Break-taking behaviour pattern of long-distance freight vehicles based on GPS trajectory data

Daxin Tian[1,3], Xiongyu Shan[1], Zhengguo Sheng[2], Yunpeng Wang[1,3], Wenzhong Tang[1], and Jian Wang[4]

[1]*School of Transportation Science and Engineering, Beijing Key Laboratory for Cooperative Vehicle Infrastructure Systems and Safety Control, Beihang University, Beijing, 100191, China*

[2] *Department of Engineering and Design, University of Sussex, Brighton, BN1 9RH, United Kingdom*

[3]*Jiangsu Province Collaborative Innovation Center of Modern Urban Traffic Technologies, No.2 SiPaiLou, Nanjing, 210096, China*

[4]*College of Computer Science and Technology, Jilin University, Changchun 130012, China*

Abstract

This paper focuses on the break-taking behaviour pattern of long-distance freight vehicles, providing a new perspective on the study of behaviour patterns and simultaneously providing a reference for transport management departments and related enterprises. Based on Global Positioning System (GPS) trajectory data, we select stopping points as break-taking sites of long-distance freight vehicles and then classify the stopping points into three different classes based on the break-taking duration. We then explore the relationship of the distribution of the break-taking frequency between the three single classifications and their combinations, on the basis of the break-taking duration distribution. We find that the combination is a Gaussian distribution when each of the three individual classes is a Gaussian distribution, contrasting with the power-law distribution of the break-taking duration. Then we experimental analysis the distribution of the break-taking durations and frequencies, and find that, for the durations, the three single classifications can be fitted individually by an Exponential distribution and together by a Power-law distribution, for the frequencies, both the three single classifications and together can be fitted by a Gaussian distribution, so that can validate the above theoretical analysis.

Key words: break-taking behaviour, long-distance freight vehicle, statistical analysis

1. Introduction

Global Positioning System (GPS) trajectory data have triggered substantial interest in the study of behaviour patterns. Various groups have been studied, including people in general [1-3] and drivers in particular, including taxi drivers [4-6] and commercial vehicle drivers [7,8]. However, with the development of the logistic distribution network, long-distance freight vehicles are necessary to be studied. The result can be expected to be useful for transport management departments and related enterprises.

There have been substantial researches related to behaviour patterns. These include time use studies [9-12], which focus on problems related to the statistical analysis of people's behaviour patterns, and behaviour pattern analyses of various types of vehicle, for example, basic taxi driver's working status and daily taxi driver's temporal and spatial activities [13-18], as well as the business patterns of commercial vehicles [19-24]. Researches on long-distance freight vehicles mainly focus on the relationship between the driving hours or the breaking hours and the commercial truck driver safety, and these researches are usually based on the field survey data[25,26]. There are also other researches related to the sleeping condition of the driver[27]. And the researchers discuss the basic classifications of the drivers' daily behaviour based on the naturalistic data collection[28].

Compared with others, long-distance freight vehicles have some special characteristics. They transport cargo from city to city and must sometimes load or unload cargo within a transport centre. Their transit time is long; in addition to the temporal and spatial characteristics of their travel, they might sometimes participate at random times and places in specific activities, such as break taking, wherein the vehicle must stop for a period of time. Although in Guangxi Province, where the GPS trajectory data was collected, the traffic police departments take a series of measures in order to control the fatigue driving, however, in reality, for the purpose of profit, the long-distance freight vehicle drivers are generally entirely free to choose the duration, frequency, and location of their breaks, so that the existence of relevant laws and regulations have no impact on the break-taking behaviour pattern. As the characteristics of break-taking activities can reflect the temporal and spatial characteristics of the vehicle, break-taking provides an opportunity to study the behaviour pattern of the vehicle driver.

In this paper, we study the break-taking behaviour pattern of long-distance freight vehicles based on GPS trajectory data during one week in Guangxi Province, China. We take note of the stopping points, classify the break-taking activities into three classes, and then determine the characteristics of the distribution of the break-taking duration and frequency for the three classes individually. The combination of the break-taking duration follows a power-law distribution, and each of the three classes individually follows an exponential distribution. And the combination of the break-taking frequency follows a Gaussian distribution based on two principal conditions as follows. Each of the three classes follows a Gaussian distribution, and the combination of the break-taking duration follows a power-law distribution.

The structure of this paper is as follows, in Section 2, we describe the data used in this research, and then based on the large amount of trajectory data, we select and classify the stopping points, in Section 3, we theoretical analyse the relationship among the combination and the three individual classes, in Section 4, we experimental validate the above conclusion based on the statistical analysis of the data, finally we summarize the research contents and prospect the future research.

2. Analysis of Stopping Points

2.1 Data Description

This research is based on GPS trajectory data of long-distance freight vehicles from a GPS Vehicle Information Management System. That system contains a series of basic stations that record the GPS trajectory data of vehicles with a GPS sensor. The frequency is recorded every 60 s, and there are nearly 36,000 vehicles in total, including nearly 12,000 and 24,000 online and offline vehicles, respectively. In this paper, the GPS trajectory data were collected from July 6 to July 12, 2015, in Guangxi, China, for almost 3,000 vehicles in total, including information such as vehicle ID, date, time, latitude, longitude, velocity, and fuel consumption. On one hand, the selected data can basically reflect the general characteristics, on the other hand, selecting data that lasts for one week can further ensure the continuity of the data records. The quality of the dataset is relatively reliable, and we perform data checking [29,30] with error records to ensure the reliability of the analysis. The error records can mainly be divided into three cases, first, there are error records in the trajectory data records, such as the vehicle ID, date, time, longitude, latitude and speed, such records must be deleted, second, there are omissions in the trajectory data records, if the problems are among the vehicle ID, date, time, such records can be added directly, if the problems are among the longitude, latitude, speed, such records can be deleted directly, third, there are deviated records in the trajectory data records, especially the longitude and latitude, such records can be corrected by dotting on the map. Guangxi Province covers an area of nearly 240000 square kilometers, the terrain is dominated by mountainous and hilly basins, and the flat occupies 27 percent of the total area. By the end of 2012, the total mileage of the highway in Guangxi Province is 107906 km, the mileage of the expressway is 2883 km, the mileage of the

First-Grade Highway is 984 km, the mileage of the Second-Grade Highway is 9720 km, and the others are the Third-Grade or Fourth-Grade Highway, the First-Grade Highway is usually the multi-lane highway, the Second-Grade Highway is usually the two-lane highway which is special for vehicles, and the Third-Grade Highway is usually the two-lane highway which is mainly for vehicles. There are three main types of long-distance freight vehicles, such as the flat vehicles, high barrier vehicles and container vehicles, and the common tonnage ranges from two tons to about thirty tons.

