Inverse Reinforcement Learning for Variable Speed Limit Systems

Josephine Grau*[1], Robin Abohariri[2], Alexander Roocroft[2], Marco Rinaldi[2], Hans van Lint[2], Marvin Baumann[1], and Peter Vortisch[1]

[1]Institute for Transport Studies, KIT, Germany
[2]Department of Transport and Planning, TU Delft, Netherlands

Short summary

This study investigates the use of Maximum Entropy Inverse Reinforcement Learning (IRL) to infer the underlying objective functions of Variable Speed Limit (VSL) systems. The objectives of such systems are often unknown or not explicitly accessible, limiting systematic analysis, comparison, and further development. Using a controlled simulation environment with a Model Predictive Control-based upstream speed reduction scheme, we evaluate whether IRL can recover the underlying objective from observed control behavior. Different reward function formulations are considered to assess the approach's robustness. The results show that IRL achieves high policy imitation accuracies across all tested settings. While the control policy can be recovered from only observed traffic-state features, incorporating domain knowledge via engineered features improves the interpretability and accuracy of the learned reward structure. Overall, the findings demonstrate that IRL is a promising tool for uncovering implicit traffic control objectives, enabling systematic analysis and simulation of VSL strategies.

Keywords: Inverse reinforcement learning, traffic management, freeway control, variable speed limits

1 Introduction

Variable Speed Limit (VSL) systems are widely used to improve traffic safety and efficiency by dynamically adjusting speed limits based on prevailing traffic conditions. Traditional VSL implementations, such as the rule-based MARZ systems in Germany (Bundesministerium für Verkehr und Infrastruktur (Pub.), 2018), rely on predefined thresholds for traffic flow or speed to trigger speed changes. Model Predictive Control (MPC) approaches treat VSL as a predictive optimization problem (Hegyi et al., 2005; Frejo et al., 2014). MPC controllers forecast traffic evolution and proactively adjust speed limits according to a predefined objective function. Despite their potential, MPC-based VSL schemes are rarely implemented in practice due to their complexity and computational demands. Recent advances leverage Reinforcement Learning (RL) to enable data-driven, adaptive VSL control. Multi-agent frameworks, such as MARVEL, learn coordinated policies for multiple VSL controllers, optimizing safety and mobility without requiring an explicit traffic model (Zhang et al., 2024).

For simulation studies, a realistic replication of VSL behavior is often necessary to evaluate traffic scenarios or assess new control strategies. However, reproducing field-installed control logic is challenging. First, the algorithms often evolve over years, shaped by local adaptations, parameter tuning, and operational experience, and are rarely fully documented. Second, such models are usually tied to specific freeway sections and system designs, making adaptation to new contexts labor-intensive. For researchers who primarily require VSL to ensure realistic background traffic, the high cost and complexity of developing accurate replicas often result in oversimplified or omitted VSL behavior.

Supervised learning can partially bridge this gap by learning policies directly from observed behavior, as we showed in previous studies (Grau et al., 2026). However, in many cases, a deeper understanding of the underlying objectives is also desirable. Even when a VSL algorithm is known, its threshold-based logic may only indirectly reflect its operational goals. Without knowledge of the existing objective, improvement strategies can conflict with embedded priorities, and benchmarking or comparative studies become problematic.

This motivates the central question of this work: Can we infer the implicit objective function of a transportation decision system solely from its observed behavior? To address this question, we use Inverse Reinforcement Learning (IRL). While RL models decision-making as the maximization of a cumulative reward, IRL seeks to recover the reward function that explains an observed expert policy. Foundational work by Ng & Russell (2000) introduced IRL as the problem of learning reward functions from demonstrations, and the Maximum Entropy IRL formulation (Ziebart et al., 2008) resolves reward ambiguity by selecting the most unbiased distribution over trajectories consistent with expert behavior.
IRL has been successfully applied in different fields, e.g., robotics (Argall et al., 2009), where movements of a robot are learned from human demonstrations. In transportation research, IRL has recently been applied to traffic signal control, for example, in Gu et al. (2023), where Relative Entropy IRL is used to reconstruct an expert's latent reward to then train an RL agent to optimize the controller.
In this study, we use IRL primarily to investigate VSL systems. We generate ground-truth traffic simulations with a control algorithm based on an explicit objective function. Recognizing that real-world rule-based algorithms often encode latent, complex multi-objectives, we employ a simplified control policy that reduces upstream speed near congestion tails. The resulting system outputs are used as demonstrations for IRL, which learns a reward function that represents the implicit objectives. We then evaluate the IRL output in terms of i) policy imitation accuracy to determine how closely the inferred policy reproduces the original actions and ii) interpretability, assessing whether the recovered reward function provides insight into the original control strategy.
This work lays the foundation for extended studies in the field of policy reconstruction. We demonstrate the feasibility of using Maximum Entropy IRL to recover the latent objectives of a VSL control system from observed behavior. This framework can be extended to support the use of IRL in transportation, enabling the evaluation, replication, and improvement of traffic management strategies without requiring full access to the underlying control algorithms.

