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	西安交通大学 同济大学，上海师范大学
参赛队号	19106980006
队员姓名	1.李旸
	2.焦玥明
	3.谢明含

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 基于运动学片段划分的行驶工况构建方法研究

摘 要：

汽车行驶工况（Driving Cycle）又称车辆测试循环，是描述汽车行驶的速度-时间曲线，体现汽车道路行驶的运动学特征，是汽车行业的一项重要的共性基础技术，是车辆能耗/排放测试方法和限值标准的基础，也是汽车各项性能指标标定优化时的主要基准。本文以已知汽车行驶数据为基础，研究构建汽车工况曲线的问题。针对粗糙复杂数据集的预处理问题，对数据中的缺失数据，利用线性函数，对缺失数据进行整体填补，针对不同的数据处理问题，合理选择问题处理的次序。在经过处理的数据的基础上，首先根据记录的速度、加速度和发动机状态确定记录的工况状态。然后根据一次怠速过程开始到另一次 怠速过程开始的标准将所有记录划分成运动学片段。

为构建工况曲线拟合体系，首先构造例如平均速度、平均行驶速度、平均加速度、平均减速度、怠速时间比等描述运动学片段的特征参数。在得到运动学片段的特征矩阵后，利用 PCA 计算特征值的重要性来确定相应的主成分。利用 k-means 聚类算法，对不同的运动学片段分类。最后在不同的分类中依据均方误差最小来确定最能代表类别的运动学片段，拟合行驶工况曲线。最终将拟合的曲线与真实工况的特征进行比较，说明了构造方式的可行性。

汽车行驶工况构建

1 问题介绍

1.1 问题背景

汽车行驶工况（Driving Circle）也被称作为运转循环，是汽车工业领域的一项共性基础技术，该技术通过对车辆的行驶状况进行调查和分析，可以反映车辆的真实操作工况，是汽车各项性能指标检测及优化的重要数据支撑，也是推动汽车技术发展的重要因素之一。依其形态，行驶工况可分为瞬态工况和模态工况两种，其中模态工况的典型代表 NEDC（New European Driving Cycle）为我国一直所采用[1]。但随着新能源车的出现，该工况在面对诸如新型车辆怠速启动等新技术的问题时也逐渐变得捉襟见肘。此外由于不同地区的交通条件不同，同一个工况运用于实际情况时会产生巨大差异，这种差异会使得同一型号的汽车在不同地区的运行过程中表现出不同的性能，特别是汽车的燃料经济性和排放性差别较大[2]。所以总体来看该工况并不符合我国国情。然而随着社会的发展，我国在未来一段时间内汽车保有量会不断增长，据交通运输部预测，截止 2020 年我国汽车保有量将超过 2 亿辆[3]。此外构建较为精准的汽车行驶工况对于交通协同控制，新车仿真，技术开发，风险评估等都具有十分重要的现实意义，因此当务之急便是开发出适合我国的汽车行驶工况。如何从数学建模的角度，减少数据误差，优化汽车工况曲线是本文要解决的问题。

1.2 国内外研究现状

自世纪 20 年代起，世界各国就开始了对汽车行驶工况的研究： Kent， Allen 及 Rule（1978）通过计算车速数据，以确定悉尼的行驶工况与车辆污染物排放量的关系[4]。Lin 及 Niemeier（2002）重新讨论了统一循环(LA92 驱动循环)施工方法,并将行车轨迹当作一个随机过程构建了代表性行驶工况[5]。Ball, Owsley（1998） 等采用微路径和速度—加速度的方法对自我调节在提高老年驾驶员安全方面的潜在作用进行了评估[6]。之后的研究者又对影响行驶工况的各因素进行了分析，如：驾驶员性别，汽车结构，天气因素，交通饱和度等，其中 Ericsson（2001）从实际采集到的 19230 种驾驶模式中计算出 62 种驾驶模式参数，并对其进行回归分析以研究驾驶模式与燃料使用及排放之间的关系[7]。

随着对行驶工况研究的深入，各国的典型工况也逐渐多了起来。1966 年，美国加利福尼亚州就提出了世界上最早的行驶工况“7 工况法”[8]，随后又推出了瞬态工况（Transient Driving Cycle）—联邦测试程序（FTP75）及负荷模拟工况（IM240）。欧洲使用的工况包括市内循环（ECE）和市郊循环（EUDC），属于稳态循环。日本主要有三种工况，即 10 工况，11 工况和 10-15 工况，日本的行驶工况相对欧美工况而言相对简单，主要包括匀速，匀加速和匀加速这三个数据。其他地区工况还有悉尼行驶工况，印度 4Mode 行驶工况，墨尔本行驶工况等[9]。

由于我国汽车工业起步较晚，相关研究也比国外落后。从 90 年代末开始，我国的科研人员才开始以国内典型城市为对象进行研究。例如赵慧和张镇顺（2000）将可以测量汽车行驶速度及尾气排放浓度的数据采集系统安装在一辆轻型车上，以中环和九龙为实验对象构建出一个基于香港实际的行驶工况[10]。杜爱民，步曦等（2006）以上海的 25 条公交路线

为研究对象，运用主成分分析及聚类分析等方法构建出上海公交车的高速，低速及综合行驶工况[11]。訾琨，黄永青等（2006）通过收集到的大量实际道路汽车行驶瞬时速度数据，利用修正的 MOBILE 模型研究得出机动车车型和汽车城市道路运行工况对机动车排放污染物的影响[12]。马冬，丁焰（2008）等研究了轻型汽车的实际行驶工况的排放问题[13]。王岐东，贺克斌等（2010）采用 GPS 及多普勒速度仪对北京，长春等 8 个不同规模的城市的汽车行驶工况进行了测试，发现中国城市特征与欧美工况相比，在行驶模式分布、平均车速等方面均存在很大差异[14]。

总体而言，目前国内的主要研究方法可以概括为四种类型：短行程法，行程片段法，定步长截取法和 V-A 矩阵分析法，而研究内容也涵盖了很多方面，主要包括道路试验与数据采集，行驶工况的构建方法研究及对污染物排放的研究。汽车行驶工况本身是一项基础研究，研究开发符合我国特定区域的汽车行驶工况尤为必要，并且极具现实意义。

1.3 待解决的问题

针对题目所提供的数据，本文依次解决以下汽车行驶工况构建问题。

问题一 针对原始采集数据中出现的时间不连续，加、减速度等数据异常异常，怠速时间异常等包含不良数据的情况，使用数据填充和删除等方法对原始数据进行补充和处理，最终得到连续、合理的数据。

问题二 在问题一得到的干净数据的基础上，将驾驶数据按照速度，加速度和发动机状态分为怠速工况，加速工况，减速工况和匀速工况。然后进一步依据怠速工况将行驶数据划分为若干运动学片段。

问题三 根据问题二得到的运动学片段，设计描述汽车行驶状态的特征参数，构建汽车行驶工况及汽车运动特征评估体系。将运动学片段按照特征参数进行聚类，最终选出具有代表性的片段，根据片段构建汽车行驶工况曲线，并比较特征与真实数据特征，证明合理性。

2 问题分析和求解

2.1 问题一 ：数据预处理

2.1.1 数据的整体概况：

样本数据主要是由三个文件构成，为了确定样本文件的采样地点，采样方法是否基本相同，以车辆 GPS 车速作为强特征，大致分析了样本文件的数据情况。

表 1. 样本 GPS 车速统计学信息

项目	样本 1	样本 2	样本 3
样本数	185725	145825	164914
均值	26.65	23.77	30.15
标准偏差	23.91	22.55	29.09
最小值	0	0	0
第一四分位数	2.1	0.1	0
中位数	24.0	20.2	25.8
第三四分位数	41.7	40.3	47.9
最大值	111.5	116.6	261.4

车速作为车辆行驶状况最直接的体现，三个样本中，样本 2 的车速均值和车速标准偏

差整体较低，车辆在样本 2 的测试状态中整体较为平稳，而样本三中车辆的测试差异性较大，应该是路况比较复杂，就整体数据对比来看，三种样本没有特别明显的数据偏差倾向，采样基本上保证了平等，随机的原则。

2.1.2 数据异常情况分析

数据异常情况主要包括两个方面：

1.测量数据的丢失，主要原因是车辆在进入隧道等建筑覆盖场所中行驶信息数据的丢失以及车辆在停车后手动关闭采集设备的情况。

2.数据异常，数据异常包括两个方面：持续时间大于 180s 的怠速数据和加速度异常情况（加速阶段加速度大于 3.96m/s^2 和减速阶段加速度大于 8m/s^2）。

下面讲主要就这两方面，提供数据预处理的方法以及相应的结果展示。

1. 缺失数据情况分析

三个样本文件中，提供的数据都是基于采样时间递增记录的，对于正常情况来看，采样周期为 1s，当存在缺失情况时，两条记录时间差值必定大于 1s，根据这个方法，我们可以定位到发生数据缺失的具体记录，同时绘制相应的数据数据缺失图（如图 Fig1，Fig2，Fig3），来为下一步数据处理做好准备。

