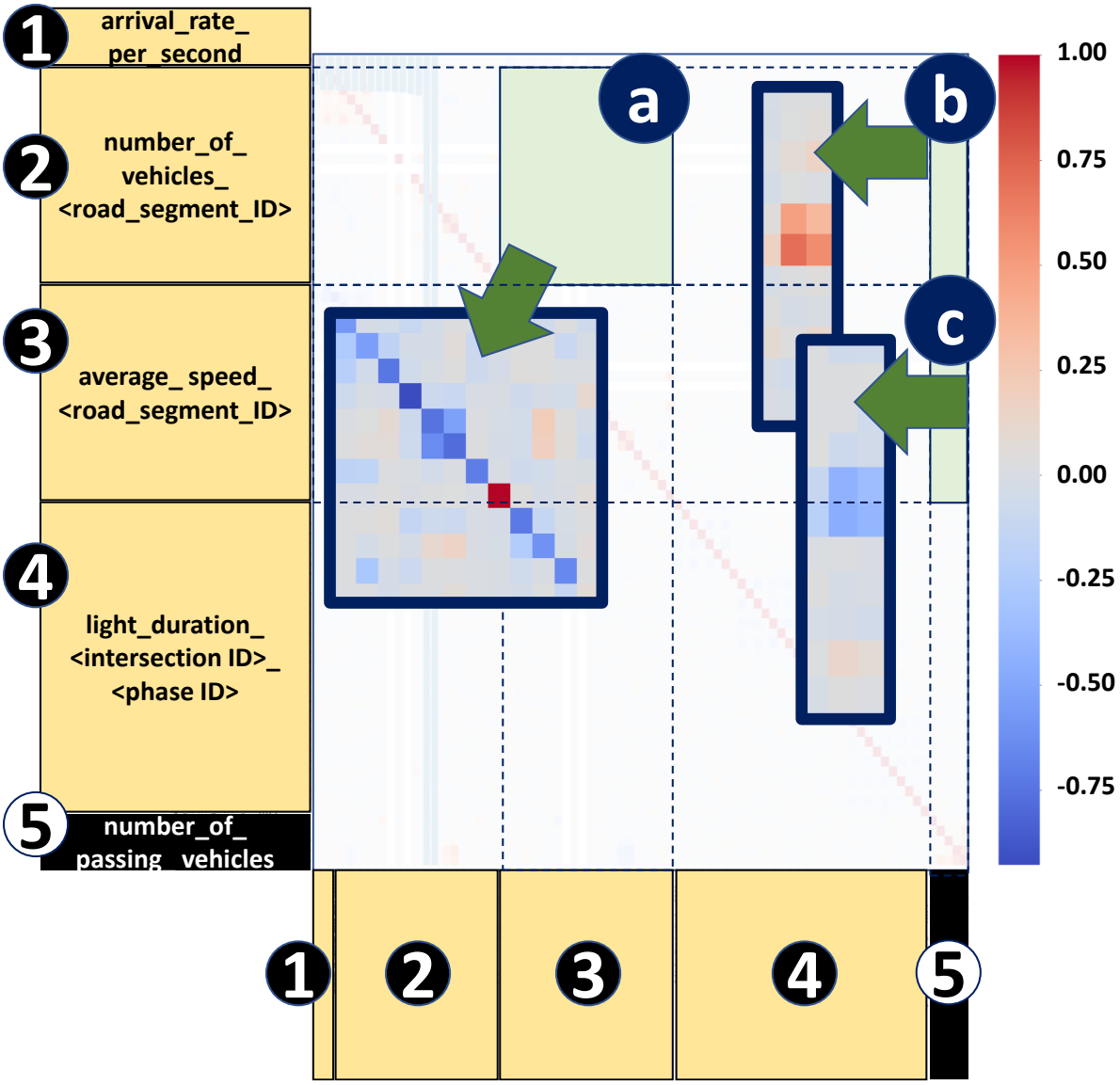

Figure 5: Correlation between the input (with yellow backgrounds) and output data (with black backgrounds).

