Accident Analysis and Prevention 132 (2019) 105256

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

# Analysis of commercial truck drivers' potentially dangerous driving behaviors based on 11-month digital tachograph data and multilevel modeling approach

Tuqiang Zhou[a,b], Junyi Zhang[b,*]

[a] The College of Transportation and Logistics, East China Jiaotong University, Nanchang, 330013, China
[b] Mobilities and Urban Policy Lab, Graduate School for International Development and Cooperation, Hiroshima University, Higashi Hiroshima, 739-8529, Japan

ARTICLE INFO

*Keywords:*
Truck driver
Digital tachograph data
Driving behavior
PCA
DBSCAN
Multilevel modeling

ABSTRACT

This study analyzed the potentially dangerous driving behaviors of commercial truck drivers from both macro and micro perspectives. The analysis was based on digital tachograph data collected over an 11-month period and comprising 4373 trips made by 70 truck drivers. First, different types of truck drivers were identified using principal component analysis (PCA) and a density-based spatial clustering of applications with noise (DBSCAN) at the macro level. Then, a multilevel model was built to extract the variation properties of speeding behavior at the micro level. Results showed that 40% of the truck drivers tended to drive in a substantially dangerous way and the explained variance proportion of potentially extremely dangerous truck drivers (79.76%) was distinctly higher than that of other types of truck drivers (14.70%~34.17%). This paper presents a systematic approach to extracting and examining information from a big data source of digital tachograph data. The derived findings make valuable contributions to the development of safety education programs, regulations, and proactive road safety countermeasures and management.

## 1. Introduction

Traffic safety is a critical issue in the transportation field. Traffic safety conditions are determined by drivers, vehicles, and driving environment. Previous research revealed that over 90% of traffic accidents were associated with unsafe driving behaviors (e.g., Petridou and Moustaki, 2001; Ellison et al., 2015; Atombo et al., 2016). Thus, an enhanced understanding of unsafe driving behaviors could provide meaningful contributions to road safety research.

Driving behavior plays an important role in driving risk analysis. However, it is difficult to measure risk in real-life situations (Eboli et al., 2017). Driving simulators are often used to investigate driving behaviors in various experimental environments (Pankok and Kaber, 2018). Some vehicle instrument technologies such as the Naturalistic Driving Study (NDS) (Guo and Hankey, 2009) and the DriveCam system (Hickman et al., 2010) have been applied to monitor driving behaviors and kinematic signatures on a large scale. The NDS programs such as the Second Strategic Highway Research Program (SHRP2), the 100-Car NDS, and Europe's UDRIVE have provided valuable insights into accident causation and driving behaviors (Guo, 2019). Most existing analyses of dangerous driving behavior have relied on crash data or self-reported questionnaire surveys (Lord and Mannering, 2010; Ellison et al., 2015; Mannering and Bhat, 2014; Huang et al., 2017). To fully explore driving behavior in traffic accidents, it is important to understand driving styles. Driving styles, or habits, are defined as the way that individuals choose to drive and are analyzed over periods of years (Constantinescu et al., 2010). Traditional attributes of driving style include choices of driving speed, acceleration, deceleration or braking, threshold for overtaking, headway, and propensity to commit traffic violations (Murphey et al., 2009; Wu et al., 2016).

Among all types of vehicles, trucks are the largest contributor to traffic accidents, injuries, and fatalities owing to their high proportion among the roadway population, as well as their size, weight, and other unique characteristics (Zhu and Srinivasan, 2011; He et al., 2019). Compared to other vehicle types, the larger size and higher center of gravity of trucks result in longer braking distances and more severe consequences when involved in accidents. Moreover, truck crashes have higher economic impacts because of the damage to high-value cargo and travel delays caused by traffic accidents. Thus, research identifying influential factors in truck accidents would facilitate the development of countermeasures that could reduce the number and severity of accidents involving trucks.

* Corresponding author.
*E-mail addresses:* zhoutuqiang@126.com (T. Zhou), zjy@hiroshima-u.ac.jp (J. Zhang).

https://doi.org/10.1016/j.aap.2019.105256
Received 22 March 2019; Received in revised form 13 July 2019; Accepted 30 July 2019
Available online 20 August 2019

Prior research has demonstrated that driving environment has a substantial influence on driver safety awareness (e.g., Kaber et al., 2012; Faure et al., 2016; Yan et al., 2017). Driving is a complex task that involves maintaining appropriate steering and speeds while accurately perceiving, identifying, and anticipating road elements such as road type or transit route, and other dynamic conditions including traffic flow, car following situation, and weather. Given these parameters, it is crucial to advance our understanding of how driving environment affects driving behavior.

The rapid rise and prevalence of mobile technologies have enabled the collection of a massive amount of passive data, e.g., big data. Effective analysis of big data provides new opportunities to advance our understanding of critical transportation problems such as road safety (Chen et al., 2016; Bao et al., 2019). Unlike small-scale data obtained via questionnaires or surveys, most big data are initially generated for other purposes, but have high potential value for research applications. Currently, this massive amount of second-by-second data have yet to be fully utilized in road safety research.

Approached with a microcosmic perspective, most road safety data are hierarchically organized to facilitate road safety research (Dupont et al., 2013). This makes multilevel models, or hierarchical linear models, technologically effective and efficient for figuring out the heterogeneity among the complicated data structures. Billot et al. (2009) investigated the influence of rain on driving behavior using a multilevel model. Huang and Abdel-Aty (2010) proposed a 5 × ST-level (S: spatial; T: temporal) hierarchy to represent the general framework of multilevel data structures in traffic safety. This is a conceptual framework for assessing structured safety data. Several case studies using Bayesian hierarchical models were summarized to improve model fitting and predictive performance over traditional models. However, all the data utilized were sourced from crash data, either frequency or severity. The authors also noted concerns about the applicability and transferability of these multilevel, or hierarchical models.

In summary, the following research gaps were identified in existing research. First, because crash data is traditionally the primary or only data source, proactive approaches to effective traffic safety measures have been ignored. Second, the influence of unobserved heterogeneities omitted by multilevel or hierarchical models on traffic safety remain unknown. Third, in Japan, installing tachograph recorders has been mandatory for trucks since 1967. Prior to then, only analogue-type recorders were available[1] . Since the 2010s, the use of digital tachograph recorders has increased and effectively replaced analogue-type recorders. However, the accumulated big data have seldom been applied to road safety research in Japan.

To fill the abovementioned research gaps, the objectives of this study are twofold. The first objective is to identify different types of truck drivers in terms of driving performance at the macro level. The second objective is to excavate the variation properties of speeding behavior across different types of truck drivers by building a multilevel model at the micro level, where the effects of multi-dimensional unobserved heterogeneities are reflected. This study used digital tachograph data collected from a major Japanese freight transport company over an 11-month period in the year 2014. Driving performance indicators included speeding, driving duration, and jerky driving indicators, which were measured using the tachograph data and further exploited to indirectly capture potentially dangerous driving behaviors. This study differs from existing studies by simultaneously considering the following aspects:

(1) Adopting a real-world commercial truck-driving data and capturing potentially dangerous driving behaviors based on multiple indicators;
(2) Using a large-scale dataset that recorded truck drivers' driving speed-related data at an interval of 0.5 s in tandem with GPS data collected at an interval of 60 s;
(3) Building a multilevel model that reflects the influences of multi-dimensional unobserved heterogeneities;
(4) Focusing on using the above large-scale driving data with detailed temporal and spatial information in the specific context of Japan.

## 2. Literature review

Many existing traffic safety studies emphasize the use of crash data to create accident indices based on metric such as crash and near-crash rates. However, such measurements neglect the importance of proactive safety countermeasures. Data are traditionally obtained from police crash reports, which include information such as date and time, age and gender of vehicle occupants, weather, road conditions, type of accident, and safety belt usage (Mannering and Bhat, 2014).

### 2.1. Naturalistic driving study (NDS)

The increasing popularity of large-scale NDSs in the past decade has filled many of the gaps between crash databases and driving behavior (Ye et al., 2017; Pantangi et al., 2019). Findings from various NDSs have made substantial contributions to the development of public policy and the improvement of vehicle safety and driver education (Guo, 2019).

Using vehicle-mounted video cameras, kinematic sensors and other recording devices, NDSs provide researchers with an in-depth and ecologically valid perspective on real-world driving. However, as stated by Pantangi et al. (2019), numerous trip-, driver-, weather-, and vehicle-specific characteristics were found to exert substantial influence on driving behaviors. Most existing NDSs have attempted to investigate one or more of the above characteristics. However, a comprehensive perspective that considers all the relationships among driving behaviors and the various factors has not yet been attempted.

### 2.2. Studies on dangerous driving behaviors

Daily driving data can provide valuable insights into road safety management. Information extracted from individual drivers can be used to identify different driving types, which are crucial to creating comprehensive road safety measures tailored to the driver population. Insurance industries could also improve their underwriting and pricing strategies based on the ability to classify drivers by risk level.

Existing studies have detected various dangerous driving behaviors. Hassan et al. (2017) identified factors that affect driver speed behaviors using crash data and self-reported questionnaire data. Radun et al. (2013) and Filtness et al. (2017) investigated fatigue-related driving behaviors. Murphey et al. (2009) classified driving styles using jerk analysis, primarily focusing on the frequency of drivers' acceleration and deceleration. However, there has not been any comprehensive analysis of various driving styles in the larger context of traffic safety issues. That is, speed, jerk, and fatigue-related driving behaviors have seldom been considered simultaneously.

