Validating a data-driven framework for vehicular traffic modeling

IOP Publishing J. Phys. Complex. 5 (2024) 025008 https://doi.org/10.1088/2632-072X/ad3ed6

Journal of Physics: Complexity

OPEN ACCESS

RECEIVED
21 December 2023

REVISED
1 April 2024

ACCEPTED FOR PUBLICATION
15 April 2024

PUBLISHED
3 May 2024

PAPER

Validating a data-driven framework for vehicular traffic modeling

Daniel Lane and Subhradeep Roy*

Embry-Riddle Aeronautical University, Daytona Beach, FL 32114, United States of America
* Author to whom any correspondence should be addressed.

E-mail: roys5@erau.edu

Keywords: data-driven modeling, complex systems modeling, information theory, model identification, SINDy, transfer entropy, traffic

Abstract

This study presents a data-driven framework for modeling complex systems, with a specific emphasis on traffic modeling. Traditional methods in traffic modeling often rely on assumptions regarding vehicle interactions. Our approach comprises two steps: first, utilizing information-theoretic (IT) tools to identify interaction directions and candidate variables thus eliminating assumptions, and second, employing the sparse identification of nonlinear systems (SINDy) tool to establish functional relationships. We validate the framework's efficacy using synthetic data from two distinct traffic models, while considering measurement noise. Results show that IT tools can reliably detect directions of interaction as well as instances of no interaction. SINDy proves instrumental in creating precise functional relationships and determining coefficients in tested models. The innovation of our framework lies in its ability to use data-driven approach to model traffic dynamics without relying on assumptions, thus offering applications in various complex systems beyond traffic.

1. Introduction

In complex systems, group-level behaviors such as self-organization and phase transitions emerge from interactions between units. Traffic systems are examples of complex systems, where 'interaction' refers to the dynamic relationships and influences between vehicles on the road [1]. Understanding these interactions is vital for developing precise models, which have practical applications in improving traffic planning, reducing travel times, fuel consumption, pollution, and congestion [2].

Traffic systems can be modeled in a variety of ways. A popular approach is agent-based models, which create a road network, add agents to it, and define their behavior and rules of interaction [3]. Traffic flow simulation software applications have been developed based on microscopic agent-based modeling, including MovSim [4, 5], SUMO [6, 7], MITSIM [8, 9]. Microscopic agent-based models consider the driver and vehicle as one entity and the movement of every driver-vehicle unit is simulated, considering car-following dynamics [9, 10], lane-changing behavior [11, 12], gap acceptance maneuvers [13], and movement at intersections [14]. Numerous models have been developed (Gipps' model [15], intelligent driver model (IDM) [16], optimal velocity model (OVM) [17], each with its own set of rules. These models exhibit sparsity which means they consider a small number of relevant features or coefficients. In traffic modeling, this sparsity aids in understanding vehicle movement and capturing underlying physics. However, these models heavily rely on assumptions about driver interaction, such as each driver being influenced only by the vehicle immediately in front of it.

In an attempt to mitigate these assumptions, some researchers have turned to artificial intelligence (AI) models that incorporate real-world data [18–21]. Nevertheless, there remains a common concern that these AI models act as black-boxes [10]. This criticism implies that neural network models can be understood solely based on their inputs and outputs, without providing insight into their internal mechanisms. For traffic systems, sparse microscopic models can be more reliable than black-box models because sparse models are based on well-defined principles that mimic human behavior by taking into account of

parameters like headway distance and speed, offering improved transparency and interpretability. Interpretability of sparse models facilitates a better understanding of traffic dynamics, whereas black-box models lack explicit rules for driving behavior. Moreover, the computational efficiency of these sparse models makes them effective in the design and control of traffic systems when implemented in real-time [22, 23].

This paper presents a two-step approach to develop sparse traffic models from data, eliminating the need for assumptions. The first step eliminates assumptions by detecting true directional relationships between vehicles based on trajectory data without having any prior knowledge of vehicle interactions. Specifically, we investigate how many preceding and following vehicles influence a subject vehicle in single-lane traffic. To achieve this, we employ information theoretic (IT) tools, which help identify the relevant candidates to be included in the model. The final step uses sparse identification of nonlinear systems (SINDy) to identify functional relationships between the candidate variables, considering only vehicles that exert influence on the subject, thus completing the model identification process.

Information theory has emerged as a valuable tool for detecting directional relationships directly from data in complex system studies. Specifically, transfer entropy (TE) and conditional transfer entropy (CTE) can quantify coupling between time-series variables, and therefore identify candidate variables for a model. These metrics have found successful applications in the study of various complex systems including human brain activity [24–26], animal collective behavior [27–31], climate modeling [32], policy-making [33–37] and financial markets [38, 39]. However, its application within vehicular traffic systems has been relatively limited [40–43]. While IT metrics can provide empirical evidence of candidate variables required for fully describing a systems dynamics, the functional relationship between the variables still needs to be discovered. Identifying the functional relationships between the candidate variables using a sparse modeling approach involves selecting a minimal combination of nonlinear functions which fully describes the system dynamics from a larger library of candidate variables, without making any prior assumptions. This relationship can be detected using the SINDy framework [44–47] and has found applications in various fields such as fluid mechanics [48], plasma dynamics [49, 50], optics [51], and power grids [52].

The present study employs a data-driven framework to construct sparse traffic models, combining IT tools with SINDy. Specifically, we evaluate the effectiveness of TE and CTE measures in detecting true coupling using synthetic data from two distinct car-following models in stop-and-go (jammed-flow) and free-flow traffic scenarios. The functional relationships among candidate variables are then determined using SINDy across varying levels of noise. Toy models serve as ground truth to evaluate the performance of TE, CTE, and the accuracy of dynamical equations identified by SINDy. The innovative aspect of this work lies in its validation of the proposed data-driven modeling framework and establishing confidence in these tools before applying them to real-world scenarios. The findings from this study additionally provide insights by comparing the two IT metrics and enhancing understanding of how the results from these metrics can be interpreted.

2. Methods

In this section, we present two microscopic traffic models used to generate toy dataset: the IDM and the OVM, as well as the data-analytic tools used for analysis: TE, CTE, and SINDy. While the IDM and OVM are structurally different, both are car-following models which determine a subject vehicle's acceleration by considering only the relative speeds and/or positions between it and the vehicle immediately ahead of it.

2.1. IDM

In IDM, the input parameters are the vehicle's speed v, bumper-to-bumper distance to the leading vehicle (distance headway) s, and the relative speed (Δv). The model outputs the acceleration of a vehicle as:

$$\frac{\mathrm{d}v}{\mathrm{d}t} = a\left(1 - \left(\frac{v}{v_0}\right)^{\delta}\right) - a\left(\frac{s^*(v,\Delta v)}{s}\right)^2 \tag{1}$$

where a is the maximum vehicle acceleration, v_0 is the desired velocity, δ is an acceleration exponent, and s^* is the desired minimum headway. The first part of the equation describes free flow, in which the acceleration a decreases to zero as the speed approaches v_0. The second part corresponds to the interaction term (braking), where the current distance headway (s) and the desired headway (s^*) are compared and deceleration is increased as the current headway decreases. The desired minimum headway s^* is given by:

$$s^* = s_0 + \max\left(0, vT + \frac{v\Delta v}{2\sqrt{ab}}\right) \tag{2}$$

where s_0 is the minimum gap allowed, T is the time headway, and b is a positive coefficient defining the rate of deceleration [16]. The ballistic method [53] is used to solve equation (1) and the speed and vehicle positions are determined as:

$$v(t+\Delta t) = v(t) + \frac{\mathrm{d}v}{\mathrm{d}t}\Delta t \quad \text{and}$$
$$x(t+\Delta t) = x(t) + v(t)\,\Delta t + \frac{1}{2}\frac{\mathrm{d}v}{\mathrm{d}t}\Delta t^2.$$

If the front vehicle is at rest, there is a possibility of calculating both a negative acceleration and velocity in the next time step. This negative velocity is prevented by implementing the following conditional statement [53]:

$$\textbf{if:} \quad v(t) + (\mathrm{d}v/\mathrm{d}t)\,\Delta t < 0$$
$$\textbf{then:} \quad v(t+\Delta t) = 0 \quad \text{and} \quad x(t+\Delta t) = x(t) - \frac{1}{2}v^2(t)\,/\frac{\mathrm{d}v}{\mathrm{d}t}. \tag{3}$$

Each vehicle is simulated identically using typical model parameter values of $T = 1.5$, $a = 0.3$, $b = 3$, $\delta = 4$, $s_0 = 2$ [16]. The circumference of the circular track is set to $L = 314$ meters and $v_0 = 30\ \mathrm{km\,h^{-1}}$ based on [54] and given an initial speed of $30\ \mathrm{km\,h^{-1}}$. Trajectory data for each vehicle are recorded at a 0.1 second interval.

