Transportation Research Part B 80 (2015) 185–201

Contents lists available at ScienceDirect

Transportation Research Part B

journal homepage: www.elsevier.com/locate/trb

A simple nonparametric car-following model driven by field data

Zhengbing He[a], Liang Zheng[b], Wei Guan[a,*]

[a] MOE Key Laboratory for Urban Transportation Complex Systems Theory and Technology, Beijing Jiaotong University, Beijing, China
[b] School of Traffic and Transportation Engineering, Central South University, Changsha, China

ARTICLE INFO

Article history:
Received 26 November 2014
Received in revised form 14 July 2015
Accepted 15 July 2015

Keywords:
Car-following model
Data-driven approach
k-nearest neighbour
Traffic oscillation
Fundamental diagram

ABSTRACT

Car-following models are always of great interest of traffic engineers and researchers. In the age of mass data, this paper proposes a nonparametric car-following model driven by field data. Different from most of the existing car-following models, neither driver's behaviour parameters nor fundamental diagrams are assumed in the data-driven model. The model is proposed based on the simple *k*-nearest neighbour, which outputs the average of the most similar cases, i.e., the most likely driving behaviour under the current circumstance. The inputs and outputs are selected, and the determination of the only parameter *k* is introduced. Three simulation scenarios are conducted to test the model. The first scenario is to simulate platoons following real leaders, where traffic waves with constant speed and the detailed trajectories are observed to be consistent with the empirical data. Driver's rubbernecking behaviour and driving errors are simulated in the second and third scenarios, respectively. The time–space diagrams of the simulated trajectories are presented and explicitly analysed. It is demonstrated that the model is able to well replicate periodic traffic oscillations from the precursor stage to the decay stage. Without making any assumption, the fundamental diagrams for the simulated scenario coincide with the empirical fundamental diagrams. These all validate that the model can well reproduce the traffic characteristics contained by the field data. The nonparametric car-following model exhibits traffic dynamics in a simple and parsimonious manner.

© 2015 Elsevier Ltd. All rights reserved.

1. Introduction

Car-following models are always of great interests of traffic engineers and researchers. As one of the most important traffic analysis tools, the first car-following models (Pipes, 1953; Chandler et al., 1958) were proposed as early as sixty years ago. From then on, a large number of car-following models were addressed. For example, the optimal velocity model (Bando et al., 1995) and the full velocity difference model (Jiang et al., 2001) were proposed based on the intuitive speed changing of vehicles, and the well-known Gipps model (Gipps, 1981) and the intelligent driver model (Treiber et al., 2000) were proposed in the perspective of driver's acceleration and deceleration strategies. More recently, to study the trigger and formation of the stop-and-go traffic oscillations, Laval and Leclercq (2010) and Chen et al. (2012b) incorporated non-equilibrium (timid and aggressive) driver behaviour into Newell's simplified car-following model (Newell, 2002). Laval et al. (2014) captured traffic oscillations by using a desire acceleration model with a white noise term. It showed that small driver errors can result in the

[*] Corresponding author.
E-mail addresses: he.zb@hotmail.com (Z. He), zhengliang@csu.edu.cn (L. Zheng), weig@bjtu.edu.cn (W. Guan).

http://dx.doi.org/10.1016/j.trb.2015.07.010
0191-2615/© 2015 Elsevier Ltd. All rights reserved.

stop-and-go oscillation. Detailed introduction about more car-following models can be found in the review articles (see e.g. Brackstone and Mcdonald, 1999; Hoogendoorn, 2001; Helbing, 2001; Saifuzzaman and Zheng, 2014). Performance comparison amongst important car-following models can be found in the recently published book (Treiber and Kesting, 2013b). In the work, the simulated time–space diagram and the fundamental diagrams were used as core elements to evaluate performance in a figure so-called "fact sheet".

A process prior to analysing or simulating real traffic dynamics is calibration, including the fundamental diagrams, driver's behaviour parameters, etc. Only calibrated models are able to approximate real traffic dynamics and conditions. Calibration, however, is a challenging topic needed to be deeply investigated (see e.g. Brockfeld et al., 2004; Kesting and Treiber, 2008; Punzo et al., 2012; Treiber and Kesting, 2013a). Even the shapes of the fundamental diagrams in assumptions are still controversial (see e.g. Kerner, 1998; Schönhof and Helbing, 2009; Treiber et al., 2010).

Recently arising high-fidelity traffic data and data-driven modelling approaches provide an opportunity to stride over the modelling and calibrating, and to extract traffic dynamics and driver's behaviour directly from mass field data. The "manual" errors produced in modelling and calibrating are not a concern of these approaches. Artificial neural networks (ANN), support vector regression (SVR), nonparametric regression, etc. are all prevailing data-driven modelling approaches. There approaches have been widely applied and investigated in the field of short-term traffic flow forecasting; see e.g. Zhang and Ge (2013), Wu et al. (2004) and Zheng and Su (2014) and a recent review article (Vlahogianni et al., 2014). In recent years more high-fidelity traffic data, such as the trajectory data provided by Next Generation Simulation Project (NGSIM, 2006), are collected and released. This makes it possible to model car-following behaviour directly from a large number of field data.

In the studies utilising the data-driven approaches to model driver's behaviour, Panwai and Dia (2007) proposed a car-following model by using ANN with the back-propagation and fuzzy ARTMAP architectures. The model classified field data into five driving modes. Before moving, each vehicle was classified into a mode first. Space headway and leader's speed were the inputs of ANN and the follower's speed were the output. The dataset for training was limited, which was collected by an instrumented following vehicle for only 300 s. Simulation results showed that the model outperformed the Gipps and psychophysical models. Colombaroni and Fusco (2014) also modelled car-following behaviour by using ANN. Car-following data were collected by using several global-positioning-system-equipped vehicles that followed each other on urban roads. More variables were added into the training inputs in order to incorporate driver's reaction delay. Nevertheless, the analyses and comparisons based on several immediate followers might be insufficient in the above two literature. Forward-propagating traffic waves triggered by leader's deceleration could be observed in Fig. 13 in Colombaroni and Fusco (2014), which might be unrealistic. In Khodayari et al. (2012), four inputs, i.e., reaction delay, relative speed, relative distance and follower's speed, were chosen to train ANN, and the follower's acceleration was output. Although the proposed model was studied and compared, the study was only limited in leader–follower pairs. Zheng et al. (2013) further presented an explicit reaction delay model based on ANN, and tested the model in a nine-vehicle platoon following a given leader. By utilising SVM, Wei and Liu (2013) investigated the asymmetric characteristic in car-following behaviour and its impact on traffic flow evolution. Space headway, follower's speed, and relative speed were taken as the inputs, and follower's speed was output. Similarly, in the model test only an immediate follower was simulated and compared with the real immediate one.

In the existing literature of data-driven modelling, most of the inputs and outputs, such as reaction delay and acceleration, are continuous variables, which are also indirect to computer simulation; only a few followers are simulated to test the model, which is insufficient; ANN and SVR are complex in architecture. These are expected to be improved, in particular when apply them in a simulation scenario. To this end, the paper proposes a simple nonparametric car-following model driven by field data. In this model, all inputs and outputs are based on vehicle positions, which are straightforward to reproduce traffic dynamics; thousands of vehicles are simulated to test and validate the model; *k*-nearest neighbour, one of the most simplest data-driven approach, is employed, which is simple and parsimonious. It is demonstrated that without making any assumption or calibration, the model is able to well replicate important traffic characteristics contained by the field data, such as the traffic oscillations from the precursor stage to the decay stage and the fundamental diagrams.

The remainder of the paper is organised as follows: The field data driving the nonparametric model are introduced in Section 2; the model is proposed in Section 3; a simulation scenario of platoons following real vehicles is presented in Section 4 in order to primarily validate the model; Section 5 simulates a 1.25 km roadway with rubbernecking, and the results are thoroughly analysed and compared; Section 6 introduces driver errors into the model, and the response of the model to small perturbations is demonstrated; conclusions are made at last.

2. Site description and the data

The nonparametric car-following model proposed in the paper is driven by a large number of field data. The well-known NGSIM providing high-fidelity trajectory dataset satisfies the demand. We propose and analyse the nonparametric car-following model based on the trajectory dataset collected on a 6-lane segment in the vicinity of Lankershim Avenue on southbound US-101 freeway in Los Angeles, California. The time period of data collection ranges from 7:35 a.m. to 8:35 a.m. on June 15, 2005. We adopt the trajectory dataset collected on Lane 1 (median lane), 2, and 3, where lane changes are relatively few. Most of the speeds are less than 60 km h^{-1} (except a few faster vehicles leaving the segment), which indicates that the segment is congested during the study period. Stop-and-go oscillations originate and propagate upstream at a constant wave speed of approximately 16 km h^{-1}. The original dataset provides a data sample every 0.1 s. This paper simply averages these data sample every 1 s (arithmetic mean) to smooth out the detection errors (for the feature of the errors, see e.g. Thiemann et al. (2008); Punzo et al. (2011); Montanino and Punzo (2013)). It is also consistent with the selected simulation time step in the upcoming simulation scenarios.

3. Nonparametric car-following model

3.1. k-nearest neighbour approach

k-nearest neighbour (*k*NN) is a nonparametric approach that is very simple to understand but works incredibly well in practice. The approach selects the most similar historical cases, and takes the average of their outputs as the estimate of this time. Specifically, the approach estimates $\mathbf{y}_0$ in focal $(\mathbf{x}_0, \mathbf{y}_0)$ as follows.

$$\hat{\mathbf{y}}_0 = \frac{\sum_{i=1}^{k} \mathbf{y}_i}{k} \tag{1}$$

where $\mathbf{x}_i$ with respect to $\mathbf{y}_i$ is one of the k most similar samples to $\mathbf{x}_0$. Distance between points or vectors in space is usually used to measure the similarity. A basic assumption of *k*NN is that history are repeating. It is reasonable to believe that drivers are repeating their behaviour in the similar circumstances including traffic and geography conditions. Thus, it is suitable for *k*NN to model car-following behaviour.

There are still several more advanced nonparametric approaches, such as kernel estimation and loess (Cleveland, 1979). However, more parameters or functions are introduced in these approaches. To keep parsimony of the model, we adopt *k*NN, where only k is needed to be specified.

