**智能垃圾箱分控系统项目**

**系统总线通信协议及通信API设计文档**

**（Ver1.0）**

|  |  |  |  |
| --- | --- | --- | --- |
| 版本 | 撰写人 | 审核人 | 时间 |
| v\_1.0 | 朱文强 | 赵蕴龙 | 2021.1.25 |
|  |  |  |  |
|  |  |  |  |

2021年1月25日

南京航空航天大学

目录

[一、 项目介绍 4](#_Toc28381192)

[1.1项目任务 4](#_Toc28381193)

[1.1.1 公交数据多维分析模型 4](#_Toc28381194)

[1.1.2 公交车乘客下车人数测算模型 4](#_Toc28381195)

[1.1.3 公交客流OD分析模型 4](#_Toc28381196)

[1.1.4 公交乘客常驻地识别模型 4](#_Toc28381197)

[1.1.5 公交乘客出行特性分析模型 5](#_Toc28381198)

[1.2 任务分析 5](#_Toc28381199)

[1.2.1 公交数据多维分析 5](#_Toc28381200)

[1.2.2 公交车乘客下车人数测算 5](#_Toc28381201)

[1.2.3 公交客流OD分析 6](#_Toc28381202)

[1.2.4 公交乘客常驻地识别 6](#_Toc28381203)

[1.2.5 公交乘客出行特性分析 6](#_Toc28381204)

[二、 方案设计 7](#_Toc28381205)

[2.1公交数据多维分析 7](#_Toc28381206)

[2.1.1 后台管理 7](#_Toc28381207)

[2.1.2 站点客流可视化 7](#_Toc28381208)

[2.1.3 线路客流可视化 10](#_Toc28381209)

[2.1.4 OD客流可视化 11](#_Toc28381210)

[2.1.5 正晚点分析 12](#_Toc28381211)

[2.1.6 线路运行时间分析 13](#_Toc28381212)

[2.2 公交车乘客下车人数测算 13](#_Toc28381213)

[2.2.1数据预处理 13](#_Toc28381214)

[2.2.2公交出行链理论 14](#_Toc28381215)

[2.2.3闭合公交出行链模型设计 15](#_Toc28381216)

[2.2.4非闭合公交出行链模型设计 15](#_Toc28381217)

[2.3 公交客流OD分析 17](#_Toc28381218)

[2.4 公交乘客常驻地识别 17](#_Toc28381219)

[2.4.1 影响因素 17](#_Toc28381220)

[2.4.2 数据预处理 17](#_Toc28381221)

[2.4.3 常驻地提取 18](#_Toc28381222)

[2.5 公交车乘客出行模式判断 18](#_Toc28381223)

[2.5.1 数据预处理 18](#_Toc28381224)

[2.5.2 K-Means聚类算法 19](#_Toc28381225)

[三、 方案实现 20](#_Toc28381226)

[3.1 公交数据多维分析 20](#_Toc28381227)

[3.1.1 后端实现 20](#_Toc28381228)

[3.1.2 前端实现 21](#_Toc28381229)

[3.2 公交车乘客下车人数测算 21](#_Toc28381230)

[3.3 公交客流OD分析 22](#_Toc28381231)

[3.4 公交乘客常驻地识别 22](#_Toc28381232)

[3.4.1 常驻地站点信息实现 22](#_Toc28381233)

[3.5 公交乘客出行特性分析 23](#_Toc28381234)

# 项目介绍

## 1.1项目任务

### 1.1.1 公交数据多维分析模型

通过对公交刷卡数据、公交行车计划数据以及GPS数据的处理分析，从客流和行车两方面进行可视化。

### 1.1.2 公交车乘客下车人数测算模型

根据给定的公交车乘客IC卡上车刷卡数据、公交站台数据、公交车到离站数据，构建概率模型，预测每位乘客下车站点。

### 1.1.3 公交客流OD分析模型

根据给定的公交车乘客IC卡刷卡数据和交车乘客下车人数测算模型得到的乘客下车站点结果，构造基于站点的公交客流OD分析模型，分析OD中密集的公交客流。

### 1.1.4 公交乘客常驻地识别模型

以公交OD分析的结果作为支撑，对乘客的常驻地地点进行识别，确定乘客最长刷卡地点是什么类型，如居住地点、工作地点、学习地点、休闲地点、访友地点等。

南京市作为一个有着840万常住人口的城市，根据2018年12月产生的250万条刷卡记录对公交常住地进行分析具有相当大的意义，常住地分析得出的地点种类对城市规划，交通建设以及推断居民的消费水平并提供个性化的产品推荐都有一定的指导意义。

### 1.1.5 公交乘客出行特性分析模型

公交客流分析是公交出行需求辨识，线网评估与优化，以及公交优先发展政策制定和实施的基础。乘客出行，公交线网客流数据的传统调查方法成本较高，受限于调查技术和调查误差，精度难以保证。现阶段，公交IC卡，GPS系统的推广使用，使得低成本的自动化采集海量数据并加以分析处理成为现实。

## 1.2 任务分析

### 1.2.1 公交数据多维分析

因为原始数据并不利于直接展示，所以需要将数据根据不同需求绘制成折线图、柱状图、饼状图、平面地图、百度地图、路径图等多种方式来进行直观的展示。

主要分为客流可视化与行车任务可视化两个大类。其中客流可视化包括站点客流、线路客流和OD客流。行车任务可视化包括正晚点分析和线路运行时间分析。每一个模块的实现首先需要后台进行数据预处理之后再发送到前台。预计后期还会加入用户权限管理。

### 1.2.2 公交车乘客下车人数测算

公交车乘客IC卡刷卡数据中包含了上车站点信息，无下车站点信息，因此需要构建合适的模型输出每位乘客的下车站点信息。项目中使用的IC卡刷卡数据包含了2018年12月一整个月的所有刷卡信息，其中包含的乘客数量较大且总的刷卡次数达百万条，因此需要对数据进行合适的预处理，方便后续进行分析。在初步分析后，发现数据中，乘客的刷卡记录存在不连续的现象，且单个乘客一个月内的刷卡记录中，每天的刷卡次数或多或少（少则1条，多则5、6条），存在不规律性，因此需要进一步考虑下车站点的判断方法。

### 1.2.3 公交客流OD分析

OD矩阵是表示乘客在区域间出行情况的记录方式，在公交系统中，OD矩阵体现了公交站点间的连通性，通过对OD矩阵的分析可以从整体公交系统上发现客流的流动特点，例如两个站点间在某个时刻有大量的密集乘客出行。

### 1.2.4 公交乘客常驻地识别

公交乘客的常驻地与公交出行轨迹是获取公交出行数据、乘客出行特征、乘客出行行为特征等的重要数据基础。目前关于用户职住地识别的研究很多，但是用户的活动主要活动地点不仅仅局限在单位和住所，还包括很多其他常去地点（如超市、饭店、商场、学校等），所以需要对公交乘客常驻地点的自动准确识别进行更加细致深入的研究。具体常驻地点分为居住地点、工作地点、学习地点、休闲地点、访友地点等。

乘客出行一般是从常住地出发经过一个或多个目的地后返回常住地, 完成一次出行。他们在目的地的停留时间要远小于在常住地停留的时间。通过分析乘客的公交出行数据从而推断旅乘客的常住地。

### 1.2.5 公交乘客出行特性分析

本次研究提取中南京市2018年12月一整个月的所有刷卡信息，包含的乘客数量较大且总的刷卡次数达两百万条，因此需要对数据进行合适的预处理，方便后续进行分析。

经过前期的工作处理，现阶段的公交车乘客IC卡刷卡数据中包含了上车站点信息和下车站点信息，但存在多次换乘记录，因此需将换乘记录整合，得到一条包含去程刷卡时间，去程刷卡站点，返程刷卡时间，返程刷卡站点的信息。

