Variable Speed Limit Control for Highway scenarios a Multi-Agent Reinforcement Learning Based Appraoch

Bálint Kővári, István Knáb and Tamás Bécsi

May 21, 2024

Variable Speed Limit Control for Highway scenarios a Multi-Agent Reinforcement Learning based approach

1st Bálint Kővári
Department of Control for Transportation and Vehicle Systems
Budapest University of Technology and Economics
Budapest, Hungary
kovari.balint@kjk.bme.hu

2nd István Gellért Knáb
Systems and Control Laboratory
Institute for Computer Science and Control
Budapest, Hungary
istvan.knab@edu.bme.hu

3rd Tamás Bécsi
Department of Control for Transportation and Vehicle Systems
Budapest University of Technology and Economics
Budapest, Hungary
becsi.tamas@kjk.bme.hu

***Abstract*—Modern road networks are critical in developing transportation infrastructures from the aspect of sustainability, thanks to the rapid increase in road users. The demand for mobility makes the existing infrastructure more crowded, boosting greenhouse gas emissions and delays in everyday commuting. Expanding the road network is only possible in some cases and is also not feasible, but Intelligent Transportation Systems (ITS) can enhance the efficiency of the existing transportation network. From a management point of view, a proven algorithm is Variable Speed Limit Control, which regulates the state of certain sectors by spatially distributing traffic in the form of dynamically varying speed limits. Combining this with a state-of-the-art predictive solution can make a big difference to performance. For the design of speed limits, this paper proposes an approach where deep reinforcement learning with the smallest industrial share not only resolves moving jams that arise during congested traffic situations, but also prevents them, thereby avoiding cumulative error from transients, all by abandoning physical equations and identified models.**

***Index Terms*—Reinforcement Learning, shockwave, traffic control, variable speed limit**

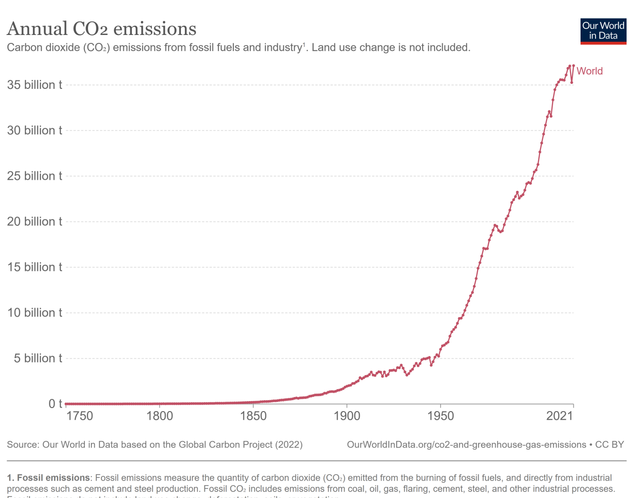

Fig. 1. Global CO2 emission [1]

I. Introduction

In the context of freeways, situations often occur, particularly during periods of heavy congestion, where the inability to effectively regulate the flow of traffic results in disruptions to the smooth movement of vehicles. Analogous to principles found in fluid dynamics, certain laws of physics must be adhered to. One such fundamental principle is the continuity theorem [2], which asserts that the cumulative rate of incoming vehicles across all lanes in the subsequent segment must be equal to the rate of outgoing vehicles across the entire cross-section of the present cell. This principle can be mathematically represented by the following equation:

$$\sum_{i=1}^{n} Q_{in_i} = \sum_{j=1}^{m} Q_{out_j}, \tag{1}$$

where n is the number of lanes on the next road segment and m means the amount of traffic lanes on the actual cell.

