中国研究生创新实践系列大赛

“华为杯”第十六届中国研究生

数学建模竞赛

学　　校		华东师范大学
参赛队号		**19102690204**
队员姓名	**1.**	傅渲铭
	2.	赵菁菁
	3.	吴一凡

中国研究生创新实践系列大赛

“华为杯”第十六届中国研究生

数学建模竞赛

题 目 汽车行驶工况构建

摘 要:

汽车行驶工况又称车辆测试循环，可以表现为能够描述特定行驶环境限制下车辆行驶的速度-时间关系的曲线。对汽车行驶的实际数据加以分析和提取可以得到汽车行驶工况，其在车辆设计、检测和维修以及车辆行驶监控等方面发挥着重要作用。我国地域辽阔，地区发展水平差距较大，针对不同地区建立汽车行驶工况具有重要意义，本文的目的即是根据汽车行驶的真实数据为某城市构建汽车行驶工况。

数据预处理（问题一）对原始数据进行预处理。首先从发动机转速分布、速度-加速度频次分布两方面出发，从宏观上分析了三份原始样本文件，对原始样本有一定把握，分析出样本中出现不良数据的 3 个最主要原因，结合相关文献，确定合理数据所具有的特征，例如加减速度的合理范围，并确定了 4 类异常数据的判断标准，制定了相应的预处理办法。准确地定位异常数据，剔除无效数据段以保证数据的准确性和连续性，并根据所制定的 3 个原则，采用拉格朗日插值或三次样条插值方法对缺失信息进行补全、对毛刺数据点进行平滑处理。最终得到的预处理后三个文件的样本数分别为 178685、143128、159495 条。同时选取和计算运动学特征参数，为后续数据的统计分析和工况构建提供数据支撑。

运动学片段的提取（问题二）根据运动学片段的定义，对数据的预处理结果进行了划分。由于长时间断点的存在，首先对每个文件划分时间片段，然后在每个较长的时间段内划分各自时间段内部的运动学片段。初步划分后，得到三个文件中的运动学片段数目分别为 1181、1026、919，共 3126 个。考虑到划分结果中存在明显不合理的片段，例如整个片段的时长仅持续几秒，我们还进一步对划分结果进行筛选，得到合理的运动学片段，三个文件的合理运动学片段数目分别为 1077、859、776，共 2712 个。选择了用于表征运动学片段的 15 个特征参数，如行驶时间、行驶路程、平均速度等，对于每个合理运动学片段，计算它们的 15 维特征参数值。

汽车行驶工况的构建（问题三）为了在复杂多样的运动学片段中提取具有代表性的片段，最终构建出合理、有效的汽车行驶工况，需要对在问题二获得的运动学片段进行分类和再次分析。根据运动学片段样本的总量和数据维度较多的特点，决定对输入数据进行降维以减少工况运算量，本文采用并对比了改进后的均值中心化 m-PCA（mean-centering PCA，主成分分析）、KPCA（Kernel PCA，核主成分分析）以及 t-SNE（t 分布随机邻域嵌入）方

法，对问题二中提取出的 15 个特征参数进行降维，根据贡献率分析选定 4 个有效维度，既保证了信息的完整性，也能够使构建的工况最大程度反映整体数据的特点；为了分析运动学片段的典型代表，需要对整体运动学片段进行聚类，本文利用 SOM（Self-organizing Maps, 自组织特征映射网络）神经网络的输出作为 FCM（Fuzzy c-means，模糊 C-均值）算法初始聚类中心，获得最终的聚类结果，改进后的算法能够有效降低 FCM 聚类法结果陷入局部最优的概率，从而更好地反映出汽车行驶的速度和加速度的分布。对不同聚类中心个数结果进行实验，计算得到相应的模糊划分系数 FPC，根据 FPC 的数值发现，将数据聚类成 3 类拟合出的模型最佳；构建汽车工况需要选取每个聚类的最典型的进行构建，本文在基于聚类结果的基础上，采用改进的马尔可夫链过程方法进行类内工况的构建，并基于蒙特卡洛模拟的方式生成符合状态转移概率矩阵的随机数组来确定下一片段，由于起始片段会影响整个马尔可夫链，采用总体特征参数偏差最小而非传统的随机方法选取起始片段，通过各类运行时间在总体数据中占的比例，确定各类中的拟合工况时间。

最后对得到的代表性工况进行了有效性验证，利用汽车行驶数据的典型特征参数对代表性工况进行了误差分析，验证了代表工况的可靠性。根据运动学参数和得到的工况结果，分析其数据和产生误差的原因，最终针对性地提出对模型改进的设想。

关键词：车辆行驶工况 主成分分析 自组织映射 模糊 C 均值聚类 马尔可夫过程 三次样条插值

目录

一、 前言

1.1 研究背景

近年来，我国乘用车保有量迅速增长，由此带来日益严重的能源危机及汽车排放物污染问题。汽车行驶工况是车辆性能测试实施标准法规中的一项重要内容，它作为汽车产品研发匹配过程中一项基础性依据，对车辆本身在特定环境条件下的实际道路中的燃料使用、排放、舒适性和可靠性等方面具有决定性影响。因此，构建出符合我国国情的、基于城市自身的汽车行驶工况尤为重要，既可以作为汽车污染物排放量和燃油消耗量的测试的依据，又可以为新车型的开发、评估提供支持，具有深刻的实际意义。

1.2 问题重述

基于上述研究背景，本文旨在研究并完成以下任务：

任务一：数据预处理

采集设备所采集的汽车行驶数据中存在一些不良情况，如因障碍物导致信号丢失造成数据不连续；汽车加、减速度不在合理范围内；长期停车时车速长时间为 0；长时间堵车造成汽车断断续续低速行驶（最高车速＜10km/h）；汽车怠速时间超过 180 秒等。上述不良情况难以合理地反映正常的汽车行驶工况，任务一的目标是：根据上述不良情况，进行冗余信息剔除以及空穴数据填充，并给出各文件经处理后的记录数。

任务二：运动学片段的提取

如图 1 所示，运动学片段呈现了汽车从怠速状态开始至下一个怠速状态开始之间的车速区间，通常包含怠速状态、加速状态、减速状态、匀速状态这个行驶状态。通过运动学片段，可以构建出汽车行驶工况的曲线。任务二的目标是：在任务一所得到的经过合理处理的数据文件的基础上，经过调研、分析、计算等步骤，划分、提取出相应的多个有一定代表性的运动学片段，并分别得到各数据文件的运动学片段数量。

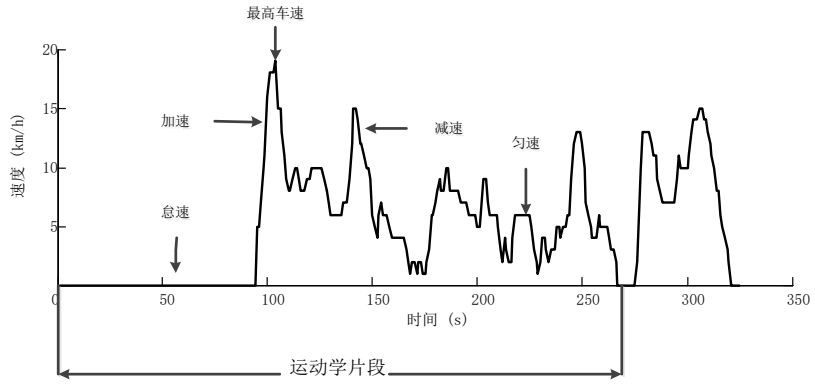


图 1.1 运动学片段的定义

任务三：汽车行驶工况的构建

车辆道路行驶工况由一系列的运动学片段组成。由于不同时间、天气和路段等因素的影响，运动学片段之间通常互有差异，因此将这些在特定情况下、代表不同交通特征的运动学片段组合起来描述代表车辆实际运行状况是一个难点。任务三旨在利用科学、有效的数学模型构建方法、采用合理的汽车运动特征评估体系构建、优化、得到一条具有代表性的汽车行驶工况曲线，并分别计算曲线的、样本的各运动特征变量的值，比较并且探究所