# 方案设计

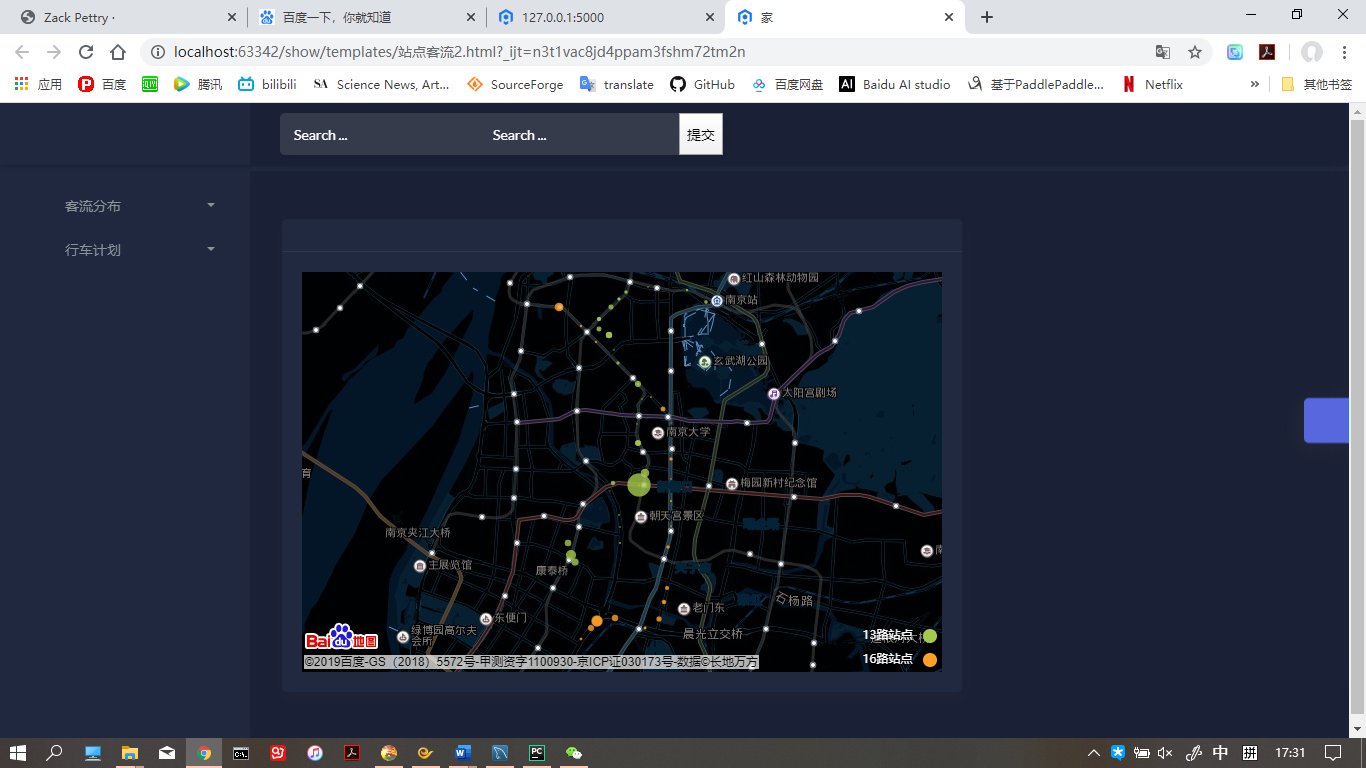
## 2.1公交数据多维分析

### 2.1.1 后台管理

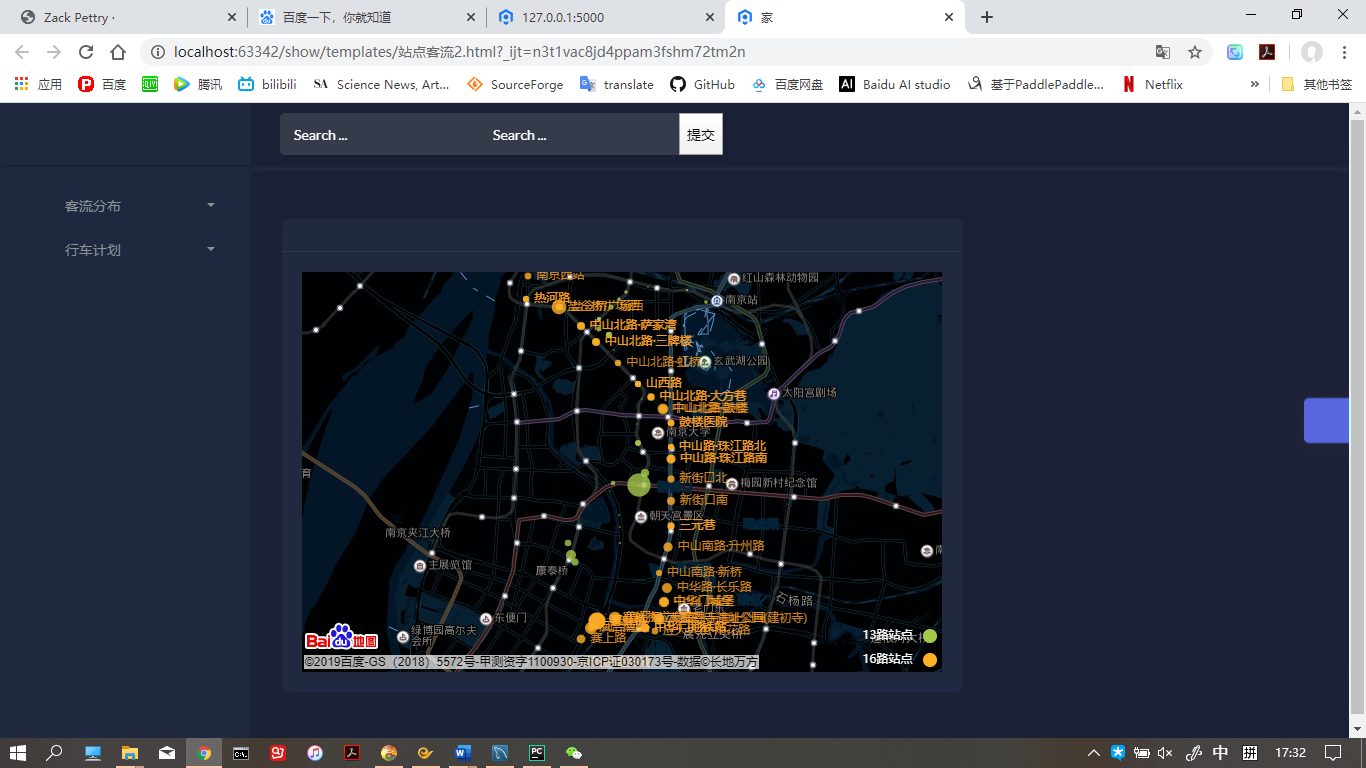
1. 数据查询。连接数据库，在前端模块发送请求之后查询数据库，将查询到的数据进行预处理，以restfulapi的形式发送给前端。
2. 用户管理。在用户登录之后返回用户所具有的权限给前端。添加与删除用户暂不做可视化配置，在系统中更改。
3. 日志查询。将日志以天的形式自动存储至服务器，内容包括用户登录、查询信息、系统运行情况等。

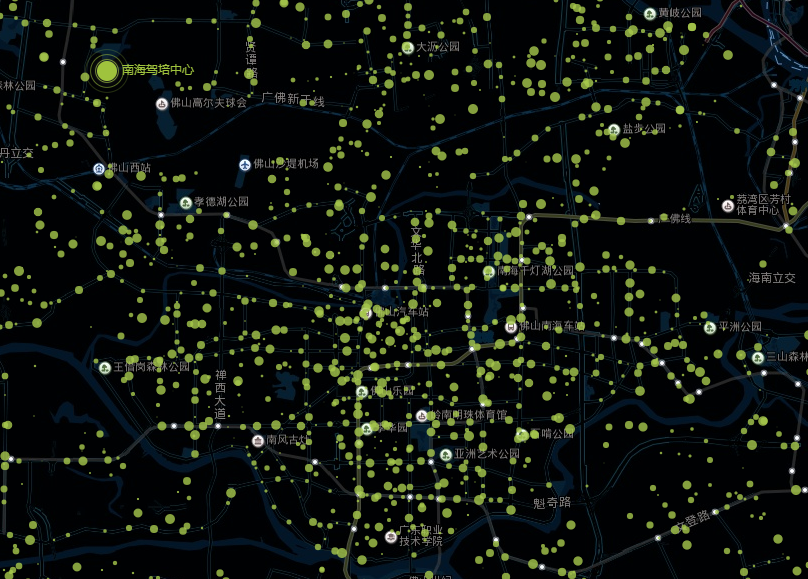
### 2.1.2 站点客流可视化

1. 按照下拉框的方式可以选择日期，路线（可不选）。根据经处理好的按照线路分类好的刷卡信息数据得到所选日期的客流量，按照其线路分类按照不同颜色别呈现到地图上（如选择线路则直接显示所选线路）。如下图所示（13路和16路）。