Active bottlenecks, which can arise due to factors such as construction, highway onramps, closed lanes due to accidents, or sections where high-speed roads intersect with urban transportation, have the potential to push a traffic network to its physical limits. In such cases, even a minor adjustment can lead to significant fluctuations in the speed profiles of individual vehicles, owing to the additional delays caused by reaction times. In the macroscopic planning of traffic control, the primary objective is to minimize changes in acceleration and prevent strongly correlated transients. This effect is particularly pronounced in highway settings compared to urban traffic situations due to the higher default maximum speed, resulting in a wider range of selectable speeds for vehicle movement. The reaction time of a single vehicle can be described by the following formula:

$$\tau = \sum_{i=1}^{n} \tau_i, \tag{2}$$

where τ marks the reaction time of the actual vehicle and n is the size of the column, which has effect on the investigated vehicle. The integration of autonomous vehicles (V2V - Vehicle to Vehicle and V2I - Vehicle to Infrastructure communication) [3] has the potential to address the aforementioned challenges by providing additional environmental information and the current state of a greater number of traffic participants. This is made possible through the utilization of sensors equipped on each vehicle and their fusion [4]. Consequently, this integration can lead to a more deterministic traffic behavior, as highlighted in [5].

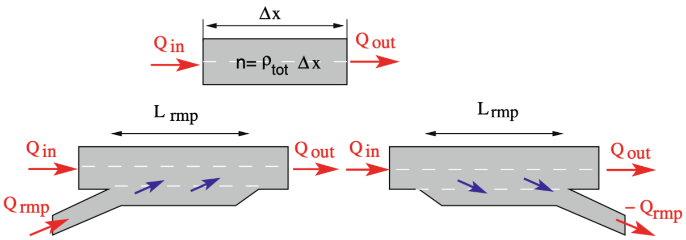

Fig. 2. Continuity in traffic flow application [2]

However, it is important to recognize that human limitations extend to both the vehicles themselves and the surrounding spatial objects. Due to the aforementioned reaction times, there is a need to align vehicle speeds with the maximum achievable velocity that ensures accident-free conditions. By doing so, the impact of deceleration is limited to only a small number of vehicles, allowing for local resolution through a reduction in the following distance via marginal adjustments.

A. Related Work

Variable Speed Limit Control (VSLC) is a method used to optimize the speed limits of specific segments within a network, taking into account the examination of both static and dynamic objects and their environmental effects on traffic flow, as explored in [6].

The fundamental concept behind this engineering solution is to proactively regulate the decision-making space available to vehicles, aiming to prevent the formation of traffic congestion. By minimizing the length of standing vehicle queues, the occurrence and propagation of shock waves can be reduced. This approach is particularly valuable in heavily congested areas and their surrounding environment, as it contributes to the reduction of fuel consumption and emissions, as highlighted in the studies by [7] and [8].

The algorithm operates by assigning certain road sections with the ability to independently regulate speed limits, aiming to influence the outflow of the active segment. The determination of the inflow into the subsequent cell is governed by the continuity law, as discussed in [9].

To describe specific road sections, similar physical indicators are employed, drawing parallels to concepts found in hydrodynamics. In accordance with the SPECIALIST article by [10], [2], states of congestion evolution and resolution are characterized using metrics such as mean speed and density within the cells under consideration.

Based on the analysis, it has been determined that the aforementioned values will be utilized to describe the current system under examination, which will be further defined in the subsequent explanation of the state representation.

The SPECIALIST algorithm, mentioned earlier, operates using four phases and six states to characterize the prevailing traffic conditions. On the other hand, the MCS algorithm, as discussed in [6], is currently implemented on Swedish and Dutch highways and makes decisions based solely on the mean speed. If the mean speed falls below 45 km/h, it imposes additional restrictions on inflow sections.

However, a key issue with these algorithms is their reliance on reacting to existing traffic conflicts without incorporating any predictive capabilities. They fail to identify the phase in which preventive measures should be taken. In contrast, the proposed solution aims to proactively address congestion by leveraging data-driven decision-making. By looking ahead and taking anticipatory actions, this approach not only resolves congestion but also mitigates its evolution.

II. Environment

A. Examined road section

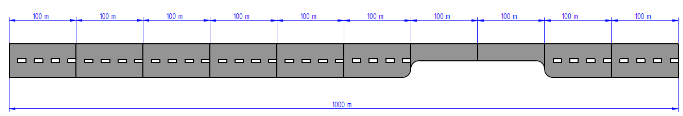

Fig. 3. Examined road section

Due to the absence of trajectory planning in the task, where vehicles follow predefined routes and randomization is implemented on the lanes they are generated on, a macroscopic model is deemed sufficient for constructing an environment suitable for applying a Reinforcement Learning-based controller. Unlike the microscopic model, which focuses on individual vehicles, the macroscopic model operates at the level of traffic flow as the basic unit. To facilitate this, the simulation is implemented in SUMO (Simulation of Urban MObility) [11], an open-source simulator that provides the necessary measurement values to describe the flow.