得曲线模型的合理性。

二、 模型假设

1. 忽略 GPS 车速与实际车速之间的误差；
2. 不考虑汽车的品牌、型号、使用时间、发动机排量等差异，认为通过设备采集到的数据均在合理范围内，不会因以上原因产生较大差距；
3. 在进行数据分析时，不考虑因驾驶员的年龄、性别、驾驶习惯等原因产生的行车状况差距；
4. 在 GPS 数据采集的间隔时间（1 秒钟）内，认为行驶车辆在进行匀速直线运动， 不考虑每个时间间隔内汽车运行速度、加速度等状态的变化；
5. 普通汽车的加减速有一定的限制，我们认为普通轿车一般情况下：0 至 100km/h 的加速时间大于 7 秒，紧急刹车最大减速度在 7.5~8 m/s^2
6. 汽车长时间怠速或停止行驶的情况会影响工况分析的准确性，因此我们认为超过 180s 的怠速为异常怠速，会进行异常处理。

三、 模型符号分析与说明

符号	描述	单位
T	行驶时间	s
S	行驶路程	m
T_a	加速时间	s
T_d	减速时间	s
T_i	怠速时间	s
T_c	匀速时间	s
V_{max}	最大速度	km/h
V_m	平均速度	km/h
V_{mr}	行驶平均速度（不含怠速）	km/h
V_{sd}	速度标准偏差	km/h
a_{max}	最大加速度	km/h/s
a_m	平均加速度	km/h/s
d_{max}	最大减速度（绝对值）	km/h/s
d_m	平均减速度（绝对值）	km/h/s
a_{sd}	加速度标准偏差	km/h/s
A	特征参数矩阵	
A'	协方差矩阵	
B	均值化的特征参数矩阵	
N	状态转移次数	
P	状态转移概率	

四、 任务一的求解

4.1 问题分析

所给的三个文件中的数据包含一些不良数据，任务一要求我们分别分析不同不良情况对在数据中的体现，对数据进行预处理。

经分析有以下三种主要原因导致不良数据的出现。其一，车辆在道路行驶过程中，由于高层建筑覆盖、过隧道等，车辆行驶记录仪容易出现短时间的未定位状态，采集数据样本中存在时间断点，根据文献[i]果断点仅为几秒钟，可以根据车辆行驶的前后状态及所采集的经纬度信息进行补点，若丢失时间片段较长则删除该片段；其二，普通轿车一般情况下从 0 加速至 100km/h 需要的时间＞7 秒，且紧急刹车最大减速度在 7.5~8m/s² 。速度突变数据的特征为：某片段时间信息完整，在某个时间节点速率突然增大或减小，处理这种数据的作法是设定合理的加减速度范围，对不符合要求的坏点作相应的处理；其三，数据样本中出现了一些包含几秒钟速度不为 0 且最高车速＜10km/h 的行驶片段，片段前后是相对较长的怠速时间，考虑车流量的影响，我们认为这种片段出现的主要原因是道路拥堵导致车辆短时间突然启停。由于速度不为 0 片段相较前后怠速时间所占比例很小，为方便后面运动学片段划分，应将其进行归零。

4.2 原始数据宏观分析

原始数据是某城市轻型汽车实际道路行驶采集的数据，由 GPS 记录仪所采集，采样频率 1Hz。文件一、文件二、文件三分别包括 185725、145825、164914 个样本数据。共有 14 维变量，分别是时间、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、扭矩百分比、瞬时油耗、油门踏板开度、空燃比、发动机负荷百分比和进气流量。对原始数据有一个宏观把握，有利于后续工作的进行。

图 4.1 是三个文件发动机转速的分布图。总体上可以看出，三个文件的发动机转速绝大部分集中在 600~1400r/min 内，其中比重最大的是 600~800r/min，即在大部分时间内，发动机转速都处于中低转速状态。

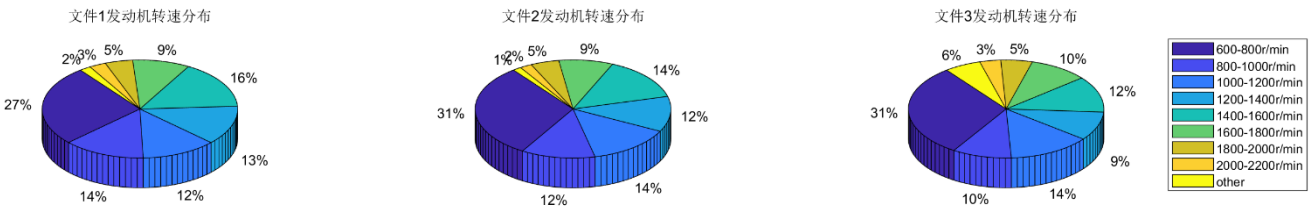
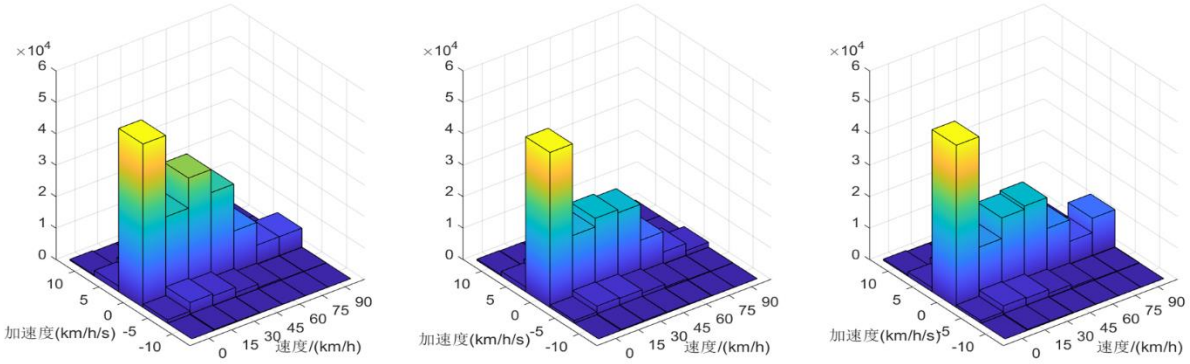


图 4.1 发动机转速分布

图 4.2 是计算了原始数据文件加速度后，三个文件的加速度-速度频次分布情况，行驶数据主要分布在 0~45km/h 车速，-2.5~2.5km/h/s 的加速度区间内，总体上具有车速较低，加减速平缓的行驶特点。