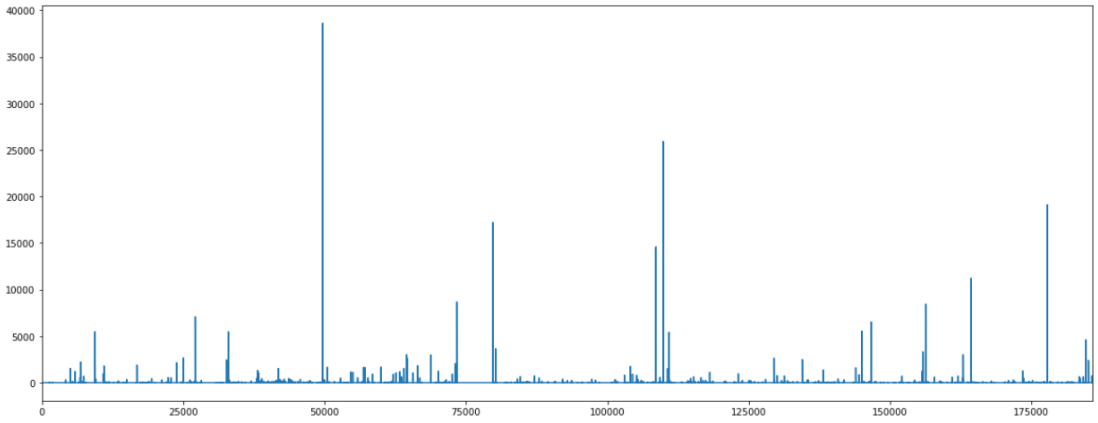


图 1. 文件 1 中数据缺失情况

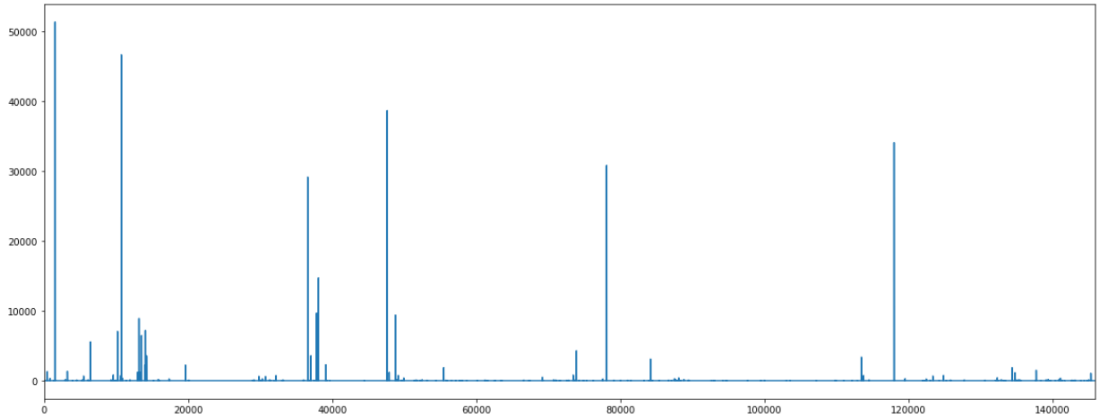


图 2. 文件 2 中数据缺失情况

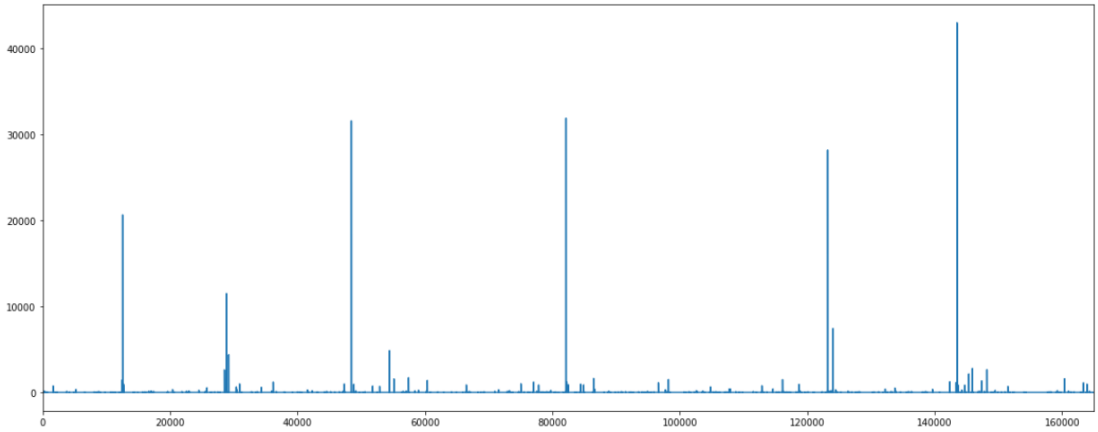


图 3. 文件 3 中数据缺失情况

数据中的 X 轴为各个文件中数据记录的行号，而 Y 轴则是出现数据缺失位置对于的数据缺失数量，即两次数据记录之间的时间差值，每个文件都有缺失上百秒数据的情况出现，缺失最大值都在 40000 以上，对于缺失值很小的情况可以采取手动填补的方法，而大于 180s 的缺失情况，对两端数据进行填补之后，当作停车处理。

2. 数据填充：

针对缺失数据的填充，主要有三个方法，1：固定值填充方法 2.基于算法的填充方法 3 .基于函数关系填充。

固定值填充方法主要是使用 0，缺失特征列值的中位数，众数，平均值等离散值来填充，在本问题中，车辆的行驶时间是一系列连续的时序序列，使用离散的数据来填充缺失值，不仅会破坏原数据连续的特性，同时部分填充值会导致测量数据的加速度或其他关联数据异常，不能作为本问题的填充方法。基于算法的填充方法主要是利用已知的监督学习或者非监督学习，通过训练模型，来对缺失值进行预测并填充。本问题中，缺失值并不是一个样本部分缺失，而是整条样本丢失，所以利用算法没有合理的模型去生成整条样本信息。基于函数关系的填充，本问题的解决主要是以靠这个基于函数关系的填充方法，在汽车行驶过程中，汽车的后一状态和前一状态联系紧密，而在丢失数据的情况下，我们假设汽车的当前状态和前一状态之间是均匀变化，也就是构建一个线性模型来对丢失数据进行补充。

1）模型假设：

假设 1：汽车在丢失数据的过程中，并没有突发时间发生，车辆状态趋于平稳。

假设 2: 汽车在行驶过程中，除了 GPS 车速变化，其他特征信息变化也是线性平稳变化，不存在突变可能。

2）符号定义

符号	意义
X_i	本样本数据集中第 i 条车辆行驶记录
x_i^j	本样本数据集中第 i 条车辆行驶记录的第 j 个特征

T_i	本样本数据集中第 i 条车辆形式记录的记录时间
a_i^j	本样本数据集第 i 条记录和第 i-1 条记录之间第 j 个特征的改变速率

对于存在数据丢失的行驶记录X_i通过与第X_{i-1}条行驶记录的相关特征作差，可以得到：

$\Delta T_i = T_i - T_{i-1}$　　ΔT_i:缺失记录的具体时长

由于题目所给的数据默认采样频率为 1Hz，即每隔一秒进行一次采样，因此本文默认时间间隔为 1 秒。所以两次记录之间需要补充的记录数量为：

$$\mathrm{N} = \Delta T_i / 1$$

根据记录数量和两次记录之间的时间差，可以计算出每个特征的变化率为

$$a_i^j = \frac{x_i^j - x_{i-1}^j}{\Delta T_i}$$

，其中x_i^j表示第i条记录的第j个特征的值。再得到新添加的样本数据为

$$X_{new} = x_i^j + a_i^j * A \quad A = 1,2 \ldots . . i+1$$

至此，数据集中丢失时间间隔小于 180s 的数据已经填充完毕，题目中由于高层建筑覆盖造成的记录丢失问题已经解决。

3. 长时间数据缺失问题：

前文介绍过，数据丢失除了高层覆盖问题，还有采集设备关闭造成的数据丢失，而因为采集设备关闭造成的数据丢失往往数据丢失数量多，持续时间长，在一般行车中，我们假设如果连续超过 180s 出现数据丢失的情况，就考虑车辆停止，或者驶入高层建筑覆盖区域之后停止，针对连续缺失情况，我们对缺失数据除进行阶段，并在两端进行填补，根据题目中的描述,以及前文提到的符号描述，汽车在加速阶段的加速度 $a_+ < 3.96 m/s^2$ 而在减速阶段汽车加速度为 $a_- < 8 m/s^2$,对于大于 180s 的连续丢失情况X_i和X_{i+1},假设他们的车速为V_i和V_{i+1}，则需要填补在X_i之后的数据数量 N+和需要填补在X_{i+1} 之前的数据两 N-分别为：

$N_- = [V_i / \max(a_+)] + 1$

$N_+ = [V_{i+1} / \max(a_-)] + 1$

则根据 N+和 N- 能够计算出对应特征 j 的变化速率a_i^j:

$a_{i-}^j = x_i^j / N_-$　　或　　$a_{i+}^j = x_i^j / N_+$

得到a_i^j，得到需要补充的数据X_{new}：

$X_{new} = x_i^j + a_{i-}^j * A$　　A = 1,2,3….N_-

至此，数据集中丢失时间间隔大于 180s 的数据完成了先切割再补充的过程，把数据分解成两个不同的运动周七，题目中长时间数据缺失的情况已经完成。

4. 加速度异常：

针对题目中加速度异常情况的处理，放在了第四步，如果在之前为对缺失值填充，以及车辆停止情况进行判断，就去计算加速度，会出现较多异常值，影响后续判断，再对其他异常问题处理之后，再处理加速度异常问题，不仅能减少工作量，也能提升数据集精确度。之前处理中，数据中缺失的数据已经填补，所以相邻两条数据之间的时间差一定是 1s，根据记录中相邻数据之间的 GPS 车速差，能够算出当前的加速度，得到关键特征加速度。根据题意，对加速度异常的数据进行标记分析。

在现实中，车辆的加速度受到车辆性能和路况的影响，并不会突然出现异常，所以在数据中加速度异常值的出现主要是由于采集设备的问题，针对这种异常问题，处理方法是将加速度异常的数据当成缺失值，使用前文提到的缺失值填补规则，对异常加速度记录进行重写，已满足要求。

5. 怠速时间处理：

怠速时间是指汽车怠速工况持续的时间，怠速工况：将发动机工作且车辆速度为 0 时的工况称为怠速工况。这种工况下，发动机转速比较低，进气少，燃烧室内的剩余废气比例较大，需要较浓的混合气体使得发动机稳定工作。所以通过数据可以发现，怠速状态下瞬时油耗时行驶中顺势油耗的百倍，可以作为判断怠速时间的一个特征。

怠速时间处理放在了最后，怠速时间主要包括两部分，一部分是车辆拥堵造成的速度缓慢从而出现怠速过程，另一部分是因为车辆停止但是发动机未关闭而出现的怠速过程，前者属于正常车辆行驶过程的基本情况，可以代 2.基本的一种路况信息，而后者则是人为因素，对汽车行驶工况构建没有帮助，题中说明，怠速过程大于 180s 既为异常情况，所以当怠速过程出现异常时候，就按停车处理，此时车辆的一次行驶周期完成，即将进入下一行驶周期，这为第二问切割运动学片段提供了理论指导。因为在 GPS 车速小于 2.8m/s 的情况下，认为车辆处于怠速工况，所以对处于怠速工况下的 GPS 车速全部置 0，为后面的问题处理提供遍历。

2.1.3 问题求解结果

通过对数据处理之后，在原有数据的基础上，添加了约 7 万条数据，添加之后各个文件对应的数目如表 2

表 2 数据处理前后记录数量统计

	文件 1	文件 2	文件 3	总和
原始数据	185725	145825	164915	496465
处理后数据	216007	161688	191255	568950

