文章编号:1005-0523(2019)02-0077-06

基于AFC数据挖掘的轨道交通站点分类研究

邓评心[1],郑长江[1],马庚华[2],李 锐[1]

(河海大学 1.土木与交通学院;2.港口海岸与近海工程学院,江苏 南京 210098)

摘要:轨道交通站点的分类对于研究不同类别站点的客流规律、周边土地利用情况以及发展趋势都有着重要影响。基于AFC数据,综合多种有效性指标确定分类数,采用主成分分析、k-means聚类、多元线性回归等方法,定性分析与定量分析相结合对站点进行类别划分。将苏州轨道交通1、2号线共58个站点分为4类,为站点分类别后续研究以及轨道交通发展研究奠定基础。

关键词:AFC数据;站点功能定位;k-means;回归分析

中图分类号:U231 **文献标志码**:A

城市轨道交通站点是整个轨道交通线网中的关键节点,站点交通的可达性使其成为城市各种社会经济活动中的集聚场所[1]。轨道交通站点的研究是轨道交通研究的一个重要方面,站点的功能定位、站内的客流特征等方面都会影响到站点的运营,进一步影响到周边用地的开发甚至整个城市的发展。

城市发展及轨道交通运营过程中产生了大量数据,数据存储和数据挖掘技术的不断发展为轨道交通站点的研究提供了有效手段。Robert Cervero和Jin Murakami Pattnaik利用香港地铁数据,通过聚类分析将站点分为5类,并给出开发强度及混合度的平均值[2-4]。国内对于站点分类的研究主要使用站点自身特性以及周边环境特征做定性分析[5-7],变量复杂冗余,且没有考虑到客流波动的影响。随着AFC系统的广泛使用,产生了大量的AFC数据,这些数据可以精确记录乘客的运输活动[8-9],并且反映站点的特征。

本文使用AFC数据进行挖掘,通过主成分提取和聚类分析的方法对轨道交通站点进行分类。然后,采用回归分析的方法定量分析站点类别和周围环境(如用地属性)之间的关系。

1 数据与对象

1.1 研究对象

苏州市位于江苏省东南部,是长三角城市群重要的中心城市之一。截至2016年底,苏州市共有2条正式运营的轨道交通线路,车站共58个,线路总长68.2 km。轨道交通1号线于2012年4月28日开通运营,轨道交通2号线于2013年12年28日开通运营,两条线路均使用AFC系统,可以收集并存储准确的进出站数据。

1.2 研究数据

为了确定苏州轨道交通1、2号线各站点的类别,采用了2016年12月5日至11日共计1周的AFC数据。数据包括每个站点的站点编号,站点名称及其每个小时进站和出站的人数,共计58个站点;其中换乘枢纽广济南路站作为一个站点,时间从6时至23时共计18 h。由于城市居民的移动模式大体上每周重复1次,

收稿日期:2018-10-21

基金项目:国家自然科学基金项目(51508161)

作者简介:邓评心(1995—),男,硕士研究生,研究方向为交通规划运输与管理。

通讯作者:郑长江(1966—),男,教授,研究方向为交通规划运输与管理。

因此采用1周的数据进行站点分类更加的科学合理。

为了分析站点类别与周边环境之间的关系，使用了2016年苏州轨道交通站点周边地区的土地利用现状数据。数据包括以站点为圆心，1 km为半径的范围内各种属性用地的面积。数据均为2016年结合网络地图与实地考察实时采集所得，与AFC数据在时间上相匹配，使得分析结果更为准确有效。

2 方法

2.1 数据准备

将AFC数据转换为一个二维矩阵的形式以供下一步处理。每个站点的数据为一行，进站客流与出站客流分开，前半部分为进站客流，后半部分为出站客流。进站和出站的客流量按小时进行划分，每天的数据作为一个整体，并以时间顺序从周一至周日排列。按照上述方法进行处理之后，每个站点有252个(进站/出站×18 h×7 d)变量用于聚类，数据的布局如图1所示。

