Examining vehicular speed characteristics through divergences from prior distributions

Caglar Kosun¹

Received: 28 July 2020 / Accepted: 8 January 2021 / Published online: 25 January 2021
© The Author(s) 2021

Abstract

A variety of approaches, within literature, has been conducted to interpret vehicular speed characteristics. This study turns the attention to the entropy-based approaches, and thus focuses on the maximum entropy method of statistical mechanics and the Kullback–Leibler (KL) divergence approach to examining the vehicular speeds. The vehicle speeds at the selected highway are analyzed in order to find out the disparities among them. However, it is turned out that the speed dynamics could not be distinguished over the speed distributions; hence the maximization of Shannon entropy seems insufficient to compare the speed distributions of each data set. For this reason, the KL divergence approach was performed. This approach displays the comparison, among the speed distributions, based on two prior distribution models, i.e., uniform and Gauss. The examination of the trends of KL divergences obtained from both distributions was made. It was concluded that the KL divergence values for the highway speed data sets ranged between about 0.53 and 0.70 for the uniform case, while for the Gaussian case the obtained values are between 0.16 and 0.33. The KL divergence trends for the real speeds were obtained analogous for both cases, but they differed significantly when the synthetic data sets were employed. As a result, the KL divergence approach proves suitable as an appropriate indicator to compare the speed distributions.

Keywords Traffic flow · Maximum entropy principle · Kullback–Leibler divergence · Speed distributions · Prior distributions

1 Introduction

Over recent decades, traffic flow has become far more complicated and complex, especially as a result of the increasing number of vehicles in cities. Moreover, there could be many other possible factors for rising traffic complexity, e.g., location of signalization, land-use characteristics, interaction among vehicles. In such a complex traffic system, all factors are in fact directly or indirectly related with traffic flow parameters. Hence, in real traffic flow, the memory is non-Markovian, time is discrete, and time-dependence is present [1, 2]. The long-range interactions also emerge, and the impact of those properties

would make the speed distributions non-Gaussian, and these distributions could be crucial guide to interpret traffic flow dynamics. It can be expressed that the distribution functions of the real traffic flow parameters, e.g., vehicular speeds, might not be well described within the classic extensive entropy formalism, i.e., the celebrated Boltzmann–Gibbs (BG) thermostatics approach [2]. In this approach, maximum entropy principle could not handle the speed dynamics, and thus not compare the speed distributions of traffic flow since it may not describe the real traffic flow sufficiently.

In literature, there are numerous approaches examining the complexity and dynamics of traffic flow and one

✉ Caglar Kosun, cglrksn@gmail.com; caglar.kosun@omu.edu.tr | ¹Department of City and Regional Planning, Ondokuz Mayıs University, İlkadım, Samsun 55100, Turkey.

of them is the use of entropy-based methods. In terms of this, recent various studies, e.g., [1–6] would be typical in the vehicular traffic arena. Of these, for example, Kosun and Ozdemir [1] focus on the platoon formation of vehicles and propose an upper and a lower limit of Tsallis q entropic index. In the study [2], vehicular speeds are considered within the framework of nonextensive thermostatics. Superstatistical analyses are performed, and the pdfs of the vehicular speeds at highway are associated with q -Gaussian distributions. Two distinct Tsallis q values are specified, and the time-dependency and time-independency of the traffic flow are discussed. Zhang and Shang [3] propose multivariate multiscale distribution entropy, and the traffic congestion at different time scales is examined. Depending on the method, the complexity of the traffic system is assessed with the complexity-entropy causality plane. Yan et al. [5] construct complex networks of multivariate traffic flow time series, and examine the dynamics of the hourly data. The authors also use network structure entropy to analyze the network structure and the calculated values are interpreted. In addition, the study [7] is related to the multivariate multiscale sample entropy in traffic. Xu et al. [8] focus on the travel time predictability based on the historical travel time series. The refined composite multiscale entropy algorithm is performed, and the connection between the upper bound of travel time predictability and entropy is given. Another study [9] considers stimulus–response equation, and the flow–concentration relationship in the traffic is given. The entropy is maximized by the Lagrange multiplier technique, and then the equation involving flow, concentration and speed parameters is defined. Silva et al. [10] examine the global, daily and hourly vehicle velocity dynamics using information theory concepts considering the causality Complexity-Entropy plane which is also utilized in [11] to characterize the global velocity behaviors where the compatibility of the velocities with correlated noise f^{-k} power spectrum with $k \geq 0$ is stated. Both in [10] and [11], the Bandt-Pompe methodology is applied considering the time series of vehicle velocities. The studies investigate the behavior of the velocities in terms of noises in detail and present the correlation degrees in the light of different events.