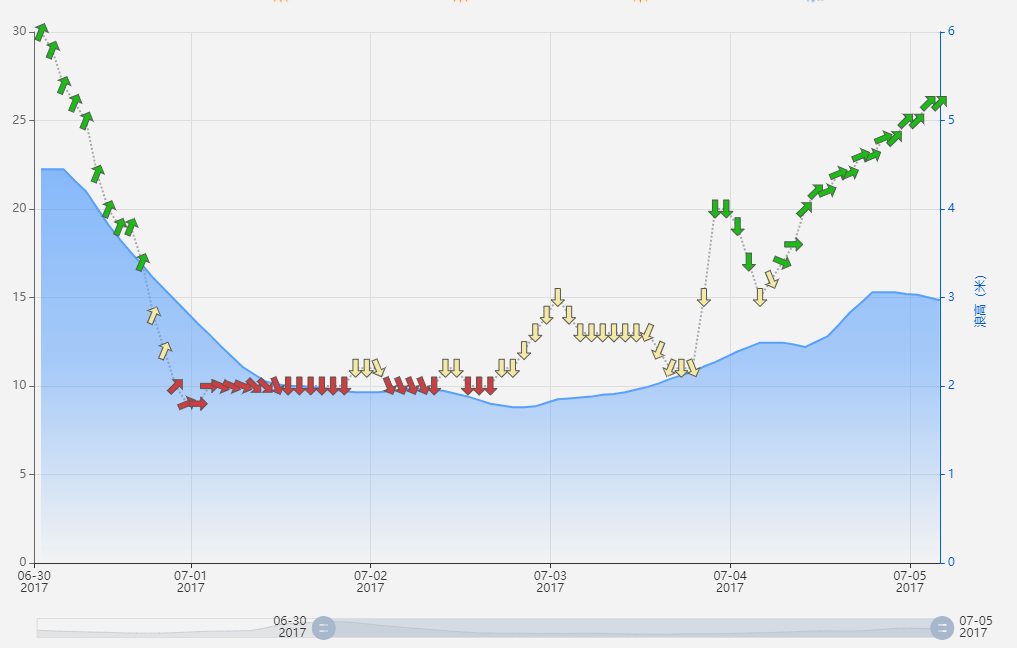


还可以显示出每一个站点的站点信息



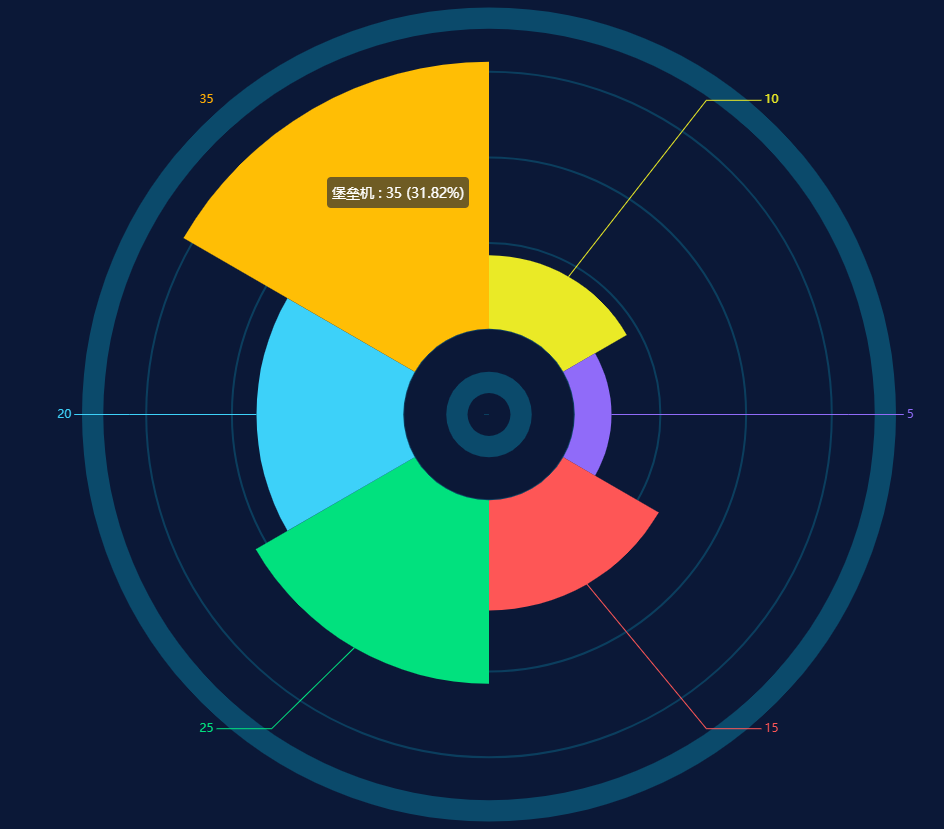
1. 以地图为背景，显示默认时间（本项目12月整月）全市所有站点（不显示线路）的流量，站点用圆圈表示，不同客流量的站点圆圈用不同的大小、颜色展示。
2. 点击地图上的某一站点后，显示该站点的具体信息，包括属于哪条线路，乘坐线路1的乘客多少人，乘坐线路2的乘客多少人，同时该线路可点，点击后，再地图上显示该线路的途径站点以及线路客流，其他站点进行隐藏。

在地图下方设置一个时间的拖动条，用于缩放一天24小时的客流流量，当缩放时间时，地图数据和柱状图数据均要一起变动。拖动条类似于：



### 2.1.3 线路客流可视化

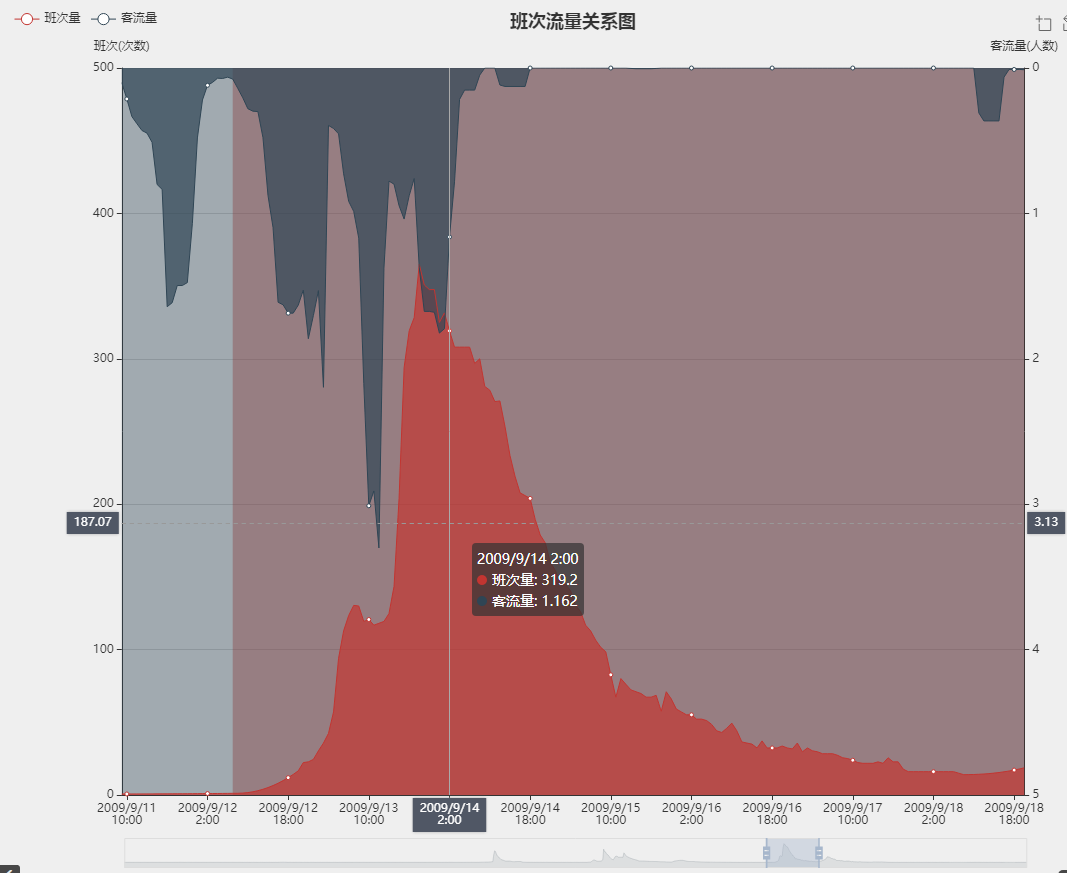
1. 页面默认显示南京市所有的线路信息，以扇形图表示，内容为各个线路的客流分布，扇形块表示线路，扇形块长度表示客流量，上方设有时间、线路选择框。



选择某一条具体的线路后，扇形图表示为该线路的所有站点信息，扇形块表示站点名称，长度表示一天中的总客流量。右侧增加条形图用于显示客流的具体信息，横坐标是时间段，纵坐标是上下车客流量。

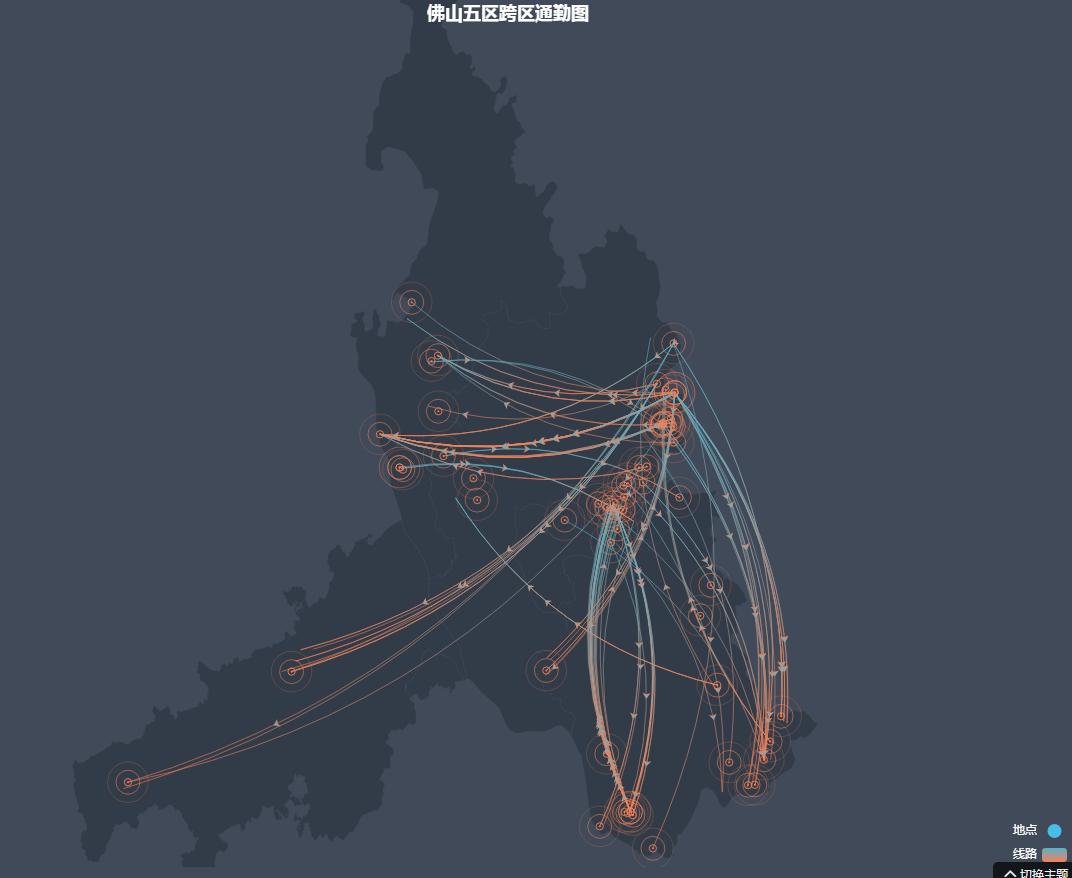


1. 选择时间和线路后，展示线路发车量和客流之间的关系，分别从到离站表和整理出来的每个线路的表获得所选日期的数据，使用下图表示，横坐标表示该线路的站点，左横坐标表示选中的时间维度下，到达各个站点的车次量，右纵坐标表示选中的时间维度下的上车客流量。

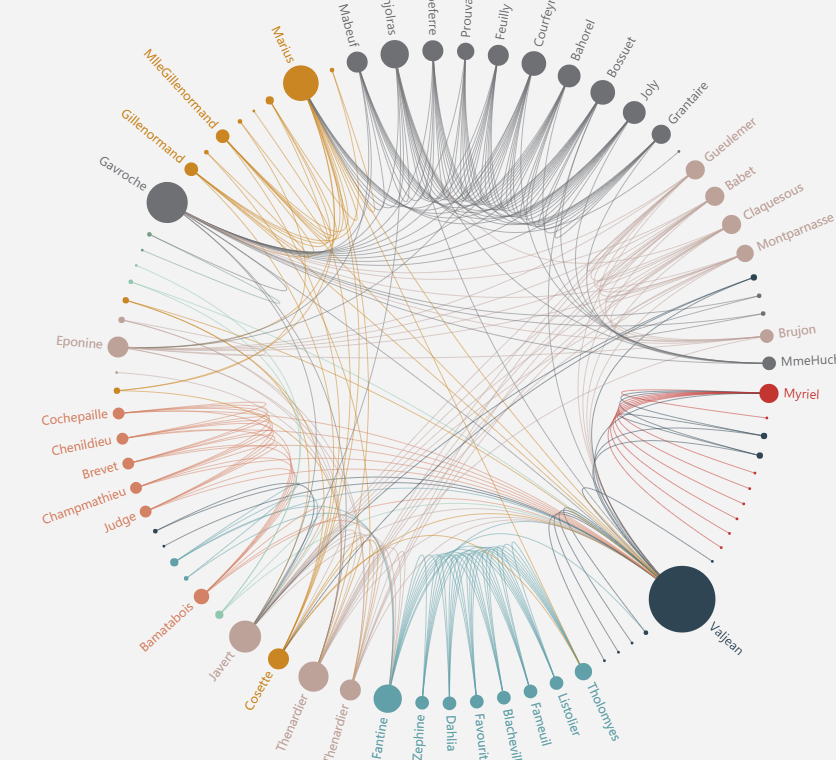


### 2.1.4 OD客流可视化

利用构建好的南京轮廓图模型，使用OD数据，根据数据的上下车站点，对全市数据范围内的OD进行可视化。一个OD点对在地图上表示为一条曲线，需要可选择时间段（月、天、时、）进行查看全市的OD。



