Beijing City Lab

Long Y, Han H Y, Yu X, 2013, Discovering Functional Zones Using Bus Smart Card Data and Points of Interest in Beijing. Beijing City Lab. Working paper # 11

基于北京公交刷卡数据和兴趣点的城市功能区识别

摘要：城市在其发展过程中逐渐形成居住区、工业区和商业区等不同的功能区。识别这些功能区并理解其分布特征，对于把握城市结构以及制定和使用科学合理的规划具有重要作用。本研究基于 2008 年 4 月北京市连续一周的 7797 余万条公交 IC 卡刷卡数据（Smart Card Data, SCD），将其转换为每个公交站台流量的二维时间序列数据，结合居民日常出行行为研究，利用数据挖掘技术，构建了基于公交刷卡数据和兴趣点（Points of Interest, POIs）的城市功能区识别(Discovering Zones of different Functions, DZoF)模型，并将识别结果在交通分析小区（TAZ）尺度上汇总。研究结果显示，利用 DZoF 模型通过恰当的数据降维和期望最大化(expectation-maximization, EM)算法进行聚类分析，可以识别出与北京市土地利用现状地图具有一定匹配度的北京市各功能区。本研究的方法可以辅助规划人员和公众有效识别和理解复杂的城市空间结构，对城市地理及规划研究具有重要的理论和实践价值。

关键词：公交IC卡刷卡数据；兴趣点；出行行为；功能区识别；北京

1、引言

城市是一个高度结构化的系统，其内部要素分布非正态，且无时无刻不处在一种自组织的临界状态。因此，不论从物质层面还是社会层面，城市都是一个复杂的空间系统 [1,2]，而基于微观对象的研究是理解这一复杂系统运作规律的重要途径。传统的对于城市要素、组织、结构等方面的研究由于受到数据条件的限制，往往局限在大尺度上（如乡镇或交通分析小区 TAZ）。而随着大数据时代的到来，遍布的（Ubiquitous）个人贡献（Volunteered）数据为描述和理解城市空间结构提供了新的渠道[3]。

各研究领域所关注的大数据，主要是由移动通讯（GSM）、全球定位系统（GPS）、社会化网络（SNS）和无线宽带热点（Wi-Fi）等基于位置服务（Location Based Services, LBS）技术所提供的公交智能卡刷卡记录、航班记录、银行卡记录、微博记录和手机通话记录等。这些数据可以形成城市居民的出行日志[4]，并被用来对城市活动进行实时监控[5]，分析城市活动的强度和时空分布特征[6]，研

究城市中人类活动模型（Human Mobility Pattern）[7]。

人类活动（Human Mobility）与城市空间结构之间有着密切的关系[8,9]。现有研究中，学者主要通过城市空间结构辅助研究人类活动，分析人们出行特征，探讨城市空间结构对人的出行影响[10]。例如，通过城市土地利用结构来研究城市通勤模型[11]，分析空间结构对于居民通勤行为的影响[12,13]。与此相反，对于如何利用已有的城市中人类活动数据，来进行城市空间结构的有关研究则鲜有学者涉及，但是这却又是非常重要的。因为伴随着城市的发展，城市的土地利用与空间结构正快速地发生变化，城市由过去的单一中心模式向多中心模式发展[14,15]，即时且明确的城市功能区块划分能够启发城市规划者对城市未来的规划，并对以往的用地规划进行验证。而传统的对于城市土地利用及空间结构的研究往往是基于遥感数据[16,17]，这些数据价格昂贵而且不能及时更新，不能满足城市规划者和学者研究与应用的需要。所以，利用 LBS 技术提供的海量数据，分析人类活动，进而开展城市空间结构的研究，将逐渐成为未来城市研究的热点。

现有城市空间结构研究中，移动通讯（GSM）和全球定位系统（GPS）数据应用最为广泛。例如，Qi 等利用杭州市出租车行驶的 GPS 信息，分析了出租车乘客上下车特征与城市区域社会功能的关系[18]，Liu 等基于上海市连续一周的 6600 多台出租车行驶的 GPS 数据，利用由 Pulliam 提出的“源-库”（source-sink）模型将日常交通模型特征化，进而用于分析上海市土地利用现状[19,20]；Yuan 等通过利用出租车 GPS 信息和城市兴趣点数据，建立语义模型，运用数据挖掘的方法研究城市不同区域的功能划分[21]。

