**摘要**

最后再写（约2页）

[问题背景与问题重述 1](#_Toc1225478369)

[1.1 问题背景 1](#_Toc99289406)

[1.2 问题重述 2](#_Toc162252923)

[基本假设与符号说明 2](#_Toc1828128818)

[问题一的分析与建模 2](#_Toc1312506497)

[问题二的分析建模 2](#_Toc344673095)

[问题三的分析建模 3](#_Toc1157311706)

[3.1 特征参数的选取 3](#_Toc1178451863)

[3.2 运动学片段的参数计算 3](#_Toc2146268807)

[3.3 聚类分析 3](#_Toc1057020590)

[3.4 工况的合成 4](#_Toc1360328146)

[模型评估 4](#_Toc924243860)

[4.1 指标计算 4](#_Toc1017336269)

[4.2 合理性分析 4](#_Toc105875669)

[参考文献 4](#_Toc1335909167)

[附录 4](#_Toc683840384)

# 1 问题背景与问题重述

## 1.1 问题背景

汽车行驶工况又称为车辆测试循环，是描述典型车辆行驶的速度-时间曲线。它可以用于确定车辆污染物排放量及燃油消耗量，体现汽车在道路行驶中的燃料经济性、排放等运动学特征，是汽车工业一项核心技术。目前，我国乘用车行驶工况采用的是欧洲的NEDC，但由于我国的交通状况、行驶状况与道路状况等因素都与欧盟国家有很大不同，因此NEDC工况与我国城市汽车的实际行驶数据偏差较大。另一方面，我国地域辽阔，不同城市之间的交通状况、发展程度与气候条件等都存在着差异，这也导致各个城市的汽车行驶工况也不尽相同。因此，基于各个城市自身的汽车行驶数据制定反映我国实际道路行驶状况的汽车行驶工况，显得越来越重要。

## 1.2 问题重述

在上述背景下，题目给出了某城市同一辆测试汽车在不同时间段采集的实际行驶数据，要求利用该数据解决以下几个问题：

问题一：由于数据采集过程中存在着不可避免的实际问题，原始数据中会包含一些不良值。要求通过合理的方法对异常数据进行预处理，并给出处理后的数据记录数。

问题二：我们将车辆从一个怠速开始到下一个怠速开始之间的运动定义为运动学片段。题目要求设计合理的方法，将处理后的数据划分为多个运动学片段，并给出每个数据文件得到的运动学片段数。

问题三：基于上面处理好的数据，采用科学有效的方法构建一条能体现采集数据源行驶特征的汽车行驶工况曲线。同时使用合理的评估体系对行驶工况曲线进行评估，说明其合理性。

# 2 基本假设与符号说明

假设普通轿车在一般情况下最大的加速度为4m/s2，紧急刹车时的最大减速度为8m/s2。

假设当汽车车速小于10km/h时，就认为是怠速状态。当怠速状态时间超过180秒就被认为是异常情况，怠速时间最长可按180秒处理。

我们定义汽车加速度大于0.1m/s2的连续过程为加速阶段，汽车加速度小于-0.1m/s2的连续过程为减速阶段。

假设采样频率为1Hz

# 3 问题一的分析与建模

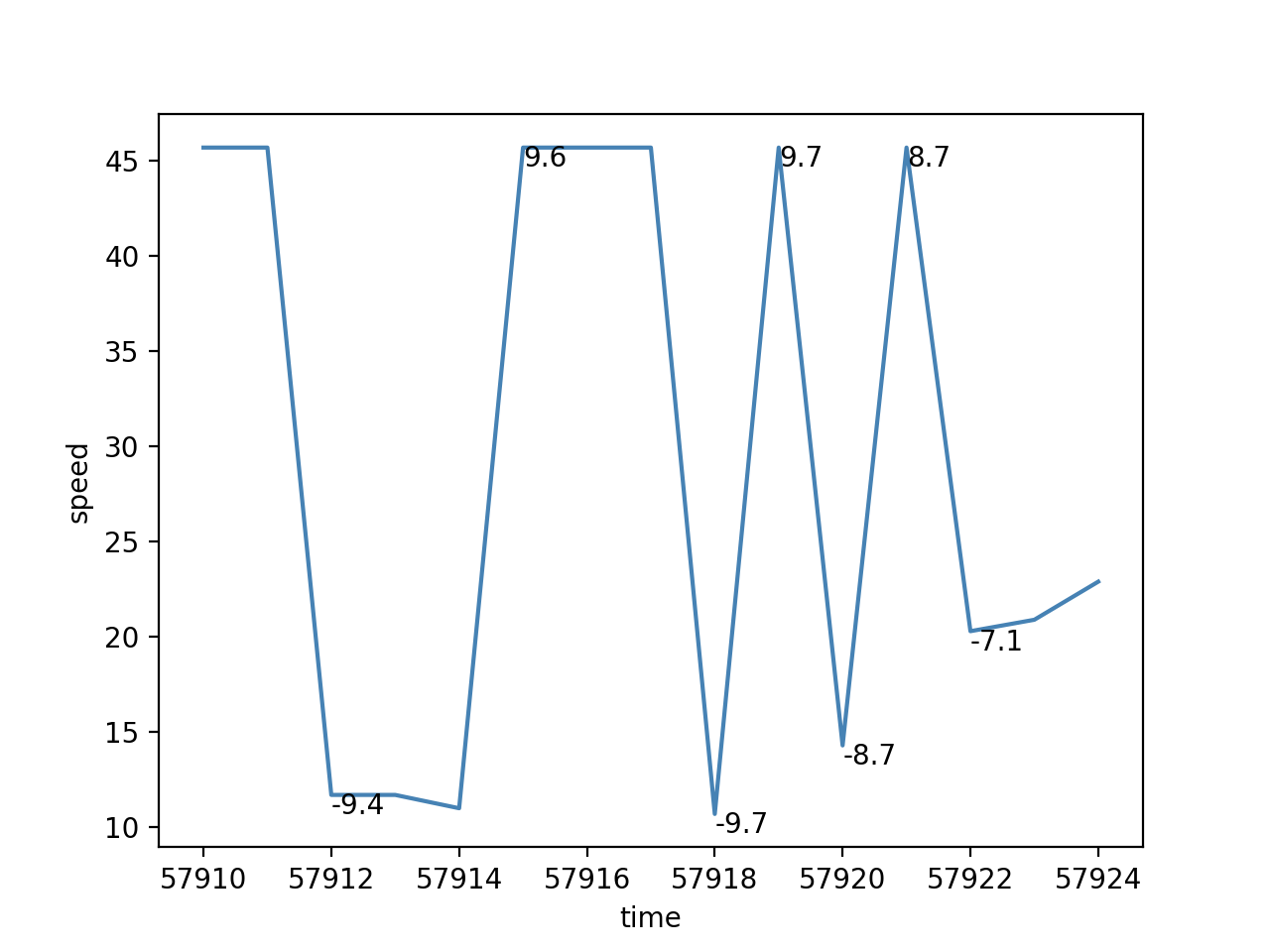
题目给出了某城市同一辆测试汽车在不同时间段采集的实际行驶数据，覆盖了工作日与节假日、高峰期与非高峰期等多时段的数据，保证了测试数据的可靠性。但由于多种主客观因素影响，原始测试数据中可能包含多种不良数据。问题一就是要求通过多种合理的处理方法对这些异常数据进行识别与预处理。

