基于自动采集数据的城市公交枢纽站识别方法
——以苏州市为例

杨阳 陈学武 李海波 许威

【摘要】公交枢纽站作为城市公共交通系统的重要组成部分，其良好运行可以提高整个公交线网的运营效率。对于一个城市的公交站点而言，需要运用科学合理的方法从所有站点中筛选出公交枢纽站。本研究利用苏州市公交自动采集数据，在数据处理的基础上，提取出可供选择的聚类变量；选用站点公交线路数、站点日上客流量和站点日换乘比例作为聚类变量，并选用 K-Means 聚类方法进行聚类；通过对聚类结果进行检验和类特征总结，最终识别出 201 个苏州市城市公交枢纽站。

【关键词】公交枢纽站；自动采集数据；聚类

1 引言

城市公交枢纽站作为公交网络的重要节点，是连接居民公交出行的重要纽带。实践证明，充分发挥公交枢纽站的运行效率，对整个公交系统运行效能的提升有重要的推动作用。

判断一个公交站是否为枢纽站，现阶段没有相对统一的标准。在规划过程中，往往基于站点区位、运营线路等级和功能，人为指定其是否为枢纽站，具有较强的主观性。对一个城市而言，公交站数量少则几百，多则上千。在实际运营过程中，部分未列入规划范围内的站点，实际上也承担了枢纽转换的功能。而指定的枢纽站在实际运营中是否真正起到了枢纽站的作用，也需要对实际数据进行分析后方可判断。

2 公交枢纽站定义

国内不同的规范、标准和著作在城市公交枢纽定义方面有一定的共性，也存在着差异。

《城市公共交通工程术语标准》[1]中对枢纽站的定义为：有多条公共交通线路汇集的客流集散量较大的起止站组合。

《城市公共交通系统规划方法与管理技术》[2]从公共客运的角度对枢纽进行了定义。公共客运交通枢纽是指公交线路之间、公共交通与其他交通方式之间客流转换相对集中的场所，分为对外交通枢纽和市内交通枢纽两种。

《城市道路公共交通站、场、厂工程设计规范》[3]中，对枢纽站也做了相应的规定。多

条道路公共交通线路共用首末站时应设置枢纽站，枢纽站可按到达和始发线路条数分类，多种交通方式之间换乘为综合枢纽站。

不难看出，城市公交枢纽站有着供给和需求两方面的要求。在供给方面，公交枢纽站应有多条线路供出行者选择；在需求方面，公交枢纽站应有较多的乘客进行集散和换乘。

3 研究数据

获得一个城市所有公交站点的运营数据，通过传统的人工调查法需要消耗大量的人力和财力，且并不能保证调查数据的完整性，几乎没有实施的可能性。近年来，随着智能公交系统的发展，公交刷卡系统、自动车辆定位系统、GPS 系统等在公共交通领域应用非常广泛。通过这些系统能够获取海量自动采集数据。自动采集数据信息量大且全面，技术简单成熟，成本低，非常适用于对整个城市公交系统的宏观研究。因此，本研究所用数据采用该类型数据。获取数据的城市是苏州市。

3.1 基础数据

自动采集数据由于数据量大，并且存在一定的数据缺失，因此在使用之前需要对数据进行处理，剔除无效部分，使之能够用于最终的数据分析。

3.1.1 IC 卡数据

利用公交 IC 卡数据可以获得站点客流量。公交 IC 卡数据信息量大，数据多，从原始数据表中可以获得每一条刷卡数据所包含的信息，如表 1 所示。

表 1 苏州市公交 IC 卡数据记录（部分）

LISTBH 卡编号	DEALRQ 交易日期	DEALSJ 交易时间	XLBH 线路编号	DEALBH 交易编号	KLX 卡类型	QCBH 汽车编号
21500030076106	20130518	130556	42	000552D8	普通	102410
21500002761120	20130518	130558	42	000552D9	普通	102410
21500002708627	20130518	130919	42	000552DA	普通	102410
21500002161064	20130518	131304	42	000552DB	普通	102410
21500000780473	20130518	131432	42	000552DC	普通	102410
21500000059233	20130518	132213	42	000552DD	普通	102410
21500001190531	20130518	132716	42	000552DD	普通	102410

在这些数据记录中，研究涉及的数据类型有卡编号、交易日期、交易时间、线路编号、卡类型和汽车编号等。由于苏州市居民乘坐公交下车不用刷卡，没有下车站点的刷卡记录，因此本研究只考虑站点的上车客流统计，不考虑下车客流。