3.2. Inputs and outputs of the model

It is known that when a vehicle follows a leader, the movement of the vehicle is closely related to the speed and acceleration of its leader, as well as its speed, acceleration, and space headway. In the model we directly input vehicle positions, and output vehicle positions. The inputs and outputs of positions avoid numerically solving differential equations that are required by many traditional parametric car-following models (Chapter 10.2 in Treiber and Kesting (2013b)), and can be directly used in simulation.

Specifically, we propose the input vector of *k*NN (refer to Fig. 1) as

$$\mathbf{x}_n(t+\tau) = (d_{n-1}(t+\tau),\ d_{n-1}(t),\ s_n(t),\ s_n(t-\tau)). \tag{2}$$

where τ is the simulation time step; $(n-1)$ is the leader of vehicle n; $d_{n-1}(t)$ is the moving distance of the leader between time $(t-\tau)$ and t given as:

$$d_{n-1}(t) = l_{n-1}(t) - l_{n-1}(t-\tau) \tag{3}$$

where $l_n(t)$ is the position of vehicle n at time t; $s_n(t)$ is the space headway of vehicle n at time t:

$$s_n(t) = l_{n-1}(t) - l_n(t) \tag{4}$$

In this paper, we set τ to be 1 s, and the data samples in the database are averaged every 1 s accordingly. Note that the moving distance corresponds to the speed of a vehicle, i.e., moving distance over a simulation step, and two successive moving distances imply the acceleration of a vehicle. Through the space headway, the speed and acceleration of both the leader and follower are included in the inputs.

The output is the moving distance of vehicle n, i.e.,

$$\mathbf{y}_n(t+\tau) = d_n(t+\tau) \tag{5}$$

Theoretically, the inputs and outputs will not introduce back-forward movement. The estimate of *k*NN is an average of selected historical cases. No negative moving distance exists in field data, and taking average of non-negative numbers will not result in a negative number.

In addition, the computational efficiency is usually concerned when we apply or train a data-driven model. In particular, the input of this model is four-dimensional, which further increases the burden. Reducing search time is a hot topic in computer science. A k-dimensional tree is a useful method, which can reduce the time complexity of *k*NN. Making use of a hash table can put the searching of four-dimensional samples off-line, and then desirable simulation speed would be achieved. These are all effective speed-up methods for a data-driven model in computer science.

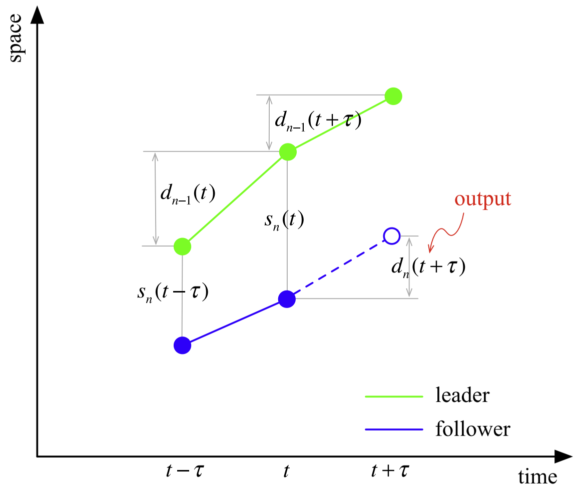


Fig. 1. Inputs and outputs of the proposed nonparametric car-following model.

3.3. Elimination of autocorrelation

Because of strong autocorrelation of time-series trajectory data, it is easy for all k samples coming from a leader–follower pair, in particular when the leader and follower move in a constant speed. Such dominance could reduce the reliability of the model. To overcome this issue, we make all k samples selected from different leader–follower pairs.

3.4. Distance between two data samples

This paper adopts the "ordinary" and simple scaled Euclidean distance, i.e., adjusting the input $x_{ji} \in \mathbf{x}_i$ by its mean $\bar{x}_j$ and standard deviation S_j before calculating Euclidean distance in a metric space, where j indicates an element in an input vector. The formula reads

$$\mathcal{D}(\mathbf{x}_i, \mathbf{x}_0) = \sqrt{\sum_{j=1}^{J} (z_{ji} - z_{j0})^2} \tag{6}$$

where

$$z_{ji} = \frac{x_{ji} - \bar{x}_j}{S_j}, \tag{7}$$

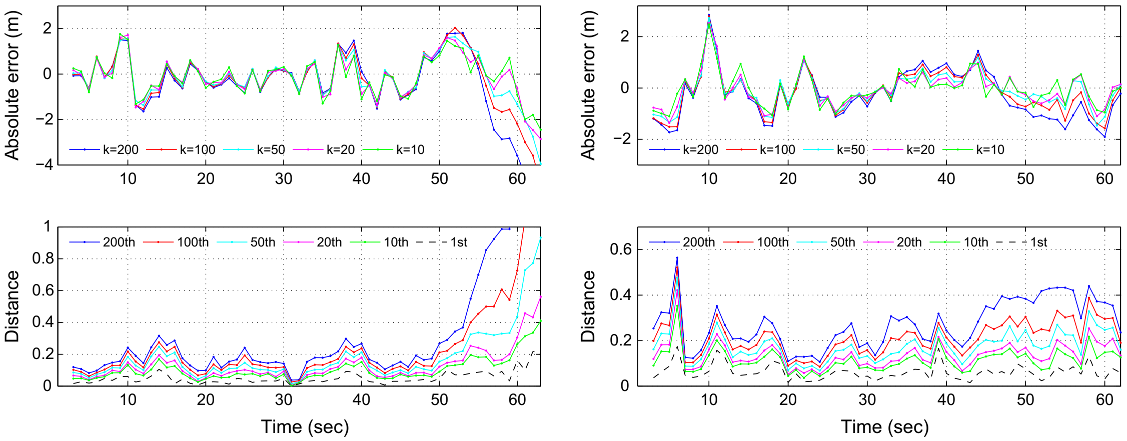
and J is the total number of all elements in an input vector.

3.5. Determination of k and similarity

The parameter k in kNN indicates the number of the historical cases that are considered to be similar to the estimated case. It apparently relates to the underlying database. Usually, k is estimated by experience. In the paper, we determine k by comparing estimation errors under different k-values. To measure the similarity of the k selected samples to the estimated case, we employ the scaled Euclidean distance of kth nearest sample to the estimated case, which is the longest distance in all k samples. Denote by $\mathcal{D}_k$ the distance.

Before introducing the determination, we first build the following three databases with different sizes by using the datasets collected on Lane 2 and 3. (i) Database 1: the small-size database containing 22,202 input–output samples, which is built by using the dataset collected on Lane 2 during the first 15 min; (ii) Database 2: the medium-size database containing 78,683 samples, which is built by using the dataset collected on Lane 2 during the all study 45 min; (iii) Database 3: the large-size database containing 152,637 samples, which is built by using the dataset collected on Lane 2 and 3 during the all study 45 min. No data from Lane 1 is included in these three databases. Such databases are employed to estimate the movement of the followers on Lane 1. The purpose is to avoid involving the empirical followers to be estimated into the database.

Now we estimate the followers of two typical vehicles (i.e., Vehicle 422 and 1989) who traverse stop-and-go oscillations on Lane 1, and present the detail in Fig. 2. In each estimation, the inputs are all real field data, and the output is estimated moving distance of the follower. Database 3 is used in this test, and the absolute error in the figure is the absolute difference between the estimated and real values. From the upper plots of Fig. 2, it can be seen that smaller k basically makes smaller



(a) Following Vehicle 422 on Lane 1

(b) Following Vehicle 1989 on Lane 1

Fig. 2. Estimation errors under different k-values, and the distance $\mathcal{D}_k$ between estimated and historical cases. The leaders are Vehicle 422 and 1989 on Lane 1 on the US-101 segment, respectively. Database 3 is used.

errors, but the errors may fluctuate more. One can compare the green and blue lines in the figure, for example. It is easily understood that the closer historical samples better reflect estimated case, but averaging a small number of samples results in instability. This is a general tendency but not always true, because we can also see a few good estimations with long distances (for example, the points with absolute errors of around zero in the blue line), which relates to the quality of database and the number of similar historical samples. The lower plots of Fig. 2 present the distance $\mathcal{D}_k$ of each estimation. The figure shows that for Database 3, when $k = 10$, the estimations are usually good with the distance smaller than 0.2, i.e. $\mathcal{D}_k < 0.2$.

To make a more general conclusion, we estimate the movement of more followers. We select automobile leader–follower pairs who move over 200 m on Lane 1 during the first 15 min. These vehicles are considered to be stable on Lane 1, i.e., having little impact from lane-changing. To be representative, a part of these pairs (50 pairs, here) are randomly selected from total 371 pairs on Lane 1. Database 3 is used, and total 2918 estimations are made. Note that using a random part of vehicles instead of all is to show that a randomly selected part of samples may be sufficient to determine a proper k. Besides, as the traffic on Lane 1 is similar to those on Lane 2 and 3, Lane 1 may be able to reflect the driving behaviour contained by the database. Similar tendency of k-values is observed when we increase the size of the part.

The statistical results are illustrated in Fig. 3. From the upper plot in Fig. 3, it can be seen that there indeed exists an optimal k-value for minimum absolute errors. For Database 3, the value is around 10. When k-value is smaller or bigger than the optimal value, the mean and standard deviation of the absolute errors become larger simultaneously. From the lower plot in Fig. 3, it can be seen that the distance is positively correlated to k-values. Moreover, for the optimal $k = 10$, the distance $\mathcal{D}_k$ (a mean plus a standard deviation) is smaller than 0.2. It implies that if we set $k = 10$, an estimation is usually good when $\mathcal{D}_k < 0.2$. It is noticed that the estimation is not very sensitive to the k-value for the database. Seeing Fig. 3, specifying $k = 9$ or $k = 11$ for Database 3 may be also able to result in satisfied results.

The size of the database also affects the estimation; see Fig. 4. In the figure, k is fixed to be 10, and Database 1, 2, and 3 are used to estimate, respectively. It is clear that the larger Database 3 results in better estimations (i.e., lower mean and standard deviation of the absolute errors), and further shortens the distance $\mathcal{D}_k$.

