RESEARCH ARTICLE

Research on construction of vehicle driving cycle based on Markov chain and global K-means clustering algorithm

Yingnan Wu, Guangzhong Liu

College of Information Engineering, Shanghai Maritime University, Shanghai, China

Abstract: Vehicle driving cycle is a time-speed curve used to describe vehicle driving rules. The research and development of vehicle driving cycle not only provide theoretical basis for the test of vehicle fuel's economy and pollutant emission level, but also guide the design and development of new models in the future. This paper adopts the actual driving data of light vehicles in Fuzhou City, Fujian Province collected by China Automotive Technology Research Center (CATC) through the data collection system, and analyzes and verifies the new data after standardized dimension reduction by combining the global K-means clustering and Markov chain principle. The specific work is divided into the following parts: 1. The global K-means clustering algorithm adopted to cluster the kinematic segment database after standardized dimension reduction; 2. Markov chain is applied to construct the working condition diagram. The basic principle of this method is to regard the short-stroke speed-time sequence as a complete random process, divide the speed intervals by lines, each of which represents a different speed state, and convert the speed into a speed state, so that the speed-time sequence becomes a state-time sequence. Since the next state is only related to the current one, a group of random state sequences can be randomly generated by the program as long as the transition probability between two adjacent states is determined and the matrix of state transition probability is established. 3. The state sequence is converted into a speed sequence, and finally a set of driving cycle conforming to the spatial characteristics of the samples is obtained.

Keywords: Vehicle driving cycle; factor analysis; global K-means clustering algorithm; Markov chain

1. Introduction

In recent years, the number of passenger cars in China has increased rapidly. This popular transportation mode provides great convenience for our life and also brings great opportunities for our industrial development. But what goes along is the trouble and harm to human life. Most of the supplementary fuels for automobiles are gasoline and diesel, and the consumption of oil is increasing along with the number of automobiles, leading to the depletion of energy. In addition, harmful substances such as CO, PM and CO_2 that produce greenhouse effect in automobile exhaust emissions not only do harm to human body, but also seriously affect air quality.

Vehicle driving cycle is a time-speed curve used to describe the rules of driving cycle of a certain type of vehicle (passenger car, bus, heavy vehicle, *etc.*) under a specific traffic environment[1], which is an important feature reflecting the emission of vehicle pollutants. Reflecting the kinematic characteristics of the vehicle driving on the road, the driving cycle is an important and common basic technology in the automobile industry, the basis of testing methods and limit standards of vehicle energy consumption/emission, and also the main benchmark for calibration and optimization of various performance indexes of the vehicle. At present, developed countries including European countries, the United States and Japan have adopted standards suitable for their own driving cycle for calibration and optimization of vehicle

Copyright © 2020 Yingnan Wu and Guangzhong Liu
doi: 10.18063/vd.v4i1.1135
This is an open-access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

performance and certification of energy consumption/emission.

From changes in road traffic conditions in our country, the government, enterprises and the public have found out that the actual fuel consumption of vehicles optimized and calibrated based on NEDC is more and more deviated from the certification results of laws and regulations, thus affecting the credibility of the government. The two most important operating driving cycles, idling time ratio and the average speed of test cycle of the world light vehicles, are more different from the actual vehicle driving cycle in China. As the most basic basis for vehicle development and evaluation, it is becoming more and more important to carry out in-depth research and formulate test cycle that reflect the actual road driving cycle in our country.

This paper, adopting the actual driving data of light vehicles in Fuzhou City, Fujian Province collected by China Automotive Technology Research Center (CATC) through the data collection system and combining the global K-means clustering and Markov chain principle, tries to construct the vehicle driving cycle and evaluates the model structure.

2. Definition of kinematic segments and determination of characteristic parameters

Kinematic segment refers to the speed interval between the start of the first idle state and that of the next, as shown in **Figure 1** (Building a curve of driving cycle based on kinematic segment is one of the most commonly used methods at present, but not a necessary step. Some methods for building a curve of driving cycle do not require dividing and extracting kinematic segments).