3.1.2 AVL 数据

本研究所指的 AVL（Automatic Vehicle Location）数据是车辆在进出公交停靠站时，记

录下的到离站信息数据。通过分析 AVL 数据，可以获得公交站点的车辆到离站信息。原始数据记录如表 2 所示。

表 2 苏州市公交 AVL 数据记录（部分）

DGUID 车辆设备标识	DBusCard 车牌号	LName 线路名称	LSGUID 站点标识	LSName 站点名称	DInTime 进站时间	DOutTime 出站时间
12812	苏 E-10437	38	9f0d5e11-2743-4307-a3de-c120d6558419	火车站	2013/5/15 10:28	2013/5/15 10:28
12812	苏 E-10437	38	70a6aced-101e-f096-e29a-f5dca3f4a152	平门	2013/5/15 10:31	2013/5/15 10:31
12812	苏 E-10437	38	08d6865e-2c51-3d0b-499a-34c32195c698	接驾桥	2013/5/15 10:33	2013/5/15 10:34
12812	苏 E-10437	38	efe9128a-640c-9467-e13f-d40f0e1c10d8	乐桥北	2013/5/15 10:37	2013/5/15 10:38
12812	苏 E-10437	38	8c8e26d8-c547-66db-9ea6-6bc0bcb430bb	饮马桥	2013/5/15 10:39	2013/5/15 10:40
12812	苏 E-10437	38	51a960f6-17ef-ec40-121e-ae7b5cc7ca91	市立医院本部	2013/5/15 10:41	2013/5/15 10:42

在这些数据记录中，对于数据筛选有作用的数据类型有车牌号、站点名称、进站时间和出站时间等。

3.2 数据匹配

苏州市公交 IC 卡数据没有刷卡站点的信息，因此需要通过 AVL 数据进行数据匹配，从而获得 IC 卡的刷卡站点，进而统计各个站点客流量。将这两个系统的数据进行匹配需要通过车辆进行识别。

IC 卡数据与 AVL 数据的匹配关系主要有：汽车编号与 AVL 数据中车牌号匹配；线路编号与 AVL 数据中线路名称匹配，交易时间与 AVL 数据中车辆进站和出站时间匹配等。

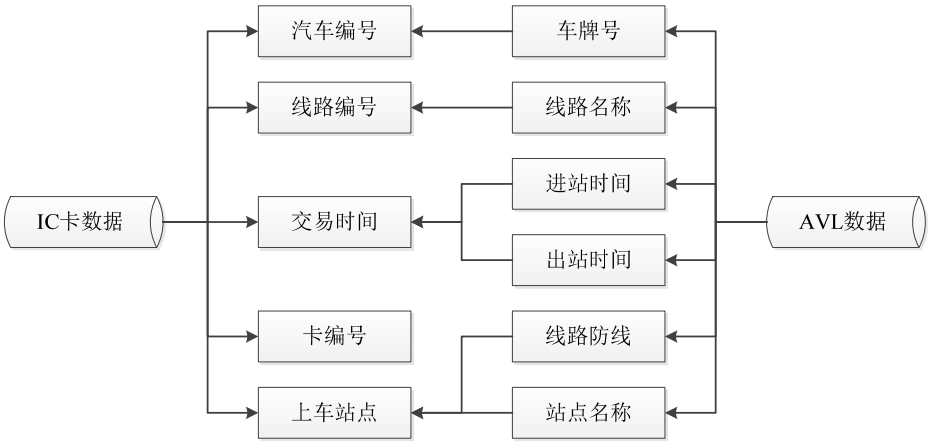

图 1 IC 卡数据与 AVL 数据匹配

在 IC 卡系统的时间与 AVL 系统的时间一致的情况下，IC 卡数据的刷卡时间和 AVL 数据的车辆进出站时间可直接匹配。但实际情况中，IC 卡系统和 AVL 系统由于技术原因，其所记录的时间会有差别，且每辆公交车的 IC 卡系统的时间与 AVL 系统的时间差各不相同，因此需要对 IC 卡系统与 AVL 系统的时间进行修正匹配。

3.3 时间修正

在通过刷卡时间判断刷卡站点前，需要通过不断循环调整两个系统之间的时间差来找出识别率最高的纪录对应的时间差。然后根据该时间差对 IC 卡数据内的刷卡时间进行一定的修正后再进行站点识别。此方法能够使最终的站点识别实现较高的成功率，两个系统的时间差范围设定在[-30min,30min]之间。