Therefore, we specify $k = 10$ for the database built on the US-101 dataset, and an estimation is considered to be satisfied if $\mathcal{D}_k < 0.2$. In practice, one can determine k by comparing the estimation results with different k-values as this subsection did. This determination is a premise to apply kNN, because the optimal k-value is highly related to the underlying database.

In a simulation scenario, any situation may occur, and we cannot guarantee that there are historical cases that are close to every estimated case. It thus requires us to monitor $\mathcal{D}_k$ in order to judge if an estimation is within the scope of the underlying database. A time–space diagram of $\mathcal{D}_k$ is designed to play the role in Section 5. Large distance means the simulation result may not be reasonable, and substitution is needed. One may add more field data in the database, or simulate this case using a traditional parametric car-following model.

3.6. Analysis on avoiding collisions

The kNN approach is based on the assumption that history repeats. Generally speaking, averaging similar collision-free historical cases would not lead to a collision. As a data-driven approach, however, the collision-free may not be mathematically proved. Instead, we can show the collision-free in statistic. To this end, we calculate the relative error of the estimated space headway as follows:

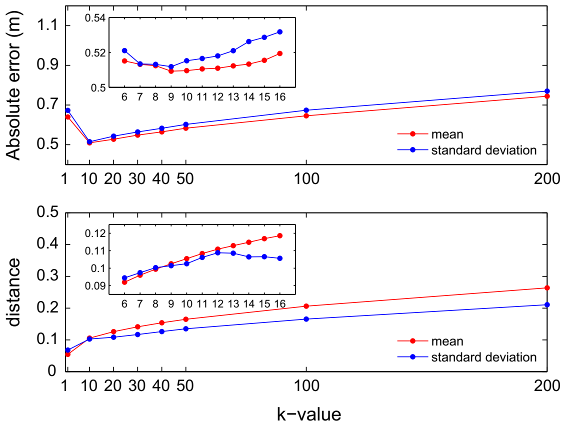


Fig. 3. Estimation results of following 50 randomly selected leaders on Lane 1 by using Database 3 and various *k*-values.

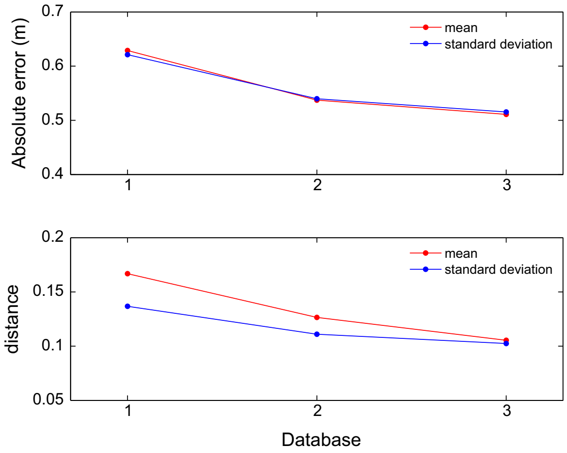


Fig. 4. Estimation results of following 50 randomly-selected leaders on Lane 1 by using Database 1–3 and *k* = 10.

$$\mathrm{RE} = \frac{s_n^{\mathrm{sim}}(t) - s_n^{\mathrm{data}}(t)}{s_n^{\mathrm{data}}(t)} \tag{8}$$

where $s_n^{\mathrm{sim}}(t)$ and $s_n^{\mathrm{data}}(t)$ are the estimated and ground-truth data, respectively. All automobile leader–follower pairs (total 371 pairs) moving over 200 m on lane 1 is estimated by using Database 3. In the results, 19,035 out of 20,733 estimations (91.8%) have the distances $\mathcal{D}_k$ smaller than 0.2, and the results are presented in Fig. 5. It shows that the relative errors are around zero and the maximum negative error is about -0.3. For a minimum space headway of around 6 m, the magnitude of the relative errors implies collision-free in a single estimation.[1]

There is still an extreme situation needed to be treated carefully. When both the leader and follower are in a standstill at time t (i.e., $d_{n-1}(t) = 0$ and $s_n(t) = s_n(t-\tau)$) and the leader has not moved yet at time $(t+\tau)$ (i.e., $d_{n-1}(t+\tau) = 0$), the movement of the follower at time $(t+\tau)$ can be predetermined to be zero, and no estimation is needed.[2] Whilst in the situation, *k*NN may make the standstill follower begin to approach its standstill leader, because *k*NN averages similar samples but not exactly same samples, and any selected historical sample outputting a (even very small) positive movement of the follower will result in the unrealistic movement. Thus, when $d_{n-1}(t) = 0, s_n(t) = s_n(t-\tau)$ and $d_{n-1}(t+\tau) = 0$, we should directly set the output $d_n(t+\tau) = 0$.

[1] Although no collision has been observed in our analysis, it may occur when the model is applied to other data due to the data-driven nature of *k*NN.

[2] In practice, it is uneasy to capture that $s_n(t) = s_n(t-\tau)$ due to the discrete nature of compute simulation. One could approximate it with $s_n(t), s_n(t-\tau) \leqslant s_{\min}$.

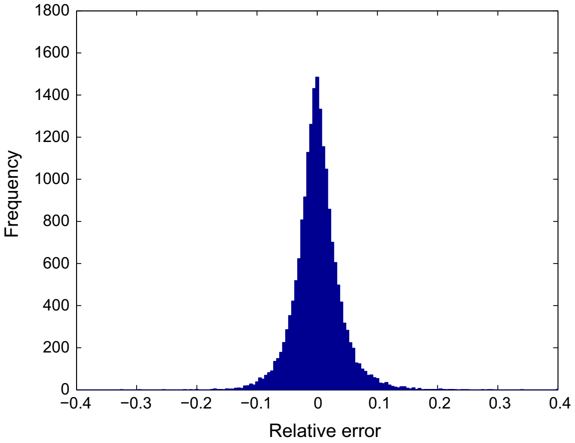


Fig. 5. A histogram plot for the relative errors of the estimated space headways when $\mathcal{D}_k < 0.2$. Total 371 vehicles from Lane 1 are estimated using Database 3.

3.7. Transferability of the model and the database

Considering the basic assumption of *k*NN, i.e., drivers repeat their behaviour in similar circumstances, an underlying database could be well transferred to any site with similar circumstances including driving habits, roadway geometry, etc. To show the transferability of the model and its database, we estimate driving behaviour from other site by using the US-101 database. Another well-known trajectory dataset collected from I-80 by NGSIM is ideal to do this. The I-80 segment is located in Emeryville, California. The length is about 500 m, and there are six lanes. From the left (median lane) to the right, we label them as Lane 1–6. Note that Lane 1 is a High Occupancy Vehicle lane. The trajectory data were collected for 45 min (4:00–4:15 p.m. and 5:00–5:30 p.m.) on April 13, 2005. We may consider the driving behaviour on the I-80 and US-101 segments are similar, because both of them are typical freeways in U.S.

This test for transferability estimates the movement of the followers on Lane 2 and 3 in the I-80 segment by using Database 3 built on the US-101 database. Similarly to the above approach, we obtain 347 and 206 automobile leader–follower pairs who move over 200 m on the lanes during the first 15 min. For Lane 2, we make 17,447 out of 18,186 (95.9%) estimations with $\mathcal{D}_k < 0.2$, and for lane 3 it is 8905 out of 9359 (95.1%) estimations. Fig. 6 presents the histogram plots for the relative errors of the estimated space headway (see Eq. (8)). It can be seen that most of the relative errors are around 0 and between -0.1 and 0.1, and the maximum deviations are about -0.2 and 0.2 (except really few estimation with large deviation). It indicates that the US-101 database could be transferred to the case of I-80.

4. A simulation scenario of a platoon following a real vehicle

To test this proposed nonparametric car-following model, this section simulates platoons following real vehicles. Two typical vehicles who traverse stop-and-go oscillations on Lane 1, i.e., Vehicle 422 and 1989, are employed as the empirical leaders. Database 3 without the data coming from Lane 1 is used.

4.1. A platoon with real boundary conditions

This subsection estimates platoons following the real leaders with real entry boundary conditions. The real entry boundary conditions mean that the entry speed and time intervals of simulated followers are the same to those of the real followers. At each simulation step, we input the current situation of the estimated vehicle into the database, search for the *k* nearest samples, and average their output as an estimate of its movement.

Total 23 and 31 followers are simulated in the platoons following the two vehicles, respectively. The time–space diagrams of the real and simulated trajectories are plotted and compared in Fig. 7. We mark the estimated cases with the distance $\mathcal{D}_k$ in order to monitor the similarity between the estimated case and the database. In the simulation scenario most of $\mathcal{D}_k$ are smaller than 0.1, which means that the database is proper to underlie the simulation scenario.

From Fig. 7, it can be seen that the time–space shapes of simulated platoons are close to the real ones. The wave speeds revealed in the simulated and real platoons are all about 16 km h^{-1}. It means that the congestion propagates in similar speed. To compare more clearly, we additionally plot the last vehicles of the real platoons in the simulated platoons (the black curves in Fig. 7(b) and (d)). In the comparison of the last vehicles, Fig. 7(d) shows satisfied accuracy, whilst a relatively large deviation occurs in Fig. 7(b). The deviation is resulted from a vehicle who fails to follow its leader along the wave not as most of the vehicles usually do; see the trajectory pointed by an arrow in Fig. 7(a). This random behaviour diminishes the wave,

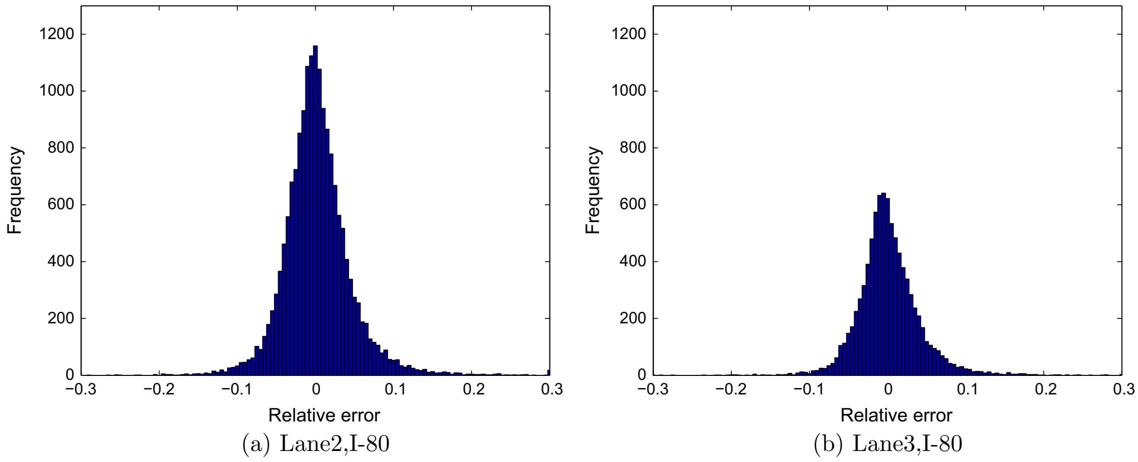


Fig. 6. Histogram plots for the relative errors of the estimated space headways when $\mathcal{D}_k < 0.2$. The estimated vehicles are from the I-80 segment, and the database for *k*NN is from the US-101 segment (Database 3).

whilst it is missed by the nonparametric model that reflects averaging behaviour. Seeing the vehicle in front of the pointed vehicle, e.g., 10th vehicle, the estimation deviation is much smaller and satisfied. It should not expect that the same wave is reproduced, because various random factors and lane changes widely exist in the real world. This can also be seen from the distribution of vehicles in the time–space diagrams. The simulated vehicles distribute more uniformly, whilst the gaps between real vehicles are quite different.

