载客事件在热点区域的分布回归分析

杨文静，20131208，ZY1321222

**摘要：**北京市出租车在空载时会趋向附近的热点区域，本文分析出租车载客事件在热点区域的分布，由此估算载客事件发生在各个热点区域的概率。首先，利用XMeans聚类方法识别载客的热点区域，然后分析在各个聚类中事件分布规律，采用非线性回归的方式建模，并进行显著性检验。最终得出合理的回归曲线。

**关键字：**非线性回归分析, 载客事件, SPSS

# 引言

现有的车载网络通信中，路由协议往往侧重于个别指标，缺乏对多项指标的综合考虑，往往在个别指标上性能优越，但无法优化多项指标，网络整体性能难以获得较大的提升。现有的车载网络中，考虑车辆的行为特征有助于路由协议的设计，提高路由效率。以及开发出预测更为准确的应用（如打车软件）有很大帮助。

城市出租车在空载时，会趋向周围的热点区域（如商业中心，火车站，飞机场等）。我们旨在获得出租车在各个热点区域的分布概率。通过分析北京市2012年11月的出租车数据，提起其中的载客事件信息。并采用X-Means方法获取其中的热点区域。基于以上热点区域，我们对载客事件在热点区域的分布建模，并进行回归分析。

本文主要分为以下X个部分，第二章将介绍本文所采用的数据集，第三章将阐述本文所用的聚类方法和聚类产生的热点的结果。回归分析及其结果将在第四章说明。最后，第五章讨论回归分析结果。

# 数据说明

北京市出租车移动轨迹数据集来源于北京智能交通系统关键技术研究与应用示范项目，该项目以支持智能交通系统工程建设、解决关键技术难题、提升交通科技发展水平和自主创新能力、为实现新北京交通体系和奥运会的顺利召开提供支持与保障为目标。其核心研发内容之一，是实时采集、存储、处理多源异构海量交通数据、形成动态交通信息以及决策支持的分布式处理系统。该项目涉及出租车为12,096辆，约占北京市出租车总数的18%，占北京市机动车总数的0.2%，对五环内（含五环）次干路以上路网的覆盖率达到90%以上。通过这些出租车上安装的GPS定位装置，每隔60s上传一次自己的经纬度位置、速度、方向信息到数据中心。每天产生的数据量约1300万条。

本文中用于分析和实验的数据是该系统在2012年11月5日早上8点到9点的数据，车辆运动范围主要集中在北京市区，面积为 25,000m\*25,000m，共计9302条实际移动轨迹记录。数据格式如下表：

表 1 北京市出租车轨迹数据集格式

|  |  |
| --- | --- |
| 列名 | 说明 |
| 车辆ID | 使用11个ASCII字符； |
| 时间标签 | 使用GMT时间格式共14个ASCII字符，格式为（YYYYMMDDHHMMSS）； |
| 84坐标系经度 | 最长为11个ASCII字符，变长； |
| 84坐标系纬度 | 最长为10个ASCII字符，变长； |
| 速度 | 单位为公里/小时，最长为3个ASCII字符，变长； |
| 事件（event） | 为1个ASCII字符，数值型字符，包括上下车、开锁车门等。 |

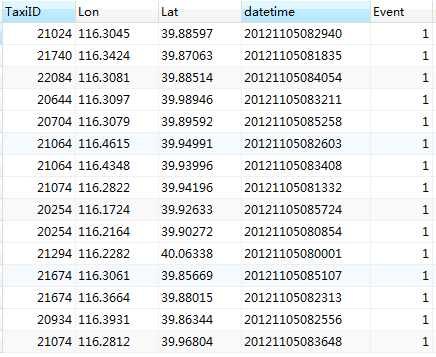


图 1 20121105早上8点到9点载客事件数据

# 载客事件聚类

我们采用Waikato Environment for Knowledge Analysis(Weka)[1]工具中实现的聚类算法X-Means对载客事件聚类。Weka是基于JAVA环境下开源的机器学习以及数据挖掘软件。X-Means[2]是经典的分类算法K-Means[3]的一种改进算法。K-Means算法的基本思想是：以空间中k个点为中心进行聚类，对最靠近他们的对象归类。通过迭代的方法，逐次更新各聚类中心的值，直至得到最好的聚类结果。X-Means解决了K-Means首初始值相关的问题,并提高了计算效率。X-Means给定分簇的范围[ minNumClusters, maxNumClusters ]，然后先执行一次K-Means过程，K值设为minNumClusters。分簇完成后，看minNumClusters个簇是否可以划分为两簇。重新设置K值，迭代执行K-Means过程。

Weka参数的参数设置如图2所示。我们采用数据中的经纬度信息，采用欧几里得距离，X-Means给定的范围为[200,250].