在处理完数据之后，部分特征起的作用比较有限，数据主要保留了部分有用的运动学特征为后续问题的解决提供遍历，保留的数据样表类似于表 3：

表 3.整理后基本运动特征数据样例

时间	车速	加速度	时间	车速	加速度
0	0	0	21	15.8	0.027778
1	6.1	1.69	22	15.2	-0.16667
2	8.5	0.666667	23	15.2	0
3	9.8	0.361111	24	14.7	-0.13889
4	12.7	0.805556	25	13.6	-0.30556
5	13.1	0.111111	26	12.6	-0.27778
6	13.6	0.138889	27	12	-0.16667
7	13.5	-0.02778	28	11.6	-0.11111

8	11.1	-0.66667	29	12	0.111111
9	8.8	-0.63889	30	13.8	0.5
10	6.5	-0.63889	31	15	0.333333
11	6	-0.13889	32	15.3	0.083333
12	7.3	0.361111	33	15.4	0.027778
13	9.3	0.555556	34	15.3	-0.02778
14	11.5	0.611111	35	14.4	-0.25
15	13	0.416667	36	13.9	-0.13889
16	14.2	0.333333	37	12.8	-0.30556
17	15.6	0.388889	38	12.1	-0.19444
18	16.2	0.166667	39	11.8	-0.08333
19	16	-0.05556	40	11.2	-0.16667
20	15.7	-0.08333	41	9.8	-0.38889
			42	9.1	-0.19444
		.	43	-10.7	-6
			44	0	.2.7222

对应片段的速度变化曲线如图 Fig4 所示。

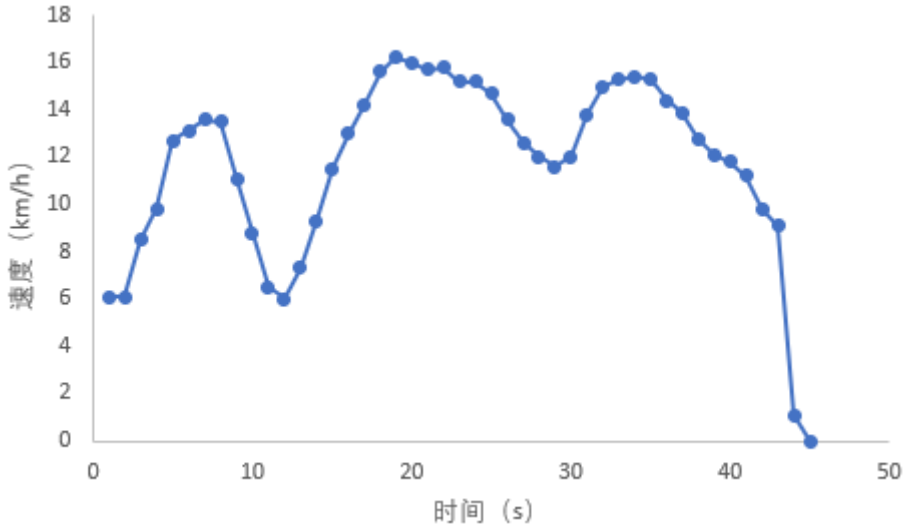


图 4. 预处理后的速度-时间变化曲线

在表 3 中展现的是随机选择的一个行驶过程片段，可以看到经过处理后的数据表比较完整，特征变化比较平整，各个环节符合固有经验，数据翔实，说明与处理方式可以有效的处理数据丢失和数值异常等情况。

2.2 问题二：运动学片段的划分

车辆运动的过程是由大量不同的怠速、加速、减速和匀速过程组成，将从一个怠速开始到下一个怠速开始的运动片段或者从一个怠速结束到下一个怠速结束的运动片段，定义为运动学特征片段，运动学片段包含了车辆的道路运行特征的全部信息。问题一中，数

据得到了合理的补充，在怠速数据处理问题中，也利用怠速时间，成功的识别了车辆是否停车等待，并已此为依据截取了完整的运动周期展现在表 3 中。所以问题二问题简化两部分，第一部分，确定在一次完整的车辆运行周期中，各个记录的工况状态（怠速，加速，匀速，减速）。第二部分，在确定工况状态之后，确定相应的运动学片段，并将运动学片段进行划分。

2.2.1 确定具体的工况状态

根据四种工况状态定义，怠速状态是指汽车行驶速度为 0 且加速度小于 0.1m/s^2 并大于-0.1m/s^2 的情况，加速状态是指汽车加速阶段加速度大于 0.1m/s^2 的情况，匀速状态是指汽车行驶速度不为 0 但加速度小于 0.1m/s^2 并大于-0.1m/s^2 的情况，而减速状态是指减速阶段加速度小于-0.1m/s^2 的情况。利用具体的工况状态定义，以及第一问中整理好的基本运动学信息表（参考表 3），对汽车的一个运动周期中的每条记录的工况给予标记，如表 3 经过标记后信息就以表 4 的标准格式储存。

表 4 带工况标记基本运动特征数据样例

时间	车速	加速度	工况
0	0	0	0
1	6.1	1.69	0
2	8.5	0.666667	0
3	9.8	0.361111	0
4	12.7	0.805556	1
5	13.1	0.111111	1
6	13.6	0.138889	1
7	13.5	-0.02778	3
8	11.1	-0.66667	2
9	8.8	-0.63889	0
10	6.5	-0.63889	0
11	6	-0.13889	0
12	7.3	0.361111	0
13	9.3	0.555556	0
14	11.5	0.611111	1
15	13	0.416667	1
16	14.2	0.333333	1
17	15.6	0.388889	1
18	16.2	0.166667	1
19	16	-0.05556	3
20	15.7	-0.08333	3
21	15.8	0.027778	3
22	15.2	-0.16667	2
23	15.2	0	3
24	14.7	-0.13889	2
25	13.6	-0.30556	2
26	12.6	-0.27778	2
27	12	-0.16667	2
28	11.6	-0.11111	2

29	12	0.111111	1
30	13.8	0.5	1
31	15	0.333333	1
32	15.3	0.083333	3
33	15.4	0.027778	3
34	15.3	-0.02778	3
35	14.4	-0.25	2
36	13.9	-0.13889	2
37	12.8	-0.30556	2
38	12.1	-0.19444	2
39	11.8	-0.08333	3
40	11.2	-0.16667	2
41	9.8	-0.38889	0
42	9.1	-0.19444	0
43	-10.7	-6	0
44	0	.2.7222	0

表 4 中，工况一列的列值就代表每条记录的具体工况类型，0 代表怠速工况，1 代表加速工况，2 代表减速工况，3 代表匀速工况。

2.2.2 算法模型建立

本文采用将从一个怠速状态开始到下一个怠速状态开始之间的片段定义为一个运动学特征片段,并以此将问题一中得到的处理后的汽车行驶数据进行划分。从表 3 中能够看到，工况一列表示了车辆在行驶过程中的各种状态。基本的运动学片段划分思路是逐条处理每个数据，依据数据处于的不同工况状态将其加入正确的运动学统计片段中。

对于每条处于怠速工况的数据，首先需要判断其是属于运动学片段起步阶段还是结束状态，判断的方法包括：

1） 是一个运动周期中的第一条怠速状态数据；

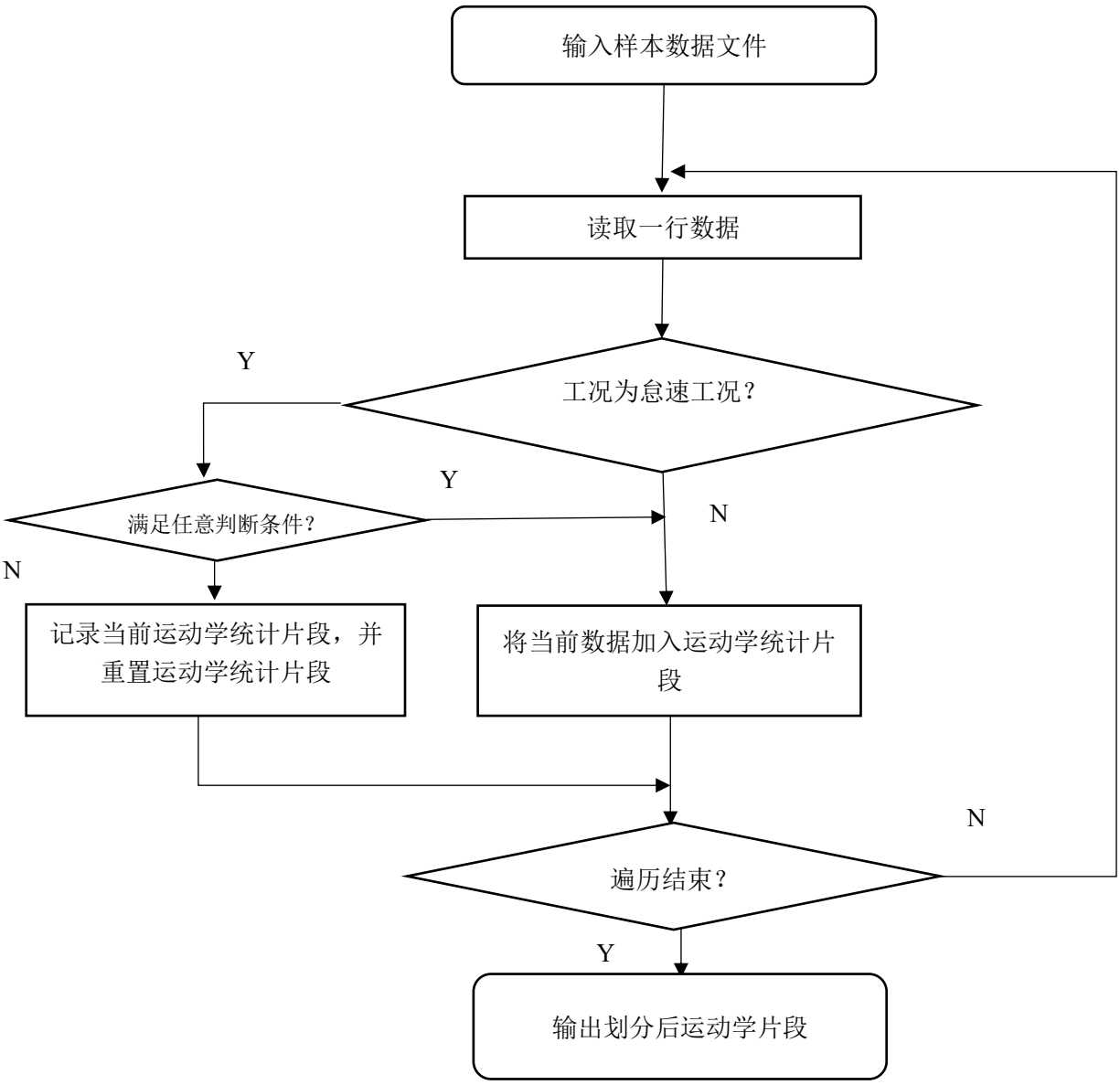
2） 后一秒的数据是也处于怠速状态；

3） 当前运动学统计片段中没有数据。

如果满足以上任意一个条件，则说明本条数据属于当前运动学片段，将其划归运动学片段中，如果均不满足，则说明这条数据属于下一个运动学片段的起步阶段，将当前运动学片段输出，并重新构建新的空运动学片段，将这条数据放入其中。

