Modeling and Simulation 建模与仿真, 2022, 11(3), 842-851
Published Online May 2022 in Hans. http://www.hanspub.org/journal/mos
https://doi.org/10.12677/mos.2022.113078

Hans汉斯

基于K-Means聚类分析的汽车行驶工况构建

苏晓雯，董平军

东华大学，上海

收稿日期：2022年4月18日；录用日期：2022年5月23日；发布日期：2022年5月30日

摘　要

为构建合理的汽车行驶工况，以给定的轻型汽车行驶数据为基础，分别运用运动学片段分析法、主成分分析法和K-均值聚类分析法对实测数据进行降维和分类，并结合相关系数法从各类运动学片段库中选取具有代表性的片段，构建反映汽车行驶特征的汽车行驶工况曲线。最后，为验证所构建的汽车行驶工况的有效性和精确性，计算作为评价体系的8个特征参数的相对误差和总体误差。结果表明，构建的汽车行驶工况曲线所反映的汽车运动特征在一定程度上可以代表数据源对应的特征，所构建的行驶工况具有有效性和精确性。

关键词

K-Means聚类分析，汽车行驶工况，主成分分析，运动学片段

Construction of Vehicle Driving Cycle Based on K-Means Cluster Analysis

Xiaowen Su, Pingjun Dong

Donghua University, Shanghai

Received: Apr. 18th, 2022; accepted: May 23rd, 2022; published: May 30th, 2022

Abstract

In order to construct a reasonable driving cycle of the car, on the basis of a given light vehicle driving data, respectively using kinematics fragment analysis, principal component analysis (PCA) and K-Means clustering analysis of measured data for dimensionality reduction and classification, combined with the correlation coefficient method and the cumulative frequency method from the various segments in the library to select representative kinematics fragments, so as to build a curve of vehicle driving cycle which can reflect the characteristics of the car's driving. Finally, in order to verify the validity and accuracy of the constructed vehicle driving cycle, the relative er-

文章引用: 苏晓雯, 董平军. 基于 K-Means 聚类分析的汽车行驶工况构建[J]. 建模与仿真, 2022, 11(3): 842-851.
DOI: 10.12677/mos.2022.113078

rors and total errors of the eight characteristic parameters of the evaluation system were calculated. The results show that motion characteristics of the vehicle reflected in the constructed vehicle driving cycle curve can represent the corresponding characteristics of the collected data sources to a certain extent, and this constructed driving cycle is effective and accurate.

Keywords

K-Means Cluster Analysis, Vehicle Driving Cycle, Principal Component Analysis, Kinematic Fragment

1. 引言

汽车行驶工况(Driving Cycle)又称车辆测试循环，是通过数据分析所构建的描述汽车行驶的速度－时间曲线(见图 1 和图 2)。它可以体现汽车道路行驶的运动学特征，模拟真实的交通状况，以测试车辆尾气排放和燃料消耗。此外，其在交通协同控制、新车评价、风险评估和车辆的设计、选型、匹配和控制策略等方面有着广泛的应用[1]。目前，汽车发达国家都有自己的汽车行驶工况标准，而国外的行驶工况与国内行驶特征存在较大差异，直接采用则导致检测结果与实际数据往往存在较大误差。因此有必要建立反映国内行驶特点的典型行驶工况，提高检测结果的准确性和可靠性。