以 AVL 系统的时间为基准，假设一个系统时差，并对 IC 卡记录时间进行修正，然后根据 IC 卡记录时间进行站点判断，同时统计站点识别率。通过不断假设和判断，当站点识别率最高时，即认为对应时差即为系统实际时差。假设系统时间差为：

$$\Delta t = T_{IC} - T_{AVL} \tag{1}$$

式中：Δt ——系统时间差，s；

T_{IC} ——IC 卡 POS 机的记录时间，s；

T_{AVL} ——AVL 系统的记录时间，s。

那么 IC 卡系统记录的时间修正为：

$$T_{IC}^{'} = T_{IC} + \Delta t \tag{2}$$

式中：$T_{IC}^{'}$ ——修正后的 IC 卡 POS 及记录时间，s。

在时间差匹配并修正之后，通过刷卡时间与车辆进出站时间的比较，若 IC 卡刷卡时间落在车辆进出站时间之间，认为在该站上车，由此识别出乘客刷卡的站点，最终获得有刷卡站点属性的公交 IC 卡数据。

3.4 换乘识别方法

利用自动采集数据容易得到站点线路数、配车数、客流量等数据，但是判断一次刷卡行为是否为换乘比较复杂，需事先给出时间界定。

公交换乘指的是出行者完成一次公交出行，从出发公交起点站连续乘坐若干不同线路公交后，最终到达公交终点站的行为[4]。如果通过公交 IC 卡数据识别出出行者在一定时间范围内存在连续乘坐公交的话，那么认为连续两次乘车之间存在一次换乘过程。这对研究公交

枢纽站的换乘有着重要的意义。

苏州公交 IC 卡只记录了乘客的上车时间。为简化识别过程，只能通过前后两次刷卡的时间差来识别换乘，忽略换乘距离等其他因素的影响，通过参数标定，最终确定若前后两次刷卡时间间隔在 3.3-86min 之间，则认为是一次换乘。

4 公交枢纽站聚类分析

利用苏州市各公交站点数据将相似特征的站点进行归类，选择常用的分析方法——聚类分析。根据聚类的结果，同一个簇中的对象具有很大的相似性，而不同簇间的对象有很大的差异性。根据此方法来从所有的公交站点中识别枢纽站。SPSS 软件提供了三种聚类方法，分别为两步聚类、K-Means 聚类和系统聚类。但选用何种方法和参数聚类，需要进一步研究。

4.1 可供选择的变量

根据初步统计结果，筛选出有效的公交站点为 2483 个。将这些站点进行聚类分析，需要选择合适的变量，使之既能反映枢纽站的总体特征，又能反映站点之间的差异。

通过自动采集数据，可以获得六类重要的聚类变量。从供给方面考虑，有站点公交线路数、日到站车辆数和平均车头时距；从需求方面考虑，有站点日上客流量、站点日换乘流量和站点日换乘比例。

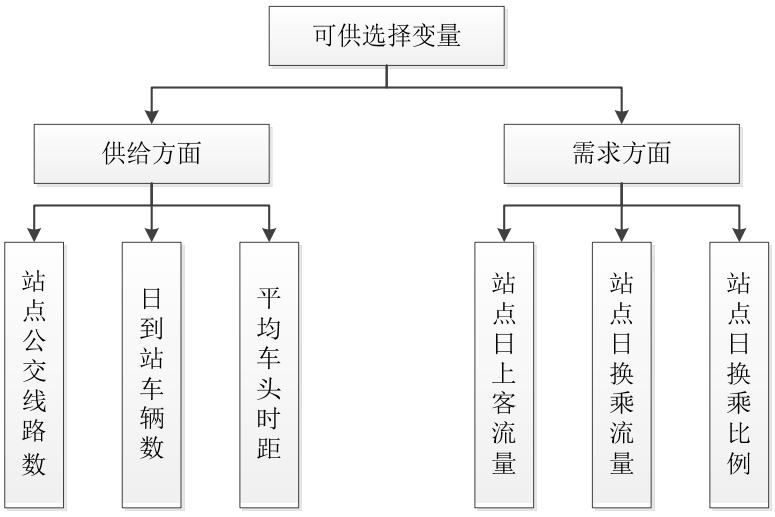

图 2 利用自动采集数据提取的变量

4.2 聚类变量选择结果

以上六个变量是否都适合用于聚类研究，需要进行进一步分析，即进行相关性检验。只有相关性较弱的变量才合适用于聚类分析。对这六个变量两两之间的相关性检验结果如表 3 所示。