2.2. OVM

Different from IDM, vehicle acceleration in OVM is dependent on the difference between the vehicle's current speed and optimal speed $V(s)$:

$$\frac{\mathrm{d}v}{\mathrm{d}t} = a_h\left[V(s) - \frac{\mathrm{d}x}{\mathrm{d}t}\right]. \tag{4}$$

The parameter a_h accounts for heterogeneity among vehicle types and drivers [17], which in our simulation is assumed to be a constant for all vehicles for all time. The optimal velocity (OV) function is a hyperbolic tangent function of distance headway s and is given by:

$$V(s) = \alpha\tanh\left[\beta\,(s - s_0)\right] + v_0 \tag{5}$$

The minimum distance headway s_0 along with constants α, β, and v_0 scale the hyperbolic tangent function and determine the response of the OV given the value of distance headway s. Here, as s approaches s_0, the OV reduces to zero to avoid collision. We choose the parameter values as $a_h = 1.8$, $\alpha = 5.5$, $\beta = 0.37$, $s_0 = 9.1$, and $v_0 = 4.9$ based on empirical evidence [55]. To avoid bias in synthetic data generation, simulation conditions (such as track length, initial conditions, etc) are kept constant to those used with the IDM.

2.3. IT tools

Pairwise TE and CTE or causation entropy are extensions of the definition of entropy described by Shannon in 1948 [56]. TE was formalized concurrently by Schreiber [57] and Palus *et al* [58] to assess the information exchange between two variables (X and Y) over time. It is a metric commonly used to detect the coupling strength and direction between time series variables. For example, to detect coupling from $Y \to X$, it quantifies how much information Y can provide about the future state of X using the present states of both variables. With a first-order Markov process assumption, pairwise TE is defined as:

$$T_{Y\to X} = \left\langle \log \frac{p\,(x_{n+1} \mid x_n, y_n)}{p\,(x_{n+1} \mid x_n)} \right\rangle \tag{6}$$

where $\langle . \rangle$ is the average computed over all the samples, n is the time index, $p(x_{n+1})$ denotes probability, and $p(x_{n+1}|x_n)$ is the probability of x_{n+1} conditioned on the present state x_n. If there is no influence from Y on X, then $p\,(x_{n+1} \mid x_n, y_n) = p\,(x_{n+1} \mid x_n)$, and $T_{Y\to X} = 0$. The unit for TE is determined by the base of the logarithm used, i.e. 'nats' for $\log_e$, and 'bits' for $\log_2$.

When three (or more) variables are involved (X, Y, and Z), pairwise TE may not distinguish the indirect couplings [59]. In such scenarios, CTE can be applied. CTE evaluates the direct influence of Y on X accounting for any indirect influence from Z, and is defined as:

$$C_{Y\to X|Z} = \left\langle \log \frac{p\,(x_{n+1} \mid x_n, z_n, y_n)}{p\,(x_{n+1} \mid x_n, z_n)} \right\rangle \tag{7}$$

where $p(x_{n+1} \mid x_n, z_n, y_n)$ is the probability of x_{n+1} conditioned on x_n, y_n and z_n. When Y does not influence X, the values of the numerator and denominator become equivalent, thus $C_{Y\to X|Z}$ equals zero.

Both TE and CTE are asymmetric by construction ($T_{X\to Y} \neq T_{Y\to X}$, and $C_{Y\to X|Z} \neq C_{X\to Y|Z}$); allowing for the dominant direction of information flow (coupling direction) to be identified [35]. When examining TE and CTE coupling from finite empirical data, a statistical significance test can be conducted using surrogate data to determine if the resultant value is statistically different than zero [40]. In the present work, the IT measurements and statistical significance tests are performed using the Java Information Dynamics Toolkit for Matlab [60]. The Kraskov estimation method is used for bias correction [61] when estimating the probability density functions.

2.4. Sparse identification of nonlinear dynamics (SINDy)

Here we present an overview of how SINDy identifies governing dynamical systems models from data. SINDy considers a dynamical system of the form:

$$\dot{\mathbf{x}}(t) = \frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = \mathbf{f}(\mathbf{x}(t)),$$

where the vector $\mathbf{x}(t) = [x_1(t); x_2(t); \ldots, x_n(t)] \in \mathbb{R}^n$ represents the state of a system at time t and the function $\mathbf{f}(\mathbf{x})$ describes the temporal evolution of the system's state. SINDy identifies the fewest terms that approximate the unknown $\mathbf{f}(\mathbf{x})$ and establishes the model based on a library of candidate basis functions $\boldsymbol{\Theta}(\mathbf{x}) = [\theta_1(\mathbf{x}), \theta_2(\mathbf{x}), \ldots, \theta_p(\mathbf{x})]$:

$$f_j = \sum_{k=1}^{p} \xi_{jk}\theta_k(\mathbf{x}),$$

where the coefficients ξ_{jk} are typically zero, and entries that are not zero indicate active terms in the dynamics. To find $\mathbf{f}$, time-series measurements of $\mathbf{x}$ and their time derivatives $\dot{\mathbf{x}}$ (measured directly or approximated numerically) are sampled at several time steps $t_1, t_2, \ldots, t_m$ and arranged into matrices such that $\mathbf{X} = [\mathbf{x}(1)\ \mathbf{x}(2)\cdots\mathbf{x}(m)]^T$ and $\dot{\mathbf{X}} = [\dot{\mathbf{x}}(1)\ \dot{\mathbf{x}}(2)\cdots\dot{\mathbf{x}}(m)]^T$ with $\mathbf{X}, \dot{\mathbf{X}} \in \mathbb{R}^{m\times n}$, where m is the sample size and n is the number of states. The library functions are next evaluated on the data by constructing $\boldsymbol{\Theta}(\mathbf{X}) = [\theta_1(\mathbf{X})\theta_2(\mathbf{X})\cdots\theta_p(\mathbf{X})] \in \mathbb{R}^{m\times p}$. Finally, SINDy uses the sparse regression technique to approximately solve:

$$\dot{\mathbf{X}} \approx \boldsymbol{\Theta}(\mathbf{X})\,\boldsymbol{\Xi},$$

where $\boldsymbol{\Xi} = [\boldsymbol{\xi_1}\boldsymbol{\xi_2}\cdots\boldsymbol{\xi_n}] \in \mathbb{R}^{p\times n}$ is a set of coefficients that determines the active terms in $\mathbf{f}$. An extension of SINDy, referred to as SINDy-PI [46], has been developed for implicit differential equations of the form:

$$\mathbf{f}(\mathbf{x}, \dot{\mathbf{x}}) = 0,$$

and then the sparse model is detected as

$$\boldsymbol{\Theta}\left(\mathbf{X}, \dot{\mathbf{X}}\right)\boldsymbol{\xi} = 0.$$

It is also possible to include control input data in the SINDy-PI algorithm.

The equations that describe the two traffic models we utilize for data generation are implicit. Creating a traffic model for a subject vehicle requires incorporating data from adjacent vehicles as control input. Given that SINDy-PI is capable of handling these requirements, we employ it in our study.

3. Results and discussion

In this section, we present the results using IT tools that can detect the true nature of interactions, i.e. whether vehicles react only to vehicles in front, or also to vehicles behind and further ahead. Next, it is necessary to identify the rules for these interactions, essentially establishing functional relationships among relevant variables. We further validate the effectiveness of the proposed framework using toy data with known ground truth.