### 2.3. Studies on factors affecting traffic crashes

Traffic crashes usually result from multiple concurrent factors. Prior studies have demonstrated that driving environment directly influences driver performance. Environmental factors include weather conditions (Brooks et al., 2011; Mueller and Trick, 2012; Peng et al., 2017), route selection (Key et al., 2017), urban versus rural regions (Islam et al., 2014), and road types (Malin et al., 2019). Temporal factors are also play a critical role in determining driving performance. Time of day, circadian rhythm, and continuous driving duration are often associated with drowsiness-related crashes (Radun et al., 2013; Filtness et al.,

[1] https://www.digitacho-efficient.net/basis/obligation.html [in Japanese; Accessed June 6, 2019]

2017; Zhang et al., 2019). Hu et al. (2013) built a dynamic time-series model in a Bayesian framework to identify temporal patterns in highway crashes. McDonald et al. (2018) developed a contextual and temporal algorithm for detecting drowsiness-related driving behaviors. One study simultaneously examined temporal and environmental influences by accident type (El-Basyouny et al., 2014). However, environmental factors were limited to weather conditions (Ahmed and Ghasemzadeh, 2018; Pantangi et al., 2019). Thus, the influences of many important environmental and temporal factors remain unknown in current traffic safety research.

### 2.4. Studies based on large-scale data

The recent availability of large-scale datasets associated with human activities has resulted in a surge of studies on human mobility (Bao et al., 2017). Multi-source big data can improve traffic flow prediction and estimated travel demand. Zhang et al. (2017) generated a taxi-passenger-demand model using roving taxicabs as real-time mobile sensors. To overcome the limitations of traditional disaggregated approaches, Zhao et al. (2018) proposed three traffic demand forecasting methods based on sources of big data. Li et al. (2018) used traditional statistics and geographic big data to determine whether a polycentric urban form could improve commuting efficiency in China. Bao et al. (2019) developed a spatiotemporal deep learning approach for citywide short-term traffic accident risk prediction. Information extracted from large-scale GPS data can serve as effective surrogate measures of traffic exposure and thus enable more effective policymaking. However, few studies have adopted this advanced technology for traffic safety analysis.

### 2.5. Methodologies

From a methodological perspective, early studies mainly rely on Poisson regression approaches (Ye et al., 2013; Li et al., 2013) to examine factors affecting accident frequency. Recently, negative binomial models (Vangala et al., 2015; Park et al., 2016; Hou et al., 2018; Rusli et al., 2018) and Bayesian models (Yu and Abdel-Aty, 2013; Shi et al., 2016) have been widely applied to manage over-dispersed data. Other advanced models such as multivariate models (Anastasopoulos et al., 2012), neural network models (Chong et al., 2013; Zeng et al., 2016), and random-effect models (Naznin et al., 2016; Hou et al., 2018) have provided additional insights on crash-frequency data.

One of the main problems of traditional statistical models is sample dependence among the observations. When samples are drawn from multiple observations and then pooled to be used in aggregate for model estimation, the sample independence required by traditional statistical models cannot be met. As a result, model estimation results are vulnerable to substantial bias due to ignorance regarding unobserved sample heterogeneities (Dupont et al., 2013). In such cases, multilevel modeling approaches are more effective because they do capture the unobserved heterogeneities.

### 2.6. Features of this study

Based on the preceding literature review, this study can be distinguished from existing studies by its use of real-world truck-driving data, its consideration of multiple factors affecting driving behavior, and its use of a large-scale dataset, 11 months of digital tachograph data that produced detailed records individual truck drivers' daily driving performance.

From a methodological perspective, this study attempted to adopt a multilevel modeling approach to investigate potentially dangerous driving behaviors. Understanding potentially dangerous driving behaviors is of great value in the creation of effective proactive safety measures.

## 3. Methodology

Numerous indicators represent various aspects of dangerous driving behaviors. To simplify the interpretation of different dangerous driving behaviors among truck drivers, this study first conducts a Principal Component Analysis (PCA). The PCA is used to derive major independent components representing the various dangerous driving behaviors. Then, a cluster analysis is executed based on the aforementioned truck driver classification components, because it is expected that truckers with different types of dangerous driving behaviors may be affected by various factors in different ways. In other words, the above two analyses use truckers as an analysis unit. On the other hand, the digital tachograph data used in this study traced individual truckers over 11 months, i.e., it is a panel data. During different trips, a driver may behave differently in terms of driving safety. To capture factors affecting dangerous driving behaviors, a multilevel model is developed and implemented at the trip level to identify potential correlations among samples because of duplicated drivers and temporal and spatial elements.

### 3.1. Principal component analysis (PCA)

Using a classic statistical approach, PCA compresses data into a group of new orthogonal variables with lower dimension by computing a linear projection of the data. The new orthogonal principal components are independent, improving the efficiency of statistical techniques, and thus simplifying the interpretation of information from original datasets (Garcia et al., 2019).

Using PCA, all the responses from the data are converted into principal components that are a linear combination of the original multi-responses. The first principal component accounts for as many of the variations in the data as possible. Then, each succeeding component accounts for as many of the remaining variations as possible. According to the Organization for Economic Cooperation and Development (OECD, 2008), PCA should be used to study the overall structure of a dataset, to assess suitability and guide certain methodological choices in the construction of a composite indicator.

Assume there are $n$ samples and $p$ factors where $p < n$, and $X = (x_1, x_2, \ldots, x_p)$ is the original data matrix with covariance matrix $\Sigma$. If the eigenvalue of the covariance matrix is $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$, then variances of new components are: $\mathrm{Var}(F_1) \geq \mathrm{Var}(F_2) \geq \ldots \geq \mathrm{Var}(F_n) \geq 0$. The corresponding unit eigenvectors are $l_1, l_2, \ldots, l_p$. Thus, the principal component $i$ of X is $Z_i = l_i' X$ ($i = 1, 2, \ldots p$).

The preceding statistical concepts may be defined as follows:

- Eigenvalue: The influence indicator of the principal component which represents how much original information can be explained by introducing this principal component. If the eigenvalue is less than one, then the explanation effect of this principal component is weaker than an original factor. Thus, introducing the component is meaningless.
- Variance contribution rate (VCR): The VCR of component $Z_i$ represents the variance proportion of $Z_i$ in total sample variance. The larger the VCR, the more original information $Z_i$ represents.
- Cumulative contribution rate (CCR): Cumulative VCR of the first several principle components.

### 3.2. Cluster analysis

Cluster analysis is an unsupervised learning technique that divides a set of physical or abstract objects into several similar clusters to gain global data figures or conduct further analysis for specific clusters (Liu et al., 2016). A cluster generated by clustering is a set of data objects that share greater similarity compared to other objects in the original set which are likewise clustered based on similarity. Similarity is determined by attribute values of the research objects. Relative distance is

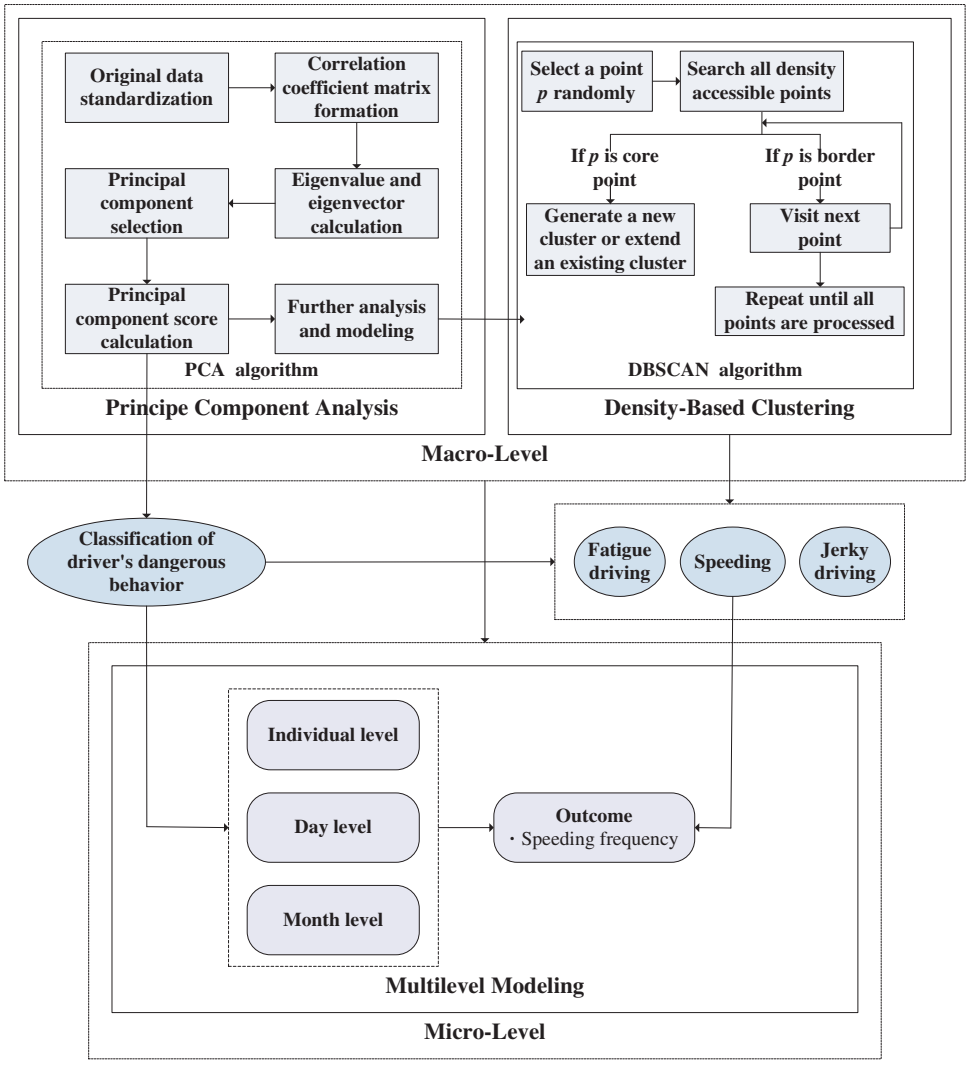


**Fig. 1.** Research procedure.

a commonly used measure.