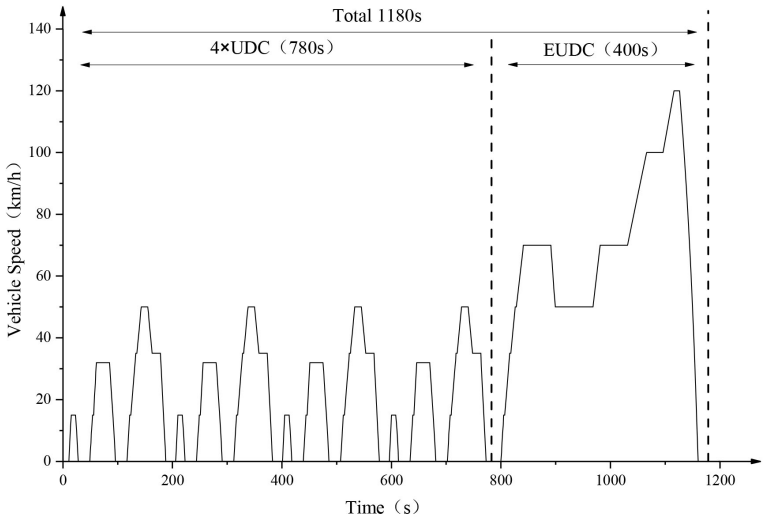

Figure 1. European NEDC driving cycle
图 1. 欧洲 NEDC 工况

目前，构建行驶工况的常用方法有单纯的短行程法、基于聚类的方法和基于马尔可夫链的方法。短行程法将数据划分为短行程片段，通过分析片段的特征参数组合，生成相应的行驶工况。国外学者 Lin 等采用短片段划分以及随机过程选择方法构建了行驶工况[2]。基于聚类的工况构建主要采用主成分分析法与聚类分析法相结合的研究方法，其中 K-均值聚类分析法在构建城市汽车行驶工况中应用较多，通过

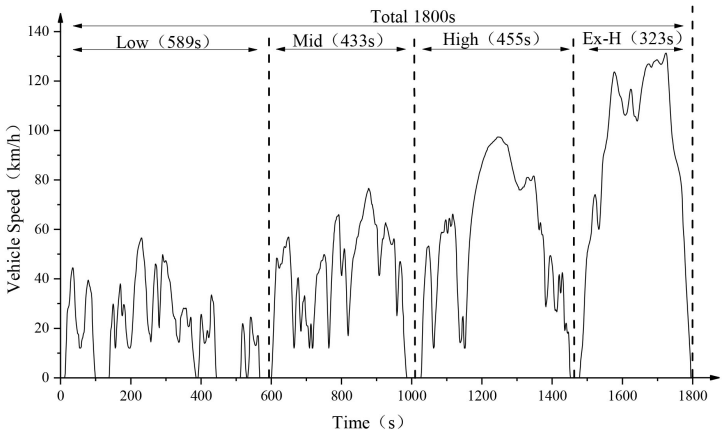

Figure 2. World WLTC driving cycle
图 2. 世界 WLTC 工况

利用聚类算法对样本数据进行分类，根据分类结果从中筛选出最佳短行程样本，最后组合为典型工况。国外学者 Fotouhi 等采用 K-Means 聚类算法构建了德黑兰的行驶工况[3]。同济大学胡志远利用短行程、主成分分析、聚类分析等方法对上海市公交车进行研究，生成了最优短行程组合[4]。彭育辉等基于 K-均值聚类方法对汽车行驶数据进行分析，提出一种以 Silhouette 函数筛选聚类结果，并根据聚类结果构建汽车行驶工况的方法[5]。刘子谭等利用短行程法、主成分分析及聚类方法，并针对 K-均值聚类稳定性较差的缺陷进行改进研究，将改进后的聚类方法应用于工况构建，生成了广州市行驶工况[6]。

然而由于交通环境的影响，汽车实际行驶工况具有较大的随机性，也有文献对基于马尔可夫链的方法进行研究。姜平等利用聚类和马尔可夫方法构建了城市汽车行驶工况[7]。苗强等采用聚类加马尔可夫链的方法构建了济南市公交车典型行驶工况[8]。曹骞等利用主成分和聚类算法对大连市乘用车行驶数据进行统计分析，并基于马尔可夫链随机过程原理构建了行驶工况[9]。李耀华等基于马尔可夫链构建了西安市城市公交线路工况[1]。

基于上述分析，本文以给定的某城市轻型汽车实际道路行驶采集的近十九万条数据为基础，结合运动学片段、主成分分析以及聚类分析的主要方法来构建汽车行驶工况，将划分的运动学片段聚成 3 类，并结合相关系数法从各类片段库中选取最优片段，从而构建出能体现汽车行驶特征的典型汽车行驶工况。通过与采集的总样本数据进行对比分析，验证了所构建行驶工况的准确性。