原始数据的加速度-速度频次分布图
从左到右分别是文件1、文件2、文件3

图 4.2 加速度-速度频次分布图

图 4.3 是原始数据的油门踏板开度数据，在原始数据中大多数时间油门踏板都处在未踏下的状态。

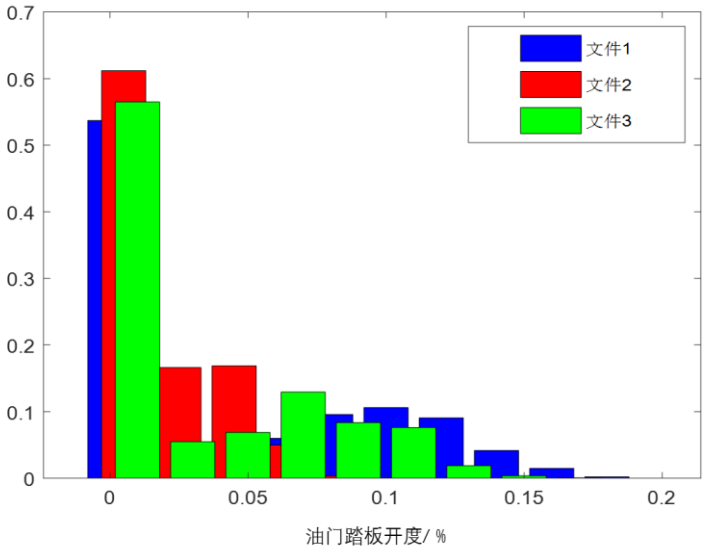


图 4.3 油门踏板开度分布

4.3 不良数据类型

根据原始数据的特性，结合汽车行驶过程的现实因素，本论文将其中的不良数据大致分为以下四类：

（1）时间断点数据：由于 GPS 记录设备的记录缺失或者传输失败而产生的时间不连续的数据；

（2）噪声数据：GPS 速度值异常的数据，例如速度大于汽车在行驶中所能到达的最

大值；

（3）尖点数据：速度处于正常范围内但加速度或减速度超出汽车正常行驶时所能达到的最大加减速度的数据；

（4）怠速异常数据：在怠速过程中偶尔出现的速度不为零但值较小的数据，或怠速时间过长可能影响工况构建的数据。

4.4 数据预处理策略及处理结果

针对上述的不良数据，分析并结合相关资料，本论文设计的数据预处理策略如下：

（1）针对时间断点数据，对于短时的时间断点，我们采用拉格朗日插值的方式对速度、加速度等数据特征进行插值。具体策略是当时间断点小于 4 秒时进行拉格朗日插值，时间断裂过长的数据不进行处理，这是因为过长的时间间隔存在多种可能的行驶状态变化，难以预测，且后续的运动学片段提取是在一个较长的时间片段内进行，时间片段之间不相互影响，所以此时无需处理。

处理示例如下图：

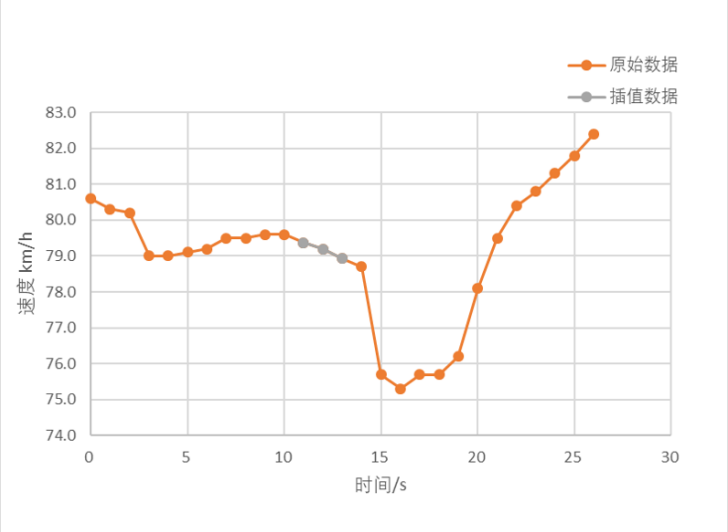


图 4.4 插值示意图

注：数据为文件 3 中 2017 年 12 月 1 号 19：59：00—19：59：26 的数据。观察这一区间内的样本可知，当车辆行驶至区间 11s~13s 时，GPS 记录仪并未记录下时间信息及速度信息等，采用拉格朗日插值法对丢失区间进行插值。

（2）针对噪声和尖点数据，我们主要采取了平滑策略，具体的平滑方案如下：

Ⅰ 加速度过大时，即加速度大于 14km/h/s：

a. 该时刻的加速度与邻近时刻的加速度差距较小时，将该点的加速度置为加速度上限；

b. 该时刻的加速度与邻近时刻的加速度差距较大时，将该点的加速度置为 0，认为该时刻的速度与上一时刻速度保持一致。

Ⅱ 减速度过大时，即减速度超过最大区间 27km/h/s-28.8km/h/s 时，我们将减速度置为减速度的上限。

Ⅲ 最后利用三次样条插值算法[ii]对加速度陡然上升和下降的区域进行平滑。

（3）针对怠速异常数据，我们主要针对怠速时间过长的数据进行了处理，对于超过

180s 的怠速时段，采取了怠速时段前半部分保留 1s，后半部分保留 180s，即保证该怠速时间段前的运动时段后有至少 1s 的怠速时间，该怠速时间段后的运动时段前有 180s 的怠速时间的方案；另外，若两个速度均为 0 的时间点序列中间出现少于 4 个速度不为 0 但≤10km/h 的时间点，则将这一不为 0 片段速度归零。

经过上述处理后，三个文件各剩余样本数分别为 178684、143127、159494。结果如下表所示：

表 4.1 数据预处理结果

文件	处理前样本数	处理后样本数	删去样本数
文件 1	185725	178684	7041
文件 2	145825	143127	2698
文件 3	164914	159494	5420

五、 任务二的求解

5.1 问题分析

运动学片段是只车辆从一个怠速状态开始直到下一个怠速状态开始的行驶片段，典型的运动学片段包括四个运动状态：加速状态、减速状态、匀速状态和怠速状态，通常情况下，当车辆行驶于不同交通环境下时所得到的运动学片段也不同。利用聚类算法将大量不同的运动学片段进行归类，同一类片段反映相似的行驶特征，最后挑选各自最优片段构建代表性最优的汽车行驶工况。

根据任务要求，原始的三个文件数据经上述任务一的处理后，根据一定的划分方式将其划分为三份运动学片段的集合，且由于划分后会存在一些不合理的无效片段，对后续聚类处理及工况构建的结果产生较大偏差，因此为了更全面地解决问题二，需要确定一定的筛选原则对各个集合进行初步片段筛选。进一步，为了便于表征和评价筛选后的运动学片段，需要选取相应的运动学特征参数并针对每一片段进行各个特征参数计算。

