Transportation Research Part D 53 (2017) 78–87

Contents lists available at ScienceDirect

Transportation Research Part D

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

Building Tianjin driving cycle based on linear discriminant analysis

Zhecheng Jing *, Guolin Wang, Shupei Zhang, Chengqun Qiu

School of Automotive and Traffic Engineering, Jiangsu University, Zhenjiang 212013, China

ARTICLE INFO

Article history:
Available online 18 April 2017

Keywords:
Driving cycle
Linear discriminate analysis
Two-class Fisher's discriminant analysis
Cluster analysis
Electric vehicle

ABSTRACT

Driving cycles are standardized measurement procedure for the certification of vehicles' economy and emission, and could help evaluate driving distance and new vehicular technologies. Thus driving cycle is always a hot research topic in vehicle industry. Linear discriminant analysis is a typical multivariate statistical method which has been used in many fields such as geology and economics in recent years, but its application to driving cycles is scarce. In this paper, Tianjin driving cycle is developed by using linear discriminant analysis. The effectiveness of the developed driving cycle is confirmed by comparing the parameter of the driving cycle and real-world driving data and evaluating the economy of electric vehicle. The uniqueness of this methodology is also discussed compared with traditional methodology in cycle development. This research could offer a new methodology for building driving cycles and has reference value to related researches.

1. Introduction

Due to the rapidly increasing number of vehicles, their emission has become the dominant contributing source of air pollution globally today. Effectiveness of efforts on quantifying the emissions depends on accurate emission models. Driving cycle is an important concept in emission estimation models (Amirjamshidi and Roorda, 2015). The main purpose of a driving cycle is to simulate the real traffic condition to test vehicle exhaust emissions and fuel consumption. A driving cycle is a series of representative speed-time data for an area (Brady and O'mahony, 2016); it covers main characteristics of traffic stream in certain region and is built by data analysis. As a core technology of automotive industry, driving cycle is used in but not limited to emission test, new car evaluation and traffic control risk assessment. In China, New European Driving Cycle (NEDC) is the official driving cycle of vehicle emission test. However, NEDC was formulated on the basis of the traffic characteristic of European cities in 1990s, which showed a low degree of similarity with China, and after decades the discrepancy is becoming larger (Ma et al., 2004). Thus, it is necessary to build a driving cycle adapted to China's traffic situation and here we develop a Tianjin driving cycle for driving cycle case study.

Cluster analysis is the main method analyzing the driving data currently (Ho et al., 2014), which is based on a distance metric measuring similarity or dissimilarity of the samples (Rosati et al., 2017). Cluster analysis is used to aggregate micro-trips with similar characteristics into groups, and the type of traffic conditions each group corresponds to is determined by analyzing the groups' characteristic parameters (Shi et al., 2011). There are two main types of cluster analysis, hierarchical cluster and quick cluster. Hierarchical cluster is usually used the number of categories the micro-trips should be classified

* Corresponding author at: School of Automotive and Traffic Engineering, Jiangsu University, Zhenjiang 212013, China.
E-mail address: jingzc@foxmail.com (Z. Jing).

http://dx.doi.org/10.1016/j.trd.2017.04.005
1361-9209/

into is unknown (Demirci et al., 2017), otherwise quick cluster is usually applied (Liu et al., 2015). However, even if the groups are obtained, the traffic condition each group corresponds to is still to be determined by analyzing the parameters of each group. However, the parameters may not correspond to any particular real traffic condition, and the groups may not be required in the study. So the cluster result may confuse the researchers if the characteristics of the groups are not clear, which may lead to time wasting and delay.