In this research, a density-based spatial clustering of applications with noise (DBSCAN) is selected for its feasibility and efficiency (Kumar and Reddy, 2016). This method estimates density by counting the number of points in a fixed-radius neighborhood. Two points are connected if they lie within each other's neighborhoods. To identify a cluster, DBSCAN begins with an arbitrary point $p$ and finds all points in the dataset that are density-reachable from $p$. If $p$ is a core point, a cluster is formed. If $p$ is a border point of some cluster, i.e., there are no points density-reachable from $p$, then DBSCAN applies the same procedure to the next unclassified point. The DBSCAN algorithm finishes when all points in the dataset have been assigned to a cluster or identified as noise (Kazemi-Beydokhti et al., 2017).

The main concepts on which DBSCAN is based are density accessibility and density connectivity. Both concepts depend on the two input parameters of neighbor $\varepsilon$ and cluster minimum point $m$. Whether two neighbor points $p_1$ and $p_2$ belong to a cluster is determined by density accessibility. If, (1) two points are close enough, i.e., distance ($p_1$, $p_2$) $< \varepsilon$, and, (2) enough points are neighbors of $p_2$, where distance ($r$, $p_2$) $> m$, and $r$ is database point, then $p_1$ and $p_2$ are density accessible. Density connectivity refers to the possibility of connecting additional points.

### 3.3. Multilevel model

The main problem associated with hierarchical data organization is that observations from the same geographical units are not independent (Huang and Abdel-Aty, 2010) and thus, multilevel models are necessary. In this study, speeding frequency is specifically set as a dependent variable. Data are split into three levels: individual level, day level, and month level. First, a *Null* model with no explanation variables is estimated to investigate the variance proportions in the total variance at different levels, as shown below:

$$y^j_{idm} = \beta_0 + \gamma_i + \gamma_d + \gamma_m + e^j_{idm}, \quad (1)$$

where, $y^j_{idm}$ is a dependent variable (speeding frequency) of trip $j$ made by driver $i$ on day $d$ of month $m$, $\beta_0$ is a constant term, $\gamma_i$, $\gamma_d$, and $\gamma_m$ are random effects at individual level, day level and month level, respectively, and $e^j_{idm}$ is the error term.

Second, a three-level model with explanatory variables (called *Full* model) is estimated as follows:

$$y^j_{idm} = \beta_0 + \beta_1 x^j_{1idm} + \beta_2 x^j_{2idm} + \cdots \beta_n x^j_{nidm} + \gamma_i + \gamma_d + \gamma_m + e^j_{idm}, \quad (2)$$

where, $x^j_{1idm}$, $x^j_{2idm}$, $\cdots$, $x^j_{nidm}$ are explanatory variables, and $\beta_1$, $\beta_2$, $\cdots$, $\beta_n$ are coefficients of the explanatory variables.

In the above two models, the random effects are assumed to be normally distributed as follows:

$\gamma_i^j \backsim N(0,\ \sigma_i^2)$, $\gamma_d^j \backsim N(0,\ \sigma_d^2)$, $\gamma_m^j \backsim N(0,\ \sigma_m^2)$, $e_{idm} \backsim N(0,\ \sigma_0^2)$,

where $\sigma_i^2$, $\sigma_d^2$, $\sigma_m^2$, and $\sigma_0^2$ are variances of random effects at individual-, month-, and day-level, and error term, respectively, and all the four random components are assumed to be uncorrelated of each other.

Then, the total variances of Eqs. (1) and (2) can be expressed below, respectively.

$$\mathrm{Var}(y_{idm}^j) = \sigma_i^2 + \sigma_d^2 + \sigma_m^2 + \sigma_0^2, \tag{3}$$

$$\mathrm{Var}(y_{idm}^j) = var(\beta_1 x_{1idm}^j) + \mathrm{var}(\beta_2 x_{2idm}^j) + \cdots + \mathrm{var}(\beta_n x_{nidm}^j) + \sigma_i^2 + \sigma_d^2 + \sigma_m^2 + \sigma_0^2. \tag{4}$$

All unknown parameters in the above models are estimated based on full maximum likelihood method, which are conducted using software *STATA*.

Generally, all the random components in the *Null* model are usually larger than those in the *Full* model because the variances of explanatory variables explain a part of the total variance.

### 3.4. Research procedure

The research procedure is designed in line with the two-fold research objectives, as shown in Fig. 1.

First, referring to various driving parameters, PCA was implemented to reduce the numerous dimensions. The main principal components selected to represent previously detailed dangerous behaviors included speeding, fatigue driving, and jerky driving. Second, based on the main principal components, truck drivers were further clustered into several smaller clusters using DBSCAN, as it is an effective and feasible tool with which to clarify difference types unsafe tendencies among different types of truck drivers. With the prior two steps, different levels of truck drivers' driving dangerousness can be identified based on massive real-time driving data at the macro perspective. Finally, truck drivers' speeding behavior, i.e., speeding frequency, was examined based on a multilevel model, where individual-, day- and month-level error components (or random effects) were introduced to evaluate the unobserved heterogeneities among different types of truck drivers classified by PCA and cluster analysis.

## 4. Data

The digital tachograph data used in this study was obtained from a major Japanese logistic company, who agreed to provide us a year of truck-driving data collected in the Chugoku region, including the five prefectures of Hiroshima, Okayama, Yamaguchi, Shimane, and Tottori. Because this study was a part of a project with a major expressway company in this region, all trucks with use of expressways were targeted and as a result, a total of 70 truck drivers' daily driving data was collected from February to December in the year of 2014. The information included in the 11-month digital tachograph data is as follows:

- Location information: longitude and latitude every minute;
- Driving speed: real-time speed every 0.5 s;
- Engine speed: revolutions per minute (RPM) every 0.5 s;
- Information of events: detailed events (e.g., ignition and stall, handle start and stop, ETC (Electronic Toll Collection) record, speeding, and overtime driving) and their corresponding timings. Overtime driving denotes an event in which a driver drove more than four and half hours without rest.

### 4.1. Driving indicators

The raw data were first cleaned and prepared for PCA. Then, corresponding indicators were selected, as introduced in Table 1. Data involving speed, driving duration, and accelerating or braking were standardized before use in subsequent analyses. In total, the proposed approach adopted sixteen driving indicators grouped into three categories based on the type of indicator: speed-related, fatigue-related, and jerk-related (Table 1). For less intuitive variables, such as Average Max Speed, Speeding Rate (%), Average Max Continuous Driving Time and Overtime Driving Rate (%), we have provided more detailed descriptions.

(1) Speed-related indicators

Previous research has emphasized correlations between accidents and speeding behavior. Average speed has been used as a measure for weighting driving behavior as it reflects overall time performance (Wang et al., 2016). Maximum speed is also an indicator of speeding behavior (Constantinescu et al., 2010; Wu et al., 2016). Atombo et al. (2016) revealed that excessive speeding and incorrect overtaking, which is associated with maximum speed violations with greater frequency than the other driving behaviors, could predetermine the severity of injuries and fatalities. Therefore, Average Max Speed was used to capture these types of speed-related violation. Regarding Speeding Rate, Wu et al. (2016) adopted a similar metric to describe speeding violations, which is derived from the proportion of vehicle speeds greater than 80% of the road speed limit. However, this proportion does not accurately reflect drivers' real speeding behaviors. Therefore, in this study, only truck drivers' actual speeding violation(s), i.e., Speeding Rate, were used for analysis.

(2) Fatigue-related indicators

There are high frequency of accidents with serious consequences from driving when operators are tired or fatigued owing to extended driving hours without sleep or rest (Radun et al., 2013). Because there is no validated and reliable device for detecting the level of a driver's sleepiness, this paper uses driving duration to describe the tendency for fatigued driving.

The detailed indicators, Max Continuous Driving Time and Overtime Driving Rate (%), effectively describe whether truck drivers obey the aforementioned driving rule. That is, to rest after driving for more than four and half hours. Similar to Speeding Rate, Overtime Driving Rate is equal to the total number of trips made by an individual truck driver divided by the number of those trips when the truck driver violated the preceding driving rule. Moreover, the commercial nature of these trips generally involves several days and nights of continuous driving. Thus, the average max continuous driving time, is used to describe the overall driving time violation for the entire trip.

(3) Jerk-related indicators

To some extent, the frequency of acceleration or braking can reflect the aggressiveness of a driver (Murphey et al., 2009). In this context, Average Acceleration, Average Braking, Acceleration Standard Deviation (SD), and Braking SD were used to classify safety-related driver styles, as described by Langari and Won (2005a; 2005b).