The performance of the regression machine learning models is measured in mean absolute error and shown in Table 2. The results are differentiated based on the predicted output data. The best prediction models are the OrthogonalMatchingPursuitCV and the HuberRegressor models with the least mean absolute error values as shown in Table 2.

When predicting the number of vehicles passing each intersection, the ten most influential input data are observed. The importance of each input data is defined based on the absolute coefficients of the best regression models. Among each set of ten input data, percentages of input data that are related to the intersection ID, which is predicted, are reported, as shown in Figure 6. For most intersections, it is shown that the predicted output is more affected by input data that are not directly connected to the intersections. The results show that the road network behavior is influenced by many areas in the network, not solely by the area close to the intersection. Such behavior could be observed through the implementation of machine learning models. Different from Figure 6, features that significantly affect the total number of vehicles passing all intersections are observed in Figure 7, instead of the ones passing each intersection. When the number of vehicles passing all intersections is predicted, the ten most influential input data are identified and observed. Figure 7 shows that a similar number of input data (features) from each intersection is used for the prediction. The result illustrates how the overall road network performance is affected by the conditions of all intersections.

Table 2: Mean absolute errors of the regression machine learning models.

Regression Machine Learning Model	***y* = Number of Vehicles Passing Intersection-1**	***y* = Number of Vehicles Passing Intersection-2**	***y* = Number of Vehicles Passing Intersection-3**	***y* = Number of Vehicles Passing Intersection-4**	***y* = Number of Vehicles Passing All Intersections**
RandomForestRegressor	4.1	41.4	22.1	40.6	71.9
LinearRegression	1.6 (10^{10})	3.7 (10^{12})	8.6 (10^{11})	3.1 (10^{12})	1.4 (10^{12})
RidgeCV	3.9	34.9	20.7	44.2	70.1
ElasticNetCV	3.7	35.5	18.9	46.7	72.0
LarsCV	3.7	35.4	19.0	36.4	65.3
LassoCV	3.7	35.4	19.0	40.2	65.3
LassoLarsCV	3.7	35.4	19.0	40.2	65.3
OrthogonalMatchingPursuitCV	3.6	35.3	19.4	**34.4**	65.5
ARDRegression	3.8	33.9	19.6	40.8	66.3
BayesianRidge	3.6	35.2	19.0	44.8	72.4
HuberRegressor	**3.0**	**25.6**	**13.3**	35.6	**64.4**
RANSACRegressor	1.9 (10^{12})	58.4	2 (10^{10})	1 (10^{11})	1 (10^{11})
TheilSenRegressor	3.7	39.0	23.8	51.3	76.9
PoissonRegressor	3.6	33.6	19.6	44.8	71.8
TweedieRegressor	3.6	35.6	19.0	52.2	80.7
PassiveAggressiveRegressor	3.6	29.5	16.7	42.2	65.9