对于非怠速工况数据，如果当前运动学片段不为空，说明数据处于一个有效运动学片段中，则将其划归当前运动学片段。

2.2.2 算法流程图



2.3.3 问题求解结果

通过对数据切割之后，将原有的基本数据切割为若干运动学片段，三个文件中切割出的运动学片段总和为 4394，各文件对应数目如下。

项目	文件 1	文件 2	文件 3	总和
运动学片段数	1749	1318	1327	4394

运动学片段是指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，下面六张图分别是从文件 1，文件 2 和文件 3 中提取的运动片片段样例，通过工况（Condition）图像可以看到，整个运动学片段符合要求，运动过程中只在开头出现了怠速状态，再第二次遇到怠速状态后，立即截断，运动轨迹也比较平滑，符合人对速度曲线的基本尝试，但是部分运动学片段因为持续时间太短，不能很好的反映车辆再行驶中的工况特性，所以不能很好的对下一问解答提供帮助，接下来仍要筛选符合要求的运动学片段。

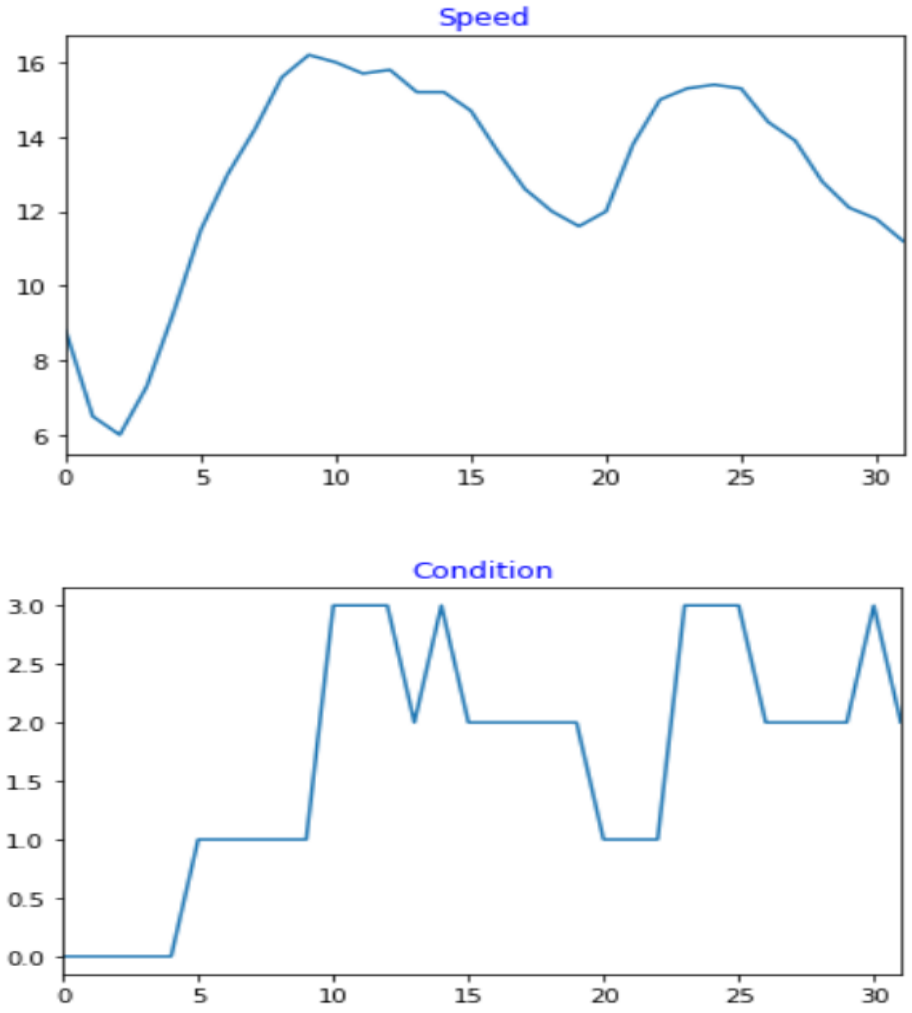
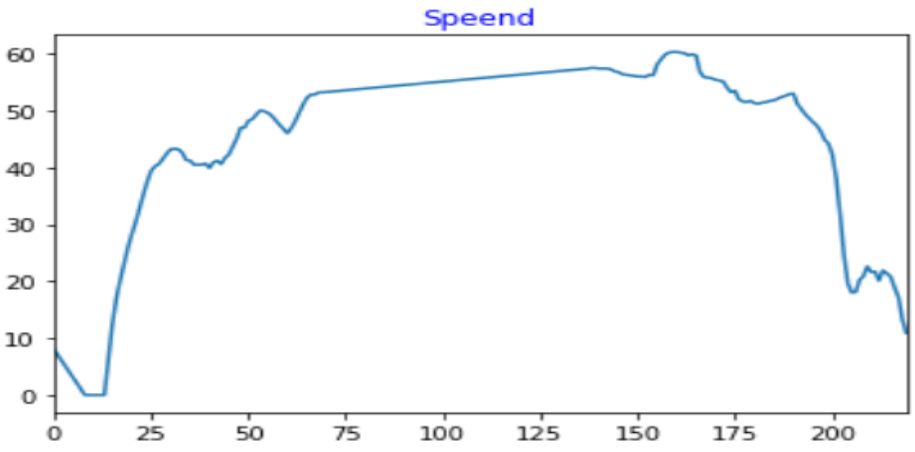


图 5. 文件 1 中切割后运动学片段样例



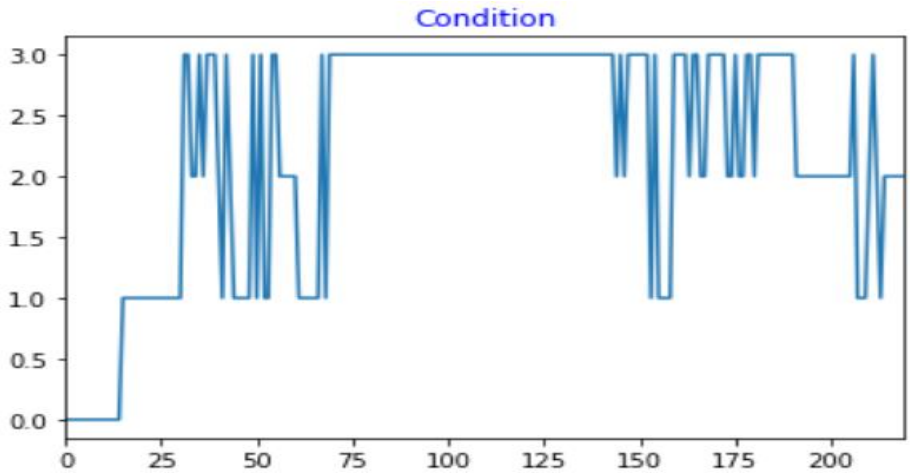


图 6. 文件 2 中切割后运动学片段样例

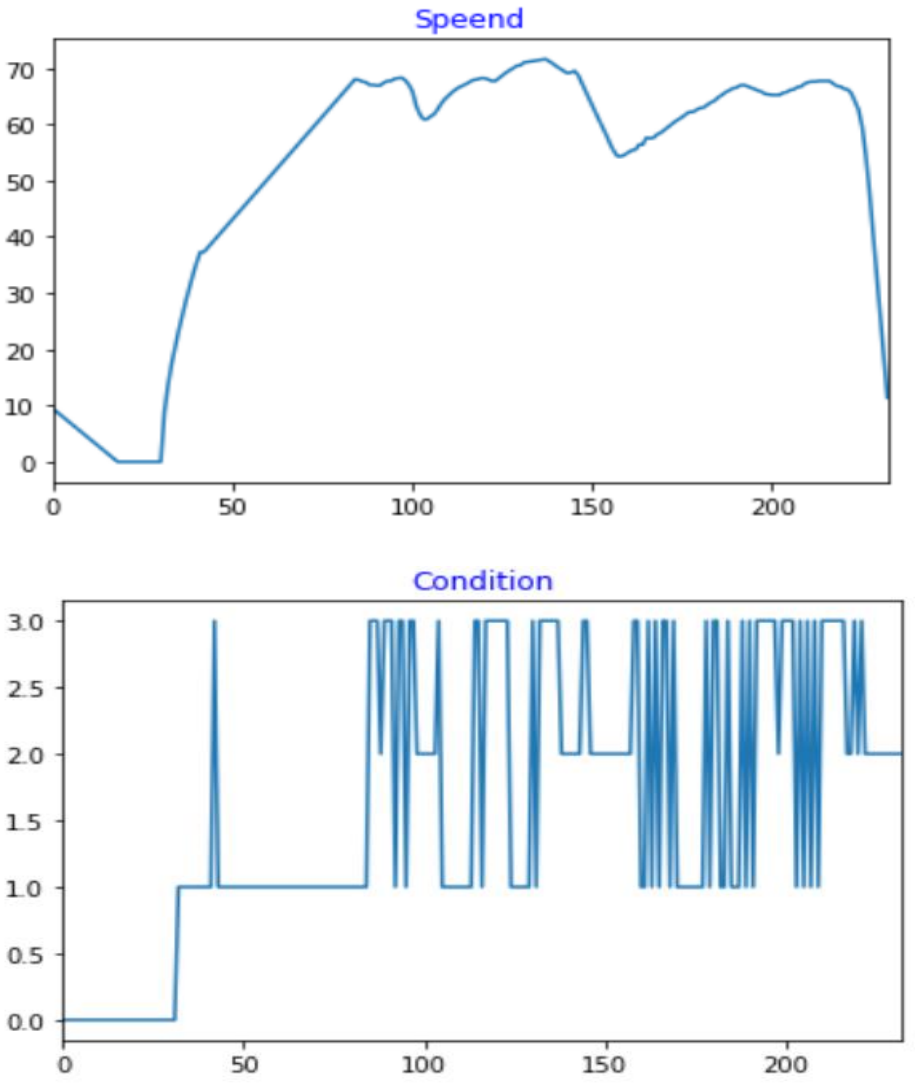


图 7. 文件 1 中切割后运动学片段样例

2.3 问题三：汽车行驶工况的构建

在第二问题中，运动学片段根据汽车行驶的工况得到了划分，第三问需要利用运动学片段实现汽车行驶工况的构建，但是仅凭第二问运动片段中的车速，加速度和工况并不能完整的去描述汽车的整个运动过程，需要构建新的运动特征，新的特征需要能够一定程度上全局性的反映一个运动片段的一些基本特征，如：加速工况比例，减速工况比例，最大加速度，最小加速度、平均速度等，这里采用 11 个统计学参数，来统计一个运动片段上的总体运动学片段特征，是个参数分别是P_a、P_d、P_c、P_i、v_m、v_{mr}、a_m、a_{max}、a_{min}、v_{sd}、a_{sd}[15]：