同时，可以选择不同的线路进行查看OD。选中线路后，可以对该线路的OD进行展示，下图：



图中圆点表示不同的站点，圆点大小表示站点的客流量。两个点之间的连线表示乘客从其中一点上车另一点下车。

除此之外，可以选中某一站点查看OD，并统计从某一站点出发，去往不同站点的数量分布，用扇形图表示。

### 2.1.5 正晚点分析

选定某一天行车计划和线路，把选中线路的预期的排班时间从数据库中获取，然后再获取这一线路当天的实际出发实际。把这两个种数据以对比柱状图的形式呈现出来。下图为例，蓝色为计划时间，黄色为实际出发时间。



### 2.1.6 线路运行时间分析

用户从下拉框选择好日期和线路，从离站的表中获取这条线路早中晚的各两条数据，一共六条的数据，采用折线图的形式，对确定时间维度的所有到站点信息的运行时间进行展示，横坐标为该线路的所有站点，纵坐标为岛之战时到站

下图为一条数据信息的折线图



## 2.2 公交车乘客下车人数测算

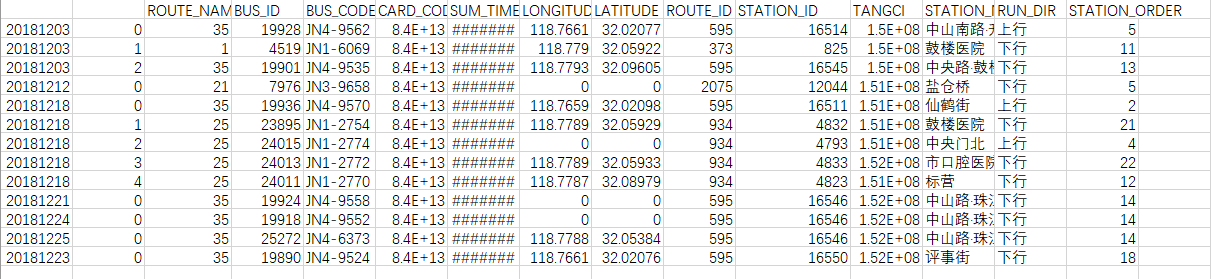
### 2.2.1数据预处理

1）在2018年12月的IC卡刷卡数据中，我们发现存在一些乘客在同一时间、同一公交线路、同一辆公交车上重复刷卡的现象，由于无法准确判定该乘客是否存在替他人刷卡的情况，且这类刷卡情况也比较多，因此，我们直接删除了这些干扰信息；

2）在2018年12月的IC卡刷卡数据中，乘客的上车站点未明确给出，仅提供了上车位置的经纬度信息，并且通过与提供的公交站台的经纬度进行对比后发现，两者的经纬度在计算和匹配上皆存在较大的误差。

进一步观察数据后发现，乘客刷卡的时间记录与公交车到离站的时间较为接近，因此，我们使用刷卡时间与到离站时间进行匹配，以此得到每位乘客具体的下车站点，将原始数据补充完整；

3）为了方便后续下车站点的预测，我们将2018年12月的IC卡刷卡数据重新导出成多个csv文件且以该乘客的IC卡ID为文件名，每个文件中仅保存了一位乘客在12月内的所有刷卡记录，每条记录涵盖以下关键信息：日期、当前的刷卡次序、乘坐的线路名称、乘坐车辆的ID和编码、该乘客的IC卡ID、刷卡时间、刷卡位置的经纬度、线路ID、上车站点ID和名称、趟次信息、站台顺序、运行方向。如下图所示：



### 2.2.2公交出行链理论

公交出行链的定义：按一定时间顺序，居民乘坐在时间、空间上前后衔接的公共交通线路出行的公交出行过程。

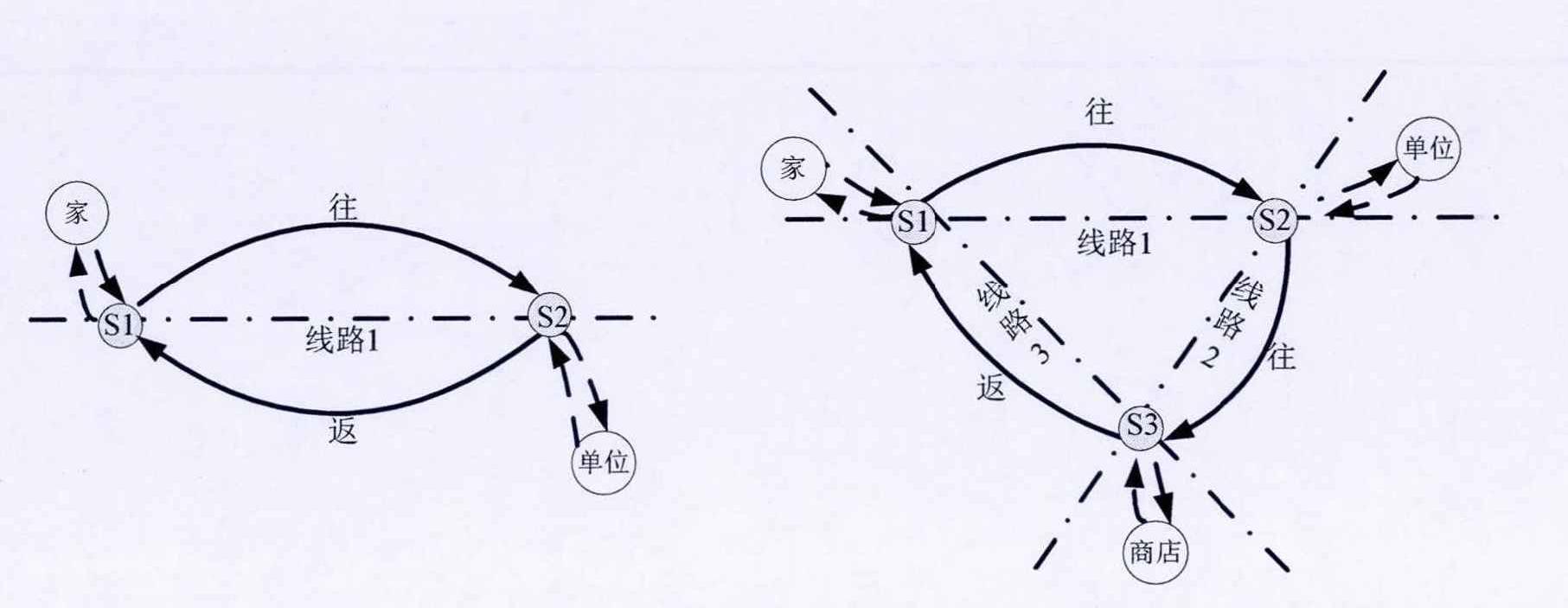
公交出行链的特点：居民在一天的出行过程中，至少乘坐了2次及以上的公共交通，乘客前一次乘车的下车时间一定小于后一次乘车的上车时间，乘客一天内乘坐的公交线路可能构成一个闭合的出行链，也可能构成一个非闭合的出行链。

### 2.2.3闭合公交出行链模型设计

闭合出行链的定义：乘客在一天内乘坐的公交线路能够通过站点间的衔接构成一个闭合的环路。

假设该乘客由家出发，最后回到家的过程都乘坐了公共交通，则乘客乘坐的公交线路构成一个闭合公交出行链，如果乘客到达目的地完成某项出行目的后，返程过程按原路返回，该类型的闭合公交出行链将其定义为原路径闭合公交出行链，如下左图所示；如果乘客在一天的出行过程中有多个目的地，乘客到达最后一个目的地后，按原路线回家的路线不再是最短路径，乘客将选择距离短、行程简单的路线回家，该类型的闭合公交出行链将其定义为环形闭合公交出行链，如下右图所示。

