4-29-2019

Driving Performances Assessment Based on Speed Variation Using Dedicated Route Truck GPS Data

Ying Li
Chang'an University

Li Zhao
University of Nebraska - Lincoln, lizhao@unl.edu

Laurence R. Rilett
Chang'an University

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 7, 2019, accepted March 17, 2019, date of publication April 9, 2019, date of current version April 29, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2909572

Driving Performances Assessment Based on Speed Variation Using Dedicated Route Truck GPS Data

YING LI[1,2], LI ZHAO[3], AND LAURENCE R. RILETT[1,3]

[1]School of Information Engineering, Chang'an University, Xi'an 710064, China
[2]School of Intelligent Systems Engineering, Sun Yat-sen University, Guangzhou 510275, China
[3]Department of Civil Engineering, University of Nebraska–Lincoln, Lincoln, NE 68588, USA

Corresponding author: Laurence R. Rilett (lrilett2@unl.edu)

This research was supported by the China Postdoctoral Science Foundation under Grant 2018M633442, the National Natural Science Foundation under Grant 71871028, and Fundamental Research Funds for the Central Universities 300102249309.

ABSTRACT It was hypothesized that a driver is not safe when travel speed is too high and also not necessarily safe when travel speed is too low. Based on this hypothesis, this paper studied the risky driving performances by measuring speed variations of a driver's recurrent trips in two perspectives: 1) driver profiles, which scored the risk on-road driving of each driver and 2) driving patterns, which reflected the risk speed patterns of a type of drivers. The proposed method was tested on a 30-day global positioning system (GPS) dataset, collected from 100 trucks. The study first split the raw dataset into trips and finds the most repeatedly traveled route. Next, the frequency and amplitude of the speed variations from trips of each truck are calculated to establish driver profiles. A risk score is used to rank the truck drivers, i.e., a higher score indicates that the truck driver is more likely to conduct risky driving performances. All trucks are featured in four pre-defined driving patterns according to the different types of speed variations. The geospatial speed distribution of several trucks is manually examined from the raw dataset to verify the results. The contribution lies in providing a method to evaluate a driver's risk performance through mass truck GPS data. The proposed method would help for monitoring on-road risky driving performances in large fleet management and also providing knowledge about driving styles among drivers which would be beneficial in study driver assistant system.

INDEX TERMS Driving pattern, dedicated route, global positioning system, trajectory, speed variation, risky driving performance.

I. INTRODUCTION

Since the first study of the Global Positioning System (GPS) data in the mid-1990s [1], [2], it has gained increasing attention in surveying individual driver behavior. GPS trace data allow researchers to discover latent heterogeneities that existed in an individual's driving behavior across time and space because it provides a complete record of day-to-day driving data at very detailed levels. Using GPS data is beneficial in improving road safety in that it can help to develop more precise methods to describe or predict drivers' risky behavior.

Even for the same driver at the same location, driver behavior would vary a lot across time and for different situations [3], and these differences would be captured by the GPS data. As such, the variation of driving behavior over time can be used as an easy way to evaluate a driver's performance on the repeated trips over the same route. Thus, driver's risky behaviors can be explored using metrics derived from GPS data, such as speed variation or speeding, as proxies for driving risk assessment [3], [4].

This paper characterized and ranked the risk of truck drivers based on their speed variations. The drivers were grouped into four pre-defined driving patterns. The method was tested on a GPS dataset that includes 100 long-haul commercial trucks traveling on their dedicated routes in China over 30 days. The trips and routes that each truck traveled on were extracted from the raw GPS data. The speed variations are obtained by comparing the speed profile of each truck and the median speed reference at different segments of the

The associate editor coordinating the review of this manuscript and approving it for publication was Zhengbing He.

geospatial route. Subsequently, the frequency and amplitude of the speed variations are measured to evaluate a driver's on-road performance. At an individual level, driver profiles are established to rank risky drivers. At a system level, driving patterns are defined and used to categorize drivers that might be considered risky or unsafe.

The remainder of this paper is structured as follows. Section 2 presents related work for mining GPS data in the study of driver behavior. Data preparation and definitions are introduced in section 3. The metrics used for driver performance is presented in section 4. Driver profiles and driving patterns are evaluated in section 5. Finally, section 6 concludes the paper and provides limitation discussions and recommendations for future work.

II. LITERATURE REVIEW OF GPS BASED DRIVER BEHAVIOR STUDIES

In the field of road safety, it is widely recognized that driver behavior is a large contributory factor, on the order of 90 percent [5], of crashes as compared to other factors such as road conditions or demographic characteristics [6]. With the help of GPS, individual driving behavior associated with safety has been widely studied in the literature. Driving risk indices can be used as indicators of risk involvement in car crashes [7]. At-risk behaviors such as improper braking and inappropriate speeding, where drivers behave more aggressively, are positively related to crashes or near-crashes, regardless of traffic conditions [8], [9]. Two key metrics for identifying risk driver behavior using multi-featured GPS trace data are driver speed profiles and driving patterns over time.

Driver profiles aim to establish a scoring system to evaluate driver behavior regarding the risk of a casualty crash. Usually, a risk index is used to score drivers which comprises of metrics such as lane changes, speeding, and hard acceleration. The higher risk index scores the more unsafe driver. Toledo *et al*. studied driver profiles and use profiling to measure behavior improvements that occur after an external policy or environmental change [10]. It was found that a statistically significant relationship between the risk index and crash history. The study provided a comprehensive method for monitoring drivers over time and suggested the need for a method which accounts for the complexity of the driving task.

Ellison *et al*. analyzed 106 drivers from a pay-as-you-drive study using GPS data from the first five weeks of the total ten weeks. Speed and acceleration were used as safety metrics [3]. The study detailed the establishment of driver risk profiles using a risk index ranging from 1 (i.e., low risk) to 100 (i.e., high risk). The risk index consisted of a risk score and a risk margin. The risk score represented how safe an individual driver is, and the risk margin represented the risk range of the same driver. It was found that road environment strongly affected driving behavior and spatiotemporal environments were reasons of drivers' various psychological responses. The study also claimed a contribution for evaluating driver behavior changes in before-and-after studies.

Driving patterns aim to classify different patterns of on-road driving across drivers. The driving behavior is usually grouped into several patterns to emphasize at-risk driving such as aggressive, erratic, or distractive driving. These studies often use pattern recognition techniques [12] such as supervised or unsupervised classification, and dimension reduction. These techniques are based on measuring the differences of the patterns, which can be expressed as a distance matrix using similarity measure [13] or other measures [14]. Zhu *et al*. studied smartphone GPS data from 12 test drivers when they are traveling [15]. The study used variabilities of speed, including speed changes, acceleration, and deceleration as indicators of at-risk driving. A less risky driver was assumed to associate with a smooth speed pattern, and a more risky driver was regarded to conduct an erratic speed pattern. Although the study only used data from two drivers at an individual level, it provided a framework to evaluate driving patterns and identify potential at-risk drivers.

Brambilla *et al*. studied 27 trips from GPS data in order to extract recurrent driving patterns from trips to detect different behaviors [14]. Three variables were used: acceleration, speed and the difference in yaw. Using K-means clustering algorithm, six clusters of driving styles were found by grouping the percentage of points within each trip that belongs to each cluster. Experts' judgment was used as ground truth. The study identified three driving patterns and was found that the six automatic identified clusters of driving styles fit well in the three patterns with a precision of 96 percent.

Many metrics can be used to measure the difference of driver profiles or driving patterns in related to the potential or actual risk of an incident. These metrics may include speed, acceleration, jerk, lane change, travel duration, the frequency of braking, and so on. Among them, the most commonly used metrics are speed and speed variations [3], [9], [15], [16]. Speed measures usually consist of some calculations such as the maximum, average, minimum and standard deviation of speeds, speed limit, and speed stability duration.

Speed measures that related to driver behavior and crash involvement, as a surrogate safety measure [17], are inclining to use GPS data techniques. GPS data provide a complete speed trajectory data before, during and after the occurrence of a crash, and that information is increasing. Other achieved data, from conventional methods (e.g., travel survey and crash report) are not typically as informative.