由于不同轨道交通车站的每日客流量差别很大，因此需要在聚类之前对变量进行标准化处理。本文采用小时比率替换每小时的客流量来进行标准化，其中，小时比率为每个车站每小时进站(或出站)的客流量与此车站当天进站(或出站)的总客流量之比。

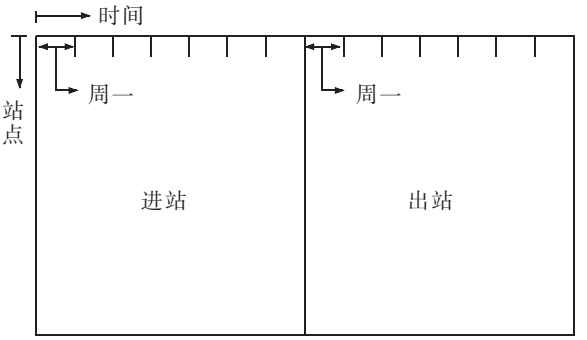


图1 AFC数据的二维矩阵形式

Fig.1 2-D matrix of the AFC data

共有252个变量可用于站点分类，但是车站客流的变化模式是重复的，因此需要减少变量数目并找出关键变量。主成分分析法(PCA)可以从多个具有相互关系的变量中提取出几个两两之间相互独立的新变量，并将原来变量所包含的信息尽可能多的反映出来[10]。可以减少变量数目，降低维数与复杂性，突出变量的特点。

2.2 聚类方法

聚类方法有很多种，应当根据数据情况灵活采用。当数据中存在离群点时，k-medoids方法更适用，因为中心点不像均值那样容易受离群点影响。然而，k-medoids算法的每次迭代的复杂度是$O(n(n-k))$，当n和k的值较大时，这种计算开支远高于k-means方法。由于数据中不存在明显的离群点，因此k-means方法更优。k-means方法易于理解，计算快速，对大数据集的效率较高，本文数据集采用k-means方法可以快速得出结果，比较适用。

2.3 聚类数的确定

k-means算法简单高效，但是难点在于聚类数k的确定。现已经提出了很多聚类有效性的判断指标，综合使用这些指标可以找出合适的k值，从而产生有效的聚类结果。3种常用的聚类有效性指标如表1所示。

表1 三种常用聚类有效性指标

Tab.1 Three cluster validity indices

指标	简称	定义
组内误差平方和	SSE	$\sum_{i=1}^{k}\sum_{x\in C_i}^{k} d^2(c_i,x)$
Dunn指数	D	$\min_i\left\{\min_{j,i\neq j}\frac{\min\limits_{x\in C_i,y\in C_j}[d(x,y)]}{\max\limits_{k\in C}[\max\limits_{x_i,y\in C_k} d(x,y)]}\right\}$
Xie-Beni指数	XB	$\frac{\sum_i\sum_{x\in C_i} d^2(x,c_i)}{n\cdot\min\limits_{j,i\neq j}[d^2(c_i,c_j)]}$

注：n为数据集中对象个数；C_i为第i个类别；c_i为C_i的聚类中心；$d(x,y)$为x和y之间的欧式距离。

SSE 是组内误差平方和，代表了所有点到相应簇中心的平方距离之和，SSE 随着 k 值的增大而减小，SSE 减幅最小时的 k 值即为较合适的聚类数。Dunn 指数是类间任意两个簇元素的最短距离与类内任意簇中的最大距离的比值[11]，D 越大，意味着类间差别大且类内差别小，此时聚类结果更有效。Xie-Beni 指数是每个点与其聚类中心之间的均方距离与聚类中心之间最小平方距离的比值[12]，当 XB 取值最小时，聚类数 k 最佳。为确定最佳聚类数 k，本文综合考虑了以上 3 个指标。