Even though such methods concerning traffic complexity and dynamics are present in the literature, KL divergence approach is not usual in this context. In terms of utilizing relative entropy (KL divergence) in vehicular traffic, the studies concerning, e.g., driving conditions [12], travel time [13–15], vehicle following [16] could be given as examples. For instance, in the study [12], transition probabilities for the vehicle speed and acceleration with regard to highway and city driving are considered. Then, KL divergence analyses are performed, driving style, i.e., smooth and aggressive, and locations are the main concerns in

the KL divergence analyses which distinguish similar and dissimilar driving conditions. Ernst et al. [14] focus on the travel time estimation, and their main aim is to obtain the divergence between the true and estimated distributions. For this purpose, KL divergence is utilized and two types of matching technologies, i.e., id matching and signature matching are also compared. In the study, the authors set forth the signature matching as the best choice since it minimizes the dissimilarity between the true and estimated travel time distributions. Furthermore, in order to quantitatively measure the similarity between the different driving styles, KL divergence approach is utilized in [17]. The study [18] examines the amount of naturalistic driving data which is required for understanding driver behaviors. The KL divergence method is used to test the similarity between density functions considering different amounts of the data. Another paper [19] focuses on understanding the characteristics of bus speed distributions. In the paper, two problems are emphasized in the investigation of speed distribution, which are briefly related with appropriate modeling for the speed distributions, and explaining the distribution characteristics, respectively. According to the authors, regarding the latter problem, there is relatively less research which focuses on performing analysis to explain the speed distribution characteristics. The KL divergence approach as a clustering measure is proposed in terms of ranking the road sections on a bus route. The accelerating behavior of driver is studied in [20] where KL divergence approach is used with Kernel Density Estimation in order to process the convergence examination algorithm. To find the required number of data, so as to obtain a convergent distribution of velocity is at issue in the algorithm. In [21], a technique identifying anomalies in flow distribution probabilities based on k -nearest neighbors and KL divergence approach is adapted to obtain the distances between distributions. The spatial and temporal traffic flow information are considered and the proposed outlier detection framework identifies the outliers for flow distributions. Although such examples could be expanded according to the traffic literature, the real speed distribution characteristics have not been examined through uniform and Gaussian prior distributions within KL divergence approach, revealing the similarities between the divergence trends based on them, hence this study aims to fill this gap. For this aim, this study examines the speed distribution characteristics of the highway in the city of Istanbul, Turkey, and points to the divergence from the randomness in the traffic flow.

To begin with, this study is posited on the analogy between molecules moving in a closed system and vehicles at the highway traffic flow. Here, it is considered that the traffic flow system is dynamic and the steady-state flow is present. It is also assumed that the kinetic energies of

the vehicles are conserved. In a few words, initially, the maximum entropy method of statistical mechanics has been tested in the study in order to reflect the speed distribution characteristics. The speed distributions are analytically obtained, grounding on the maximum entropy principle, and the uncertainty parameter (beta) is identified with respect to Gauss distribution. However, it is inferred that this seems insufficient to compare the vehicular speeds since their characteristics could not be distinguished through the pdfs. Hence, this requires a different approach. In order to find out the characteristics of the speed distributions and compare them, the KL divergence approach is utilized. To perform the KL divergence approach, two reference (prior) distributions are considered, and thus for the lucidity of comparisons, one is determined as uniform distribution which represents the steady distribution of vehicle speeds, and the other is selected as Gauss distribution. KL divergence approach computes the deviations of speed distributions from both uniform and Gaussian distributions, and then compares them. A set of distinctive inferences is drawn as a result of employing the reference distributions in KL divergence analyses.

2 Method

Entropy is basically a measure of uncertainty and disorder in the system which is under scrutiny. Also, entropy exhibits an ignorance about the subject that one considers. In this paper, to examine the vehicular speed characteristics, two methods are utilized. These are the maximum entropy method of statistical mechanics and KL divergence approach, which are linked up with the speed distributions and the divergences of the distributions, respectively. These are briefly outlined as follows.