原始的测试数据包含3个数据文件，其中数据文件1的原始数据记录数为**185725**条，数据文件2的原始数据记录数为**145825**条，数据文件3的原始数据记录数为**164914**条。

## 3.1 加减速度异常的数据处理

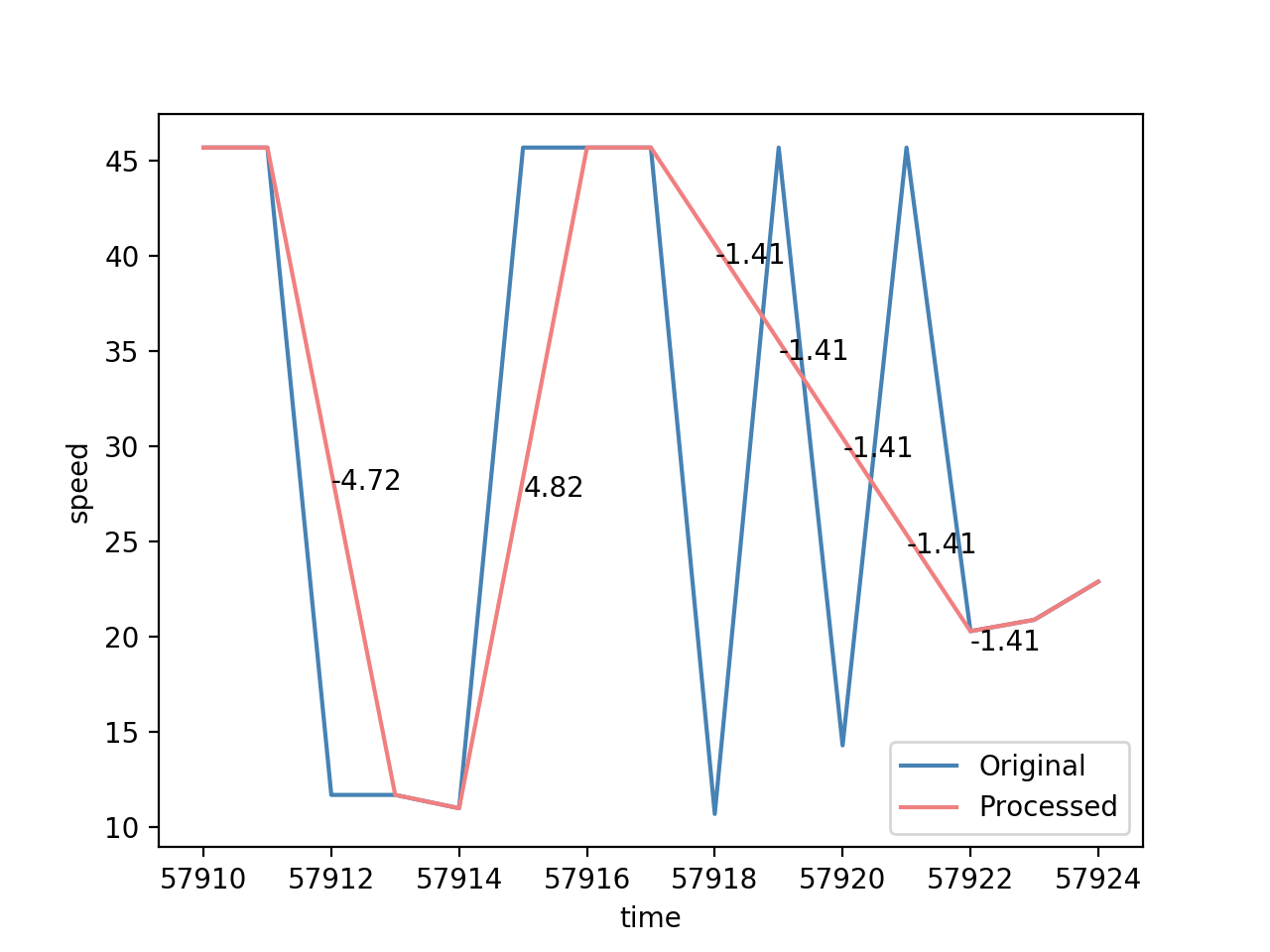
从上文的基本假设中我们可知，在一般情况下普通轿车从0加速至100km/h的时间应大于7秒，即最大加速度约为4m/s2，而轿车紧急刹车的最大减速度约为8m/s2。在实际道路行驶过程中可能会出现车辆急加速与急刹车的情况，这种情况下往往会导致车辆加速度过大。而超出最大加减速度阈值的不良数据显然会对我们构建行驶工况曲线造成影响，因此需要对这些不良数据进行处理。