3.1. Generating synthetic traffic data

We generate synthetic data from two fundamentally different car-following models, IDM and OVM, to validate the proposed data-analytic framework for identifying traffic models. The utilization of synthetic data is motivated by its known ground truth, providing a benchmark for evaluating the performance of the tools employed in this study. We generate data by simulating vehicles on a circular track with a single lane of traffic as illustrated in figure 1. This setup allows for large samples of data to be generated by tracking

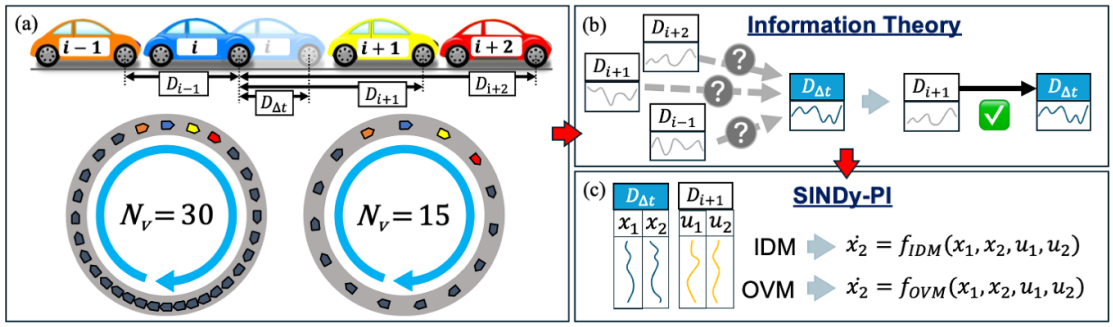


Figure 1. (a) Visualization of the simulation setup for jammed ($N_v = 30$) and free ($N_v = 15$) traffic, and the time-series variables used in the IT analysis to detect the influence of adjacent vehicles on a subject vehicle i. We use two distinct car-following models, IDM and OVM, to generate synthetic data. (b) Upon testing the influence of immediate front (using time series data of D_{i+1}), next preceding (D_{i+2}), and rear (D_{i-1}) vehicles on the subject ($D_{\Delta t}$), IT analysis accurately identifies that the sole existing influence is from $(i+1)$th vehicle. (c) IT analysis thus reveal that the candidate variables for the model of the ith vehicle are solely associated with the $(i+1)$th vehicle. Subsequently, SINDy-PI uncovers the functional relationships among these variables, utilizing both position and speed. We validate the proposed framework for model identification in both the presence and absence of noise.

vehicles within a fixed arena. For IT measures, large datasets are necessary to estimate probability density functions accurately. We simulate vehicles to drive on a single lane which eliminates lateral interactions (influence from adjacent lanes). For each model, we simulate jammed-flow and free-flow traffic conditions by adjusting the vehicle density within a constant track length of 314 meters. For jammed-flow, the number of vehicles (N_v) is set at 30, while for free-flow, N_v is set to 15. The circular track is simulated by imposing a periodic boundary condition. Periodic boundary conditions are applied to the simulation, treating the last vehicle ($i = N_v$) separately. The position of the vehicle in front of it is considered as the position of the first vehicle ($i = 1$) plus the track length, achieving the periodic boundary condition. For visualization, vehicle positions are wrapped within the track length.

3.2. Detect interaction using IT tools

Using synthetic data with known ground truth, we evaluate the ability of IT tools to identify coupling direction accurately. For IT analysis, we compute the observables as shown in figure 1. For ith vehicle at a given instant, we compute the distance headway between vehicle i and its immediate front vehicle (D_{i+1}), the distance from its immediate rear vehicle (D_{i-1}), the distance from $i+2$th vehicle (D_{i+2}), and the distance the vehicle i travels in the next Δt time interval ($D_{\Delta t}$). Subsequently, we employ the IT measures to quantify coupling between these time-series variables to detect the influence of the adjacent vehicles on a subject vehicle. For IT analysis, the simulated vehicle trajectory data was resampled to obtain data points at one second intervals to match human driver reaction time [40]. This resulted in 29 000 samples per vehicle after excluding first 1000 samples to eliminate initial transient. The table 1 provides a summary of the variables and samples utilized in IT and SINDy-PI analysis, along with the interpretation of the corresponding analysis.

The results of IT analysis are presented in figure 2. Sub-figures in the left column correspond to jammed-flow ($N_v = 30$) while those in the center correspond to free-flow ($N_v = 15$). Results of IT analysis of OVM data is shown in the first three rows, and results of IT analysis of IDM data are shown in the last row.

3.2.1. IT analysis on OVM data

As a first step, we use pairwise TE to identify the influence from the vehicle directly in front to the ith vehicle ($T_{i+1\to i}$), as well as the influence from the vehicle directly behind ($T_{i-1\to i}$). Figure 2(a) shows that for all 30 vehicles, pairwise TE accurately identifies statistically significant coupling from the front vehicle. However, it detects false coupling from the rear vehicle since the OVM dynamics do not incorporate information from the rear vehicle. A common issue with pairwise TE measure is its challenge in discerning indirect coupling. This arises because when the front vehicle alters its position, it exerts an influence on the ith vehicle to also change its position. This, in turn, impacts the distance with its immediate rear vehicle—the variable used in TE to detect the influence of the rear. Therefore, in the presence of unidentified confounding variables, the pairwise TE measure is often used to identify the dominant coupling direction. For instance, in this scenario, TE accurately identifies that the dominant coupling direction is always from the front vehicle, as $T_{i+1\to i}$ consistently greater than $T_{i-1\to i}$. These measures also exhibit consistent values across vehicles, as data is generated for all vehicles from the same exact model.

The homogeneity is lost in the pairwise TE analysis of free flow data (figure 2(b)), where, for some vehicles, TE detects statistically significant influence from the front but not from the rear. The cross symbols

Table 1. Summary of tools, variables, samples used along with their corresponding interpretation.

Tools	Time series variables	Sample size	Interpretation of the analysis
$T_{i+1\to i}$	$D_{\Delta t}, D_{i+1}$	29 000	Detect influence of front vehicle on a subject (i)
$T_{i-1\to i}$	$D_{\Delta t}, D_{i-1}$	29 000	Detect influence of rear vehicle on a subject (i)
$C_{i+1\to i\|i-1}$	$D_{\Delta t}, D_{i+1}, D_{i-1}$	29 000	Detect front-to-subject influence eliminating indirect influence from rear
$C_{i-1\to i\|i+1}$	$D_{\Delta t}, D_{i+1}, D_{i-1}$	29 000	Detect rear-to-subject influence eliminating indirect influence from front
$C_{i+1\to i\|i+2}$	$D_{\Delta t}, D_{i+1}, D_{i+2}$	29 000	Detect front-to-subject influence eliminating indirect influence from lead
$C_{i+2\to i\|i+1}$	$D_{\Delta t}, D_{i+1}, D_{i+2}$	29 000	Detect lead-to-subject influence eliminating indirect influence from front
SINDy-PI	x_1, x_2, u_1, u_2	4000 (train) 2000 (test)	Detect functional relationships

indicate statistically non-significant results, implying a lack of evidence for coupling. At first glance, it may seem that TE performed better for free flow data by accurately identifying coupling from the front and rejecting coupling from the rear. To investigate this further, we plot in figure 2(c) the OV function (equation (5)) versus distance headway s from our simulation. The plot demonstrates that vehicles maintain their maximum speed if their distance headway exceeds a threshold. The OV function determines how a vehicle responds when headway drops below the threshold as it reduces its speed. Using the distance headway of all the vehicles measured from our simulation data, we compute the interaction regimes corresponding to free and jammed traffic. Vertical lines represent the average headway of all vehicles over the entire simulation time (dotted line for jammed-flow, dashed line for free-flow), and shaded regions indicate one standard deviation. Notably, we observe that in free-flow conditions, with distance headway exceeding the threshold, vehicles follow the maximum desired speed. As a result, they are not influenced by their immediate front cars, resulting in the absence of interactions.