近些年来，作为一种大规模的具有地理标识和时间标签的公交 IC 卡刷卡数据（SCD），也逐渐被用于城市研究中。Sun 根据新加坡的 SCD 数据分析了乘客时间空间密度（Spatio-temporal Density）和活动轨迹[22]；Joh 和 Hwang 利用首尔大都市区 1000 万条的 SCD 数据，分析了公交卡持卡人的出行轨迹与市区土地利用的特征[23]；龙瀛等人利用北京市公交 IC 卡刷卡数据进行了北京职住关系及通勤流向分析[24]。此外，一些学者开始尝试利用兴趣点(Point of Interests，POIs)数据进行城市空间的研究。兴趣点是根据城市特点增加的一类基本地点，主要包括局部范围内具有地理标志作用的建筑物，兴趣点描述了实体的空间和属性信息，如实体的名称、类别、坐标等。由于兴趣点能很大程度地增强对实体位置的描述

能力，提高地理定位的精度和速度，其已经被广泛地应用于城市空间结构的研究中。赵卫锋等依据显著度的差异从城市兴趣点数据中提出分层地标，获得能够用于智能化路径引导的层次性知识空间[25]。由于 LBS 技术具有定位精度高、交互性强、数据量大的特点，而兴趣点在识别地域模式方面具有明显优势，因此将二者整合用于城市空间结构研究具有重要意义。

由 LBS 技术获得的信息尚处于它的原始状态：数据，由于数据量太大，采用传统的数据分析方法和技术实现有效信息的提取往往是难以实现的[26]。近年来，随着计算机技术飞速发展，数据库管理系统和人工智能技术逐步走向成熟，两者的融合促成了数据挖掘这一新技术的产生，有效地实现了对大数据的表征和分析[27]。学者们也开始尝试利用数据挖掘技术中的分类、关联分析、聚类分析从海量的地理信息数据中挖掘出潜在有用的信息。近些年，聚类分析在国内外被广泛用于基于 GPS、GSM、SCD 等数据，对城市中人类工作、居住、上学、交通和购物等日常活动进行分析[28,29]，进而识别出城市的时空结构[30]以及即时且详细的城市土地利用现状[31]的研究中。

在实际研究中，研究者使用的由 LBS 技术提供的数据在时间上往往是连续的，这类数据被称作时间序列数据[32]。时间序列数据通常是海量数据，数据中可能存在大量的噪声，直接对原始数据进行聚类分析效率很低，甚至不可行。因此，需要对多维时间序列数据进行降维及特征变换。常用的方法有离散傅里叶变换(Discrete Fourier Transform，DFT)[32] 、主成分分析（Principal Component Analysis, PCA）[33]、奇异值分解(Singular Value Decomposition, SVD)[34]等。对时间序列进行聚类的算法有基于相似性、基于特征、基于模型和基于分割的聚类分析。在算法的选择主要取决于数据的类型、聚类的目的和应用[35]。传统的聚类分析大多是基于向量的，它们不能很好地解决时间序列聚类问题。近年来，对时间序列的聚类研究更多地使用基于模型的聚类分析。

本研究正是基于 SCD 和兴趣点的原始数据，建立城市功能区识别(Discovering Zones of different Functions, DZoF）模型。在此模型中，构建公交站台（platform）流量数据模型，利用基于模型的期望最大化(expectation-maximization，EM)算法，对北京市 8691 个公交站台进行聚类分析，并基于传统的居民出行行为研究、居民的普遍认知、兴趣点数据模型建立 SCD

数据挖掘的模式识别规则，对聚类所得的簇（cluster）进行功能诠释。根据 DZoF 模型，本研究最终确定了北京市各个公交站台的功能，并在交通分析小区（TAZ）尺度上进行汇总，实现不同区域功能的识别。为了验证 DZoF（Discovering Zones of different Functions）模型识别结果的有效性，研究结果还与北京市城市总体规划（2004-2020）的用地现状图及该地区谷歌地图进行对比分析。