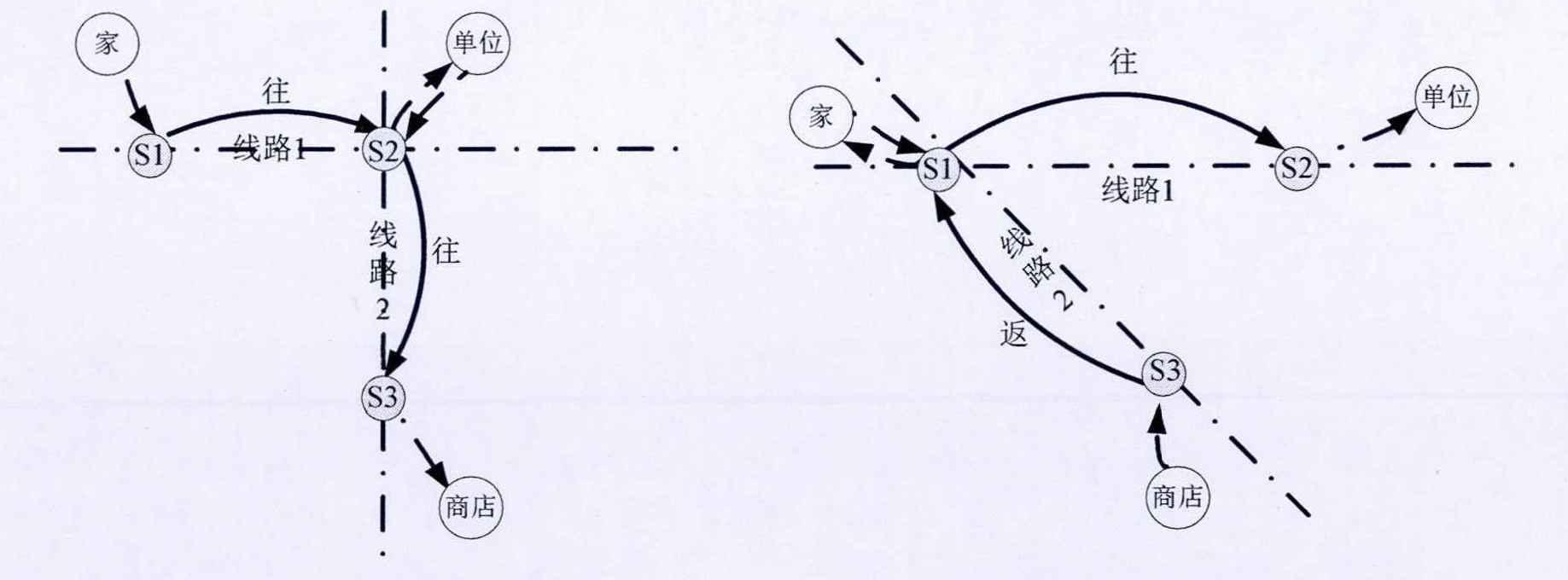
能够行程闭合公交出行链的公交出行，每次公交乘车的下车站点存在以下特点：根据乘客乘车的先后顺序，除最后一个乘车外，任何一次乘车过程的下车站点与下次乘车的上车站点能够形成空间衔接关系，乘客前后两次乘坐的是同一线路的公交车辆时，前后2次乘车过程一定在同一站点形成衔接。



### 2.2.4非闭合公交出行链模型设计

非闭合公交出行链的定义：乘客在一天内多次乘坐公共交通过程中，存在这样的情况，乘客在一天的出行过程中，选择了公共交通以外的交通方式，使得乘客一天内乘坐的线路或者部分线路能够相互衔接，但相互衔接的公交线路不能构成一个闭合的环路。

非闭合公交出行链与闭合公交出行链比较，主要区别在相对闭合公交出行链的某个环节出现断链，断链的形式可能为首尾断开，也可能为中部断开，首尾断开以及中部断开的非闭合公交出行链如下图所示。



在非闭合的公交出行链中，对于任何一次不是出现断链的乘车过程与闭合公交出行链的乘车过程具有相同的下车站点特征，为此对出现断链的公交出行链，采用不同的概率模型进行判断。

针对项目的目标数据，首先将上面的非闭合公交出行链分为两种情况：

第一种，出行链断裂，即上图所示，但历史出行记录中存在多日相似的起始地目的地重合的出行记录；

第二种，当前公交出行不连续，且没有相似的可供参考的出行记录。

对于第一种情况，乘客存在历史相似出行记录，因此其概率是站点上车次数与线路乘坐次数的比值，即，乘客在站点上车，则在该线路下行站点下车的概率为：

，为乘客乘坐该线路在各站点的上车次数，最后选取概率最高的站点为其下车站点。

对于第二种情况，出行链较为残缺，没有历史出行记录，因此其概率算法为，即，乘客在站点上车，则在该线路下行站点下车的概率为：

其中，为该站点上车人数，为该站点所有下游站点的上车总人数。

## 2.3 公交客流OD分析

### 2.3.1 数据预处理

在2018年12月的IC卡刷卡数据中，通过公交车乘客下车人数测算中下车站点识别模型推测每次刷卡数据所对应的下车站点，构成一个OD数据集，每条OD数据包含一个刷卡数据的刷卡时间，刷卡数据中的上车站点和对应的预测下车站点。

在刷卡数据得到的下车站点预测结果中，存在部分由于闭合出行链而预测得到的刷卡站点与下车站点相同和距离很近的站点，通过遍历找到这些数据并删除。

### 2.3.2 基于密度的聚类算法

DBSCAN是基于密度聚类算法的简称，这个算法旨在发现空间平面上任意形状的高密度对象簇。在公交客流中，DBSCAN可以把每次公交出行看为一个对象，把出行的上车站点和下车站点作为位置数据，结合乘客的刷卡时间，可以把每次出行视为一个在空间中分布的三维数据点，从而体现出公交客流的分布样式，并且发现密集的公交客流。

DBSCAN的具体步骤如下：

步骤 1.初始化核心对象集合，种子集合，初始化类别数，初始化大小等于样本集大小的标签集合。

步骤 2. 遍历样本集D中的元素，如果是核心对象，则将其加入至种子集中，判断核心对象的方法见定义2与公式3。

步骤 3. 如果核心对象集中的元素都已经被访问，则算法结束，输出结果，否则转入步骤4。

步骤 4. 在核心对象集合中，随机选择一个未被访问过的核心对象，将标记为已访问，然后根据在中的索引对应地将中的标记为类别，最后将的时空邻域内的所有未访问的对象存放至种子集合中，的时空邻域见定义1与公式3。

步骤 5. 如果种子集合，则当前第k类生成完毕，k=k+1，跳转至步骤3。否则，从种子集合中任选一个种子点seedj，将其标记为已访问，根据索引对应地将中的标记为类别，判断seedj是否为核心对象，如果是，则将seedj的时空邻域中的所有点加入到种子集合中。跳转至步骤5。

## 2.4 公交乘客常驻地识别

### 2.4.1 影响因素

利用乘客公交出行数据推断常住地信息, 受出行数据质量影响, 主要有:

(1) 出行次数过少。部分乘客在统计周期内的出行次数过少, 不能形成有效的出行链, 无法在出行起点与出行终点之间确定常住地, 这些旅客的常住地不能通过公交出行数据进行识别。

(2) 行程不连续。综合交通背景下, 乘客可组合多种交通方式完成出行, 导致公交出行数据在整个行程上是不连续的, 该类型旅客需要结合其他交通方式的出行数据进行判断。 此时单独使用公交数据进行分析就显得力不从心。

(3) 目的地最大停留时间。根据不同的出行目的, 乘客在目的地的停留时间一般会有一个时间上限, 当乘客在目的地的停留时间过长时乘客可能存在多个常住地, 比如一定期间内出差在外的乘客，该情况有效无法识别乘客常住地。

### 2.4.2 数据预处理

针对上述的影响因素对数据进行相应的预处理。

对于出现频繁刷卡情况的IC卡数据进行删除，因为这些出行数据符合出行影响因素一。这类乘客出行太过频繁，不适合作为推断常住地的数据集。

对于按照刷卡时间排序的两次刷卡数据，如果他们通过经纬度计算的位置相隔甚远，此类符合影响因素2的数据也需进行删除。

对于在多地停留时间相似的数据也需进行删除，这满足影响因素3的情况。

执行数据预处理的流程图如下：



### 2.4.3 常驻地提取

首先根据出行特性为每条数据打上标签，之后根据出行特性对该IC卡的出行数据进行常住地的区分。并将常住地信息按种类显示。

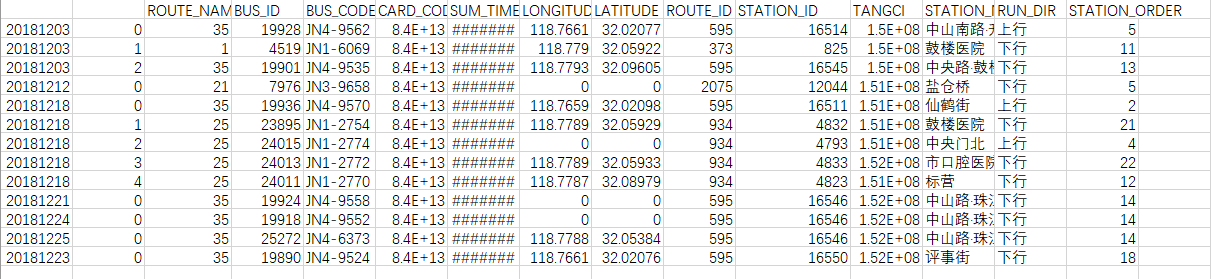
## 2.5 公交车乘客出行模式判断

### 2.5.1 数据预处理

1）在2018年12月的IC卡刷卡数据中，同一IC卡在一天内可能存在多次刷卡记录。乘客为到达某一目的地会在短时间内进行多次的转乘，换乘的数据增加了数据量并对乘客出行模式的判断有一定的影响。因此需整合所有的转乘信息，将多次乘车记录统计为一条涵盖出发点与目的地的记录，将中间的换乘记录进行略去。

2）对于如IC卡号，乘车时间，乘车站点，这些主要信息缺失的数据进行删除处理。

3）为了方便对出行模式进行判别，将2018年12月的IC卡刷卡数据按日期进行划分，以天为单位并将数据重新导出。每个文件中仅保存了这一天内的所有刷卡记录，每条记录涵盖以下关键信息：日期、当前的刷卡次序、乘坐的线路名称、乘坐车辆的ID和编码、该乘客的IC卡ID、去程刷卡时间、刷卡位置的经纬度、线路ID、去程上车站点ID和名称、趟次信息、站台顺序、运行方向，返程刷卡时间，返程上线站点ID和名称。如下图所示：