4.2. A platoon with different boundary conditions

In this subsection, the following platoons of the real leaders are estimated given different boundary conditions. A new vehicle is sent on the roadway when the gap between the entrance and the last vehicle on the roadway exceeds a given value. 30 m, 40 m, and 50 m for the gap are selected to test. It is clear that the smaller the gap is, the higher the demand is. The initial moving distance is set to be 15 m (i.e., 54 km h^{-1} for the initial speed).

Fig. 8 presents the time–space diagrams of the trajectories of the simulated following platoon. It can be seen that the waves originated from the real leaders propagate with a speed of -16 km h^{-1}, and the shape of the waves are consistent with the real ones in the US-101 dataset. The impact of sudden deceleration of the leaders is relatively small when the arrival demand is low (see Fig. 8(c) and (f)), and the waves thus fail to stably propagate. It also demonstrates that the model can well reflect the impact of realistic speed drop. Meanwhile, a few monitored distances $\mathcal{D}_k$ are within [0.2,0.3], but it seems that the limited cases with relatively large distance do not make large impact on the simulation results.

4.3. Necessity of each input

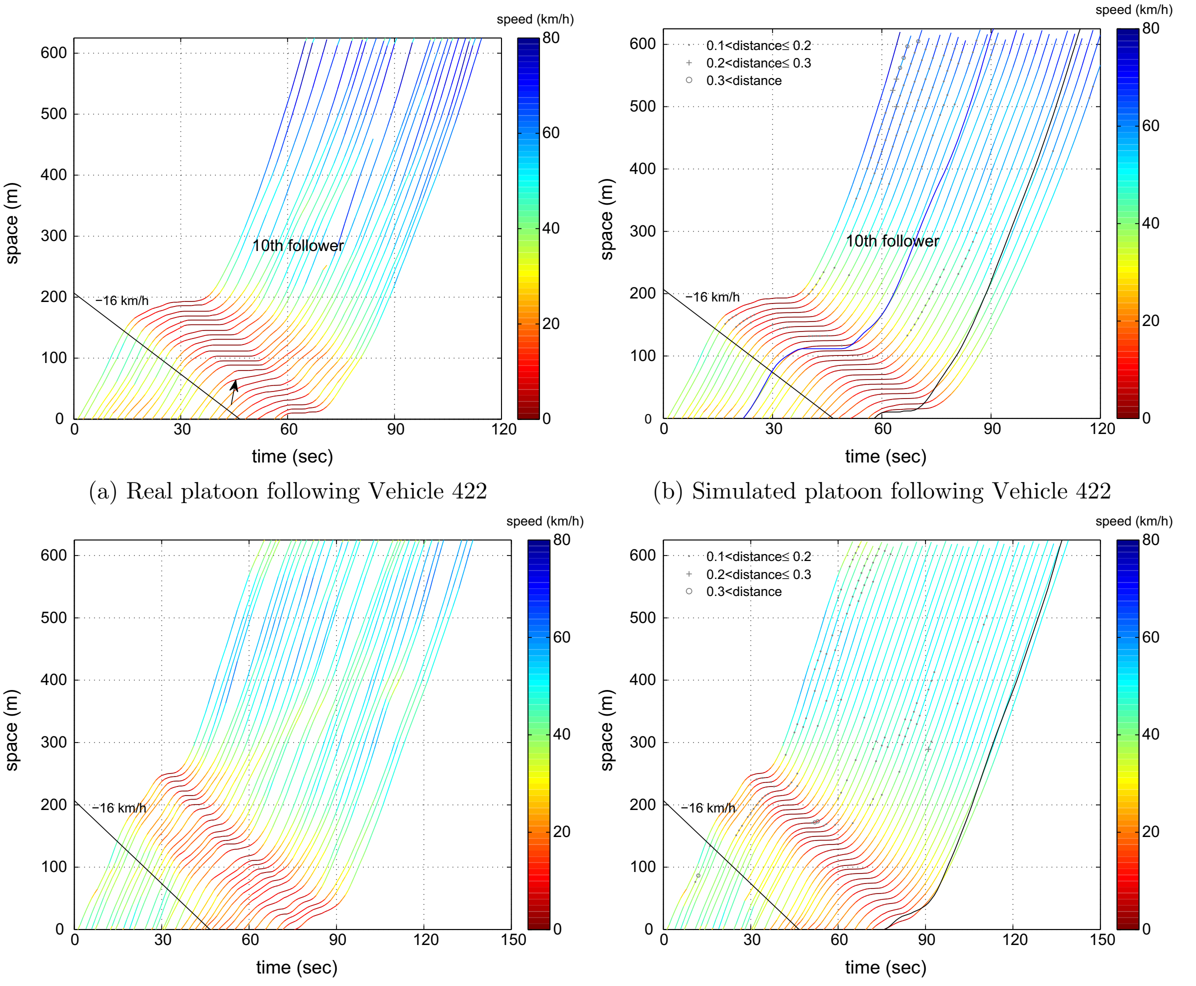
Along with this simulation scenario, we demonstrate the necessity of each input proposed in Eq. (2). To this end, we remove one input from the total four inputs, and conduct the same simulation presented in Fig. 7(d). See Fig. 9 for the results. Comparing with Fig. 7(d) resulted by all four inputs in the same scenario, Fig. 9 well demonstrates the necessity of all four inputs, because the traffic waves become unrealistic or even collisions occur without inputing any one of the four.

5. A rubbernecking simulation scenario

In Chen et al. (2012b, 2014), rubbernecking caused by the clean-up work is reported on Lane 1 between 7:50 a.m. and 8:05 a.m. on the study segment of US-101. In order to test the nonparametric car-following model, this section simulates the rubbernecking scenario by using the nonparametric car-following model.

Before presenting the scenario, we briefly introduce four distinct stages of traffic oscillations addressed in Chen et al. (2014). Our simulation results will be analysed based on the stages. The four stages are (i) precursor: the speeds of the deceleration and acceleration waves are close to zero, indicating localised slow-and-go driving motions; (ii) growth: the waves propagate backward in space at the speed of 16–24 km h^{-1}, and the minimum speed of vehicles decreases significantly as the waves propagate; (iii) stable: the amplitude remains relatively constant as waves propagate backward in space; (iv) decay: the amplitude diminishes as waves propagate.

In the rubbernecking simulation scenario, a 1.25 km one-lane roadway is simulated for 1 h. A new vehicle enters the roadway with an initial speed of 54 km h^{-1}, when its leader has left the entrance 30 m away. The rubbernecking zone is located at section [1,1.05] km. When vehicles enter the zone, they have a probability r to rubberneck and then slow down by a



(a) Real platoon following Vehicle 422

(b) Simulated platoon following Vehicle 422

(c) Real platoon following Vehicle 1989

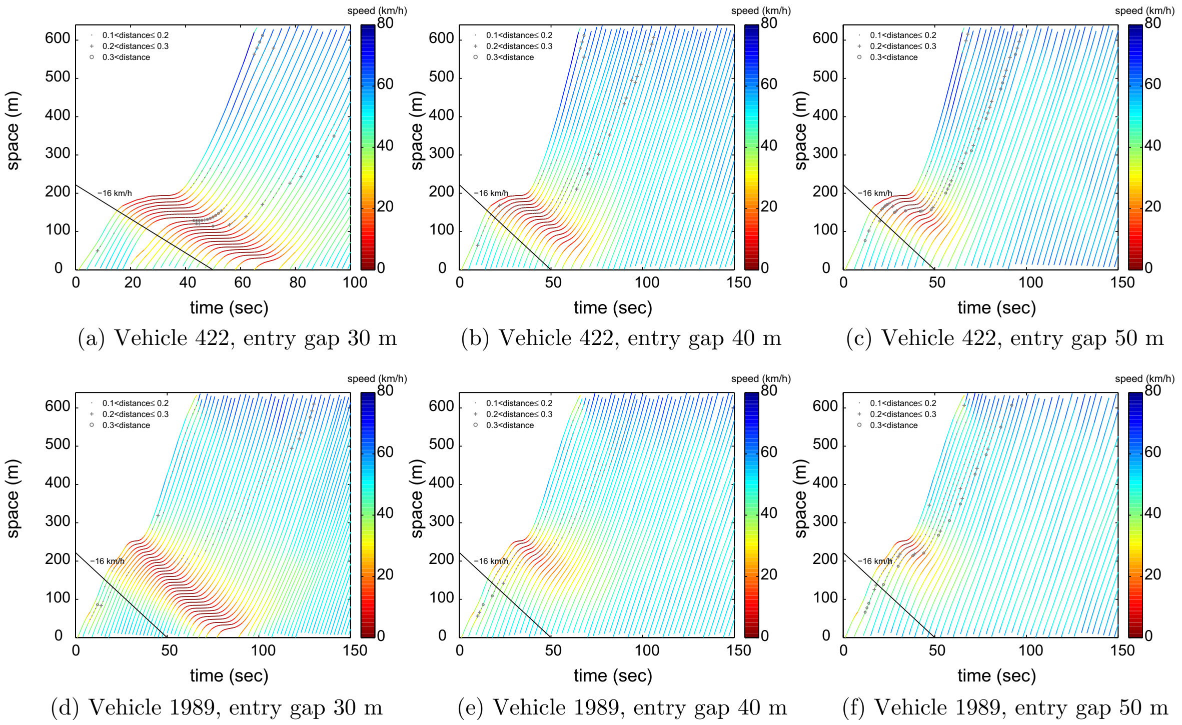
(d) Simulated platoon following Vehicle 1989