2.1 Maximum entropy method

Shannon [22] has made a connection between information/uncertainty about an event and probability distribution, and elaborated the concept of the entropy in the context of the theory of communication. Once the complete information is provided, then the entropy would be zero [23]. In this study, the maximum entropy principle is utilized in the framework of vehicular speeds. Accordingly, in order to understand the vehicular speed characteristics, the author examines the properties of the speed distributions of the molecules (vehicles) instead of dealing with each individual speed.

The basic idea behind the maximum entropy method is the maximization of Shannon entropy Eq. (1) under a set of constraints. Conventionally, the constraints employed

in the maximization of the entropy could be given as in Eqs. (2) and (3).

$$S = -k \sum_i p_i \ln p_i \quad (1)$$

$$\sum_i p_i = 1 \quad (2)$$

$$\sum_i p_i \epsilon_i = \bar{E} \quad (3)$$

where S is the entropy, k is a positive constant, p_i is the probability of occurrence of the i th state, ϵ_i is the state energy, $\bar{E}$ is the mean energy.

Equations (2) and (3) express the normalization of the probabilities and conservation of the mean energy, respectively.

For equal probabilities, the entropy is defined as in Boltzmann's celebrated formula $S = k \ln W$, where W designates a set of discrete states, and $p_i = 1/W$. There is a connection between the entropy of a system and probability of its state [24]. One of the properties of this entropy is additivity, i.e., the total entropy is the sum of the entropies of the systems. For instance, the entropies of the two systems could be denoted $S_{BG}(A)$ and $S_{BG}(B)$, respectively. Then, the total entropy is defined as $S_{BG}(A + B) = S_{BG}(A) + S_{BG}(B)$, note that this is applicable for Boltzmann–Gibbs thermostatics domain [25].

In the thermodynamic equilibrium state, the Shannon entropy equation is maximum under the given constraints [26]. The method of Lagrange multiplier is utilized in the maximization of the entropy function. The maximization of the entropy for the vehicular speed problem is given in detail under Sect. 3.

2.2 Kullback–Leibler (KL) divergence approach

Let $p(x)$ and $q(x)$ are two probability distributions [23] over a discrete random variable and the KL divergence between $p(x)$ and $q(x)$ is defined as in Eq. (4).

$$D_{KL}(p||q) = \sum_{x \in X} p(x) \ln \frac{p(x)}{q(x)} \quad (4)$$

where $p(x)$ is the posterior distribution, $q(x)$ is the prior distribution.

KL divergence [27] is an information-led measure of distance between the given probability distributions $p(x)$ and $q(x)$, and it indicates the divergence between the two distributions. However, this is not related with the metric measure.

The fundamental properties of KL divergence can be listed as below [23].

- KL divergence is nonnegative, i.e., $D_{KL}(p||q) \geq 0$.
- KL divergence is not symmetric, i.e., $D_{KL}(p||q) \neq D_{KL}(q||p)$, and thus KL divergence is not a true distance.
- If and only if $p(x) = q(x)$ for all x , the KL divergence $D_{KL}(p||q) = 0$.

In this paper, KL divergence approach is utilized for calculating the divergences between the real vehicular speed probability distributions and the selected reference distributions as displayed in Sect. 4.

3 Revisiting the maximum entropy method for the vehicular speed problem

Vehicular speed data are in fact discrete. The author intends to adapt this discrete data to the continuous case. For this purpose, distinct from the conventional expression of mean energy constraint, the new constraint is defined over mean of speed squares. This assumption also leads to the identification of beta Eq. (12). Accordingly, the entropy $S = -k \sum_i p_i \ln p_i$ is maximized subject to the norm and the mean of speed squares constraints, as, respectively, denoted in Eq. (5), and Eq. (6).

$$\sum_i p_i = 1 \quad (5)$$

$$\sum_i p_i v_i^2 = \overline{v^2} \quad (6)$$

where v_i is the speed of the state i , and p_i is the probability of the state i , $\overline{v^2}$ is the mean of speed squares.

Let us define a $F(p_i, \lambda, \beta)$ function covering the entropy and the aforementioned constraints for the vehicular speed problem, by Eq. (7).

$$F(p_i, \lambda, \beta) = - \sum_i p_i \ln p_i - \lambda \left(\sum_i p_i - 1 \right) - \beta \left(\sum_i p_i v_i^2 - \overline{v^2} \right) \quad (7)$$

Let us derive $F(p_i, \lambda, \beta)$ in terms of variable p_i , as indicated in Eq. (8).

$$\frac{\partial}{\partial p_i} \left(- \sum_i p_i \ln p_i - \lambda \left(\sum_i p_i - 1 \right) - \beta \left(\sum_i p_i v_i^2 - \overline{v^2} \right) \right) = 0 \quad (8)$$

where λ and β are the Lagrange undetermined multipliers.