As depicted in Figure 3, the environment consists of a two-lane straight highway section with 100-meter-long edges. At the bottleneck location, a closed lane is modeled, extending over a 200-meter-long section. By setting up the network, the road sections are separated from each other, allowing them to be treated as distinct entities and utilized as agents in subsequent applications.

B. Physical limitations

In order to accurately model the conditions of a real traffic network, several factors needed to be considered. These include permitted lane changes, the minimal time interval for decision-making, speed limits, and the flow of vehicles.

To account for permitted lane changes, the SUMO network configuration file allows for specifying the rules regarding lane changing behavior. Additionally, a minimal time interval of 10 seconds was selected on each road segment to represent the legal decision-making timeframe. This was implemented as a state machine, with the choice between idle and active states determined by monitoring the overflow of a counter. The timing modes of microcontroller units, such as the "Clear Timer on Compare Match," were utilized for this purpose [12].

Speed limits in the simulation can only be displayed on real traffic signs to enable drivers to easily adhere to them. To ensure realistic transitions, speed changes were limited to a maximum of 10 km/h, resulting in the selection of velocity stairs with this value as the increment. The range of allowed speed limits was carefully considered, taking into account the impact of cross-sectional narrowings. For instance, if the number of lanes was halved, the allowed speed would need to be doubled. The chosen upper limit for speed was the Hungarian maximal value of 130 km/h, which is widely used in many European countries. The lower limit was determined empirically as 30 km/h, serving as a reasonable threshold.

Another crucial aspect was determining the flow of vehicles in the network. To create an environment that reflects real-world traffic conditions, a controlless traffic scenario was used as a baseline for comparison. The flow rate was set to the lower threshold of a shock wave, as described in the SPECIALIST model (1500 vehicles/hour/lane) [6].

III. Reinforcement Learning

Reinforcement Learning (RL) is a data-driven algorithm that uses experience to make decisions based on rewards or penalties. RL involves an Agent interacting with an Environment, where the Agent receives feedback in the form of the current state and a reward. By combining trial-and-error learning with optimal behavior, RL enables the Agent to learn a policy for making decisions in specific environmental conditions. This is achieved through Q-learning [13], which assigns a quality value (Q-value) to state-action pairs based on a finite horizon of actions. In our case, the RL model is trained by determining the weights and biases of a neural network, offering computational efficiency compared to a lookup-table approach. This allows for real-time solving of complex problems without exponential increases in computational demand. The Q-value of an action is influenced by previous rewards, weighted by a discount factor (gamma), and the learning rate (alpha). The Bellman equation is used to derive the Q-value, ensuring optimal behavior. The extent of discounting and learning rate greatly impacts the performance of the RL model. This update equation is further described in the epsilon-greedy policy, which determines the exploration-exploitation trade-off. Please note that the specific form of the update equation is not provided in the text.

$$Q(s,a) = Q(s,a) + \alpha(r + \gamma max_{a'} Q(s,a') - Q(s,a)) \quad (3)$$

where r means the reward and the $Q(s,a)$ is the computed action quality, that is compared to the estimated quality indicator (Q_{pred}), which comes from a prediction based on choosing random samples from the teaching data and results the mean-squared error to compute loss.

$$L_{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Q_{pred} - Q(s,a))^2 \quad (4)$$

RL algorithms require training data obtained from random actions and their evaluation to differentiate positive and negative experiences. The ϵ-greedy policy is used to balance exploration and exploitation. By comparing a decreasing scalar value ϵ to a random value between 0 and 1, the agent determines whether to take random exploratory actions or choose the currently perceived optimal actions. This policy is crucial when the agent lacks prior experiences and aims to gather information while gradually focusing on exploiting the best opportunities. Learning is an iterative process, where

