Construction of Vehicle Operating Mode Based on K-Means Algorithm

Liping Yuan¹* and Yang Li²

¹School of Information Engineering, Wuhan HuaXia University of Technology, Wuhan 430223, China.
²Wuhan Maritime Communication Research Institute, Wuhan 43025, China.
Email: 94524948@qq.com

Abstract. The vehicle driving condition (Driving Cycle), also called the vehicle test cycle, is to describe the speed-time curve of the vehicle driving process, which can reflect many kinematic characteristics of the vehicle’s exercise process on the road. At present, the operating conditions of automobile used in China can not meet the requirements of certification and calibration of automobile products, so it is urgent to develop the test conditions that can reflect the characteristics of automobile exercise in China. Using the method of neural network training to amplify the data, we can accurately fill the missing data in the course of driving find out the principal component of the kinematics segment of the vehicle by using principal component analysis, select some of the most expressive kinematics fragments by using the K-means clustering (K-Means Cluster) analysis method, and construct the model of the Vehicle working condition with little error from the real operating condition, thus effectively reduce the error caused by various environment, traffic, road surface and other external factors.

1. Introduction
The driving condition of automobile can be used to study the road condition, weather and other factors in the driving process, which has far-reaching significance for vehicle oil consumption, pollution generation, transportation economy and road traffic safety [1]. Therefore, it is urgent to study and establish an effective driving condition of urban vehicles. At present, most regions, including China, adopt European driving conditions and formulate relevant laws. However, ECE (European urban conditions) is a single process, which can not perfectly deal with the complex and changeable driving conditions of Chinese urban vehicles and carry out corresponding vehicle development. At present, many domestic scholars have built their own car conditions suitable for China. In this paper, the main idea is to combine the principal component analysis method and K-means method to complete the task of modeling the car driving conditions from the collected driving data, and complete the preprocessing of the bad data in the original data and the extraction of kinematic segments.

2. Data Preprocessig
In the process of data acquisition, some bad data may be generated due to human factors, road conditions, instrument accuracy or GPS drift and other factors. The deletion or completion of these dirty data is of great significance for the later kinematic segment extraction and further construction of vehicle driving conditions, so as to facilitate the subsequent statistical analysis of sample data. "Bad" data mainly includes several types: when a car passes through a high-rise building or a tunnel, the time discontinuity in the data caused by GPS signal loss [2]. Due to the existence of GPS drift and other factors, the acceleration is abnormal. The speed or vehicle speed caused by long-term parking is
abnormal. In addition, the situation of long-time traffic jam and intermittent low-speed driving can usually be treated as idle speed, which is abnormal if the idle time exceeds 180 seconds. A car was used to drive on the roads of Fuzhou and Putian in Fujian Province, and the data were collected according to different routes and different times.

2.1. Pretreatment of Time Discontinuity
Based on the analysis of the geographical location of GPS data, it is found that there is no tunnel more than 60 seconds' drive in the test site, while the high-rise building has an impact on GPS, but the impact should be less than 60 seconds. It can be assumed that when the signal is lost for more than 60 seconds, the vehicle is in the phase of lights out and stops. When the signal of the vehicle is lost within 3 seconds, since the vehicle can be considered to make a uniform acceleration or deceleration movement within 3 seconds, the speed in the disappearance period can be inferred from the speed in the first second before the signal loss and the speed in the second after the signal loss. When the car signal disappears for more than 3 seconds and less than 60 seconds, the missing data can be amplified by neural network learning.

2.2. Pretreatment of Abnormal Acceleration and Deceleration of Automobile
When the data acceleration exceeds the general performance index of the vehicle, it is usually due to GPS drift, resulting in abnormal acceleration of the vehicle. Generally, the drift is instantaneous, and the corresponding algorithm can be designed, that is, the abnormal acceleration data can be corrected by the acceleration at the front and back time. Due to the constant acceleration of the car in 3 seconds, the abnormal value can be repaired with the normal acceleration of the car in front and back seconds.

2.3. Pretreatment of Long-term Parking Problems
In the data, there is a period of time when the speed is 0 for a long time, but the car is still running, which shows that the car is in a process of parking without turning off, in reality, it is often a process of waiting for traffic lights. For this state, it can be considered to be in idle state.

3. Kinematical Segment Extraction
Kinematical segment refers to the speed range between the beginning of idling state and the beginning of next idling state. Accurate kinematical segment extraction is often helpful to build the working condition curve of automobile. The key to extract kinematic segment[3] is to be able to preprocess the original data reasonably and judge the idling range of the vehicle. Firstly, find out the point where the vehicle's GPS speed is 0 and the vehicle's speed is not 0 in the previous second, and judge whether the speed is 0 in consecutive seconds. If it is 0, it is regarded as the idle state, the beginning of a kinematic segment and the end of the previous kinematic segment. Otherwise, continue to search until the conditions are met.