Fig. 7. Comparisons of the real and simulated platoons following Vehicle 422 and 1989 on Lane 1 of the US-101 segment. The time and initial speed of the simulated followers entering the road are the same to the real platoons. Database 3 is used. The black curves are the last vehicles in the real platoons, and the blue curve is the 10th follower in the real platoon following Vehicle 422. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

percentage of $(1 - p)$. The output, i.e., follower's moving distance in Eq. (5), is multiplied by p in order to describe the influence of the speed drop in the model. If rubbernecking occurs, it will occur at most once. The influence of rubbernecking lasts for h simulation steps, i.e., h sec in the scenario. The database contains all data collected from Lane 1, 2, and 3 during all 45 min. Neither the fundamental diagrams nor parameters about driver behaviour is needed to be specified or calibrated. Note that only congested traffic could be well simulated here, because the underlying database only contains congested traffic data.

The simulation results are demonstrated in the time-diagrams of trajectories, which are considered as very important tools to validate models (see e.g. Laval and Leclercq (2010) and Chen et al. (2012b), and Chapter 10–13 in Treiber and Kesting (2013b)). See Fig. 10 for the diagrams, in which $r = 0.05,\ p = 0.8$, and $h = 5$. In the figure, Fig. 10(a) is a regular time–space diagram of trajectories, in which different speeds are highlighted. Whilst Fig. 10(b) is a time–space diagram of the distance $\mathcal{D}_k$, and the colours represent different distances of the estimations. We employ this type of diagram to monitor the similarity between the database and the estimation cases. From Fig. 10(b), we can see that almost all distances of the estimations are less than 0.2. It means that all the estimated cases in the simulation scenario could find similar historical cases in the database, and the results highly likely have a high quality as stated in Section 3.5.

We first focus on important and representative traffic characteristics in Fig. 10. It can be seen from the figure that the simulated oscillations transform from localised disturbances to well-developed ones. The wave speed in the stable stage is around -16 km h^{-1}, which is consistent with the wave speed of the US-101 dataset. The period and amplitude of oscillations at the location of 400 m are around 5 min and 58 km h^{-1}. The observations are all consistent with the existing findings



(a) Vehicle 422, entry gap 30 m

(b) Vehicle 422, entry gap 40 m

(c) Vehicle 422, entry gap 50 m

(d) Vehicle 1989, entry gap 30 m

(e) Vehicle 1989, entry gap 40 m

(f) Vehicle 1989, entry gap 50 m

Fig. 8. Simulated platoons following real vehicles on the US-101 segment. The "distance" in the legend is the distance from the *k*th (i.e., 10th) sample to the estimated case, i.e., $\mathcal{D}_k$.

and empirical observations, such as wave speed empirically ranging from -10 to -20 km h^{-1} and the period ranging from 2 to 15 min (Mauch and Cassidy, 2002; Ahn and Cassidy, 2007; Schönhof and Helbing, 2007; Laval and Leclercq, 2010; Zheng et al., 2011a; Zheng et al., 2011b; Chen et al., 2012a,b, 2014; Jiang et al., 2014).

To see more detail, some regions in Fig. 10(a) are zoomed in and analysed. Fig. 10(c)–(e) present three patterns that a stable wave forms, which could be also observed in the US-101 dataset. In Fig. 10(c), several slowdowns of two vehicles make the traffic in the precursor stage grow. In Fig. 10(d), a slowdown primarily triggers slow traffic, whilst it is amplified by a series of slowdowns afterwards at the same location. In Fig. 10(e), several waves join together and finally form a stable wave propagating backward in spaces. Fig. 10(f) presents an oscillation in the decay stage. Few empirical study focuses on the traffic in this stage probably due to lack of empirical data. Considering the good performance of the nonparametric model, it may be able to be studied with the aid of the model.

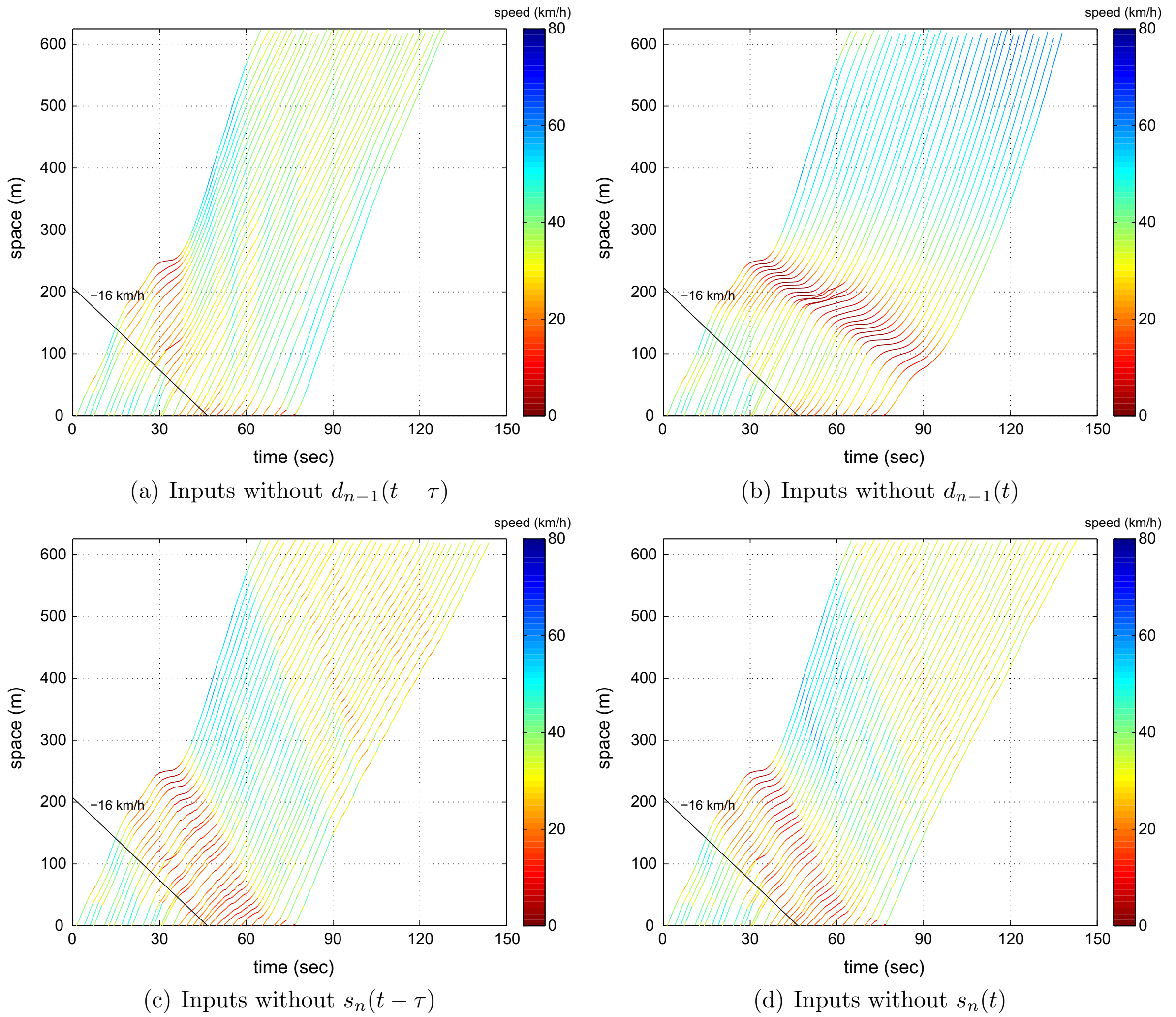
We further present the fundamental diagrams to show the relations amongst the fundamental traffic parameters. Suppose that virtual detectors are installed in the roadside, and the traffic flow q, density ρ, and speed v within a time period T are measured as follows.

$$q = \frac{N}{T}, \ \rho = \frac{\sum_{n=1}^{N} \frac{1}{v_n}}{T}, \text{ and } v = \frac{q}{\rho} = \frac{N}{\sum_{n=1}^{N} \frac{1}{v_n}} \tag{9}$$

where N is the count of the vehicles passing the detection location within the time period T, and v_n is the passing speed of a detected vehicle. Fig. 11 compares the fundamental diagrams for the empirical and simulated traffic, in which the aggregation time period is set to be 60 s. It is observed that the shapes of the simulated fundamental diagrams are quite similar to the empirical ones for the US-101 segment. It means that the simulated traffic has the similar relations amongst flow, density, and speed contained by US-101 dataset. We can thus believe that the simulation scenario is conducted based on the fundamental diagrams similar to the real ones. It demonstrates that the simulated traffic evolves similarly to the real traffic, and the nonparametric car-following model is able to well replicate the fundamental diagrams.

Meanwhile, hysteresis is observed in the fundamental diagrams for the simulation scenario. Particularly see the flow-density relations in Fig. 11(a). The lower branch is produced by the upstream traffic far away from the rubbernecking location (i.e., 200 m, 400 m, and 600 m). At these locations, the oscillations have been in the stable stage, during which lower flow-density relations are generated. This is consistent with the empirical findings given in Chen et al. (2014), in which the hysteresis during the stable stage of an oscillation is demonstrated in a perspective of the trajectories of vehicle platoons.