***written in bold**: the best mean absolute error

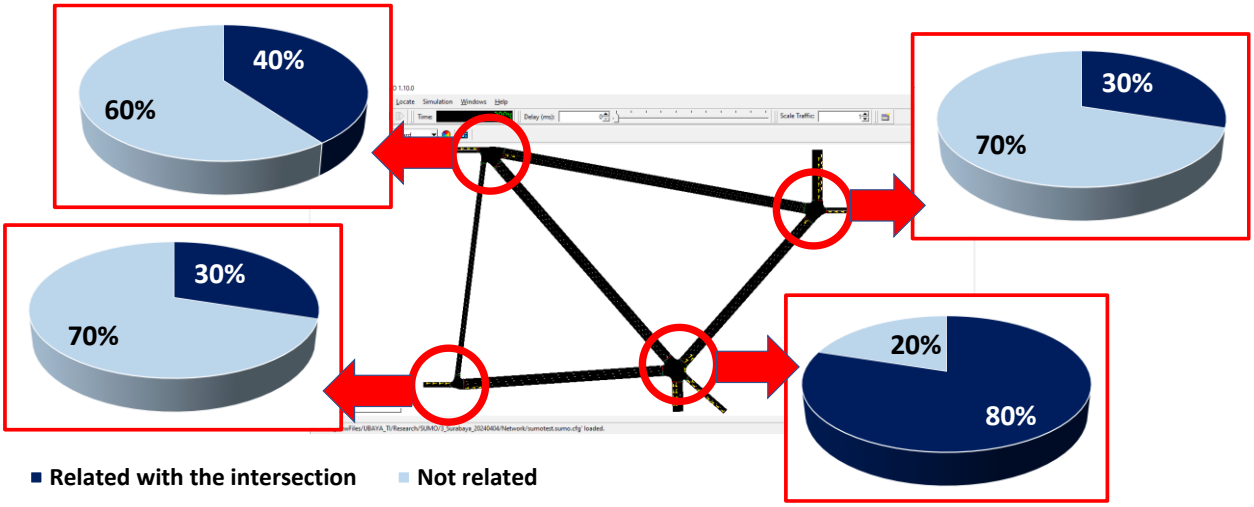

Figure 6: Ten most influential input data for each intersection.

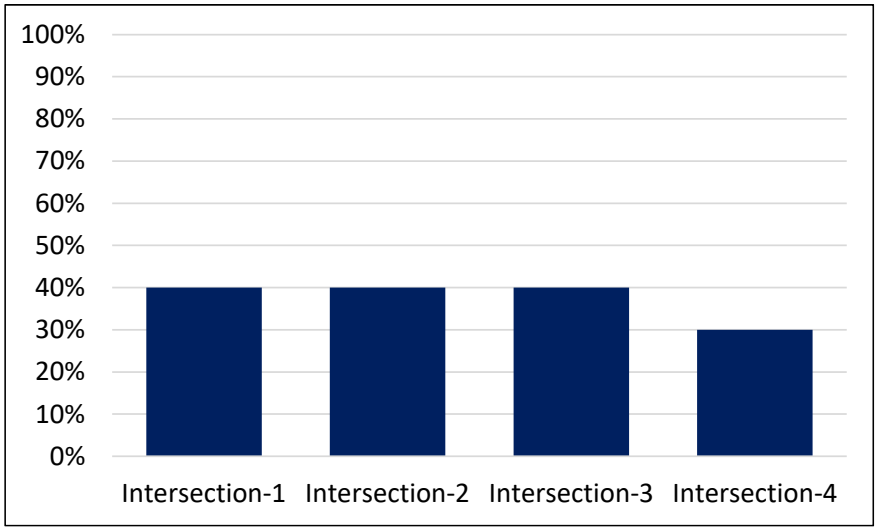

Figure 7: Effects of input data of any intersection on the predicted number of vehicles passing all intersections.

7 CONCLUSIONS

This study discussed how discrete event simulation and regression machine learning models could be used to understand the traffic of a road network. The numerical experiments showed how all parts of the road network are interconnected and the best regression machine learning models to conduct the prediction for all intersections. For further research, it is necessary to utilize the predicted behavior to make traffic management decisions and implement the proposed methodology for solving real cases.

REFERENCES

Arboretti, R., R. Ceccato, L. Pegoraro, and L. Salmaso. 2022. "Design of Experiments and Machine Learning for Product Innovation: A Systematic Literature Review". *Quality and Reliability Engineering International* 38:1131–1156.

Bakir, D. K. Moussaid, Z. Chiba, and N. Abghour. 2024. "A Comprehensive Review of Traffic Congestion Prediction Models: Machine Learning and Statistical Approaches". In *2015 IEEE International Conference on Computing (ICOCO)*, 539–545 https://doi.org/10.1109/ICOCO62848.2024.10928197.