In the data preprocessing process, the long-time traffic jam or continuous low-speed driving (the maximum speed is less than 10km/h) is regarded as idle speed, and in the process of low-speed driving, if the speed changes to 10km/h for only one second or several seconds, it is still regarded as idle speed. While the idle speed is generally higher than 10km/h in the first second, and lower than 10km/h in the next few seconds, and the acceleration is lower than 0, it can be regarded as the beginning of an idle interval, and the starting position of each two idle intervals is a kinematic segment, and its extraction flow is shown in Figure 1.
Figure 1. Flow chart of kinematic segment extraction.

Through this method, all kinematic segments can be found out, and the acceleration and deceleration caused by traffic jam can be eliminated to a large extent by judging whether the vehicle speed in consecutive seconds is less than 10km/h or equal to 0, which is regarded as idle speed. However, there are some defects in this method. For example, when waiting for a traffic light intersection with a long queue, there may be a situation that the vehicle speed is longer than the judgment time and more than 10km/h, but after the time, the vehicle will stop to wait for the traffic light. Therefore, for roads with different complexity, it is necessary to select an appropriate continuous judgment time.

4. The Construction of Automobile Driving Condition

4.1. Establishment of Vehicle Driving Mode Model

The kinematic segments extracted from the data file are not all in line with the requirements of building the working condition curve. It is necessary to extract the representative kinematic segments, and then combine a working condition curve through the combination of working conditions, compare with the actual measured data, and judge the rationality of the driving condition after calculating the error.

In the process of vehicle driving, some basic characteristic parameters can be selected to reflect the driving characteristics of each kinematic segment. Because the dimension of each feature parameter is not uniform, it may cause the problem of too much data dispersion, so it is necessary to standardize the feature parameters extracted from each kinematic segment. In the process of using feature parameters to describe kinematic segments, multiple feature parameters ensure that the feature information of segments will not be lost to some extent, but will inevitably bring about information overlap, so principal component analysis is needed[4]. By using principal component analysis, we can get the principal component which can reflect most of the parameters in the model, thus greatly reducing the dimension of the data, that is, to reduce the dimension of the characteristic parameters.

Through clustering analysis of these principal components[5], all kinematic segments can be divided into several categories, and the distance between each kinematic segment and its focus can be
obtained. Then, some kinematic segments with the smallest focus distance in each category are selected to form a working condition to synthesize an alternative set, because the distance between each kinematic segment and the focus reflects the distance between the kinematic segment and this type of row. The smaller the distance is, the better the data can reflect the characteristics of the class.

Finally, select a few suitable segments from the selection and combination set of motion conditions, which can well reflect the overall characteristics of this exercise. The algorithm flow is shown in Figure 2.

4.2. Solution of Vehicle Driving Condition Model

In the calculation process of characteristic parameters, due to different dimensions, the value of each variable will lead to a large difference in the distribution interval of the numerical value, which is relatively scattered. In the process of processing, different segments are taken as rows, and different characteristic parameter variables are taken as columns, which are put into a matrix for standardization, so that the variables with large standard deviation will not be taken into account in the later principal component analysis and cluster analysis, the result is stable.

\[
A_{mv} = \begin{pmatrix}
  a_{11} & \cdots & a_{1n} \\
  \vdots & \ddots & \vdots \\
  a_{m1} & \cdots & a_{mn}
\end{pmatrix}
\]  

(1)
Equation (1) is the characteristic parameter matrix of the kinematic segment, \(a_j (i = 1, 2, \cdots m; j = 1, 2, \cdots n)\) refers to the jth characteristic parameter variable of the ith kinematic segment. The standardized matrix \(X\) is obtained by standardizing the matrix \(A_{\text{max}}\).

\[
X_{\text{max}} = \begin{pmatrix}
  x_{1j} & \cdots & x_{mj} \\
  \vdots & \ddots & \vdots \\
  x_{nj} & \cdots & x_{nn}
\end{pmatrix}
\]

(2)

\[
x_{ij} = a_j - \frac{\bar{a}_j}{S_j} \]

(3)

\[
\bar{a}_j = \frac{1}{m} \sum_{i=1}^{m} a_j
\]

(4)

\[
S_j^2 = \frac{1}{m-1} \sum_{i=1}^{m} (a_{ij} - \bar{a}_j)
\]

(5)

\(x_{ij}\) represents the element value after standardization, \(S_j^2\) represents the standard deviation, \(\bar{a}_j\) represents the average value of each column in \(A\) matrix.