2.2 Selection of stopping points

To find stopping points, we must determine whether GPS trajectory data ever have zero velocity. A GPS measurement always contains some errors; for example, a stopping point might have ten pieces of GPS trajectory data, whose velocity might not be exactly zero. Hence, we need to specify a velocity threshold to judge whether a piece of GPS trajectory data belongs to the stopping points. We can find the threshold using the three-sigma rule. Here we can illustrate the three-sigma rule as follows,

The three-sigma rule [31,32] is a basic law of mathematical statistics that can be applied to data that follow a normal distribution; hence, data must be checked to determine whether they fit such a distribution using a quantile–quantile (Q–Q) plot. Specifically in this study, it is the velocity of the GPS trajectory data that might or might not fit a normal distribution, and the data must be processed to follow that distribution. The confidence interval of the data can be assured using the three-sigma rule with x – 3*sigma and x + 3*sigma being the lower and upper threshold limits, respectively.

If the data do not fit a normal distribution, then the data must be transformed to a pseudonormal distribution to enable the three-sigma rule to be applied. The formula is

$$x^{(\gamma)} = \begin{cases} \dfrac{x^{\gamma}-1}{\gamma}, \gamma \neq 0, \\ \ln(x), \gamma = 0. \end{cases} \tag{1}$$

For the values x_1, x_2, x_3 …, and x_M, the optimization of the index r can be determined by calculating the maximum of the following formula.

$$l(\gamma) = \max\left(-\frac{M}{2}\ln\left(\frac{1}{M}\sum_{i=1}^{M}\left(x_i^{(\gamma)} - \bar{x}^{(\gamma)}\right)^2\right) + (\gamma-1)\sum_{i=1}^{M}\ln(x_i)\right) \tag{2}$$

where

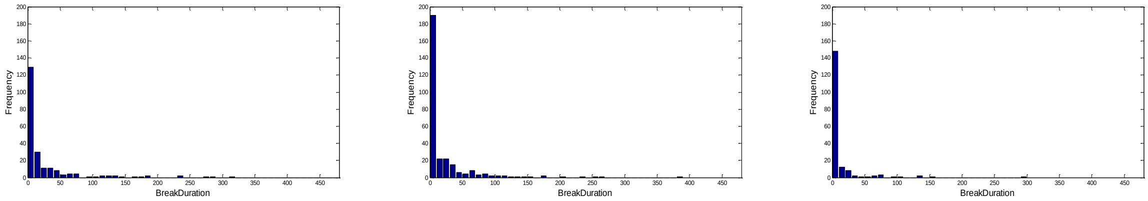
$$\bar{x}^{(\gamma)} = \frac{1}{M}\sum_{i=1}^{M}\frac{x^{\gamma}-1}{\gamma} \tag{3}$$

According to the three-sigma rule which has been mentioned above, we calculate the mean value x and the standard deviation sigma based on the data records of the velocity, and then set the x – 3sigma as the lower threshold and the x + 3sigma as the upper threshold[33]. Then, based on the threshold, we can select the GPS trajectory data whose velocity is less than the lower threshold as the possible stopping points. In addition, if the data records of the velocity can't be fitted to the normal distribution, then it can be transformed to the pseudonormal distribution according to the method mentioned above, after that the three-sigma rule can also be applied to it.

After the selection of the GPS trajectory data belonging to the stopping points, there are always some pieces of GPS trajectory data belonging to the same stopping point, recorded before and after the adjacent periods. As has been mentioned above, data regarding a stopping point includes vehicle ID, date, time, latitude, longitude, and velocity. Some pieces of GPS trajectory data can be combined in the following manner. Retain the time of the initial portion of the records as the time of the stopping point, as well as the duration. Consider the mean latitude and longitude of all portions of the records to be the latitude and the longitude of the stopping point, and the difference between the first and last portions of a record’s times to be the duration of the selected stopping point.

2.3 Classification of stopping points

Then, we detect stopping points, select three vehicles’ stopping point data, and check their break duration distribution. We divide the break duration from 0 to 480 into 10-min time intervals and statistically analyse the distribution of the break durations. As shown in Figure 1, the three single distributions have similar distribution characteristics; when the break duration changes from short to long, the frequency change from high to low with a long tail.



(a) Vehicle ID = 63014579 (b) Vehicle ID = 63014852 (c) Vehicle ID = 63016248

Figure 1. Break duration distribution of the vehicles. From the above three plots, we can determine that the three can almost be fit to the similar long-tailed distribution.

Based on the three vehicles mentioned above, we further statistically analyse the distribution of the break duration of all the vehicles. The Project “Stardriver” took advantage of a smartphone app to support driver self-observation, and finally offered six mainly measurement categories of drivers’ activities, such as drive, wait, load, unload, break and service (including fueling), combined with the actual situation[34], in general, we can divide the classification of a long-distance freight vehicle driver’s daily break-taking activities into four categories: waiting for a traffic light, having a meal (as the service’s representative), sleeping, and loading or unloading cargo. On this basis, to determine the characteristics of the distribution of the break duration of all the vehicles, we must ensure that the distributions of the break durations of every vehicle have similar characteristics using the t-test method as the significance test. Here we can illustrate the t-test method as follows,

The t-test [35,36], also known as the student's t-test, can be used to determine whether two sets of data follow the same distribution on the basis of whether there is a significant difference in the squared deviations of the two sets of data in statistical analysis. The statement of the t-test in MATLAB is [H,P,C] = ttest(x,y,ALPHA), where x and y represent the two sets of data being compared, and ALPHA is the significance level of the t-test that is always set to 0.05. H is the judgment standard of the t-test, reflecting the result; when H = 0, the null hypothesis is not rejected, and when H = 1, the null hypothesis is rejected with a confidence level of 0.05. P is the probability of the expected result, whether the null hypothesis is to be accepted or rejected. We sequentially perform a comparison of the distribution on the break duration of every vehicle. That is, if veh1, veh2, and veh3, are three vehicles, then we compare veh1 with veh2, and then veh2 with veh3, continuing in this manner until we traverse every vehicle. Here, t is the dispersion statistic, x is the sample mean, u is the population mean, δ is the sample standard deviation, and n is the sample size. The formula is as follows:

$$t = \frac{\bar{x} - \mu}{\frac{\delta}{\sqrt{n}}} \tag{4}$$

According to the t-test method which has been mentioned above, the result shows that the values of H are all zero and that the values of P are all above 0.9. Based on the t-test, we can validate that the distributions of the break duration of every vehicle have similar characteristics.