2.4　回归分析

回归分析是一种研究因变量和自变量之间关系的方法。在本研究中，因变量为轨道交通站点类别，自变量为站点周围 1 km 范围内各种用地属性的占比，用地属性包括行政、教科、医疗、商业、枢纽、设施、居住、景区等 8 种。对于每种类别的站点建立一个线性回归模型，如果一个站点包含在该类中，则响应变量为 1，否则为 0。通过回归分析可以从定量方面判断出何种用地属性对站点类别产生显著影响，从而验证分类的科学性与合理性。回归模型如公式 1 所示

$$y=\varepsilon+\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p \tag{1}$$

其中：x 为自变量；y 为因变量；$\beta_0,\beta_1,\cdots,\beta_p$ 为回归系数；ε 为随机误差项。

3　结果与分析

3.1　聚类结果

采用 k-means 方法对数据进行聚类，k 值从 2 取到 7，计算出不同 k 值情况下的 SSE，D，XB 3 项指标，然后通过最大最小归一化对指标进行标准化处理，得出各项指标的变化曲线，如图 2 所示。

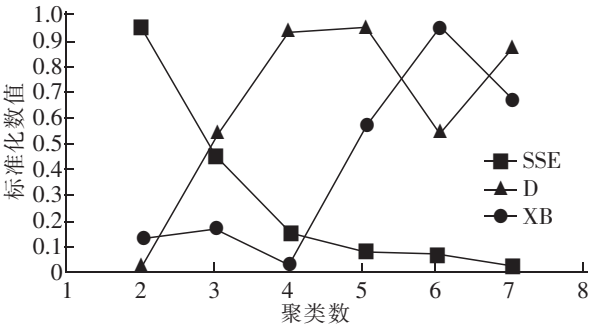


图 2　不同聚类数下的有效性指标曲线

Fig.2　Validity indices curve with the number of clusters

通过上图可以看出，SSE 曲线关键在寻找“肘点”，其在 4 类之后下降速度明显变慢；D 值越大表示聚类效果越好，D 曲线在聚类数为 4，5 和 7 时具有较大值；XB 值越小表示聚类效果越好，XB 曲线在聚类数为 4 时出现最小值。综合考虑这 3 个指标，确定聚类数 k 值为 4，即将 58 个站点分为 4 类，聚类结果如表 2 所示。类别 1 共有 15 个站点，包括汾湖路、滨河路等。类别 2 共有 9 个站点，包括乐桥、临顿路等。类别 3 共有 32 个站点，包括木渎、金枫路等。类别 4 包括苏州火车站、高铁苏州北站共 2 个站点。

表 2　站点聚类结果

Tab.2　Results of station classification

类别	代表站点	站点数/个
类别 1	汾湖路、滨河路、广济南路、养育巷、中央公园、星海广场、时代广场、星湖街、南施街、钟南街、大湾、平河路、月亮湾、松涛街、桑田岛。	15
类别 2	苏州乐园、乐桥、临顿路、相门、东方之门、山塘街、石路、三香广场、郭巷。	9
类别 3	木渎、金枫路、玉山路、塔园路、西环路、桐泾北路、东环路、文化博览中心、星塘街、骑河、富翔路、富元路、蠡口、徐图港、阳澄湖中路、陆慕、平泷路东、劳动路、胥江路、桐泾公园、友联、盘蠡路、新家桥、石湖东路、宝带桥南、尹中路、郭苑路、尹山湖、独墅湖南、独墅湖邻里中心、金谷路、金尚路。	32
类别 4	苏州火车站、高铁苏州北站。	2