However, the huge sample size and high dimensionality of GPS trajectory data bring statistical and computational challenges that hinder their widespread adoption for travel behavior studies [18]. First, raw GPS data require more than multiple dimensions to represent cannot be used directly. A considerable amount of efforts needs to be put into processing the raw GPS data into a trip-log format that describes travel behavior regarding a related set of origins, destinations, trips, journeys, and routes. Second, the data quality caused by system error or random error [19], missing data, and the

different sampling rate is common in various forms of GPS data. What is more, critical information is sometimes not captured in the GPS data, including speed reference (i.e., speed limit), driver characteristics, and trip purpose. This makes the extraction of driver behavior features even harder to achieve. In such a context, the amount of GPS traced vehicle samples used in previous driver behavior studies is very limited. Most of these studies focus on establishing complex models (14), analyzing a few individual cases collected from experiments [14], [15], or studying the distribution of travel behavior such as trip generation and purpose [18]. Therefore, although the use of GPS data is promising, there is still a need for new methodologies for extracting valuable information from the large GPS datasets, and a great room for research on driver behavior patterns based on GPS data.

This paper firstly explains the challenges in converting geocoded raw GPS data points into a meaningful database that describes the trips and journeys of a truck. Then, the frequency and amplitude of speed variations in spatial trips are used as measures of risky driving behaviors, which was tested on a dataset of 100 truck GPS data. It tries to characterize the driving variations of each driver through repeated trips on the same route and recognize recurrent behaviors shared between drivers using a criteria-based method. The proposed method in this paper helps to 1) further understand on-road driving behavior at both individual level and system level; and 2) identify risk drivers or risky driving patterns that be interested in fleet manager, vehicle insurance investor, or driver education officer.

III. DEFINITIONS AND DATA

A. TRAJECTORY DATA PROCESSING

In general, a trajectory depicts a continuous motion history of an object over time in the Euclidean space. In the transportation field, vehicle trajectory is a sequence of consecutive geo-referenced coordinates that are recorded at a specific frequency over a period of time. Using positioning devices such as GPS, vehicles can be tracked over both space and time. A GPS-recording vehicle trajectory P, with R data points, is mathematically defined as:

$$P_R = [(p_1, t_I), (p_2, t_2), \ldots, (p_r, t_r), \ldots, (p_R, t_R)] \quad (1)$$

where P represents a trajectory and p represents a data point on the trajectory P, r is the r^{th} data point, and R is the total number of the data points of trajectory P. Each data point of the trajectory (i.e., $\forall p_r \in P_R$) is recorded at timestamp t, therefore (p_r,t_r) represents the r^{th} data point which is collected at time t_r. If the GPS recording interval (i.e., t_r - t_{r-1}) remains the same for all trajectories of all vehicles in the study, the timestamp feature can be excluded from the dataset for the purpose of dimensionality reduction. In this situation, the general identifiable features for a moving vehicle are geographic coordinates (i.e., latitude and longitude), speed, and other features (e.g., yaw, altitude, mileage, etc.). A simplified data point (without associated timestamp) from GPS with the same recording rate may be expressed as:

$$p_r = (x_r, y_r, v_r), r \in [1, 2, \ldots, R] \quad (2)$$

where x_r, y_r, and v_r represent latitude, longitude, and speed of r^{th} data point, respectively, and R is the total number of data points. Note that equation (2) only illustrates three features. However, it is readily generalizable to more dimensions.

In general, an analyst will follow the 4-step below to transform the raw GPS data into an informative dataset that may be used to study safety.

1) **Exclude short trips or system errors** (i.e., unrealistic records of GPS data). A valid trajectory needs to maintain a certain length, and the values should be reasonable [20]. A trajectory that lasts several minutes or less and extraordinary outliers, for example, negative speed or speed over 200 km/h [21] is excluded from the dataset. Note that thresholds should not be too rigorous in order to avoid removing valid speeds;
2) **Organize the direction of trajectory data**. The trajectory derived from GPS data is stored as a manner of time order in the data stack. It would be misaligned when comparing trajectories in the same route however in different directions. For the ease of calculation, the direction feature should be reorganized by comparing the latitude and longitude coordinates of the beginning and end of trajectories;
3) **Smooth the speed**. Systematic errors can be readily identified and removed, as discussed in step 1. However random errors are more difficult to address. Thus, a filter, such as the Kalman filter, is used for speed denoising [19]; and
4) **Divide a trip into several segments**. A long trip may experience several different traffic conditions (i.e., different speed limits). Therefore, trips are divided based on their geospatial locations. Within a small segment, it is assumed that the traffic condition keeps the same.

B. DATA PREPARATION

To study the driving performance, it is expected that the GPS trajectory represents a trip that may have the following features:

1) Run on the same road segment or the same route repeatedly. Thus, the driving pattern can be identified through multiple ''experiments'';
2) Less influenced by external traffic flow such as traffic jam or road condition such as work zone. Due to the frequent speed changes, influenced by external traffic conditions may exist; and
3) Less influenced by traffic control such as traffic signal.

By observing the above requirements, truck data from the Chinese road freight vehicle monitoring and service platform (hereafter refer as ''the platform'') seems ideal. The platform was established in 2014 and is the world's largest commercial vehicle networking platform. To the authors' knowledge, it is also the only national-level monitoring platform for commercial trucks (e.g., heavy trucks and semi-trailer tractors

over 12 tons) in China. Before entering service, all trucks are equipped with a GPS. The GPS is out of reach of the driver and is designed to continuously transfer log data to the platform (i.e., GPS cannot be turned off). In another word, the embedded GPS in the truck records information and transmits it to the platform at a 30-second interval, throughout the life of the truck (e.g., collects data even when the engine is turned off). The GPS log data include timestamped vehicle ID, geographic location (latitude and longitude), altitude, speed, angle, mileage and warning information such as fatigue.

In this paper, we focus on trucks which traveled on their repeated routes. The trucks transport goods back and forth between two locations along the same route on a regular basis. Although truck drivers can be assigned several routes, a dedicated truck driver is committed to only a few routes most of the time. Note the specific geospatial information of the dedicated routes is estimated from the dataset.

The GPS dataset used in this paper includes 100 trucks collected on April 1-30, 2016 that followed their routes across China (i.e., 100 trucks run on 100 routes). There would be more than one driver that is assigned to operate a truck, and the truck is run by which driver(s) is unknown. Thus, it was assumed in this study that each truck was operated by one driver from the dataset. The implications of this assumption will be discussed later in the paper.

It should be noted that the platform does not contain other useful information (e.g., routes, speed limit, driver information, and driving conditions) and this information could not be obtained from other sources. A preprocessing procedure, which the general steps have been discussed before, was developed to clean the data. The critical steps of the data preparation are shown below.

1) TRIP DEFINITION

The most challenging part of the data sorting is breaking the raw GPS data into individual trips [22]. The raw GPS dataset contains locations and speed information at 30-second intervals. The raw data include those times when the truck is stopped and the engine is turned off. In this study two metrics, the time duration for speed dwell (t_{sd}) and time duration for speed gap (t_{sg}), are used to define a trip. Speed dwell measures the period when the truck's speed is greater than zero. In contrast, the speed gap measures the period when a truck's speed is equal to zero. Both durations should be greater than the predefined time thresholds in order for the data to be considered valid for future analysis. The relationships are expressed in (3) and (4), respectively.

$$t_{sdj} \geq T_{sd} \tag{3}$$

$$t_{sg,j} \geq T_{sg} \tag{4}$$

where, $t_{sd,j}$ and $t_{sg,j}$ represent the time of speed dwell and the time of speed gap for trip j ($j = 1, 2, \ldots, J$), respectively. T_{sd} and T_{sg} represent the threshold for speed dwell and the threshold for speed gap, respectively.

Using equation (3), a trip is defined as a set of points where the speeds are all non-zero, and the length is greater than the speed dwell threshold. That said, any speed dwell period t_{sd} that is greater than a predefined threshold T_{sd} is considered a valid trip in this paper. Equation (4) defines the speed gap, which consists of a set of zero in speed indicating there is no motion in a period longer than the speed gap threshold. Speed gap is critical to split the continuous dataset into trips. The determination of the threshold is discussed later.