2. 数据预处理

本文使用的数据集是某城市中给定的一辆轻型车辆采集的实际道路行驶数据，数据集中共有 189,725 条数据，如下表 1 所示。

Table 1. Vehicle driving data (partial)
表 1. 汽车行驶数据(局部)

编号	时间	GPS 车速	X 轴加速度	Y 轴加速度	Z 轴加速度	经度	纬度	发动机转速	瞬时油耗
1	2017/12/18 13:42:20	0	0	−0.324	−0.936	119.3678	25.99242	900	0.36
2	2017/12/18 13:42:21	4.5	0	−0.324	−0.918	119.3678	25.99241	1025	0.44
3	2017/12/18 13:42:22	6.9	0	−0.324	−0.936	113.3678	25.9924	1137	0.46
…	…	…	…	…	…	…	…	…	…

采集设备直接记录的原始采集数据通常会包括一些不良数据值，为了使得数据集中的汽车行驶数据更加合理有效，需要对原始数据进行处理与清洗。将不良数据分为缺失数据、尖点数据、毛刺数据以及怠速数据几个类型，数据处理主要包含以下四个方面：

1) 缺失数据插值处理。对于数据时间不连续的缺失数据，通过 Matlab 程序找出相邻两个时间间断，但是速度 > 0 的节点进行插值处理。综合考虑各类因素，选择 Hermite 插值法。

2) 尖点数据的平滑处理。尖点数据是指两个相邻速度之间存在较大差异的数据，在行驶过程中，会因为各种原因造成异常加减速的数据。采用线性插值法对尖点数据选择进行平滑处理，令其等于前一个车速和后一个车速的平均数。

3) 毛刺数据删除处理。在车辆较长的怠速期间内，突然出现的个别汽车速度不为 0 的数据片段称作毛刺数据。而在汽车实际行驶状况中这种情况是不可能会出现的。而这种数据对后续的运动片段划分有较大的影响，因此需要清洗这些数据。

4) 怠速数据删除处理。汽车怠速是指当汽车停止运动但发动机保持最低转速时的状态。当怠速时间超过 180 s 时，一般视为异常情况。因此，需要将这种异常数据剔除掉。处理方法为：判断怠速的起点和终点，若连续怠速时间超过 180 s，则保留靠近怠速终点的最后 180 秒数据，删除前面的异常数据。

使用 Matlab 软件进行数据处理，经处理后的数据集中的数据量如表 2 所示。

Table 2. Comparison of data volume before and after processing
表 2. 处理前后数据量对比

原始数据量	插值拟合后数据量	最终处理完成后数据量	原始数据与最终数据的差额
185,725	217,506	178,667	−7085

3. 主成分分析与 K-Means 聚类分析

3.1. 运动学片段划分及特征参数提取

运动学片段也称为短行程，是指汽车从怠速状态开始到下一个怠速状态开始之间的速度范围，如图 3 所示，主要包括怠速阶段、加速阶段、匀速阶段和减速阶段[10]。根据运动学片段来构建汽车行驶工况曲线，需要先对运动学片段进行划分和提取。将数据集中按时间排序的数据划分为多个运动学片段。通过编写对应的 Python 程序，划分出 1421 个运动学片段。

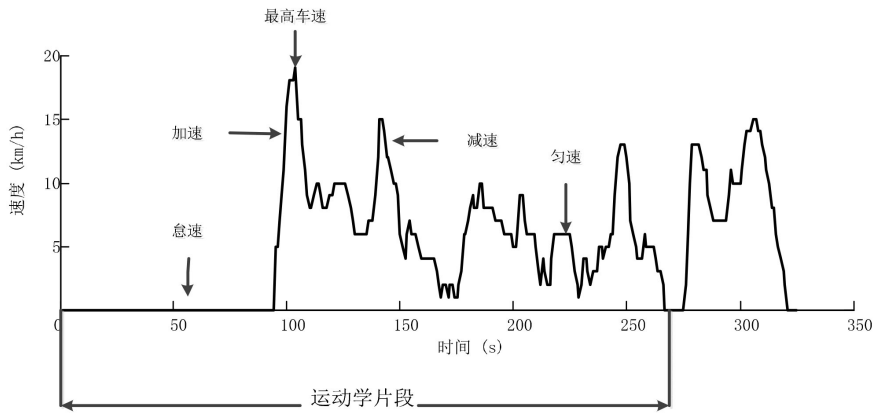

Figure 3. Definition of kinematic fragment
图 3. 运动学片段的定义

汽车行驶工况是由多个具有代表性的运动学片段组合而成，选取具有代表性的运动学片段需要有计算出其特征参数作为标准与依据。在运动学片段的评估分析中，特征参数能够体现该运动学片段的交通特征。因此，对划分后的运动学片段进行特征参数提取和特征值计算。所选用的描述运动学片段的 17 个特征参数如下表 3 所示。