### 4.2. Explanatory variables for speeding behavior

The multilevel analysis used data collected from 4373 trips. Factors at spatial, temporal, and trip levels were defined based on previous studies (Huang et al., 2008; Familar et al., 2011; Ellison et al., 2015). Details are provided in Table 2.

1) Spatial variables: Because most driving routes were within Hiroshima Prefecture, the origins and destinations were divided geographically into five categories: (1) Hiroshima City urban area, (2) Hiroshima City suburban area, (3) eastern Hiroshima City urban area, (4) northern Hiroshima City urban area, and (5) outside the Hiroshima Prefecture. Three road types were distinguished; namely, expressways, national roads, and prefectural roads.
2) Temporal variables: According to Huang et al. (2008), speeding behavior varies on weekdays and weekends. Thus, speeding frequencies of truck drivers on weekdays and holidays, which included weekends, are counted, respectively.

**Table 1**
Description and measurement of data of driving indicators for principal component analysis (PCA) and cluster analysis.

| No. | Driving indicators | Description | Measurement |
|---|---|---|---|
| *Speed-related indicators* | | | |
| 1 | Average Max Speed (km/h) | Average of maximum speeds of all trips made by a truck driver during the survey period | Sum of maximum speeds of all trips that a truck driver made during the survey period divided by the number of all the trips |
| 2 | Max Speed (km/h) | Maximum speed of all trips made by a truck driver during the survey period | Recorded by an on-board digital tachograph data recorder equipped inside each truck |
| 3 | Average Speed (km/h) | Average speed of all trips made by a truck driver during the survey period | Sum of average speeds of all trips that a truck driver made during the survey period divided by the number of all the trips |
| 4 | Average Speed Standard Deviation (SD) | Standard deviation of average speeds of all trips made by a truck driver during the survey period | |
| 5 | Speeding Rate (%) | Frequency that a truck driver over-speeded in all trips | Number of trips that a truck driver is speeding/ the truck driver's total trips |
| *Fatigue-related indicators* | | | |
| 6 | Average Max Continuous Driving Duration (s) | Average of maximum continuous driving durations of all trips made by a truck driver during the survey period | Sum of maximum continuous driving durations of all trips that a truck driver made during the survey period divided by the number of all the trips |
| 7 | Max Continuous Driving Duration (s) | Maximum continuous driving duration among all trips made by a truck driver during the survey period | Recorded by an on-board digital tachograph data recorder equipped inside each truck |
| 8 | Average Driving Duration (s) | Average of driving durations of all trips made by a truck driver during the survey period | Sum of driving durations of all trips that a truck driver made during the survey period divided by the number of all the trips |
| 9 | Max Driving Duration (s) | Maximum driving duration among all trips made by a truck driver during the survey period | Recorded by an on-board digital tachograph data recorder equipped inside each truck |
| 10 | Overtime Driving Rate (%) | Frequency that a truck driver violated the driving time rule during the survey period | The number of all trips made by the truck driver during the survey period divided by the number of trips that a truck driver drove more than four and half hours without rest |
| *Jerk-related indicators* | | | |
| 11 | Average Acceleration (km/hs) | Average of a truck driver' accelerated speed of overall trips | Sum of a truck driver' accelerated speed /the truck driver's total trips |
| 12 | Max Acceleration (km/hs) | Maximum acceleration in all trips made by a truck driver during the survey period | Recorded by an on-board digital tachograph data recorder equipped inside each truck |
| 13 | Average Acceleration Standard Deviation (SD) | Standard deviation of average accelerations in all trips made by a truck driver during the survey period | |
| 14 | Average Braking (km/hs) | Average of decelerations in all trips made by a truck driver during the survey period | Sum of a truck driver' decelerated speed /the truck driver's total trips |
| 15 | Max Braking (km/hs) | Maximum deceleration in all trips made by a truck driver during the survey period | Recorded by an on-board digital tachograph data recorder equipped inside each truck |
| 16 | Average Braking Standard Deviation (SD) | Standard deviation of average decelerations in all trips made by a truck driver during the survey period | |

3) Trip-level variables: Whether it rained, driving duration, and driving distance were considered at trip level. Rain often results in low visibility and slippery road conditions, and thus is regarded as a major factor in many road safety studies.

## 5. Results

### 5.1. PCA results

The variance contribution rates, and cumulative contribution rates are listed in Table 3. The first three components explain approximately 72% of total variance, which is enough to warrant further analysis. A rotation is needed to better explain the original variables. Rotated results are detailed in the last three columns, in which the three main components with eigenvalues greater than 1.0 are presented.

Hereafter, the above three rotated (principal) components are used to replace the original 16 variables. And, the practical meaning of each component is given based on the correlations between diving indicators and rotated components. Table 4 shows the rotated component matrix.

Looking at strong correlations between rotated components and driving indicators, it is observed that Rotated Component 1 (RC1) is related to most speeding behaviors (four out of five indicators: Average Max Speed, Max Speed, Average Speed, and Speeding Rate) and most fatigue-driving behaviors (four out of five indicators: Average Max Continuous Driving Duration, Max Continuous Driving Duration, Average Driving Duration, and Overtime Driving Rate). So, RC1 can be interpreted to represent strong speeding and fatigue-driving behaviors. Rotated Component 2 (RC2) has a strong correlation with acceleration

**Table 2**
Explanation variables used in the multilevel analysis.

| Explanatory variables | Description |
|---|---|
| **Spatial variables** | |
| origin | 1: 19.73%, 2: 15.89%, 3: 17.47%, 4: 1.66%, 5: 45.25% |
| destination | 1: 20.54%, 2: 16.53%, 3: 17.22%, 4: 1.67%, 5: 44.04% |
| route (dummy variable for each type) | 1: Sanyo expressway (67.85%), 2: Chugoku expressway (23.55%), 3: National road No.2 (2.90%), 4: National road No.182 (1.97%), 5: National road No.486 (1.58%), 6: National road No.54 (2.15%) |
| road type (dummy variable for each type) | 1: expressway (91.40%), 2: national road (6.45%), 3: prefectural road (2.15%) |
| **Temporal variables** | |
| holiday (including weekend) | 0: No (73.59%), 1: Yes (26.41%) |
| **Trip-level variables** | |
| rain | 0: No (77.49%), 1: Yes (22.51%) |
| driving duration (s) | total length of driving time during a trip |
| driving distance (km) | total driving distance during a trip |

**Table 3**
Total variance explained by different principal components.

| Principal Component | Initial Eigenvalues | | | Rotation Sums of Squared loadings | | |
|---|---|---|---|---|---|---|
| | Total | % of Variance | Cumulative % | Total | % of Variance | Cumulative % |
| 1 | 7.436 | 46.474 | 46.474 | 5.555 | 34.717 | 34.717 |
| 2 | 3.033 | 18.953 | 65.428 | 4.896 | 30.597 | 65.314 |
| 3 | 1.015 | 6.345 | 71.773 | 1.033 | 6.459 | 71.773 |
| 4 | 0.989 | 6.183 | 77.956 | | | |
| 5 | 0.775 | 4.841 | 82.791 | | | |
| 6 | 0.690 | 4.311 | 87.108 | | | |
| 7 | 0.512 | 3.197 | 90.305 | | | |
| 8 | 0.468 | 2.923 | 93.228 | | | |
| 9 | 0.351 | 2.191 | 95.419 | | | |
| 10 | 0.251 | 1.571 | 96.990 | | | |
| 11 | 0.196 | 1.225 | 98.215 | | | |
| 12 | 0.114 | 0.711 | 98.926 | | | |
| 13 | 0.076 | 0.476 | 99.402 | | | |
| 14 | 0.069 | 0.431 | 99.833 | | | |
| 15 | 0.019 | 0.117 | 99.951 | | | |
| 16 | 0.008 | 0.049 | 100.000 | | | |

**Table 4**
Rotated component matrix.

| No. | Driving variables | Rotated Components | | |
|---|---|---|---|---|
| | | RC1 | RC2 | RC3 |
| 1 | Average max speed (km/h) | 0.881* | −0.137 | 0.120 |
| 2 | Max speed (km/h) | 0.712* | 0.150 | 0.074 |
| 3 | Average speed (km/h) | 0.825* | −0.406 | 0.024 |
| 4 | Average speed STD | 0.547 | 0.390 | −0.334 |
| 5 | Speeding rate (%) | 0.860* | −0.134 | 0.006 |
| 6 | Average max continuous driving duration (s) | 0.775* | −0.447 | 0.530 |
| 7 | Max continuous driving duration (s) | 0.719* | −0.156 | −0.243 |
| 8 | Average driving duration (s) | 0.734* | −0.408 | 0.790* |
| 9 | Max driving duration (s) | 0.186 | −0.060 | 0.270 |
| 10 | Overtime driving rate (%) | 0.735* | −0.107 | 0.701* |
| 11 | Average acceleration (km/hs) | −0.349 | 0.881* | 0.139 |
| 12 | Max acceleration (km/hs) | −0.197 | 0.879* | −0.009 |
| 13 | Average acceleration STD | −0.030 | 0.276 | 0.841* |
| 14 | Average braking (km/hs) | −0.304 | 0.909* | 0.086 |
| 15 | Max braking (km/hs) | −0.125 | 0.913* | −0.081 |
| 16 | Average braking STD | −0.220 | 0.390 | 0.099 |

*Note.*
* Strong correlations (> 0.7) between rotated components and driving indicators.