表 3 双变量相关性检验

	公交线路数	日到站车辆数	平均车头时距	日上客流量	日换乘流量	日换乘比例
公交线路数	--	0.944	-0.736	0.609	0.715	0.431
日到站车辆数	--	--	-0.766	0.637	0.728	0.401
平均车头时距	--	--	--	-0.504	-0.465	-0.282
日上客流量	--	--	--	--	0.932	0.322
日换乘流量	--	--	--	--	--	0.531
日换乘比例	--	--	--	--	--	--

从上表可以看出，公交线路数与日到站车辆数，日上客流量与日换乘流量之间存在着非常大的相关性，而日换乘比例与其他五个变量之间的相关性均较小。综合比较，选取最能体现公交枢纽站特征的变量，即公交线路数、站点日上客流量和站点日换乘比例作为聚类的变量。

对公交线路数而言，由于一条线路不存在换乘情况，与枢纽站的功能特征不符，因此将线路数为 1 的站点剔除，共 780 个站点；对日上客流量而言，客流量过小也不符合枢纽站的特征，从数量级上考虑，将流量在 100 以下的站点剔除，共 562 个站点；同样方法将换乘比例在 10%以下的站点剔除，共 266 个站点。剔除部分数据之后，最终有效的聚类样本量为 875 个。

4.3 聚类结果分析

本研究选取的三个变量均为连续变量，在没有离散变量且变量种类较小的情况下，选用 K-Means 聚类方法有很好的效果。本研究选用该聚类方法。

4.3.1 分类结果及检验

在 SPSS 中将聚类变量进行数据标准化。通过不断调试，最终确定聚类数为 6。聚类结果如表 4 所示。

表 4 聚类结果汇总

聚类	个数	聚类	个数
1	447	4	133
2	51	5	141
3	17	6	86
有效数	875		
缺失数	0		

从方差检验表 5 看出，3 个变量中的任一类间均方远大于类内误差均方。从概率值来看，3 个变量使类间无差异的假设成立的概率均小于 0.1%。方差分析结果表明，参与聚类分析的 3 个变量能很好地区分各类，类间差异足够大。

表 5 方差检验

	聚类		误差		f	Sig.
	均方	df	均方	df		
站点公交线路数	124.415	5	.290	869	429.163	.000
站点日上客流量	144.111	5	.177	869	816.135	.000
站点日换乘比例	115.087	5	.344	869	334.974	.000

类中心变量值反映了在三个指标下，类中心的空间坐标。对标准化数据的类中心变量值统计如图 3 所示。可以看出，各类中心存在较大的差异，说明变量特征存在较大的差异性。

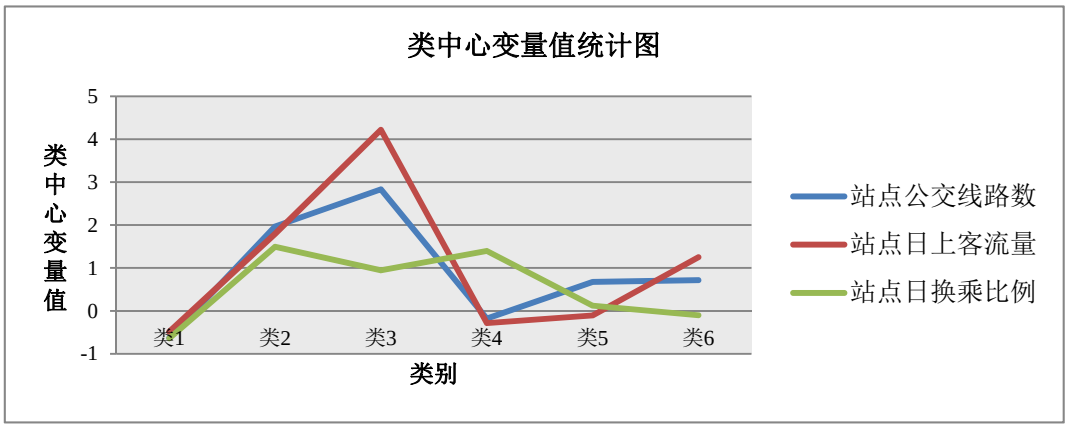

图 3 类中心变量值统计图

4.3.2 类特征总结

对聚类结果进行分类汇总，统计不同类别下的各变量类均值，如表 6 所示。

表 6 各变量类均值统计

	聚类均值					
	类 1	类 2	类 3	类 4	类 5	类 6
站点公交线路数	4	15	18	6	9	10
站点日上客流量	324	2092	3968	481	619	1673
站点日换乘比例	15.9%	34.2%	29.5%	33.4%	22.4%	20.4%