When p_i is solved, the formulation is given below

$$p_i = \frac{1}{Z} e^{-\beta v_i^2} \quad (9)$$

where $\frac{1}{Z}$ is the normalization constant which is defined in terms of Z in Eq. (10).

$$\sum_i e^{-\beta v_i^2} = Z \quad (10)$$

where β is the inverse temperature in the thermodynamics sense.

The Gaussian probability distribution is given as below in regard to speed variable v

$$p(v) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2} \left(\frac{v-\bar{v}}{\sigma} \right)^2} \approx C e^{-\frac{1}{2} \left(\frac{v}{\sigma} \right)^2} \quad (11)$$

where $\bar{v}$ is the mean speed and $\bar{v} = 0$, C is the constant and $\frac{1}{\sigma \sqrt{2\pi}} = C$.

In order to identify β , the Eqs. (9) and (11) are considered. As a result, β is determined as

$$\beta = \frac{1}{2\sigma^2} \quad (12)$$

Hence, the probability distribution for the vehicular speed data is expressed as below.

$$p(v) = C e^{-\frac{v^2}{2\sigma^2}} \quad (13)$$

3.1 Data processing and analyses

Traffic studies may require different kinds of data; in terms of this, one may notice the references, e.g., [28, 29] for some descriptions and applications. Concerning the data usage in more recent studies, for example, low-frequency probe vehicle data are considered, e.g., in [30] to identify the network-wide turn-level intersection congestion. In another study [31], the complementary usage of traffic counter and probe car data is taken into account in estimating the path travel time. Besides, for instance in the car following models, trajectory dataset is considered, e.g., when investigating the impact of signal status on headway in [32], and generating Markov chain and training neural network in [33]. Regarding the purpose of this study, only vehicle speed data measured at a fixed point on the road is needed. Hence, for the analysis of vehicle speeds, one month speed data based on 24 h' measurement obtained from the selected surveillance point at İstanbul highway is considered. The data were measured at two-minute intervals, corresponding to the average values. For this study, three lane (left lane, middle lane, right lane) speed data in one direction are separated in two groups, each belonging to the fifteen days. Then, these six different data sets are employed, as denoted on Table 1.

After determining the data sets, the inconsistent speed values (null values, and zero values) were identified and discarded from each data set. By means of the box plot (Fig. 1), the outliers were tried to be visualized. It is easily noticed that those are more pronounced in Data set #4.

Table 1 Describing data set numbers for the analyses

15 days (first half)		15 days (second half)			
Data set #1	Data set #2	Data set #3	Data set #4	Data set #5	Data set #6
Right lane	Middle lane	Left lane	Right lane	Middle lane	Left lane

Fig. 1 Box plot for each speed data set

However, some of them are not omitted due to the characteristics of the data since it is surmised that those stem from the nature of the speeds (e.g., interactions). Only the Data set #4 involves very low values such as 1, and 2 km/h, which would be the potential outliers. On examining all the data sets, the values less than 12 km/h were discarded so that the data sets are consistent among each other.

As expected that the dynamics of the vehicular speeds varies depending on the each data set, and in order to compare them, the maximum entropy method of statistical mechanics is first tested. The histograms of the speeds extracted from the considered six data sets of traffic flow are plotted in Fig. 2.

The traffic flow involves non-random correlations, long-term memory and these could be associated with the speeds. According to the speed data sets, as the author expected that the probabilities could not follow an inverse relationship with energies (speeds) as is the case with Boltzmann–Gibbs thermostatics [34].

It is evident that the data does not comply with the fitting plots of a selected standard Gaussian distribution even if it will be scaled for each obtained histogram. Further, it is vague that how far the distributions are different from each other. Thus, this situation could not allow providing a reliable interpretation for making a quantitative comparison of the speed characteristics. As a result, to compare the vehicular speed characteristics by this method, i.e., maximum entropy method of statistical mechanics seems insufficient. For this reason, the paper

puts forward the KL divergence approach which provides divergences and these become quantitative indicators for the comparison of the distribution characteristics. In this approach, two common distributions, i.e., uniform and Gauss are selected as references. It is also wondered in this study whether there is a similarity between the trends of divergences from both reference distributions considering each data set. Let us analyze the speed data sets through the KL divergence approach in the following section.