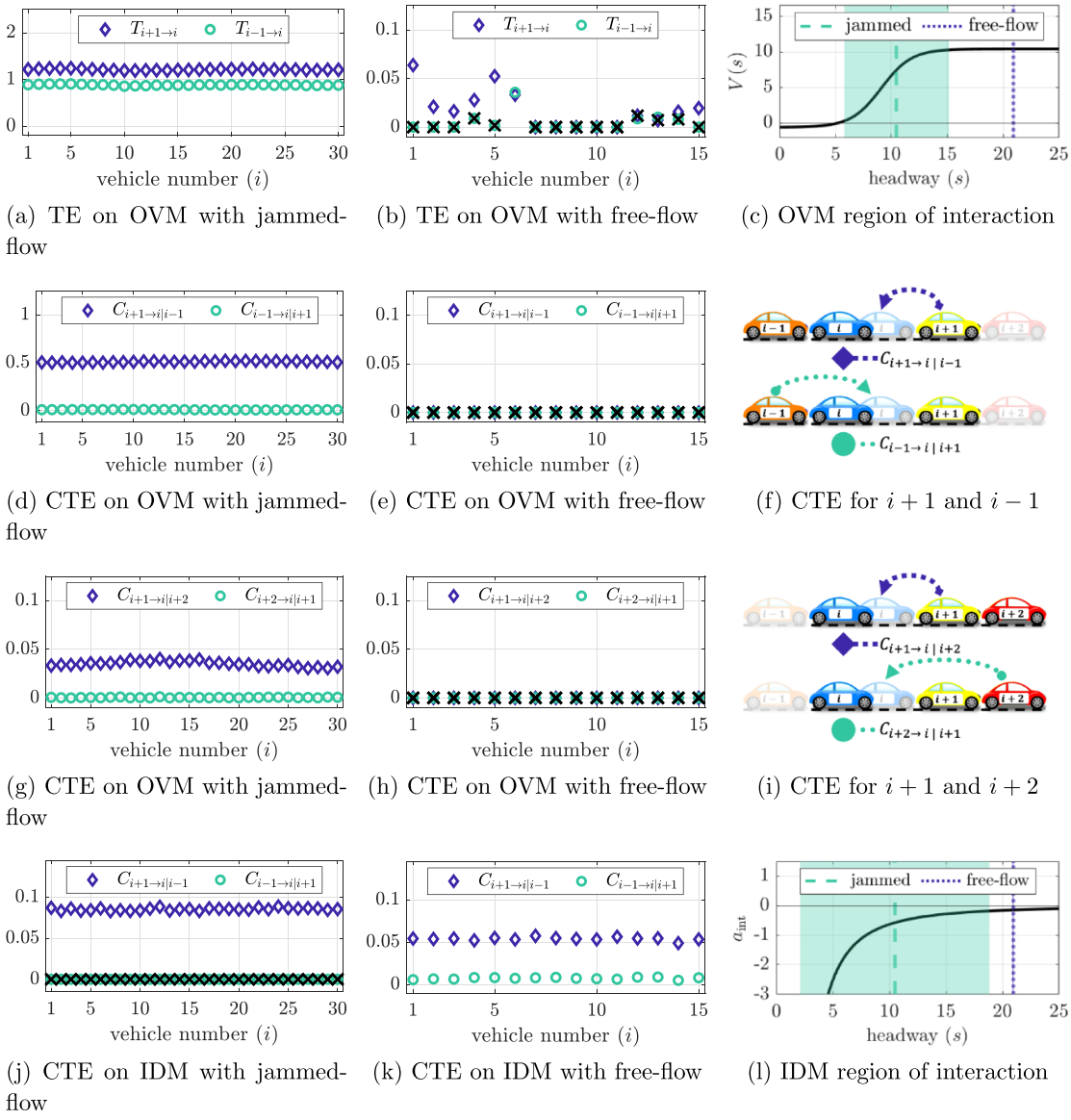
Next, we employ the CTE measure on OVM data to determine its ability to accurately discern that the coupling is solely present in jammed traffic and that the direction of influence is only from the front vehicle. The results of CTE analysis with OVM data and the corresponding schematic illustrations are presented in the figures 2(d)–(i). We conduct a thorough analysis by examining the influence of two preceding vehicles ($i+1$ and $i+2$) and the immediate rear vehicle ($i-1$) on the ith vehicle.

Observing figure 2(d), CTE correctly detects the significant coupling is from the immediate front car for each ith vehicle in jammed traffic. However, CTE values from the rear vehicle are identified as statistically significant, albeit with magnitudes much smaller when compared to those from the front vehicle. This is evident from the averages computed over 30 vehicles, with $\langle C_{i+1\to i\|i-1}\rangle = 0.5120$ nats and $\langle C_{i-1\to i\|i+1}\rangle = 0.0109$ nats. This finding indicates it is necessary to account for both the significance of the test results and actual CTE values in order to infer coupling. The value of rear-to-target coupling being almost near zero indicates that, despite its statistical significance, such couplings are not present, and can therefore be ignored. Similarly, in figure 2(g), the results of the CTE analysis involving two preceding vehicles indicate a negligible influence from the $i+2$th vehicle, thus can be disregarded. Therefore, combining knowledge from figures 2(d) and (g), the CTE results accurately indicate that the only influence in jammed traffic originates from the immediate front car ($i+1$).

The figures 2(e) and (h) illustrate the results of the CTE analysis on free-flow data, correctly revealing non-significant coupling from either direction. The results demonstrate that CTE has the ability to accurately detect the lack of interaction. The discussion on the lack of interaction in free-flow OVM is already elaborated above using the OV function in figure 2(c).

3.2.2. IT analysis on IDM data

As for the IDM, we only use CTE analysis, which is in line with the conclusions of IT analysis on OVM data, which indicates that CTE is better than pairwise TE for our study because it takes indirect coupling into account. Figures 2(j) and (k) present the results of CTE analysis of IDM data. For both jammed and free-flow, CTE results accurately identify the only coupling is from the immediate front vehicle. Note that,



(a) TE on OVM with jammed-flow (b) TE on OVM with free-flow (c) OVM region of interaction

(d) CTE on OVM with jammed-flow (e) CTE on OVM with free-flow (f) CTE for $i+1$ and $i-1$

(g) CTE on OVM with jammed-flow (h) CTE on OVM with free-flow (i) CTE for $i+1$ and $i+2$

(j) CTE on IDM with jammed-flow (k) CTE on IDM with free-flow (l) IDM region of interaction

Figure 2. Results and figures related to computing TE (a), (b) and CTE (d)–(l). Sub-figures (a), (d), (g), (j) in the left column correspond to jammed flow ($N_v = 30$), and sub-figures (b), (e), (h), (k) in the center column correspond to free flow ($N_v = 15$). In the first two columns, IT units on the vertical axis are in nats and the black symbol **x** denotes that the empirical measurements are not statistically different from zero. In the third column: sub-figure (c) shows the regions of interaction experienced during the jammed and free flow simulation of OVM, sub-figures (f) and (i) are schematics showing the conditions being tested for their respective rows, and (l) shows the regions of interaction experienced during the jammed and free flow simulation of IDM.

unlike in OVM, IDM has interaction occurring even in the free-flow. This is evident from the figure 2(l), which displays the interaction regions similarly computed as in figure 2(c) from the interaction term in equation (1). The interaction regions show that, for the simulation parameters used, there is strong coupling in jammed-flow and weak coupling in free-flow, which is also determined correctly by the CTE measure. Specifically, upon comparing $C_{i+1\to i|i-1}$ values between the figures 2(j) and (k), we find that the influence from the front vehicle is more pronounced in jammed flow compared to free-flow. The stronger influence in jammed traffic occurs because vehicles go through a series of stop-and-go events, requiring them to respond more frequently to the front vehicle compared to when traffic is flowing freely.

In summary, IT analysis of OVM and IDM data shows that CTE is a more effective approach than pairwise TE to distinguish indirect influences and infer whether coupling exists or not. To accurately infer the presence of coupling both statistical significance and actual values must be considered; and finally CTE measure is found to be sensitive to coupling strength (e.g. for IDM, the front vehicle exerts more influence in jammed flow than in free-flow.). In this study, the CTE analysis identifies coupling only from the immediate front vehicle, thus matching the ground truth of both the car-following models. This tool is thus validated by two distinct traffic models to accurately infer directional influence from adjacent vehicles, proving its usefulness for real-world applications. This validation is crucial as accurately identifying the range of

vehicles, whose variables should be incorporated into the model, is essential for the development of sparse traffic models.

3.3. Model identification using SINDy-PI

The IT analysis of IDM and OVM data suggests that to create a traffic model for the vehicle i, only variables from the immediate front vehicle $(i+1)$ need to be included, as the other vehicles have no influence. With this knowledge, we proceed to employ SINDy-PI to investigate its capability for discerning functional relationships among variables, thereby facilitating the completion of model identification. We use data from jammed-flow scenarios where interaction is present in both OVM and IDM models. When applying SINDy-PI, we randomly select a vehicle from the set of 30 vehicles. Notably, we observe that this random selection does not impact the performance of SINDy.