本文方法部分定义了用于城市功能识别的 DZoF 模型，并对利用 DZoF 模型实现多维时间序列聚类分析，最终识别出由聚类结果的实际意义的具体方法进行了介绍；应用部分将以北京市为例，基于北京市 2008 年四月连续一周的公交 IC 卡刷卡数据和城市兴趣点数据对该城市的各功能区进行识别，并对实验结果进行检验；最后，对整个研究进行总结和讨论。

2 研究区概况及数据

2.1 研究区概况

本研究的研究区域为北京市域，总面积为 16410 平方公里，常住人口 2069.3 万[1]。北京市拥有四通八达的现代化、立体交通网络。2008 年，全市公共电汽车线路共有 648 条，运营里程 1.7 万公里，运营车辆 2.06 万辆；轨道交通线路为 8 条，运营里程达到 200 公里；运出租车运营车辆 6.6 万辆[2]。

2.2 数据

2.2.1 公交线路和公交站台

本研究的数据主要为北京市 2008 年 4 月连续一周（4 月 7 日—4 月 13 日）的公交 IC 卡刷卡数据(该数据不包括轨道交通刷卡数据)，其中共涉及超过 600 条公交线路（上下行计算共计 1287 条，其中一票制线路 566 条，分段计价线路 721 条），约 3.7 万个公交站点(stop)[3]和 8691 个公交站台（platform）。图 1 为北京市公交站台（Platform）的分布图。

[1]数据来自北京市 2012 年统计年鉴 http://www.bjstats.gov.cn
[2]数据来自北京市公交网统计数据 http://www.bjbus.com/
[3]该数字为所有公交线路站点数的综合，而不是站台数量，站点为某一公交线路上某站台的名称。