我们通过对题目给出的原始采集数据的各条记录进行加速度计算后，发现数据集中也存在这种加减速度异常的数据值。数据文件1的异常数据的记录数为**28**条，数据文件2的异常数据的记录数为**401**条，数据文件3中**无**加减速度异常的记录。图x展示了数据文件2中的一段异常记录，从图中我们可以明显看出测量到的GPS速度变化异常，几乎每秒都在进行急加速或急减速，加减速度均超出了正常范围。使用这样的异常数据来构建行驶工况曲线无疑会出现较大的偏差。



图x 加减速度异常的速度曲线图

对于这类加减速度异常的原始数据，我们的处理方法是对异常数据记录进行剔除，并使用**线性插值法**对原本的异常数据点进行修正，从而消除离群点数据对于构建行驶工况曲线带来的不利影响。图x展示了修正后的数据记录，可以看出通过线性插值法修正后加减速度均已在正常范围内，数据异常问题得以解决。

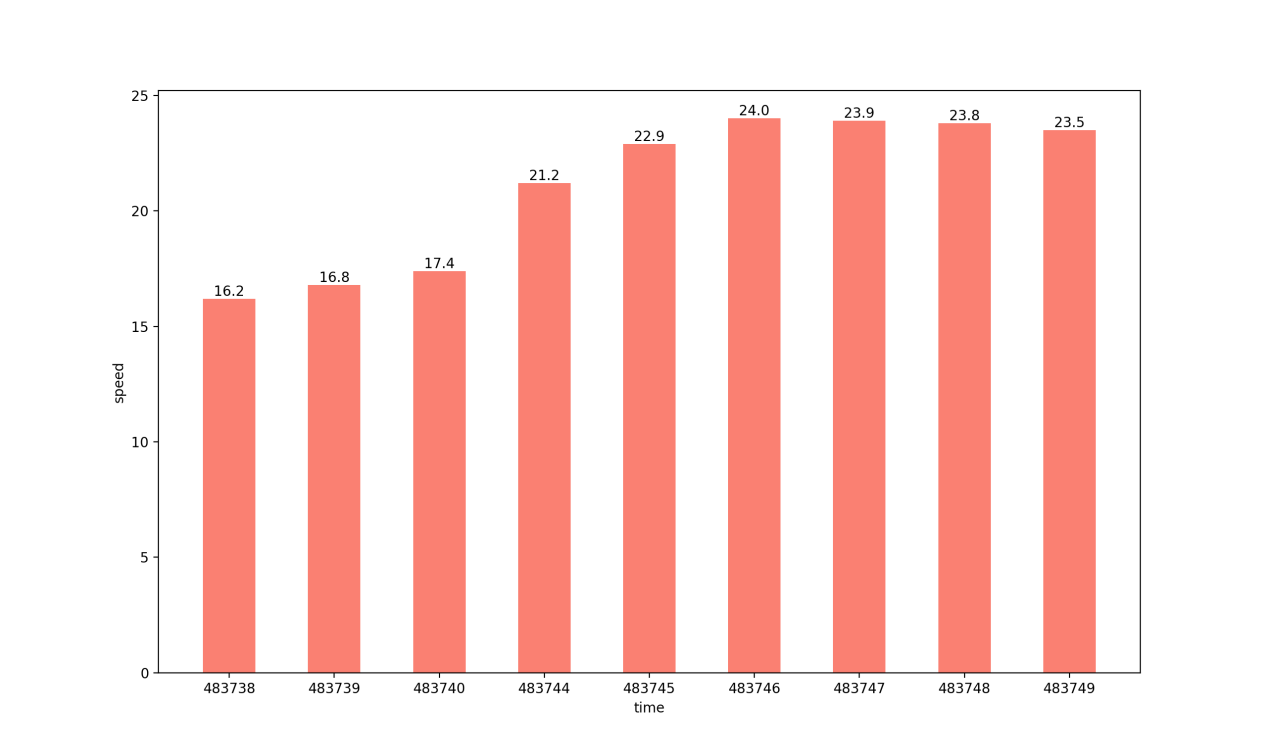


图x 修正后的速度曲线图

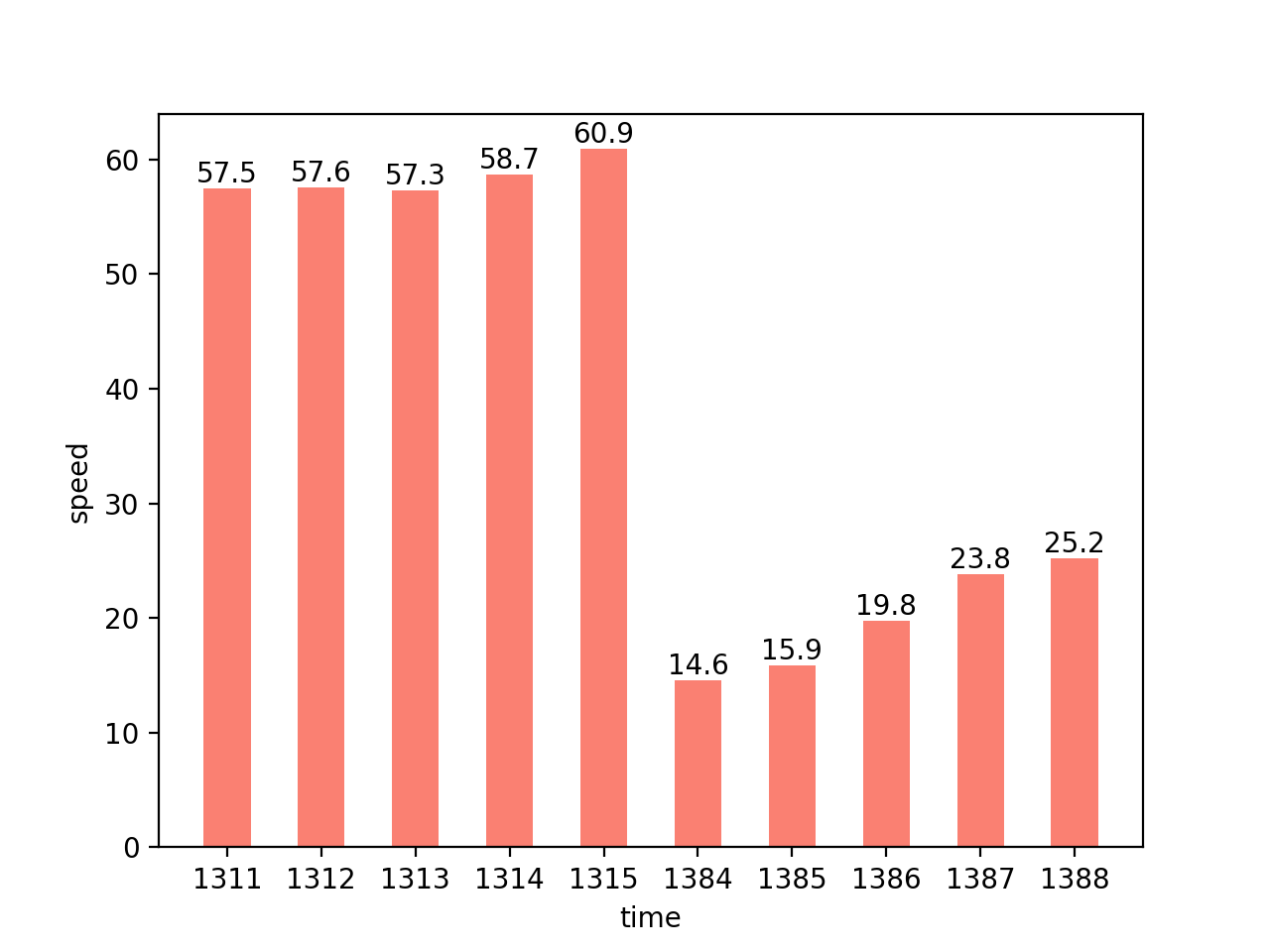
## 3.2 时间不连续的缺失数据处理

汽车行驶数据通过车载的终端采集设备直接采集，采集过程中可能会出现采集设备异常等问题，导致数据丢失。此外，现实路况也可能给数据采集带来一定的影响，例如在汽车过隧道时，会导致GPS信号丢失，无法进行正常的数据采集。这些问题造成了提供的数据中时间不连续，部分数据丢失的问题。

在对缺失数据段进行观察后，我们发现不同的数据段数据缺失的严重程度不同。图x展示了两种不同的缺失程度，图x a中仅缺失了3秒的数据，而图x b中缺失了超过60秒的数据。结合原始数据质量与现实生活的考量，我们设置允许最大不连续时间段的阈值为**30**秒。连续缺失数据未超过阈值的数据段（如图x a仅缺失3秒的数据段）我们通过**线性插值法**对缺失数据进行填充修正处理，而超过阈值的数据我们则进行删除。在构建行驶工况曲线的过程中，我们选择基于运动学片段进行构建。考虑到缺失数据过多会对整个运动学片段带来影响，因此对于那些超过阈值需要被删除的异常数据，我们选择删除该段缺失数据所在的**整个运动学片段**的数据记录。