The position and velocity of vehicle i are considered as primary states $\mathbf{x} = [x_1, x_2]$, position and velocity from vehicle $i+1$ are considered as a control input $\mathbf{u} = [u_1, u_2]$, and velocity and acceleration are represented as the time derivatives of $\dot{\mathbf{x}} = [\mathrm{d}x_1, \mathrm{d}x_2]$. For both training and testing, we utilize 400 and 200 seconds of data, respectively, with a time resolution of $\Delta t = 0.1$. This accounts for a total of 4000 data points for training and 2000 data points for testing. The testing and training sets remain consistent throughout the evaluations of SINDy-PI. In SINDy-PI, the candidate basis functions included in the library $\boldsymbol{\Theta}$ dictate the form of the final model. We assume the final model to be in the form of a polynomial which allows a direct comparison between each coefficient obtained through SINDy-PI and the model itself. Consequently, IDM in equation (1) and OVM in equation (4) are transformed into polynomials (details on these polynomial expansions are provided in appendix). The OVM equation is expanded using a Pade approximation of order [1,2], where this order denotes the power of the polynomials present in the numerator and denominator, respectively, of the series approximation [62, 63].

To evaluate the performance of SINDy-PI, we construct the library of candidate basis functions $\boldsymbol{\Theta}$ that include the compulsory terms ($\boldsymbol{\Theta}_C$) derived from the relevant models, as well as we add redundant terms ($\boldsymbol{\Theta}_R$) to create a complete polynomial of a given degree. The compulsory terms can be regarded as the initial hypothesis of the model. The performance of SINDy-PI will be assessed by its ability to accurately retain the compulsory terms, determine their correct coefficients, and reject the redundant terms. When constructing the library, we set the maximum exponent for $\mathbf{x}$ to 4 and for $\mathbf{u}$ to 2, resulting in terms with highest power of 6 $(x_1^4 u_1^2, x_1^4 u_1 u_2, \ldots)$ and yielding a total of 250 terms

$$\boldsymbol{\Theta} = \overbrace{\begin{bmatrix} | & | \\ \boldsymbol{\Theta}_C & \boldsymbol{\Theta}_R \\ | & | \end{bmatrix}}^{250\ \text{terms}},$$

where $\boldsymbol{\Theta}_C$ which is the set of all compulsory terms required to compose either the IDM or OVM and are presented in table 2 and $\boldsymbol{\Theta}_R$ contains the redundant terms.

3.3.1. SINDy-PI analysis of IDM and OVM

We systematically evaluate the performance of SIDNy-PI starting with using only the compulsory terms in the library for each model, denoted as $\boldsymbol{\Theta}_0 = \boldsymbol{\Theta}_C$. Next, we incrementally introduce redundant terms by selecting from $\boldsymbol{\Theta}_R$ up to a specified polynomial degree N:

$$\boldsymbol{\Theta}_N = \left[\boldsymbol{\Theta}_C\ \boldsymbol{\Theta}_R^N\right], \text{for } N = \{0, 1, 2, \ldots, 6\} \text{ such that:}$$
$$\begin{cases} \boldsymbol{\Theta}_0 = \boldsymbol{\Theta}_C \\ \boldsymbol{\Theta}_1 = \left[\boldsymbol{\Theta}_C\ \boldsymbol{\Theta}_R^1\right], \text{ terms up to degree 1} \\ \boldsymbol{\Theta}_2 = \left[\boldsymbol{\Theta}_C\ \boldsymbol{\Theta}_R^2\right], \text{ terms up to degree 2} \\ \vdots \\ \boldsymbol{\Theta}_6 = \left[\boldsymbol{\Theta}_C\ \boldsymbol{\Theta}_R^6\right], \text{ terms up to degree 6} \end{cases}$$

We evaluate SINDy-PI corresponding to each library $\boldsymbol{\Theta}_N$ and compare these results to the true model coefficients of the IDM and OVM. The performance of SINDy-PI is quantified in terms of sensitivity, specificity, and accuracy. Terms identified by SINDy-PI are labeled as true positive (TP) if they are present in the true model and false positive (FP) if they are not.

Table 2. We present the true model coefficients obtained from the polynomial expansion of IDM and OVM alongside sample SINDy results for comparison. The SINDy coefficients presented correspond to library of $\mathbf{\Theta}_3$ for IDM and library of $\mathbf{\Theta}_6$ for OVM. Rows displaying 'n/a' indicate the absence of the corresponding term in the library of the respective model.

Model coefficients and coefficients estimated by SINDy-PI				
Library	IDM		OVM	
θ_i	Model	SINDy-PI	Model	SINDy-PI
1	3.6000	3.4265	26.4497	26.4497
x_1	2.4000	2.4000	10.9867	10.9867
x_2	−1.8000	−1.8533	−25.8060	−25.8060
x_1x_2	n/a	n/a	−4.4848	−4.4848
x_1^2	0.3000	0.3089	1.2075	1.2075
x_2^2	−1.3074	−1.3462	n/a	n/a
$x_1^2x_2$	n/a	n/a	−0.2464	−0.2464
x_2^3	−0.4743	−0.4884	n/a	n/a
x_2^4	−0.0843	−0.0868	n/a	n/a
$x_1x_2^4$	−0.0005	−0.0005	n/a	n/a
$x_1^2x_2^4$	−0.0001	−0.0001	n/a	n/a
u_1	−2.4000	−2.4000	−10.9867	−10.9867
u_1x_1	−0.6000	−0.6178	−2.4149	−2.4149
u_1x_2	n/a	n/a	4.4848	4.4848
u_2x_2	0.6324	0.6512	n/a	n/a
$u_1x_1x_2$	n/a	n/a	0.4928	0.4928
$u_2x_2^2$	0.4743	0.4884	n/a	n/a
$u_2x_2^3$	0.1666	0.1716	n/a	n/a
$u_1x_2^4$	0.0005	0.0005	n/a	n/a
$u_1x_1x_2^4$	0.0001	0.0001	n/a	n/a
u_1^2	0.3000	0.3089	1.2075	1.2075
$u_1^2x_2$	n/a	n/a	−0.2464	−0.2464
$u_2^2x_2^2$	−0.0833	−0.0858	n/a	n/a
$u_1^2x_2^4$	−0.0001	−0.0001	n/a	n/a
dx_2	16.0000	15.5400	14.3367	14.3367
dx_2x_1	8.0000	8.0000	2.4916	2.4916
$dx_2x_1^2$	1.0000	1.0296	0.1369	0.1369
dx_2u_1	−8.0000	−8.0000	−2.4916	−2.4916
$dx_2u_1x_1$	−2.0000	−2.0592	−0.2738	−0.2738
$dx_2u_1^2$	1.0000	1.0296	0.1369	0.1369

As shown in table 2, the total number of positives (active terms) for the IDM and OVM are $P = 25$ and $P = 18$ respectively, and the number of negatives, N, depends on the library used for testing. Sensitivity, specificity, and accuracy are then measured as follows:

$$\text{sensitivity} = \frac{\text{TP}}{P}$$
$$\text{specificity} = \frac{N - FP}{N}$$
$$\text{accuracy} = \frac{\text{TP} + N - \text{FP}}{P + N}.$$

Tables 3 and 4 present the results in percentage. Additionally, we measure error computed by comparing the coefficient of each term identified by SINDy-PI with the true coefficient of the respective model. We calculate the error for each FP term (FP_e) by computing the absolute difference between the estimated coefficient returned by SINDy-PI and the true value, which is zero since those terms are absent in the model. The error for each TP term (TP_e) is determined as the ratio of the absolute difference between the estimated coefficient and true coefficient, divided by the true coefficient. Tables 3 and 4 shows the maximum values for both types of errors across all six evaluated libraries.

From table 3, we observe that SINDy-PI accurately identified all TP terms for IDM data across all libraries, achieving a sensitivity of 100%. Notably, the library $\mathbf{\Theta}_5$ led to the highest number of FP, resulting in a specificity of 64.1% and an overall accuracy of 70.42%. No FP terms are identified when evaluated with $\mathbf{\Theta}_3$, resulting in specificity and accuracy of 100%. The estimated coefficients for the TP terms fall within a range of 8% of the actual coefficients. The coefficients for the FP terms are on the order of 1×10^{-10} or lower,

Table 3. presents the sensitivity, specificity, accuracy, and maximum error for the coefficients identified by SINDy-PI for the IDM evaluated at each size of library $\boldsymbol{\Theta}_N$. There are 25 terms required to represent the IDM that SINDy-PI identified in each evaluation for a sensitivity of 100%. Additional terms identified by SINDy-PI are considered false positives.