Moreover, we conduct the experiment with various settings and present the periods and amplitudes in Fig. 12. Notice that the periods stay around 2–4 min when $r > 0.08$, r is negatively correlated with the period, and p is positively correlated with



(a) Inputs without $d_{n-1}(t-\tau)$

(b) Inputs without $d_{n-1}(t)$

(c) Inputs without $s_n(t-\tau)$

(d) Inputs without $s_n(t)$

Fig. 9. Simulated platoons taking three inputs amongst the total four inputs. The leader is Vehicle 1989 on Lane 1 of the segment of US-101.

the period. The relationship between a period and r appear to be convex and rapidly approaching a stable value. Although the period become stable, the amplitudes continue to drop. These observations are quite consistent with the outcome of the variants of Newell's model (Newell, 2002) proposed in Laval and Leclercq (2010) and Chen et al. (2012b).

6. A simulation scenario with driver errors

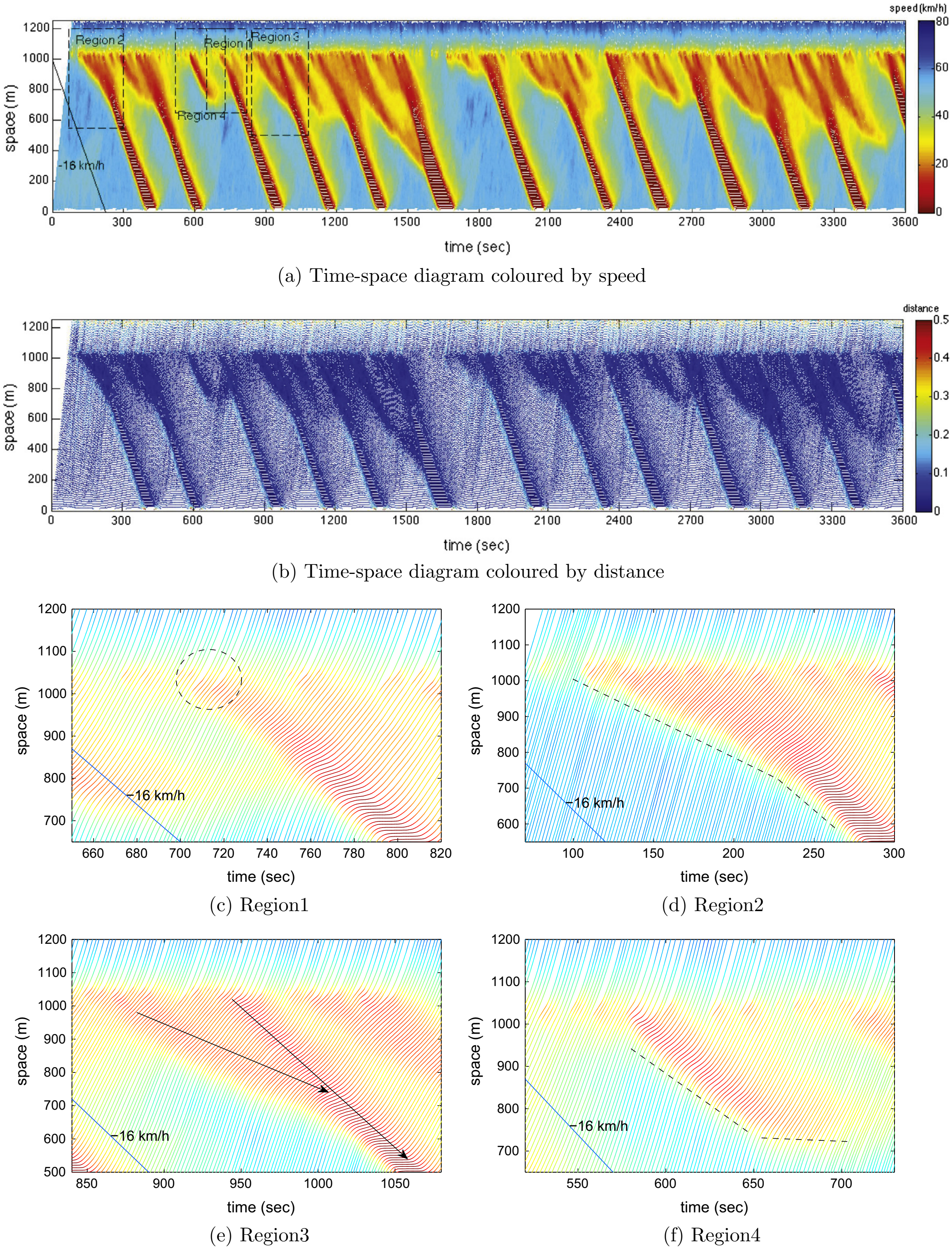
In the study of the trigger of traffic oscillations in the absence of lane changes, it has been shown that even a small perturbation, e.g., driver errors, may result in traffic instability, i.e., the oscillation (see e.g. Treiber et al., 1999; Igarashi et al., 2001; Wilson, 2008; Orosz et al., 2009; Laval et al., 2014). To test the response of the nonparametric model to perturbations, this section introduces driver errors to the model and simulate in a condition without external disturbance. By conducting the experiments with various driver errors, we observe a critical error intensity to trigger traffic instability.

We model the driver errors in a form of a white Gaussian noise with diffusion coefficient σ^2 (Laval et al., 2014). Thus, the moving distance of a follower with driver errors is written as follows:

$$\tilde{d}_n(t+\tau) = d_n(t+\tau) + W(\varepsilon) \tag{10}$$

where

$$W(\varepsilon) = \begin{cases} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{\varepsilon^2}{2\sigma^2}} & , \ d_{\text{jam}} < d_n(t+\tau) < d_{\text{free}} \\ 0 & , \ \text{otherwise} \end{cases} \tag{11}$$



(a) Time-space diagram coloured by speed

(b) Time-space diagram coloured by distance

(c) Region1

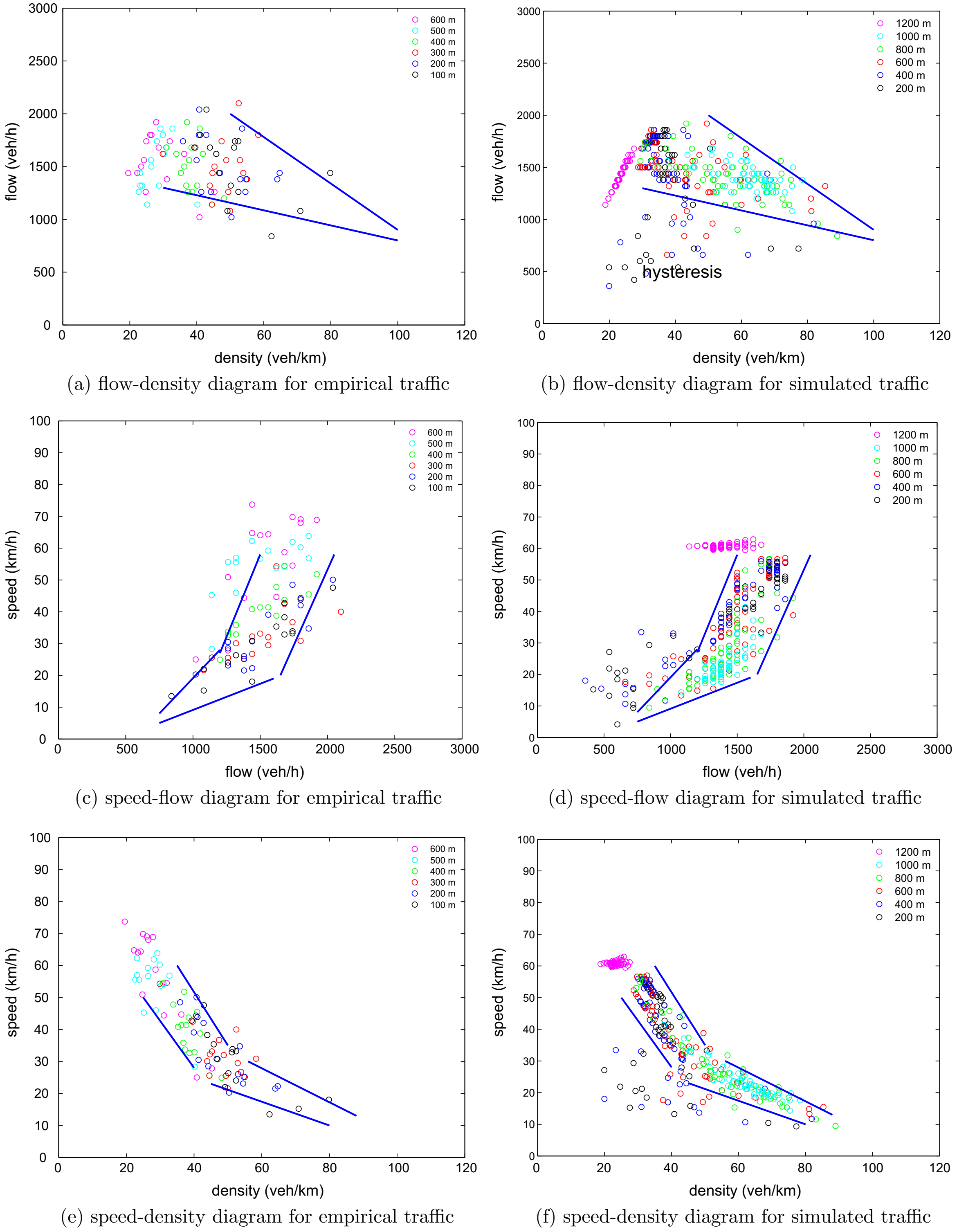
(d) Region2

(e) Region3

(f) Region4

Fig. 10. Time–space diagrams of trajectories in the rubbernecking scenario ($r = 0.05,\ p = 0.8$, and $h = 5$).

where d_{jam} and d_{free} are the moving distance/speed around jam density and free-flow conditions, respectively. Limiting the perturbation only occurring when $d_n(t+\tau) \in (d_{\text{jam}}, d_{\text{free}})$ roughly reflects the fact that human errors (i.e., speed variability) is maximal at the critical density, and is minimal at the free-flow condition and jam density (Jia et al., 2011; Laval et al., 2014).