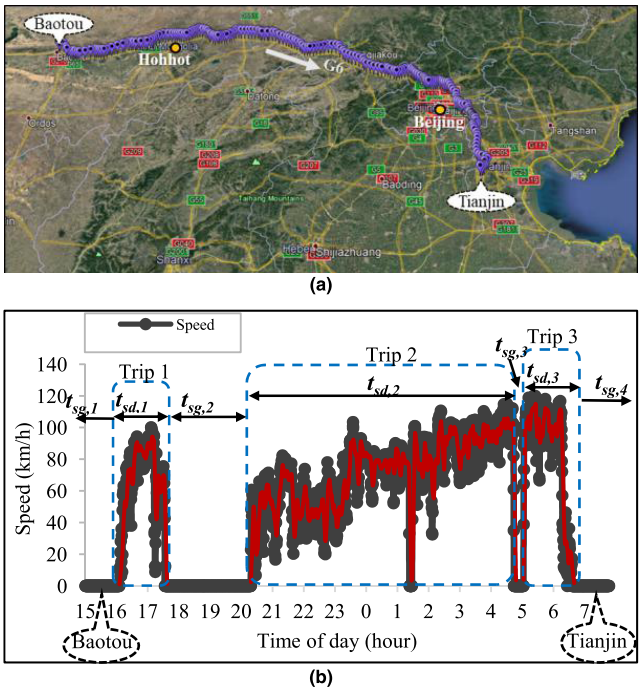


FIGURE 1. An example of the manually identified trip. (a) Geospatial coordinates (for demonstration, here show the coordinates at a 5-minute interval). (b) Speed versus time for a truck's entire journey.

If several trips belong to one travel purpose, a journey is then defined as the total trips from the origin to the destination through a route. In other words, a journey from start to finish may include several stops, which breaks the journey into several trips. Figure 1(a) shows an example of a truck that traveled from Baotou to Tianjin, China. The speed gaps before the start and the ending data points were more than 24 hours indicating this was a single journey. The total mileage of the journey was 783 km. The route that the truck followed was mainly on the Beijing-Tibet expressway (G6). Figure 1(b) shows the speed versus time of the truck's entire journey. It may be seen that the truck left Baotou at approximately 15:30 in the afternoon and arrived at Tianjin at approximately 6:30 the next morning.

In figure 1(b), the journey was split into three trips by two speed gaps in between and two speed gaps at the beginning and the end of the journey. Both thresholds for T_{sd} and T_{sg} are used as 15 minutes for this example. Note that approximately 1:20 a.m. the truck speed was zero and it lasted

about 3 minutes. It was found by carefully checking the geo-coordinates that the truck did not deviate from its route during this short time (i.e., 3 minutes gap). Therefore, the 3-minute of zero speed was not identified as a valid speed gap, therefore, it did not split trip 2 into two parts. It is hypothesized that the truck was trapped in temporary congestion or conducted a roadside stop. Note that a Kalman Filter algorithm was used for speed denoising [23], where the grey line represents the raw GPS speed, and the red line represents the smoothed speed for each trip.

2) SPEED GAP THRESHOLD

In order to identify a trip, a key parameter is the speed gap threshold, T_{sg}, which is set by the user and is a function of the application. The speed gap threshold values that had been identified in the literature include 2 min [24], 3 min [21], 5 min [25], and 15 min [22], which depended on the characteristics of the applications such as travel mode classification and trip purposes.

There are several situations where the speed may register as zero over a given period:

1) Driver intentional stop (e.g., rest, eating and maintenance);
2) Traffic congestion;
3) Traffic control(e.g., traffic signal, stop sign); and
4) Temporary roadside parking (e.g., stopping to make a phone call, changing drivers, etc.).

For category 1 the speed gaps tend to be longer, all else being equal, and the driver deviates from the geospatial route (e.g., drives to a restaurant) and then returns. For categories 2, 3, and 4, the speed gaps tend to be shorter, all else being equal, and the driver often stays on the geospatial route. Since this study is focused on driving pattern recognition at the level of the trip unit, the chosen speed gap threshold T_{sg} should be long enough to differentiate between the driver's intentional stops and other stops.

To identify T_{sg}, a sensitivity analysis using eight thresholds (e.g., 5 min, 10 min, 15 min, 20min, 30 min, 40 min, 50min, and 60 min) was conducted. Intuitively, as T_{sg} decreases as the number of separate trips identified will increase. Take the first truck in the dataset as an example, it is not surprising to get the largest number of trips identified (e.g., 213) occurred for 5 min threshold and the smallest number of trips (e.g., 69) occurred for the 60 min threshold. The number of trips identified for thresholds of 20 min, 30 min, 40 min, 50min, and 60 min resulted in approximately 81 trips on average and the deviation among them was on the order of 10 percent. In other words, the number of trips did not change appreciably for this range of speed thresholds.

Based on the sensitivity analysis, T_{sg} was set 15 min for the dataset in this study. It was also felt that given the long-distance travel of the tucks and the fact that stops were relatively few, the 15 minutes made intuitive sense. Based on this threshold, 10294 trips were identified for all 100 trucks.

3) DEDICATED ROUTE

It should be noted that given a dedicated route, the trips on it may vary over time. For example, drivers may take slight deviations in the routes from day to day, and there are also uncertainties introduced by discrete sampling and sampling error [26]. Therefore, a method was developed to measure the similarity among routes by comparing the degree of overlap between trips. The similarity measure chosen was the longest common subsequence (LCS) method [27].

LCS measures the number of matched data points between two trajectories. It has the advantage of allowing some points, or outliers, to be left unmatched. This is a great feature because it allows for slight deviations in a route that, as described above, exist in the dataset in this study. It is also useful for comparing trips that have different lengths, which exist in the dataset as well. The following parts describe the similarity measure of two routes using LCS method.

Let $H(P_{r,j})$ and $H(P'_{r',j'})$ be the first r and r' points from trajectories $P_{R,j}$ and $P'_{R',j'}$ ($r = 1, 2, \ldots, R$ and $r' = 1, 2, \ldots, R'$), respectively. $H(*)$ is defined as a head function which returns a sequence of data points from the head of a given trajectory, such that $H(P_{r,j}) = \{p_{1,j}, p_{2,j}, \ldots, p_{r,j}\}$ and $H(P'_{r',j'}) = \{p'_{1,j'}, p'_{2,j'}, \ldots, p'_{r',j'}\}$ for trajectory j ($j = 1, 2, \ldots, J$) and trajectory j'($j' = 1, 2, \ldots, J'$), respectively. Given two data points, the Euclidian distance between them may be readily calculated as follows:

$$dist\left(p_{r,j}, p_{r',j'}\right) = \sqrt{\left(x_{r,j} - x_{r'j'}\right)^2 + \left(y_{r,j} - y_{r',j'}\right)^2} \quad (5)$$

where $dist\left(p_{r,j}, p'_{r',j'}\right)$ calculates the Euclidian distance between data point p_r in trajectory j and data point $p'_{r'}$ in trajectory j' with the geo-coordinates of x and y are latitude and longitude in each data point respectively.

A predefined constant ε is used as the distance threshold of two data points that are classified as being matched points. That said, if this distance is less than the matching threshold ε, then the two points are considered matched. Therefore, a recursive function $LCS_\varepsilon\left(P_{R,j}, P'_{R',j'}\right)$ is defined to calculate the LCS score for any two trajectories. Thus, when $R = 0$ or $R' = 0$, $LCS_\varepsilon\left(P_{R,j}, P'_{R',j'}\right) = 0$. When $R \neq 0$ and $R' \neq 0$:

$$LCS_\varepsilon\left(P_{R,j}, P'_{R',j'}\right) = \begin{cases} 1 + LCS_\varepsilon\left(P_{R-1,j}, P'_{R'-1,j'}\right) & dist\left(p_{r,j}, p'_{r',j'}\right) > \varepsilon \\ \max\left(LCS_\varepsilon\left(P_{R-1,j}, P'_{R',j'}\right), \right. & \\ \left. LCS_\varepsilon\left(P_{R,j}, P'_{R'-1,j'}\right)\right) & dist\left(p_{r,j}, p'_{r',j'}\right) > \varepsilon \end{cases} \quad (6)$$

Not surprisingly, a key input of the LCS measure methodology is to choose an appropriate distance threshold ε. In general, ε is a function of the granularity of the data points in the trajectory. This granularity can be measured by a linear function of the standard deviation of the Euclidian

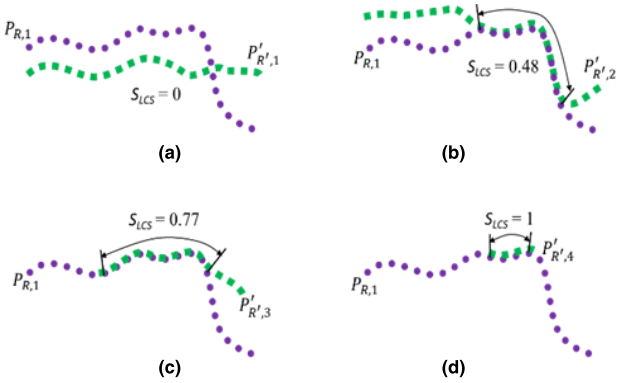


FIGURE 2. Similarity measure of two trajectories.

distance between the trajectories being examined. In this paper, a distance metric of three times the standard deviation (i.e., $\varepsilon = 3{*}std.$) was used to identify whether two data points from two different trajectories "match" or not.