Then, we can determine the characteristics of the distributions of the break duration of all the vehicles. The results can be seen in Figure 2, where the x-axis represents the break duration and the y-axis represents the break frequency. Then, we choose four kinds of original functions for curve fitting: the exponential, Gaussian, power, and lognormal distributions. The performance of the fitting can be evaluated using the parameter R-squared, the coefficient of determination, whose value ranges from zero to one, with one representing the best curve fitting. The closer the value is to one, the stronger is the ability of x to explain y. The R-squared values of the above four distributions are 0.9875, 0.9872, 0.9971, and 0.1316, respectively, showing that the power-law distribution has the best fitting. The results are shown in Table 1.

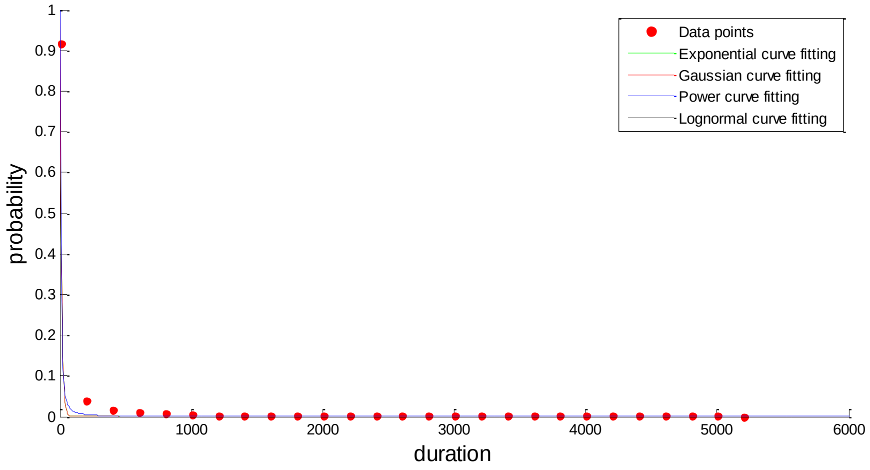


Figure 2. Break duration distribution of long-distance freight vehicles. This plot shows the distributions of the break durations fitted to four general kinds of distribution.

Table 1. Fitting results

Function	Parameters	R squared

Exponential
$f(x)=a\cdot e^{b\cdot x}$
	a = 1.046
	b = −0.08136
		0.9875

Gaussian
$f(x)=a\cdot e^{-\frac{(x-b)^2}{c^2}}$
	a = 2.866e+166
	b = −9445
	c = 482.5
		0.9872

Power
$f(x)=a\cdot x^b$
	a = 15.86
	b = −1.523
		0.9971

Lognormal
$f(x)=\frac{a}{x}\cdot e^{-\frac{(\ln x-b)^2}{c^2}}$
	a = 0.2935
	b = 0.6731
	c = 0.02893
		0.1316

As mentioned above, break-taking activities can involve waiting for a traffic light, having a meal, sleeping, and loading and unloading cargo, however, the waiting time for a traffic light is too short and the frequency of this activity is too frequent, so that we won't take it into account. Based on the distributions of the break duration, we can divide the records of the stopping points into three classes, less than 1 h, between 1 and 3 h, and more than 3 h, which corresponding to the first, second, and third classes, respectively.

3. Theoretical analysis about relationship among the combination and the three individual classes

We discuss the relationship among the combination of the three classes of the distributions on the break-taking frequency and the three individual classes themselves in the background of the distribution of the break-taking duration.

First, we should introduce some background information. $f(x)$ is the overall distribution of x, and $f(x,n)$ is the individual distribution of x. $f(x,n)$ is related to the distribution of parameter n, which can be represented as $r(n)$. The formula is

$$f(x)=\int_0^\infty r(n)\cdot f(x,n)dn \qquad (5)$$

In addition, we discuss the integral of the Gaussian distribution, whose formula is

$$\int a\cdot e^{-\frac{(x-b)^2}{c^2}}dx$$

$$\begin{aligned}&\int_0^\infty a\cdot e^{-\frac{(x-b)^2}{c^2}}dx\\ &=\int_0^b a\cdot e^{-\frac{(x-b)^2}{c^2}}dx+\int_b^\infty a\cdot e^{-\frac{(x-b)^2}{c^2}}dx\end{aligned} \qquad (6)$$

The integral from 0 to ∞ can be divided into two partitions: the integral from 0 to b and the integral from b to ∞. The fitting curve of the distribution is axisymmetric, and on the basis of the symmetry, the result of the former is non-integrable but can have a narrow value. The result of the latter is ½; hence, we can set the sum of the former and the latter to V, with no effect on the following discussion.

We also discuss the integral of the product of the exponential and power distributions, whose formula is

$$\int x^{n} \cdot e^{x} dx$$

$$\begin{aligned}&\int_{0}^{\infty} x^{n} \cdot e^{x} dx \\ &= x^{n} \cdot e^{x} - n \cdot x^{n-1} \cdot e^{x} + \cdots + \left(-1\right)^{n-1} n! \cdot x \cdot e^{x} + \left(-1\right)^{n} n! \cdot e^{x} \\ &= \left[x^{n} - n \cdot x^{n-1} + \cdots + \left(-1\right)^{n-1} n! \cdot x + \left(-1\right)^{n} n!\right] \cdot e^{x}\end{aligned}$$

$$\begin{aligned}&\int_{0}^{\infty} a_{1} \cdot x^{b_{1}} \cdot a_{2} \cdot e^{b_{2} \cdot x} dx \\ &= a_{1} \cdot a_{2} \cdot \left[x^{b_{1}} \cdot e^{b_{2} \cdot x} - b_{1} \cdot x^{b_{1}-1} \cdot e^{b_{2} \cdot x} + \cdots \left(-1\right)^{b_{1}-1} b_{1}! \cdot x \cdot e^{b_{2} \cdot x} + \left(-1\right)^{b_{1}} b_{1}! \cdot e^{b_{2} \cdot x}\right] \\ &= a_{1} \cdot a_{2} \cdot \left[x^{b_{1}} - n \cdot x^{b_{1}-1} + \cdots + \left(-1\right)^{b_{1}-1} b_{1}! \cdot x + \left(-1\right)^{b_{1}} \cdot b_{1}!\right] \cdot e^{b_{2} \cdot x}\end{aligned} \tag{7}$$

The integral from 0 to ∞ takes advantage of the basic method of integration by parts, and we can set the result to W, with no effect on the following discussion.

a) Gaussian distribution in the three individual classes of the break-taking frequency