任务二的总体解决思路如下图所示：

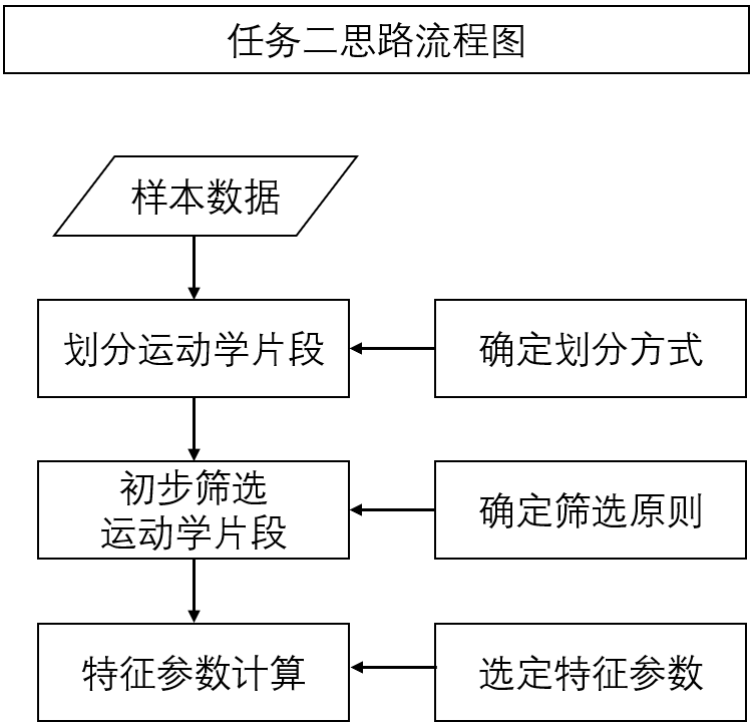


图 5.1 任务二流程图

5.2 划分方式

根据文献[1]目前国内运动学片段四种运动学状态的划分有三种方式：
第一种划分方式：
怠速状态：速度≤1km/h 且|加速度|≤0.1m/s² 的连续运转过程；
加速状态：速度≠0 且加速度≥0.1m/s² 的连续运转过程；
匀速状态：速度≥1km/h 且|加速度|≤0.1m/s² 的连续运转过程；
减速状态：速度≠0 且加速度≤-0.15m/s² 的连续运转状态；
第二种划分方式：
怠速状态：速度＝0 且发动机工作的连续运转过程；
加速状态：速度≠0 且加速度≥0.15m/s² 的连续运转状态；
匀速状态：速度≠0 且|加速度|≤0.15m/s² 的连续运转状态；
减速状态：速度≠0 且加速度≤-0.15m/s² 的连续运转状态；
第三种划分方式：
怠速状态：速度＝0 且发动机工作的连续运转过程；
加速状态：速度≠0 且加速度≥0.36m/s² 的连续运转状态；
匀速状态：速度≠0 且|加速度|≤0.36m/s² 的连续运转状态；
减速状态：速度≠0 且加速度≤-0.36m/s² 的连续运转状态。
本论文综合考虑以上三种划分方式及现实因素，本文采用第二种划分方式。

5.3 筛选原则

按上述方式划分后的运动学片段中，存在一些不良片段，如时长很短的片段、怠速区

间相对很长的片段等，会对后续处理的准确性产生较大影响，因此按照一下规则对其进行初步筛选：

（1）与相邻运动学片段时间间隔超过 180s 的片段，予以剔除；

（2）运动学片段的完整时长＜20s，予以剔除[iii]；

（3）认为怠速时段超过 180s 属于异常情况，予以剔除。

5.4 车辆运动特征参数的选择

对运动学片段的表征主要考虑一下几个方面：行程长度、行驶时间、怠速时间、行驶速度、平均加速度以及描述瞬时速度和加速度的标准偏差等等。表 5.1 列出了本论文所选择的 15 个反映片段特征的参数以及它们的意义。通过计算片段的这 15 个特征值，可以进行主成分分析和聚类分析。

表 5.1 车辆行驶特征参数

符号	描述	单位
T	行驶时间	s
S	行驶路程	m
T_a	加速时间	s
T_d	减速时间	s
T_i	怠速时间	s
T_c	匀速时间	s
V_{max}	最大速度	km/h
V_m	平均速度	km/h
V_{mr}	行驶平均速度（不含怠速）	km/h
V_{sd}	速度标准偏差	km/h
a_{max}	最大加速度	km/h/s
a_m	平均加速度	km/h/s
d_{max}	最大减速度（绝对值）	km/h/s
d_m	平均减速度（绝对值）	km/h/s
a_{sd}	加速度标准偏差	km/h/s

5.5 求解过程及结果

由于预处理后文件中的数据由若干较长的时间段组成，所以先将各个文件经过预处理后依然存在的时间断点找出，并且记录每个文件最后一个样本处的时间断点为无穷大；再在每个时间段内根据运动学片段划分原理划分出短行程。前者处理由 Python 编程实现，后者处理由 Matlab 编程实现。处理后得到的结果如下表所示：

表 5.2 运动学片段划分结果

文件（已预处理）	时间片段数	运动学片段数目
文件1	612	1181
文件2	416	1026
文件3	668	919
总计		3126

对划分后的运动学片段，根据上述的筛选原则进行初步筛选，由 Matlab 编程实现，筛选后的结果如下表所示：

表 5.3 运动学片段划分结果

文件（已预处理）	筛选前片段数	筛选后片段数
文件1	1181	1077
文件2	1026	859
文件3	919	776
总计	3126	2712

剔除掉不合理的运动学片段后，对各个运动学片段计算 15 个特征值，该过程由 Python 编程实现，部分结果如下表所示：

表 5.4 文件一部分特征参数计算结果

片段序号	T	V_m	V_{mr}	a_{max}
1	68	7.0	7.8	4.5
2	369	26.2	35.4	3.9
3	119	20.4	25.3	4.3
4	218	27.3	40.8	6.1
…	…	…	…	…
534	166	20.1	30.3	10.6
535	28	1.1	3.4	2.7
536	80	19.4	25.0	3.1
537	126	12.1	21.8	9.2
…	…	…	…	…
1074	154	27.4	35.8	5.1
1075	194	39.0	40.9	5.7
1076	285	33.5	34.4	7.2
1077	486	51.7	53.1	6.7

六、 任务三的求解

6.1 问题分析

任务三要求利用科学、有效的数学模型构建方法，整合三个文件所得到的运动学片段，采用合理的汽车运动特征评估体系构建、优化、得到一条具有代表性的汽车行驶工况曲线，并分别计算曲线的、样本的各运动特征变量的值，比较并且分析所得曲线模型的合理性。由于运动学片段之间的差异性，应先将各特征较为相似的运动学片段归为一类。聚类分析是用于将数据对应的研究对象（客体）进行分类的统计方法。在搜索了相关文献后，本论文选择使用 FCM 聚类方法[2]对运动学片段进行聚类。基于所得到的三个文件运动学片段的特征值，如果仅使用一两种片段特征值进行聚类分析，很可能会丢失一些重要信息造成结果偏差，而若完全使用 15 种片段特征值，会很大程度上增加计算的复杂度，且容易造成过度拟合。因此，在对片段聚类前，应先进行特征降维，本论文使用 PCA 方法实现。最后使用马尔科夫过程选择出最具代表性的汽车行驶工况。

6.2 特征参数降维

6.2.1 基于主成分分析法的特征参数降维

由于不同特征参数的量纲不同，导致各参数取值离散化，方差大的变量在后续分析时权重较大而得到优先考虑，降低了计算的准确性。为了消除量纲不同所造成的影响，在进行特征降维前应先作标准化处理[iv]。

对运动学片段构造包含 15 个特征参数的矩阵 A 为：

$$A_{\mathrm{m\times n}}=\begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix} \tag{1}$$

注：a_{ij}表示第 i 个运动学片段的第 j 个参数。

首先求出矩阵 A 的协方差矩阵A'，对A'做均值化处理，得到均值化处理后的特征参数矩阵 B，如下式（2）所示：

$$B_{\mathrm{ij}}=\frac{A_{\mathrm{ij}}^{'}}{A_{\mathrm{j}}^{'}},\qquad \overline{A_{j}^{'}}=\frac{1}{\mathrm{m}}\sum_{k=1}^{m}a_{\mathrm{kj}} \tag{2}$$

对该矩阵 B 进行主成分分析。利用线性变换构造出新变量，满足各主成分之间两两独立且线性无关，以此消除各特征参数对运动学片段描述时具有的重叠信息，同时保留它们的绝大部分有用信息。相关研究表明，若主成分的特征值＞1，且累计贡献率＞80%，则符合汽车行驶工况的构建要求[2]。上述过程由 Python 编程实现，得到的结果如下表所示：

表 6.1 主成分贡献率及累积贡献率

主成分	特征值	贡献率/%	累计贡献率/%
1	7.818	52.12	52.12
2	2.594	17.29	69.41
3	1.766	11.78	81.19
4	1.151	7.68	88.87
5	0.645	4.30	93.17
6	0.472	3.14	96.31
7	0.208	1.39	97.70
8	0.144	0.97	98.67
9	0.116	0.79	99.46
10	0.064	0.47	99.93
11	0.03	0.02	99.95
12	0.02	0.01	99.96
13	0.01	约为 0	100
14	约为 0.	约为 0	100
15	约为 0	约为 0	100

从上表可以看到，前四个主成分的特征值均＞1，前三个主成分的累计贡献率为81.19%，前四个主成分的累计贡献率为 88.87%，因此可以选取前三个主成分或选取前四个主成分来表征 15 个行驶特征参数所包含的信息，既达到降维目的，又有效利用绝大部分有用信息。本论文仅选取前三个主要成分来表征。