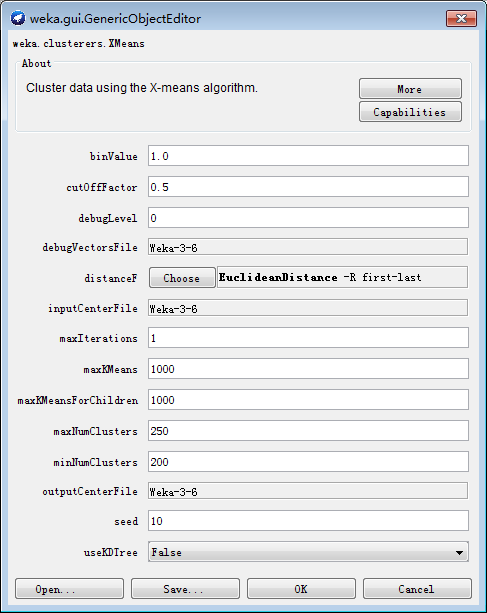


图 2 X-Means参数设置

输出结果如图3所示，并给出了每个簇的中心点以及每个簇中元素的个数，其分簇效果如图4所示。共被分为200簇。

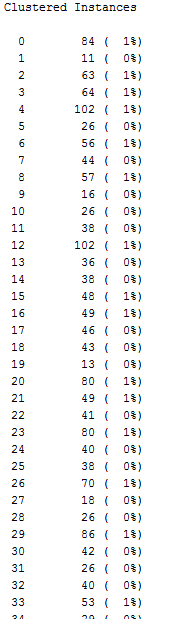


图 3 每个簇的事件数

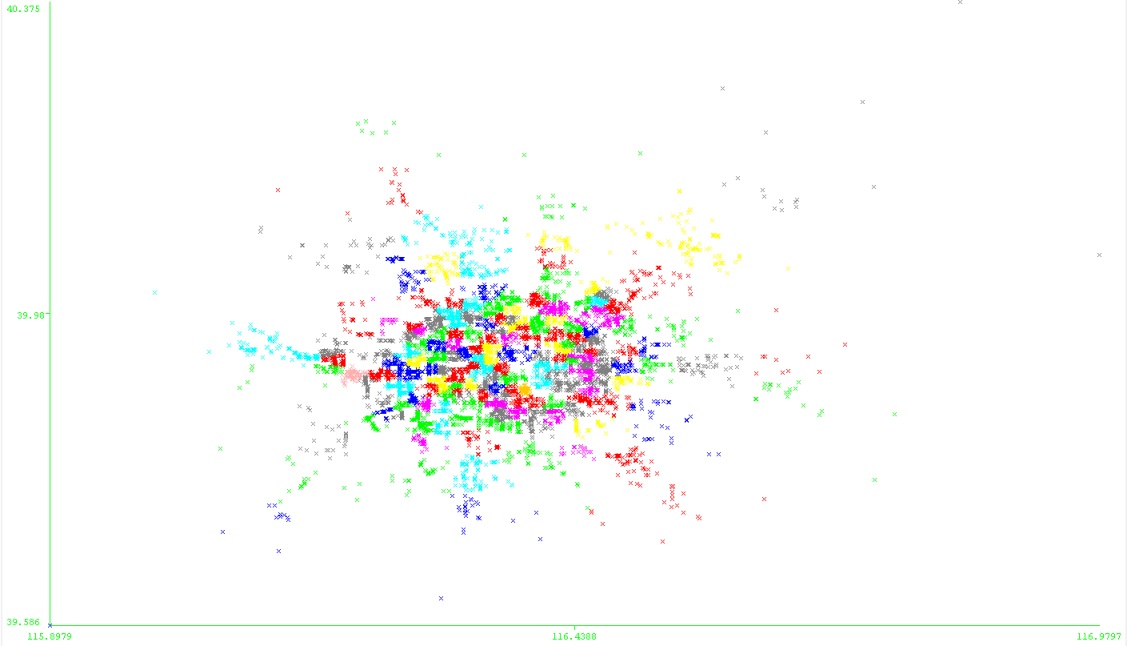


图 4 XMeans 分簇效果

# 解决问题的方法和计算结果

## 数据初步分析

对各个簇按载客事件排序，并得到载客事件在各个区域的分布图5 如下：

图 5 载客事件分布图

对横坐标中的摘客事件去对数后，所得的结果如图6所示。其中横坐标为事件，纵坐标为对事件排名取对数。



图 6 转化后的载客事件分布图

将数据导入SPSS中，如图7 所示。

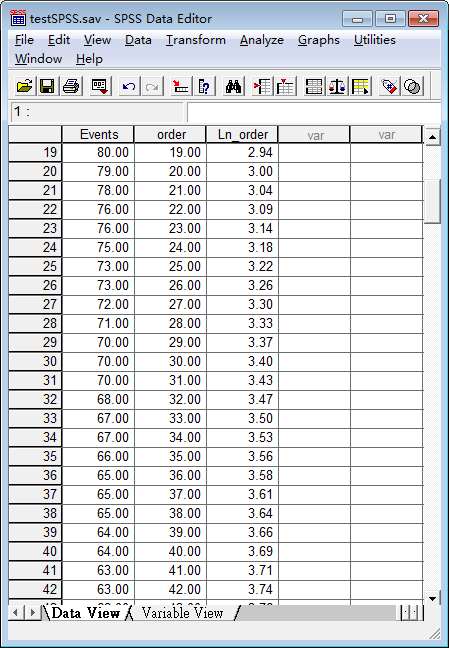


图 7 回归分析数据示意

## 回归模型的建立和分析

以上数据分布形式类似于抛物线，对数据采用Cubic模型建模：

其中为常数项，为待估参数，x为事件数目，y为排名的取log值。z为簇的排名序号。

由此看见如表XXX,回归的参数估计为:

其中，R2=0.995,R2为0-1的值，当其越接近1时，理论值与实际值越接近。结果表明其回归程度好。

表 3 模型摘要和参数估计

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Equation | Model Summary | | | | | Parameter Estimates | | | |
| R Square | F | df1 | df2 | Sig. | Constant | b1 | b2 | b3 |
| Cubic | .995 | 13862.868 | 3 | 192 | .000 | 5.181 | .018 | -.001 | 3.64E-006 |

表 2 变量处理摘要

|  |  |  |  |
| --- | --- | --- | --- |
|  | | Variables | |
|  | | 因变量 | 自变量 |
|  | | Ln\_order | Events |
| Number of Positive Values | | 198 | 196 |
| Number of Zeros | | 1 | 0 |
| Number of Negative Values | | 0 | 0 |