The stopping points have been classified into three individual classes based on the distributions of the break-taking duration, and the distribution on the break-taking duration of every individual vehicle can be fitted to a similar distribution, the Power distribution. At the same time, we can suppose that the first, second, and third classes fit a similar distribution tendency, might to be the Exponential distribution. Then, based on the relationship among $\mathrm{f(x)}$, $\mathrm{r(n)}$, and $\mathrm{f(x,n)}$, corresponding here to $\mathrm{E(t)}$, $\mathrm{r(f)}$, and $\mathrm{P(t,f)}$, $\mathrm{E(t)}$ represents the break-taking duration distribution of every class, $\mathrm{r(f)}$ represents the break-taking frequency distribution of the three classes, and $\mathrm{P(t,f)}$ represents the break-taking duration distribution of every individual vehicle, relating to the break-taking frequency distribution. The relationships among $\mathrm{E(t)}$, $\mathrm{r(f)}$, and $\mathrm{P(t,f)}$ are as follows, a_1, b_1, a_2, and b_2 are the corresponding parameters of the different distributions, and the specific values of the parameters have no effect on the discussion. Thus, we can determine the expression of $\mathrm{r(f)}$ as follows. V has been discussed above; a_3, b_3, and c_3 are the corresponding parameters of the distribution, and t can be seen as a constant. We can determine that $\mathrm{r(f)}$ is the Gaussian distribution, corresponding to the Gaussian distribution in the three individual classifications of the break-taking frequency:

$$E(t)=\int_0^\infty r(f)\cdot P(t,f)df$$

$$a_1\cdot e^{b_1\cdot t}=\int_0^\infty r(f)\cdot a_2\cdot t^{b_2}df$$

$$a_1\cdot e^{b_1\cdot t}=\int_0^\infty r(f)df\cdot a_2\cdot t^{b_2}$$

$$\int_0^\infty r(f)df=\frac{a_1\cdot e^{b_1\cdot t}}{a_2\cdot t^{b_2}}$$

$$\int_0^\infty a_3\cdot e^{-\frac{(f-b_3)^2}{c_3^2}}\cdot\frac{1}{V}\cdot\frac{a_1\cdot e^{b_1\cdot t}}{a_2\cdot t^{b_2}}df=\frac{a_1\cdot e^{b_1\cdot t}}{a_2\cdot t^{b_2}}$$

$$r(f)=a_3\cdot e^{-\frac{(f-b_3)^2}{c_3^2}}\cdot\frac{1}{V}\cdot\frac{a_1\cdot e^{b_1\cdot t}}{a_2\cdot t^{b_2}} \quad (8)$$

b) Gaussian distribution in the combination of the three classifications of the break-taking frequency

Because the distribution on the break-taking duration of every individual vehicle can be fitted to the Power distribution, we can suppose that the distribution of the break-taking duration of the combination of the three classifications also can be fitted to the power distribution. Similarly, based on the relationship among $\mathrm{f(x)}$, $\mathrm{r(n)}$, and $\mathrm{f(x,n)}$, corresponding here to $\mathrm{G(f)}$, $\mathrm{r(t)}$, and $\mathrm{G(f,t)}$, $\mathrm{G(f)}$ represents the distribution of the break-taking frequency of the combination of the three classes, $\mathrm{r(t)}$ represents the break-taking duration distribution of every individual vehicle, and $\mathrm{G(f,t)}$ represents the break-taking frequency distribution of the three classes. The relationship among $\mathrm{G(f)}$, $\mathrm{r(t)}$, $\mathrm{G(f,t)}$ is as follows, a_1, b_1, a_2, b_2, a_3, and b_3 are the parameters, as we have discussed, and the specific values of the parameter have no effect on the discussion. Thus, we can determine the expression of $\mathrm{G(f)}$ as follows. W has been discussed previously; a_4 and b_4 are the corresponding parameters of the distribution, and t can also be seen as a constant. We can determine that $\mathrm{G(f)}$ is the Gaussian distribution, corresponding to the Gaussian distribution of the combination of the three classes of the break-taking frequency:

$$G(f)=\int_0^\infty r(t)\cdot G(f,t)dt$$

$$G(f)=\int_0^\infty a_4\cdot t^{b_4}\cdot a_3\cdot e^{-\frac{(f-b_3)^2}{c_3^2}}\cdot\frac{1}{V}\cdot\frac{a_1\cdot e^{b_1\cdot t}}{a_2\cdot t^{b_2}}dt$$

$$G(f)=\int_0^\infty a_4\cdot t^{b_4}\cdot\frac{a_1\cdot e^{b_1\cdot t}}{a_2\cdot t^{b_2}}dt\cdot\frac{1}{V}\cdot a_3\cdot e^{-\frac{(f-b_3)^2}{c_3^2}}$$

$$G(f)=\frac{W}{V}\cdot a_3\cdot e^{-\frac{(f-b_3)^2}{c_3^2}} \tag{9}$$

In conclusion, we can demonstrate that the distribution of the combination of the three classes of the break-taking frequency can be fitted to the Gaussian distribution. At the same time, the facts that the distribution of the three individual classes of the break-taking frequency can be fitted to the Gaussian distribution and that the distribution of the break-taking durations can be fitted to the power distribution explain the relationship among the combination and the three classes of the break-taking frequency as well as the reason for the emergence of the combination.