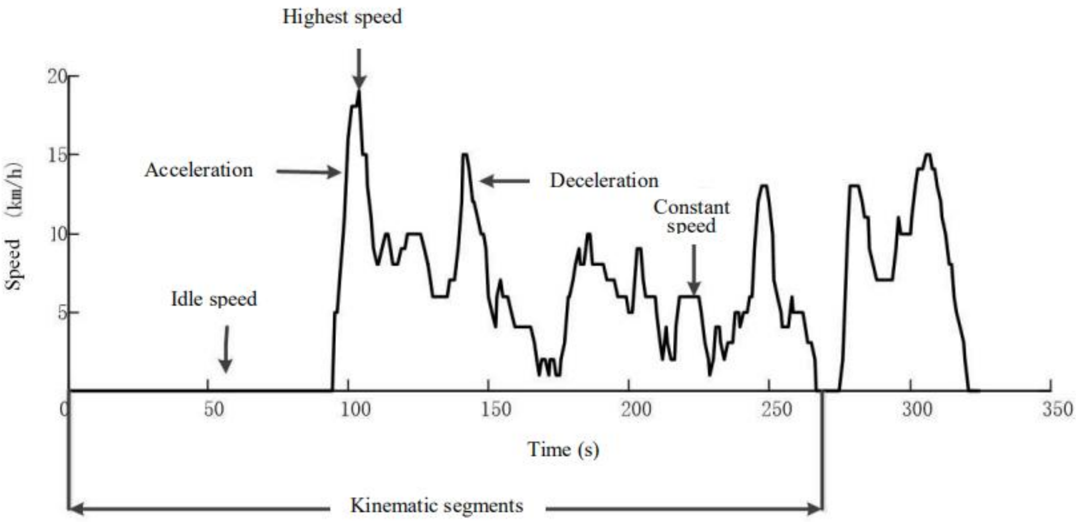

Figure 1. Definition of kinematic segments

In order to select a suitable representative driving cycle and truly reflect the actual driving cycle on the road, nine characteristic parameters are defined as evaluation criteria (as shown in **Table 1**), and the representative driving cycle with the smallest average relative error with the characteristic parameters of the experimental data is selected from all combined short trips.

No.	Characteristic Parameter	Unit	Meaning
1	V_a	km/h	Average speed
2	V_b	km/h	Average travel speed
3	a_i	m/s^2	Average acceleration
4	a_j	m/s^2	Average deceleration
5	P_a	%	Idle speed time ratio
6	P_b	%	Acceleration time ratio
7	P_c	%	Deceleration time ratio
8	V_{sd}	km/h	Standard deviation of speed
9	a_{sd}	m/s^2	Standard deviation of acceleration

Table 1. Characteristic parameters of kinematic segment

The characteristic parameters of each kinematic segment are extracted to obtain a characteristic parameter matrix.

3. Theoretical basis of global K-means and Markov chain

3.1 Global K-means clustering principle

The K-means algorithm tries to find the K partitions that minimize the square error function through fast iteration. When the result clusters are compact and obviously separated from each other, its effect is better. K-means can quickly converge to the local minimum through iteration, which is simple in structure and short in operation time. When the processed data set is large, it is of flexibility and high efficiency. However, the main disadvantage of this method is that it is sensitive to the initial points of clustering[2].

The main principle of global K-means algorithm is to realize global optimization based on increment. Instead of depending on the initial value, the clustering center point selected by the algorithm calculates the function value of clustering square error of the known clustering center point and other data points, and then selects the data point with the smallest function value as the best initial clustering center for the next cluster[3].

3.2 Markov chain

Markov property was first proposed by the Russian mathematician Markov, which means that given the current state, the conditional probability distribution of the future state of a system or stochastic process depends only on the current state. The property unrelated to the previous state is Markov property or without aftereffect. The discrete-time stochastic process with Markov property is Markov process. Markov process has three characteristics: discreteness, randomness and free of aftereffect. Markov prediction method applies probability to predict random time sequence[4].