The similarity S_{Lcs} between two trajectories j and $j^{'}$ can be found as shown in (7). It may be seen that the metric is the quotient of the longest common subsequence score and the minimum length of the two trajectories, where length refers to the number of data points in a trajectory.

$$S_{LCS}\left(P_{R,j}, P'_{R',j'}\right) = \frac{LCS_{\varepsilon}\left(P_{R,j}, P'_{R',j'}\right)}{min\left(R, R'\right)} \tag{7}$$

By definition, the similarity S_{Lcs} takes a value between 0 and 1. The metric represents how much two trajectories overlap each other. As the S_{Lcs} value increases, so too does the amount of overlap, or similarity, between the two trajectories. Figure 2 illustrates four examples of calculating the similarity S_{Lcs} of two trajectories of $P_{R,j}$ and $P^{'}_{R',j'}$, where $j = 1$ and $j^{'} = 1$, 2, 3, and 4 in the four examples, respectively. The base trajectory $P_{R,1}$ is shown in purple dots and consists of 23 data points. Four trajectories: $P^{'}_{R',1}$ in figure 2(a), $P^{'}_{R',2}$ in figure 2(b), $P^{'}_{R',2}$ in figure 2(c), and $P^{'}_{R',4}$ in figure 2(d) are shown in green squares and consists of 19, 23, 13, and 4 data points, respectively.

Figure 2(a) shows an example of two trajectories that do not share any common data pairs. Mathematically, it means that the distance between any points $\boldsymbol{p_{r,1}}$ in trajectory $\boldsymbol{P_{R,1}}$ and any points $\boldsymbol{p'_{r',1}}$ in trajectory $\boldsymbol{P^{'}_{R',1}}$ is greater than the threshold $(\boldsymbol{dist}\,(\boldsymbol{p_{r,1}}, \boldsymbol{p'_{r',1}}) > \boldsymbol{\varepsilon}$, assuming $\boldsymbol{\varepsilon}$ is very small in these examples). In this case the similarity value is 0 (i.e., 0/*min*(19, 23)). Figure 2(b) shows an example where 11 of the data points in trajectory $\boldsymbol{P^{'}_{R',2}}$ that match with data points in trajectory $\boldsymbol{P_{R,1}}$(i.e., the distances between the green and purple dots are within the distance threshold). Therefore, the similarity value is 0.48 (i.e., 11/*min*(23, 23)). Similarly, the trajectories in Figure 2 (c) have a similarity value of 0.77 (i.e., 10/*min*(13, 23)). If one trajectory overlaps completely with another trajectory, as shown in Figure 2 (d), then the similarity between the two trajectories is 1 (i.e., 4/*min*(4, 23)).

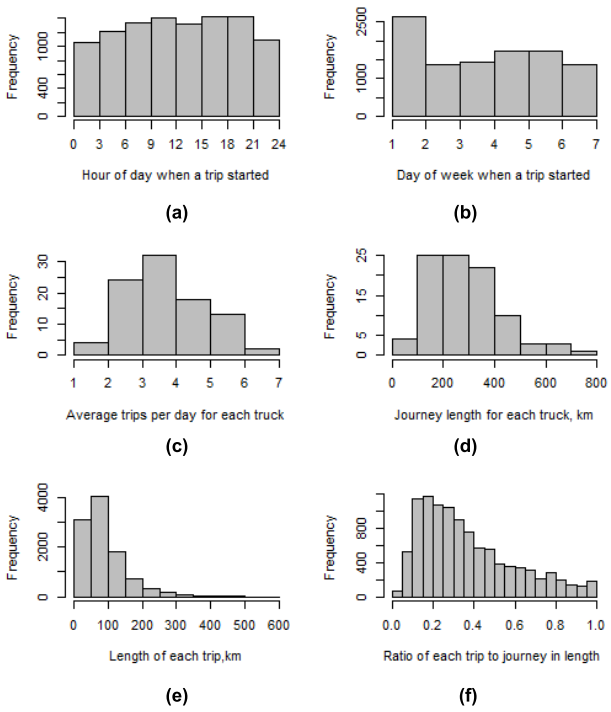


FIGURE 3. Summary of the trips.

Using the similarity measure method, trips with high similarity values (e.g., $\geq 85\%$) indicate that those trips are from the same route. In this study, the dedicated route for each truck is identified by classifying the most repeated trips (i.e., the greatest number of trips with high similarity values). Therefore, trips on the dedicated route for each truck will be studied with respect to their speed variations.

A dynamic programming model was developed to automate the comparison of the different trips from the study. Figure 3 shows statistics related to all those trips and journeys from all 100 trucks (i.e., 100 routes) that are identified.

Figure 3(a) shows the start time of the trips. It can be seen that the start time of the trips is relatively uniform distributed with only a weak bimodal trend. Figure 3(b) shows the day of the week when a trip begins, and it may be seen that Monday has considerably more trips than the other days of the week. On average, each truck experiences between 1 to 7 trips per day, as shown in figure 3(c). The mean trips per day is 3.7 trips and the standard deviation is 1.1 trips/day. In figure 3(d), it may be seen that the total route distance ranges from 53 km to 779 km. The mean is 303 km and the standard deviation is 165 km. Because a journey may have multiple trips, the distribution of the trip length is a useful metric, which is shown in figure 3 (e). It may be seen that the majority of trips are less than 100 km and most of the trips are less than 200 km. Figure 3(f) shows the ratio of each trip length to its respective journey length. For example, if a trip is 100 km and its journey is 400 km, then the ratio is 0.25. In other words, this trip represents one-quarter of the total journey.

IV. PERFORMANCE METRICS

Many studies have demonstrated that the speed variation, over both a given trip and across different trips, are an excellent surrogate measure of at-risk driving [9], [15]–[17]. Speed variation in this paper is defined as the speed differences between the average speed for a given trip and the average speed across all trips for a given truck. In addition, it is important to compare the speed of the truck at the ''same'' location across various trips. It is hypothesized that the lower variation in the speed at a particular location and the less risky driving behavior. In summary, the frequency and amplitude (including mean and standard deviation) of the truck speed variations will be used to categorize the truck drivers' driving performances into several distinct patterns. The mathematics behind this approach are discussed below.

A. MEDIAN SPEED REFERENCE

As mentioned before, the speed limit is often used as a reference value for identifying speeding behavior [3]. Unfortunately, the speed limit is not recorded in the dataset. Thus, this study uses Median Speed Reference (MSR) to represent the speed reference for a given segment of the trip. It is assumed that the segments are small enough that the speeds in the segment are homogeneous. A benefit of using the median as a measure of central tendency is that it is less affected by outliers and skewed data. The MSR for the segment k_{i,j_i} in trip j_i of truck i is calculated as follows:

$$\begin{aligned} MRS_{i,j_i,k_{i,j_i}} &= Median\left(s_{i,j_i,r_{i,j_i}} \mid k_{i,j_i}, j_i, i\right), \\ i &\in [1, 2, \ldots, I], j_i \in [1, 2, \ldots, J_i], \\ k_{i,j_i} &\in \left[1, 2, \ldots, K_{i,j_i}(\omega_{i,j_i})\right], r_{i,j_i} \in \left[1, 2, \ldots, R_{i,j_i}\right] \end{aligned} \tag{8}$$

where

i, I	A truck i from the dataset,with the total truck number I.
j_i, J_i	A trip j_i of a truck i, with the total trip number of each truck J_i.
k_{i,j_i}, K_{i,j_i}	A segment k_{i,j_i} in trip j_i of a truck i, with the total segment number of each truck K_{i,j_i}. In this study, the number of segments K_{i,j_i} is a function of the segment ratio ω_{i,j_i} for each trip j_i of each truck i, $0 < \omega_{i,j_i} < 1$.
r_{i,j_i}, R_{i,j_i}	A data point r_{i,j_i} in aa trip j_i of a truck i, with the total number of data points R_{i,j_i}.
$s_{i,j_i,r_{i,j_i}}$	The median speed from all data points r_{i,j_i}, given the segment k_{i,j_i} in a trip j_i of a truck i.