4. Experimental validation about the distribution regulation of the break-taking durations and frequencies

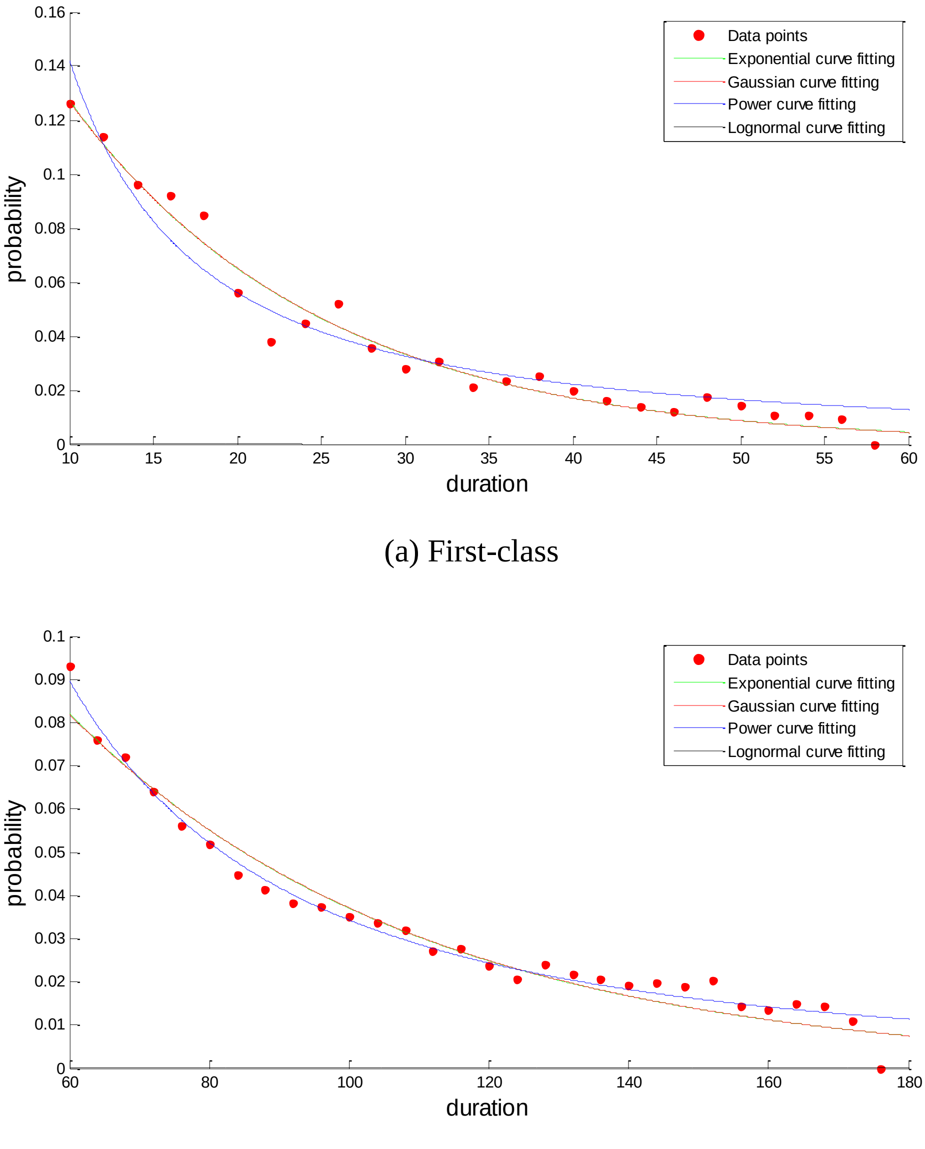
In the previous section, we theoretical demonstrate the relationship between the distribution of the combination of the three classes and the distribution of the three classes about the break-taking frequencies. And in this section, we will provide the experimental validation based on the actual GPS trajectory data.

4.1 Analysis of the distributions of the break-taking durations

In this research, the break-taking activities of vehicles has been divided into three classes based on the break duration, and we statistically analysed the characteristics of the distributions of the duration based on the combination of the three classifications and the first-, second-, and third-class break-taking activities; then, the corresponding fitting results were provided.

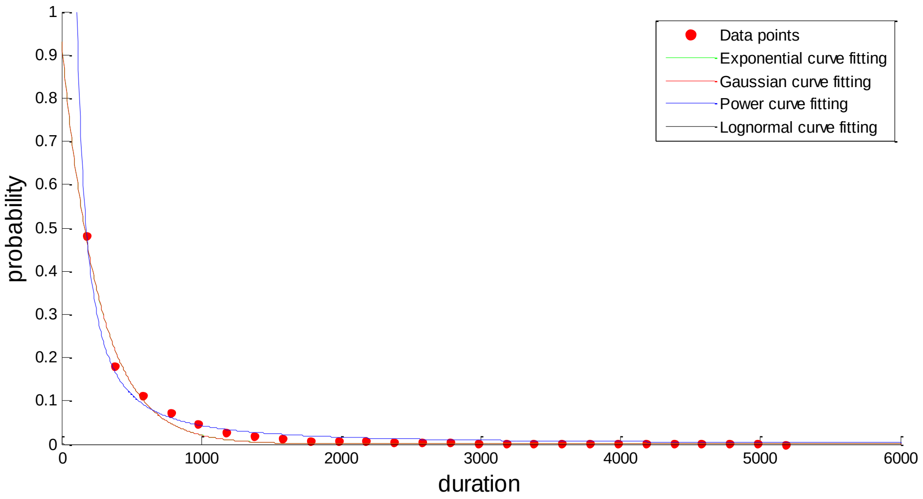
Through the statistical analysis, we can see that of the three classes, the most common break-taking durations are 10 to 22 min, 60 to 84 min, and 180 to 980 min, respectively, based on the threshold of no less than 50% of the total. Moreover, no more than 5% of the total break-taking duration is more than 48 min, 160 min, and 1,780 min, and no more than 15% of the break-taking duration is more than 40 min, 144 min, and 1,180 min. In addition, to reflect the degree of dispersion of the break-taking duration, we can determine the standard deviations, which are 4.1092, 17.7993, and 409.6710, respectively, for the three classes.

Based on the three classes, we fit the distributions of the break-taking duration of the three classifications and their combination [37,38] to exponential, Gaussian, power, and lognormal distributions. The fitting results are as follows. As discussed above, the distribution of the combination can be fitted to the power distribution with the R-squared nearly equal to one. Then, we determine that the distributions of the three classes can be fitted to the exponential distribution, as shown in Figure 3. The fitting results of the exponential and Gaussian distributions show no significant difference. However, the corresponding R-squared values reveal that the exponential distribution is better, which is consistent with the previous theoretical demonstration. And the fitting results of the others can be seen in Table 2, with a 95% confidence bound. From the fitting result above, we can determine that parameters a and b are increasing over the three classes, showing an increase over the break-taking duration.



(a) First-class

(b) Second-class



(c) Third-class

Figure 3. Vehicle driver's break-taking duration distribution of the three classes. These plots show the distributions of the break-taking durations of the three classes, respectively, that are fitted to four general kinds of distribution.

Table 2. Fitting results

First class		
Function	**Parameters**	**R squared**

Exponential	a = 0.2481
	b = −0.06688
		0.9672

Gaussian	a = 2.636e+23
	b = −1681
	c = 226
		0.9654

Power	a = 3.107
	b = −1.34
		0.9416

Lognormal	a = 0.9355
	b = 0.4618
	c = 0.4541
		0.5160

Second class		
Function	**Parameters**	**R squared**

Exponential	a = 0.2711
	b = −0.0199
		0.9608

Gaussian	a = 5.707e+25
	b = −6212
	c = 797.7
		0.9588

Power	a = 196.1
	b = −1.878
		0.9504

Lognormal	a = 0.92
	b = 0.3832
		0.5326

Third class		
Function	**Parameters**	**R squared**

Exponential	a = 0.9304
	b = −0.003792
		0.9843

Gaussian	a = 5.076e+245
	b = −2.99e+05
	c = 1.257e+04
		0.9836

Power	a = 757.1
	b = −1.415
		0.9634

Lognormal	a = 0.9161
	b = 0.6693
	c = 0.4174
		0.2549

4.2 Analysis of the distributions of break-taking frequencies

4.2.1 Distribution of the mixture

Based on the analysis of the stopping points, we fitted the distribution of the break-taking frequency, regardless of the three classes, to exponential, Gaussian, power, and lognormal distributions and determined that the combination can be fitted to the Gaussian distribution better than the other three classes individually. The results reveal the break-taking frequency distribution and the best-fitted distribution, as well as the other three fitted distributions, as shown in Figure 4. Data points are represented by solid points, and distributions are represented by solid lines.