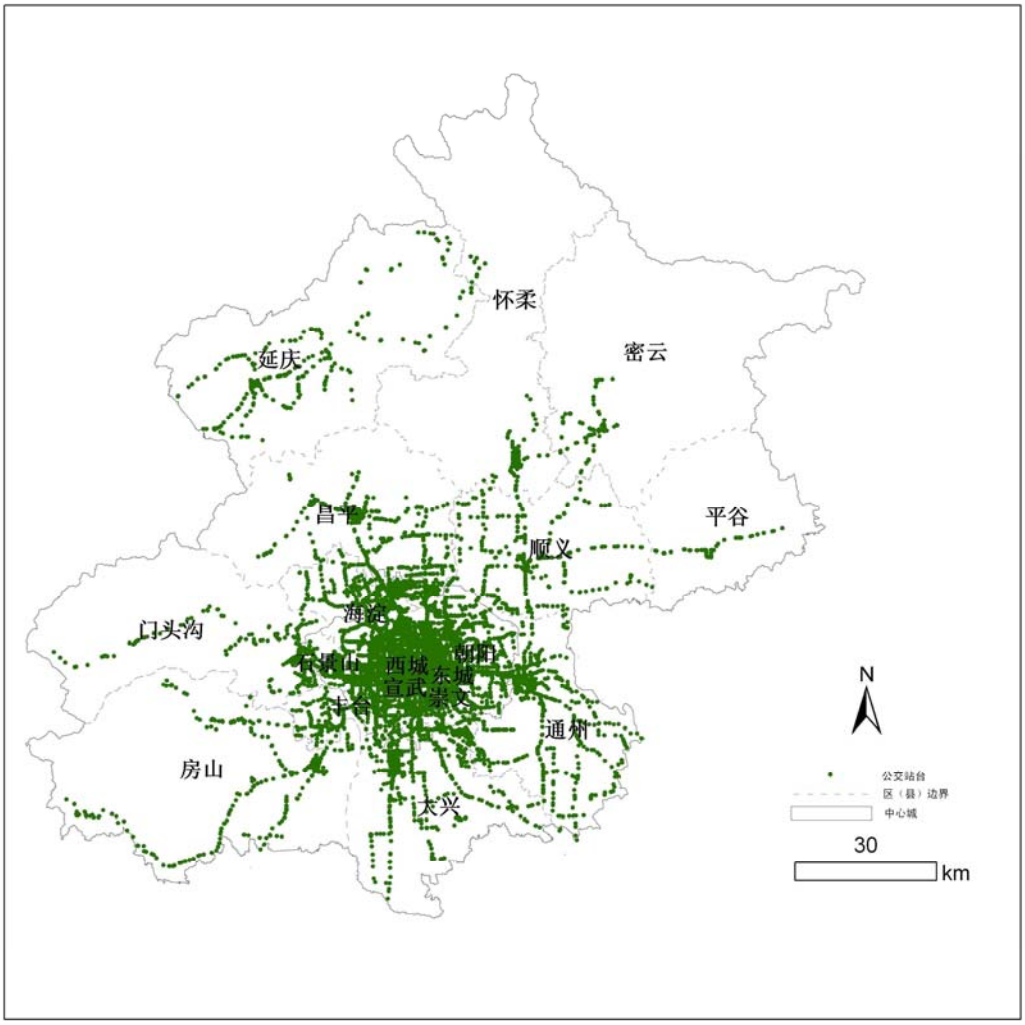

图 1 北京市公交站台分布图

Fig.1 The platform of the Beijing Metropolitan Area(BMA)

2.2.2 公交 IC 卡刷卡数据

鉴于非技术原因，本文所利用的公交 IC 卡刷卡数据不包括祥龙公司的运通线线路和轨道交通数据。记录包涵的基本信息包括：每个持卡人刷卡的时间和地点（其中地点以线路号和站点号表示）、卡类型（普通卡、学生卡或工作人员卡等）、交易序号（表示持卡人累计刷卡次数）、司机编号和车辆编号等。该一周数据共有 77976010 次刷卡记录。

北京市公交线路按照计价方式分为两种：（1）短距离的一票制线路，主要位于中心城区，此类线路 SCD 仅记录上车刷卡时间，无下车信息。（2）分段计价线路，这一类线路的路线一般比较长，一端或起始站点都位于五环之外，计价方