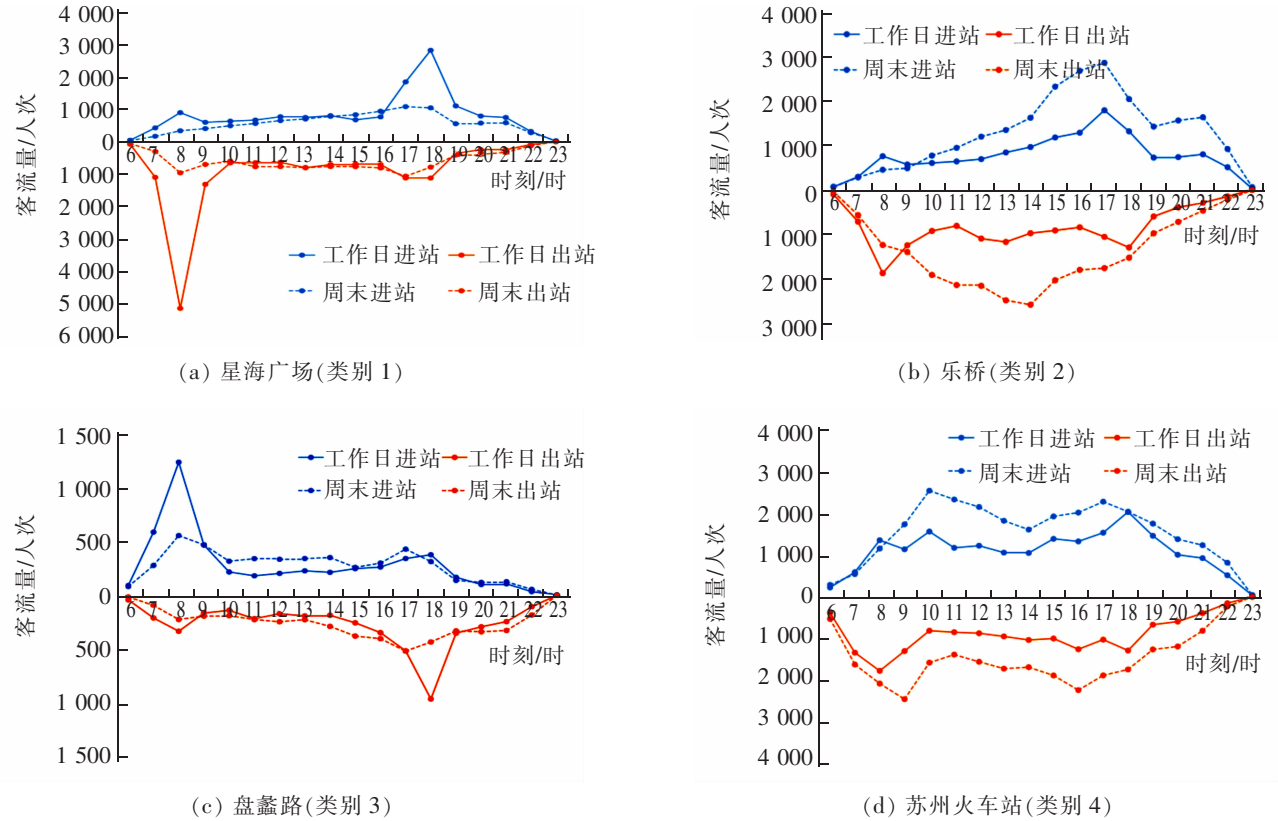
3.2 站点时空特征

从站点的空间分布来看,类别1的站点主要位于老城区、金鸡湖周边以及城市外围工业区,这些地区行政办公场所或工厂密集,工作岗位较多。类别2的站点主要位于老城区之内以及城市其他大型娱乐场所周边,商业发达,是人们主要的购物娱乐区域。类别3的站点主要位于老城区外围,以居住用地为主。类别4的站点位于苏州市两个综合交通枢纽——苏州火车站和高铁站附近。

在时间上,不同类别的站点客流变化的差异性较大,同一类别站点客流变化规律大致相同,本文选取各个类别中具有代表性的站点分析其在工作日和周末的客流变化规律。类别1中的代表站点为星海广场,客流变化如图3(a)所示。在工作日,早高峰以出站客流为主,晚高峰以进站客流为主,其他时间客流量较少且比较均衡。这表明许多乘客早上在此处下车,傍晚在此处上车,一整天都在站点附近工作。在周末,全天的客流波动不大,没有出现明显的高峰,客流量与工作日平峰时段的客流量大致相同。