表 5 运动学片段运动学特征表

序号	特征参数	意义	单位
1	P_a	加速比例	%
2	P_d	减速比例	%
3	P_c	匀速比例	%
4	P_i	怠速比例	%
5	v_m	平均速度	km/h
6	v_{mr}	平均运行速度	km/h
7	a_m	平均加速度	m/s^2
8	a_{max}	最大加速度	m/s^2
9	a_{min}	最小加速度	m/s^2
10	v_{sd}	速度标准偏差	km/h
11	a_{sd}[	加速度标准偏差	m/s^2

每个运动学特征的计算方法[15]：

1） 加速比例 ： 同一运动学片段中全部加速工况记录数/运动学片段总记录数;

2） 减速比例 ： 同一运动学片段中全部减速工况记录数/运动学片段总记录数;

3） 匀速比例 ： 同一运动学片段中全部匀速工况记录数/运动学片段总记录数;

4） 怠速比例 ： 同一运动学片段中全部怠速工况记录数/运动学片段总记录数;

5） 平均速度 ： 同一运动学片段中全部速度和/运动学片段总记录数;

6） 平均运行速度 = 同一运动学片段中汽车运行距离/全部非怠速记录的时间之和

7） 平均最大加速度 = 同一运动学片段中全部记录的加速度之和/总记录时间

8） 最大加速度 = 同一运动学片段中最大的加速度

9） 最大加速度 = 同一运动学片段中最小的加速度

10）速度标准偏差 $\mathrm{vsd} = \sqrt{\frac{1}{k-1}\sum_{i=1}^{k}(v_i - V_m)^2}$

11）加速度标准偏差 $\mathrm{asd} = \sqrt{\frac{1}{k-1}\sum_{i=1}^{k}a_i{}^2}$

其中：v_i为第 i 秒的瞬时速度，单位为 km/h；a_i为第 i 秒的加速度，单位为 m/s^2。第二问的结果解析中提到过，在运动学片段切割之后，有很多散碎的运动学片段，对构建汽车行驶工况不仅没有帮助，很可能还会加入噪音，在剔除了部分不复合要求的运动学片段，并尽量选取持续时间长的运动片段之后，文件 1 中的运动学片段为 161，文件 2 中的运动学片段为 75，文件 3 中的运动学片段为 102，对三个文件计算上述十一个特征后，可以得到相应的特征表，如图 5 所示。

index	Pi	Pa	Pd	Pc	Vm	Vmr	amax	amin	am	vsd	asd
1	0.34632	0.121212	0.134199	0.398268	15.71934	24.04747	0.5	-1.72222	0.001684	10.90332	0.240181
2	0.14433	0.2	0.160825	0.494845	32.93287	38.48781	1.333333	-2.47222	0.001718	15.96767	0.333588
3	0.005464	0.418033	0.311475	0.265027	32.49071	32.66923	1.277778	-1.88889	0.003871	8.627765	0.441181
4	0.016409	0.251931	0.208494	0.523166	41.90429	42.60338	1.833333	-2.66667	0.000536	13.72447	0.282546
5	0.011628	0.229651	0.424419	0.334302	43.58111	44.09383	0.916667	-0.94444	0.002915	14.76052	0.207976
6	0.191011	0.277154	0.228464	0.303371	19.37528	23.95	1.277778	-2.05556	0.009988	8.61268	0.337605
7	0.089231	0.289231	0.224615	0.396923	33.47662	36.75642	1.194444	-1.97222	0.006838	13.37371	0.33009
8	0.028716	0.224662	0.204392	0.54223	27.57137	28.38652	0.777778	-1.33333	0.001455	7.198904	0.23705
9	0.020665	0.193172	0.178796	0.607367	33.55634	34.26441	2.166667	-1.72222	0.001922	7.066365	0.212215
10	0.183236	0.19883	0.167641	0.450292	34.98638	42.83535	1.694444	-3.22222	0.001083	19.07022	0.370918
11	0.131498	0.431193	0.159021	0.278287	44.07001	50.74259	1.916667	-4.88889	0.003823	22.58944	0.589709
12	0.079602	0.368159	0.273632	0.278607	55.81277	60.63982	2.916667	-2.47222	0.00152	22.04806	0.436211
13	0.26455	0.224868	0.148148	0.362434	28.44381	38.6754	2.333333	-2.16667	0.009406	19.80065	0.372332
14	0.839024	0.092683	0.043902	0.02439	4.517865	28.06552	1.694444	-2.19444	0.00813	8.96038	0.299297
15	0.142857	0.147619	0.133333	0.57619	15.7094	18.32763	1.333333	-3	0.000246	9.005241	0.413791

图 8　运动片段特征参数值样表

目前，车辆行驶工况构建的方法有三种。第一种利用非监督算法对运动学片段进行分析。第二种通过截取采样的方法，分别对运动学片段中截取 1-2 个与该运动学片段最相关的子片段，然后拼接成一个总的片段。第三种，马尔可夫理论，利用状态转移概率矩阵选取片段构建矩阵。本方法主要是使用 PCA 对原数据进行降维，并利用 k-means 聚类，来拟合工况。

2.3.理论依据：

主成分分析（PCA）是机器学习领域很常用的降维算法，它计算速度快捷，算法简单，成为特征选择和降维的热门算法。样本特征之间两两之间构成样本的协方差矩阵，通过求解协方差矩阵的特征值和特征向量，对特征向量由大到小排序就得到了原样本特征的重要性排序。

k-means 算法作为非监督学习的一种重要算法，在各个领域都有广泛应用。K-means 算法首先设置 K 个中心，通过计算样本和中心的距离（欧式距离，曼哈顿距离或其他距离），将与中心距离低于阈值的点归到该类，同时根据新加入的点的坐标，修改中心的坐标，最终完成算法聚类。

汽车行驶的运动学片段，是一个无标记的样本集，适合使用 PCA 和 k-means 这类的无监督学习，方案有相应的可行性。

2.4 汽车行驶工况的构建

前文构建了文件 1，2，3 的运动学特征，现在分别对三个文件的运动学特征进行 PCA 特征选择。

2.4.1 样本数据分析

就文件 1 中的运动学特征数据而言，图 6 中的特征值大小反映了成分序号对主成分方差的贡献，可以看到前五个特征值的累计百分比达到了 86.732%且它们的特征值均大于 1，这五个特征可以全面反映信息

成分	初始特征值		
	特征值	方差百分比	累积 %
1	3.464	31.494	31.494
2	2.006	18.241	49.735
3	1.658	15.073	64.808
4	1.383	12.572	77.380
5	1.029	9.352	86.732
6	0.661	6.011	92.743
7	0.339	3.084	95.827
8	0.302	2.742	98.569
9	0.136	1.239	99.808
10	0.021	0.192	100.000
11	3.010E-16	2.736E-15	100.000

图 9.1　各主成分特征值和贡献度

图 7 为主成分的碎石图，纵轴是成分对应的特征值，横轴为相应组件的序号

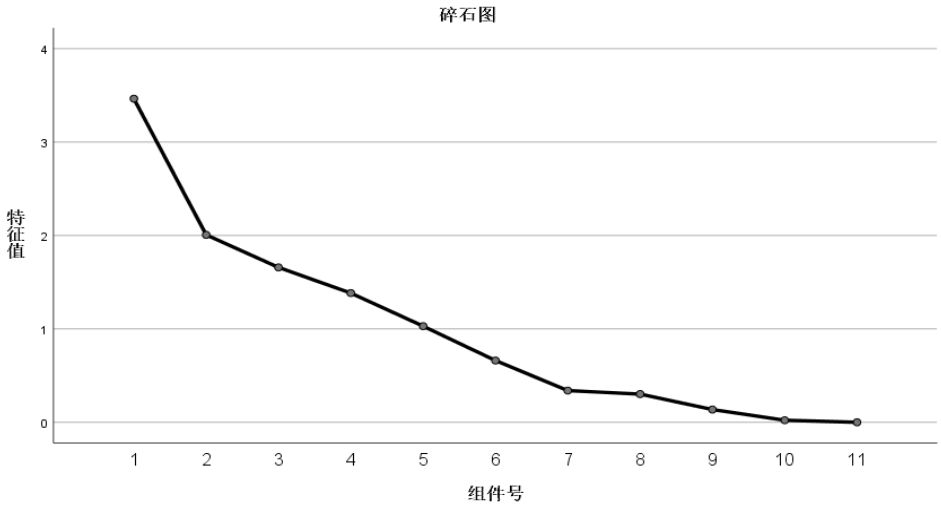


图 9.2 主成分碎石图

通过 PCA 分析，可以选定将文件 1 中的运动学特征降维到五维效果最佳，在提取 5 个成分之后，通过图 9 成分矩阵，可以看到原数据各个特征对成分的贡献。

成分矩阵[a]					
	成分				
	1	2	3	4	5
Pi	-0.494	-0.446	0.727	0.055	-0.037
Pa	0.698	-0.464	-0.443	0.038	0.058
Pd	0.586	-0.331	-0.466	0.218	-0.399
Pc	-0.214	0.893	-0.235	-0.183	0.193
Vm	0.791	0.533	-0.019	0.232	0.055
Vmr	0.697	0.390	0.387	0.326	0.034
amax	0.505	0.138	0.354	-0.206	-0.356
amin	-0.503	0.020	-0.211	0.635	-0.389
am	-0.181	-0.231	-0.219	0.500	0.670
vsd	0.515	-0.074	0.539	0.471	0.142
asd	0.654	-0.412	0.051	-0.468	0.277