2 Methodology

Inverse Reinforcement Learning

In this study, we make use of the Maximum Entropy Inverse Reinforcement Learning (IRL) as described by (Ziebart et al., 2008). In Reinforcement Learning (RL), an agent learns a policy by maximizing a reward function R that depends on the agent's state s and the action a taken. The reward is commonly modeled as a linear combination of functions of state features $\phi(s)$:

$$R(s, a) = \theta_a^T \phi(s). \tag{1}$$

In our case, the function ϕ incorporates feature engineering and min-max scaling to ensure comparability.

In IRL, the reward function is unknown and must be inferred from observed state-action pairs, typically provided as expert trajectories. This approach is particularly useful in settings where exact decision-making rules are unknown, but an underlying objective can be assumed. Given a set of feature functions $\phi(s)$, the task is to find suitable weights θ such that the resulting reward function induces a policy that reproduces the observed behavior.
As several reward functions may lead to the same policy, Ziebart et al. (2008) uses the principle of maximum entropy, selecting the distribution over trajectories that matches the observed behavior while making the least additional assumptions. The parameters θ are then estimated using maximum likelihood.
The evaluation of the IRL is based on policy imitation accuracy, comparing the output of the learned policy induced by the reward function with the ground-truth policy of the test set.

Experimental Setup

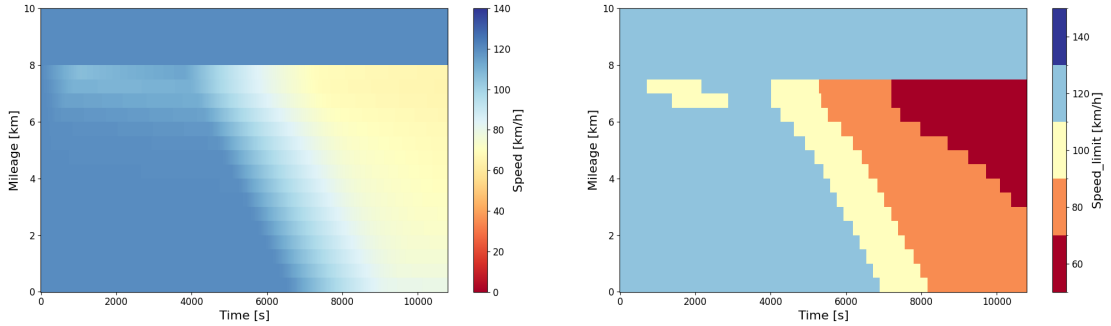
The aim of this exploratory study is to evaluate how accurately an objective function can be recovered using IRL. For this purpose, a known ground truth objective function is required. Model Predictive Control (MPC) is employed, as it explicitly relies on an objective function to determine its control policy. To generate data, a CTM simulation is combined with an MPC control loop that optimizes a predefined objective function. This setup allows full control over the underlying

objective and ensures that the true optimization target is known. On the other hand, the training data may still prove noisy, as MPC does not always find the optimal solution due to computational constraints and the fixed control interval of 1 min.
For this initial study, a VSL scheme that reduces speeds upstream of congestion tails for safety reasons, akin to that currently operational on a majority of Dutch motorways, is used. For N cross-sections with collocated speed measurements v and speed limit displays vsl, the operational objective function is the root mean-squared error between downstream measurements and VSL, given by

$$\min_{vsl} \sqrt{\frac{1}{N}\left(\sum_{i=1}^{N-1}\left(vsl_i - v_{i+1}\right)^2 + \left(vsl_N - v_N\right)^2\right)}. \tag{2}$$