类别2中的代表站点为乐桥,客流变化如图3(b)所示。在工作日,早高峰出站客流量较大,晚高峰出站客流量较大,但与类别1相比,高峰不像类别1那么突出。并且进站客流在19时至21时期间持续具有较高的客流量,甚至在21时前后形成一个新的高峰,这表明即使在晚上,许多人仍然停留在类别2站点附近的区域。在周末,几乎各个时段的客流量都大于工作日,且进站客流在17时和21时出现两个高峰,这意味着乘客下午或晚上购物娱乐结束后回家。从这种客流变化模式可以推断出类别2站点周围是商业区。

类别3中的代表站点为盘蠡路,客流变化如图3(c)所示。与其他类别相比,类别3站点的客流量比较少。在工作日,站点客流变化规律与类别1中的站点恰好相反,早高峰以进站客流为主,晚高峰以出站客流为主,其他时间客流量较少且比较均衡。这表明乘客上午从此处上车前往工作地点,下午从此处下车返回家中。在周末,站点客流变化规律与类别1中的站点大致相同,客流量与工作日平峰时段相当。从这种客流变化模式可以推断出类别3站点周围是居住区。



(a) 星海广场(类别1)　(b) 乐桥(类别2)

(c) 盘蠡路(类别3)　(d) 苏州火车站(类别4)

图3 站点客流变化图

Fig.3 Diurnal pattern of subway stations

类别 4 中的站点较少，仅有苏州火车站和高铁苏州北站 2 个，苏州火车站的客流变化如图 3(d)所示。工作日和周末的客流变化规律基本一致，客流分布比较平均，全日的客流量都比较大，没有明显的客流高峰和低谷，且周末的客流量要大于工作日的客流量。此类站点客流量与周围交通枢纽密切相关，会受到枢纽规模、车辆运营组织的影响。

从空间和时间的规律可以推断出各类别站点的特征：类别 1 位于行政办公区域以及行政办公主导的具备一定商业功能的区域；类别 2 位于具有购物娱乐功能的商业区域；类别 3 位于居住区域以及居住主导的混合区域；类别 4 位于大型交通枢纽区域。

3.3　回归结果

分类别回归分析得出非标准化系数 *B*、标准化系数 *Beta*、显著性 *Sig.*等指标，结果如表 3 所示。

在类别 1 中，行政与科教用地呈正相关且显著，居住用地呈负相关，这说明此类站点周边的工作区域较多而居住区域较少，行政办公功能处于主导地位。在类别 2 中，商业用地呈正相关且显著，景区呈负相关，这说明此类站点周边多为商业区域，人们到此处购物娱乐而不是前往景区游览。在类别 3 中，居住用地影响极为显著且是正相关关系，其他部分属性如行政、设施也有较为显著的影响，这说明此类站点周边为居住区域或以居住区为主导。在类别 4 中，仅有枢纽用地影响极为显著且是正相关，其他属性影响不显著，这说明站点主要受到周边大型交通枢纽的影响。

通过回归分析得出的结论与通过站点空间分布和客流变化规律所得出的结论相符合，这不仅揭示了各类别站点的特征，而且验证了 k-means 聚类所得出的结果，说明本文站点分类方法具有可操作性和科学合理性。