In the process of dimensionality reduction, the covariance matrix \(T\) is calculated from the matrix \(X\), as shown in formula (6).

\[
T = \begin{pmatrix}
  S_1^2 & \text{cov}(1,2) & \cdots & \text{cov}(1,n) \\
  \text{cov}(1,2) & S_{21} & \cdots & \text{cov}(2,n) \\
  \vdots & \vdots & \ddots & \vdots \\
  \text{cov}(n,1) & \text{cov}(n,2) & \cdots & S_n^2
\end{pmatrix}
\]

(6)

Where \(S_x^2 = \text{cov}(x, x)\), and \(\text{cov}(x, y) = \text{cov}(y, x) = \frac{1}{m-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})\).

Then the correlation matrix \(R\) is obtained from the matrix \(A\).

\[
R = \frac{1}{m-1} \begin{pmatrix}
  r_{11} & r_{12} & \cdots & r_{1n} \\
  r_{21} & r_{22} & \cdots & r_{2n} \\
  \vdots & \vdots & \ddots & \vdots \\
  r_{n1} & r_{n2} & \cdots & r_{nn}
\end{pmatrix}
\]

(7)

\[
r_{ij} = \frac{\text{cov}(x, y)}{S_x S_y}
\]

(8)

The correlation matrix \(R\) is obtained from the matrix \(A\), and \(P_i\) is the characteristic parameter of the matrix \(R\). Suppose \(\frac{\sum_{i=1}^{m} P_i}{\sum_{j=1}^{n} P_j}\) is the contribution rate of the \(i\)th principal component, the greater the contribution rate is, the more information it carries about the kinematic segment. When several principal components are selected and its cumulative contribution rate is greater than 80%, it is considered that it roughly contains all the information of the kinematic segment. By using SPSS software to analyze the dimensionality reduction factors of the standardized characteristic parameters, the table of variance contribution shown in Table 1 can be obtained.
Table 1. Extract the sum of load squares.

| component | characteristic value | Variance percentage | Cumulative contribution rate (%) |
|-----------|----------------------|---------------------|-----------------------------------|
| 1         | 3.797                | 29.210              | 29.210                            |
| 2         | 2.269                | 17.453              | 46.663                            |
| 3         | 1.620                | 12.460              | 59.123                            |
| 4         | 1.346                | 10.351              | 69.474                            |

It can be seen from table 1 that the cumulative contribution rate of variance of the first four components has reached 69%, which can better represent most of the characteristic information. Therefore, the first four components are taken as the main components.

In the matrix of principal components, principal component 1 The parameter values of average speed, acceleration time, deceleration time and standard deviation of speed are relatively high, which means that principal component 1 can well represent the characteristic parameters of average speed, acceleration time, deceleration time and standard deviation of speed; similarly principal component 2 can well represent idle time ratio, acceleration time ratio and deceleration time ratio; principal component 3 can well represent acceleration and deceleration; principal component 4 can well represent acceleration and deceleration. The characteristic parameters, such as average velocity and standard deviation of acceleration, are well expressed.

Table 2. Composition matrix.

| Composition | 1       | 2       | 3       | 4       |
|-------------|---------|---------|---------|---------|
| Average speed $v_a$ | .738    | -.075   | -.113   | -.471   |
| Idle speed time ratio $p_i$ | -.634   | .589    | .220    | -.058   |
| Acceleration time ratio $p_a$ | .390    | -.713   | -.193   | -.014   |
| Deceleration time ratio $p_d$ | .168    | -.682   | -.229   | .342    |
| Average running speed $v_{ma}$ | .528    | .157    | -.180   | -.628   |
| Acceleration time $T_a$ | .812    | .085    | .377    | .145    |
| Idle Time $T_i$ | -.085   | .440    | .379    | -.145   |
| Deceleration Time $T_d$ | .781    | .087    | .398    | .192    |
| Standard deviation of speed $v_{a\sigma}$ | -.697   | .283    | -.090   | -.263   |
| Acceleration $a$ | -.065   | .387    | -.673   | .165    |
| Standard deviation of acceleration $a_{a\sigma}$ | .503    | .450    | -.151   | .518    |
| Deceleration $v_d$ | -.011   | -.353   | .708    | .064    |
| Deceleration Standard Deviation $d_{a\sigma}$ | .610    | .389    | -.084   | .401    |