The lengths of segments are set to the same value for a given trip. However, segment length can vary across different trips because it depends on the entire length of a trip. In general, a long trip has more segments than a short trip. The parameter ω_{i,j_i} is used to control the number of segments for each trip. It is calculated as the quotient of the number of data points in the segment dividing and the total number of data points in a trip. The number of segments should be chosen with care. If it is too large, there may be too few data points, which may cause errors and unstable results in the similarity calculation. If it is too small, there may be too many data points and the segments may not be homogeneous. In addition, the computational cost is increased exponentially. In this study, given the sampling interval between two successive data points is 30 seconds, a segment length of 5 min (i.e., 10 data points) is used, i.e., $\omega_{i,j_i} = 30/300 = 0.1$.

The speed difference $D_{i,j_i,r_{i,j_i}}$ for data point r_{i,j_i} and the corresponding $MRS_{i,j_i,k_{i,j_i}}$ in segment k_{i,j_i} in trip j_i of truck i can be calculated as follows:

$$D_{i,j_i,r_{i,j_i}} = s_{i,j_i,r_{i,j_i}} - MRS_{i,j_i,k_{i,j_i}} \tag{9}$$

Note the first and last segments of a trip are removed (i.e., $2 < k_{i,j_i} < K_{i,j_i}$-1) due to the possible large external influences when the trip starts or ends.

B. FREQUENCY OF SPEED VARIATION FOR DRIVER PROFILES

The performance of a given trip is calculated by the frequency of speed variations. In words, it is the percentage of time that a vehicle's spend difference (between speeds and the MSR) is above or below the predefined DS_m ($m = 1$ and 2). Two values of DS_m are used: $DS_1 = 8$ km/h and $DS_2 = 16$ km/h. The frequency of speed variation higher than DS$_m$, i.e., F^h_{i,DS_m} and the frequency of speed variation lower than DS_m, i.e., F^l_{i,DS_m} for each truck i ($i = 1, 2, \ldots, I$) are shown in (10) – (11) and (12) – (13), respectively.

$$F^h_i = \frac{1}{J_i}\sum_{j_i=1}^{J_i} \frac{\sum_{r_{i,j_i}=1}^{R_{i,j_i}} C^h_{i,j_i,r_{i,j_i}}}{R_{i,j_i}}, \tag{10}$$

$$C^h_{i,j_i,r_{i,j_i}} = \begin{cases} 1, & if\ D_{i,j_i,r_{i,j_i}} > DS_m \\ 0, & otherwise \end{cases} \tag{11}$$

$$F^l_i = \frac{1}{J_i}\sum_{j_i=1}^{J_i} \frac{\sum_{r_{i,j_i}=1}^{R_{i,j_i}} C^l_{i,j_i,r_{i,j_i}}}{R_{i,j_i}} \tag{12}$$

$$C^l_{i,j_i,r_{i,j_i}} = \begin{cases} 1, & if\ D_{i,j_i,r_{i,j_i}} < -DS_m \\ 0, & otherwise \end{cases} \tag{13}$$

where $C^h_{i,j_i,r_{i,j_i}}$ and $C^l_{i,j_i,r_{i,j_i}}$ are dummy variables. If the speed variation is higher or lower than DS_m, then $C^h_{i,j_i,r_{i,j_i}}$ and $C^l_{i,j_i,r_{i,j_i}}$ are set to 1, otherwise $C^h_{i,j_i,r_{i,j_i}}$ and $C^l_{i,j_i,r_{i,j_i}}$ are set to 0. For each data points $r_{i,j_i} \in \left[1, 2, \ldots, R_{i,j_i}\right]$ in a trip $j_i \in [1, 2, \ldots, J_i]$ of a truck $i \in [1, 2, \ldots, I]$, the sums of $C^h_{i,j_i,r_{i,j_i}}$ and $C^l_{i,j_i,r_{i,j_i}}$ count the number of data points with a speed much higher or lower in a trip. These data points associated with speeds either too high or too low are regarded as potentially risky speeds and will be used to construct the risky metrics.

For each truck i, four metrics of the frequencies of risky speeds are calculated, i.e., $F^h_i(16)$, $F^h_i(8)$, $F^l_i(8)$, and $F^l_i(16)$.

Note these metrics are inclusive of each other, i.e., the calculations of $F_i^h(8)$ and $F_i^l(8)$ include the speed differences between 8 km/h and 16 km/h, and also speed differences greater than 16 km/h (i.e., $F_i^h(16)$ and $F_i^l(16)$), respectively.

These metrics are combined as a frequency score to provide a surrogate measure for risky driving. The frequency score is based on a weighted average of the four metrics whose weights are the severest. In this case, speed variation greater than 16 km/h (i.e., $F_i^h(16)$) is more prone to be risky driving than speed variation between 8 km/h and 16 km/h, i.e., $F_i^h(8-16)$. Similarly, the rank of risky speed variation is assumed as $F_i^h(16) > F_i^h(8) > F_i^l(16) > F_i^l(8)$. Thus, the overall score, for each truck i, is calculated by the weighted frequency of speed variation:

$$Fp_i = \frac{N \sum_n \beta(n) (F_i^h, F_i^l)}{\sum_n \beta(n)} \tag{14}$$

where Fp_i is the driver profile score for truck i. $\beta(n)$ is a weight function of the risky speed variation, $n = 1, 2, 3,$ and 4, thus the total number of $N = 4$. The weights $\beta(1)$ is for $F_i^h(16)$, $\beta(2)$for $F_i^h(8)$, $\beta(3)$for $F_i^l(16)$, and $\beta(4)$ for $F_i^l(8)$ for all trucks respectively. Because the importance or contributions of the four metrics to the risky are different, the frequency score of the speed variations is weighted by the importance (i.e., severity) of each metric that is contributed to the risky behavior. The more severe the risk the higher the weight. Note that this paper proposes a general methodology. The specific values for the function of $\beta(n)$ will be chosen by the user. In this paper, it was assumed that the relative risks followed as β(n = 1,2,3 and 4) of 4, 3, 2, and 1, respectively, according to the relationship of $F_i^h(16) > F_i^h(8) > F_i^l(16) > F_i^l(8)$.

C. FREQUENCY AND AMPLITUDE OF SPEED VARIATION FOR DRIVING PATTERNS

In addition to the frequency, and amplitude of the speed variation is also used to measure the speed variations. The mean and standard deviation of the amplitude is calculated using (15) – (17) for high-speed differences and (18) – (20) for low-speed differences, respectively.

$$Mn_i^h = \frac{1}{J_i} \sum_{j_i=1}^{J_i} \frac{\sum_{r_{i,j_i}=1}^{R_{i,j_i}} S_{i,j_i,r_{i,j_i}}^h}{R_{i,j_i}} \tag{15}$$

$$Sd_i^h = \sqrt{\frac{1}{J_i - 1} \sum_{j_i=1}^{J_i} \left(\frac{\sum_{r_{i,j_i}=1}^{R_{i,j_i}} S_{i,j_i,r_{i,j_i}}^h}{R_{i,j_i}} - Mn_i^h \right)^2} \tag{16}$$

$$S_{i,j_i,r_{i,j_i}}^h = \begin{cases} D_{i,j_i,r_{i,j_i}} - DS_m, & if\ D_{i,j_i,r_{i,j_i}} > DS_m \\ 0, & otherwise \end{cases} \tag{17}$$

$$Mn_i^l = \frac{1}{J_i} \sum_{j_i=1}^{J_i} \frac{\sum_{r_{i,j_i}=1}^{R_{i,j_i}} S_{i,j_i,r_{i,j_i}}^l}{R_{i,j_i}} \tag{18}$$

$$Sd_i^l = \sqrt{\frac{1}{J_i - 1} \sum_{j_i=1}^{J_i} \left(\frac{\sum_{r_{i,j_i}=1}^{R_{i,j_i}} S_{i,j_i,r_{i,j_i}}^l}{R_{i,j_i}} - Mn_i^l \right)^2} \tag{19}$$

$$S_{i,j_i,r_{i,j_i}}^l = \begin{cases} DS_m - D_{i,j_i,r_{i,j_i}}, & if\ D_{i,j_i,r_{i,j_i}} < -DS_m \\ 0, & otherwise \end{cases} \tag{20}$$

where $S_{i,j_i,r_{i,j_i}}^h$ and $S_{i,j_i,r_{i,j_i}}^l$ calculate the speed differences between a data point $D_{i,j_i,r_{i,j_i}}$ and the DS_m ($m = 1$ and 2). For each data points $r_{i,j_i} \in [1, 2, \ldots, R_{i,j_i}]$ in a trip $j_i \in [1, 2, \ldots, J_i]$ of a truck $i \in [1, 2, \ldots, I]$, it sums all the speed variations of data points with a speed much higher or lower in a trip. Then, mean and standard deviation are used to represent the amplitude of this amount of speed variations for each truck. In this way, the extent of the risky speeds is measured in a different perspective.