Results from SINDy-PI for IDM							
Evaluated Library $\boldsymbol{\Theta}_N$	$\boldsymbol{\Theta}_0$	$\boldsymbol{\Theta}_1$	$\boldsymbol{\Theta}_2$	$\boldsymbol{\Theta}_3$	$\boldsymbol{\Theta}_4$	$\boldsymbol{\Theta}_5$	$\boldsymbol{\Theta}_6$
No. of terms in $\boldsymbol{\Theta}_N$	38	38	48	72	104	142	180
Sensitivity (%):	100	100	100	100	100	100	100
Specificity (%):	76.92	76.92	86.96	100	92.41	64.10	70.97
Accuracy (%):	92.11	92.11	93.75	100	94.23	70.42	75.00
Max (TP_e) (%):	4.904	7.753	7.556	7.511	4.820	4.820	4.820
Max (FP_e):	3×10^{-15}	2×10^{-13}	3×10^{-15}	n/a	1×10^{-14}	4×10^{-10}	9×10^{-11}

Table 4. presents the sensitivity, specificity, accuracy, and maximum error for the coefficients identified by SINDy-PI for the OVM evaluated at each size of library $\boldsymbol{\Theta}_N$. There are 18 terms required for the OVM that SINDy-PI identified in each evaluation for a sensitivity of 100%. Additional terms identified by SINDy-PI are considered false positives.

Results from SINDy-PI for OVM							
Evaluated Library $\boldsymbol{\Theta}_N$	$\boldsymbol{\Theta}_0$	$\boldsymbol{\Theta}_1$	$\boldsymbol{\Theta}_2$	$\boldsymbol{\Theta}_3$	$\boldsymbol{\Theta}_4$	$\boldsymbol{\Theta}_5$	$\boldsymbol{\Theta}_6$
No. of terms in $\boldsymbol{\Theta}_N$	24	26	34	56	94	136	180
Sensitivity (%):	100	100	100	100	100	100	100
Specificity (%):	100	100	100	100	100	100	100
Accuracy (%):	100	100	100	100	100	100	100
Max (TP_e) (%):	0.008	0.007	0.008	0.008	0.012	0.004	0.008
Max (FP_e):	n/a	n/a	n/a	n/a	n/a	n/a	n/a

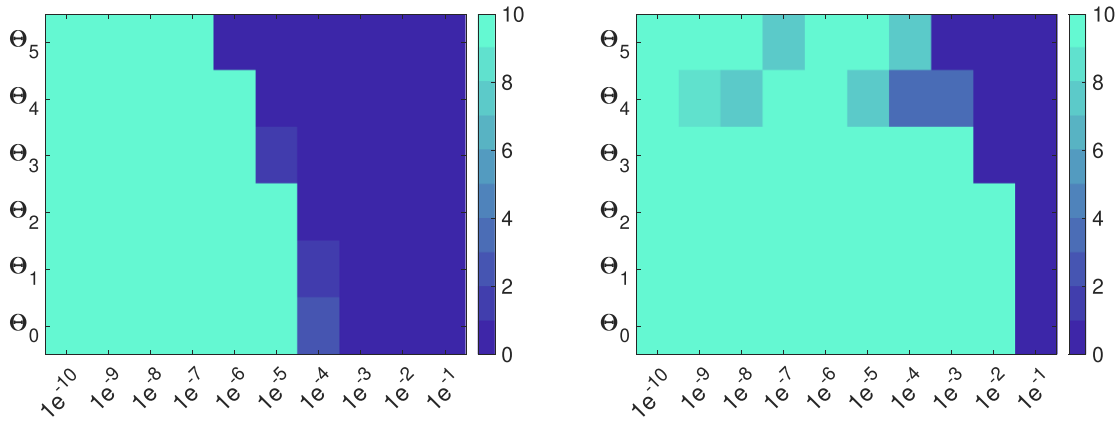
markedly smaller than any coefficients found in the actual set. Consequently, when the model is not known in advance, these can be confidently disregarded. When applied to the OVM, SINDy-PI achieves a perfect accuracy of 100% across all evaluated libraries, as shown in table 4. Furthermore, the error in the TP coefficients is below 0.01%.

These results indicate that SINDy-PI accurately identified the TP terms and their corresponding coefficients for the respective model. The reduced specificity observed in the IDM results is attributed to the ballistic method used in generating data. If the front vehicle is at rest, the ballistic update method to solve IDM (section 2.1) introduces discontinuities. SINDy-PI accommodates these discontinuities by incorporating additional terms. Moreover, we perform additional tests using varying sampling rates with CTE and SINDy-PI. The supplementary analysis produced comparable results, affirming the robustness of our findings.

3.3.2. SINDy-PI analysis of IDM and OVM with noise

Next, we evaluate the robustness of SINDy-PI's performance in the presence of measurement noise, which is common with real-world data, including traffic systems [64]. We add Gaussian noise $\mathcal{N}(0, \sigma_i^2)$ of ten increasing magnitudes where σ_i ranges from 1×10^{-10} to 1×10^{-1}. Additionally, at each σ_i and library size $\boldsymbol{\Theta}_N$, we generate ten independent datasets for the evaluation of SINDY-PI to account for stochastic noise. A model returned by SINDy-PI is considered successful if it meets three criteria: sensitivity $= 100\%$, $TP_e \leqslant 10\%$, and $FP_e \leqslant 1$. The summary of these results from ten evaluations for a given σ_i and library size $\boldsymbol{\Theta}_N$ are presented in figure 3.

Looking at the results with the IDM data in figure 3(a), we see that the accuracy of SINDy-PI falls sharply with added noise of $\sigma \gtrsim 1 \times 10^{-4}$ for smaller libraries of $\boldsymbol{\Theta}_0$ to $\boldsymbol{\Theta}_2$ which increases to $\sigma \gtrsim 1 \times 10^{-6}$ at the largest library of $\boldsymbol{\Theta}_5$. A similar trend is observed for OVM data shown in figure 3(b), where SINDy-PI performance deteriorates with both increased noise levels and an increased library size. As the noise level increases, performance declines because the added noise magnitude is comparable to the smallest coefficient present in each model, as outlined in table 2. When the library size is large, this effect is more pronounced at smaller magnitudes of σ. These results are consistent with previous observations on SINDy's performance with noisy data [46, 65].



(a) SINDy-PI on IDM data with noise (b) SINDy-PI on OVM data with noise

Figure 3. Success rate of SINDy-PI for identifying (a) IDM and (b) OVM models for different library $\boldsymbol{\Theta}_N$ and measurement noise. The color scale represents the number of times SINDy-PI met the three criteria from the ten trials.

4. Conclusions and future work

The significance of the current study lies in the introduction of a novel framework for identifying sparse models of complex systems from data, eliminating the need for assumptions. This retention of sparsity is crucial as it preserves the underlying physics, rendering the discovered equations interpretable. This stands in contrast to other existing data-driven approaches such as neural networks, which often lack sparse functional relationships between variables. We demonstrate the effectiveness of our proposed data-driven framework using traffic modeling as an application, where traditional traffic models rely on assumptions. The framework operates through a two-step process. In the first step, IT metrics are used to gain insights into the range of interactions extracted from data, accurately discerning the directionality and extent of vehicle interaction. This initial step eliminates the need for assumptions commonly employed in sparse traffic models. The results obtained from the IT metrics assist in isolating the variables, narrowing them down to interactions between two adjacent vehicles for both the IDM and OVM models used in this study. Subsequently, these identified variables are used to establish a functional relationship, completing the traffic modeling process. Using synthetic data with a known ground truth, this study validates the framework. The validation of this framework holds significance in gaining insights into the anticipated behavior of these data-analytic tools and instilling confidence in their real-world application.

Beyond traffic modeling, the framework holds broader implications for modeling complex systems, as it can be adapted for various domains where variables influencing dynamics are unknown. The framework's sparsity-promoting system identification technique ensures the retention of the physics of the dynamics, facilitating the accurate discovery of relationships among variables while preventing overfitting.

Data availability statement

All data that support the findings of this study are included within the article (and any supplementary files).

Acknowledgments

This work is supported by the National Science Foundation CAREER Award (CMMI-2238359). The authors express their gratitude to Kadierdan Kaheman for the support provided in implementing SINDy-PI within Matlab.