The values of R-squared are shown in Table 3. We can see that the Gaussian distribution is significantly better fitted than the other three because its R-squared is very close to one, which is consistent with the previous theoretical demonstration. However, in addition, we can also discuss three other parameters that relate to curve fitting and that can reflect the correctness of data fitting, sum squared error (SSE), mean squared error (MSE), and root MSE (RMSE). Here we can illustrate the three parameters as follows,

The three parameters are the sum of squares, the mean sum of squares, and the root mean sum of squares, respectively; y_i represents the actual data, and $\hat{y_i}$ represents the fitted data. The number of corresponding data points is n, and w_i is the weight coefficient specifying the extent to which the data volume of a particular corresponding data point accounts for the data volume overall. The formulas are

$$SSE = \sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2 \tag{10}$$

$$MSE = SSE/n = \frac{1}{n}\sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2 \tag{11}$$

$$RMSE = \sqrt{MSE} = \sqrt{SSE/n} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2} \tag{12}$$

According to the three parameters which have been mentioned above. They can also measure the error between the original data and the fitting data; the closer the value is to zero, the better the curve fits the data, and the values can also be seen in Table 3.

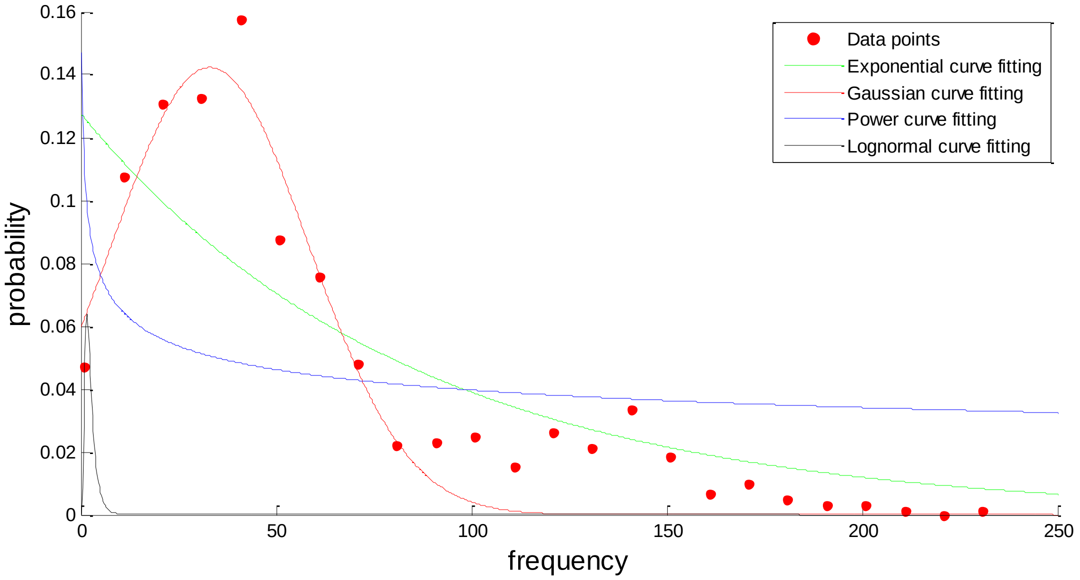


Figure 4. Vehicle driver's break-taking frequency distribution for the combination. This plot shows the distribution of the break-taking frequency fitted to four general kinds of distributions.