(two out of three indicators: Average Acceleration, Max Acceleration) and braking (two out of three indicators: Average Braking, Max Braking). Thus, RC2 represents strong jerk-driving behaviors. Rotated Component 3 (RC3) is slightly related to fatigue-driving behaviors (two out of five indicators: Average Driving Duration, Overtime Driving Rate) and acceleration (one out of three indicators: Average Acceleration SD), and it indicates weak fatigue-driving behaviors.

By interpreting the real meaning of rotated components, we can further calculate their scores for a comprehensive assessment. The scores are calculated using ratio of variance contribution. And then, a cluster analysis is conducted by using the component scores.

### 5.2. Cluster analysis results

Based on the rotated components from the above PCA, we derived five clusters, as summarized in Table 5. The safety levels can be interpreted to be very safe (Cluster 1), slightly safe (Cluster 2), slightly dangerous (Cluster 3), dangerous (Cluster 4), and very dangerous (Cluster 5), each of which includes 9, 6, 27, 9, and 19 truck drivers, respectively. The safety levels are assigned by interpreting the speeding, fatigue, and jerky driving indicators. In Table 5, RC1_s, RC1_w, RC2_s, and RC3_2 represent the relationship between clusters and specific driving behaviors, where "_s" represents a strong correlation and "_w" a weak correlation. After the normalization of component scores, the detailed relationships are described as Low (0-0.2), Slight low (0.2-0.4), Neutral (0.4-0.6), Slight high (0.6-0.8) and High (0.8–1).

Cluster 1 accounts for 12.86% of the truck drivers, who are identified as *very safe truck drivers*. These truck drivers seldom experience speeding, fatigue driving, and jerky driving. Cluster 2 (8.57%) includes *slightly safe truck drivers*, because their frequencies of speeding, fatigue driving, and jerky driving are slightly low. Truck drivers in Cluster 3 can be interpreted to be *slightly dangerous truck drivers*, because frequencies of speeding and acceleration are slightly high, but frequency of fatigue driving is neutral. Cluster 4 includes *dangerous truck drivers*, who exhibit high frequency speeding, and slightly higher than average frequencies of fatigued and jerky driving. Truck drivers in Cluster 5 are *very dangerous truck drivers* with high frequencies of speeding and jerky driving, and slightly elevated frequencies of fatigue driving. Thus, dangerous and very dangerous truck drivers, Clusters 4 and 5, account for 40% of operators, and including truck drivers in Cluster 3, 78.57% of drivers exhibited dangerous driving behaviors in the 11-month period.

Fig. 2 depicts clusters of interactions between different figures and displays combination of different rotated component scores. For instance, the figures in column 1 and row 2 divide truck drivers into two different groups: the blue points represent truck drivers with a potentially high risk of jerky driving and a low risk of speeding, while the red blocks show a different pattern. This is because RC1 represents speeding behavior and fatigue-driving behavior, and RC2 indicates jerky driving behavior. Fig. 2 gives us a more intuitive information about different truck drivers' potentially dangerous driving tendencies.

The major findings from Fig. 2 are summarized below.

(i) The combination of strong speeding and fatigue-driving behaviors (RC1) and strong jerky driving behaviors (RC2): Most truck drivers who infrequently drove over speed and or with jerky behaviors, were safe drivers overall. The remaining truck drivers' scores were located dispersedly with nearly half of the truck drivers displaying a strong tendency to over-speed, and accelerate or brake frequently, thus demonstrating very dangerous driving behaviors.
(ii) Combination of strong speeding and fatigue-driving behaviors (RC1) and weak fatigue-driving behaviors (RC3): There was no distinct tendency of fatigue-driving behavior because truck drivers performed in a balanced way in terms of driving time with few truck drivers driving for extended periods of time without rest.
(iii) Combination of strong jerky driving behaviors (RC2) and weak fatigue-driving behaviors (RC3): Truck drivers were located

**Table 5**
Interpretation of the clusters.

| Cluster | | Truck drivers (proportions) | Speeding (RC1_s) | Fatigue (RC1_w, RC3_w) | Jerk (RC2_s, RC3_w) |
|---|---|---|---|---|---|
| 1 | Very safe truck drivers | 9 (12.86%) | Low | Low | Low |
| 2 | Slightly safe truck drivers | 6 (8.57%) | Slight low | Slight low | Slight low |
| 3 | Slightly dangerous truck drivers | 27 (38.57%) | Slight high | Neutral | Slight high |
| 4 | Dangerous truck drivers | 19 (27.14%) | High | Slight high | Slight high |
| 5 | Very dangerous truck drivers | 9 (12.86%) | High | Slight high | High |

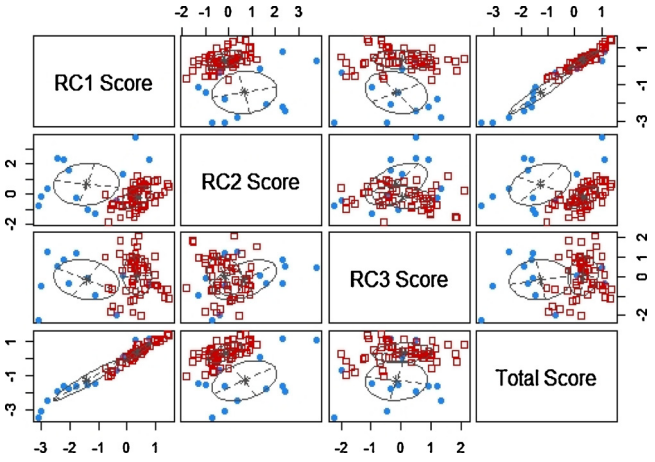


**Fig. 2.** Cluster of interaction between different figures.

randomly in this case, and only four truck drivers (5.7% of all truck drivers) were recognized for frequent acceleration or braking.

Furthermore, to capture location characteristics of truck drivers' dangerous behaviors, we created a density estimation diagram, depicted in Fig. 3, which identifies areas with high concentrations of the truck drivers.

### 5.3. Multilevel modeling results

Here, speeding frequencies based on all data collected from 4373 trips made by 70 truck drivers over 11 months were used and analyzed for different clusters. Sample sizes are 547 for Cluster 1, 393 for Cluster 2, 1762 for Cluster 3, 919 for Cluster 4, and 752 for Cluster 5. Differing from the Speeding Rate (%) discussed in Section 4.1, for clustering truck drivers' driving styles at the macro level, speeding frequencies reflect truck drivers' detailed speeding behaviors, that is how many times a truck driver drove over the speed limit during a trip, which was measured at the micro level. The samples for each of the five clusters were composed of individual truck drivers and their trips made over the eleven months of data collection. Thus, we suggest that it is not logical to assume independence across the samples. Because of this concern, this study adopted the multilevel model introduced in Section 3, and addressed speeding frequency during each trip as a dependent variable.

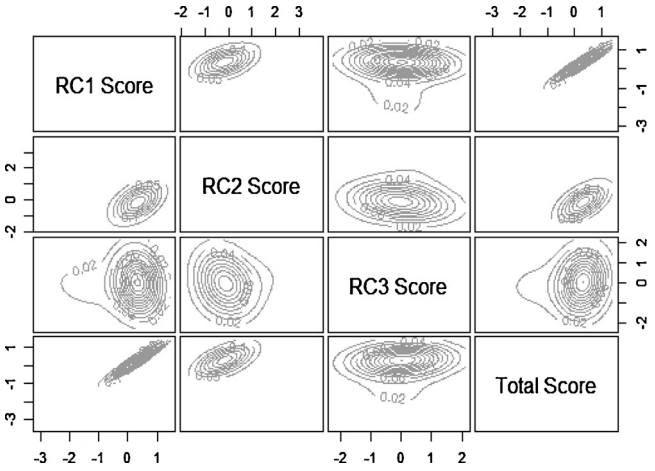


**Fig. 3.** Density estimation diagram.

First, a *Null* model with no explanatory variable was estimated for each of the five clusters. Next, a *Full* model with explanatory variables was estimated for each cluster. With such a *Full* model, it was possible to know how much of the unobserved variances of random components could be explained by observed variables. Results of variance proportions captured by both the *Null* and *Full* models are shown in Table 6 and detailed parameter estimation results of the *Full* models are displayed in Table 7.

The major findings of Table 6 are summarized below.

First, the *Null* model describes the dependent variable as a function of an average value, i.e., error component or random effect, and the average value was specified to vary across the levels, to enable the investigation of variance proportion at different levels.

(i)

(i) The individual-level random effect showed the strongest influence on speeding behavior among almost all the clusters, except for Cluster 5, which refers to *very dangerous truck drivers* identified in the prior cluster analysis. This suggests that the actual speed-related violations are mainly determined by clustered individual truck drivers, excluding Cluster 5.
(ii) The variance related to the month-level random effect accounts for the second largest segment (0.00%˜55.79%). Specifically, the month-level random effect of Cluster 5 shows the largest impact, and those of Clusters 2 and 3 have the second largest impact.
(iii) The day-level random effect had the weakest impact on speeding behavior (0.00%˜19.35%).
(iv) The variances explained by the day-level random effect of Cluster 3, and month-level random effects of Clusters 1 and 5 may be ignored.