4 Exerting KL divergence approach

As mentioned, the distribution function which is given by Eq. (13) cannot describe the fluctuations in the real vehicular speeds sufficiently, due to loss of information. That is, the analytical solution of this function cannot handle the analysis of speed data sets efficiently. The distribution function in fact may well suit for the mean speed region in the traffic flow. However, the real traffic data is time-dependent. At first glance, much as the appearances of some data sets seem close on the plots (Fig. 2), they have different dynamical behavior. The differences among dynamical behaviors of the speed data sets are outlined through the KL divergence approach. It is expected that the KL divergence values vary for each data set. Notice that albeit the existence of this variation, the flow does not exhibit nonequilibrium, i.e., the dynamic equilibrium is present overall, thus the system does not break down. In the following sections, one could find the results of the KL divergence analyses in the light of the two prior distributions.

4.1 The use of Gaussian prior distribution for KL divergence analysis: gaussian case

The fact that the distribution of vehicle speeds could not be described with Gaussian distribution most particularly stems from the tail region. Besides, the vehicular speed distributions have inherently different internal dynamics at each data set. To this end, the KL divergence technique tries to compare the dynamics of the speed distributions over Gaussian reference distribution in this section. This thereby allows for obtaining the differentiation from the Gaussianity.

Fig. 2 **a** Histogram of Data set #1 **b** Histogram of Data set #2 **c** Histogram of Data set #3 **d** Histogram of Data set #4 **e** Histogram of Data set #5 **f** Histogram of Data set #6

To be comparable, the KL divergences are calculated for six data sets based on a standard Gaussian prior distribution by Eq. (4). If the observed probabilities are considerably different from the probabilities of the Gaussian prior distribution, the related regions on the histograms could not be characterized with the reference Gaussian distribution. One can investigate the KL divergences on Fig. 3.

It is also inferred from the Gaussian case that, when the D_{KL} values tend to zero, the stability of the system would rise, and the influence of the tail region becomes negligible. In this respect, one would state that the system is steady-state and time-independent. It bears a relation to Gaussian-like dynamics, and this results in near zero values in the KL divergences, one can investigate Fig. 5 in terms of obtaining the divergences. As a result, the KL divergence analysis based on Gaussian prior distribution could suffice to describe the tail effects of the real data. Thus, the large deviations from Gaussianity in

Fig. 3 KL divergence values for each data set based on Gauss distribution

the tail region could inform us about the stability and indicate the time-dependency of the traffic flow.

4.2 The use of uniform prior distribution for KL divergence analysis: uniform case

The uniform distribution is also selected as a reference since it has certain probabilities for the given speed intervals, and it could be easily comparable, hence it is meaningful to select the uniformity as a prior distribution. By all means, uniform prior could be employed to indicate the divergences of speed distributions from the uniformly distributed probabilities. This also shows us the degree of divergence from the evenness of the distribution, as well as exhibits the uncertainty about the incident. Another reason to select uniform prior is to inquire whether there is a similarity between the trends of the KL divergence values obtained through Gaussian and uniform prior distributions. In order to definitely understand the grounds of the divergences, one should examine the speed histogram bins one by one, while considering Eq. (4).

For the uniform case, the fluctuations in D_{KL} values (Fig. 4) mainly hinge on the probability values in the mean speed region (if the dominant probabilities do not emerge in the tails) this is because larger divergences stem from there.

The trends of D_{KL} values obtained from both cases would be also relevant to surmise the dynamics of speeds at each data set in regard to crosschecking the D_{KL} trends of the cases. That is, should there occur an apparent distinction between the trends of the divergences obtained from the two cases, one may suspect, e.g., the influential peaks in the distribution, which is clarified in the following section. Furthermore, in divergence diagrams (Fig. 5), in the mean speed and its nearby region, there is a similar trend of the divergences of both cases within each data set. It is also inferred that the influences of the tail region on D_{KL} values are certainly significant in Gaussian case rather than in uniform case.