Bonela, S. R., and B. R. Kadali. 2023. "Examining the Effect of Vehicle Type on Right-Turn Crossing Conflicts of Minor Road Traffic at Unsignalized T-Intersections". *IATSS Research* 47:545–556.
de Barros, O. M., C. L. Marte, C. A. Isler, L. R. Yoshioka, and E. S. da Fonseca Junior. 2023. "Spatial Matrices for Short-Term Traffic Forecasting Based on Time Series". *Latin American Transport Studies* 1:100007.
Dammak, B., F. Ciari, A. Joaua, and H. Naseri. 2025. "Investigating of Machine Learning's Capability in Enhancing Traffic Simulation Models". *Transportation Research Procedia* 82:1229–1243.
Dutta, P., S. Khatua, and S. Choudhury. 2023. "Fast move: A Prioritized Vehicle Rerouting Strategy in Smart City". *Vehicular Communications* 44:100666.
Eom, M., and B.-I. Kim. 2020. "The Traffic Signal Control Problem for Intersections: A Review". *European Transport Research Review* 12:50.
Gomes, B., J. Coelho, and H. Aidos. 2023. "A Survey on Traffic Flow Prediction and Classification". *Intelligent Systems with Applications* 20:200268.
Goyani, J., N. Gore, and S. Arkatkar. 2024. "Crossing Conflict Models for Urban Un-signalized T-intersections in India". *Transportation Letters* 16:829–837.
Groth, S. 2019. "Multimodal Divide: Reproduction of Transport Poverty in Smart Mobility Trends". *Transportation Research Part A* 125:56–71.
Han, G., X. Liu, Y. Han, X. Peng, and H. Wang. 2024. "CycLight: Learning Traffic Signal Cooperation with a Cycle-Level Strategy". *Expert Systems with Applications* 255:124543.
Kafy, M. A., S. I. Faisal, M. L. Rahman, R. Moni, H. Shanmuganathan, and D. M. Raza. 2024. "Traffic Congestion Prediction using Machine Learning". In *2024 11th International Conference on Computing for Sustainable Global Development (INDIACom)*, 1290–1295 https://doi.org/10.23919/INDIACom61295.2024.10498418.
Lee, S., S. C. Wong, and P. Varaiya. 2017. "Group-based Hierarchical Adaptive Traffic-Signal Control Part I: Formulation". *Transportation Research Part B: Methodological* 105:1–18.
Lin, P., Y. He, M. Pei, and R. Yang. 2022. "Data-driven Spatial-Temporal Analysis of Highway Traffic Volume Considering Weather and Festival Impacts". *Travel Behaviour and Society* 29:95–112.
Lin, Y., A. Tiwari, B. Fabien, X. Ban, and S. Devasia. 2023. "Increasing Traffic Capacity of Mixed Traffic at Signalized Traffic Intersections Using Delayed Self Reinforcement". *Transportation Research Part C* 157:104403.
Luitel, S., P. A. Shrestha, and H. Tiwari. 2025. "Travel Time Prediction for Two-Lane Two-Way Undivided Carriageway Road Section- A Case Study". *Transportation Research Interdisciplinary Perspectives* 31:101386.
Matković, K., D. Gračanin, M. Jelović, and H. Hauser. 2015. "Interactive Visual Analysis of Large Simulation Ensembles". In *2015 Winter Simulation Conference (WSC)*, 517–528 https://doi.org/10.1109/WSC.2015.7408192.
Min, W., and L. Wynter. 2011. "Real-time Road Traffic Prediction with Spatio-Temporal Correlations". *Transportation Research Part C* 19:606–616.
Mirchandani, P. and L. Head. 2001. "A Real-Time Traffic Signal Control System: Architecture, Algorithms, and Analysis". *Transportation Research Part C: Emerging Technologies* 9(6):415–432.
Qiu, Z., T. Zhu, Y. Jin, L. Sun, and B. Du. 2023. "A Graph Attention Fusion Network for Event-Driven Traffic Speed Prediction". *Information Sciences* 622:405–423.
scikit-learn. 2024. 1. Supervised learning. https://scikit-learn.org/stable/supervised_learning.html, accessed 13th April.
Singgih, I. K. 2021. "Production Flow Analysis in a Semiconductor Fab Using Machine Learning Techniques". *Processes* 9(3): 407.
SUMO. 2024. Simulation of Urban MObility. https://eclipse.dev/sumo/, accessed 12th April.
van Dijk, N. M., and E. van der Sluis. 2008. "Practical Optimization by OR and Simulation". *Simulation Modelling Practice and Theory* 16:1113–1122.
Weng, W., J. Fan, H. Wu, Y. Hu, H, Tian, F. Zhu, and J. Wu. 2023. "A Decomposition Dynamic Graph Convolutional Recurrent Network for Traffic Forecasting". *Pattern Recognition* 142:109670.
Xu, M., T. Z. Qiu, J. Fang, H. He, and H. Chen. 2023. "Signal-control Refined Dynamic Traffic Graph Model for Movement-Based Arterial Network Traffic Volume Prediction". *Expert Systems With Applications* 228:120393.
Zheng, G., W. K. Chai, J. Zhang, and V. Katos. 2023. "VDGCNeT: A Novel Network-wide Virtual Dynamic Graph Convolution Neural Network and Transformer-based Traffic Prediction Model". *Knowledge-Based Systems* 275:110676.