Table 3. Definition table of characteristic parameters
表 3. 特征参数定义表

序号	特征参数	意义	单位
1	T	运行时间	s
2	T_a	加速时间	s
3	T_d	减速时间	s
4	T_c	匀速时间	s
5	T_i	怠速时间	s
6	P_a	加速时间比	%
7	P_d	减速时间比	%
8	P_c	匀速时间比	%
9	P_i	怠速时间比	%
10	a_a	平均加速度	m/s^2
11	a_d	平均减速度	m/s^2
12	S	运行里程	km
13	V_{max}	最大速度	km/h
14	V_m	平均速度	km/h
15	V_{me}	平均行驶速度	km/h
16	V_{sd}	速度标准差	km/h
17	a_{sd}	加速度标准差	m/s^2

计算得到各个运动片段的每个运动特征参数值，部分数据如下表 4 所示。

Table 4. Characteristic values of each kinematic fragment (partial)
表 4. 各运动学片段特征值(局部)

(a)

片段编号	运行时间	加速时间	减速时间	匀速时间	怠速时间	加速时间比	减速时间比	匀速时间比	怠速时间比
0	68	25	19	16	8	0.3676	0.2794	0.2353	0.1176
1	369	110	86	76	97	0.2961	0.2331	0.2060	0.2329
...	...	...	...	...	...	...	...	...	...

(b)

片段编号	平均加速度	平均减速度	运行里程	最大速度	平均速度	平均行驶速度	速度标准差	加速度标准差
0	68	25	19	16	8	0.3676	0.2794	0.1176
1	369	110	86	76	97	0.2961	0.2331	0.2329
...	...	...	...	...	...	...	...	...

3.2. 主成分分析

原始数据常常存在量纲不一致的特点，首先采用最大最小标准化方法来进行数据标准化。其次，由于过大的数据量会大大降低计算效率，不利于聚类分析等各种问题，因此经过数据处理和运动学片段划分后，将利用主成分分析法对标准化处理之后的数据进行降维处理，目的是减少变量，提高后续计算能力。

运用 Python 软件进行主成分分析，贡献率从大到小依次排列并计算累计贡献率，输出结果见表 5 所示。通常选择累计贡献率小于 85%的主成分，前 4 个主成分累计贡献率达 88.59%，因此将用 4 个主成分替换 17 个特征参数用于工况的构建，将数据由 17 维降到了 4 维。

Table 5. Contribution rate and cumulative contribution rate of each principal component
表 5. 各主成分贡献率及累计贡献率

主成分排序	贡献率	累计贡献率
主成分 1	54.702379%	54.702379%
主成分 2	14.875836%	69.578215%
主成分 3	14.223835%	83.802050%
主成分 4	4.791009%	88.593059%

3.3. K-Means 聚类分析

主成分分析为聚类分析做准备，聚类分析是本文的重要方法，本文选择 K-Means 聚类方法对主成分评分数据进行处理。首先，采用手肘法来确定 k 值，也就是将数据分为几类。在 K-Means 算法中，k 值的选择往往对聚类结果具有一定的影响，为了评估聚类效果，选用平方误差和(SSE)作为聚类结果的评估指标。通过每个簇点与其质心之间距离的平方来计算 SSE [11]。手肘法可以用来反映 k 的不同取值和 SSE 的关系趋势，便于找到最佳 k 值。

对降维后的数据进行 K-Means 聚类分析，首先需要确定 k 值，手肘图如图 4 所示。根据手肘图的拐点，选择聚类簇个数 k 值为 3，将所有的运动学片段分为 3 类，聚类效果图如图 5 所示。其中横纵坐标分别是降维后的数据的前两列。