So far, the frequency (i.e., proportion) and amplitude (i.e., mean and standard deviation) of the speed variation is normalized respectively into the range of 0 to 1 and then averaged in the low-speed score (LSS) and high-speed score (HSS) for each truck i ($i = 1, 2, \ldots, I$):

$$LSS_i = \frac{Fp_i^l + Mn_i^l + Sd_i^l}{3} \tag{21}$$

$$HSS_i = \frac{Fp_i^h + Mn_i^h + Sd_i^h}{3} \tag{22}$$

Based on LSS and HSS, four driving patterns are defined by using speed variations as shown in figure 4. Thus, a truck's on-road driving style can be categorized into one of the four patterns.

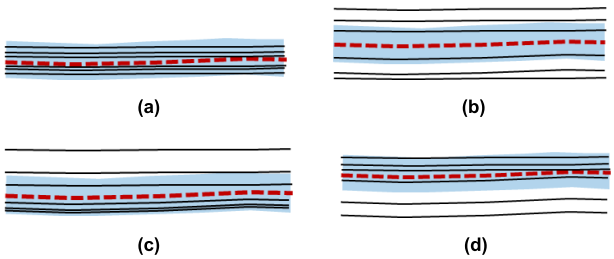


FIGURE 4. Four types of pattern. The solid black line represents a trip; dashed red line represents the MSR of all trips, and the shaded blue area represents DS boundaries. Overall, it indicates speed (vertical axis) over time or distance (horizontal distance).

Pattern 1 refers to both LSS and HSS are small. Pattern 2 refers to both LSS and HSS are large. Pattern 3 refers to small LSS and large HSS. Pattern 4 refers to large LSS and small HSS. Pattern 1 describes trucks that experience more stable traveling condition with more records with minor speed changes. Pattern 2 describes a driver with both large HSS and LSS, which indicates the driver who frequently engages in risky driving (off from the driver's "median" driving style). Similarly, Pattern 3 describes a driver with a small LSS and a large HSS, which indicates the driver who mostly drives safely but occasionally engages in risky driving. Pattern 4 describes trucks that may experience more severe traffic congestion.

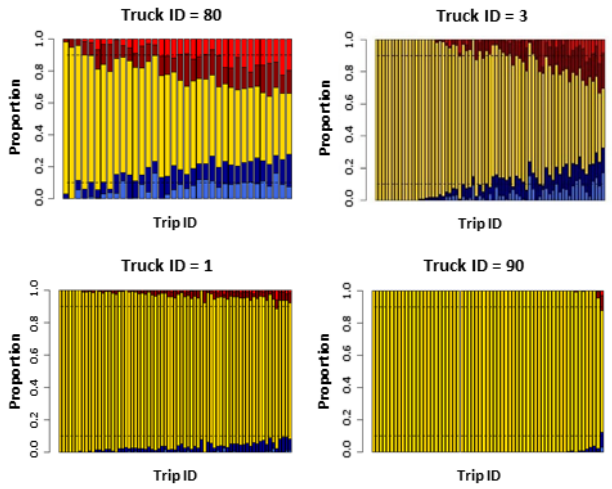


FIGURE 5. Frequency proportions of speed variations for four example trucks. The color code is F_i^h (16) in cardinal, F_i^h (8) in dark red, F_i^l (8) in navy, F_i^l (16) in royal blue. The gold color represents the speed difference within 8 kph. Colors are following R color palette.

V. RESULTS

A. DRIVER PROFILES

All the trucks are scored by the frequency of the speed variations using equations (10) - (13). Figure 5 provides four sample trucks, which shows the frequency proportions (e.g., higher speed portion, lower speed portion, etc.) for each trip. The horizontal axles for the four trucks are the trip ID (numbering is not shown), in which a vertical bar (may contain red, yellow and blue colors) represents a trip generated by the truck, and the total number of trips (i.e., vertical bars) for each truck differs. In a vertical bar, the trip is plotted in a way that accumulates proportions of the frequencies of speed variations, with different colors indicating different levels of speed variations, i.e., $F^h(16)$ in cardinal,$F^h(8)$ in dark red, F^l (16) in royal blue and $F^l(8)$ in navy. The remaining part (the middle part) is in yellow, which essentially represents the proportion of steady speed (i.e., the speed differences are within DS of 8 km/h).

As shown in figure 5, the speed of truck 80 fluctuates greatly with a considerable proportion of speed differences that is beyond $\pm$8 km/h or $\pm$16 km/h in many trips. The speed fluctuation is reduced in truck 3 as more trips contain larger proportions of speed difference within $\pm$8 km/h (i.e., yellow colored area), and Truck 1 increased the proportion of steady speed (indicated by speed difference within the $\pm$8 km/h). Finally, for truck 90, speed differences of almost all trips are within $\pm$8 km/h, indicating a much steadier driving performance among trips.

Potential at-risk drivers would be identified through a surrogate safety measure of driver behavior, which is the frequency of speed variations. This measure assumes that the greater the frequency of speed variations, the more unstable the driving behavior, thus the greater the risk of crashes. The above proportion of frequency in speed variation within different levels ($DS_1 = 8$ km/h and $DS_2 = 16$ km/h) are combined to establish each driver's risk profile, as calculated in equation (14). The calculation results are summarized in table 1. Note that due to the limited space, table 1 only lists the driver profile scores of a few trucks and their rankings among all trucks in the dataset.

TABLE 1. Driver profile score.

Truck ID	$F^l(16)$	$F^l(8)$	$F^h(8)$	$F^h(16)$	Score (Fp)	Ranking
80	0.058	0.110	0.124	0.092	0.386	1
76	0.052	0.090	0.077	0.040	0.213	17
57	0.068	0.108	0.075	0.025	0.190	19
60	0.023	0.076	0.076	0.047	0.211	25
15	0.022	0.072	0.059	0.030	0.152	44
82	0.009	0.037	0.034	0.037	0.128	70
85	0.069	0.077	0.018	0.000	0.069	77
30	0.003	0.020	0.029	0.012	0.062	92
1	0.001	0.022	0.020	0.004	0.034	94
90	0.000	0.003	0.001	0.001	0.005	98

In table 1, the overall score Fp indicates the total speed variations that are out of the predefined DS. For example, there is 38.6% of the time that the speed differences of truck 80 are beyond $\pm$8 km/h (including $\pm$16 km/h), and this amount of the time is 0.5% for truck 90. A higher Fp value represents a higher potential risk associated with a specific truck, which has a higher ranking (the highest ranking is 1, and the lowest ranking is 100).

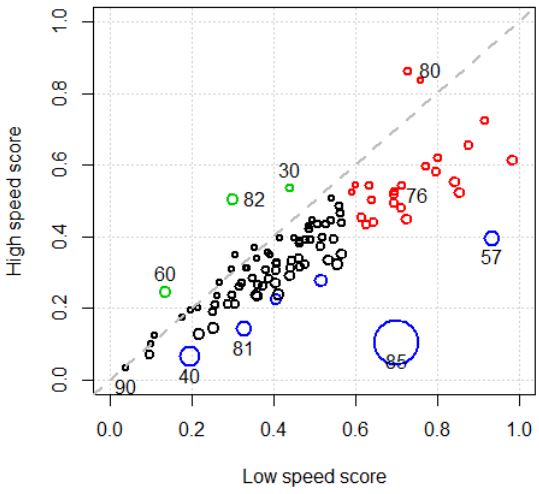


FIGURE 6. Four driving patterns classified by the relationship between HSS and LSS.

B. DRIVING PATTERN

As shown in figure 6, each circle represents a truck with its HSS and LSS as calculated in equations (21) and (22). Some trucks are marked with their identification numbers (i.e., Truck ID). As defined previously, a stable driving pattern (i.e., pattern 1) would be those associated with both small HSS and LSS, as shown in the black circle in figure 6.

TABLE 2. Examples of driving patterns.