AUTHOR BIOGRAPHIES

IVAN KRISTIANTO SINGGIH is the Head of Systems Engineering Laboratory and Assistant Professor in the Study Program of Industrial Engineering at University of Surabaya (Ubaya), Indonesia. He received his B.S. and M.S. degrees in Industrial Engineering from Bandung Institute of Technology, Indonesia, and Ph.D. degree in Industrial Engineering from Pusan National University, Busan, Korea, in 2009, 2010, and 2017, respectively. He was a Postdoctoral Researcher and Research Assistant Professor under the Department of Industrial and Management Engineering at Pohang University of Science and Technology (POSTECH), and a Postdoctoral Researcher at Korea Advanced Institute of Science and Technology (KAIST) and Korea University. His research interests include operations research, smart systems, machine learning, and simulation. He is highly

interested in collaborations between academia, practitioners, government, etc. He received the Best Student Paper Award in IJIE 2013 Conference, Korean Government Scholarship Excellent Academic Achievement Award in 2014, and the Best Paper Award in Institute of Supply Chain and Logistics Indonesia National Seminar 2024. He is the Head of Transportation Group of Institute of Supply Chain and Logistics Indonesia (ISLI) and was the Director of Education and Expertise in The Association of Indonesian Researchers in South Korea (APIK). His email address is ivanksinggih@staff.ubaya.ac.id and his website is https://www.researchgate.net/profile/Ivan-Singgih.

MOSES LAKSONO SINGGIH is a Professor at the Department of Industrial and Systems Engineering, Institut Teknologi Sepuluh Nopember (ITS), Surabaya, Indonesia. He received his Bachelor's and Master's degree from Industrial Engineering Department, Institut Teknologi Bandung (ITB), Indonesia. He received Ph.D. from the University of Queensland, Australia. His research interests are productivity, quality, and smart manufacturing systems. Currently, he supervises postgraduate students with topics: a design for manufacturing and assembly (DFMA); quality management; lean six sigma; internet of things; sharing economy; circular economy and product-service systems. His email address is moseslsinggih@ie.its.ac.id and his publications can be found at https://www.researchgate.net/profile/Moses-Singgih and https://www.smartmoses.com/.

DANIEL NATHANIEL is an undergraduate student in the Study Program of Industrial Engineering at University of Surabaya (Ubaya), Indonesia. He is highly interested in Python-based optimization, e.g., genetic algorithms, image processing, etc. His email address is danieln0903@gmail.com.