(a) flow-density diagram for empirical traffic

(b) flow-density diagram for simulated traffic

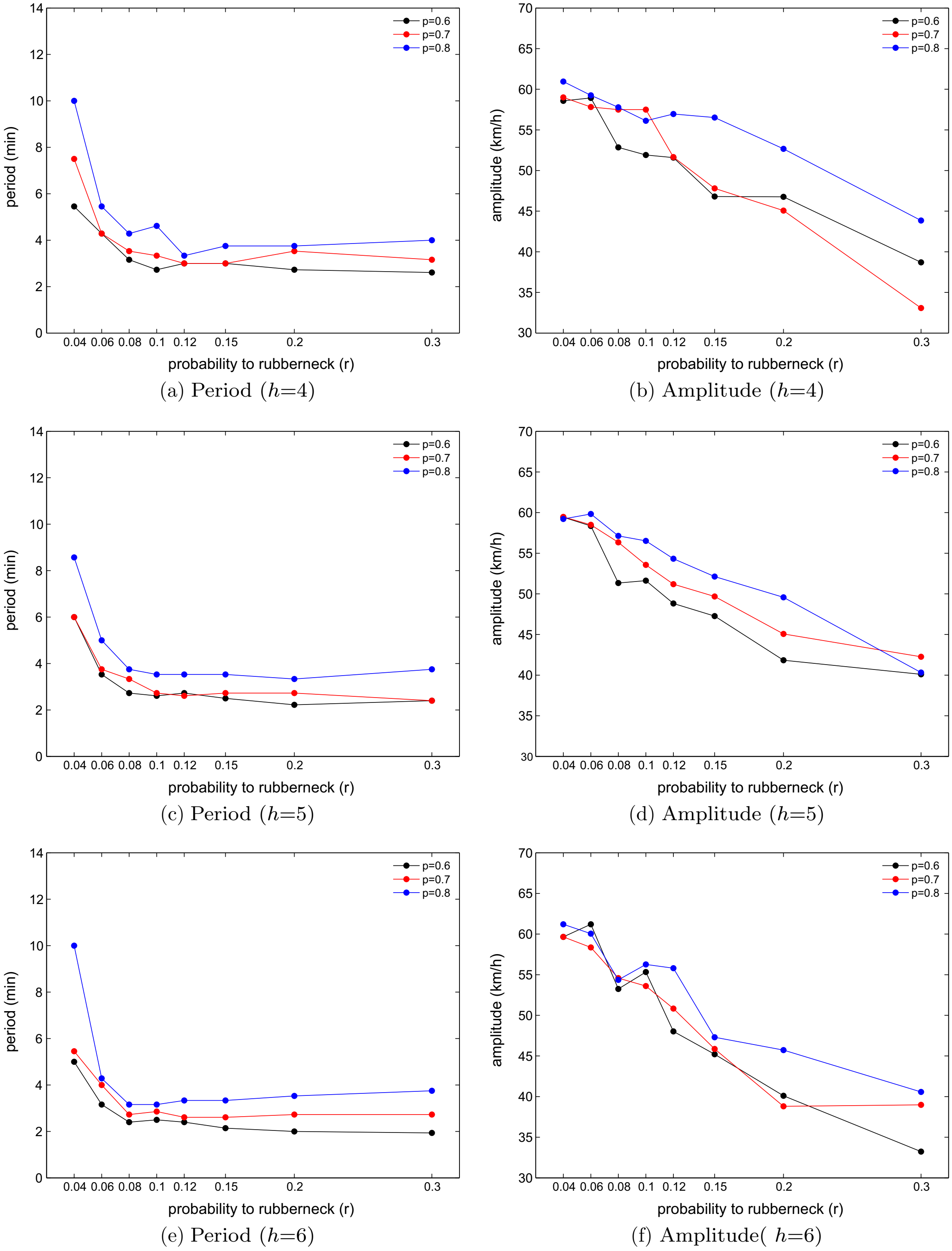
(c) speed-flow diagram for empirical traffic

(d) speed-flow diagram for simulated traffic

(e) speed-density diagram for empirical traffic

(f) speed-density diagram for simulated traffic

Fig. 11. The fundamental diagrams for the rubbernecking scenario and for the empirical US-101 segment. The empirical fundamental diagrams are drawn using the NGSIM dataset between 7:50 a.m. and 8:35 a.m., and the data are aggregated every 60 s.



(a) Period (h=4)

(b) Amplitude (h=4)

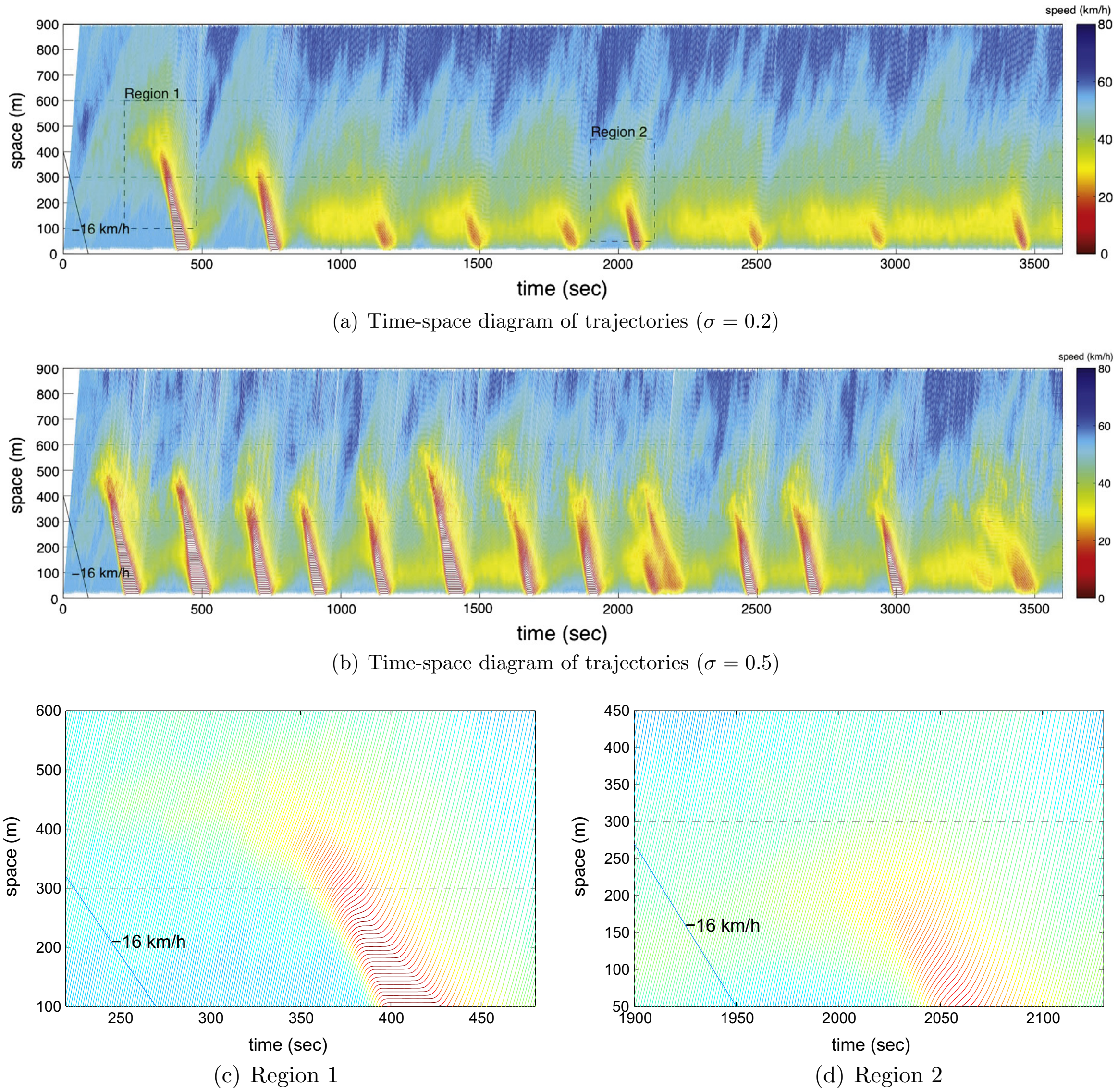
(c) Period (h=5)

(d) Amplitude (h=5)

(e) Period (h=6)

(f) Amplitude(h=6)

Fig. 12. The periods and amplitudes of traffic oscillations at the location of 400 m.



(a) Time-space diagram of trajectories ($\sigma = 0.2$)

(b) Time-space diagram of trajectories ($\sigma = 0.5$)

(c) Region 1

(d) Region 2

Fig. 13. Time–space diagrams of trajectories in the simulation scenario with driver errors.

Although driver errors are directly added to the outputted moving distance, collisions will not be introduced if the diffusion coefficient is small.

In the simulation scenario, a 900 m one-lane roadway is simulated for 3600 s. The white Gaussian noise is only added when a vehicle is moving in the section between the locations of 300 m and 600 m. This is analogous to an uphill section. It is set that d_{free}=54 km h^{-1}, d_{jam} = 15 km h^{-1}, and the entry speed and gap are 54 km h^{-1} and 20 m, respectively.

We first test the model with different values of the diffusion coefficient. The results[3] show that the stop-and-go oscillation is only triggered by the noise greater than $\sigma = 0.2$. No oscillation occurs in the experiment when $\sigma = 0$ and 0.1. Thus, there is a threshold of the error leading to traffic instability. It also shows that even without external disturbance, driver errors can also result in traffic oscillations, and the required error may be quite small. It is consistent with the existing findings in physics and transportation studies.

Fig. 13 presents a part of the simulation results. In Fig. 13(a) and (b) the periodic waves and their speeds are all consistent with the real oscillation. The precursor stages of oscillations in Fig. 13(c) and (d) have realistic shapes. It is also noticed in Fig. 13(a) and (d) that stable oscillations form in a section without driver errors, i.e., in the upstream of the section with

[3] All $\mathcal{D}_k$ are within satisfied ranges, i.e., smaller than about 0.2. It is not shown to save space.

driver errors. It implies a situation that the traffic slowly moves in the uphill section whilst stop-and-go waves occur in its upstream without a clear trigger.