Truck ID	F^l	F^h	Mn^l	Mn^h	Sd^l	Sd^h	LSS	HSS	Pattern
90	0.018	0.012	-0.01	0.01	0.08	0.08	0.04	0.03	1
1	0.126	0.110	-0.08	0.08	0.09	0.11	0.10	0.10	1
15	0.505	0.410	-0.40	0.34	0.56	0.41	0.49	0.39	1
76	0.764	0.540	-0.72	0.51	0.61	0.49	0.70	0.52	2
80	0.904	1.000	-0.78	1.00	0.50	0.59	0.73	0.86	2
82	0.252	0.326	-0.23	0.48	0.42	0.71	0.30	0.50	3
30	0.128	0.186	-0.09	0.17	0.19	0.38	0.14	0.52	3
85	0.787	0.085	-0.76	0.05	0.55	0.17	0.70	0.10	4
57	0.948	0.462	-0.92	0.38	0.93	0.35	0.93	0.40	4

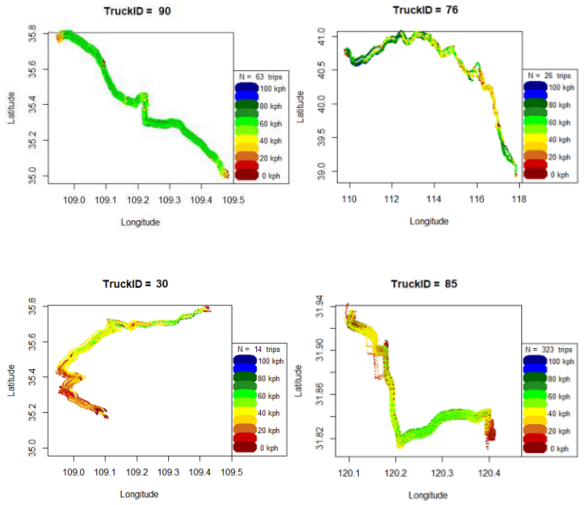


FIGURE 7. Trips with speed display in four sample trucks. In order to avoid a complete overlap, a small random value is added to the geospatial coordinates of each trip. Color represents the level of speed as shown in the legend.

Besides, pattern 2 is in red, pattern 3 in green, and pattern 4 in blue.

In figure 6, the size of the points indicates how big of the difference between the HSS and the LSS values, which can be expressed in HSS/LSS or LSS/HSS (physically it is the distance to the gray dashed diagonal). For example, the location of truck 30 in figure 6 is calculated with an HSS value of 0.52 and an LSS value of 0.14, thus the size of the circle (i.e., radius) is HSS/LSS = 0.52/0.14 = 3.71. Therefore truck 30 is determined as pattern 3 (in green color).

Due to the limited space, table 2 only listed the LSS and HSS scores of a few trucks and their patterns among all trucks in the dataset.

Pattern 1 is preferable as its speed is tightly distributed, indicating the driving performances of all trips are very close to each other. Using speed variation as a surrogate index to the risky driving, patterns 2, 3, and 4 are all defined as potential risky driving. The larger the HSS, the riskier driving it would be. Therefore, pattern 2 is comparatively riskier than pattern 3, and pattern 3 is comparatively riskier than pattern 4.

Four sample trucks (truck IDs of 90, 76, 30, and 85) are selected as representatives of the four driving patterns 1, 2, 3 and 4, respectively, and are plotted in figure 7 to illustrate their on-road speed variations.

As shown in the legends in figure 7, the numbers of trips for truck ID 90, 76, 30, and 85 are 63, 26, 14 and 323, respectively. These trips are used to calculate speed variations for each truck and then compared among trucks. As can be seen in figure 7, the speed of truck 90 is very uniform, which indicates a "tight" speed variation. The speed range of truck 76 is wide, with a high speed 90 km/h and a low speed of 40 km/h on different trips. Truck 30 goes with relatively considerably low speed at some locations. One of the possibilities is that truck 30 may have experienced recurrent traffic jams at those road segments. For truck 85, most of the trips keep speed in the range between 40 km/h and 50 km/h. However, several trips are below 20 km/h, which could be caused by a temporary traffic jam (i.e., accidents).

In summary, 100 trucks are all featured in the four driving patterns. The statistical summary of LSS and HSS for each pattern is listed in table 3.

TABLE 3. Statistical summary of the four driving patterns.

		Pattern 1	Pattern 2	Pattern 3	Pattern 4
LSS	mean (std.)	0.41 (0.166)	0.72 (0.089)	0.22 (0.116)	0.52 (0.268)
HSS	mean (std.)	0.33 (0.117)	0.58 (0.125)	0.51 (0.183)	0.20 (0.123)
Number of trucks		73	18	3	6

VI. CONCLUSIONS AND LIMITATIONS

A. CONCLUSIONS AND DISCUSSIONS

The main goal of the paper is to study the driver profiles and driving patterns to reflect the risk driver behavior. For driver profile at an individual level, drivers (assuming each truck was operated by one driver throughout the data collection period) are scored regarding their risky driving, where the frequency of speed variations is measured. For driving pattern at an overall level, drivers are categorized into four patterns. Three patterns (i.e., patterns 2, 3 and 4) are regarded as

risky driving, which frequency and amplitude of the speed variations are taken into consideration. The method is tested on a GPS dataset with 100 trucks under dedicated routes. The results provide a glimpse of driver behavior from a perspective of speed variations.

However, it should be acknowledged that the method using speed variation as a metric of risk driver behavior [28], [29] is rather unitary. In fact, many variables may be related to risky driving [30]. For example, previous studies have found that acceleration is an important variable in evaluating risk driver behavior. However, the acceleration rate calculated from consecutive speeds with a time interval of 30 seconds in this paper may wipe off many behavior details, in which case the acceleration reflected driving behavior may not be reliable. That is why acceleration was not included in this study. It is suggested to take more risk related features, such as acceleration, braking, yaw rate and so on, into consideration in the future when the data is richer and the quality is improved (e.g., second-by-second data).

The speed limit is also an important feature as it can be a universal reference to the risky driving speed. Given the speed limit, travel speed can be easily categorized as speeding or not speeding. Henceforth, driving patterns can be grouped roughly into aggressive driving (speeding a lot and frequent speed changes) and conservative driving (within the speed limit and fewer speed changes). This paper proposed a suboptimal method that in the case speed limit information is missing, which uses the median as a speed reference (i.e., MSR) as it measures a tendency of ''variation'' including travel speed higher and lower than a predefined threshold. It is hypothesized that travel speeds both too high and too low are not safe driving behaviors.

The method in this paper can be used to explore the aggressive driver behavior, fuel efficiency, performance of autonomous vehicle with trajectory tracking [31] at-risk indexing, etc. Determining driving patterns is important in the field of driving safety, in which it can be used to assist in holistic sensing for intelligent driver assistance systems. The usefulness of study driving pattern in this paper is also for testing and screening purposes and especially for large-scale monitoring of professional truck drivers [16]. Evaluation score of individual truck or fleet will encourage drivers to improve their driving habits, reducing the risk of accidents and tickets. The study may attract interest from fleet companies for monitoring fleet of commercial vehicles [10], or insurance companies for assessing the insured drivers [11], in that safe and reliable on-road driving performances are priority traits that drivers are expected. What is more, a difference in driver behavior may be found for those who seek potential risky drivers of re-training and education programs.

B. LIMITATIONS AND FUTURE WORK

The study benefits from using GPS trajectory data of a dedicated route as it provides repeated experimental trips of a truck. This way, the speed variation is measured in spatial rather than temporal terms. However, without considering the temporal variation, speed variation as a risky driving indicator may be biased, because the assumption here is that the influence of external traffic is constant over time.

As on-road driving is a combination of vehicle operation, driver characteristics, and traffic environment, this paper, however, mainly and merely focuses on the output of vehicle operation and tries to use the data from vehicle operation (i.e., speed variation) to reflect driver behaviors. Admittedly, the lack of information on traffic conditions is critical for driver behavior studies. For example, if a driver exhibits abrupt accelerating or braking behavior, it is not easy to discern whether it is because the driver behaves aggressively or because road traffic impacts the driver. What is more, this paper did not consider individual driver characteristics (e.g., demographics, personality), which may lead to bias as it fails to connect driver behavior to those endogenous causes.

The information above should be taken into account in the future with the richness of more comprehensive data and the maturity of more advanced feature extraction technology. By doing so, the next step is to generalize the applications of the driver profile and driving pattern in assessing risky behaviors to various traffic conditions. The reliability of the risky driving behavior measures is also one of the directions of future research.

ACKNOWLEDGMENT

The authors would like to thank Beijing Sinoiov Iov Technology Company for providing the data. The contents reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein and are not necessarily representative of the sponsoring agencies.

REFERENCES

[1] D. P. Wagner, ''Lexington area travel data collection test: GPS for personal travel surveys,'' Office Highway Policy Inf. Office Technol. Appl., Federal Highway Admin., Battelle Transp. Division, Washington, DC, USA, Final Rep. 00811205, 1997.