6.2.2 基于 kernel PCA 核主成分分析法的特征参数降维

PCA 存在它的局限性，它是一种线性算法。 它不能解释特征之间的复杂多项式关系。于是我们考虑采用 kernel PCA 算法，去解决线性不可分的问题，通过 Kernel 函数将非线性相关转为线性相关。

KPCA 的大致思路是：对于输入空间(Input space)中的矩阵 X ，我们先用一个非线性映射把 X 中的所有样本映射到一个高维甚至是无穷维的空间(称为特征空间，Feature space)，使其线性可分，然后在这个高维空间进行 PCA 降维。

6.2.3 基于 t-SNE 算法的特征参数降维

t-SNE 是基于在邻域图上随机游走的概率分布，可以在数据中找到其结构关系。当数据是强非线性时，这种技术也能很好地工作。它对可视化效果也非常好。

6.2.4 降维结果选择

通过最终结果的特征值比较，我们最终选择 PCA 降维后的四维数据进行后续的数据处理。

6.3 对运动学片段进行聚类分析

6.3.1 改进的 FCM 聚类法

在大数据挖掘处理过程中，聚类分析方法占据着重要的位置。目前主流的聚类分析方法有以下几种：划分聚类、层次聚类、基于密度或网格的聚类，查阅相关资料，对这些聚类方法的各种性质总结，如下表所示：

表 6.2 聚类分析的性质比较

聚类方法	优点	缺点	适用情况
划分聚类	对于大的数据集简单高效，时间空间复杂度均较低	容易得到局部最优解；K 值需预设，结果对 K 点个数选取敏感；对噪声和离群值敏感	中等数量数据集
层次聚类	可以产生高质量聚类；可解释性好；可以用于非球形族	时间复杂度高	小数量数据集
基于密度的聚类	对噪声不敏感；可以发现任意形状的聚类	聚类结果与参数关系紧密；稀疏的聚类或距离近的聚类效果不好	任意族形状
基于网格的聚类	聚类速度快	参数敏感，无法处理不规则分布的数据等；算法效率以聚类结果的准确性高为代价	底层数据密度小

根据文献[2]所述，自组织映射（self-organizing maps，SOM）算法可以通过训练函数得到权值分布，增强样本的可聚类性。在给定聚类中心和聚类数的情况下，FCM 算法可以接近最优聚类。对比上述几种聚类方法与文献所述方法，本文选择使用改进的 FCM 聚类法对运动学片段进行聚类。通过将 SOM 网络所得权值中心作为 FCM 算法初始质心，使得聚类结果更加接近最有聚类，再采用改进 FCM 聚类法得到最终聚类结果。

本文对得到的运动学片段进行聚类分析的总体思路如下图所示：

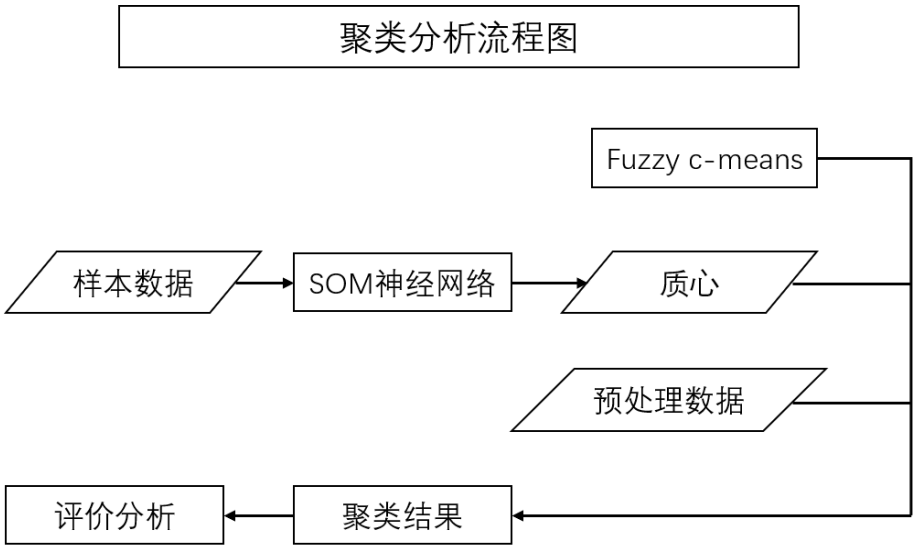


图 6.1 聚类分析流程图

6.3.2 聚类个数

对不同聚类中心个数(2-9 个)的结果进行实验，计算得到相应的模糊划分系数 FPC（Fuzzy partition coefficient），FPC 的数值越接近 1，说明聚类效果越好。根据 FPC 的数值发现，将数据聚类成 3 类拟合出的模型最佳。

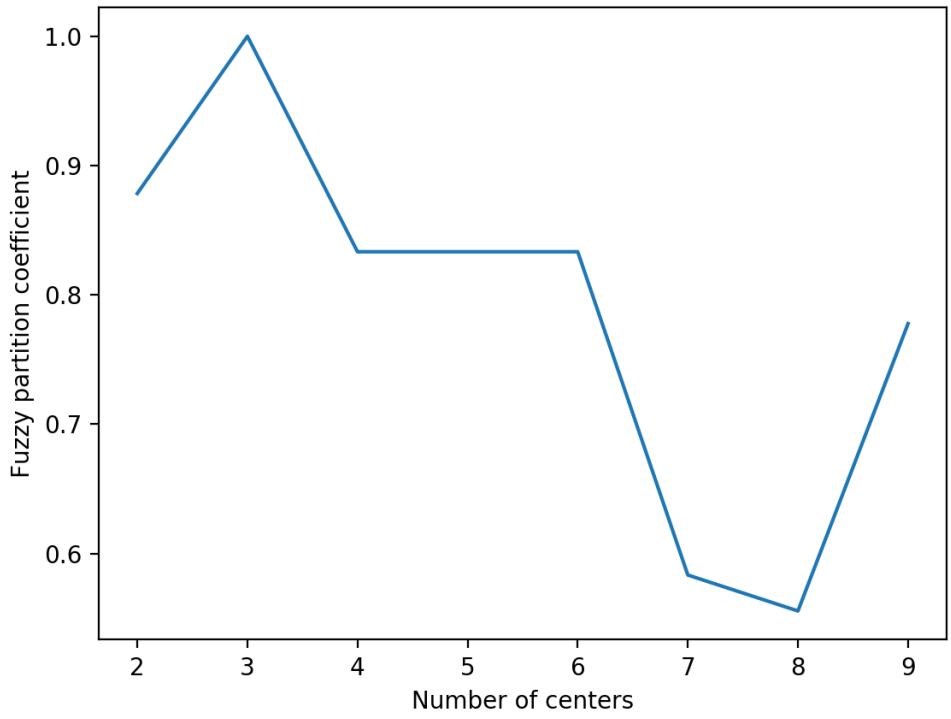


图 6.2 运动学片段聚类结果

6.3.3 聚类结果

由于运动学片段分布在三个文件中，首先将各运动学片段的序号按文件一、文件二、文件三的顺序重新定位，共得到编号 1 到 2712 的 2712 个运动学片段。使用 Python 与 Matlab 编程实现整个聚类过程。

在三维主成分分析的基础上，可以得到个运动学片段在前 3 个主成分上的得分，见下表 6.3.