Different from cluster analysis, discriminant analysis is a statistical technique for classifying an observation into one of several a priori groups that depend upon the observation's individual characteristics (Glen, 2001), it is a very popular and reliable statistical method in classifying where the dependent variable appears in qualitative form and has been successfully applied to a variety of practical problems, such as economy (Altman, 1968), geography (D'archivio and Maggi, 2017), chemistry (Sheini et al., 2016) and biology (Petrescu and Ilia, 2016), since its first application in the 1930s (Fisher, 1936). Researchers could design targeted experiments to collect cycle data corresponding to the traffic type they are aiming at, so that traffic characteristics can be obtained with less data. And as a result, a driving cycle could be established with lower cost and less time. In this study, a driving cycle from real-world driving characteristics in Tianjin suburb area is developed after analyzing the cycle data by discriminate analysis, and is validated by testing the economy of electric vehicles.

2. Methodology

Cycle data classification is a process of allocating the database into two or more classes, and it is a key step to develop a driving cycle. Linear discriminant analysis (LDA) is a multivariate statistical method to discriminate samples. The basic idea of LDA is to find the optimal discriminant vectors so that the projected sets of samples have the largest possible between-class separation and the smallest distance ratio of within-class (Safo and Ahn, 2016). Common LDA includes distance discrimination, Bayesian discrimination analysis and Fisher's discrimination. The driving cycle focused in this paper includes two states of traffic which are congestion and smooth, thus the two-class Fisher's discrimination is most suitable for this issue, and the two classes are congestion traffic condition and smooth traffic condition respectively.

Two-class Fisher's discrimination uses all variables of different attributes of two classes to establish a discriminant function: $y = c_1x_1 + c_2x_2 + \cdots + c_px_p$, substitute the p variables of a new sample into the function to get the y_x, and then compare y_x with the discriminant critical value y_0 to determine which class this new sample belongs to. The detailed procedure of two-class Fisher's discrimination is as follows:

2.1. Establishment of discriminant function

n_1 and n_2 samples are sampled from two populations G_1 and G_2, with each sample having p variables $x_{ki}^{(1)}$ and $x_{ki}^{(2)}$ $(k = 1, 2, \ldots, n_1(n_2); i = 1, 2, \ldots, p)$. For a new sample x, it is difficult to use $x_{ki}^{(1)}$ and $x_{ki}^{(2)}$ to determine its attribution. Therefore, we need to construct a new variable y, so that y can separate the two classes G_1 and G_2. For a sample x with p variables, the variable y is generally expressed as

$$y = c_1x_1 + c_2x_2 + \cdots + c_px_p \tag{1}$$

y is a composite indicator composed of linear combinations of x_i $(i = 1, 2, \ldots, p)$, and it is defined as a linear discriminant function, while $c_1, c_2, \ldots c_p$ are the discriminant coefficients.

Assume that the average values of the k th variable of class G_1 and G_2 are $\overline{x_k^{(1)}}$ and $\overline{x_k^{(2)}}$, and we can get the average discriminant value of class A and B

$$\begin{cases} \overline{y^{(1)}} = \sum_{k=1}^{p} c_k \overline{x_k^{(1)}} \\ \overline{y^{(2)}} = \sum_{k=1}^{p} c_k \overline{x_k^{(2)}} \end{cases} \tag{2}$$

In order to make the discriminant function distinguish the samples from different population accurately, the difference between $\overline{y^{(1)}}$ and $\overline{y^{(2)}}$ should be as big as possible, and for each class, the deviation sum of squares $\sum_{i=1}^{n_1}(y_i^{(1)} - \overline{y^{(1)}})^2$ and $\sum_{i=1}^{n_2}(y_i^{(2)} - \overline{y^{(2)}})^2$ should be as small as possible, which means:

$$\mathrm{I} = \frac{\left(\overline{y^{(1)}} - \overline{y^{(2)}}\right)^2}{\sum_{i=1}^{n_1}\left(y_i^{(1)} - \overline{y^{(1)}}\right)^2 + \sum_{i=1}^{n_2}\left(y_i^{(2)} - \overline{y^{(2)}}\right)^2} \rightarrow \mathrm{MAX} \tag{3}$$

Let:

$$Q = Q(c_1, c_2, \ldots c_p) = \left(\overline{y^{(1)}} - \overline{y^{(2)}}\right)^2 \tag{4}$$