通过各变量类均值的统计结果，可以总结各类的特征。

类 1 代表的是公交线路数、日上客流量和日换乘比例均最小的一类；

类 2 代表的是公交线路数和日上客流量较多，日换乘比例最高的一类；

类 3 代表的是公交线路数和日上客流量最多，日换乘比例较大的一类；

类 4 代表的是公交线路数和日上客流量较小，但是日换乘比例较大的一类；

类 5 和类 6 代表的是公交线路数、日上客流量和日换乘比例均适中的一类。

5 公交枢纽站识别结果

根据前文对公交枢纽站的定义，在线路数、日上客流量和日换乘比例三个指标反映下，

公交枢纽站可呈现出以下三种特征：

（1）集散型：这种类型的枢纽站以客流的集散为主，公交日上客流量非常大，但是换乘比例较小或适中；

（2）换乘型：这种类型的枢纽站以客流的换乘为主，典型特征为日换乘比例非常高，但是公交线路数和日上客流量不一定大；

（3）混合型：公交枢纽站的线路数、日上客流量和日换乘比例均较高，集散和换乘的客流均占据一定数量。

将聚类得出的6类站点与以上三种特征对应，可以看出，类3属于集散特征明显的一类，类 4 属于换乘特征明显的一类，类 2 属于特征混合型的一类。因此，将类 2、类 3、类 4 的站点作为筛选出的公交枢纽站，共 201 个站点。

为验证三个变量作用下筛选结果的可靠性，将任一变量剔除，只用两个变量进行聚类，比较结果的差异。由两个变量聚类出的公交枢纽站在三个变量聚类出的 6 类中的分布情况如表 7 所示。

表 7 两个变量下的聚类结果分布

聚类变量	类 1	类 2	类 3	类 4	类 5	类 6	总数
站点公交线路数 站点日上客流量	--	22	17	--	--	1	40
站点公交线路数 站点日换乘比例	--	24	10	59	--	--	93
站点日上客流量 站点日换乘比例	--	41	17	82	10	1	151

可以看出，无论是选用三个变量中的哪两个进行聚类，其筛选出的公交枢纽站，绝大多数都落在了类 2、类 3、类 4 中，只有少量落在了类 5 和类 6 中。因此可以说明，使用站点公交线路数、站点日上客流量和站点日换乘比例三个变量进行聚类，筛选出的公交枢纽站更加全面，具有更广的代表性。

6 结语

城市公交枢纽站作为公交线网连接的关键节点，其运行效率的好坏对整个公交系统的效能大小有重要影响。本研究数据来源为公交系统自动采集数据（公交 IC 卡数据、AVL 数据等），这种类型的数据可以提供整个城市全天的刷卡和车辆到离站数据。基于大数据的特征提取，可以从宏观层面对城市所有站点进行研究。从公交系统实际运营数据中筛选出起到枢纽转换功能的站点，从而避免枢纽站只根据规划或人为指定进行选择。以往自动采集数据多

用于对乘客出行 OD 判别及换乘的研究，本文将其应用于苏州市城市公交枢纽站的识别，一整套完整的研究方法可为其他城市的相关研究提供思路。

【参考文献】

[1] CJJ/T 119-2008.城市公共交通工程术语标准[S].北京:中国建筑工业出版社,2008.

[2] 王炜,杨新苗,陈学武等.城市公共交通系统规划方法与管理技术[M].北京:科学出版社,2002.

[3] CJJ/T 15-2011.城市道路公共交通站、场、厂工程设计规范[S].北京:中华人民共和国住房和城乡建设部,2011.

[4] 吴美娥.对公交 IC 卡数据处理分析及应用的探索[D].北京:北京交通大学,2010.

【作者简介】

杨阳，男，硕士，悉地（苏州）勘察设计顾问有限公司，助理工程师。电子信箱：543619636@qq.com

陈学武，女，博士，东南大学城市智能交通江苏省重点实验室、现代城市交通技术江苏高校协同创新中心，教授，博士生导师。电子信箱：chenxuewu@seu.edu.cn

李海波，男，本科，东南大学城市智能交通江苏省重点实验室、现代城市交通技术江苏高校协同创新中心，博士研究生。电子信箱：jslihaibo@foxmail.com

许威，男，硕士，悉地（苏州）勘察设计顾问有限公司，助理工程师。电子信箱：978355728@qq.com