**Table 6**
Variance proportions.

| Models | Individual-level error component | Day-level error component | Month-level error component | Variations captured by explanatory variables |
|---|---|---|---|---|
| *Null* model | | | | |
| Cluster 1 | 93.61% | 6.39% | 0.00% | |
| Cluster 2 | 60.78% | 18.42% | 20.80% | |
| Cluster 3 | 77.50% | 0.00% | 22.50% | |
| Cluster 4 | 80.65% | 19.35% | 0.00% | |
| Cluster 5 | 35.00% | 9.21% | 55.79% | |
| *Full* model | | | | |
| Cluster 1 | 45.77% | 20.06% | 0.00% | 34.17% |
| Cluster 2 | 29.07% | 8.55% | 36.01% | 26.37% |
| Cluster 3 | 70.39% | 0.00% | 14.91% | 14.70% |
| Cluster 4 | 61.04% | 9.64% | 0.00% | 29.32% |
| Cluster 5 | 0.00% | 0.00% | 20.24% | 79.76% |

**Table 7**
*Full* model estimation results.

| Variable | Cluster 1 | | | Cluster 2 | | | Cluster 3 | | | Cluster 4 | | | Cluster 5 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\beta$ | SE | P > \|z\| | $\beta$ | SE | P > \|z\| | $\beta$ | SE | P > \|z\| | $\beta$ | SE | P > \|z\| | $\beta$ | SE | P > \|z\| |
| Intercept | 0.170 | 0.391 | 0.664 | 0.229 | 0.284 | 0.419 | 2.223 | 0.320 | 0.000 | 0.037 | 0.530 | 0.945 | 0.061 | 0.265 | 0.818 |
| Origin = 1 | 0.783 | 0.471 | 0.096 | −0.747 | 0.842 | 0.375 | 0.432 | 0.600 | 0.472 | 3.736 | 1.355 | 0.006 | 0.387 | 0.423 | 0.360 |
| Origin = 2 | −0.412 | 0.400 | 0.304 | 0.623 | 0.857 | 0.467 | −1.217 | 0.565 | 0.031 | −3.369 | 1.345 | 0.012 | 0.001 | 0.171 | 0.993 |
| Origin = 3 | −0.239 | 0.950 | 0.801 | −0.327 | 0.323 | 0.313 | −0.835 | 0.331 | 0.012 | −3.306 | 1.068 | 0.002 | −0.030 | 0.510 | 0.953 |
| Origin = 4 | – | – | – | −0.945 | 0.631 | 0.134 | – | – | – | 3.031 | 1.364 | 0.026 | 0.113 | 0.225 | 0.614 |
| Origin = 5 | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – |
| Destination = 1 | 0.494 | 0.364 | 0.175 | −0.409 | 0.291 | 0.160 | −0.109 | 0.202 | 0.591 | −0.538 | 0.381 | 0.159 | −0.141 | 0.142 | 0.317 |
| Destination = 2 | 0.230 | 1.278 | 0.857 | 1.172 | 0.304 | 0.000 | −0.175 | 0.244 | 0.475 | 0.201 | 0.321 | 0.532 | 0.024 | 0.151 | 0.871 |
| Destination = 3 | 0.197 | 0.909 | 0.828 | −0.074 | 0.297 | 0.803 | −0.164 | 0.250 | 0.511 | −3.102 | 1.067 | 0.004 | 0.083 | 0.318 | 0.795 |
| Destination = 4 | – | – | – | – | – | – | – | – | – | −0.005 | 0.419 | 0.991 | 0.501 | 0.195 | 0.797 |
| Destination = 5 | −0.316 | 1.415 | 0.327 | 0.199 | 0.813 | 0.807 | −1.336 | 0.559 | 0.017 | −3.19 | 1.240 | 0.01 | 0.117 | 0.197 | 0.554 |
| O2D1 | 0.720 | 0.825 | 0.383 | – | – | – | −2.585 | 0.844 | 0.002 | −0.649 | 1.431 | 0.650 | −1.456 | 0.518 | 0.005 |
| O3D1 | −0.717 | 1.009 | 0.478 | – | – | – | −1.304 | 0.585 | 0.026 | −3.148 | 1.167 | 0.007 | −0.003 | 0.566 | 0.996 |
| Route = 1 | 0.674 | 0.993 | 0.497 | −1.183 | 0.422 | 0.005 | 0.432 | 0.429 | 0.314 | – | – | – | 2.938 | 0.629 | 0.000 |
| Route = 2 | – | – | – | – | – | – | – | – | – | – | – | – | 2.845 | 0.658 | 0.000 |
| Route = 3 | – | – | – | – | – | – | – | – | – | – | – | – | 2.159 | 0.489 | 0.000 |
| Route = 4 | – | – | – | – | – | – | −1.081 | 0.638 | 0.090 | – | – | – | 3.114 | 0.763 | 0.000 |
| Route = 5 | 0.439 | 0.233 | 0.059 | – | – | – | −0.726 | 0.725 | 0.317 | – | – | – | – | – | – |
| Route = 6 | – | – | – | – | – | – | −0.694 | 0.667 | 0.298 | – | – | – | 1.296 | 0.756 | 0.087 |
| Road type = 1 | −0.482 | 0.922 | 0.601 | 1.856 | 0.407 | 0.000 | −0.556 | 0.508 | 0.274 | 1.697 | 0.432 | 0.000 | −0.307 | 0.583 | 0.598 |
| Road type = 2 | – | – | – | 0.136 | 0.318 | 0.669 | −0.367 | 0.626 | 0.558 | – | – | – | – | – | – |
| Road type = 3 | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – |
| Rain | −0.163 | 0.094 | 0.082 | −0.020 | 0.086 | 0.814 | 0.012 | 0.042 | 0.770 | −0.041 | 0.756 | 0.585 | −0.144 | 0.052 | 0.005 |
| Holiday | −0.172 | 0.121 | 0.154 | −0.146 | 0.172 | 0.396 | −0.077 | 0.054 | 0.154 | −0.876 | 0.088 | 0.322 | −0.138 | 0.070 | 0.047 |
| Driving time | 0.11 | 0.000 | 0.000 | 0.188 | 0.000 | 0.000 | 0.432 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.001 | 0.000 | 0.000 |
| Distance | 0.004 | 0.001 | 0.002 | 0.003 | 0.001 | 0.000 | 2.223 | 0.000 | 0.001 | 0.005 | 0.001 | 0.000 | 0.002 | 0.001 | 0.000 |
| Random effects | | | | | | | | | | | | | | | |
| Individual | 0.044 | 0.032 | 0.000 | 0.055 | 0.051 | 0.000 | 0.225 | 0.064 | 0.000 | 0.137 | 0.051 | 0.000 | 0.000 | 0.000 | 0.000 |
| Day | 0.019 | 0.027 | 0.000 | 0.016 | 0.022 | 0.000 | 0.000 | 0.000 | 0.000 | 0.022 | 0.018 | 0.000 | 0.000 | 0.000 | 0.000 |
| Month | 0.000 | 0.000 | 0.000 | 0.068 | 0.046 | 0.000 | 0.048 | 0.015 | 0.000 | 0.000 | 0.000 | 0.000 | 0.369 | 0.037 | 0.000 |

Second, the explanatory variables were added into the model and expressed as the *Full* model. This step captured changes in variances explained by unobserved heterogeneities, captured via random effects, due to the introduction of explanatory variables.

(i) The most distinct and meaningful finding is that the variance proportion of explanatory variables for the truck drivers in Cluster 5 (79.76%), i.e., *very dangerous truck drivers*, was substantially higher than that of truck drivers from the other clusters (14.70% ˜34.17%). This demonstrates that explanatory variables exert substantial influence on speeding behavior among *very dangerous truck drivers*. This phenomenon proves the efficacy and validates the results from the PCA and cluster analysis, which revealed the existence of different types of truck drivers in terms of dangerous driving behaviors.
(ii) The second important finding concerns the proportion of variance reduction. Comparing the *Null* model with the *Full* model, we observe that the individual-level random effects of Clusters 1, 2 and 5, and the month-level random effect of Cluster 5 realized the largest reductions in variance, exceeding 30%. This indicates that the unobserved variations were accurately reflected by the introduced explanatory variables.
(iii) The variations of the individual-level random effect of Cluster 5, the day-level random effects of Clusters 3 and 5, and month-level random effects of Clusters 1 and 4 were less meaningful in the *Full* model.

Looking at Table 7, the most distinct result is the positive impact of driving time and driving distance on speeding behavior among all driving types. As for the implications, the origin-destination (OD) information suggests trips departing from Hiroshima City suburban area and the eastern Hiroshima urban area, and those arriving at Hiroshima City urban area were less likely to over-speed. This phenomenon may have been due to spatial or traffic conditions. In terms of road types and routes, traveling on expressways was inherently more dangerous. The truck drivers in Cluster 5 were associated with an increased speeding when they choose Route 1 (Sanyo expressway) while the truck drivers of Cluster 2 exhibited the opposite trend.

Interestingly, temporal effects indicate that potentially extremely dangerous truck drivers (Cluster 4) were less likely to speed on holidays, that is, weekends and statutory holidays, or rainy days. The former may be associated with traffic flow, while the latter indicates that although these truck drivers were more likely to speed in normal situations, they demonstrated good safety awareness in adverse weather conditions.