表 3　分类别回归结果

Tab.3　Results of different clusters by regression

类别	指标	行政	教科	医疗	商业	枢纽	设施	居住	景区
类别 1	*B*	0.479	0.999	4.061	0.978	−0.014	−7.763	−0.063	−0.083
	Beta	0.285	0.333	0.154	0.391	−0.001	−0.285	−0.067	−0.008
	Sig.	0.031	0.007	0.247	0.067	0.993	0.049	0.725	0.953
类别 2	*B*	−0.084	−0.031	1.884	1.027	−0.941	−0.425	0.002	−0.384
	Beta	−0.064	−0.013	0.092	0.530	−0.089	−0.020	0.002	−0.047
	Sig.	0.657	0.919	0.534	0.028	0.483	0.899	0.991	0.753
类别 3	*B*	0.625	0.021	−5.265	−0.906	−2.678	9.545	1.032	1.380
	Beta	0.254	0.005	−0.137	−0.248	−0.134	0.240	0.750	0.089
	Sig.	0.007	0.955	0.146	0.098	0.096	0.019	0.000	0.342
类别 4	*B*	−0.020	0.011	0.319	−0.099	4.633	−0.357	0.029	0.087
	Beta	−0.033	0.010	0.033	−0.109	0.926	−0.036	0.085	0.022
	Sig.	0.629	0.866	0.635	0.329	0.000	0.632	0.396	0.748

4　结束语

本文基于苏州轨道交通 1、2 号线 AFC 数据，使用 k-means 聚类并结合回归分析验证，定性分析与定量分析相结合，将 58 个站点分为 4 类。不仅提供了一种易于操作、科学合理的站点分类方法，可以推广应用到其他城市，而且可以为后续研究，如客流预测、新开通线路影响评价等方向打下基础，在土地开发和可持续规划等方面发挥重要作用。

参考文献：

[1] 傅搏峰，吴娇蓉，陈小鸿. 郊区轨道站点分类方法研究[J]. 铁道学报，2008，30(6)：19-23.

[2] CERVERO R，DUNCAN M. Residential self selection and rail commuting：A nested logit analysis[C]//Cali fornia：University of California Transportation Center Working Papers，2002.

[3] CERVERO R，JIN M. Rail+property development：A model of sustainable transit finance and urbanism[C]//California：University of California Transportation Center Working Papers，2008.

[4] KUBY M，BARRANDA A，UPCHURCH C. Factors influencing light-rail station boardings in the united states[J]. Transportation Research Part A，2004，38(3)：223-247.

[5] 李向楠. 城市轨道交通站点分类的聚类方法研究[J]. 铁道标准设计，2015(4)：19-23.

[6] 贺鑫，李科. 基于聚类分析法的城市轨道交通站点分类[J]. 信息通信，2015(7)：36-37.

[7] 金昱. 城市轨道交通站点客流时变特征及其影响因素研究——以上海为例[J]. 现代城市研究，2015(6)：13-19.

[8] BAGCHI M，WHITE P R. The potential of public transport smart card data[J]. Transport Policy，2005，12(5)：464-474.

[9] KIM K，OH K，LEE Y K，et al. Discovery of travel patterns in seoul metropolitan subway using big data of smart card transaction systems[J]. The Journal of Society for E-Business Studies，2013，18(3)：211-222.

[10] KIM K W，LEE D W，CHUN Y H. A comparative study on the service coverages of subways and buses[J]. Ksce Journal of Civil Engineering，2010，14(6)：915-922.

[11] DUNN J C. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters[J]. Journal of Cybernetics，1973，3(3)：32-57.

[12] XIE X L，BENI G. A validity measure for fuzzy clustering[J]. IEEE Trans Pami，1991，13(13)：841-847.

Classification of Rail Transit Stations Based on AFC Data Mining

Deng Pingxin[1], Zheng Changjiang[1], Ma Genghua[2], Li Rui[1]

(1. College of Civil Engineering and Transportation Engineering, Hohai University, Nanjing 210098, China; 2. College of Harbor, Coastal and Offshore Engineering, Hohai University, Nanjing 210098, China)

Abstract：The classification of rail transit stations has an important impact on the study of passenger flow pat terns, surrounding land use and development trends of different types of stations. Based on AFC data, a variety of effectiveness indicators were used to determine the number of classifications. By way of principal component analysis, k-means clustering and multiple linear regression, qualitative and quantitative analysis were combined to classify the sites. A total of 58 stations of Suzhou Rail Transit Line 1 and 2 were divided into 4 categories, which would lay the foundation for the follow-up study of the classified sites and the development of rail transit.

Key words：AFC data; function orientation for station; k-means; regression analysis