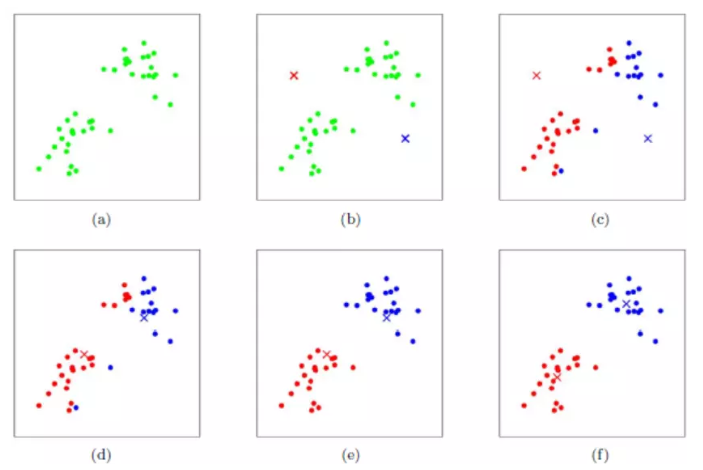
### 2.5.2 K-Means聚类算法

聚类算法的目的是依据已知的数据，将相似度高的样本集中到各自的簇中。聚类算法不仅可以将数据实现分割，还可以用于异常点的监控，所谓异常点就是远离任何簇的样本。

Kmeans利用距离远近的思想将目标数据聚为指定的K个簇，簇类样本越相似表明聚类效果越好。该算法的实现方式为不断的计算各样本点与簇中心之间的距离，直到收敛为止。其具体步骤如下：

1. 从数据中随机挑选K个样本点作为原始额簇中心。
2. 计算机剩余样本与簇中心的距离，并把各样本标记为离K个簇中心最近的类别。
3. 重新计算各簇中样本点的均值，并以均值作为新的K个簇中心。
4. 不断重复(2)和(3)，直到簇中心变化趋于稳定，形成最终的K个簇。

算法的聚类过程示意图如下：



关于聚类中最佳K值得确定，采用轮廓系数法，该方法综合考虑了簇的密集性与分散性两个信息。如果数据集被分割为理想的K个簇，那么对应的簇内样本会很密集，而簇间样本会很分散。轮廓系数的计算公式可以表示为：

其中，a(i)体现了簇内的密集性，代表样本i与同簇内其他样本点距离的平均值；b(i)反映了簇间的分散性，它的计算过程是，样本i与其他非同簇样本点距离的平均值，然后从平均值中挑选出最小值。

# 方案实现

## 3.1 公交数据多维分析

### 3.1.1 后端实现

后端采用spring boot框架进行开发，使用到的模块有spring mvc、spring security、hibernate等，与前端通信使用restfulapi的形式。容器采用tomcat。

具体的硬件需求需要根据实际的使用情况而定。

数据库连接组件采用spring data jpa进行连接，默认设置连接有效时间为一个月。每一个前端模块对应后端唯一一个控制器模块以便于权限管理。返回数据以json的形式进行封装。

### 3.1.2 前端实现

前端采用HTML5，CSS结合VUE框架进行实现。图表组件主要使用百度开源插件Echarts。

各个模块以独立页面的形式进行展示，由于Echarts绘图所需的系统开销较大，所有的数据处理尽量交由后端处理，每个接口另整理成文档以便查看。

## 3.2 公交车乘客下车人数测算

由于乘客刷卡次数不等，出行链也不同，因此考虑分刷卡不同刷卡次数的情况。

1. 刷卡次数为1次。这种情况下由于无后续刷卡做参考，因此使用第二种公交出行不连续的概率模型进行判断下车站点；
2. 刷卡次数为2次。这种情况下，如果该乘客一个月内不同天中，存在2条及以上的相似刷卡记录，则说明构成完整出行链，于是将最后的刷卡记录作为前一次刷卡记录的下车站点，第一次刷卡的站点作为第二次刷卡后的下车站点。

如果一个月内不存在相似的刷卡记录，则说明该出行链断裂，此时如果两次刷卡在同线路上，根据该乘客两次刷卡所乘坐车辆的线路方向（上行还是下行），若两次刷卡皆为上（或者下）行，就通过第一种公交出行断裂的概率模型来计算最后一次刷卡的下车站点；如果存在第一次上（下）行，第二次下（上）行，就通过第二种出行不连续概率模型进行判断第二次刷卡的下车站点。

又如果一个月内不存在相似的刷卡记录，两次刷卡不在同一线路，此时需要通过第二种出行不连续的概率模型来判断第一次刷卡的换乘下车站点。

3）刷卡次数为3次。这种情况下，如果该乘客一个月内不同天中，存在2

条及以上的相似刷卡记录，则说明构成完整出行链，因此将最后的刷卡

记录作为前一次刷卡记录的下车站点，第一次刷卡的站点作为第二次刷

卡后的下车站点。

如果一个月内不存在相似的刷卡记录，则说明该出行链断裂，此时如果

三次刷卡在同线路上，根据该乘客三次刷卡所乘坐车辆的线路方向（每

一次是上行还是下行），若三次刷卡皆为上（或者下）行，就通过第一

种公交断裂的概率模型来计算最后一次刷卡的下车站点；如果存在至少

有一次刷卡与另两次不同方向的情况（即上上下、下下上等），此时第

二种出行不连续的概率模型来判断第一次刷卡的换乘下车站点。

又如果一个月内不存在相似的刷卡记录，三次刷卡中有不在同一线路上

的情况，此时需要通过第二种出行不连续的概率模型来判断中间的换乘

下车站点。

4) 刷卡次数为3次以上时，判定下车站点的方法与上面两种方法类似。

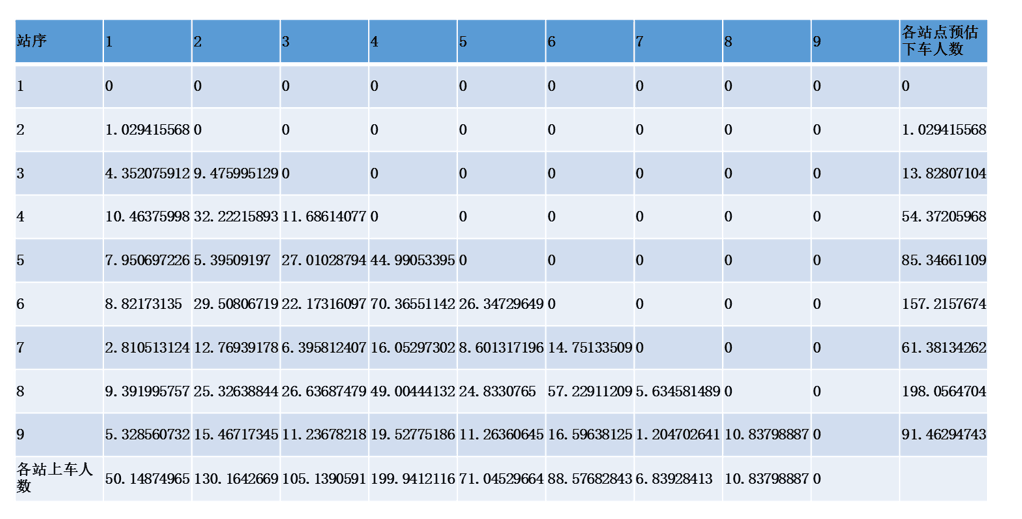
## 3.3 公交客流OD分析

### 3.3.1 数据预处理实现

通过获取公交车下车站点识别得到的结果，提取刷卡时间，刷卡上车站点，下车站点组成一条OD数据，具体样式如下图

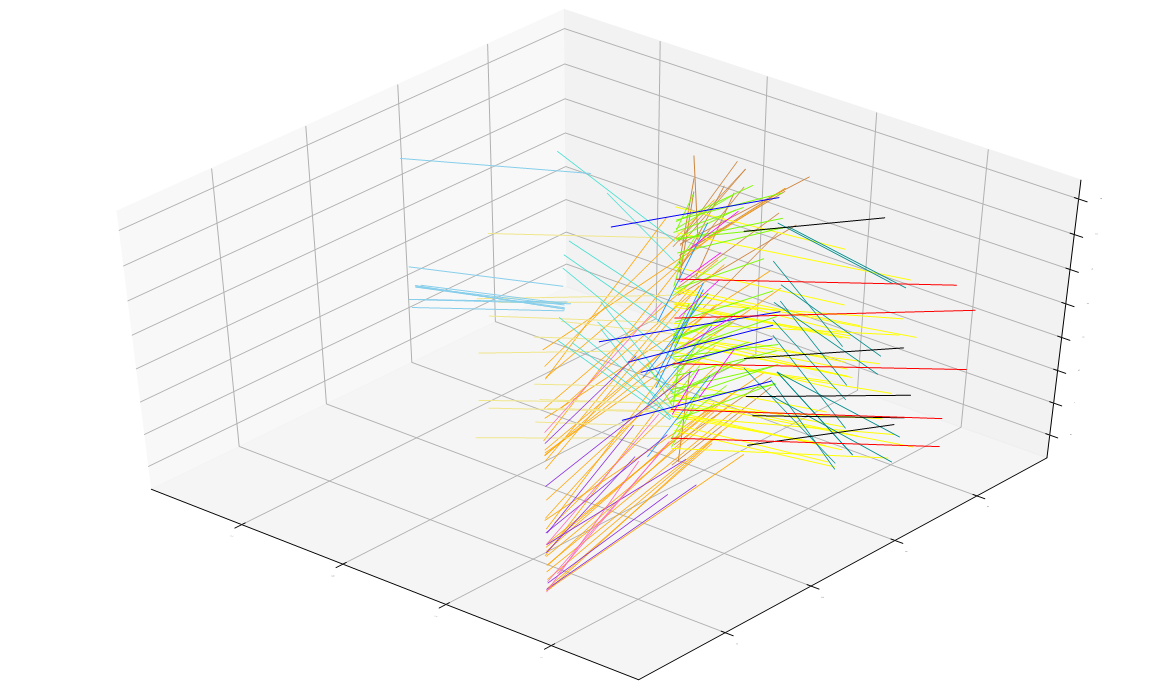


通过遍历OD数据，统计所有出行中出现的站点，记站点数为M，建立一个MxM的矩阵，其中行表示刷卡站点，列表示预测的下车站点。得到的矩阵的一部分如图所示：



### 3.3.2 聚类方法实现

从OD矩阵中提取刷卡数据，遍历地将他们通过站点坐标投影到平面上，把时间投影到纵轴上，可以为公交客流绘制一个三维的连线图，进行聚类后得到的结果如图所示，从而可以更直观地找到有密集连通性的公交站点，并可以发现他们之间客流产生的高峰期。