According to the scores of main components in Table 2, K-means cluster analysis[6] was carried out to select representative kinematic segments. After clustering the principal components, the distance between the kinematic segments and the clustering centers in each classification can be sorted according to the size of the kinematic segments and the clustering centers contained in each classification and the calculated Euclidean distance $d_{ij}$ of the degree of affinity. The smaller the distance between the kinematic segment data and the assembly point is, the more representative the kinematic
segment is to the assembly point class data. According to the sorting results, the most representative 10 kinematic segment data of all classes can be added to the vehicle tooling synthesis alternative set. When there are more kinematic segments in a category, the 10 segments with the smallest \( d_{ij} \) in the category are taken as candidates. When there are fewer kinematic segments in a category, all the kinematic segments are taken as candidates. Finally, it is necessary to select representative segments from candidate segments for synthesis to generate driving conditions. In this paper, the acceleration time ratio, deceleration time ratio, uniform speed time ratio, idle speed time ratio, average speed, average acceleration and other variables are taken as the parameters in the screening process, and the simulation \([7-8]\) is carried out within the allowable error range to synthesize the driving condition of the car as shown in Figure 3.

According to figure 3, 12 characteristic parameters are selected for comparison between the actual driving condition of the vehicle and the driving condition established by the model. As shown in Table 3, the relative error between the established driving condition model and the actual driving condition is almost 5%, which indicates that the similarity of the model is very high.

### Table 3. Error table of model establishment.

| Characteristic parameter                  | Modeling condition | Actual working | relative error |
|------------------------------------------|--------------------|----------------|----------------|
| Average speed (km/h)                    | 26.97              | 25.65          | 0.049          |
| Average running speed (km/h)            | 40.64              | 39.62          | 0.025          |
| Average acceleration \((m/ s^{2})\)     | 1.39               | 1.29           | 0.072          |
| Average deceleration \((m/ s^{2})\)     | 1.64               | -1.65          | 0.007          |
| Idle time ratio(%)                      | 32                 | 34             | 0.059          |
| Acceleration time ratio(%)              | 37                 | 35             | 0.054          |
| Deceleration time ratio(%)              | 31                 | 32             | 0.034          |
| Standard deviation of speed (km/h)      | 8.02               | 7.62           | 0.052          |
| Standard deviation of acceleration \((m/ s^{2})\) | 2.45               | 2.67           | 0.089          |

### 5. Conclusion

In this paper, the construction method of urban vehicle driving condition is studied. Urban vehicle driving condition is the investigation of real-time vehicle data status, and the analysis of experimental data. Firstly, the approximate driving route of the vehicle is determined on the map, so that we can
more accurately preprocess the data, including time invariant continuity, abnormal acceleration, long-term parking and idle processing. Then, using the classical K_means clustering method, 12 eigenvalues of the main kinematics segments are selected for calculation, which can construct a curve of vehicle driving conditions in a short time and better reflect the characteristics of vehicle movement.

6. Acknowledgments
The authors wish to thank the support of the Huaxia University laboratory.

7. References
[1] PENG HanRui. A construction method suitable for urban mixed road driving conditions[J]. Tianjin: Tianjin Science and Technology, 2019: 46(12):39-44
[2] Mahayadin A R, Shahriman A B, Hashim M S M, et al. Efficient methodology of route selection for driving cycle development[C]. Journal of Physics Conference Series. 2017:012082.
[3] SHI Qin, ZHENG YuBo, JIANG Ping. Research on Driving Conditions of Urban Roads Based on Kinematics Fragments[J]. Automotive Engineering, 2011, 33(3):256-261.
[4] ZENG Xiaorong, KONG LingWen, YANG XueYi. Application of Principal Component Analysis in Vehicle Driving Conditions[J]. Automotive practical technology, 2014(5):5-9.
[5] PENG YuHui, YANG HuiBao, LI MengLiang. Research on Construction Method of Urban Road Vehicle Driving Conditions Based on K-means Cluster Analysis[J]. Automotive Technology, 2017(11):13-18.
[6] Bauer C G, Forest T W, Bauer C G, et al. Effect of hydrogen addition on the performance of methane-fueled vehicles. Part II: Driving cycle simulations[J]. International Journal of Hydrogen Energy, 2001, 26(1):71-90.
[7] Adak P, Sahu R, Elumalai S P. Development of emission factors for motorcycles and shared auto-rickshaws using real-world driving cycle for a typical Indian city.[J]. Science of the Total Environment, 2016, 544:299-308. D.V. Annenko. MIPRO, Opatija, Croatia, 2013[C]. Croatia: MIPRO, 2013.
[8] Perhinschi M G, Marlowe C, Tamayo S, et al. Evolutionary algorithm for vehicle driving cycle generation[J]. Journal of the Air & Waste Management Association, 2011, 61(9):923-31. Suryadiputra Liawatimena, Agung Trisetyarso. 2018 Indonesian Association for Pattern Recognition International Conference, Indonesia, 2018[C]. Indonesia: IEEE, 2018.