表 6.3 各运动学片段的 3 个主成分得分

运动学片段序号	第一主成分得分	第二主成分得分	第三主成分得分
1	1.739	-0.759	-0.371
2	-4.787	-1.735	-0.867
3	-0.858	0.257	0.538
4	-3.423	0.200	1.125
…	…	…	…
…	…	…	…
…	…	…	…
…	…	…	…
…	…	…	…
…	…	…	…
2709	3.038	2.849	-3.397
2710	2.947	-0.820	-0.434
2711	-0.618	-0.370	1.143
2712	-1.962	-0.049	2.580

将片段的主成分得分作为新的变量，进行聚类分析。由于奇异样本点有可能会导致后续神经网络训练时间增加或导致网络无法收敛[2]，因此将新的综合变量归一化处理。使用 Matlab 中的 SOM 神经网络工具箱创建 SOM 神经网络。经过改进后的 FCM 聚类，将 2712 个运动学片段分成 3 类，第一类、第二类、第三类分别有 1029、484、1199 个运动学片段。结果如下表所示：

表 6.4 运动学片段聚类结果

	类 1	类 2	类 3
片段序号	1	2	3
	15	4	10
	17	5	11
	18	6	12
	19	7	14
	22	8	16
	•	•	•
	•	•	•
	•	•	•
	2696	2703	2701

	2702 2704 2709	2706 2707 2708	2705 2711 2712
片段总数	1029	484	1199
中心点	[2.413 -0.807, -0.142]	[-3.857, -0.353, 0.0813]	[-0.383, 0.746, 0.179]
类间平均距离	6.29	3.21	3.64
类内平均距离	2.30	4.27	2.44

图 6.3 所示分别为成分 1、成分 2、成分 3 的聚类结果。图 6.4 所示为运动学片段的聚类结果，从图中可以看出，聚类效果较好。

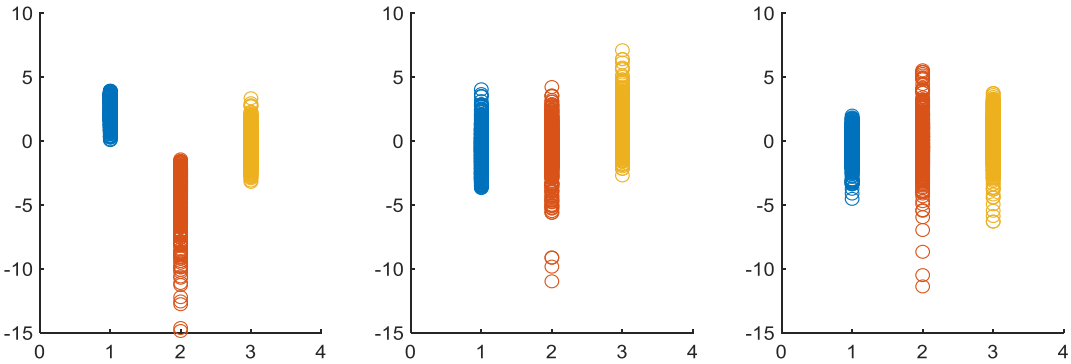


图 6.3 聚类结果

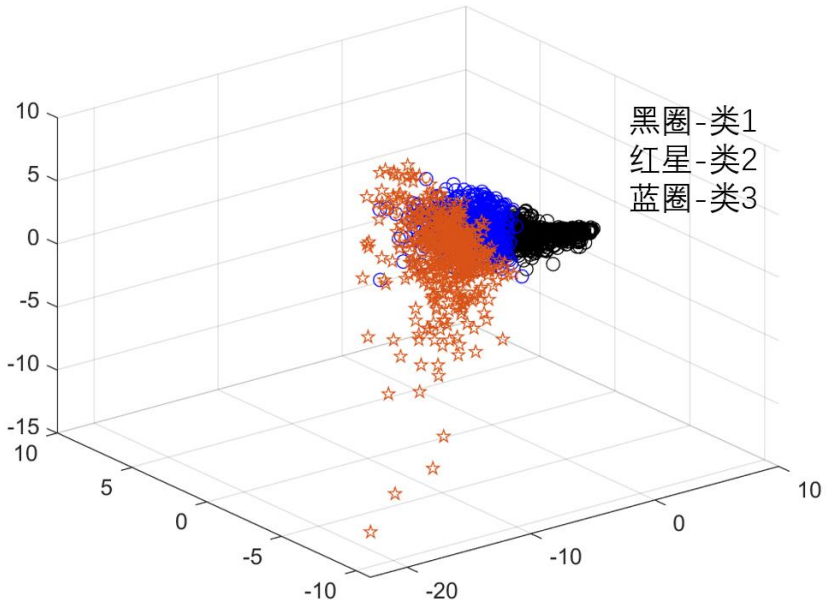


图 6.4 运动学片段聚类结果

三类片段样本集的各特征参数值对比如下表所示，可以看到，三类片段样本集的特征参数差异明显，分别代表了三类不同的行驶特征。类 1、类 2、类 3 的样本时间长度比例分别为 19.4%、40.1%、40.5%；类 1 的怠速占比最高、匀速占比最低，分别为 52.8%和 10.6%，代表了道路拥塞且堵车的汽车行驶状况；类 2 的怠速占比最低、匀速占比最高，分别为 12.6%和 17.8%，代表了道路通畅情况下的汽车行驶状况；类 3 的怠速占比及匀速占比处于中间水平， 代表道路道路繁忙但并未严重拥堵的情况下的汽车行驶状况。

表 6.5 聚类结果分析

聚类	时长占比/%	怠速占比/%	加速占比/%	减速占比/%	匀速占比/%
类 1	19.4	52.8	19.7	16.9	10.6
类 2	40.1	12.6	40.5	29.1	17.8
类 3	40.5	31.0	33.6	24.6	10.8

6.4 基于马尔科夫理论过程构建工况

汽车行驶工况使用速度-时间曲线来表现汽车行驶状况，汽车行驶的规律与道路状况、驾驶员状态等随机因素有关，因此我们可以考虑引入随机过程的概念来构建工况。经过调研和讨论，本文使用文献[3]中构建马尔科夫过程的方法并构建工况。

6.4.1 马尔科夫基本理论

马尔科夫过程是一种无后效性的随机过程，也就是说马尔科夫过程中，下一时刻的状态仅依赖于当前时刻的状态，与之前的状态无关。可用如下公式表示：

$$\mathrm{P}(S_{t+1}|S_t) = \mathrm{P}(S_{t+1}|S_1S_2 \cdots S_t) \tag{6.1}$$

在实际生活中，随机过程随处可见，因此马尔科夫过程的应用十分广泛，在例如价格预测、出行方式预测、人口的增长模型建立等应用中都起到重要作用，由于其时效性等诸多有点，马尔科夫过程已经在生物、物理、化学、计算机等领域获得了广泛应用。汽车行驶规律的预测也存在一些随机因素，因此我们认为构建汽车行驶工况可以使用马尔科夫理论进行分析。

某随机过程在 t0 时刻所处状态，只与 t0 前一时刻所处状态有关，而与再之前的状态无关，即具有无后效性，则称该随机过程为马尔可夫过程。本文对总体采样车速数据在 1、5、100s 不同时间尺度下进行相关性验证分析，结果如图 6.4 所示。图 6.4 表明，间隔 1 s 的车速数据近似于线性，间隔 5 s 的数据相关性减弱，而间隔 100s 的数据则基本不具有相关性。

由下面公式式可定量计算不同时间尺度下的车速相关系数，其中 X 为速度向量，Y 为与 X 间隔一定时间尺度后的速度向量。1、5、100s 不同时间尺度下的车速相关系数见表 1 。

$$\rho_{\mathbf{XY}} = \frac{\mathrm{Cov}(\mathbf{X},\mathbf{Y})}{\sqrt{D(\mathbf{X})}\cdot\sqrt{D(\mathbf{Y})}} \tag{6.2}$$