[2] L. Yalamanchili, R. M. Pendyala, N. Prabaharan, and P. Chakravarty, ''Analysis of global positioning system-based data collection methods for capturing multistop trip-chaining behavior,'' *Transp. Res. Rec., J. Transp. Res. Board*, vol. 1660, pp. 58–65, 1999.

[3] A. B. Ellison, S. P. Greaves, and R. Daniels, ''Profiling drivers' risky behaviour towards all road users,'' in *Proc. Australas. College Road Saf. Conf.*, Sydney, NSW, Australia, Aug. 2012, pp. 9–10.

[4] T. P. Hutchinson and L. N. Wundersitz, ''Road safety mass media campaigns: Why are results inconclusive, and what can be done?'' *Int. J. Injury Control Saf. Promotion*, vol. 18, no. 3, pp. 235–241, Sep. 2011.

[5] E. Petridou and M. Moustaki, ''Human factors in the causation of road traffic crashes,'' *Eur. J. Epidemiol.*, vol. 16, no. 9, pp. 819–826, Sep. 2000.

[6] J. Jun, R. Guensler, and J. Ogle, ''Differences in observed speed patterns between crash-involved and crash-not-involved drivers: Application of in-vehicle monitoring technology,'' *Transp. Res. C, Emerg. Technol.*, vol. 19, no. 4, pp. 569–578, Aug. 2011.

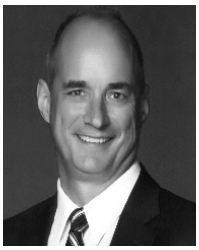
[7] T. Lotan and T. Toledo, ''In-vehicle data recorder for evaluation of driving behavior and safety,'' *Transp. Res. Rec., J. Transp. Res. Board*, vol. 1953, pp. 112–119, Jan. 2006.

[8] S. G. Klauer, T. A. Dingus, V. L. Neale, J. D. Sudweeks, and D. J. Ramsey, ''Comparing real-world behaviors of drivers with high versus low rates of crashes and near-crashes,'' Nat. Highway Traffic Saf. Admin., Washington, DC, USA, Tech. Rep. DOT-HS-811-091, 2009.

[9] T. A. Dingus *et al.*, "The 100-car naturalistic driving study, phase II—Results of the 100-car field experiment," Dept. Transp., Nat. Highway Traffic Saf. Admin., Washington, DC, USA, Tech. Rep. FHWA-JPO-06-056, 2006.
[10] T. Toledo, O. Musicant, and T. Lotan, "In-vehicle data recorders for monitoring and feedback on drivers' behavior," *Transp. Res. C, Emerg. Technol.*, vol. 16, no. 3, pp. 320–331, Jun. 2008.
[11] A. B. Ellison, S. P. Greaves, and M. C. J. Bliemer, "Driver behaviour profiles for road safety analysis," *Accident Anal. Prevention*, vol. 76, pp. 118–132, Mar. 2015.
[12] A. Laureshyn, K. Åström, and K. Brundell-Freij, "From speed profile data to analysis of behaviour: Classification by pattern recognition techniques," *IATSS Res.*, vol. 33, no. 2, pp. 88–98, 2009.
[13] Y. Wang, K. Qin, Y. Chen, and P. Zhao, "Detecting anomalous trajectories and behavior patterns using hierarchical clustering from taxi GPS data," *ISPRS Int. J. Geo-Inf.*, vol. 7, no. 1, 2018.
[14] M. Brambilla, P. Mascetti, and A. Mauri, "Comparison of different driving style analysis approaches based on trip segmentation over GPS information," in *Proc. IEEE Int. Conf. Big Data*, Dec. 2017, pp. 3784–3791.
[15] X. Zhu, X. Hu, and Y.-C. Chiu, "Design of driving behavior pattern measurements using smartphone Global Positioning System data," *Int. J. Transp. Sci. Technol.*, vol. 2, no. 4, pp. 269–288, Dec. 2013.
[16] A. E. A. Wåhlberg, "Driver celeration behaviour and accidents—An analysis," *Theor. Issues Ergonom. Sci.*, vol. 9, no. 5, pp. 383–403, Sep. 2008.
[17] S. Boonsiripant, "Speed profile variation as a surrogate measure of road safety based on GPS-equipped vehicle data," Ph.D. dissertation, Georgia Inst. Technol., Atlanta, GA, USA, 2009.
[18] J. Grengs, X. Wang, and L. Kostyniuk, "Using GPS data to understand driving behavior," *J. Urban Technol.*, vol. 15, no. 2, pp. 33–53, 2008.
[19] J. Jun, R. Guensler, and J. H. Ogle, "Smoothing methods to minimize impact of global positioning system random error on travel distance, speed, and acceleration profile estimates," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 1972, no. 1, pp. 141–150, 2006.
[20] Z. Xu, T. Wei, S. Easa, X. Zhao, and X. Qu, "Modeling relationship between truck fuel consumption and driving behavior using data from Internet of vehicle," *Comput.-Aided Civil Infrastruct. Eng.*, vol. 33, no. 3, pp. 209–219, Mar. 2018.
[21] W. Bohte and K. Maat, "Deriving and validating trip purposes and travel modes for multi-day GPS-based travel surveys: A large-scale application in The Netherlands," *Transp. Res. C, Emerg. Technol.*, vol. 17, no. 3, pp. 285–297, 2009.
[22] N. Schuessler and K. W. Axhausen, "Identifying trips and activities and their characteristics from GPS raw data without further information," presented at the 8th Int. Conf. Survey Methods Transp., vol. 502, Annecy, France, May 2008.
[23] H. Xu, H. Liu, C.-W. Tan, and Y. Bao, "Development and application of an enhanced Kalman Filter and global positioning system error-correction approach for improved map-matching," *J. Intell. Transp. Syst.*, vol. 14, no. 1, pp. 27–36, Feb. 2010.
[24] P. Stopher, C. FitzGerald, and J. Zhang, "Search for a global positioning system device to measure person travel," *Transp. Res. C, Emerg. Technol.*, vol. 16, no. 3, pp. 350–369, 2008.
[25] J. Wolf, S. Schönfelder, U. Samaga, M. Oliveira, and K. W. Axhausen, "Eighty weeks of global positioning system traces: Approaches to enriching trip information," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 1870, pp. 46–54, 2004.
[26] J. He, Y. Zhang, G. Huang, and P. De Souza, "CIRCE: Correcting imprecise readings and compressing excrescent points for querying common patterns in uncertain sensor streams," *Inf. Syst.*, vol. 38, no. 8, pp. 1234–1251, Nov. 2013.
[27] L. Zhu, J. R. Holden, and J. D. Gonder, "Trajectory segmentation map-matching approach for large-scale, high-resolution GPS data," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2645, no. 1, pp. 67–75. 2017.
[28] L. Aarts and I. van Schagen, "Driving speed and the risk of road crashes: A review," *Accident Anal. Prevention*, vol. 38, no. 2, pp. 215–224, 2006.
[29] N. J. Garber and R. Gadiraju, "Factors affecting speed variance and its influence on accidents," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 1213, pp. 64–71, 1989.
[30] C. Ma, W. Hao, X. Wang, and W. Yan, "The impact of aggressive driving behavior on driver-injury severity at highway-rail grade crossings accidents," *J. Adv. Transp.*, vol. 2018, 2018, Art. no. 9841498.
[31] Z. Xu, M. Wang, F. Zhang, S. Jin, J. Zhang, and X. Zhao, "PaTAVTT: A hardware-in-the-loop scaled platform for testing autonomous vehicle trajectory tracking," *J. Adv. Transp.*, vol. 2017, 2017, Art. no. 9203251.

YING LI received the M.Sc. degree in transport studies from Imperial College London, U.K., in 2011, and the Ph.D. degree in transport from University College London, U.K., in 2016. She is currently a Lecturer with the School of Information Engineering, Chang'an University, China. Her research interest includes robust traffic modeling and optimization.

LI ZHAO has been a Research Associate with the Department of Civil Engineering, University of Nebraska–Lincoln, USA, since 2017. Her research interests include computational intelligence and intelligent transportation, statistical analysis, and data mining.



LAURENCE R. RILETT is a Distinguished Professor of civil engineering with the University of Nebraska–Lincoln, USA. He also serves as the Director of the UNL Mid-America Transportation Center. His research interests include multimodal transportation systems analysis, intelligent transportation systems, transportation planning and operations, and dynamic network modeling and optimization.