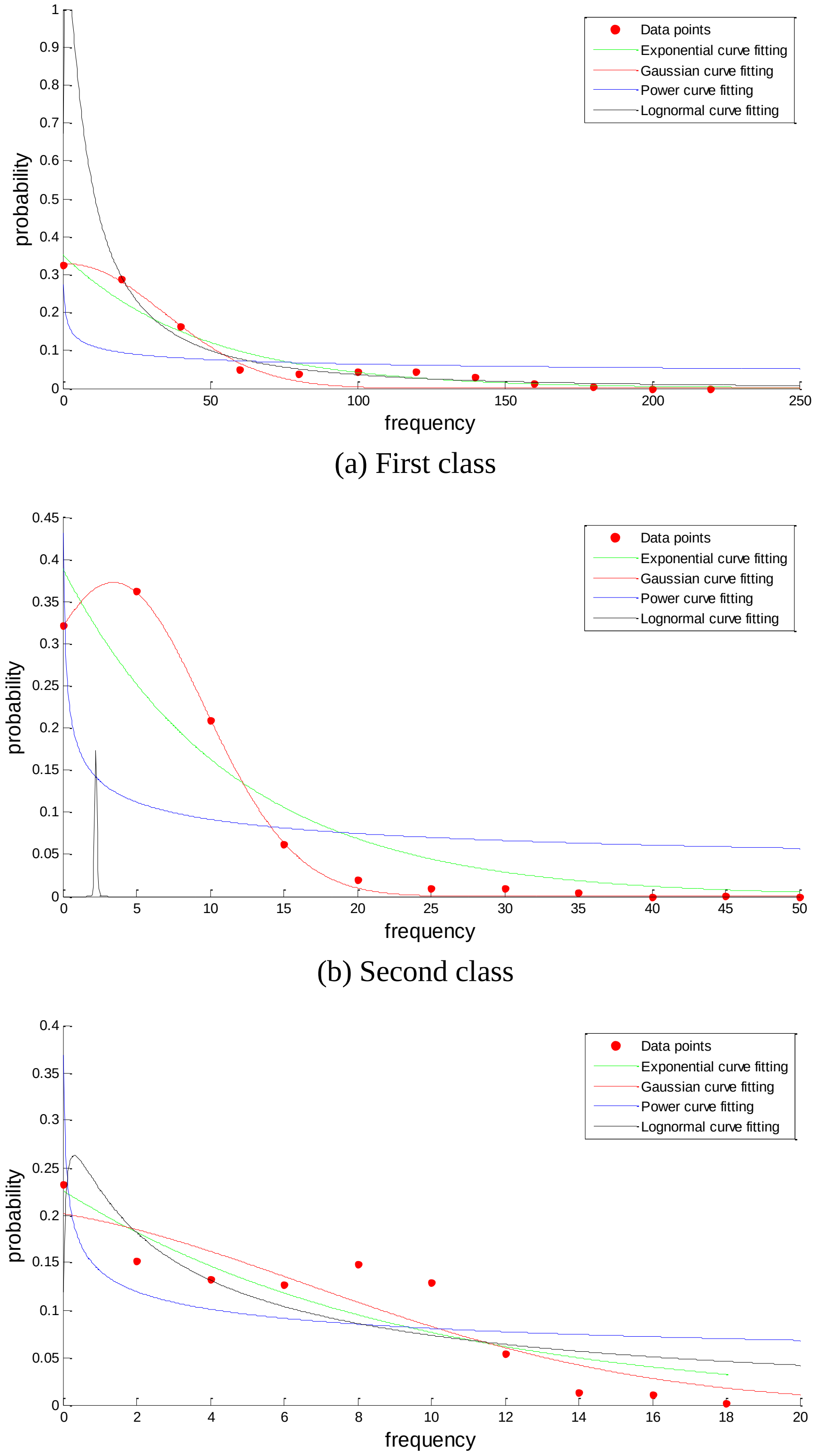
Table 3. Fitting results

Function	Parameters	R squared	SSE	MSE	RMSE
Exponential	a = 0.1276 b = −0.01185	0.6267	0.0185	0.1703	0.0290
Gaussian	a = 0.1423 b = 33.1 c = 35.49	0.8932	0.0051	0.1245	0.0155
Power	a = 0.1085 b = −0.2185	0.1475	0.0423	0.2095	0.0439
Lognormal	a = 0.1156 b = 0.7496 c = 0.7956	0.0734	0.0914	0.2569	0.0660

4.2.2 Distributions of the three individual classes

The distributions of the break-taking frequency of the three individual classes can also be fitted to the Gaussian distribution significantly better than to the other three distributions discussed above, which is consistent with the previous theoretical demonstration. The fitting results can be seen in Figure 5. The best-fitted distribution is represented by the red line, and the other three distributions are represented by lines of other colours.

The fitting parameters can be seen in Table 4 and the value of R-squared for the Gaussian distribution is obviously better than that for the other three distributions, as shown in Table 4, where the three parameters SSE, MSE, and RMSE are also listed.



(a) First class

(b) Second class

(c) Third class

Figure 5. Vehicle driver's break-taking frequency distribution of the three single classes. The above three plots show distributions of the break-taking frequency of the three classes, respectively, fitted with four general kinds of distribution.

Table 4. Fitting results

First class					
Function	**Parameter**	**R squared**	**SSE**	**MSE**	**RMSE**
Exponential	a = 0.3523 b = −0.02123	0.9389	0.0078	0.1673	0.0280
Gaussian	a = 0.3282 b = 1.801 c = 46.2	0.9528	0.0054	0.1568	0.0246
Power	a = 0.1961 b = −0.2426	0.5481	0.0579	0.2759	0.0761
Lognormal	a = 5.867 b = 2.993 c = 2.323	0.9461	0.0028	0.1323	0.0175
Second class					
Function	**Parameter**	**R squared**	**SSE**	**MSE**	**RMSE**
Exponential	a = 0.3888 b = −0.08662	0.8569	0.0247	0.2289	0.0524
Gaussian	a = 0.3726 b = 3.434 c = 8.719	0.9981	0.0003	0.0781	0.0061
Power	a = 0.1797 b = −0.2927	0.4117	0.1017	0.3260	0.1063
Lognormal	a = 0.3928 b = 0.8071 c = 0.05323	0.1584	0.2830	0.4337	0.1881
Third class					
Function	**Parameter**	**R squared**	**SSE**	**MSE**	**RMSE**
Exponential	a = 0.2262 b = −0.1084	0.7865	0.0099	0.1876	0.0352
Gaussian	a = 0.2106 b = −2.77 c = 13.21	0.8004	0.0081	0.1847	0.0341
Power	a = 0.1416 b = −0.2447	0.5182	0.0224	0.2302	0.0530
Lognormal	a = 0.8649 b = 3.554 c = 3.074	0.6596	0.0139	0.2110	0.0445

5. Conclusion and Discussion

In this study, we find that it is of value to perform a statistical analysis of the behaviour pattern of break-taking activity based on GPS trajectory data of long-distance freight vehicles. We take advantage of various indices that enable us to determine the characteristics of the break-taking behaviour pattern, including break-taking duration and frequency, and to simultaneously deepen our understanding of the vehicle's work status, which is significantly different from the work status of taxis or other vehicles. The results show that the distribution of the break-taking duration of the combination can be fitted to a power distribution and that the distribution of the break-taking duration of the three individual classes can be fitted to an exponential distribution. In addition, the results show that the distribution of the break-taking frequency of the combination can be fitted to a Gaussian distribution and that the distribution of the break-taking frequency of the three individual classifications can be fitted to a Gaussian distribution.

Similarly, the above mentioned distributions show a degree of normality, reflecting the clustering of the two vehicle's indices, and the explanation for the emergence of the break-taking frequency distribution of the combination might be that the distributions of the three classes on the basis of the power distribution for the break-taking duration distribution can also reflect the basic relationship among them.

In conclusion, the study shows a process of data mining from GPS trajectory data and reflects a series of characteristics of the vehicles' break-taking behaviour pattern, which can be helpful for the study of behaviour patterns and traffic management.

On the premise of ensuring that the overall characteristics can be fully reflected, here we only utilise 3000 vehicles for this study, in the future research, it is possible to further extend the amount of data. If the research takes advantage of the data that collected in the case of different geographical environment or different laws and regulations, then there may be a distribution with different characteristics, and in the analysis of its formation causes, the factors of the geographical environment and related laws and regulations can be considered, and this will be a very interesting research. As the development of the technology, the data types of the GPS trajectory data of the long-distance freight vehicles will be more and more abundant, such as the vehicle's condition and land-use attribution, so that the future work can be to bridge the knowledge of the behaviour pattern of the vehicles with these reflected by the new data types. And the related researches will have important significance for the intelligent transport systems, such as assist traffic management departments to better design the national network of rest stops, predict vehicle drivers' behaviour pattern or optimise traffic management systems.

Acknowledgments

This research was supported in part by the National Natural Science Foundation of China under Grant nos. 61672082 and U1564212.