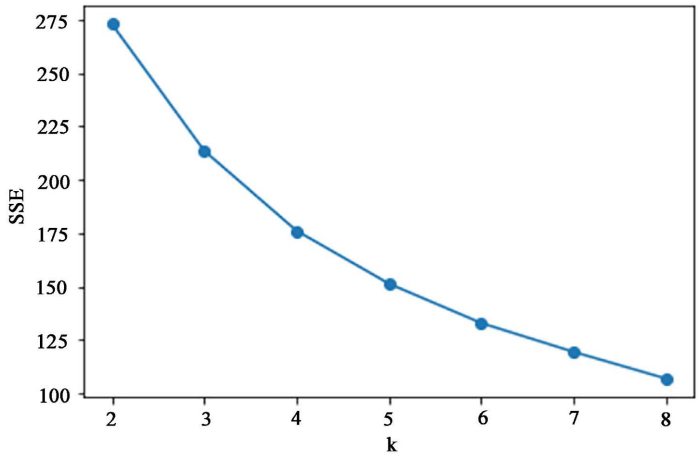

Figure 4. Elbow diagram
图 4. 手肘图

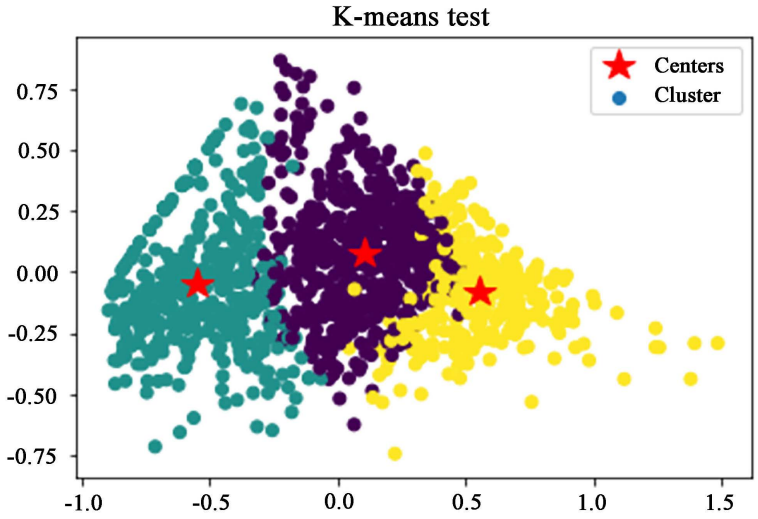

Figure 5. Clustering diagram of kinematic fragments
图 5. 运动学片段聚类图

通过评估聚类中心的特征值，找到每一类数据的特点，聚类中心的特征值如表 6 所示。

Table 6. Eigenvalues of cluster centers
表 6. 聚类中心的特征值

	第 1 类	第 2 类	第 3 类
运行时间	0.016645	0.049312	0.006694
加速时间	0.046925	0.170074	0.007414
减速时间	0.043017	0.133798	0.009014
匀速时间	0.005981	0.018844	0.000666
怠速时间	0.093908	0.126895	0.150542
加速时间比	0.358979	0.449658	0.074917
减速时间比	0.385482	0.359781	0.169062
匀速时间比	0.232071	0.296445	0.13232
怠速时间比	0.220116	0.104162	0.699957
平均加速度	0.154096	0.130917	0.084096
平均减速度	0.70828	0.70739	0.779982
运行里程	0.006225	0.05314	0.000661
最大速度	0.228996	0.491348	0.065986
平均速度	0.140942	0.383227	0.018834
平均行驶速度	0.181354	0.423988	0.056988
速度标准差	0.179724	0.432003	0.03754
加速度标准差	0.339284	0.297921	0.196263

从以上的特征值表中，发现第三类运动学片段中怠速比例最高，达到 69.9%，说明第三类片段代表的是堵车严重的路段，而第一二类数据的怠速时间比相对于第三类则明显减少，并且第二类数据的平均速度是三类数据中最高的，是道路通畅的路段。