Fig. 4 KL divergence values for each data set based on uniform distribution

5 Results and discussion

This study attains the examination of the speed distribution characteristics based on Gaussian and uniform prior distributions through the KL divergence approach. For the Gaussian case, the significant probabilities in the tail region have much more influence on D_{KL} values since the prior probability is lower in this region, comparing with the uniform case. The results of both cases in the KL divergence approach allow comparing the speed distribution dynamics of each data set. In Fig. 5, the divergences obtained from both the Gaussian and uniform distributions considering histograms are displayed. In the plots, the more deviation from zero, the more those regions have divergences from the defined Gaussian or uniform probabilities. The divergences locate in different ranges for each data set, which as seen on the vertical axes. On the tail region, near zero values are prevalent for uniform case, while the higher values and fluctuations are more pronounced in Gaussian case, and these are rather noticeable in Data set #4. As a common inference for both cases, a discernable variation in the values emerges in the vicinity of mean speed. Please note that such a variation may be basically associated with the height of the distribution, and thus, the use of the specified prior distributions.

In this study, the proposed approach is tested with two ways. To do this, firstly next generation simulation (NGSIM) data are utilized. Considering five freeway lanes, a 24 h' speed data are extracted from the Interstate 80 (I-80) Freeway NGSIM data set [35]. For this study, each lane speed data of the freeway are designated as Data set #7, Data set #8, Data set #9, Data set #10, and Data set #11, respectively. The zero values were identified and discarded from each data set. The data sets were scaled and the histograms are depicted in this respect (Fig. 6). The analyses were repeated for Gaussian and uniform case by employing these data sets. The results show that the KL divergence trends obtained from the two cases agree once again (Figs. 8 and 9).

Secondly, four synthetic data sets (named as Data set #12, Data set #13, Data set #14, and Data set #15) are generated so as to pursue the possible differentiation between the KL divergence trends of the two cases. Some extreme cases of distributions are also tested in this way. A heavy-tailed data set is generated at first (Data set #12). After employing this data set along with the real data sets, the similarity of the trends of each case disappears considering a fast decline in the uniform case. Further, a sharp peak is placed in the left tail, then in mean speed region, and then in both regions, which are seen in the histograms of Data set #13, 14, and 15,

Fig. 5 The obtained divergences for each bin based on uniform and Gauss distribution; **a** and **b** Data set #1, **c** and **d** Data set #2, **e** and **f** Data set #3, **g** and **h** Data set #4, **i** and **j** Data set #5, **k** and **l** Data set #6. The plots on the left and right side correspond to uniform and Gaussian case, respectively

respectively (Fig. 7). The analyses are performed again for the two cases. With regard to all the data sets, the depiction of the KL divergence results is found in Figs. 8 and 9 for Gaussian and uniform case, respectively.

It is found that the trends do not remain similar after employing the synthetic data sets (Figs. 8 and 9). This finding would satisfy the aforementioned wonder of the study. Thus, it is inferred that the dissimilarity between the trends may emerge under certain circumstances.

To sum up, the analyses are conducted through the real speed and synthetic data sets. In the light of this, it is tentatively suggested that the KL divergence trends obtained through uniform and Gaussian prior distributions are analogous; however, the trends would differ as the data distribution approaches uniformity or influential peak(s)/outliers occur in the distribution.

The following results could be derived from the analyses.

- Once a heavy-tailed distribution is processed, the D_{KL} value obtained from uniform case would show a conspicuous decline (Data set #12).
- If a peak is divided into more than one peak in the given distribution, the D_{KL} value obtained from uniform case would decrease. However, switching the locations of those peaks in the distribution does not affect the D_{KL} value due to the fixed prior probabilities of uniform case, one may delve into the histograms of Data set #14 and Data set #15.
- The influence of the tail region on D_{KL} values is rather noticeable in Gaussian case; Data set #13 can be given as an extreme example.
- For Gaussian case, the locations of bin probabilities in the histograms are pivotal since they could affect the D_{KL} values. In terms of this, one may examine the D_{KL} results of Data set #13 and Data set #15.

Fig. 5 (continued)

6 Conclusions

In this study, the vehicular speed characteristics of the highway is examined through Kullback–Leibler (KL) divergence approach, addressing the insufficiency of maximum entropy method of statistical methods. By utilizing two prior distributions, i.e., Gaussian and uniform in this approach, the KL divergences are obtained. According to the divergences, the fluctuations among the speed distributions are revealed. Moreover, the results speak to non-random behavior of vehicular speeds. It is also found that the trends of the divergences based on the

two distributions are obtained analogous with regard to the real speed data sets. Through the employment of the synthetic data sets, the emergence of the dissimilarity between the trends is exhibited. Another result in this approach is that the influence of the tail region on the divergences is notable in Gaussian case. Consequently, the Gaussian case seems more sensitive to describe the dynamics of the tail region. Besides, the outcomes of the KL divergence analyses become quantitative indicator in order to make a relevant comparison of the speed distributions and better understand the dynamics of the vehicular speeds. As a final remark, the proposed approach may be adapted to different multi-particle complex systems.