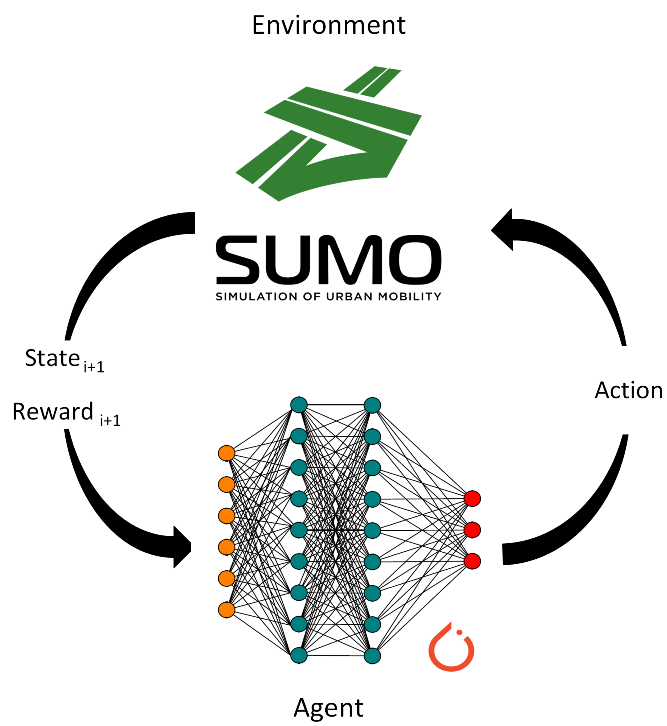

Fig. 4. Reinforcement Learning in use

the experiences are stored in a buffer in form of $(s,a,r,s^{'})$ and as above mentioned some data will randomly choosen in order to calculate stochastic gradient descent [14].

As shown on "Fig. 4", the training loop is implemented as the above mentioned SUMO environment which communicates with the python gym environment through the TraCI interface, and the agent is a Deep Q Network, which has densities and velocities of certain road segments as input and has a 3 element vector as output, containing 10 km/h increase and decrease of enabled velocity and as third action let the current state active also in the next training episode.

While there have been RL algorithms implemented for onramp traffic control using variable speed limits on specific sections of the network, the performance of these algorithms depends heavily on how the state vector and reward are defined. Even with a fixed output space, the selection of appropriate abstractions and representations is crucial. One example of controlling onramps using RL is described in the work by [15], where three density parameters were utilized as input variables for the algorithm.

It is important to note that the success of RL algorithms in onramp traffic control relies on carefully designing the state representation and reward system, taking into account the specific dynamics and objectives of the traffic system under consideration.

- **MARL:** To ensure the resolution of shock waves throughout the entire network, it is divided into sections that can independently take actions. In the implemented pseudocode, each agent operates in every training episode without knowledge of the actions taken by other agents. Experiences are stored in a buffer for learning. It is important to note that all agents share the same neural network, enabling a self-play approach where agents contribute to environmental changes without explicit cooperation or competition.
  In our case, the agents have a common goal, and their actions are guided by the same reward function. This framework allows for an indirect form of cooperative behavior, where agents optimize their behavior with single actions, leading to an optimized state for the entire network. This approach aligns with the concept of identical payoff in multi-agent games, where agents share the same reward function [16].
  By designing the system in this way, the network does not need to be aware of which section is the active agent. This flexibility allows for easy expansion of the model to incorporate additional sections without affecting the state representation.
- **State:** The main metrics of flow description are density and average speed, which determine how may vehicles are abiding at the same time in a certain road section, and provides feedback of the inhibition of the flow.

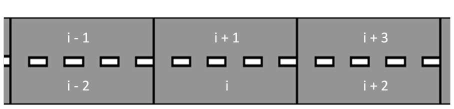

Fig. 5. State representation

A key distinction between this multi-agent solution and recently implemented algorithms is the absence of predefined thresholds to activate specific states. In this approach, the neural network receives a 10-element vector