Simulation setup

The traffic dynamics in this study are based on the classical Cell Transmission Model (CTM) introduced by Daganzo (1995). The triangular fundamental diagram is parametrized by a free-flow speed of 130 km/h, a maximum (backwards) congestion wave speed of 15 km/h, a critical density of 18 veh/km/lane, and a jam density of 150 veh/km/lane. A minor extension of the original formulation of the CTM is employed. In addition to density, cell-wise speeds are tracked and acceleration is explicitly bounded to ensure physically realistic vehicle behavior and prevent unrealistic speed adjustments.
The baseline model considers a homogeneous, single-lane freeway discretized into 20 cells of length 0.5 km with an on-ramp at cell 15. The simulation is conducted over a total duration of 3 hours, with hourly-varying upstream demand (1500–1800–1500 veh/h) and on-ramp inflow (300–600–300 veh/h) for varying traffic conditions. A simulation time step of 10 seconds and a control update interval of 1 minute are used. The MPC scheme predicts the system's evolution over a 3-minute horizon. Figure 1 shows the speed evolution and the corresponding speed limits of the above-defined scheme.



(a) Space-time diagram of Speeds

(b) Space-time diagram of VSL

Figure 1: CTM space-time diagrams

3 Results and discussion

We evaluated the policy imitation accuracy and the objective fulfillment (RMSE) using Equation 2 for the learned policy. This not only answers whether a policy learned from the training data mimics it, but also whether it follows the underlying objective function. The feature space consists of traffic features and engineered features depending on the action, i.e., the chosen speed limit. To determine a VSL action at a certain cross-section ($cs0$), speed (v) and density (k) of past timesteps ($-5t, \ldots, -1t$) and neighboring cross-section (upstream $-2cs$, $-1cs$, current $cs0$ and downstream $1cs$, $2cs$) are considered. This results in $2 \times 5 \times 5 = 50$ features (or $2 \times 1 \times 5 = 10$ when only the last timestep is used). Timesteps correspond to simulation timesteps at 10-second intervals here. Additionally, we introduced a VSL Tracking Error (VTE) feature, defined as the squared difference between the imposed speed limit and the observed speed at a cross-section downstream (feature "$-1t_1cs_v$"), which closely aligns with the VSL scheme (see Equation 2).
We evaluated the robustness of the IRL approach by testing different configurations of reward

function features. These feature sets are i) all available features (54 features), ii) the speed and denisty traffic features without the added VTE features (50 features), iii) speed and density features of the last timestep (ts) only (10 features), iv) features of iii) together with VTE features (14 features) and v) the VTE features only (4 features). Table 1 summarizes the policy imitation accuracy and objective fulfillment, measured by RMSE (Equation 2).

Features	**Policy imitation acc**	**RMSE y_{test}**	**RMSE learned policy**
i) all features	0.9875	4.8815	4.8902
ii) traffic features	0.9669	4.8815	5.6881
iii) last ts features	0.9778	4.8815	5.1223
iv) last ts + VTE	0.9963	4.8815	4.8804
v) VTE features	0.9954	4.8815	4.8787
no speed limit	1.0000	35.7050	35.7050

Table 1: IRL-based VSL policy reconstruction: feature set comparison with reference baseline "no speed limit"