## 3.4 公交乘客常驻地识别

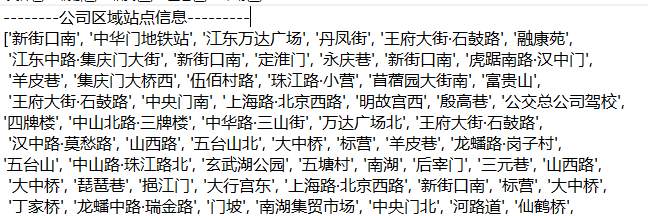
### 3.4.1 常驻地站点信息实现

进行数据处理之后将聚类数据与原数据合并，根据聚类的标签获取对应的站点信息作为常驻地。

刷卡时间满足早高峰与晚高峰的出行记录作为上班族进行处理，所得住宅区域与公司区域站点信息如下：



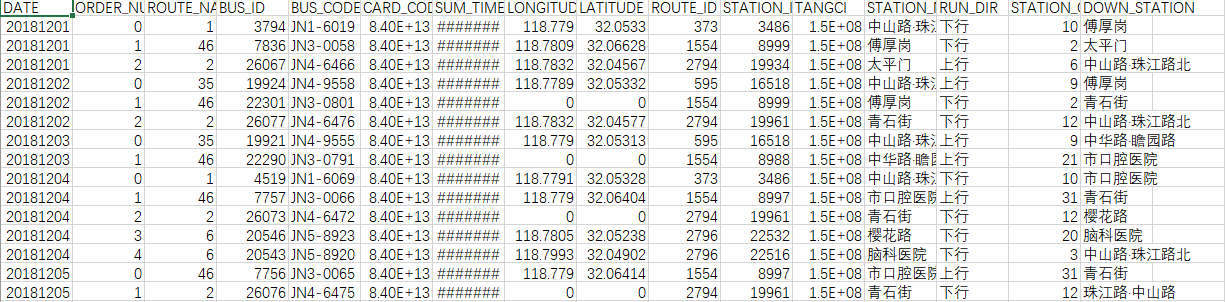
公司区域的站点信息由刷卡数据中的返程站点信息获得：



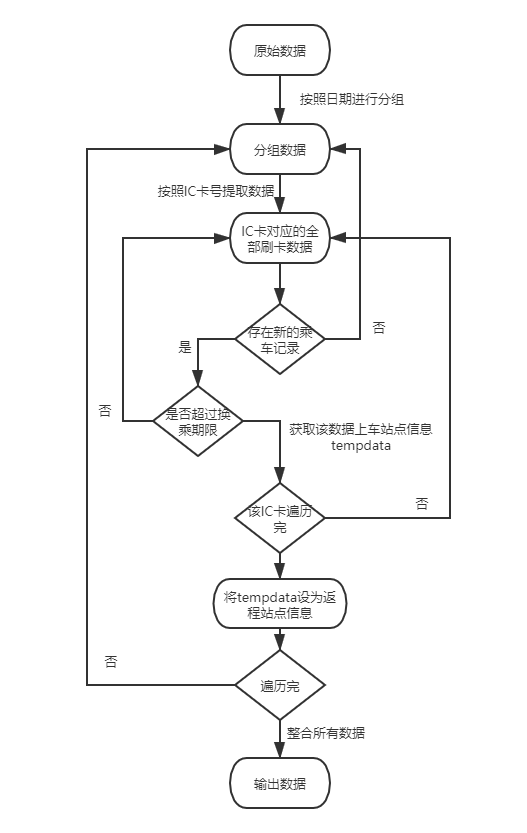
## 3.5 公交乘客出行特性分析

### 3.5.1 数据预处理实现

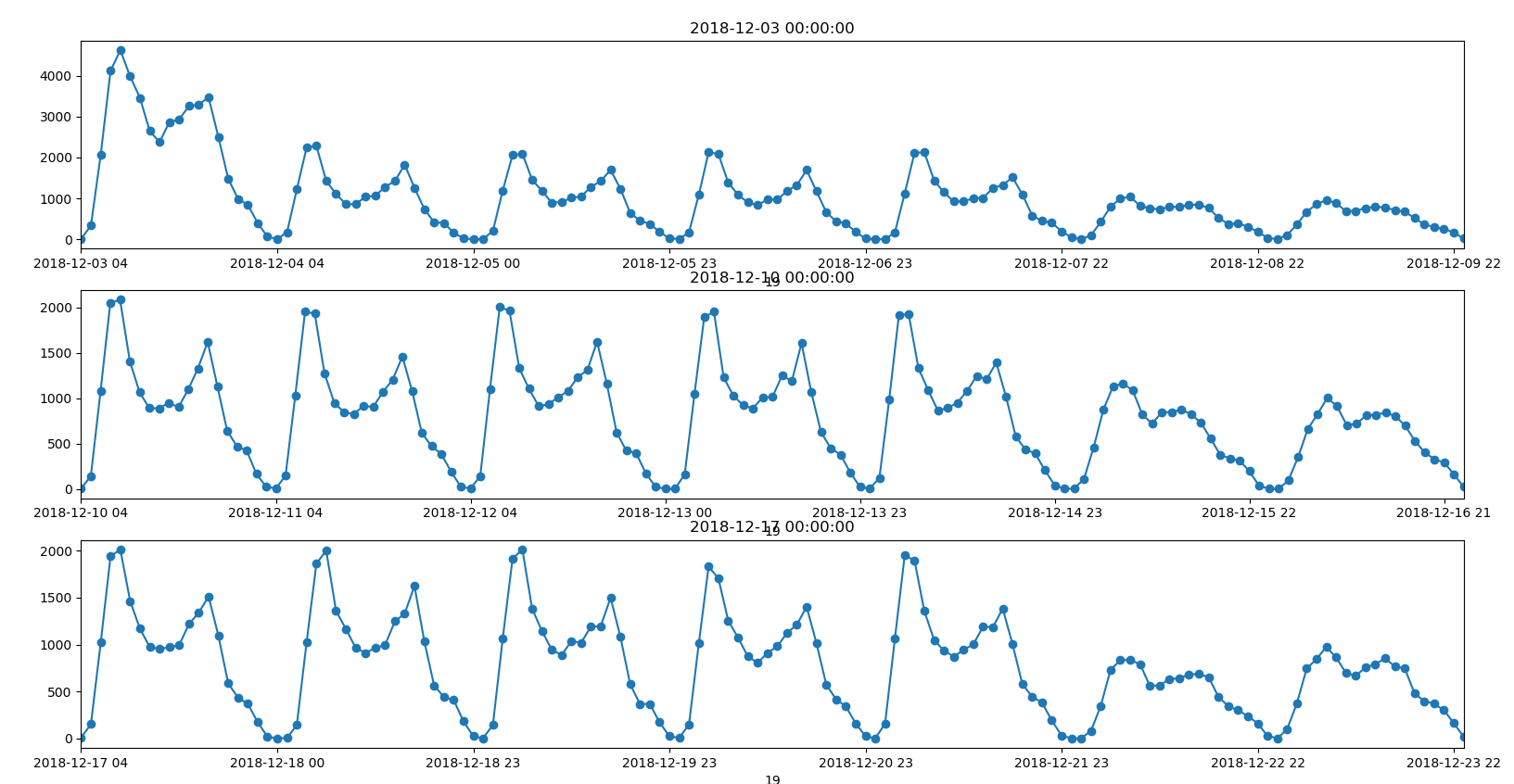
2018年12月的IC卡刷卡数据共有两百多万条，从中选取了50万条数数据用于模型的设计。经过之前环节的处理，现有的数据中已经包含日期、当前的刷卡次序、乘坐的线路名称、乘坐车辆的ID和编码、该乘客的IC卡ID、去程刷卡时间、刷卡位置的经纬度、线路ID、上车站点ID和名称、趟次信息、站台顺序、运行方向，下车站点名称。具体数据样式如下图：



针对该数据集进行如下流程的数据处理：

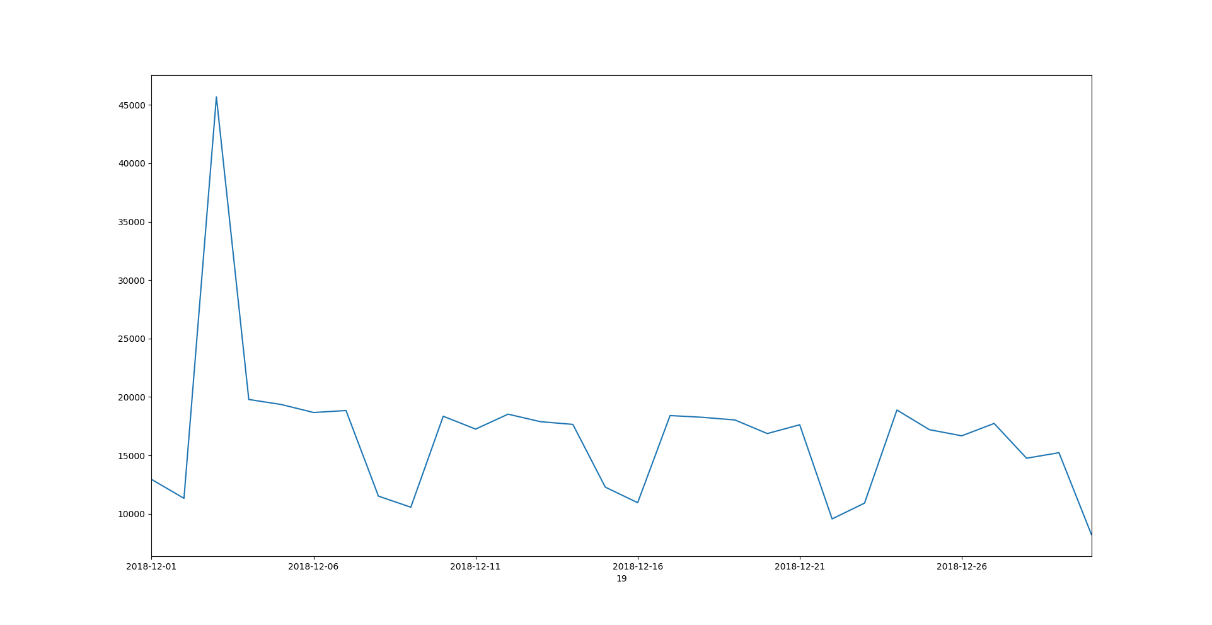


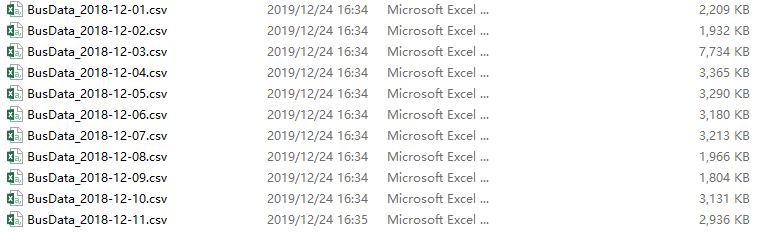
1）为分析乘客的出行模式，先将所有数据按照精确到时刻的时间进行分组，并统计当天的全部乘车次数，将此信息使用图表展示：