图x a 缺失较少的异常数据段



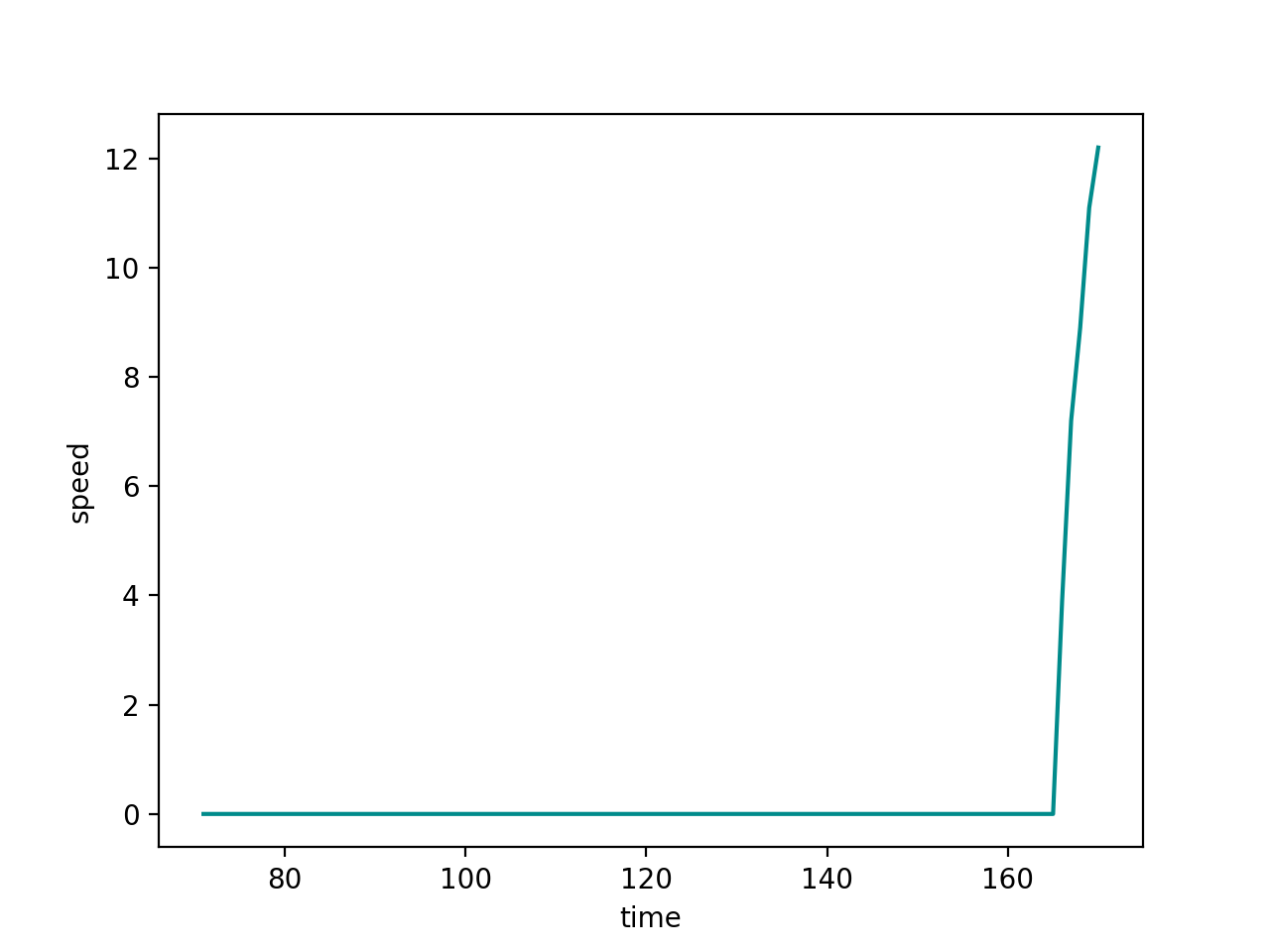
图x b 缺失较多的异常数据段

经过对缺失数据进行处理后，数据文件1剩余总记录数为**136810**条。数据文件2剩余总记录数为**110395**条。数据文件3剩余总记录数为**108808**条。

## 3.3 怠速异常的数据处理

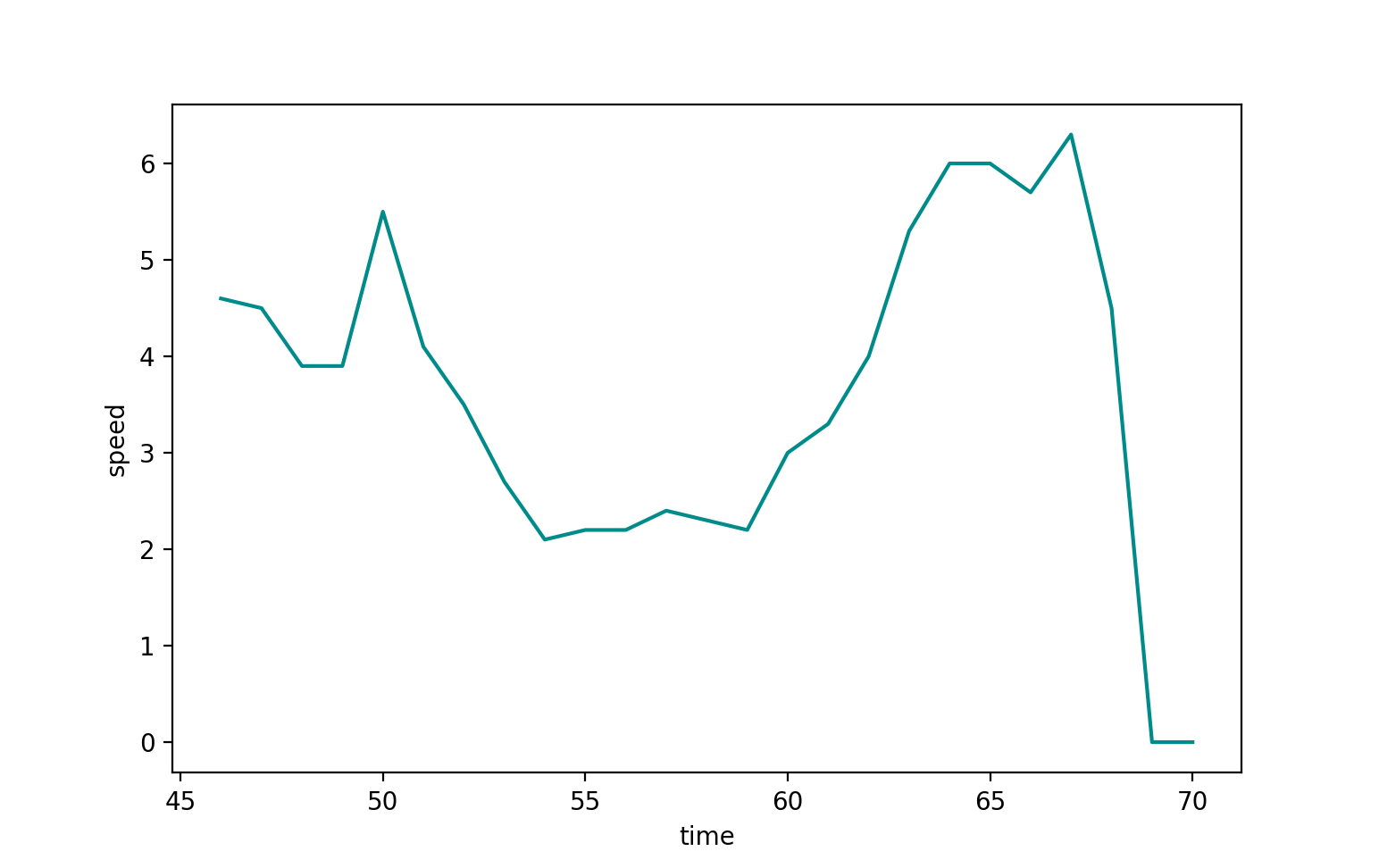
由于汽车行驶工况受交通状况、道路状况、行驶状况等多种主客观因素影响，在数据采集的过程中往往会出现多种类型的行驶数据。我们的原始数据覆盖了工作日与节假日、高峰期与非高峰期等多时段的数据，这其中也包含了几种怠速异常的不良数据。