Algorithm 2: DQN-algorithm [17] extended on MARL

```
Initialize replay memory D to capacity N
Initialize action-value function Q with random weights
  θ
Initialize target action-value function Q̂ with weights
  θ⁻ = θ
for episode = 1, M do
    state_t ←reset()
    done ←FALSE
    reward_t ←0
    for time step = 1, T do
        for agents = 1, A do
            With probability ϵ select a random action
            otherwise select a_t = argmax_a Q(s,t)
            state_{t+1}, reward_t, done← step(a_t)
            Store transition (s_t, a_t, s_{t+1}, r_t, d_t) in D
            Sample random minibatch of transitions
              (s_t, a_t, s_{t+1}, r_t, d_t) from D
            s_t = s_{t+1}
        end
        if d' = TRUE then
            y_j =r'_j
        end
        else
            y_j =r'_j + γ · max_a Q̂(s'_{t+1}, θ̂)
        end
        Perform a gradient descent step on
          (y_j − Q(s_t, a_j, θ))² with respect to the
          network parameters θ
        Every C steps reset Q̂ = Q
    end
end
```

consisting of average velocity and density data from the active agent and its two neighboring cells in both directions. The decision to include two cells in each neighboring direction is based on the identification numbering scheme shown in Fig. 5 (referring to the figure in the document). The inclusion of measures from the section before and after the lane jump ensures that the neighboring cell is shifted by two positions.

$$state_i = \begin{bmatrix} velocity_{i-2} \\ density_{i-2} \\ . \\ . \\ . \\ velocity_{i+2} \\ density_{i+2} \end{bmatrix} \quad (5)$$

It is evident that modifying speed limits on the first and last sections cannot be represented in the same way as the intermediate nodes. This is because these sections do not have neighbors on both sides. Additionally, their impact on the network is not as significant as the sections in the middle. In the case of the last road section, there is no cross-sectional change, so the maximum speed that can be achieved is already the optimal speed at that location. Therefore, no actuation is necessary to empty that zone. Similarly, the first section is not suitable for actuation because the inflow rate is determined by the spawning of vehicles directly in that area. Hence, these restricted sections are not considered for actuation during training and testing. However, their metrics will still be computed for the purpose of creating the state representation.

- **Action:** As mentioned earlier, the action space consists of a three-element vector representing speed increase, speed decrease, and idle mode. It is essential to enforce environmental constraints such as minimum actuation time and predefined maximum speed according to traffic regulations. These constraints ensure that VSLC is implemented in a manner that prioritizes safety. Without these constraints, the speed in the bottleneck could exceed the maximum allowable speed, which contradicts the goal of controlled regulation.

$$action = \begin{bmatrix} +10 & km/h \\ 0 & \\ -10 & km/h \end{bmatrix} \tag{6}$$

In order to determine the level of velocity perception, real traffic situations were used as a baseline. Considering that the segments are 100 meters long, it is important to actuate as quickly as possible, similar to human processing time. To ensure a smooth transition between speeds, the speed differences will be set at 10 km/h, as mentioned earlier, for both increasing and decreasing speed limitations. This value is chosen based on the scale of speed indicators in personal cars and the commonly used motorway speed signs. Additionally, a constraint is implemented to avoid speed limits below 30 km/h in any section, to prevent the neural network from finding a local optimum by sacrificing one lane to achieve maximum speed in the other.

- **Reward:** When defining the reward function, it is important to focus on the achievable task at hand. In this case, the task is to minimize waiting time and maximize the average speed that can be realistically achieved. Therefore, the reward function should be designed to incentivize actions that lead to reduced waiting time and increased average speed on the road network.

$$R = \frac{v_{avg}}{1+w}, \tag{7}$$

The initial implementation of the reward function did not yield satisfactory results as the network converged to a local optimum where the outer lane was stopped to achieve maximum speed in the inner lane. To address this issue, the weights between the two variables in the reward function were reconsidered. The denominator was squared to increase its influence on the result, and a lower bound for speed limitations was implemented. The revised reward function is as follows:

$$R = \frac{v_{avg}}{(1+w)^2}, \tag{8}$$

where summarized latency on the network is described with w, and the average speed of vehicles on the whole network is represented with v_{avg}. Through a quick dimension, analysis is definable, that the output dimension is $[\frac{m}{s^3}]$, which equals the metrics of the above-mentioned third derivative of displacement by time, that equals the change in acceleration which is above defined as realizable goal parameter.