图 9.3 成分矩阵

同理，文件 2 和文件 3 的运动学特征参数的各主成分特征值和贡献度，主成分碎石图以及成分矩阵也可以得到，见图 9 和图 10。

	初始特征值			提取载荷平方和		
成分	总计	方差百分比	累积 %	总计	方差百分比	累积 %
1	2.876	26.148	26.148	2.876	26.148	26.148
2	2.349	21.352	47.500	2.349	21.352	47.500
3	2.013	18.303	65.803	2.013	18.303	65.803
4	1.549	14.081	79.884	1.549	14.081	79.884
5	0.800	7.274	87.159			
6	0.635	5.772	92.931			
7	0.349	3.173	96.104			
8	0.257	2.336	98.440			
9	0.169	1.536	99.976			
10	0.003	0.024	100.000			
11	4.302E-16	3.911E-15	100.000			

图 10.1 文件 2 各主成分贡献度

	1	2	3	4
Pi	-0.736	0.333	0.487	0.228
Pa	0.616	0.536	-0.401	0.043
Pd	0.436	0.541	-0.561	0.257
Pc	0.149	-0.883	0.030	-0.366
Vm	0.833	-0.467	0.031	0.272
Vmr	0.682	-0.418	0.305	0.465
amax	0.425	0.306	0.478	-0.099
amin	-0.315	-0.218	-0.593	0.406
am	0.203	-0.068	0.042	-0.682
vsd	0.077	0.099	0.726	0.475
asd	0.507	0.574	0.313	-0.372

图 10.2 文件 2 各主成分对特征参数的影响

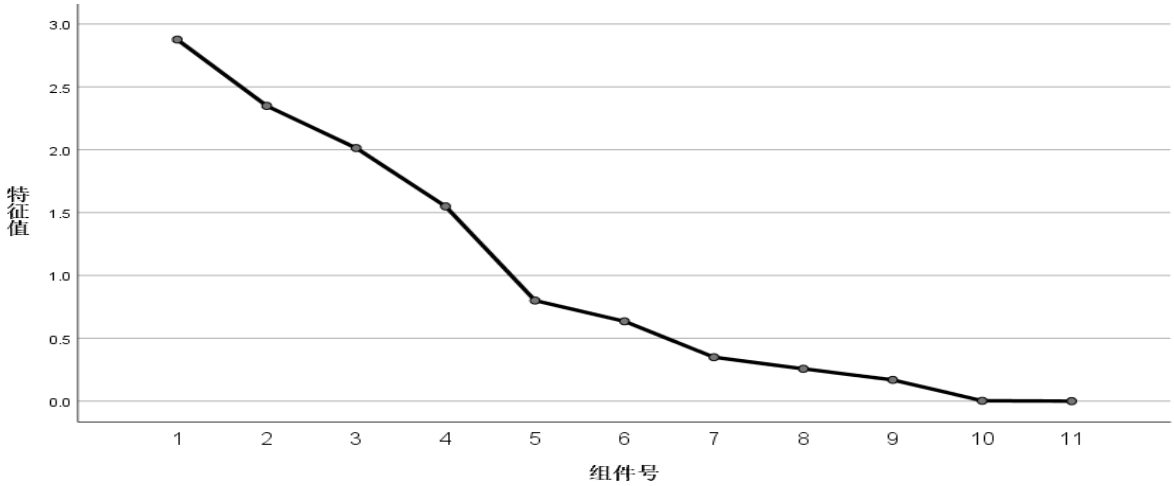


图 10.3 文件 2 主成分碎石图

图 10 文件 2 运动学特征数据分析结果

	初始特征值			提取载荷平方和		
成分	总计	方差百分比	累积 %	总计	方差百分比	累积 %
1	3.442	31.295	31.295	3.442	31.295	31.295
2	2.382	21.657	52.953	2.382	21.657	52.953
3	1.453	13.205	66.158	1.453	13.205	66.158
4	1.352	12.288	78.445	1.352	12.288	78.445
5	1.024	9.310	87.755	1.024	9.310	87.755
6	0.540	4.907	92.662			
7	0.393	3.573	96.235			
8	0.280	2.544	98.779			
9	0.132	1.204	99.982			
10	0.002	0.018	100.000			
11	9.021E-17	8.201E-16	100.000			

图 11.1 文件 3 主成分贡献度

	1	2	3	4	5
Pi	-0.748	-0.135	0.521	0.313	0.068
Pa	0.697	-0.576	-0.211	0.130	0.039
Pd	0.635	-0.510	-0.215	0.099	-0.299
Pc	-0.086	0.813	-0.255	-0.456	0.065
Vm	0.769	0.582	-0.148	0.141	-0.022
Vmr	0.670	0.638	0.024	0.318	0.025
amax	0.286	0.118	0.715	-0.349	0.099
amin	-0.574	-0.069	-0.457	0.440	0.113
am	0.077	-0.221	-0.277	-0.204	0.886
vsd	0.381	0.204	0.368	0.681	0.325
asd	0.619	-0.508	0.265	-0.278	0.102

图 11.2 文件 3 主成分对特征参数的影响程度

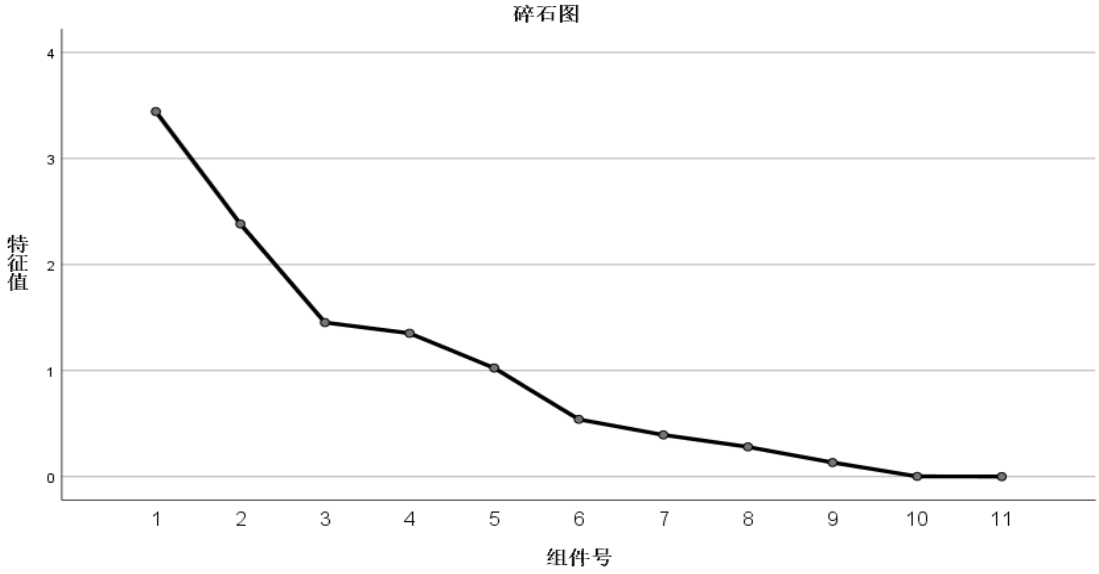


图 11.3 文件 3 主成分碎石图

2.4.2 样本模型拟合

通过刚才的主成分分析方法，可以看到不同文件的运动学特征的重要性各有不同，为了统一计算和便于展示，在模型拟合环节，统一都选择三个主成分来作为对数据整体的一个表示，利用 k-means，将各个文件的运动学结果聚类划分为三类，划分前后的对比结果见图 12，由上到下分别是文件 1，2，3 的聚类前后对比。