## 6. Discussion

Traffic accidents are relatively rare events. So, surrogates were needed to compensate for the insufficient number of real-word incidents required to execute an accurate risk assessment. Existing studies have mainly focused on accident indices such as crash rate or near-crash rate. However, more a proactive approach is required to identify and mitigate potential risks. A comprehensive analysis that considers truck drivers, vehicles, and driving environment, as well as temporal and spatial effects is also needed.

Accordingly, in this study, truck drivers were divided into five clusters using PCA and cluster analysis, based on their daily performance. Additional cluster analysis was applied obtain different truck drivers' potential driving tendency in terms of speed, driving time, and jerky performance. The advantage of adopting the DBSCAN algorithm was to provide intuitive information about inter-cluster interactions across different dangerous behaviors. For instance, although some truck drivers seldom drove over-speed, they did show a tendency for fatigue-driving. While other truck drivers demonstrated stable control of acceleration and braking, but exhibited high frequencies of speeding. The

complexity of dangerous behaviors was accurately reflected by this density-based cluster analysis. The combination of DBSCAN clustering and PCA was a first attempt to classify truck drivers' driving styles for safety studies. The results of this macro-level analysis will help improve the management of truck drivers for transport and logistics companies. The insurance and actuarial industries can also benefit from this classification methodology.

Driving risk varies distinctly among drivers. Thus, from the micro-level viewpoint, multilevel modeling was introduced to reveal the influence of various unobserved factors across different truck drivers with different driving behaviors. The modeling results confirm that speeding behaviors vary largely across the spectrum of dangerous and safe truck drivers. This supports the use of multilevel models for safety research. Moreover, detailed information about spatial, temporal and trip attributes were obtained from the multilevel modeling estimation.

Guo (2019) conducted a review of statistical methods for NDS and pointed out that data mining and machine learning methods are promising approaches for analysis of high-resolution and high-frequency NDS driving data. Also, there is a need for alternative and/or additional statistical analysis methods that employ traditional regression tools. Pantangi et al. (2019) investigated aggressive driving behavior and further raised concerns about distinguishing among various types of speeding violations such as drastic speeding in a short period of time or marginal speeding sustained for long period of time. The proposed research framework and multiple specified driving indicators, in addition to the explanatory variables, constitute a comprehensive response to the above concerns.

Various implications for traffic safety management and continuing research on commercial truck drivers can be derived from this study. First, truck drivers identified for dangerous tendencies must be monitored more closely during daily company operations. Safety education or regulations would reduce inappropriate behaviors, thus mitigating some of the more serious consequences. Second, different driving characteristics were obtained from the density-based cluster analysis. Further study focusing on the complex driving behavior patterns is required. Moreover, the results of the multilevel model indicate that as driving time and distance increase, all truck drivers are more likely to speed, regardless of behavioral type. Countermeasures addressing such behaviors, such as reasonable operation arrangements and schedules, could reduce the potential risks.

Finally, fully-utilized large-scale GPS data effectively served as a surrogate measure of truck drivers' potential tendencies in terms of speed, driving time, and frequency of acceleration or braking. The incorporation of big data with data mining techniques, as well as multilevel modeling approaches provided deeper insights on various potentially dangerous behaviors from both macro and micro perspectives.

## 7. Conclusion

The paper proposes an innovative approach to extracting useful information about commercial truck driver behavior from widely available big data sources. This study used eleven months of digital tachograph data that comprised 4373 trips made by 70 truck drivers in Japan. Results suggest that 40% of truck drivers exhibit substantially dangerous driving tendencies. The explanatory variables introduced in this study accurately expressed the influence of unobserved conditions and phenomena for potential extremely dangerous truck drivers, especially in comparison with other types of truck drivers.

Although many factors may affect driving behavior, because of data constraints, truck drivers' personal attributes (e.g., age, gender, education) and psychological conditions (e.g., upset, tired, angry) were ignored, and GPS data in this research was only used for extracting OD information. Furthermore, although the intuitive aspects of multilevel modeling are appealing, many challenges remain to its practical application and interpretation. Despite these limitations, this paper provides a systematic approach to identifying potential risks among different truck drivers that considers both macro and micro perspectives. We suggest that future work focus on the following four aspects. First, big data should be integrated with questionnaire survey data about truck drivers' personal characteristics and psychological factors. Second, more spatial information could be derived from GPS data. Third, effective data fusion approaches should be developed to support the above data integration and maximize statistical and epidemiological methods, deep learning techniques, and behavioral models. Finally, the above efforts should be extrapolated to implement practical safety improvements by developing effective decision support systems.

## Declaration of Competing Interest

None.

## Acknowledgements

This research was funded by the Grants-in-Aid for Scientific Research (A), Japan Society for the Promotion of Science (No.15H02271). The authors would also like to thank the China Scholarship Council for their financial support to this research and the three anonymous reviewers for their valuable suggestions.

## References

Ahmed, M.M., Ghasemzadeh, A., 2018. The impacts of heavy rain on speed and headway behaviors: an investigation using the SHRP2 naturalistic driving study data. Transp. Res. Part C Emerg. Technol. 91, 371–384.
Anastasopoulos, P., Shankar, V., Haddock, J., Mannering, F., 2012. A multivariate tobit analysis of highway accident-injury-severity rates. Accid. Anal. Prev. 45 (1), 110–119.
Atombo, C., Wu, C., Zhong, M., Zhang, H., 2016. Investigating the motivational factors influencing drivers intentions to unsafe driving behaviours: speeding and overtaking violations. Transp. Res. Part F Traffic Psychol. Behav. 43, 104–121.
Bao, J., Liu, P., Yu, H., Xu, C., 2017. Incorporating twitter-based human activity information in spatial analysis of crashes in urban areas. Accid. Anal. Prev. 106, 358–369.
Bao, J., Liu, P., Ukkusuri, S.V., 2019. A spatiotemporal deep learning approach for city-wide short-term crash risk prediction with multi-source data. Accid. Anal. Prev. 122, 239–254.
Billot, R., Faouzi, N.E.E., Vuyst, F.D., 2009. Multilevel assessment of the impact of rain on drivers' behavior. Transport. Res. Record J. Transport. 2107, 134–142.
Brooks, J.O., Crisler, M.C., Klein, N., Goodenough, R., Beeco, R.W., Guirl, C., Tyler, P.J., Hilpert, A., Miller, Y., Grygier, J., Burroughs, B., Martin, A., Ray, R., Palmer, C., Beck, C., 2011. Speed choice and driving performance in simulated foggy conditions. Accid. Anal. Prev. 43 (3), 698–705.
Chen, C., Ma, J., Susilo, Y., Liu, Y., Wang, M., 2016. The promises of big data and small data for travel behavior (aka human mobility) analysis. Transp. Res. Part C Emerg. Technol. 68, 285–299.
Chong, L., Abbas, M.M., Flintsch, A.M., Higgs, B., 2013. A rule-based neural network approach to model driver naturalistic behavior in traffic. Transp. Res. Part C Emerg. Technol. 32, 207–223.
Constantinescu, Z., Marinoiu, C., Vladoiu, M., 2010. Driving style analysis using data mining techniques. Int. J. Comput. Commun. Control. 5, 654–663.
Dupont, E., Papadimitriou, E., Martensen, H., Yannis, G., 2013. Multilevel analysis in road safety research. Accid. Anal. Prev. 60, 402–411.
Eboli, L., Guido, G., Mazzulla, G., Pungillo, G., Pungillo, R., 2017. Investigating car users' driving behavior through speed analysis. PROMET-Traffic Eng. 29, 193–202.
El-Basyouny, K., Barua, S., Islam, T., Li, R., 2014. Assessing the effect of weather states on crash severity and type by use of full Bayesian multivariate safety models. Transp. Res. Record. 2432, 65–73.
Ellison, A.B., Greaves, S.P., Bliemer, M.C.J., 2015. Driver behaviour profiles for road safety analysis. Accid. Anal. Prev. 76, 118–132.
Familar, R., Greaves, S., Ellison, A., 2011. Analysis of speeding behavior: multilevel modeling approach. Transport. Res. Record J. Transport. 2237, 67–77.
Faure, V., Lobjois, R., Benguigui, N., 2016. The effects of driving environment complexity and dual tasking on drivers' mental workload and eye blink behavior. Transp. Res. Part F Traffic Psychol. Behav. 40, 78–90.
Filtness, A.J., Armstrong, K.A., Watson, A., Smith, S.S., 2017. Sleep-related crash characteristics: implications for applying a fatigue definition to crash reports. Accid. Anal. Prev. 99, 440–444.
Guo, F., 2019. Statistical methods for naturalistic driving studies. Annu. Rev. Stat. Appl. 6, 309–328.
Guo, F., Hankey, J.M., 2009. Modeling 100-Car Safety Events: a Case-Based Approach for Analyzing Naturalistic Driving Data. The National Surface Transportation Safety Center for Excellence.
Garcia, D.P., Caraschi, J.C., Ventorim, G., Vieira, F.H.A., Protásioc, T.P., 2019. Assessment of plant biomass for pellet production using multivariate statistics (PCA