Appendix A. Expressing IDM as rational function

The IDM equation from section 2.1 is given by

$$\frac{\mathrm{d}v}{\mathrm{d}t} = a\left(1 - \frac{v^{\delta}}{v_0^{\delta}}\right) - a\left(\frac{s^*\left(v\Delta v\right)^2}{s^2}\right), \tag{A.1}$$

where $s^* = s_0 + \max(0, vT + v\Delta v(2\sqrt{ab})^{-1})$. The expression $\max(0, vT + v\Delta v(2\sqrt{ab})^{-1})$ accounts for the scenario where the front vehicle is at rest. Utilizing a conditional statement with the ballistic update method, as discussed in section 2.1 (equation (3)), enables us to identify data points where the front vehicle is at rest. In our analysis, we exclude such terms from the dataset and exclusively consider instances when the vehicles are in motion. Consequently, s^* can be rewritten as $s^* = s_0 + vT + v\Delta v(2\sqrt{ab})^{-1}$.

We obtain a common denominator from equation (A.1):

$$\frac{\mathrm{d}v}{\mathrm{d}t} = \frac{a\left(v_0^\delta - v^\delta\right)s^2 - as^*\left(v, \Delta v\right)^2 v_0^\delta}{v_0^\delta s^2}. \tag{A.2}$$

Next, we select the states x_1 and u_1 as the positions of vehicle i and $i+1$, respectively, and write states as $\dot{x_1} = x_2 = v$ and $\dot{x_2} = \mathrm{d}v/\mathrm{d}t$. The bumper-to-bumper vehicle headway is $s = u_1 - x_1 - 4$, where the length of each vehicle 4 meters, and $\Delta v = u_2 - x_2$. Making these substitutions in equation (A.2), and evaluating with the parameter values $a, \delta, v_0, s_0, T, a, b$ as defined in section 2.1 we obtain:

$$\begin{aligned}\dot{x_1} &= x_2\\ \dot{x_2} &= \frac{\mathrm{N_{IDM}}}{\mathrm{D_{IDM}}}\end{aligned}$$

where

$$\begin{aligned}\mathrm{N_{IDM}} = {} & 3.6 + 2.4x_1 - 1.8x_2 + 0.3x_1^2 - 1.3074x_2^2 - 0.4743x_2^3 - 0.084\,33x_2^4\\ & - 0.000\,4977x_1x_2^4 - 0.000\,062\,21x_1^2x_2^4 - 2.4u_1 - 0.6u_1x_1\\ & + 0.6325u_2x_2 + 0.4743u_2x_2^2 + 0.1667u_2x_2^3 + 0.000\,4977u_1x_2^4\\ & + 0.000\,1244u_1x_1x_2^4 + 0.3u_1^2 - 0.083\,33u_2^2x_2^2 - 0.000\,062\,22u_1^2x_2^4\\ \mathrm{D_{IDM}} = {} & 16 + 8x_1 + x_1^2 - 8u_1 - 2u_1x_1 + u_1^2.\end{aligned}$$

The implicit form of the IDM equation is obtained as $\mathrm{D_{IDM}}\,\dot{x_2} = \mathrm{N_{IDM}}$, aligning with the coefficients presented in table 2. The coefficients provides the basis for comparing the results from SINDy-PI.

Appendix B. Expressing OVM as rational function

The OVM equation from section 2.1 is given by:

$$\frac{\mathrm{d}v}{\mathrm{d}t} = a_h\left[V(s) - \frac{\mathrm{d}x}{\mathrm{d}t}\right], \tag{B.1}$$

where OV function $V(s) = \alpha\tanh[\beta(s - s_0)] + v_0$. We use the Pade approximation of order [1,2] to approximate the hyperbolic tangent function as:

$$\tanh_{[1,2]}(x) = \frac{3x}{3 + x^2}.$$

Substituting $\tanh_{[1,2]}(\beta(s - s_0))$ into the OV function, we obtain:

$$V(s) = \frac{3\alpha\beta\left(s - s_0\right)}{3 + \beta^2\left(s - s_0\right)^2} + v_0. \tag{B.2}$$

Substituting equation (B.2) in equation (B.1) and obtaining a common denominator yields:

$$\frac{\mathrm{d}v}{\mathrm{d}t} = \frac{a_h\left[3\alpha\beta\left(s - s_0\right) + \left(v_0 - \mathrm{d}x/\mathrm{d}t\right)\left(3 + \beta^2\left(s - s_0\right)^2\right)\right]}{3 + \beta^2\left(s - s_0\right)^2}. \tag{B.3}$$

Similar to IDM, we select the states x_1 and u_1 as the positions of vehicle i and $i+1$, respectively, and write states as $\dot{x_1} = x_2 = v$ and $\dot{x_2} = \mathrm{d}v/\mathrm{d}t$. In OVM, the distance headway is $s = u_1 - x_1$. Substituting these values into equation (B.3) and evaluating with the parameters $a_h, \alpha, \beta, s_0, v_0$ as defined in section 2.2 we obtain:

$$\begin{aligned}\dot{x_1} &= x_2\\ \dot{x_2} &= \frac{\mathrm{N_{OVM}}}{\mathrm{D_{OVM}}}\end{aligned}$$

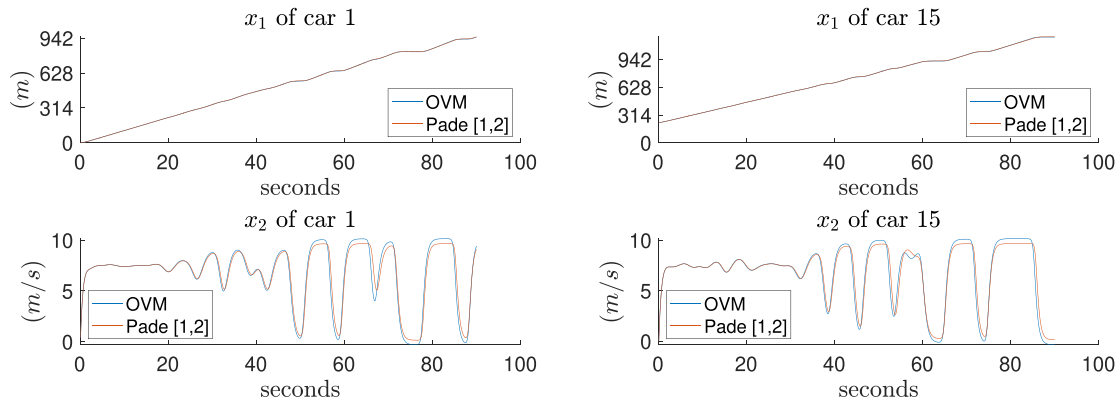


Figure B1. Comparison of OVM data generated using the original hyperbolic tangent function and its corresponding Pade approximation for vehicles 1 and 15 from the simulation of jammed traffic.

where

$$\begin{aligned}\mathrm{N_{OVM}} &= 26.4497 + 10.9867x_1 - 25.806x_2 - 4.484\,84x_1x_2 \\ &\quad + 1.207\,46x_1^2 - 0.246\,42x_1^2x_2 - 10.9867u_1 - 2.414\,92u_1x_1 \\ &\quad + 4.484\,84u_1x_2 + 0.492\,84u_1x_1x_2 + 1.207\,46u_1^2 - 0.246\,42u_1^2x_2 \\ \mathrm{D_{OVM}} &= 14.3367 + 2.491\,58x_1 + 0.1369x_1^2 - 2.491\,58u_1 - 0.2738u_1x_1 + 0.1369u^2.\end{aligned}$$

The implicit form of the OVM equation is obtained as $\mathrm{D_{OVM}}\,\dot{x_2} = \mathrm{N_{OVM}}$, aligning with the coefficients presented in table 2. The coefficients provides the basis for comparing the results from SINDy-PI.

B.1. Comparison of the OVM using hyperbolic tangent and Pade approximations

The simulation data used to assess SINDy's performance is generated using the hyperbolic tangent representation, consistent with the true OV model as visualized in figure B1. We utilize the Pade approximation of the OVM model as a basis for comparison with the SINDy-PI results, which utilizes a polynomial basis library.