Given the initial distribution of Markov chain, its statistical characteristics are completely determined by the conditional transition probability. Therefore, the transition probability P_{ij} between states is calculated according to Markov principle, which indicates the conditional probability of state i of state j, thus the statistical characteristics of experimental data can be determined. The state transition probability matrix can be formed by the state transition probability:

$$p_{k\times k}=\begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1k} \\ p_{21} & p_{22} & \cdots & p_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ p_{k1} & p_{k2} & \cdots & p_{kk} \end{bmatrix}_{kxk}$$

According to the maximum likelihood function, for a fixed Markov process, the formula of state transition probability is as follows:

$$P_{ij}=\frac{N_{ij}}{\sum_{s} N_{ij}}$$

where N_{ij} is the number of events that the state i at time of τ -1 changes to the state j at time of τ [5].

4. Analysis on principal components

In the process of constructing vehicle driving cycle, kinematic segments need to be clustered according to the characteristic parameter values above. However, among the selected characteristic parameters, some variables are not independent from each other and of certain redundancy, which makes the information expressed overlap. On the other hand, if only one or two characteristic parameters are analyzed, some information will be lost. Therefore, the factor analysis model must be substituted into the characteristic parameter matrix to reduce the dimension of the data. Factor analysis is a multivariate statistical analysis method, the core idea of which is data transformation and dimensionality reduction. First, complex variables are integrated into a few main factors, and then the problem is interpreted or comprehensively evaluated[6]. It can explain most information of original variables with a few potential factors[7]. According to **Table 2** and **Table 3**, three principal components, average travel speed V_b, average acceleration a_i and deceleration time ratio P_c are obtained.

Ingredients	Initial characteristic value			Extract squares and load			Rotate squares and load		
	Total	Variance (%)	Accumulation (%)	Total	Variance (%)	Accumulation (%)	Total	Variance (%)	Accumulation (%)
1	4.009	44.534	44.534	4.009	44.543	44.543	3.310	36.780	36.780
2	3.416	37.954	82.497	3.416	37.497	82.497	2.581	28.672	65.453
3	1.042	11.573	94.071	1.042	94.071	94.071	2.576	28.618	94.071
4	0.534	5.929	100						
5	4.371E-16	4.856E-15	100						
6	2.811E-16	3.123E-15	100						
7	2.589E-17	2.877E-16	100						
8	-2.961E-17	-3.290E-16	100						
9	-1.628E-16	-1.809E-15	100						

Table 2. The total variance interpreted

	Ingredients		
	1	2	3
Average speed (*km/h*)	0.473	0.824	-0.311
Average travel speed (*km/h)*	0.529	0.839	-0.079
Average acceleration (m/s^2)	0.944	-0.082	0.287
Average deceleration (m/s^2)	-0.514	0.519	0.368
Idle speed time ratio (%)	-0.390	0.687	0.519
Acceleration time ratio (%)	-0.926	-0.115	-0.316
Deceleration time ratio (%)	0.290	-0.791	0.537
Standard deviation speed (*km/h*)	0.917	0.387	0.025
Standard deviation of acceleration (m/s^2)	0.645	-0.705	-0.251

Table 3. Coefficient matrix of component score

5. Construction of vehicle driving cycle

5.1 Specific model

Step One: Firstly, the extracted kinematic segments are clustered by global K-means algorithm. The specific steps are as follows[8]:

Step1 [Initialization] Calculate the mean value of all sample data as the center of the first cluster, $m_1 = \frac{1}{n}\sum_{i=1}^{n} x_i$,set $q = 1$;

Step2 [Termination conditions], $q = q + 1$,if $q > k$, then the algorithm is terminated;

Step3 [Find the next cluster center] Taking $m_1, m_2, \cdots, m_{q-1}$ as the centers of the previous $q-1$ clusters, each sample x_i in the data set as the initial center of the q_{th} cluster respectively, the K-means algorithm is executed. The sample x_i making E smallest is selected as the best initial center of the q_{th} . cluster to obtain the new centroid of each cluster $y_1, y_2, \cdots, y_q$;

Step4. Making $m_i = y_i, i = 1,2,\cdots, q$, turn to Step2.