4. 行驶工况构建与验证

4.1. 行驶工况构建

原始数据经过清洗、运动学片段的划分、降维与分类后，考虑到每个分类中的数据量还会较大，因此需要从每类运动学片段中抽取合适的片段来合成行驶工况。在选取运动学片段时考虑它们的相关性，即相关系数。在每个聚类中选取与该类特征值相关系数最大的若干代表性运动学片段构建车辆的行驶工况。计算出各类运动学片段和它所在的类别中心的相关系数，列取出相关系数较大的前 10 个运动片段如表 7 所示。

Table 7. Correlation coefficients of various kinematic fragments
表 7. 各类运动学片段相关系数大小

片段编号	第 1 类	片段编号	第 2 类	片段编号	第 3 类
1105	0.997226	1263	0.999974	1369	0.999746
1101	0.996365	905	0.999971	1418	0.999603
446	0.996007	322	0.999968	832	0.999082
38	0.995294	993	0.999923	1220	0.998655
418	0.994683	1077	0.998996	591	0.998048
358	0.99411	352	0.998889	13	0.997702
1028	0.994041	797	0.998862	1355	0.997099
1255	0.993906	306	0.998754	194	0.996822
24	0.993869	1237	0.998617	916	0.996668
610	0.989902	1111	0.998466	1178	0.994999

由于拟定构建的行驶工况的时间长度为 1200~1300 s，分别从第一类工况中选取 3 个运动学片段，从第二类工况中选取 4 个运动学片段，从第三类工况中选取 2 个运动学片段。将所选运动学片段首尾相连，最终构建的长度为 1265 s 的行驶工况如图 6 所示，横坐标为时间(s)，纵坐标为速度(km/h)。

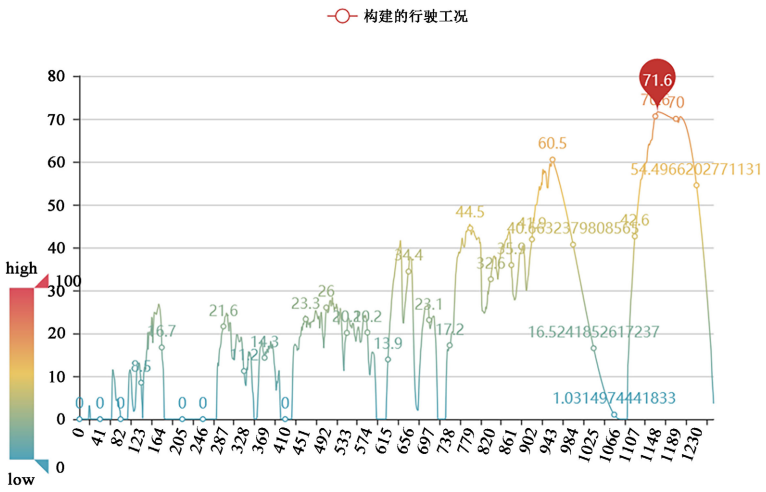

Figure 6. Driving cycle diagram
图 6. 行驶工况图

4.2. 行驶工况验证

由于前期行驶工况的构建是由对于试验数据采取统计的方法得到的，所以要验证其与原始数据之间的差异[11]。选取能反映总体统计情况的特征参数作为分析指标，计算所构建道路行驶工况及相应的总体行驶工况的相对误差和绝对误差，并根据计算结果分析所建道路行驶工况的有效性。所选取的指标特征参数如下表 8 所示。σ_i 是构建出道路行驶工况的特征参数，σ_j 是工况所对应总体的特征参数。相对误差 $\delta=\dfrac{\left|\sigma_i-\sigma_j\right|}{\sigma_j}$，绝对误差 $\Delta=\left|\sigma_i-\sigma_j\right|$。

Table 8. Parameters used in error analysis
表 8. 误差分析所用参数

序号	特征参数	意义	单位
1	P_a	加速时间比	%
2	P_d	减速时间比	%
3	P_c	匀速时间比	%
4	P_i	怠速时间比	%
5	a_a	平均加速度	m/s^2
6	a_d	平均减速度	m/s^2
7	V_m	平均速度	km/h
8	V_{me}	平均行驶速度	km/h