and HCA). Renew. Energy 139, 796–805.
Hassan, H.M., Shawky, M., Kishta, M., Garib, A.M., Al-Harthei, H.A., 2017. Investigation of drivers' behavior towards speeds using crash data and self-reported questionnaire. Accid. Anal. Prev. 98, 348–358.
He, Y., Yan, X., Lu, X.-Y., Chu, D., Wu, C., 2019. Rollover Risk Assessment and Automated Control for Heavy Duty Vehicles Based on Vehicle-to-infrastructure Information. IET Intel. Transport Syst. 13 (6), 1001–1010. https://doi.org/10.1049/iet-its.2018.5495.
Hickman, J.S., Hanowski, R.J., Bocanegra, J., 2010. Distraction in Commercial Trucks and Buses: Assessing Prevalence and Risk in Conjunction with Crashes and Near Crashes. Report No. FMCSA-RRR-10-049. Federal Motor Carrier Safety Administration, Washington, DC.
Hou, Q., Tarko, A.P., Meng, X., 2018. Investigating factors of crash frequency with random effects and random parameters models: new insights from Chinese freeway study. Accid. Anal. Prev. 120, 1–12.
Hu, S., Ivan, J.N., Ravishanker, N., Mooradian, J., 2013. Temporal modeling of highway crash counts for senior and non-senior drivers. Analysis and Prevention 50, 1003–1013.
Huang, H., Abdel-Aty, M., 2010. Multilevel data and Bayesian analysis in traffic safety. Accid. Anal. Prev. 42, 1556–1565.
Huang, H., Chin, H.C., Mazharul, H., 2008. Severity of driver injury and vehicle damage in traffic crashes at intersections: a Bayesian hierarchical analysis. Accid. Anal. Prev. 40, 45–54.
Huang, H., Zhou, H., Wang, J., Chang, F., Ma, M., 2017. A multivariate spatial model of crash frequency by transportation modes for urban intersections. Anal. Methods Accid. Res. 14, 10–21.
Islam, S., Jones, S.L., Dye, D., 2014. Comprehensive analysis of single- and multi-vehicle large truck at-fault crashes on rural and urban roadways in Alabama. Accid. Anal. Prev. 67, 148–158.
Kaber, D., Zhang, Y., Jin, S., Mosaly, P., Garner, M., 2012. Effects of hazard exposure and roadway complexity on young and older driver situation awareness and performance. Transp. Res. Part F 15, 600–611.
Kazemi-Beydokhti, M., Abbaspour, R.A., Mojarab, M., 2017. Spatio-temporal modeling of seismic provinces of iran using DBSCAN algorithm. Pure Appl. Geophys. 174, 1937–1952.
Key, C.E.J., Morris, A.P., Mansfield, N.J., 2017. A study investigating the comparative situation awareness of older and younger drivers when driving a route with extended periods of cognitive taxation. Transp. Res. Part F 49, 145–158.
Kumar, K.M., Reddy, A.R.M., 2016. A fast DBSCAN clustering algorithm by accelerating neighbor searching using Groups method. Pattern Recognit. 58, 39–48.
Langari, R., Won, J.S., 2005a. Intelligent energy management agent for a parallel hybrid vehicle-part I: system architecture and design of the driving situation identification process. IEEE Trans. Veh. Technol. 54 (3), 925–934.
Langari, R., Won, J.S., 2005b. Intelligent energy management agent for a parallel hybrid vehicle-part II: torque distribution, charge sustenance strategies, and performance results. IEEE Trans. Veh. Technol. 54 (3), 935–953.
Li, X., Mou, Y., Wang, H., Yin, C., He, Q., 2018. How does polycentric urban form affect urban commuting? Quantitative measurement using geographical Big Data of 100 Cities in China. Sustainability 10 (12), 4566.
Li, Z., Wang, W., Liu, P., Bigham, J., Ragland, D., 2013. Using geographically weighted Poisson regression for county-level crash modeling in California. Saf. Sci. 58, 89–97.
Liu, X., Li, M., Qin, S., Ma, X., Wang, W., 2016. A predictive fault diagnose method of wind turbine based on K-Means clustering and neural networks. J. Int. Technol. 17 (7), 1521–1528.
Lord, D., Mannering, F., 2010. The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. Transportation Research Part A 44, 291–305.
Malin, F., Norros, I., Innamaa, S., 2019. Accident risk of road and weather conditions on different road types. Accid. Anal. Prev. 122, 181–188.
Mannering, F., Bhat, C.R., 2014. Analytic methods in accident research: methodological frontier and future directions. Anal. Methods Accid. Res. 1, 1–22.
McDonald, A.D., Lee, J.D., Schwarz, C., Brown, T.L., 2018. A contextual and temporal algorithm for driver drowsiness detection. Accid. Anal. Prev. 113, 25–37.
Mueller, A., Trick, L., 2012. Driving in fog: the effects of driving experience and visibility on speed compensation and hazard avoidance. Accid. Anal. Prev. 48, 472–479.
Murphey, Y.L., Milton, R., Kiliaris, L., 2009. Driver's style classification using jerk analysis computational intelligence in vehicles and vehicular systems. CIVVS' 09. IEEE Workshop on 23–28.
Naznin, F., Currie, G., Logan, D., Sarvi, M., 2016. Application of a random effects negative binomial model to examine tram-involved crash frequency on route sections in Melbourne, Australia. Accid. Anal. Prev. 92, 15–21.
OECD, 2008. Handbook on Constructing Composite Indicators. Methodology and User Guide. OECD Publications, Paris.
Pankok, C., Kaber, D., 2018. The effect of navigation display clutter on performance and attention allocation in presentation- and simulator-based driving experiments. Appl. Ergon. 69, 136–145.
Pantangi, S.S., Fountas, G., Sarwar, M.T., Anastasopoulos, P.C., Blatt, A., Majka, K., Pierowicz, J., Mohan, S.B., 2019. A preliminary investigation of the effectiveness of high visibility enforcement programs using naturalistic driving study data: a grouped random parameters approach. Anal. Methods Accid. Res. 21, 1–12.
Park, M., Lee, D., Jeon, J., 2016. Random parameter negative binomial model of signalized intersections. Math. Probl. Eng. 2016, 1–8.
Peng, Y., Abdel-Aty, M., Shi, Q., Yu, R., 2017. Assessing the impact of reduced visibility on traffic crash risk using microscopic data and surrogate safety measures. Transp. Res. Part C 74, 295–305.
Petridou, E., Moustaki, M., 2001. Human factors in the causation of road traffic crashes. Eur. J. Epidemiol. 16 (9), 819–826.
Radun, I., Ohisalo, J., Radun, J., Wahde, M., Kecklund, G., 2013. Driver fatigue and the law from the perspective of police officers and prosecutors. Transp. Res. Part F 18, 159–167.
Rusli, R., Haque, M.M., Afghari, A.P., King, M., 2018. Applying a random parameters Negative Binomial Lindley model to examine multi-vehicle crashes along rural mountainous highways in Malaysia. Accid. Anal. Prev. 119, 80–90.
Shi, Q., Abdel-Aty, M., Yu, R., 2016. Multi-level Bayesian safety analysis with unprocessed Automatic Vehicle Identification data for an urban expressway. Accid. Anal. Prev. 88, 68–76.
Vangala, P., Lord, D., Geedipally, S.R., 2015. Exploring the application of the negative binomial–generalized exponential model for analyzing traffic crash data with excess zeros. Anal. Methods Accid. Res. 7, 29–36.
Wang, X., Fan, T., Li, W., Yu, R., Bullock, D., Wu, B., Tremont, P., 2016. Speed variation during peak and off-peak hours on urban arterials in Shanghai. Transp. Res. Part C 67, 84–94.
Wu, C., Sun, C., Chu, D., Huang, Z., Ma, J., Li, H., 2016. Clustering of several typical behavioral characteristics of commercial vehicle drivers based on GPS data mining. Transportation Research Record: Journal of Transportation 2581, 154–163.
Yan, L., Huang, Z., Zhang, Y., Zhang, L., Ran, B., 2017. Driving risk status prediction using Bayesian networks and Logistic regression. Iet Intell. Transp. Syst. 11 (7), 431–439.
Ye, M., Osman, O.A., Ishak, S., Hashemi, B., 2017. Detection of driver engagement in secondary tasks from observed naturalistic driving behavior. Accid. Anal. Prev. 106, 385–391.
Ye, X., Pendyala, R., Shankar, V., Konduri, K., 2013. A simultaneous model of crash frequency by severity level for freeway sections. Accid. Anal. Prev. 57, 140–149.
Yu, R., Abdel-Aty, M., 2013. Multi-level Bayesian analysis for single- and multi-vehicle freeway crashes. Accid. Anal. Prev. 58, 97–105.
Zeng, Q., Huang, H., Pei, X., Wong, S.C., 2016. Modeling nonlinear relationship between crash frequency by severity and contributing factors by neural networks. Anal. Methods Accid. Res. 10, 12–25.
Zhang, D., He, T., Lin, S., Munir, S., 2017. Taxi-Passenger-Demand Modeling based on big data from a roving sensor network. Ieee Trans. Big Data 3 (3), 362–374.
Zhang, M., Zhang, D., Goerlandt, F., Yan, X., Kujala, P., 2019. Use of HFACS and fault tree model for collision risk factors analysis of icebreaker assistance in ice-covered waters. Safety Science 111, 128–143. https://doi.org/10.1016/j.ssci.2018.07.002.
Zhao, Y., Zhang, H., An, L., Liu, Q., 2018. Improving the approaches of traffic demand forecasting in the big data era. Cities 82, 19–26.
Zhu, X., Srinivasan, S., 2011. A comprehensive analysis of factors influencing the injury severity of large-truck crashes. Accid. Anal. Prev. 43, 49–57.