7. Conclusions

Based on the truth that drivers repeat their behaviour in similar circumstances, this paper proposes a data-driven car-following model by using the simple *k*NN. The approach takes the average of the most similar cases as its output, which implies the most likely driving behaviour under the current circumstance. Four inputs are selected, i.e., leader's moving distances and follower's space headways in the latest two time steps, and the output is follower's moving distance. The only parameter *k* is determined through analysing estimation errors with different *k*-values. A time–space diagram of the distance of *k*th sample is designed to monitor the simulation scope. Three simulation scenarios are conducted, i.e., following real leaders, rubbernecking, and driver errors. Time–space diagrams containing thousands of vehicle trajectories, fundamental diagrams, and periods and amplitudes of oscillations are presented. Explicit analysis demonstrates that the model is able to reproduce realistic stop-and-go oscillations. We conclude the advantages of the proposed nonparametric car-following model as follows:

(i) Neither mathematical equation nor calibration is needed to be concerned in the model.
(ii) Neither the fundamental diagrams nor driver's behaviour parameters is assumed.
(iii) The model is simple and parsimonious particularly in the conceptual point of view, and the only parameter is *k*.
(iv) All inputs and outputs are based on vehicle positions, which are straightforward to reproduce traffic dynamics in computer simulations.
(v) The model is able to well reproduce traffic characteristics contained by the underlying database, such as all stages of stop-and-go oscillations, fundamental diagrams, periods and amplitudes of oscillations.

Based on field data, the nonparametric car-following model can be rapidly deployed and replicate reliable traffic characteristics without model calibration. However, the simplicity, in return, may also limit the freedom to set simulation scenarios due to lack of mechanism description.

Due to no mathematical formula, it may be easy to incorporate various factors for specific purposes or extend to multi-lane situations. One could just concentrate on the inputs and outputs, and does not need to worry about the mathematical expression. These are future work directions.

Acknowledgements

The authors are grateful to three anonymous referees whose comments greatly improved the quality of this paper, and to Prof. Rui Jiang and Prof. Ning Jia for their helpful suggestions. This research has been funded by NFSC (71471014) and 973 Program of China (2012CB725403).

References

Ahn, S., Cassidy, M., 2007. Freeway traffic oscillations and vehicle lane-change maneuvers. In: Allsop, R., Bell, M., Heydecker, B. (Eds.), 17th International Symposium of Transportation and Traffic Theory. Elsevier, Amsterdam, pp. 691–710.

Bando, M., Hasebe, K., Nakayama, A., 1995. Dynamical model of traffic congestion and numerical simulation. Physical Review E 51, 1035–1042.

Brackstone, M., Mcdonald, M., 1999. Car-following: a historical review. Transportation Research Part F: Traffic Psychology and Behaviour 2, 181–196.

Brockfeld, E., Kühne, R.D., Wagner, P., 2004. Calibration and validation of microscopic traffic flow models. Transportation Research Record, 62–70.

Chandler, R., Herman, R., Montroll, E., 1958. Traffic dynamics: studies in car following. Operations research, 1–17.

Chen, D., Ahn, S., Laval, J., Zheng, Z., 2014. On the periodicity of traffic oscillations and capacity drop: the role of driver characteristics. Transportation Research Part B 59, 117–136.

Chen, D., Laval, J., Ahn, S., Zheng, Z., 2012a. Microscopic traffic hysteresis in traffic oscillations: a behavioral perspective. Transportation Research Part B 46, 1440–1453.

Chen, D., Laval, J., Zheng, Z., Ahn, S., 2012b. A behavioral car-following model that captures traffic oscillations. Transportation Research Part B 46, 744–761.

Cleveland, W., 1979. Robust locally weighted regression and smoothing scatterplots. Journal of the American Statistical Association 74, 829–836.

Colombaroni, C., Fusco, G., 2014. Artificial neural network models for car following: experimental analysis and calibration issues. Journal of Intelligent Transportation Systems 18, 5–16.

Gipps, P., 1981. A behavioural car-following model for computer simulation. Transportation Research Part B 15, 105–111.

Helbing, D., 2001. Traffic and related self-driven many-particle systems. Reviews of Modern Physics 73, 1067–1141.

Hoogendoorn, S., 2001. State-of-the-art of vehicular traffic flow modelling. Proceedings of the Institution of Mechanical Engineers, Part I: Journal of Systems and Control Engineering 215, 283–303.

Igarashi, Y., Itoh, K., Nakanishi, K., Ogura, K., Yokokawa, K., 2001. Bifurcation phenomena in the optimal velocity model for traffic flow. Physical Review E 64, 047102.

Jia, B., Li, X., Chen, T., Jiang, R., Gao, Z., 2011. Cellular automaton model with time gap dependent randomisation under Kerner's three-phase traffic theory. Transportmetrica 7, 127–140.

Jiang, R., Wu, Q., Zhu, Z., 2001. Full velocity difference model for a car-following theory. Physical Review E 64, 017101.

Jiang, R., Hu, M., Zhang, H.M., Gao, Z., Jia, B., Wu, Q., Wang, B., Yang, M., 2014. Traffic experiment reveals the nature of car-following. PloS ONE 9, e94351.

Kerner, B., 1998. Experimental features of self-organization in traffic flow. Physical Review Letters 81, 3797–3800.

Kesting, A., Treiber, M., 2008. Calibrating car-following models by using trajectory data. Transportation Research Record, 148–156.

Khodayari, A., Ghaffari, A., Kazemi, R., Braunstingl, R., 2012. A modified car-following model based on a neural network model of the human driver effects. IEEE Transactions on Systems, Man, and Cybernetics 42, 1440–1449.

Laval, J., Leclercq, L., 2010. A mechanism to describe the formation and propagation of stop-and-go waves in congested freeway traffic. Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences 368, 4519–4541.

Laval, J., Toth, C.S., Zhou, Y., 2014. A parsimonious model for the formation of oscillations in car-following models. Transportation Research Part B 70, 228–238.

Mauch, M., Cassidy, M., 2002. Freeway traffic oscillations: observations and predictions. In: 15th International Symposium of Transportation and Traffic Theory, pp. 653–673.

Montanino, M., Punzo, V., 2013. Making NGSIM data usable for studies on traffic flow theory. Transportation Research Record 2390, 99–111.

Newell, G., 2002. A simplified car-following theory: a lower order model. Transportation Research Part B 36, 195–205.

NGSIM, 2006. The Next Generation Simulation Program. <http://ops.fhwa.dot.gov/trafficanalysistools/ngsim.htm>.

Orosz, G., Wilson, R., Szalai, R., 2009. Exciting traffic jams: nonlinear phenomena behind traffic jam formation on highways. Physical Review E 80.

Panwai, S., Dia, H., 2007. Neural agent car-following models. IEEE Transactions on Intelligent Transportation Systems 8, 60–70.

Pipes, L., 1953. An operational analysis of traffic dynamics. Journal of Applied Physics 24, 274–281.

Punzo, V., Borzacchiello, M.T., Ciuffo, B., 2011. On the assessment of vehicle trajectory data accuracy and application to the Next Generation SIMulation (NGSIM) program data. Transportation Research Part C: Emerging Technologies 19, 1243–1262.

Punzo, V., Ciuffo, B., Montanino, M., 2012. Can results of car-following model calibration based on trajectory data be trusted? Transportation Research Record 2315, 99–111.

Saifuzzaman, M., Zheng, Z., 2014. Incorporating human-factors in car-following models: a review of recent developments and research needs. Transportation Research Part C: Emerging Technologies 48, 379–403.

Schönhof, M., Helbing, D., 2007. Empirical features of congested traffic states and their implications for traffic modeling. Transportation Science 41, 135–166.

Schönhof, M., Helbing, D., 2009. Criticism of three-phase traffic theory. Transportation Research Part B 43, 784–797.

Thiemann, C., Treiber, M., Kesting, A., 2008. Estimating acceleration and lane-changing dynamics from next generation simulation trajectory data. Transportation Research Record, 90–101.

Treiber, M., Hennecke, A., Helbing, D., 1999. Derivation, properties, and simulation of a gas-kinetic-based, non-local traffic model. Physical Review E 59, 239–253.

Treiber, M., Hennecke, A., Helbing, D., 2000. Congested traffic states in empirical observations and microscopic simulations. Physical Review E 62, 1805–1824.

Treiber, M., Kesting, A., 2013a. Microscopic calibration and validation of car-following models – a systematic approach. In: 20th International Symposium on Transportation and Traffic Theory. Elsevier B.V., pp. 922–939.

Treiber, M., Kesting, A., 2013b. Traffic Flow Dynamics: Data, Models and Simulation. Springer.

Treiber, M., Kesting, A., Helbing, D., 2010. Three-phase traffic theory and two-phase models with a fundamental diagram in the light of empirical stylized facts. Transportation Research Part B 44, 983–1000.

Vlahogianni, E.I., Karlaftis, M.G., Golias, J.C., 2014. Short-term traffic forecasting: where we are and where were going. Transportation Research Part C: Emerging Technologies 43, 3–19. http://dx.doi.org/10.1016/j.trc.2014.01.005.

Wei, D., Liu, H., 2013. Analysis of asymmetric driving behavior using a self-learning approach. Transportation Research Part B 47, 1–14.

Wilson, R.E., 2008. Mechanisms for spatio-temporal pattern formation in highway traffic models. Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences 366, 2017–2032.

Wu, C.H., Ho, J.M., Lee, D., 2004. Travel-time prediction with support vector regression. IEEE Transactions on Intelligent Transportation Systems 5, 276–281.

Zhang, Y., Ge, H., 2013. Freeway travel time prediction using Takagi–Sugeno–Kang fuzzy neural network. Computer-Aided Civil and Infrastructure Engineering 28, 594–603.

Zheng, J., Suzuki, K., Fujita, M., 2013. Car-following behavior with instantaneous drivervehicle reaction delay: a neural-network-based methodology. Transportation Research Part C: Emerging Technologies 36, 339–351.

Zheng, Z., Ahn, S., Chen, D., Laval, J., 2011a. Applications of wavelet transform for analysis of freeway traffic: bottlenecks, transient traffic, and traffic oscillations. Transportation Research Part B 45, 372–384.

Zheng, Z., Ahn, S., Chen, D., Laval, J., 2011b. Freeway traffic oscillations: microscopic analysis of formations and propagations using Wavelet Transform. Transportation Research Part B 45, 1378–1388.

Zheng, Z., Su, D., 2014. Short-term traffic volume forecasting: a *k*-nearest neighbor approach enhanced by constrained linearly sewing principle component algorithm. Transportation Research Part C: Emerging Technologies 43, 143–157.