由图 6.4 可知，随着时间尺度的增加，车速采样数据的相关性降低，即汽车下一时段的运行状态仅与当前时段运行状态有关，而与之前的历史运行状态无关，符合马尔可夫链定义。由此证明在小时间尺度下，本文采样的汽车运行工况数据具备马尔可夫特性，可采用马尔可夫链方法来构建城市公交线路工况。

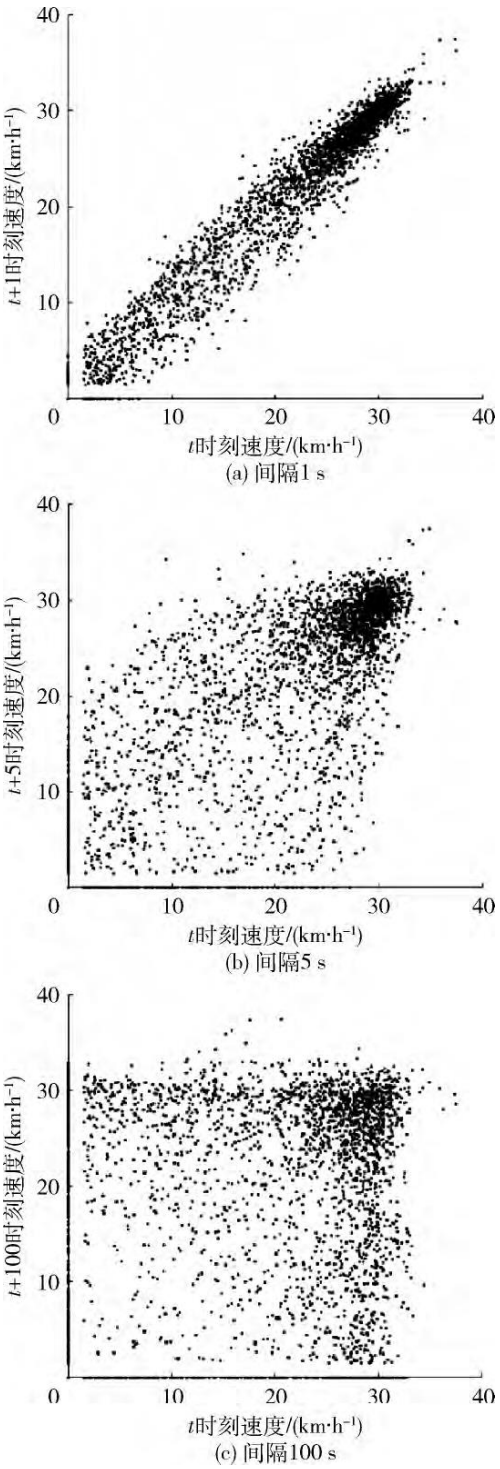


图 6.4 不同时间尺度下的车速相关图

6.4.2 状态划分以及转移概率

参考文献[3]中的做法，我们将运动学片段进行二次划分，以平均速度为划分标准并将划分结果归入马尔科夫链的状态空间，划分标准如下：

1. 平均速度区间为 0-10km/h
2. 平均速度区间为 10-20km/h

3. 平均速度区间为 20-30km/h
4. 平均速度区间为 30-40km/h
5. 平均速度区间为 40-50km/h
6. 平均速度超过 50km/h

获得状态空间后，我们按照以下公式来进行状态转移矩阵的计算：

$$p_{ij} = \frac{N_{ij}}{\sum N_{ij}} \tag{6.3}$$

$$i = 1,2,\cdots 6; j = 1,2,\cdots 6$$

其中N_{ij}表示当前时刻属于状态 i 的模型下一时刻转移到状态 j 的次数。

6.4.3 工况起始片段的选取

基于马尔可夫链的汽车运行工况构建过程中，后一片段的状态取决于当前片段的状态和对应的状态转移概率，故需要构建汽车起步初始段来进行后续工况的构建。车辆起步的运动趋势是从静止开始加速的状态。本文选取与整体试验数据偏差最小的一个运动学片段作为工况起始片段，既保证了信息的完整性，也能够使构建的行驶工况最大程度地反映整体试验数据的数据特点。

首先，对运动学片段所有特征参数进行无量纲化，具体计算公式如下：

$$z_{ij,k} = \frac{\left(x_{ij,k} - \min x_{j,k}\right)}{\left(\max x_{j,k} - \min x_{j,k}\right)} \tag{6.4}$$

$$i = 1,2,\cdots m_k; j = 1,2,\cdots p; k = 1,2,\cdots n$$

其次对整体数据的所有特征参数求平均值，同样进行无量纲化，然后计算每一个运动学片段的特征参数相对于整体试验数据平均水平的误差的平方和，计算公式如下：

$$M = \sum_{j=1}^{p}\left(z_{ij} - \overline{z_j}\right)^2, i = 1,2,3\cdots 3148 \tag{6.5}$$

6.4.4 工况中间片段的选取

通过上一小节选出的工况起始片段，以及状态转移方程，可以确定中间片段的起始状态和起始点的速度，即工况起始片段的最终数据点所处的状态为中间片段的起始状态，而起始片段的最终数据点的速度值就是中间片段的起始速度。

后续片段的选取采用随机事件模拟来确定，即产生一个随机数，使它服从一个已知的概率分布，这里指的是马尔可夫过程中的状态转移概率矩阵，根据随机数的取值范围，来判断下一片段所属状态。

假设当前工况片段所属状态为 i，则可以得到该状态转移到各个状态的概率值，也就是转移矩阵中第 i 行的转移概率。因此，在$(0,1)$区间内取一个随机数 r，若 r 满足：

$$\sum_{j=1}^{k-1} P_{ij} < r < \sum_{j=1}^{k} P_{ij} \tag{6.5}$$

则可判定下一片段的状态为 k。

6.4.5 工况结尾片段的选取

先从所有运动学片段中选取备选结尾片段，然后利用最小二乘法原理挑选与整体试验数据的误差平方和较小的片段作为目标结尾片段。

6.4.6 代表性工况的构建和结果分析

通常构建的典型城市循环工况时长T 约1 200 s，可通过各类总运行时间在总体数据中所占的时间比确定各类在最终拟合工况中所占的时间，定义为

$$T_k = \frac{T_{duringcondition}}{T_{all}} \sum_{i=1}^{N_k} T_{k,i'} \tag{6.6}$$

式中: Tk为第k类在最终合成车辆行驶工况中所占时间; Tall为整体样本所持续时间; Tduringcondition为所要构建车辆行驶工况持续的时间; Nk为第k类中短行程的总条数; Tk，i为第k类中第i条运动学片段的运行时间。

保证马尔可夫链选取的工况片段必须能够足够各类总运行时间在总体数据中所占的时间比，不然需要重新构建汽车工况。

通过上述公式计算取整后，获得5条I类低速片段，3条 II 类中速片段，3条III类高速片段，最终构建出持续时间为1258s、最高车速为61. 9 km/(h*s)的典型车辆行驶工况，如图所示。