由图表可见，2018年的12月份涵盖了三个完整的星期。刷卡次数与时刻有着明显的关系，存在明显的早高峰与晚高峰，工作日与周末之间的区分也很明显。可见可从时间入手对乘客的行为模式进行分析。

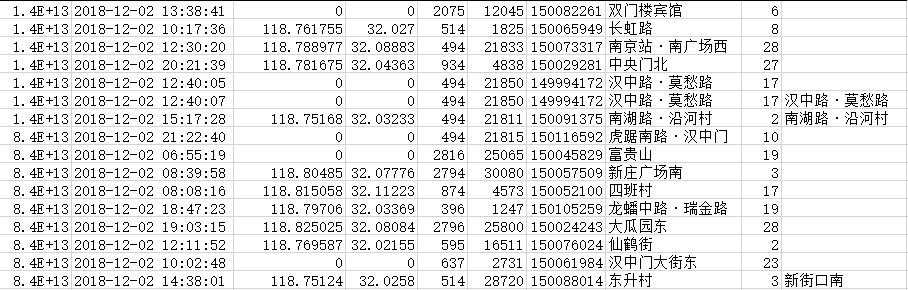
2）样本数据集涵盖了当月每天的刷卡次数，为方便进行处理，将数据按照日期划分，分成30张表，每张表包含当天的所有刷卡次数。同时，将样本按照刷卡日期和IC卡号进行升序排序。样本数据中刷卡次数与日期的对应关系如下图所示：





3）同一IC卡在一天内可能存在多次刷卡记录。乘客为到达某一目的地会在短时间内进行多次的转乘，换乘的数据增加了数据量并对乘客出行模式的判断有一定的影响。因此需整合所有的转乘信息，将多次乘车记录统计为一条涵盖出发点与目的地的记录，将中间的换乘记录进行略去。

本次操作以12月2日的刷卡数据为例，对其进行关于转乘数据的处理。处理前的数据如下：



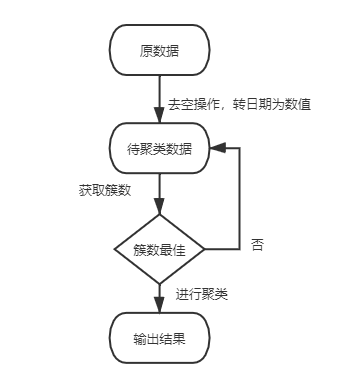
设定50分钟为换乘时间，超过此限制的下一条刷卡记录作为一条新的公交路线予以记录，并将此条信息的刷卡地点和刷卡时间作为该IC卡对应刷卡记录的返程的刷点地点和返程刷卡时间。换乘时间小于此限制的记录则进行删除。如果不存在大于此限制的新刷卡记录则将该IC卡的返程刷卡地点和返程刷卡时间置空。

表中存在返程刷卡信息的数据鉴于其刷卡次数不存在规律，视为异常数据进行删除。最终数据如下：



### 3.5.2 聚类处理实现

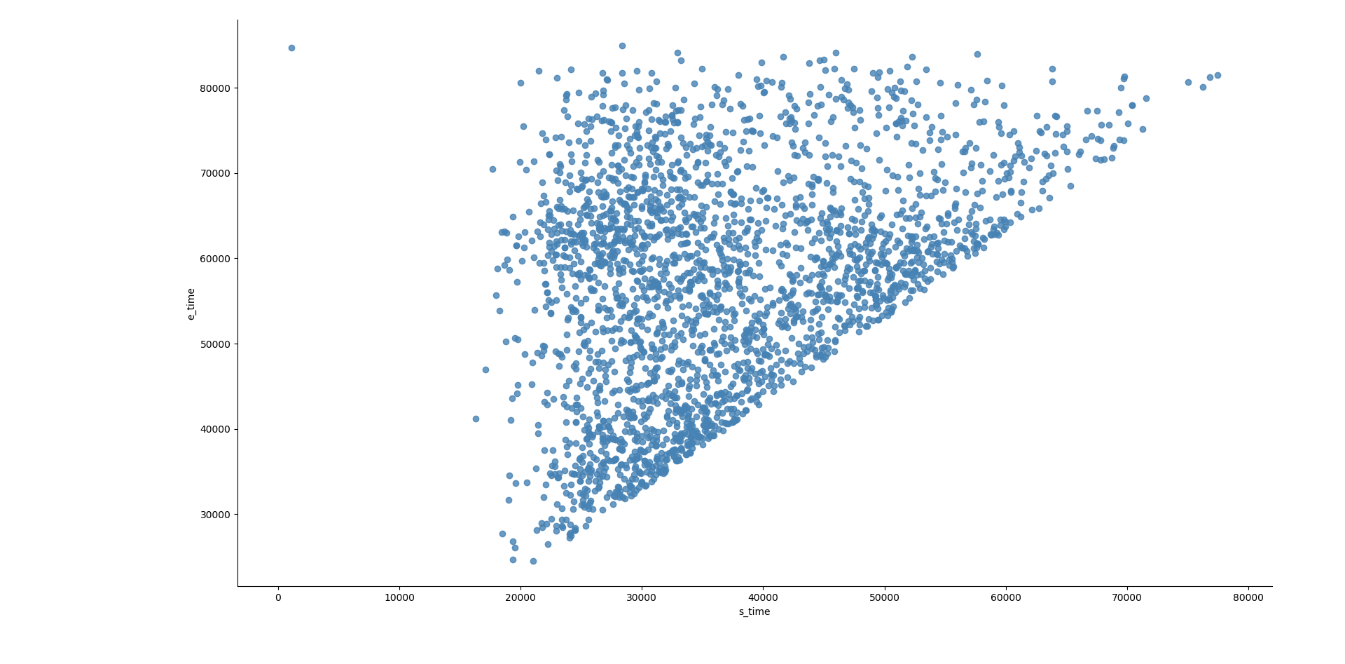
此时数据表中包含日期、当前的刷卡次序、乘坐的线路名称、乘坐车辆的ID和编码、该乘客的IC卡ID、去程刷卡时间、刷卡位置的经纬度、线路ID、去程上车站点ID和名称、趟次信息、站台顺序、运行方向，下车站点，返程刷卡时间，返程刷卡站点名称，返程刷卡站点编号。根据之前的思路，由每个IC卡号的启程刷卡时间和返程刷卡时间进行聚类分析。其流程图如下：



首先进行数据的处理，将IC卡编号提出作为每条的数据的新索引。并删除诸多对出行模式不造成影响的信息，比如：乘坐车辆的ID和编码，线路ID，趟次信息，站台顺序，运行方向。

之后为满足聚类的算法中各参数为数值类型的要求，需再讲去程刷卡时间和返程刷卡时间转换为便于计算的数值。

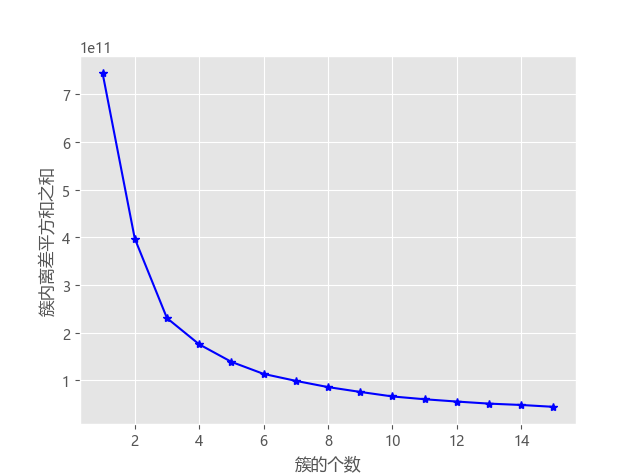
由于采用的K-means聚类算法需要现指定聚类的簇数，为此先根据去程刷卡时间和返程刷卡时间绘制散点图，从而确定大致的簇类划分。



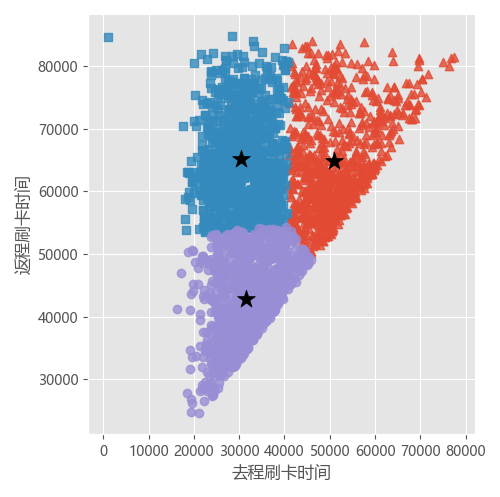
散点图中存在一定的噪声，只是因为存在诸多出行不规律的刷卡记录。

由图中的数据分布大致可推出有3个簇。其中图中左上方的刷卡记录满足去程刷卡时间早，返程刷卡时间晚，且数据密集，可推测此为满足上班族的出行模式类。

在通过计算簇内离差平方和确定簇类的最佳数量为3。



最终的聚类效果如下：



蓝色部分乘客的乘车时间满足早高峰与晚高峰的时间划分，可以推断为上