Across all feature configurations, the policy imitation accuracy remains consistently high, exceeding 96%. The best performance is achieved with last timestep features combined with the VTE features, achieving an accuracy of 99.63%. The lowest accuracy (96.69%) is observed when restricting the model to traffic features from the last timestep only, specifically excluding the VTE features. This indicates that while traffic features alone are sufficient to capture most policy behavior, knowledge of the system improves policy imitation. By removing correlated features when only looking at traffic features from the last timestep, policy imitation accuracy can be slightly increased from 96.69% to 97.78% (from ii) to iii)), or from 98.75% to 99.63% (i) to iv)).
In terms of objective fulfillment, the RMSE of the learned policies is remarkably stable across all scenarios, with values around 5. These values are significantly lower than the no-speed-limit baseline (RMSE = 35.7), demonstrating the effectiveness of the learned policies. Furthermore, the RMSE of the learned policies is only marginally higher than that of the test set, indicating that the IRL approach successfully recovers policies that generalize well and follow the underlying objective. The use of VTE features improves object fulfillment and can even surpass MPC optimization.
A closer examination of the learned reward weights reveals distinct structural patterns. When using the full feature set, the weights are distributed across many features and are therefore relatively small in magnitude. Therefore, the added VTE features have larger weights than the large number of traffic features. Regarding the VTE features, the most negative weight consistently corresponds to the VTE of the selected action, reflecting the intended objective of minimizing the deviation between the imposed speed limit and the observed downstream speed. This is shown in Figure 2 for the reduced feature set with only last-timestep features. The VTE features associated with alternative actions often have positive weights, suggesting that larger deviations from other speed limits indirectly favor the selected action.

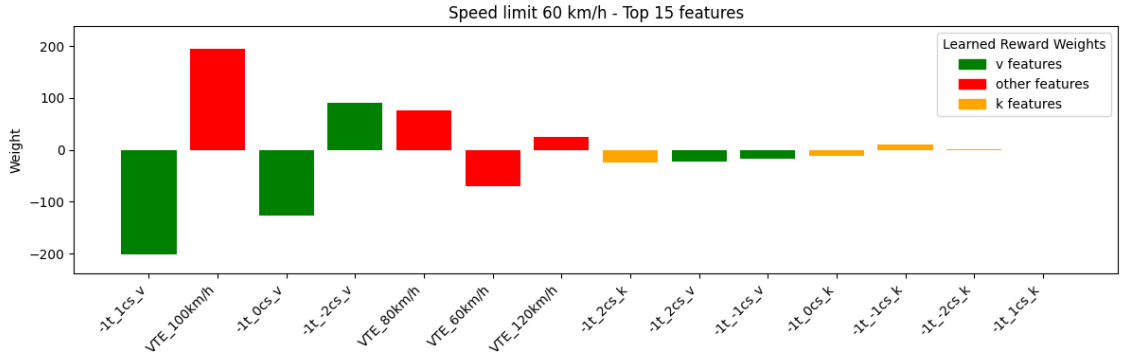


Figure 2: Weights for action "Speed limit 60 km/h" with reduced traffic features

When considering only traffic features, the learned weights exhibit interpretable spatio-temporal patterns, as illustrated in Figure 3. For density-related features, lower speed limits are associated with positive weights, whereas higher speed limits are associated with negative weights. This suggests that as traffic density increases, the model favors more restrictive speed limits, reflecting the

typical inverse relationship between density and speed. For the speed-related features, we clearly see the influence of downstream speeds, especially in the measurement of the next downstream cross-section, which aligns with the used policy. Results show low negative weights for a speed limit of 60km/h and high positive weights for a speed limit of 120km/h. The timesteps do not seem to play an important role, which could be explained by the relatively small simulation timesteps of 10s compared to the control timestep of one minute.
Overall, these results demonstrate that the IRL framework not only achieves high imitation accuracy but also recovers meaningful, interpretable reward structures that align well with the ground-truth VSL strategy.