根据相对误差和绝对误差的公式，总体数据与构建的汽车行驶工况的误差分析如表 9 所示。从表中可知，所构建的行驶工况的特征参数中，多数参数小于 10%，尤其平均加速度的相对误差为 0.75%，从特征参数的验证结果来看，行驶工况的构建方案是比较合理的。

Table 9. Error analysis of vehicle driving cycle
表 9. 汽车行驶工况的误差分析

特征参数	加速比例 P_a	减速比例 P_d	匀速比例 P_c	P_i 怠速比例	a_a 平均加速度	a_{sd} 平均减速度	V_m 平均速度	V_{mr} 平均行驶速度
构建工况	0.283	0.3075	0.2055	0.204	0.3754	−0.3714	6.6598	8.3661
采集总体	0.2646	0.2222	0.3338	0.1794	0.3726	−0.4948	7.134	8.6935
相对误差	6.95%	38.39%	38.44%	13.71%	0.75%	24.94%	6.65%	3.77%

5. 总结与展望

本文以某城市轻型汽车实际道路行驶采集的近十九万条数据作为实验数据集，划分出 1421 个运动学片段，提取描述运动学片段的 17 个特征参数，采用主成分分析和 K-Means 聚类算法对特征参数矩阵进行降维和分类处理，将运动学片段分为 3 类，并结合相关系数法从 3 类运动学片段库中选取代表性片段，从而构建出了符合数据源中汽车行驶特征，时长 1265 s 的车辆行驶工况。通过分析实验数据与行驶工况的特征参数，验证了工况的准确性。

实验结果表明，本文中行驶工况的构建方案比较合理，拟合出的轻型汽车行驶工况能反映真实数据

特征，符合城市道路实际工况。然而本文对于工况构建的指标没有考虑到现实生活中存在的地形、环境、温度等不确定性因素，这些因素均可能影响运动学片段的走势，未来可以进一步结合马尔科夫链方法进行研究。

参考文献

[1] 李耀华, 任田园, 邵攀登, 宋伟萍, 李忠玉, 苟琦智. 基于马尔可夫链的西安市城市公交工况构建[J]. 中国科技论文, 2019, 14(2): 121-128.

[2] Lin, J. and Niemeier, D.A. (2003) Regional Driving Characteristics, Regional Driving Cycles. *Transportation Research Part D*, **8**, 361-381. https://doi.org/10.1016/S1361-9209(03)00022-1

[3] Fotouhi, A. and Montazerigh, M. (2013) Tehran Driving Cycle Development Using the K-Means Clustering Method. *Scientia Iranica*, **20**, 286-293.

[4] 胡志远, 秦艳, 谭丕强, 楼狄明. 基于大样本的上海市乘用车行驶工况构建[J]. 同济大学学报(自然科学版), 2015, 43(10): 1523-1527.

[5] 彭育辉, 杨辉宝, 李孟良, 乔学齐. 基于 K-均值聚类分析的城市道路汽车行驶工况构建方法研究[J]. 汽车技术, 2017(11): 13-18.

[6] 刘子谭, 朱平, 刘旭鹏, 刘钊. K 均值聚类改进与行驶工况构建研究[J]. 汽车技术, 2019(11): 57-62.

[7] 姜平, 石琴, 陈无畏. 聚类和马尔科夫方法结合的城市汽车行驶工况构建[J]. 中国机械工程, 2010, 21(23): 2893-2897.

[8] 苗强, 孙强, 白书战, 闫伟, 李国祥. 基于聚类和马尔可夫链的公交车典型行驶工况构建[J]. 中国公路学报, 2016, 29(11): 161-169.

[9] 曹骞, 李君, 曲大为. 大连市乘用车典型行驶工况的构建[J]. 上海交通大学学报, 2018, 52(11): 1537-1542.

[10] 李杰, 王晓佳, 朱建军, 张翠平, 汪洋. 太原市公交车行驶工况的构建[J]. 中国科技论文, 2018, 13(19): 2223-2227.

[11] 宋怡帆. 基于聚类和 Python 语言的深圳市城市道路车辆行驶工况构建[D]: [硕士学位论文]. 西安: 长安大学, 2018.