- **Network architecture:** Considering the nature of the task, it would be appropriate to use a convolutional neural network, because from the point of view of implementation, a sliding window-type problem must be implemented. Nevertheless, the model was applied to an MLP network, which contains fully connected layers, in the size 10 - 64 - 16 - 3 to try, if it can solve the problem.

IV. Results

A. Computation of emission

As shown in “Fig. 6, 7, 8”, the metrics has improved an order of magnitude compared to uncontrolled speed limits. The correlation between the individual trends (marginal, maximal 3% discrepancy) is due to the fact that each emission was calculated using the same model in SUMO (HBEFA v2.1) [18] along the following equation, only assuming the the scalar multipliers:

$$E = c_0 + c_1 va + c_2 va^2 + c_3 v + c_4 v^2 + c_5 v^3, \tag{9}$$

where E is the emission in $[\frac{mg}{s}]$ and c_i values are constants depending on the target emission and vehicle type.
Data was collected from the entire network and plotted to gain a comprehensive understanding of the long-term trend. The significant decrease observed in the data can be attributed to reduced travel time, which has an indirect impact on emissions calculations.

B. Comparison

The model demonstrates a significant reduction in greenhouse gas emissions, ranging from 24% to 28%. This reduction is particularly crucial in highway and suburban environments, where air pollution is a major concern. By minimizing time spent in reactive traffic and reducing standstill situations, the model contributes to creating a more vibrant and environmentally friendly city.

TABLE I
Performance

	Emissions in 10 Episodes of test		
	HC [kg]	***CO2 [kg]***	***NOx [kg]***
Fix speed limit	0.048359	302.90999	0.12805
RL model	0.0365435	220.047116	0.09402
Emission reduction [%]	24,43	27,36	26,58

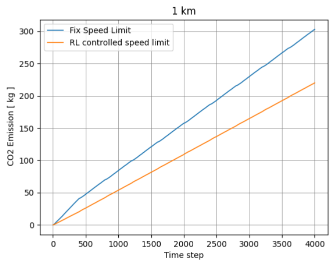

Fig. 6. CO2

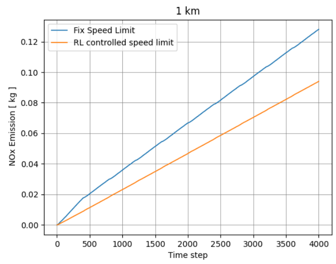

Fig. 7. NOx

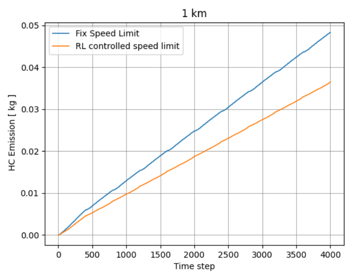

Fig. 8. HC

V. Conclusion

The demonstrated success of Multi-Agent Reinforcement Learning (MARL) in solving the traffic control problem shows the potential for further improvements using more complex architectures, such as convolutional neural networks. This type of software support can play a significant role in reducing air pollution in densely populated and high-traffic areas by classifying vehicles and controlling their access to certain zones.

Future research will involve comparing the MARL approach with existing threshold algorithms to determine its effectiveness in highway decision-making and evaluate the significance of not only resolution but also prevention, which is one of the main benefits of machine learning algorithms, thus underlining the relevance of this field of research. It will also explore other network architectures to further enhance performance. One important aspect is defining the domain of the Variable Speed Limit Control (VSLC) area and filtering out non-performance factors to minimize sensor requirements while maintaining performance and safety [19], the need for which increases with the volume of traffic [20].

Additional research directions include implementing cross-sectional narrowings and studying the impact of optimal distances between onramps or constructions in the environment. These investigations will contribute to a more comprehensive understanding of the potential applications and benefits of MARL in traffic control.

Acknowledgment

The research was supported by the European Union within the framework of the National Laboratory for Autonomous Systems. (RRF-2.3.1-21-2022-00002)
Supported by the ÚNKP-22-5-BME-272 New National Excellence Program of the Ministry for Culture and Innovation from the source of the National Research, Development and Innovation Fund

References

[1] H. Ritchie, M. Roser, and P. Rosado, “Co and greenhouse gas emissions,” *Our World in Data*, 2020, https://ourworldindata.org/co2-and-greenhouse-gas-emissions.