References

[1] Gonzalez, M. C., Hidalgo, C. A., Barabasi, A. L.: Understanding individual human mobility pattern, Nature, 2008, 453, pp. 779-782

[2] Williams, N. E., Thomas, T. A., Dunbar, M., Eagle, N., Dobra, A.: Measures of human mobility using mobile phone records enhanced with GIS data, PLoS One, 2015, 10(7)

[3] Gao, S.: Spatio-temporal analytics for exploring human mobility patterns and urban dynamics in the mobile age, Spatial Cognition & computation, 2015, 15(2), pp. 86-114

[4] Zheng, Y., Xie, X.: Inferring a behavioral state of a vehicle. United States Patent US 8543320 B2, 2013

[5] Gong, L., Liu, X., Wu, L., Liu, Y.: Inferring trip purposes and uncovering travel patterns from taxi

trajectory data, Cartography and Geographic Information Science, 2016, 43(2), pp. 103-114

[6] Hoque, M. A., Hong, X. Y., Dixon, B.: Analysis of mobility patterns for urban taxi cabs. Proc. Int. Conf. Computing, Networking and Communication, Maui, Hawaii, January 2012, pp. 956-760

[7] Joubert, J. W.: Inferring commercial vehicle activities from GPS data. Proc. Int. Conf. Swiss Transport Research, Monte Verità, May 2012, pp. 1-10

[8] Figliozzi, M. A.: Analysis of the efficiency of urban commercial vehicle tours data collection, methodology, and policy implications, Transportation Research Part B Methodological, 2007, 41(9), pp. 1014-1032

[9] Joyce, M., Stewart, J.: What can we learn from time use data, Monthly Labor Review, 1999, 122(8), pp. 3-6

[10] Rogger, K. K.: Variations in time use at stages of the life cycle, Monthly Labor Review, 2005, 128(9), pp. 38-45

[11] Kalenkoski, C. M., Pabilonia, S. W.: Time to work or time to play: the effect of student employment on homework, housework, screen time, and sleep, Bureau of Labor Statistics, 2009, 19(2), pp. 29

[12] Reifschneider, M. J., Hamrick, K. S., Lacey, J. N.: Exercise, eating patterns, and obesity: evidence from the ATUS and its eating & health module, Social Indicators Research, 2011, 101(2), pp. 215-219

[13] Liu, L., Andris, C., Ratti, C.: Uncovering cabdrivers' behavior patterns from their digital traces, Computers Environment & Urban Systems, 2010, 34(6), pp. 541-548

[14] Xu, Y., Shaw, S. L., Chen, J.-L., Li, Q.-Q., Fang, Z.-X., Li, Y.-G.: Uncover repeated spatio-temporal behavioral patterns embedded in GPS-based taxi tracking data. Proc. Int. Conf. GIScience, OH, USA, 2012, pp. 1-7

[15] Zong, F., Sun, X., Zhang, H.-Y., Zhu, X.-M., Qi, W.-T.: Understanding taxi drivers' multi-day cruising patterns, Traffic & Management, 2015, 27(6), pp. 467-476

[16] Yuan, J., Zheng, Y., Xie, X., Sun, G.-Z.: T-Drive: enhancing driving directions with taxi drivers' intelligence, IEEE Transactions on Knowledge & Data Engineering, 2013, 25(1), pp. 220-232

[17] Yuan, J., Zheng, Y., Zhang, L.-H., Xie, X.: T-Finder: a recommender system for finding passengers and vacant taxis, IEEE Transactions on Knowledge & Data Engineering, 2013, 25(10), pp. 2390-2403

[18] Ma, S., Zheng, Y., Wolfson, O.: T-Share: a large-scale dynamic taxi ridesharing service. Proc. Int. Conf. Data Engineering, Brisbane, Australia, 2013, pp. 1-12

[19] Sharman, B. W., Sc, M. A., Roorda, M. J.: Multilevel modelling of commercial vehicle inter-arrival duration using GPS data, Transportation Research Part E Logistics & Transportation Review, 2013, 56(9), pp. 94-107

[20] Gliebe, J., Cohen, O., Hunt, J. D.: Dynamic choice model of urban commercial activity patterns of vehicles and people, Transportation Research Report Journal of the Transportation Research Board, 2007, 2003(1), pp. 17-26

[21] Sharman, B. W., Roorda, M. J.: Analysis of freight GPS data: a clustering approach for identifying trip destinations, Transportation Research Record, 2011, 1277, pp. 83-91

[22] Roorda, M. J., Kwan, H., Maccabe, S.: Comparison of GPS and driver-reported urban commercial vehicle tours and stops. Proc. Int. Conf. Transportation Research Board, Denver, Colorado, 2008, pp. 1-20

[23] Holguin, V. J., Thorson, E.: Trip length distributions in commodity-based and trip-based freight demand modeling investigation of relationships, Transportation Research Record, 2000, 1707(1), pp. 37-48

[24] Sharman, B. W., Roorda, M. J., Habib, K. N.: Comparison of parametric and non-parametric hazard models of stop durations on urban commercial vehicle tours, Transportation Research Record, 2012, 2269, pp. 117-126

[25] Chen, C., Xie, Y.: Modeling the safety impacts of driving hours and rest breaks on truck drivers

considering time-dependent covariates, Journal of Safety Research, 2014, 51, pp. 57–63

[26] Chen, C., Xie, Y.: The impacts of multiple rest-break periods on commercial truck driver's crash risk, Journal of Safety Research, 2014, pp. 48:87

[27] Pylkkönen, M., Sihvola, M., Hyvärinen, H. K., Puttonen, S., Hublin, C., Sallinen, M. Sleepiness, sleep, and use of sleepiness countermeasures in shift-working long-haul truck drivers, Accident; analysis and prevention, 2015, 80, pp. 201-210

[28] Soccolich, S. A., Blanco, M., Hanowski, R. J., et al.: An analysis of driving and working hour on commercial motor vehicle driver safety using naturalistic data collection, Accident; analysis and prevention, 2013, 58(5), pp. 249

[29] Famili, F., Shen, W.-M., Weber, R., Simoudis, E.: Data preprocessing and intelligent data analysis, Intelligent Data Analysis, 1997, 1(1), pp. 3-23

[30] Runkler, T. A.: Data analytics models and algorithms for intelligent data analysis (Springer Vieweg press, Heidelberg, 2012)

[31] Zhang, Z.-S., Yang, D.-G., Zhang, T., He, Q.-C., Lian, X.-M.: A study on the method for cleaning and repairing the probe vehicle data, IEEE Transactions on Intelligent Transportation Systems, 2013, 14(1), pp. 419-427

[32] Box, G. E. P., Cox, D. R.: An analysis of transformations. Journal of the Royal Statistical Society Series B Methodological, 1964, 26(2), pp. 211-252

[33] Pukelsheim, F.: The Three Sigma Rule, American Statistician, 1994, 48(48), pp. 88-91

[34] Prockl, G., Sternberg, H.: Counting the Minutes—Measuring Truck Driver Time Efficiency, Transportation Journal, 2015, 54(2), pp. 275-287

[35] Surhone, L. M., Timpledon, M. T., Marseken, S. F.: Student's T-Test (Betascript press, 2010)

[36] Delaney, H. D.: The Effect of Nonnormality on Student's Two-Sample T Test, Monte Carlo Methods, 2000, pp. 30

[37] Motulsky, H., Christopoulos, A.: Fitting models to biological data using linear and nonlinear regression, a practical guide to curve fitting (GraphPad Software Inc press, California, 2003)

[38] Newman, M.: Power laws, Pareto distributions and Zipf's law, Contemporary Physics, 2004, 46(5), pp. 323-351