一种情况如图x所示，数据中存在着一段长时间内行驶速度均为0km/h的怠速异常情况，在现实中长期停车熄火等人，或者停车熄火了但仍在采集数据等都会造成这种异常情况的发生。从上文的基本假设中我们可知，怠速时间超过180秒被认为是异常情况，怠速时间最长可按180秒进行处理。因此对于这类异常我们的处理方法是识别出整段怠速片段，删除多余的异常数据，只保留180秒（即最长怠速时间）内的数据。



图x 长期停车的异常数据

另外一种异常情况与实际更加相关。在现实生活中，高峰期或节假日时期的路况一般较为拥堵，极易出现长时间堵车的情况。这时汽车一般处于断断续续行驶且只能低速行驶的情况。这对应于数据中存在着一段长时间行驶速度不为0但均小于10km/h的异常情况，如图x所示。对于这类不良数据，我们的处理方法是将这类数据按照怠速情况进行处理。若连续低速的数据记录超过60秒（说明是**长时间**堵车）则将其速度记录全部置为0km/h，认为这是怠速状况，并按照上述怠速异常的方法进行处理。若连续低速的数据记录未超过60秒，则认为是正常情况，保留原始的速度记录。



图x 长时间低速行驶的异常数据

经过上面3部分全部预处理之后，数据文件1剔除的异常记录数为54551条，总记录数为**131174**条。数据文件2剔除的异常记录数为37748条，总记录数为**108077**条。数据文件3剔除的异常记录数为58177条，总记录数为**106737**条。

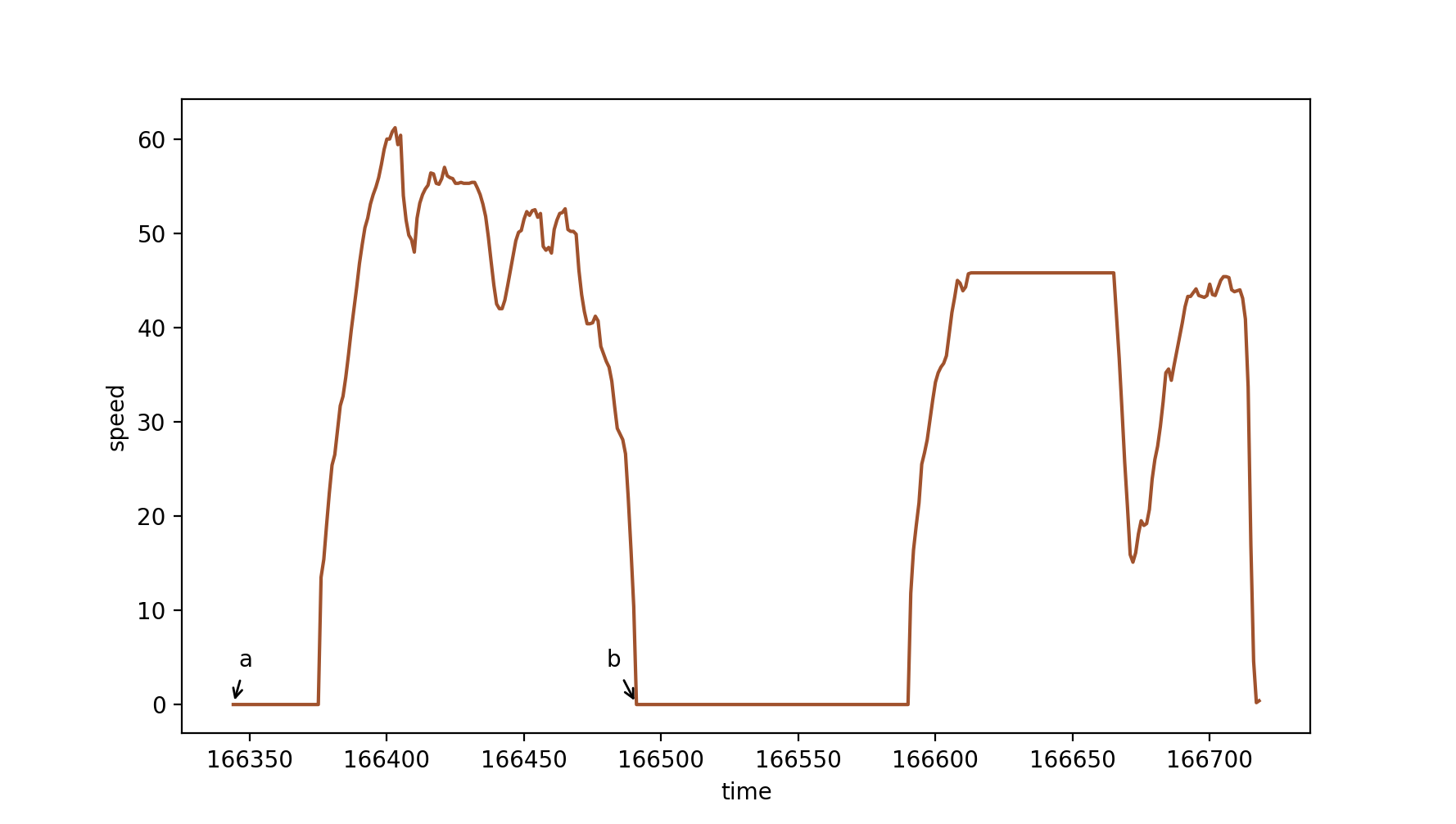
# 4 问题二的分析建模

在现有的构建行驶工况曲线的方法中，基于运动学片段进行构建是最常用的方法之一。在车辆行驶的过程中，会出现频繁的启动、加速、减速等过程，运动学片段就是指车辆从一个怠速状态开始至下一个怠速状态开始之间的运动过程。如下图所示便是一个运动学片段的定义说明。



图x 运动学片段的定义

在经过问题一的预处理后我们得到了一份处理后的数据源，图x展示了处理后的部分数据，图中可以清晰地看出在这段数据中存在着多个运动学片段。我们基于上述运动学片段定义对数据进行划分。自一个怠速状态开始至下一个怠速状态开始被认为是一个运动学片段，图中a-b段即是一个运动学片段。