| Number of Missing Values | User-Missing | 0 | 0 |
|  | System-Missing | 0 | 3 |

表 4 情况处理摘要

|  |  |
| --- | --- |
|  | N |
| Total Cases | 199 |
| Excluded Cases(a) | 3 |
| Forecasted Cases | 0 |
| Newly Created Cases | 0 |

具有缺失值的数据将被排除，由表4 可知，共有199组数据，被排除的数据为3.

其回归结果如图8所示。



图 8 Cubic回归结果

# 讨论

本文通过对聚类后的簇内事件数据排序，建立Cubic回归模型，从而进行回归分析。由回归结果可知，所建立模型

与实际数据的回归特性明显，R2达到0.995. 且排除值少，仅为3. 变量过程中系统的缺失值仅为3.

那么，

由此可知，载客事件热点区域的排名与区域中事件数得关系。同理，已知z，即可解出载客事件在某一热点区域的发生数目。在某区域载客事件发生概率的计算在现有的车载网络中，考虑车辆的行为特征有助于路由协议的设计，提高路由效率。以及开发出预测更为准确的应用（如打车软件）有很大帮助。

此模型与实际值符合结果较好，但是也存在一些缺陷。

1. 模型不够直观，我们最终需要由区域热度（排名）获得载客事件在此区域的发生数目。现在则相反。
2. 参数过多，现有模型中具有一个常数项和三个参数，使得此模型的适用性受限。即当数据变化时，参数变化的规律不容易发现。

**参考文献**

[1]Mark Hall,Eibe Frank,Geoffrey Holmeset al. The WEKA data mining software: an update[J]. ACM SIGKDD Explorations Newsletter. 2009. 11（1）: 10-18

[2]Dan Pelleg,Andrew W. Moore. X-means: Extending K-means with Efficient Estimation of the Number of Clusters.[M]，2000: 727-734

[3]James MacQueen. Some methods for classification and analysis of multivariate observations[M]California, USA，1967: 14