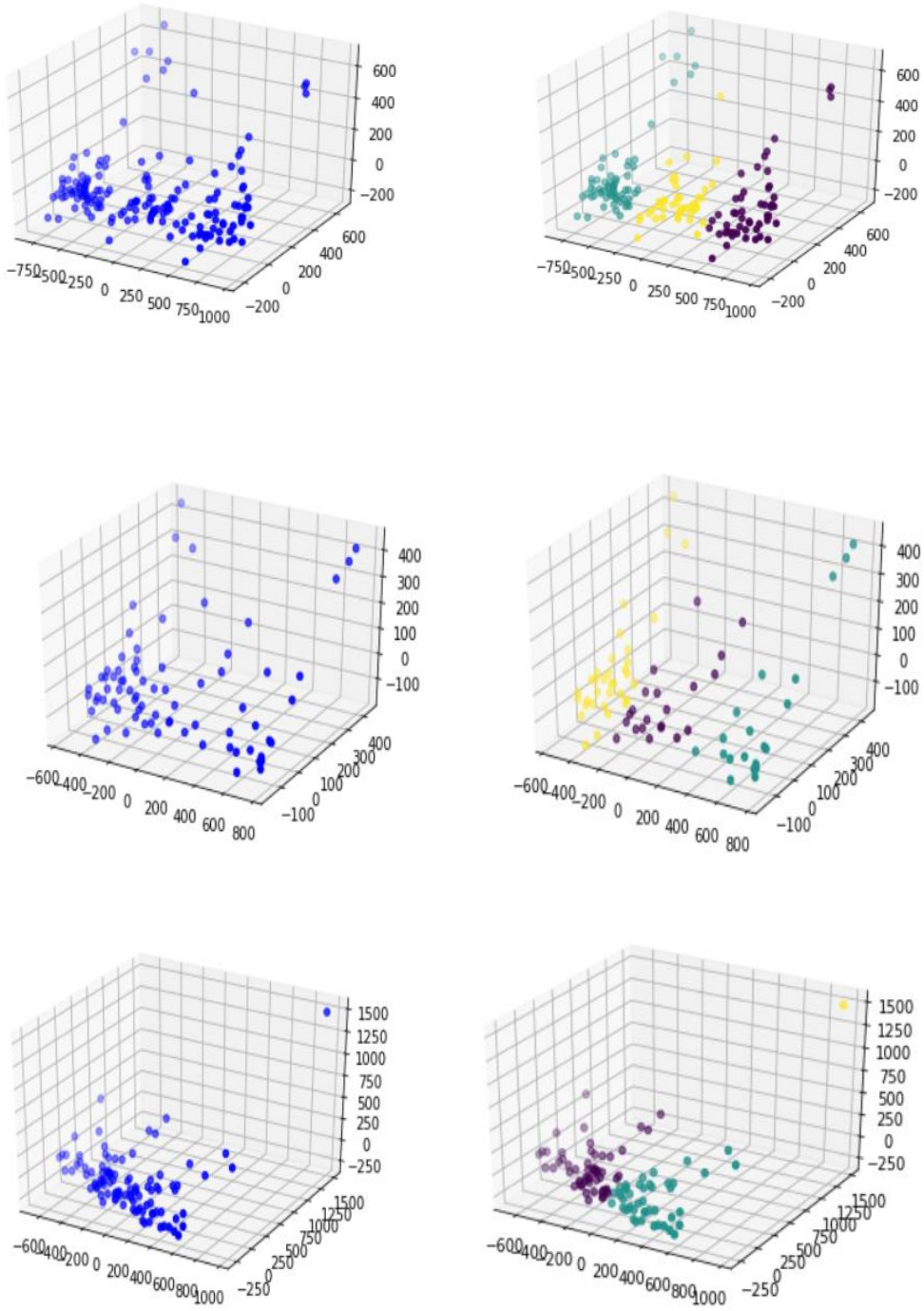


图 12 聚类前后结果对比

2.4.3 工况拟合曲线

之前的工作对三个文件中的各个样本进行了聚类分析，将样本归为三类，在拟合曲线方面，需要分别从三个类中各取 1-2 个片段作为拼接工况曲线的子曲线。取片段的具体方法如下，以样本 1 为例：

1）计算样本 1 的总体平均运动学参数，平均最大加速度a_{avgmax}，平均加速比例P_{avga}，平均减速比例P_{avgd} 等平均参数，设他们为X_{avgi} i = 1,2,3….11. 样本 1 中的三类为 A,B,C.

2) 从 A 中每次抽取一个样本，根据公式（1）计算均方差，选取均方差最小的样本作为选择片段，在 B 类和 C 类中重复操作，每类选出 1-2 个片段。

$$\text{Rmse} = \sqrt{\sum_{i=1}^{11}(x_i - X_{avgi})^2}$$

3）将片段拼接作为工况拟合曲线

根据上面方法可以分别得到文件 1，2，3 的三条拟合曲线

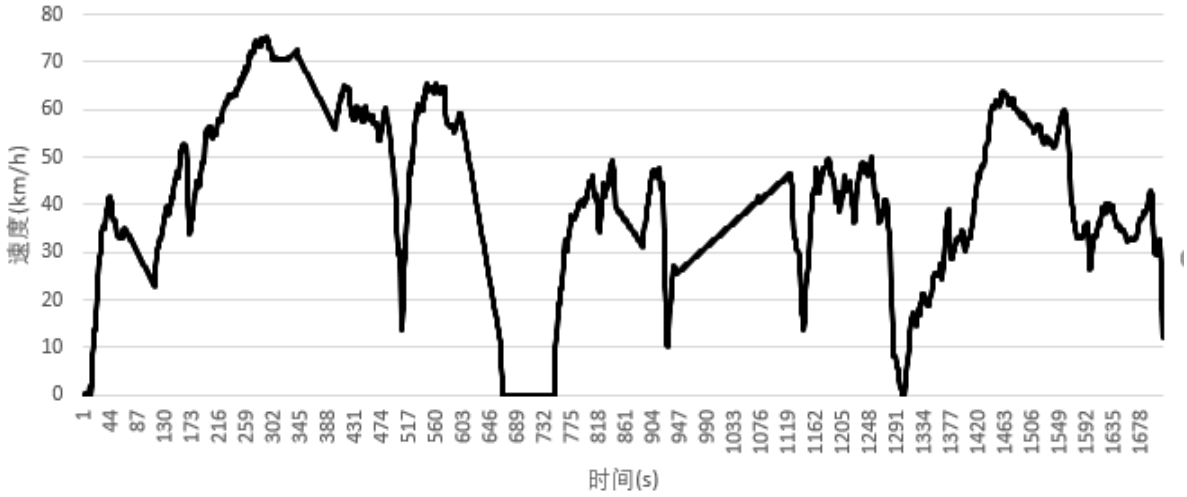


图 13　文件 1 工况拟合曲线

图 14 文件 2 工况拟合曲线

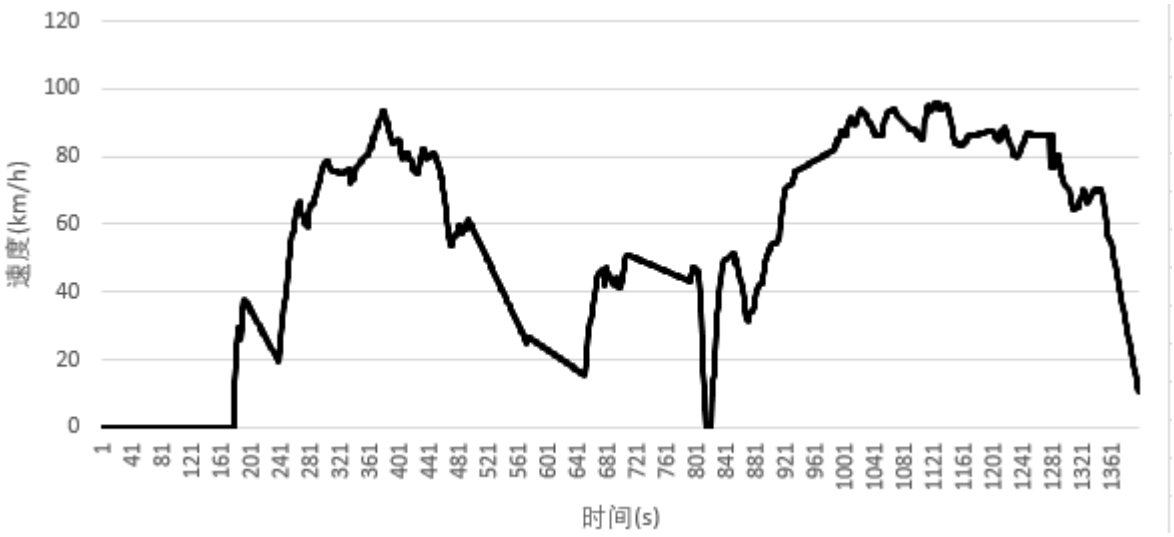


图 15 文件 3 工况拟合曲线

分别计算三个文件的工况拟合曲线的运动学特征，和文件原数据做性能分析，可以得到如下图的结果。

	拟合工况	实际数据	误差
Pi	0.073555	0.107789	0.317603
Pa	0.27087	0.270641	0.000846
Pd	0.224752	0.212554	0.057387
Pc	0.430823	0.409015	0.053317
Vm	40.82689	42.56654	0.040869
Vmr	44.06834	47.70907	0.076311
amax	2.833333	3.96	0.284512
amin	-2.77778	-8	0.652778
am	0.001762	0.002588	0.319319
vsd	17.99755	23.89512	0.246811
asd	0.32578	0.361697	0.099302

	拟合工况	实际工况	误差
Pi	0.06008	0.099294	0.394928
Pa	0.338677	0.298508	0.134567
Pd	0.281563	0.251631	0.118954
Pc	0.339679	0.350567	0.031058
Vm	39.2878	42.13856	0.067652
Vmr	40.9282	46.78393	0.125165
amax	2.277778	3.916667	0.41844
amin	-4.41667	-5.69444	0.22439
am	0.001225	0.002576	0.524498
vsd	13.64897	20.84544	0.34523
asd	0.413975	0.408303	0.013892

图 16 文件 1,2 汽车工况拟合与实际数据对比(左 1，右 2)

	拟合工况	实际工况	误差
Pi	0.135387	0.117407	0.153141
Pa	0.243438	0.278777	0.126763
Pd	0.26361	0.226692	0.162856
Pc	0.397564	0.397124	0.00111
Vm	53.27609	50.60833	0.052714
Vmr	61.61842	56.06993	0.098956
amax	3.411111	3.96	0.138608
amin	-6.61111	-7.92	0.165264
am	0.001751	0.002142	0.182673
vsd	30.39285	27.3301	0.112065
asd	0.309774	0.383257	0.191734

图 17 文件 3 汽车工况拟合与实际数据对比

3 方法的优缺点分析

优点：方法主要使用的场合是问题三中使用了 PCA 和 K-means，一个是机器学习常用的降维算法，另一个是最容易理解的聚类算法，针对数据集本身的特性—不具有标签，这两个无监督的方法起到了出人意料的作用，模型简单易于理解和实现，同时受数据变化影响较小，可以得到较为稳定的结果。

缺点：数据中有很多的异常值出现，通过聚类对比图也可以看出，而 k—means 不仅对异常值敏感，还需要手动选择 k 值。在测试过程中，手动设置参数势必会牺牲部分精度，而穷举所有参数又会耗费大量时间。

改进方案：因为时间问题，本文只使用了一个 k-means 算法，这一无监督学习的效果收到一定先验信息的影响，比如 K 值的设定。在后续可以多尝试几个算法，甚至融合多个算法的结果以提升聚类精准度，最终结果的精度有可能得到进一步提升。

4 参考文献

[1] 余志生. 汽车理论[M]. 机械工业出版社, 2000.

[2] 姜平. 城市混合道路行驶工况的构建研究[D]. 合肥工业大学博士学位论文, 2011.

[3] 清华大学“城市可持续交通”课题组. 中国城市可持续交通: 问题、挑战与实现途径[M]. 中国铁道出版社, 2007.

[4] Kent J H, Allen G H, Rule G. A driving cycle for Sydney[J]. Transportation Research, 1978, 12(3):147–152.

[5] Lin J, Niemeier D A. Exploratory analysis comparing a stochastic driving cycle to California's regulatory cycle[J]. Atmospheric Environment, 2002, 36(38):5759-5770.

[6] Ball K, Owsley C, Stalvey B, et al. Driving avoidance and functional impairment in older drivers[J]. Accid Anal Prev, 1998, 30(3):313-322.

[7] Ericsson E. Independent driving pattern factors and their influence on fuel-use and exhaust emission factors[J]. Transportation Research Part D, 2001, 6(5):325-345.

[8] 杨众凯. 基于行驶工况的严寒地区公共汽车通用动力系统配置研究[D]. 吉林大学博士学位论文，2015.