Step Two: The basic principle of constructing representative driving cycle based on Markov chain random process is to regard the short-stroke speed-time sequence as a complete random process, divide the speed intervals by lines, each of which represents a different speed state, and convert the speed into a speed state, so that the speed-time sequence becomes a state-time sequence. Since the next state is only related to the current one, as long as the transition probability between two adjacent states is determined, a set of random state sequences is randomly generated by the program, then the state sequences are converted into speed sequences, and finally a group of driving cycles conforming to the spatial characteristics of the samples is obtained[9].

After obtaining all probability matrices of state transition, representative driving cycles are respectively generated for each kind, and then combined into representative driving cycles with a total duration of 1200-1300 s. The time proportion of each kind is the same as that in the database. The specific operation steps are as follows:

Step1. Taking the beginning of the state as the current state, the next state value is determined by the probability matrix of state transition, generating sufficient time in the same way and returning to the state 1.

Step2. Convert the state sequence of step one into a speed sequence according to the following formula:

$$v_i = [(S_i - 1) + rd]\Delta d$$

Among them:

Δd : The speed space of the current state;

rd : Random numbers obeying evenly distribution, the value of which ranges [0,1];

S_i : The state value of time i;

v_i : the speed value corresponding to the state S_i .

Step3. Calculate characteristic values for the obtained speed sequence, compare the results with the class statistical results. If the deviation value of average absolute value is less than 10%, it is qualified, and the iteration is pushed out; otherwise, it returns to step1.

In the paper, there are nine characteristic parameters selected, including average speed, average travel speed, average acceleration, average deceleration, idle speed time ratio, acceleration time ratio, deceleration time ratio, standard deviation of speed and standard deviation of acceleration. The formula for deviation of average absolute value is as follows:

$$b = \frac{\sum_{k=1}^{n} |C_k - G_k|}{n}$$

Among them:

C_k : The k-th characteristic value of the candidate driving cycle;

G_k : The statistical result of the k-th characteristic value of the class database;

n: The number of characteristic variables, n = 9;

b: The deviation value obtained after calculation.

5.2 Experimental results

Using the global K-means clustering algorithm model mentioned above, the data matrix with 3 principal components is clustered and analyzed by MATLAB, and the short-run samples can be classified into 3 categories. The classification effect can be directly observed from **Figure 2**.