图x 运动学片段示例

按上述方法划分后，最终数据文件1我们得到了**1181**个运动学片段，数据文件2我们得到了**725**个运动学片段，数据文件3我们得到了**649**个运动学片段。

# 5 问题三的分析建模

## 5.1 特征参数的选取

对于每一个运动学片段，我们从其中提取9种运动学特征参数，并在之后对其进行计算和筛选。详细特征参数见表X。

表X 特征参数

|  |  |  |  |
| --- | --- | --- | --- |
| 编号 | 特征参数 | 编号 | 特征参数 |
| 1 | **平均速度** | 6 | **加速时间比** |
| 2 | **平均行驶速度** | 7 | **减速时间比** |
| 3 | **平均加速度** | 8 | **速度标准差** |
| 4 | **平均减速度** | 9 | **加速度标准差** |
| 5 | **怠速时间比** |  |  |

其中，平均速度为在一个运动学片段内，行驶的距离与时间的比值，而平均行驶速度是不包括怠速阶段在内的平均速度，显然平均速度小于等于平均行驶速度。平均加速度、平均减速度为该运动学片段内正负加速度各自的平均值。怠速时间比，加速时间比，减速时间比则是各行驶状态在整个运动学片段内所占比例，三者之和为1。最后，速度标准差，加速度标准差则是用来衡量速度、加速度的离散程度。

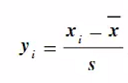
## 5.2 归一化处理与主成分分析

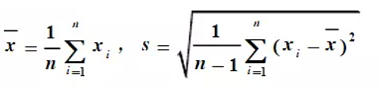
在对每一个运动学片段提取了特征参数后，我们得到了一个2555 \* 9的特征参数矩阵。

为了避免使用所有特征参数作为分类指标，导致信息重叠，我们对这2555 \* 9 的特征参数矩阵进行了主成分分析。

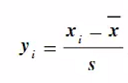
在进行主成分分析之前，为了消除特征参数间不同量纲的影响，我们对特征参数进行了归一化处理。在进行归一化的算法选择上，我们使用了两种归一化算法。

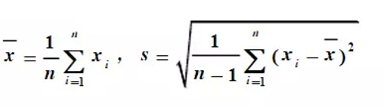
我们选取的第一种归一化方法是Min-Max Normalization，将各个特征参数落到了[0,1]区间内，得到的新的特征值序列均值为0，方差为1，且无量纲。具体公式如下图所示：





我们选择的第二种归一化方法是Zero-mean Normalization，即标准差标准化，基于原特征序列的均值和标准差来进行序列的归一化，具体公式如下：



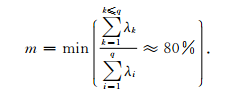


我们对每一张的结果都进行了主成分分析。

第一种方法的分析结果如下表所示：

|  |  |  |  |
| --- | --- | --- | --- |
| 编号 | 贡献率 | 编号 | 贡献率 |
| 1 | **0.55971444** | **6** | **0.01751926** |
| 2 | **0.1947529** | **7** | **0.01433069** |
| 3 | **0.09489261** | **8** | **0.00540254** |
| 4 | **0.07485658** | **9** | **0.00159128** |
| 5 | **0.0369397** |  |  |

鉴于第二种方法区分度较低，我们最终选择了Min-max标准化后的结果。在其累计贡献率约为80%时应保留的主成分为：



基于此公式，本次保留前三个主成分进行聚类。

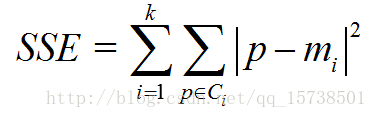
## 5.3 聚类分析

我们使用K均值聚类算法对之前所获取到的2555个运动学片段进行聚类。

k均值聚类是最著名的划分聚类算法，由于简洁和效率使得他成为所有聚类算法中最广泛使用的。给定一个数据点集合和需要的聚类数目k， k均值算法根据某个距离函数反复把数据分入k个聚类中。它是一种迭代求解的聚类分析算法，其步骤是随机选取K个对象作为初始的聚类中心，然后计算每个对象与各个种子聚类中心之间的距离，把每个对象分配给距离它最近的聚类中心。聚类中心以及分配给它们的对象就代表一个聚类。每分配一个样本，聚类的聚类中心会根据聚类中现有的对象被重新计算。这个过程将不断重复直到满足某个终止条件。终止条件可以是没有（或最小数目）对象被重新分配给不同的聚类，没有（或最小数目）聚类中心再发生变化，误差平方和局部最小。

由于K均值聚类算法需要指定K的值，因此我们首先需要决策出合理的K值。在决策K值时我们所采取的方法是“手肘法”。

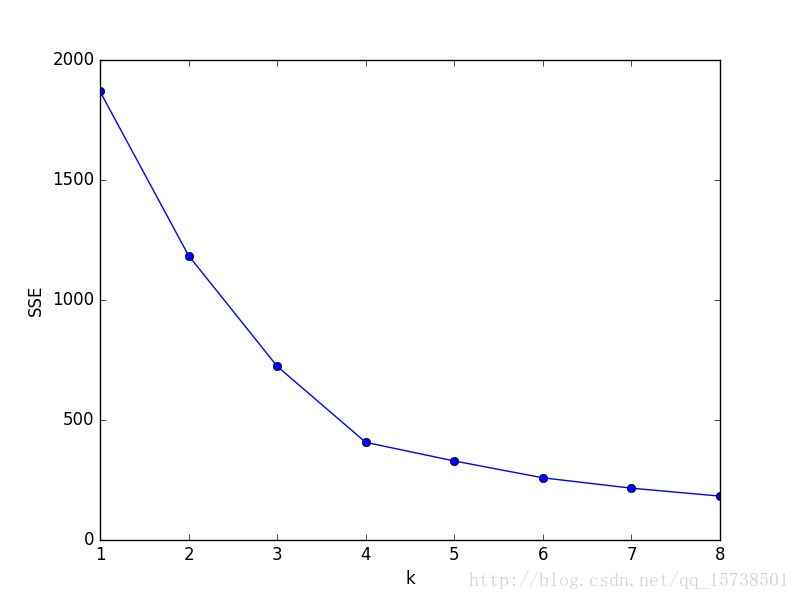
手肘法的核心思想是：随着聚类数k的增大，样本划分会更加精细，每个簇的聚合程度会逐渐提高，那么误差平方和SSE（sum of the squared errors，误差平方和）自然会逐渐变小。并且，当k小于真实聚类数时，由于k的增大会大幅增加每个簇的聚合程度，故SSE的下降幅度会很大，而当k到达真实聚类数时，再增加k所得到的聚合程度回报会迅速变小，所以SSE的下降幅度会骤减，然后随着k值的继续增大而趋于平缓，也就是说SSE和k的关系图是一个手肘的形状，而这个肘部对应的k值就是数据的真实聚类数。这也是该方法被称为手肘法的原因。计算SSE的具体公式如下：