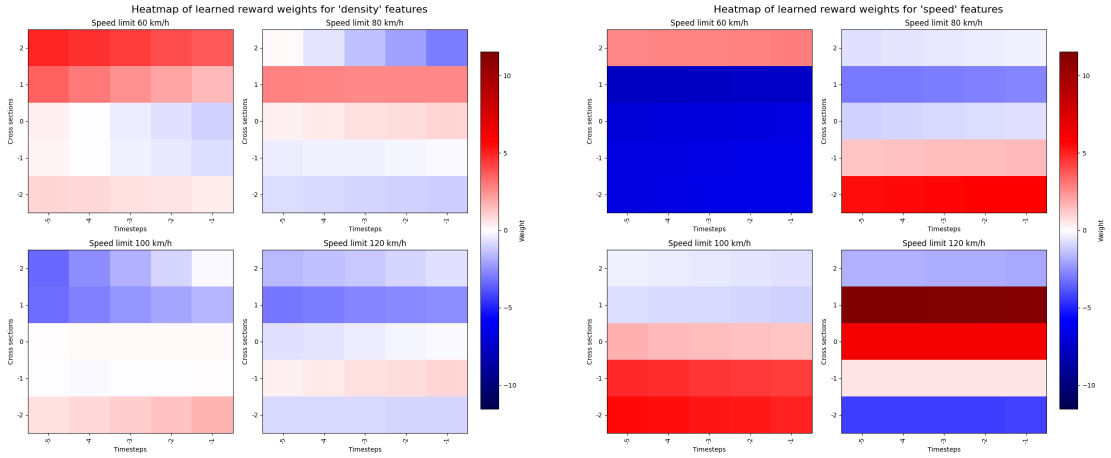


Figure 3: Space-time diagram of weight distribution for density features (left) and speed features (right)

4 Conclusions and Outlook

This study was motivated by the need to better investigate and assess variable-speed limit (VSL) control schemes to understand their underlying objectives and decision-making logic. By applying Inverse Reinforcement Learning (IRL), we demonstrated that meaningful reward structures can be recovered that accurately reproduce observed VSL policies. The results show consistently high policy imitation accuracy above 97% across all feature configurations. The learned reward weights exhibit interpretable patterns according to the used VSL scheme. The inclusion of engineered features can further improve interpretability and accuracy.
For future work, the proposed framework can be extended to analyze and compare different classes of VSL strategies, such as traffic flow harmonization, shock wave resolution, and reactive control schemes. Additionally, applying the approach to microscopic models and real-world data would improve practical relevance. Accounting for heterogeneous driver behavior, including compliance modeling and speed distributions, would enable more realistic evaluation of VSL systems and enhance the applicability of IRL in traffic control.

Acknowledgements

This paper is based on work funded by the German Federal Ministry for Digital and Transport (BMDV) within the research initiative mFUND. The research collaboration was financially supported by the Karlsruhe House of Young Scientists (KHYS). The authors further acknowledge partial financial support of the EU Horizon Europe Research and Innovation Programme (Grant Agreement No. 101103808 ACUMEN). Large Language Models were used to enhance readability. The authors reviewed and edited all content as needed and take full responsibility for the final publication.

References

Argall, B. D., Chernova, S., Veloso, M., & Browning, B. (2009). A survey of robot learning from demonstration. *Robotics and Autonomous Systems*, *57*(5), 469–483.

Bundesministerium für Verkehr und Infrastruktur (Pub.). (2018). *Merkblatt für die Ausstattung von Verkehrsrechnerzentralen und Unterzentralen.*

Daganzo, C. F. (1995). The cell transmission model, part ii: Network traffic. *Transportation Research Part B: Methodological*, *29*(2), 79-93.

Frejo, J. R. D., Núñez, A., De Schutter, B., & Camacho, E. F. (2014). Hybrid model predictive control for freeway traffic using discrete speed limit signals. *Transportation Research Part C: Emerging Technologies*, *46*, 309–325.

Grau, J., Weyland, C. M., Baumann, M. V., & Vortisch, P. (2026). Reconstruction of a freeway control systems' algorithm based on convolution neural networks. *Transportation Research Procedia*, *95*, 433-440.

Gu, S., Zhang, T., & Zhang, Y. (2023). Inverse reinforcement learning integrated reinforcement learning for single intersection traffic signal control. In *Proceedings of the 2023 5th International Conference on Industrial Artificial Intelligence (IAI).*

Hegyi, A., De Schutter, B., & Hellendoorn, H. (2005). Model predictive control for optimal coordination of ramp metering and variable speed limits. *Transportation Research Part C: Emerging Technologies*, *13*(3), 185–209.

Ng, A., & Russell, S. (2000). Algorithms for Inverse Reinforcement Learning.

Zhang, Y., Quinones-Grueiro, M., Zhang, Z., Wang, Y., Barbour, W., Biswas, G., & Work, D. (2024). MARVEL: Bringing Multi-Agent Reinforcement- Learning Based Variable Speed Limit Controllers Closer to Deployment. *IEEE Access*, *12*, 161995–162014.

Ziebart, B. D., Maas, A., Bagnell, J. A., & Dey, A. K. (2008). Maximum Entropy Inverse Reinforcement Learning. In *Proceedings of the twenty-third AAAI Conference on Artificial Intelligence.*