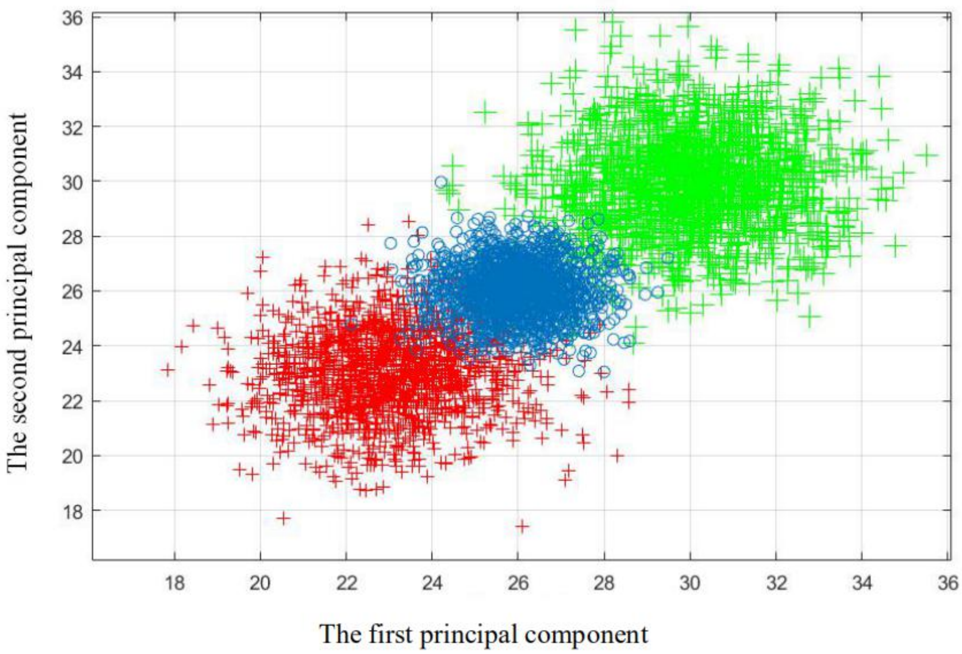

Figure 2. Scatter distribution of samples after clustering

As seen from the figure, the boundary between the adjacent two types is clear, sample points of the first two classes are relatively concentrated, while the third class is slightly loose and of a large span. On the whole, all kinds of scatters can be distinguished well, and the classification result is acceptable.

Category	V_a	V_b	a_i	a_j	P_a	P_b	P_c	V_{sd}	a_{sd}
Class one	22.2	24.2	0.32	-0.39	15	42	33	3.44	0.45
Class two	26.1	28.6	0.49	-0.44	13	45	28	3.26	0.42
Class three	32.5	35.7	0.55	-0.52	11	46	24	2.94	0.36

Table 4. Average characteristic values of various classes after clustering

According to the above table, speed and acceleration of the class one are all low, so they are low-speed segments; data of the class two is in the middle, which are medium-speed segments; speed and acceleration of the class three are all high, which are high-speed segments.

Subsequently, the clustering results are substituted into Markov state transition model, obtaining the results by running the program with MATLAB.

The obtained representative driving cycles are combined in order of class 1, class 2 and class 3 to finally obtain **Figure 3,** the joint probability distribution diagram of speed and acceleration of the representative driving cycles, and **Figure 4**, the representative driving cycles.

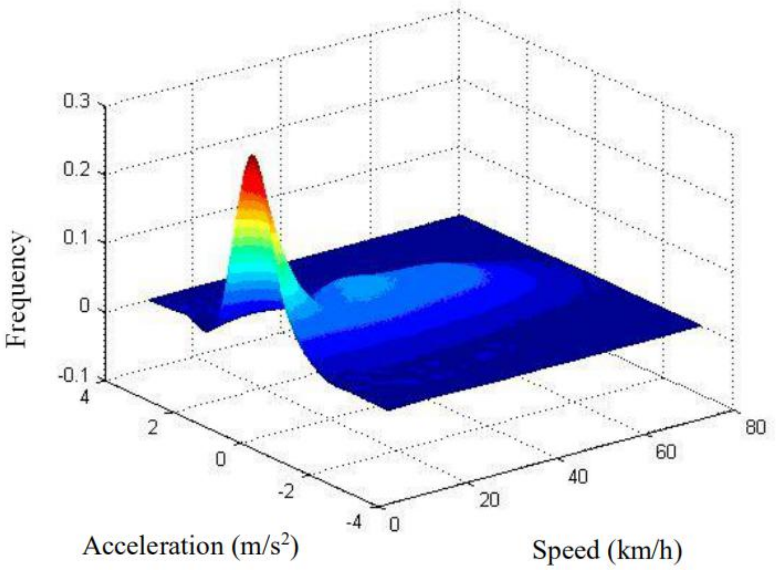

Figure 3. The joint probability distribution diagram of speed and acceleration of the representative driving cycle