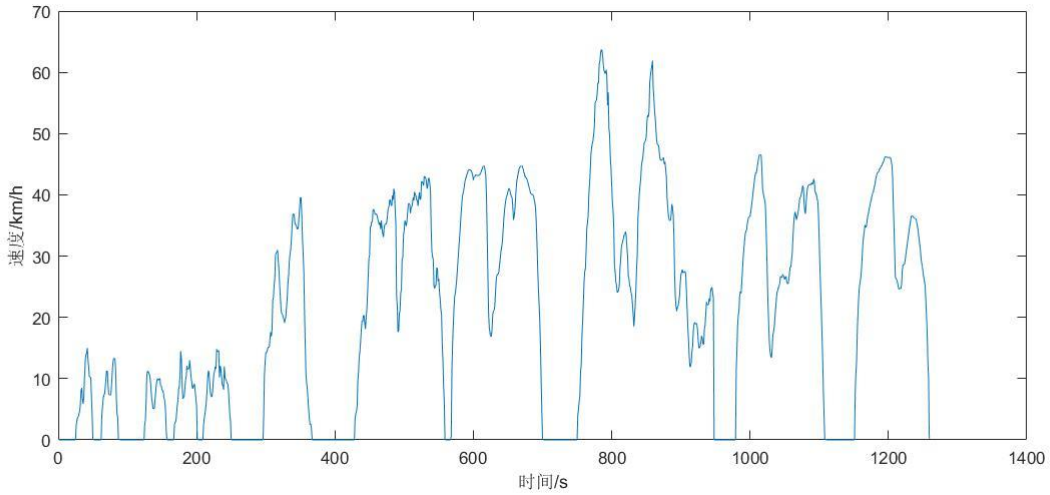


图6.5 最终代表性工况

我们对该代表性工况还进行了加速度的时间分布分析：

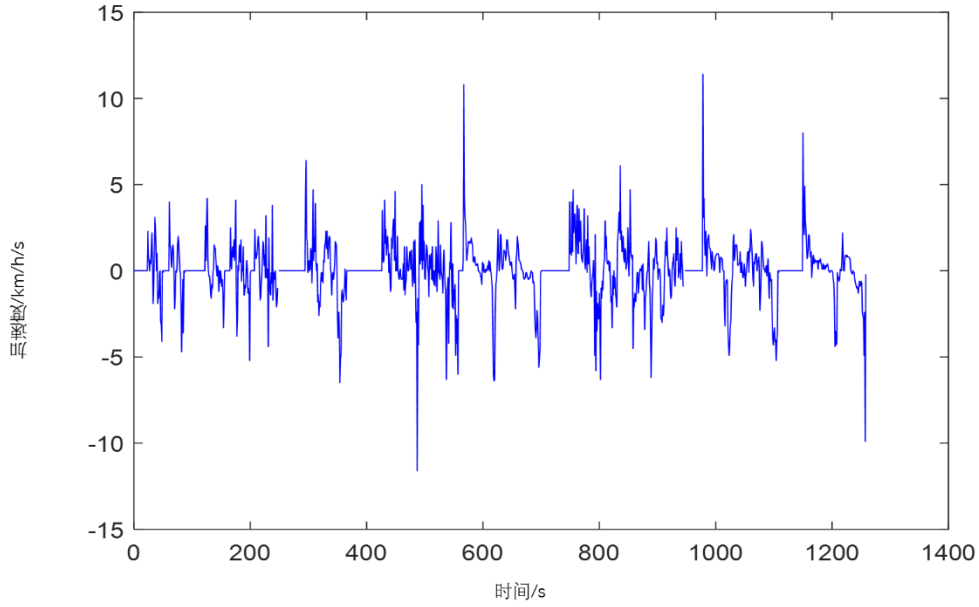


图 6.6 代表性工况的加速度分布

可以看到，我们的代表性工况整体的加速度均分布在-14km/h 至 14km/h 之间，没有异常的加速或减速行为，这也符合我们对城市道路状况的分析和理解。除了少数加速度的尖锐变化，多数时间加速度变化较小，汽车速度趋向平稳，这也是城市道路中汽车行驶的表征之一。

除此之外，我们还对该代表性工况的加速度-速度的联合分布情况进行了分析，如图 6.7 所示。

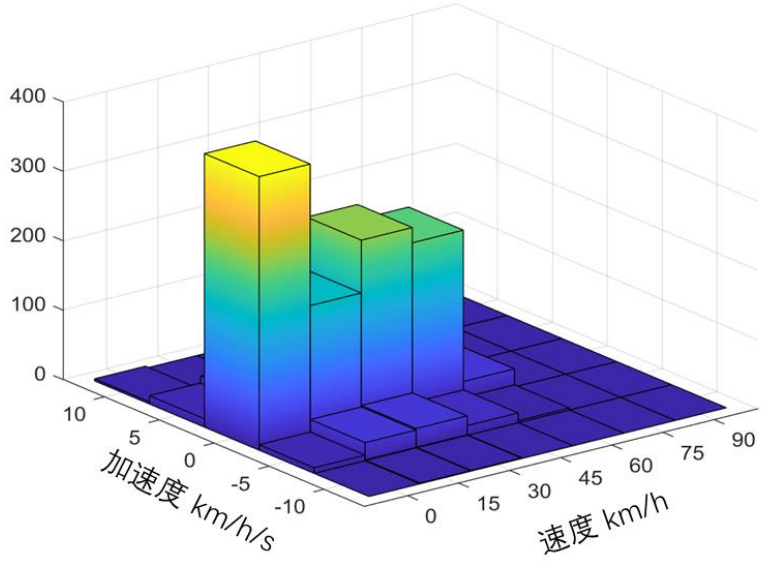


图 6.7 代表性工况的加速度-速度的联合分布

可以看出，该分布与原始数据特征一致。

6.4.7 误差分析

为了验证本文获得的工况的代表性，我们对代表性工况及原始数据的特征参数进行了计算和误差分析，如下表所示：

表 6.6 代表性工况与原始数据误差分析

	平均速度（km/h）	平均加速度（km/h/s）	平均减速度（km/h/s）	怠速占比	加速占比	减速占比	匀速占比	平均值
实际	27.3	1.26	1.26	23.85	32.29	26.89	16.62	
工况	25.77	1.28	1.22	24.12	32.84	26.45	16.11	
误差(%)	5.604	1.587	3.175	1.132	1.703	1.636	3.069	2.558

根据上表可以看到，代表性工况与实际数据误差最大的特征值为平均速度，误差仅为5.604%，而平均加速度、平均减速度、怠速占比、加速占比、减速占比以及匀速占比的误差均在 5%以下，同时平均加速度、怠速占比、加速占比和减速占比的误差都低于 2%，其中怠速占比的误差最小，仅为 1.132%。代表性工况与实际数据的平均误差仅为 2.558%，这说明我们的代表性工况可以比较好的代表实际数据的特征，能够体现本城市汽车的行驶状况。

参考文献

[1] 李宁. 城市道路车辆行驶工况的构建与研究[D]. 河北农业大学, 2013.

[2] 陈弘, 刘海, 乔胜华, et al. 基于三次样条插值的车辆行驶数据分析[J]. 汽车技术, 2013(8):54-57.

[3] 李洋. 基于聚类算法的汽车行驶工况研究[D].

[4] 刘应吉，夏鸿文，姚羽，等. 组合主成分分析和模糊 c 均值聚类的车辆行驶工况制定方法 [J] .

[5] 高建平, 丁伟, 孙中博, et al. 基于 PCA 和 FCM 的汽车行驶工况研究与构建[C]// 第十三届河南省汽车工程科技学术研讨会论文集. 2016.

[6] 苗强, 孙强, 白书战, et al. 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(11).

[7] 高建平, 任德轩, 郗建国. 基于全局 K-means 聚类算法的汽车行驶工况构建[J]. 河南理工大学学报(自然科学版), 2019, 38(01):117-123.

[8] 张富兴, 李孟良, 乔维高, et al. 车辆行驶工况运动学水平的研究[J]. 武汉理工大学学报(交通科学与工程版), 2005, 29(5):667-670.

[9] 丛孟营, 李庆磊, 徐永平. 城市客车行驶工况数据分析[J]. 客车技术与研究, 2014(3):59-62.

[10] 李宁. 城市道路车辆行驶工况的构建与研究[D]. 河北农业大学, 2013.

[11] 石琴, 郑与波, 姜平. 基于运动学片段的城市道路行驶工况的研究[J]. 汽车工程, 2011, 33(3):256-261.

[12] 刘应吉, 夏鸿文, 姚羽, et al. 组合主成分分析和模糊 c 均值聚类的车辆行驶工况制定方法[J]. 公路交通科技, 2018.