ORCID iD

Subhradeep Roy https://orcid.org/0000-0002-5740-188X

References

[1] Haman I T, Kamla V C, Galland S and Kamgang J C 2017 *Proc. Comput. Sci.* **109** 887–92
[2] Pasquale C, Sacone S, Siri S and Ferrara A 2019 *Annu. Rev. Control* **48** 312–24
[3] Kagho G O, Balac M and Axhausen K W 2020 *Proc. Comput. Sci.* **170** 726–32
[4] Ali F, Khan Z H, Altamimi A B, Khattak K S and Gulliver T A 2023 *Appl. Sci.* **13** 7234
[5] Treiber M and Kesting A 2010 *IEEE Intell. Transp. Syst. Mag.* **2** 6–13
[6] Krajzewicz D, Erdmann J, Behrisch M and Bieker L 2012 *Int. J. Adv. Syst. Meas.* **5** 128–38 (available at: www.iariajournals.org/systems_and_measurements/tocv5n12.html)
[7] Chai R, Hou Y, Fu B and He Y 2023 Simulation modeling of typical urban traffic congestion areas based on SUMO *6th Int. Conf. on Information Communication and Signal Processing* pp 853–9 (available at: https://ieeexplore.ieee.org/abstract/document/10390622?casa_token=wIb5ZuMOy1sAAAAA:30ZdZXAvH-CTdCagGfRJTgfXXe4Aojs6fQ5OECjBiCQos5yAZa9ejvieGrkHHI9_fQbEwiQGYd8)
[8] Yang Q I and Koutsopoulos H N 1996 *Transp. Res.* C **4** 113–29
[9] Wang Z, Shi Y, Tong W, Gu Z and Cheng Q 2023 *J. Transp. Eng.* A **149** 04023075
[10] Aghabayk K, Sarvi M and Young W 2015 *Transp. Rev.* **35** 82–105
[11] Wang Z, Shi X and Li X 2019 *J. Indian Inst. Sci.* **99** 589–99
[12] Rahman M, Chowdhury M, Xie Y and He Y 2013 *IEEE Trans. Intell. Transp. Syst.* **14** 1942–56
[13] Akcelik R 2007 A review of gap-acceptance capacity models *29th Conf. of Australian Institutes of Transportation Research (Adelaide, South Australia, Australia)* (available at: https://trid.trb.org/Results?q=&serial=%22CONFERENCE%20OF%20AUSTRALIAN%20INSTITUTES%20OF%20TRANSPORT%20RESEARCH%20(CAITR),%2029TH,%202007,%20ADELAIDE,%20SOUTH%20AUSTRALIA,%20AUSTRALIA%22)
[14] Zhao J, Knoop V L and Wang M 2023 *Transp. Sci.* **57** 135–55
[15] Gipps P G 1981 *Transp. Res.* B **15** 105–11
[16] Treiber M, Hennecke A and Helbing D 2000 *Phys. Rev.* E **62** 1805–24
[17] Bando M, Hasebe K, Nakayama A, Shibata A and Sugiyama Y 1995 *Phys. Rev.* E **51** 1035

[18] Kikuchi S and Chakroborty P 1992 *Transp. Res. Rec.* **1365** 82–91 (available at: https://onlinepubs.trb.org/Onlinepubs/trr/1992/1365/1365-009.pdf)
[19] McDonald M, Wu J and Brackstone M 1997 Development of a fuzzy logic based microscopic motorway simulation model *Proc. Conf. on Intelligent Transportation Systems* pp 82–87
[20] Hongfei J, Zhicai J and Anning N 2003 Develop a car-following model using data collected by "five-wheel system" *IEEE Int. Conf. on Intelligent Transportation Systems* vol 1 pp 346–51
[21] Panwai S and Dia H 2007 *IEEE Trans. Intell. Transp. Syst.* **8** 60–70
[22] Nagel K and Schreckenberg M 1992 *J. Physique* I **2** 2221–9
[23] Waraich R A, Charypar D, Balmer M and Axhausen K W 2015 *Computational Approaches for Urban Environments* (Springer) pp 211–33
[24] Wibral M, Lizier J, Vogler S, Priesemann V and Galuske R 2014 *Front. Neuroinform.* **8** 1
[25] Jirsa V and Sheheitli H 2022 *J. Phys. Complex.* **3** 015007
[26] Varley T F, Pope M, Faskowitz J and Sporns O 2023 *Commun. Biol.* **6** 451
[27] De Lellis P, Marin M R and Porfiri M 2022 *J. Phys. Complex.* **4** 015001
[28] Roy S, Howes K, Muller R, Butail S and Abaid N 2019 *Entropy* **21** 42
[29] Zhang P, Rosen M, Peterson S D and Porfiri M 2018 *J. Fluid Mech.* **848** 968–86
[30] Butail S, Mwaffo V and Porfiri M 2016 *Phys. Rev.* E **93** 042411
[31] Das R and Porfiri M 2023 *J. Phys. Complex.* **4** 025020
[32] Hlinka J, Hartman D, Vejmelka M, Runge J, Marwan N, Kurths J and Palus M 2013 *Entropy* **15** 2023–45
[33] Sattari S, Basak U S, James R G, Perrin L W, Crutchfield J P and Komatsuzaki T 2022 *Sci. Adv.* **8** eabj1720
[34] Basak U S, Sattari S, Hossain M, Horikawa K and Komatsuzaki T 2021 *Biophys. Physicobiol.* **18** 131–44
[35] Butail S and Porfiri M 2019 *Chaos* **29** 011102
[36] Barak-Ventura R, Marin M R and Porfiri M 2022 *Patterns* **3** 100546
[37] Roy S and Abaid N 2017 *R. Soc. Open Sci.* **4** 170130
[38] Shaffer I and Abaid N 2020 *Entropy* **22** 1176
[39] Marschinski R and Kantz H 2002 *Eur. Phys. J.* B **30** 275–81
[40] Roy S 2020 *Chaos* **30** 113125
[41] Liu Z, Wang Y, Cheng Q and Yang H 2022 *IEEE Trans. Intell. Transp. Syst.* **23** 18012–23
[42] Assadi A, Tkachenko P and Del Re L 2021 Interaction models for merging and cut-in scenarios *2021 European Control Conf. (ECC)* pp 2346–51
[43] Lane D and Roy S 2023 Using information theory to detect model structure with application in vehicular traffic systems *IFAC-PapersOnLine* **56** 367–72
[44] Brunton S L, Proctor J L and Kutz J N 2016 *Proc. Natl Acad. Sci.* **113** 3932–7
[45] Champion K, Lusch B, Kutz J N and Brunton S L 2019 *Proc. Natl Acad. Sci.* **116** 22445–51
[46] Kaheman K, Kutz J N and Brunton S L 2020 *Proc. R. Soc.* A **476** 20200279
[47] Boninsegna L, Nuske F and Clementi C 2018 *J. Chem. Phys.* **148** 241723
[48] Loiseau J C, Noack B R and Brunton S L 2018 *J. Fluid Mech.* **844** 459–90
[49] Alves E P and Fiuza F 2022 *Phys. Rev. Res.* **4** 033192
[50] Kaptanoglu A A, Hansen C, Lore J D, Landreman M and Brunton S L 2023 *Phys. Plasmas* **30** 033906
[51] Sorokina M, Sygletos S and Turitsyn S 2016 *Opt. Express* **24** 30433–43
[52] Lakshminarayana S, Sthapit S and Maple C 2022 *Sustainability* **14** 2051
[53] Treiber M and Kanagaraj V 2015 *Physica* A **419** 183–95
[54] Tadaki S I, Kikuchi M, Fukui M, Nakayama A, Nishinari K, Shibata A, Sugiyama Y, Yosida T and Yukawa S 2013 *New J. Phys.* **15** 103034
[55] Nakayama A, Kikuchi M, Shibata A, Sugiyama Y, Tadaki S-I and Yukawa S 2016 *New J. Phys.* **18** 043040
[56] Shannon C E 1948 *Bell Syst. Tech. J.* **27** 379–423
[57] Schreiber T 2000 *Phys. Rev. Lett.* **85** 461–4
[58] Palus M, Komarek V, Hrncir Z and Sterbova K 2001 *Phys. Rev.* E **63** 046211
[59] Sun J and Bollt E M 2014 *Physica* D **267** 49–57
[60] Lizier J T 2014 *Front. Robot. AI* **1** 11
[61] Kraskov A, Stogbauer H and Grassberger P 2004 *Phys. Rev.* E **69** 066138
[62] Baker G A and Gammel J 1961 *J. Math. Anal. Appl.* **2** 21–30
[63] Perev K 2022 *Proc. Tech. Univ. Sofia* **72** 7
[64] Li L, Jiang R, He Z, Chen X M and Zhou X 2020 *Transp. Res.* C **114** 225–40
[65] Didonna M, Stender M, Papangelo A, Fontanela F, Ciavarella M and Hoffmann N 2019 *Lubricants* **7** 64