Fig. 6 **a** Histogram of Data set #7 **b** Histogram of Data set #8 **c** Histogram of Data set #9 **d** Histogram of Data set #10 **e** Histogram of Data set #11, which are extracted from the NGSIM data set

Fig. 7 **a** Histogram of Data set #12 **b** Histogram of Data set #13 **c** Histogram of Data set #14 **d** Histogram of Data set #15 which are based on the generated synthetic data sets

Fig. 8 KL divergence values for all the data sets based on Gaussian distribution

Fig. 9 KL divergence values for all the data sets based on uniform distribution

Compliance with ethical standards

Conflict of interest The author states that there is no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Kosun C, Ozdemir S (2017) Determining the complexity of multi-component conformal systems: a platoon-based approach. *Phys A* 471:688–695. <https://doi.org/10.1016/j.physa.2016.12.027>
- Kosun C, Ozdemir S (2016) A superstatistical model of vehicular traffic flow. *Phys A* 444:466–475. <https://doi.org/10.1016/j.physa.2015.10.042>
- Zhang Y, Shang P (2019) The complexity–entropy causality plane based on multivariate multiscale distribution entropy of traffic time series. *Nonlinear Dyn* 95:617–629. <https://doi.org/10.1007/s11071-018-4586-2>
- Kosun C, Ozdemir S (2017) An entropy-based analysis of lane changing behavior: an interactive approach. *Traffic Inj Prev* 18(4):441–447. <https://doi.org/10.1080/15389588.2016.1204446>
- Yan Y, Zhang S, Tang J, Wang X (2017) Understanding characteristics in multivariate traffic flow time series from