[9] 步曦. 上海市市区乘用车行驶工况的研究[D]. 同济大学硕士学位论文,2004

[10] 赵慧, 张镇顺. 香港城区汽车行驶工况的研究[J]. 环境科学学报, 2000, 20(3):312-315.

[11] 杜爱民, 步曦, 陈礼璠, et al. 上海市公交车行驶工况的调查和研究[J]. 同济大学学报(自然科学版), 2006, 34(7):943-946.

[12] 訾琨, 黄永青, 涂先库等. 宁波市区汽车行驶工况和污染物排放调查研究[J]. 内燃机工程, 2006(1):81-84.

[13] 马冬,丁焰, 刘志华. 轻型汽车实际行驶工况的排放研究[J]. 安全与环境学报, 2008(5): 66-68.

[14] 王岐东, 贺克斌, 姚志良等. 中国城市机动车行驶工况研究[J]. 环境污染与防治, 2007(10): 745-748.

[15] 王楠楠.城市道路行驶工况构建及油耗研究[D].合肥工业大学硕士学位论文，2012.

附录 I 数据预处理代码

```
import time
from datetime import datetime
import warnings
warnings.filterwarnings('ignore')

data = pd.read_csv('new3.csv',encoding='gbk')
del data['之前时刻']

data['差值'] = data['差值'].apply(lambda x: int(x[7:9])*3600+int(x[10:12])*60+int(x[13:15]))
out = pd.DataFrame([],columns=['车速','加速度','发动机'])
ver = 1

for i in range(len(data['差值'])):
    item = data['差值'][i]
    if item == 1:
        out = out.append([{'车速':data['GPS 车速'][i],'加速度':0,'发动机':1}],ignore_index=True)
    elif item > 1:
        print(i,item)
        delta_v = (data['GPS 车速'][i]-data['GPS 车速'][i-1])/(3.6*item)
        print('  ',data['GPS 车速'][i],data['GPS 车速'][i-1],delta_v)
        if item < 180:
            j = 0
            speed = data['GPS 车速'][i-1]/3.6
            while abs(data['GPS 车速'][i]-speed*3.6) >= abs(delta_v*3.6) and j < item:
                print('    ',abs(data['GPS 车速'][i]-speed*3.6),abs(delta_v*3.6),item-j)
                out = out.append([{'车速':speed*3.6,'加速度':0,'发动机':1}],ignore_index=True)
                speed += delta_v
                j += 1
            if j > 1:
                out = out.append([{'车速':data['GPS 车速'][i],'加速度':0,'发动机':1}],ignore_index=True)

        else:
            print('  last v',data['GPS 车速'][i-1])
            speed = data['GPS 车速'][i-1]/3.6
            while speed > 0:
                speed += -8
                out = out.append([{'车速':speed*3.6,'加速度':0,'发动机':1}],ignore_index=True)
```

```
                out = out.append([{'车速':0,'加速度':0,'发动机':1}],ignore_index=True)
                out.to_csv('./new3_process/new3_'+str(ver)+'.csv',encoding='gbk')
                out = pd.DataFrame([],columns=['车速','加速度','发动机'])
                ver += 1
                speed = data['GPS 车速'][i]/3.6
                mod = speed % 3.96
                if mod > 0:
                    out = out.append([{' 车 速 ':speed*3.6,' 加 速 度 ':0,' 发 动 机 ':1}],ignore_index=True)
                speed = mod
                while speed*3.6 < data['GPS 车速'][i]:
                    speed += 3.96
                    out = out.append([{' 车 速 ':speed*3.6,' 加 速 度 ':0,' 发 动 机 ':1}],ignore_index=True)
                out = out.append([{' 车 速 ':data['GPS 车 速 '][i],' 加 速 度 ':0,' 发 动 机 ':1}],ignore_index=True)

out.to_csv('./new3_process/new3_'+str(ver)+'.csv',encoding='gbk')

for j in range(1,80):
    df = pd.read_csv(r'D:\test\new3_process\new3'+'_'+str(j)+'.csv',encoding = 'gbk')
    row = df.shape[0]
    df['工况'] = 0
    df['异常'] = 0
    for i in range(1,row):
        df.iloc[i,2] =(float(df.iloc[i,1])- (df.iloc[i-1,1]))/3.6
        if float(df.iloc[i,2]) > 3.96 or float(df.iloc[i,2])< -8:
            df.iloc[i,5] = 1
        if float(df.iloc[i,1]) >= 10:
            if float(df.iloc[i,2]) > 0.1:
                df.iloc[i,4] = 1
            elif float(df.iloc[i,2]) < -0.1:
                df.iloc[i,4] = 2
            else:
                df.iloc[i,4] = 3
    df.rename(columns={'Unnamed: 0':'时间'}, inplace = True)
    df.to_csv(r'D:\test\new3_process\new3'+'_'+str(j)+'.csv',index = 0)
    print('已完成:',j/(79)*100)
```

附录 III 运动学片段划分代码

```
import pandas as pd
import datetime
```

```
import time
import warnings

warnings.filterwarnings('ignore')

f = [i for i in range(1,160)]
ver = 1 #  用于标识是第几个文件
for n in f:
    print(n)
    data = pd.read_csv('./new1_process/new1_'+str(n)+'.csv',encoding='gbk')
    print(len(data['工况']))
    del data['发动机']
    cut = pd.DataFrame([],columns=['时间','车速','加速度','工况','异常'])
    flag = True #  用于标识怠速工况是否是下一片段的开始

    i = 0
    d_count = 0 #  怠速持续时间
    while i < len(data['工况']):
        if data['异常'][i] == 1:
            del cut
            while i < len(data['工况']) and (data['工况'][i] != 0 or data['异常'][i] == 1):
                i += 1
            cut = pd.DataFrame([],columns=['时间','车速','加速度','工况','异常'])
            flag = True
        else:
            if data['工况'][i] == 0:
                d_count += 1
                if d_count < 180:
                    if i == 0 or data['工况'][i-1] == 0 or flag == True or cut.shape[0] == 0:
                        # print('i == 0')
                        cut = cut.append([{'时间':data['时间'][i],
                                           '车速':data['车速'][i],
                                           '加速度':data['加速度'][i],
                                           '工况':data['工况'][i],
                                           '异常':data['异常'][i]}],ignore_index=True)
                        flag = False
                    else:
                        if 1 not in cut['异常'].values:
                            print('save it at:',i)

cut.to_csv('./new1_slice/new1_'+str(ver)+'.csv',encoding='gbk')
                            ver += 1
                        # else:
```

```
                    # print('not proper at:',i)
                cut = pd.DataFrame([],columns=['时间','车速','加速度','工况','异
常'])
                flag = True
        else:
            if cut.shape[0] > 0: # 说明切片中有数据，即前面有怠速工况
                # print('other gongkuang:',i)
                cut = cut.append([{'时间':data['时间'][i],'车速':data['车速'][i],
                        '加速度':data['加速度'][i],'工况':data['工况'][i],
                        '异常':data['异常'][i]}],ignore_index=True)
            d_count = 0 # 怠速计数归 0
        i += 1
```

附录 III 特征构建代码

```
import pandas as pd
import warnings
import os

warnings.filterwarnings('ignore')

dirname = '.\\problem2_1'
f_out = pd.DataFrame([],columns=['index','Pi','Pa','Pd','Pc',
                                 'Vm','Vmr','amax','amin','am',
                                 'vsd','asd','ni','na','nd','nc','num'])
i = 1
for maindir,subdir,file_name_list in os.walk(dirname):
    for  filename in file_name_list:
        apath = os.path.join(maindir,filename)
        print(apath)
        data = pd.read_csv(apath,encoding='gbk')

        ser = data['工况'].value_counts()
        if 0 in ser and 1 in ser and 2 in ser and 3 in ser:
            Pi = ser[0]/data.shape[0]
            Pa = ser[1]/data.shape[0]
            Pd = ser[2]/data.shape[0]
            Pc = ser[3]/data.shape[0]
            Vm = sum(data['车速'].values)/data.shape[0]
            Vmr = sum(data['车速'].values)/(ser[1]+ser[2]+ser[3])
            amax = max(data['加速度'].values)
            amin = min(data['加速度'].values)
            am = sum(data['加速度'].values)/data.shape[0]
```

```
            vsd = sum([(v-Vm)**2 for v in list(data['车速'].values)])/(data.shape[0]-1)
            vsd = vsd ** 0.5

            asd = sum([a**2 for a in list(data['加速度'].values)])/(data.shape[0]-1)
            asd = asd ** 0.5

            num = filename.split('.')[0].split('_')[1]

            f_out = f_out.append([{'index':i,'Pi':Pi,'Pa':Pa,'Pd':Pd,'Pc':Pc,'Vm':Vm,'Vmr':Vmr,

'amax':amax,'amin':amin,'am':am,'vsd':vsd,'asd':asd,'ni':ser[0],
                                        'na':ser[1],'nd':ser[2],'nc':ser[3],'num':num}],
                                    ignore_index=True)
            i += 1

f_out.to_csv('problem2_1_feature.csv',index=0)
```

附录Ⅳ 特征构建代码

```
import pandas as pd
import numpy as np
from sklearn.decomposition import PCA
pca = PCA(n_components = 3)
dt = pd.read_csv(r'D:\test\new1_split\problem3_3_feature.csv',encoding = 'gbk')
newX = pca.fit_transform(dt.drop('index',axis = 1 ).values)
import matplotlib.pyplot as plt
from mpl_toolkits.mplot3d import Axes3D
fig = plt.figure()
ax = fig.add_subplot(111, projection='3d')
ax.scatter(newX[:,0],newX[:,1],newX[:,2],c='r')
plt.show()
#plt.scatter(newX[:,0],newX[:,1],marker='o')
#plt.show()
def kmean(newX):
    from sklearn.cluster import KMeans
    y_pred = KMeans(n_clusters=3, random_state=9).fit_predict(newX)
    fig = plt.figure()
    ax = fig.add_subplot(111, projection='3d')
    #ax.scatter(newX[:,0],newX[:,1],newX[:,2])
    #plt.show()
    ax.scatter(newX[:, 0], newX[:, 1],newX[:,1],c = y_pred)
    plt.show()
```

```
 #y_pred
return y_pred
a = pd.DataFrame(kmean(newX))
a = pd.DataFrame(a)
a.columns = ['label']
df = pd.concat( [dt,a], axis=1 )
df.to_csv(r'D:\test\new1_split\problem3_3_label.csv',index = 0)
```