[2] M. Treiber and A. Kesting, “Traffic flow dynamics,” *Traffic Flow Dynamics: Data, Models and Simulation, Springer-Verlag Berlin Heidelberg*, pp. 983–1000, 2013.

[3] D. W. Matolak, “V2v communication channels: State of knowledge, new results, and what’s next,” in *Communication Technologies for Vehicles: 5th International Workshop, Nets4Cars/Nets4Trains 2013, Villeneuve d’Ascq, France, May 14-15, 2013. Proceedings 5*. Springer, 2013, pp. 1–21.

[4] J. Ližbetin, J. Pečman, J. Hanzl, and O. Stopka, “Modeling the intelligent transport systems elements functionality testing plan: A case study,” *Communications - Scientific letters of the University of Zilina*, vol. 24, 09 2022.

[5] F. Vrbanić, E. Ivanjko, S. Mandžuka, and M. Miletić, “Reinforcement learning based variable speed limit control for mixed traffic flows,” in *2021 29th Mediterranean Conference on Control and Automation (MED)*. IEEE, 2021, pp. 560–565.

[6] E. F. Grumert, A. Tapani, and X. Ma, “Characteristics of variable speed limit systems,” *European transport research review*, vol. 10, no. 2, pp. 1–12, 2018.

[7] B. Kővári, L. Szőke, T. Bécsi, S. Aradi, and P. Gáspár, “Traffic signal control via reinforcement learning for reducing global vehicle emission,” *Sustainability*, vol. 13, no. 20, p. 11254, 2021.

[8] M. Loman, B. Šarkan, and T. Skrúcanỳ, “Influence of cruise control use on vehicle´ s consumption,” *LOGI–Scientific Journal on Transport and Logistics*, vol. 12, no. 1, pp. 13–24, 2021.

[9] B. Khondaker and L. Kattan, “Variable speed limit: an overview,” *Transportation Letters*, vol. 7, no. 5, pp. 264–278, 2015.

[10] A. Hegyi, S. P. Hoogendoorn, M. Schreuder, H. Stoelhorst, and F. Viti, “Specialist: A dynamic speed limit control algorithm based on shock wave theory,” in *2008 11th international ieee conference on intelligent transportation systems*. IEEE, 2008, pp. 827–832.

[11] M. Behrisch, L. Bieker, J. Erdmann, and D. Krajzewicz, “Sumo–simulation of urban mobility: an overview,” in *Proceedings of SIMUL 2011, The Third International Conference on Advances in System Simulation*. ThinkMind, 2011.

[12] D. Camera, “Newbie’s guide to avr timers,” 2012.

[13] C. J. Watkins and P. Dayan, “Q-learning,” *Machine learning*, vol. 8, pp. 279–292, 1992.

[14] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[15] Z. Li, P. Liu, C. Xu, H. Duan, and W. Wang, “Reinforcement learning-based variable speed limit control strategy to reduce traffic congestion at freeway recurrent bottlenecks,” *IEEE transactions on intelligent transportation systems*, vol. 18, no. 11, pp. 3204–3217, 2017.

[16] L. Buşoniu, R. Babuška, and B. De Schutter, “Multi-agent reinforcement learning: An overview,” *Innovations in multi-agent systems and applications-1*, pp. 183–221, 2010.

[17] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, “Human-level control through deep reinforcement learning,” *nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[18] D. Krajzewicz, S. Hausberger, P. Wagner, M. Behrisch, and M. Krumnow, “Second generation of pollutant emission models for sumo,” in *SUMO2014 - Second SUMO User Conference*, ser. Reports of the DLR-Institute of Transportation Systems, May 2014. [Online]. Available: https://elib.dlr.de/89398/

[19] S. Yadav and R. Rishi, “An optimal security framework based on driver authentication for intelligent transportation systems,” *LOGI - Scientific Journal on Transport and Logistics*, vol. 14, pp. 24–34, 01 2023.

[20] J. Hudec, B. Šarkan, J. Caban, and O. Stopka, “The impact of driving schools’ training on fatal traffic accidents in the slovak republic,” *Zeszyty Naukowe. Transport/Politechnika Śląska*, 2021.