其中，Ci是第i个簇，p是Ci中的样本点，mi是Ci的质心（Ci中所有样本的均值），SSE是所有样本的聚类误差，代表了聚类效果的好坏。

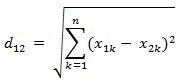
我们假定最合理的聚类结果产生的类别数量不会太多，因此我们计算了K从1到8时的SSE分布曲线，如图X所示。并最终选取了4作为本次K均值聚类算法的K值。

图X SSE-K分布曲线



在进行聚类之前，我们还需要确定用于评估“两个样本间距离”的指标。我们选取了最常用的四种计算距离的公式进行调研，分别是欧氏距离、曼哈顿距离、夹角余弦距离和相关距离。

欧氏距离是这4种方法中最简单直观，也是最常使用的一种方法，其具体公式如下：



曼哈顿距离为标准坐标系上的绝对轴距综合，其结果是两个坐标差的绝对值，是一种“城市区块距离”，其计算公式如下：

C:\Users\Administrator\Desktop\22220180517132100191.png

夹角余弦距离是通过从原点出发，并指向两个位置坐标所产生的两条向量间的夹角来测距，其计算公式如下：

C:\Users\Administrator\Desktop\3333333320180517151015797.png

相关距离依据相关系数来计算距离。相关系数是衡量给定随机变量X与Y的相关性的一种方法，取值范围是[-1,1]。相关系数和相关距离的计算公式分别如下所示：

C:\Users\Administrator\Desktop\44420180517152200717.png

C:\Users\Administrator\Desktop\555520180517152239902.png

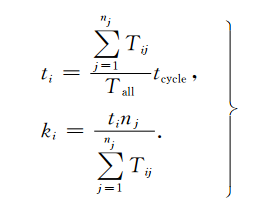
经过调研，我们认为相关距离和夹角余弦距离更适合用于文本数据聚类的计算。而曼哈顿距离与欧式距离较为接近，但通常使用欧氏距离聚类质量更高，且各类聚类算法实现中均默认实现了欧氏距离。因此我们选择了使用欧式距离作为评估样本间相似性的指标。

最后，我们为K均值聚类算法编写了程序进行聚类分析。聚类结果如图X所示。

如果不好弄图，就再改改这句。

## 5.4 工况的合成

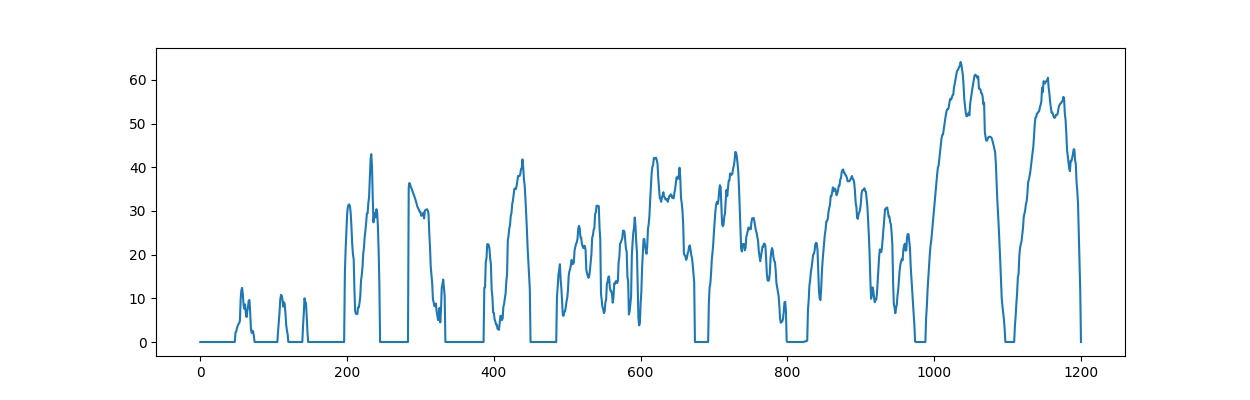
本次我们需要拟合出时间长度在1200~1500s的运动学片段。我们根据各类在整个数据集中所占的时间比例，来确定各类片段在最终工况合成中所占的时间和个数。具体公式如下：



其中：Ti为第i类在工况合成中所占的时间，Ni为第i类片段中片段的个数，Tij为第i类中第j个片段所持续的时间，Tall为样本中所有片段持续的总时间，Tcycle为工况合成所需的时间，Ki为第i类片段在拟合的工况中的片段个数。

依据前一阶段得到的聚类结果，我们拟合出了时间长度为1201s的用车行驶工况，如图X所示。

图X 行驶工况



# 6 模型评估

## 6.1 指标计算

对于我们合成的工况，计算各个运动特征值，如平均速度等

对于处理后的数据集全体进行计算各个特征值

较为简单

## 6.2 合理性分析

看4.1的结果是否接近了。

吹一下不合理的原因

与其他标准比较

改进的思路等

# 参考文献

# 附录