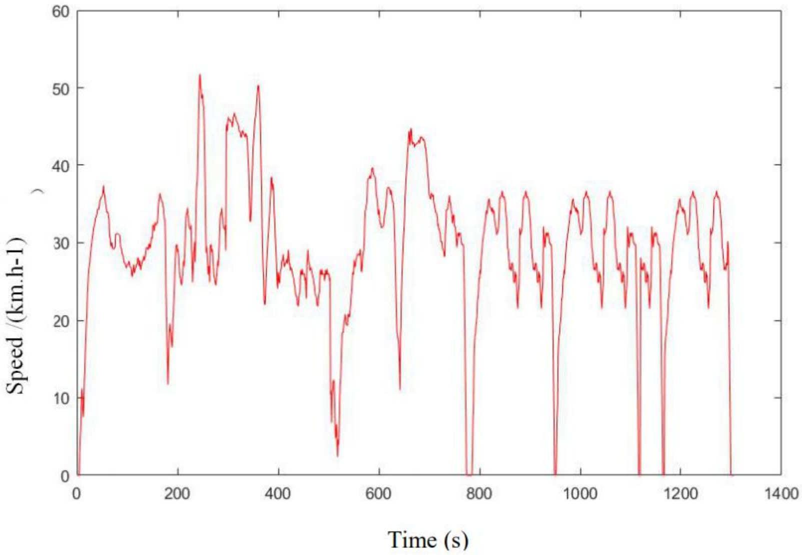

Figure 4. The representative driving cycles

6. Verification and analysis of driving cycles

In order to verify that the constructed driving cycles can represent the corresponding characteristics of the collected data sources (processed data), the statistical results of the characteristic parameters of the two are compared, and the results are shown in **Table 5**.

Index of characteristic value	Constructed driving cycles	Original data	Error(%)
Average speed (km/h)	48.86	49.89	2.06
Average travel speed (km/h)	52.34	53.44	2.06
Average acceleration (m/s^2)	0.438	0.448	2.23
Average deceleration (m/s^2)	-0.436	-0.427	2.11
Idle speed time ratio (%)	11.8	11.2	5.36
Acceleration time ratio (%)	42.6	41.7	2.16
Deceleration time ratio (%)	30.2	29.5	2.37
Standard deviation of speed (km/h)	3.29	3.34	1.50
Standard deviation of acceleration (m/s^2)	0.472	0.479	1.46

Table 5. Comparison of constructed driving cycle and collected data

The maximum deviation value is 5.36%, far less than 10%. Therefore, the constructed driving cycles can accurately reflect the overall characteristics.

References

1. Wan X, Huan W, Qiang M. Construction of driving cycle for passenger vehicles in Shenzhen. Journal of Shenzhen University (Science & Engineering) 2016; 33(3): 281-287.
2. Xie H, Zhang P, Luo S. Spectral clustering based on global K-means. Journal of Computer Applications 2016; 33(3): 281-287.
3. Gu X, Xu F, Yang Y, Qu F. Analysis of college students' achievement based on global K-means algorithm. Journal of Changchun University of Science and Technology: Natural Science Edition 2019; 42(5): 93-97.

4. Xiang X, Zhang L. Forecasting cotton price based on Markov chain. China Cotton 2016; 43(10): 1-6.
5. Gao J. Construction of vehicle driving cycle based on K-means clustering algorithm. Journal of Henan Polytechnic University (Natural Science) 2019; 38(1): 112-118.
6. Liu Y, Lu Y. Evaluation of road traffic safety and decision-making research based on factor analysis. Safety and Environmental Engineering 2009; 16(6): 112-114.
7. Zhang L, Zhang X. Weighted clustering based on improved CRITIC. Statistics and Decision 2015; 22: 65-68.
8. Xie J, Jiang S, Wang C, Zhang Y, Xie W. An improved global K-means clustering algorithm. Journal of Shaanxi Normal University (Natural Science Edition) 2010; 38(2): 18-22.
9. Cao Q, Li J, Liu Y, Qu D. Construction of driving cycle based on Markov chain for passenger car in Changchun City. Journal of Jilin University 2018; 48(5): 1366-1373.