- complex network structure. *Phys A* 477:149–160. <https://doi.org/10.1016/j.physa.2017.02.040>
6. Shang D, Xu M, Shang P (2017) Generalized sample entropy analysis for traffic signals based on similarity measure. *Phys A* 474:1–7. <https://doi.org/10.1016/j.physa.2017.01.061>
 7. Yin Y, Shang P (2016) Multivariate multiscale sample entropy of traffic time series. *Nonlinear Dyn* 86:479–488. <https://doi.org/10.1007/s11071-016-2901-3>
 8. Xu T, Xu X, Hu Y, Li X (2017) An entropy-based approach for evaluating travel time predictability based on vehicle trajectory data. *Entropy* 19(4):165. <https://doi.org/10.3390/e19040165>
 9. Karmakar (Sur) K, Majumder SK, (2008) Maximum entropy approach in a traffic stream. *Appl Math Comput* 195(1):61–65. <https://doi.org/10.1016/j.amc.2007.04.089>
 10. Silva MJ, Cavalcante TS, Rosso OA, Rodrigues JJ, Oliveira RA, Aquino AL (2019) Study about vehicles velocities using time causal information theory quantifiers. *Ad Hoc Netw* 89:22–34. <https://doi.org/10.1016/j.adhoc.2019.02.009>
 11. Aquino AL, Cavalcante TS, Almeida ES, Frery AC, Rosso OA (2015) Characterization of vehicle behavior with information theory. *Eur Phys J B*. <https://doi.org/10.1140/epjb/e2015-60384-x>
 12. McDonough K, Kolmanovsky I, Filev D, Yanakiev D, Zwabowski S, Michelini J, Abou-Nasr M (2011) Modeling of vehicle driving conditions using transition probability models. In: 2011 IEEE International Conference on Control Applications. <https://doi.org/10.1109/CCA.2011.6044388>
 13. Prokhorchuk A, Dauwels J, Jaillet P (2020) Estimating travel time distributions by Bayesian network inference. *IEEE Trans Intell Transport Syst* 21(5):1867–1876. <https://doi.org/10.1109/TITS.2019.2899906>
 14. Ernst JM, Krogmeier JV, Bullock DM (2012) Kullback-Leibler comparison framework for the evaluation of travel time distribution estimates. In: 2012 15th International IEEE Conference on Intelligent Transportation Systems. <https://doi.org/10.1109/ITSC.2012.6338645>
 15. Ernst JM, Krogmeier JV, Bullock DM (2014) Estimating required probe vehicle re-identification requirements for characterizing link travel times. *IEEE Intell Transport Syst Mag* 6(1):50–58. <https://doi.org/10.1109/ITS.2013.2288648>
 16. Ng TC, Adams M, Ibanez-Guzman J (2007) An information theoretic vehicle following system. In: 2007 IEEE Intelligent Vehicles Symposium. <https://doi.org/10.1109/IVS.2007.4290279>
 17. Zhu B, Jiang Y, Zhao J, He R, Bian N, Deng W (2019) Typical-driving-style-oriented personalized adaptive cruise control design based on human driving data. *Transport Res Part C Emerg Technol* 100:274–288. <https://doi.org/10.1016/j.trc.2019.01.025>
 18. Wang W, Liu C, Zhao D (2017) How much data is enough? A statistical approach with case study on longitudinal driving behavior. *IEEE Trans Intell Veh* 2(2):85–98. <https://doi.org/10.1109/TIV.2017.2720459>
 19. Du Y, Deng F, Liao F, Ji Y (2017) Understanding the distribution characteristics of bus speed based on geocoded data. *Transport Res Part C Emerg Technol* 82:337–357. <https://doi.org/10.1016/j.trc.2017.07.004>
 20. Liu, R, Zhu X (2019) Statistical characteristics of driver accelerating behavior and its probability model. ArXiv.
 21. Djenouri Y, Belhadi A, Lin JCW, Cano A (2019) Adapted k-nearest neighbors for detecting anomalies on spatio-temporal traffic flow. *IEEE Access* 7:10015–10027. <https://doi.org/10.1109/ACCESS.2019.2891933>
 22. Shannon CE (1948) A mathematical theory of communication. *Bell Syst Tech J* 27(3):379–423. <https://doi.org/10.1002/j.1538-7305.1948.tb01338.x>
 23. Karmeshu, (ed) (2003) Entropy measures, maximum entropy principle and emerging applications. Springer-Verlag, Berlin Heidelberg
 24. Bowley R, Sánchez M (1999) Introductory statistical mechanics. Oxford University Press, New York
 25. Tsallis C (2009) Introduction to nonextensive statistical mechanics: approaching a complex world. Springer, New York
 26. Niven RK (2004) The constrained entropy and cross-entropy functions. *Phys A* 334(3–4):444–458. <https://doi.org/10.1016/j.physa.2003.11.011>
 27. Kullback S, Leibler RA (1951) On information and sufficiency. *Ann Math Stat* 22(1):79–86
 28. Barceló J, Kuwahara M (eds) (2010) Traffic data collection and standardization. Springer-Verlag, New York
 29. Treiber M, Kesting A (2013) Traffic flow dynamics: data, models and simulation. Springer-Verlag, Berlin Heidelberg
 30. He Z, Qi G, Lu L, Chen Y (2019) Network-wide identification of turn-level intersection congestion using only low-frequency probe vehicle data. *Transport Res Part C Emerg Technol* 108:320–339. <https://doi.org/10.1016/j.trc.2019.10.001>
 31. Tani R, Owada T, Uchida K (2020) Path Travel Time Estimating Method by Incomplete Traffic Data. *Int J Intell Transport Syst Res* 18(1):43–52. <https://doi.org/10.1007/s13177-018-0168-4>
 32. Zhang J, Tang TQ, Wang T (2019) Some features of car-following behaviour in the vicinity of signalised intersection and how to model them. *IET Intel Transport Syst* 13(11):1686–1693. <https://doi.org/10.1049/iet-its.2018.5510>
 33. Tang TQ, Gui Y, Zhang J, Wang T (2020) Car-Following model based on deep learning and Markov theory. *J Transport Eng Part A Syst* 146(9). <https://doi.org/10.1061/JTEPBS.0000430>
 34. Kosun C, Ozdemir S (2020) The modified Brandeis dice for the traffic states dependency: The I-dice problem. *Comput Res Progress Appl Sci Eng (CRPASE)* 6(2):76–83
 35. U.S. Department of Transportation Federal Highway Administration. (2016). Next Generation Simulation (NGSIM) Vehicle Trajectories and Supporting Data. [Interstate 80 Freeway Dataset]. Provided by ITS Data Hub through Data.transportation.gov. Accessed 2020–12–06 from <https://doi.org/10.21949/1504477>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.