式为分段计价。这类线路的 SCD 则记录了持卡人的完整刷卡时空信息。本研究为了统计上下车流量，所用数据主要为分段计价刷卡数据，共 37649207 条。

2.2.3 兴趣点（POI）

本研究中所使用的兴趣点数据为北京市 2010 年兴趣点数据，共有 113810 条，来自新浪微博地理服务平台[4]。如图 2 所示。

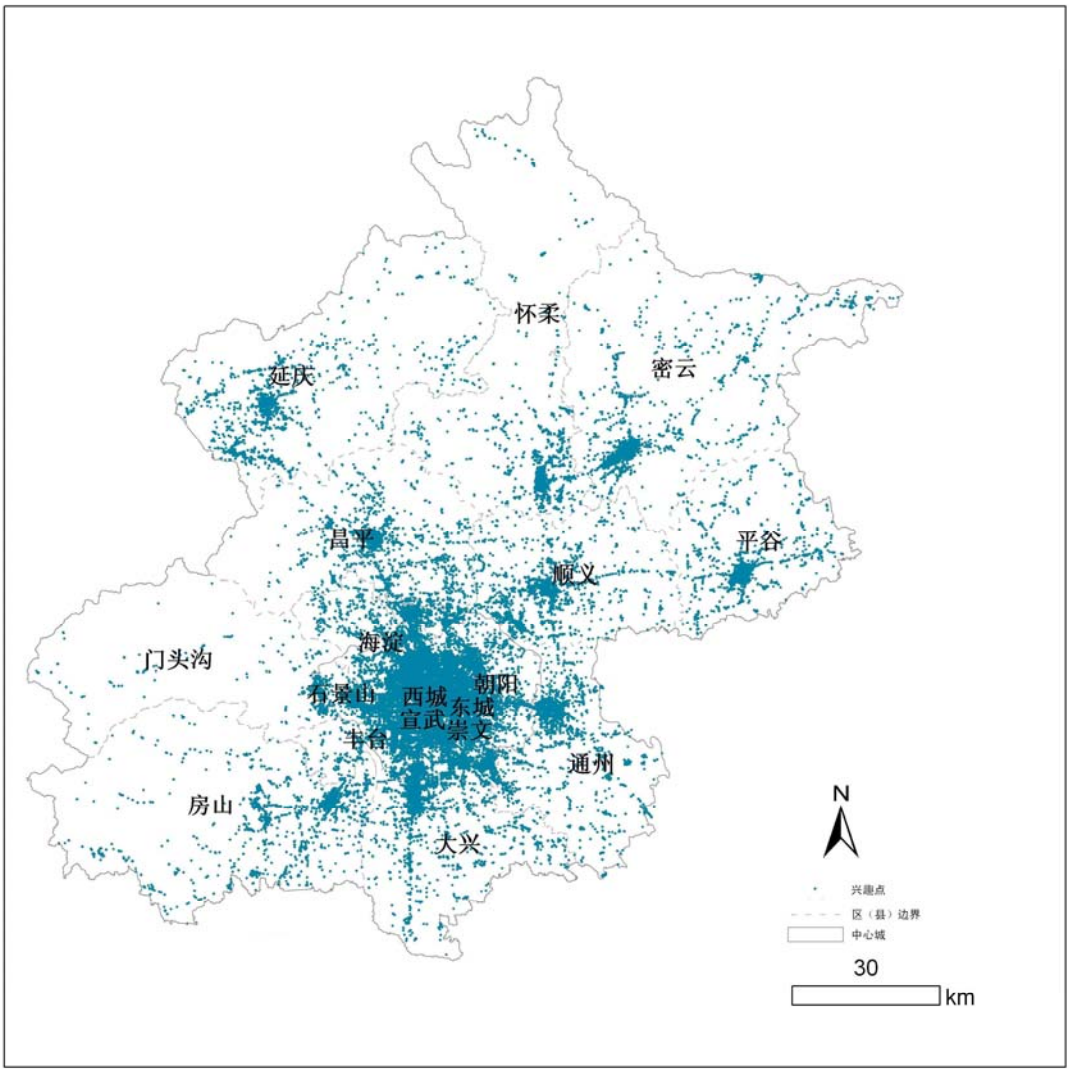

图 2　2010 年北京市兴趣点空间分布图

Fig.2 The POIs of the Beijing Metropolitan Area(BMA)

兴趣点数据包含三级编码，其中，一级编码分类及解释说明如表 1 所示。

[4]新浪开放平台，2012 年 4 月新浪微博 LBS 平台正式开放，第三方开发者可免费接入新浪位置服务。新浪微博 LBS 平台最具特色的是基于用户及基于 POI(具体位置点)的接口两个功能，基于用户的相关接口，使用户能获取单个人的时间线动态；POI 接口是基于某个具体位置的接口。http://open.weibo.com/

表 1 兴趣点一级编码及类别说明

Tab.1 Codes，Categories and Description of POIs

一级编码	兴趣点类别	解释说明	一级编码	兴趣点类别	解释说明
01	汽车服务	加油站、加气站、汽车养护、洗车场、汽车租赁等	11	风景名胜	公园广场、风景名胜
02	汽车销售	大众销售、丰田销售、本田销售、通用销售、宝马销售等	12	商务住宅	产业园区、楼宇相关、住宅区等
03	汽车维修	汽车综合维修、大众维修、本田维修等	13	政府机构及社会团体	政府机关、外国机构、社会团体等
04	摩托车服务	摩托车销售、摩托车维修	14	科教文化服务	博物馆、图书馆、文化馆、学校、科研机构等
05	餐饮服务	中餐厅、外国餐厅、快餐厅、咖啡厅等	15	交通设施	飞机场、火车站、港口码头、地铁站等
06	购物服务	商场、便民商店、家电卖场、超市、家具建材市场等	16	金融保险服务	银行、保险公司、证券公司、财务公司
07	生活服务	旅行社、邮局、物流速递、人才市场、电力营业厅、美容美发点等	17	公司企业	知名企业、公司、工厂、其他农林牧渔基地
08	体育休闲服务	运动场所、娱乐场所、休闲场所、影剧院相关等	18	道路附属设施	收费站、加油站服务区
09	医疗保健服务	综合医院、专科医院、诊所等	19	地名地址信息	交通地名、城市中心
10	住宿服务	酒店宾馆、旅馆招待所	20	公共设施	报刊亭、公共厕所、经济避难场所

不同类别的兴趣点数量如图 3 所示：

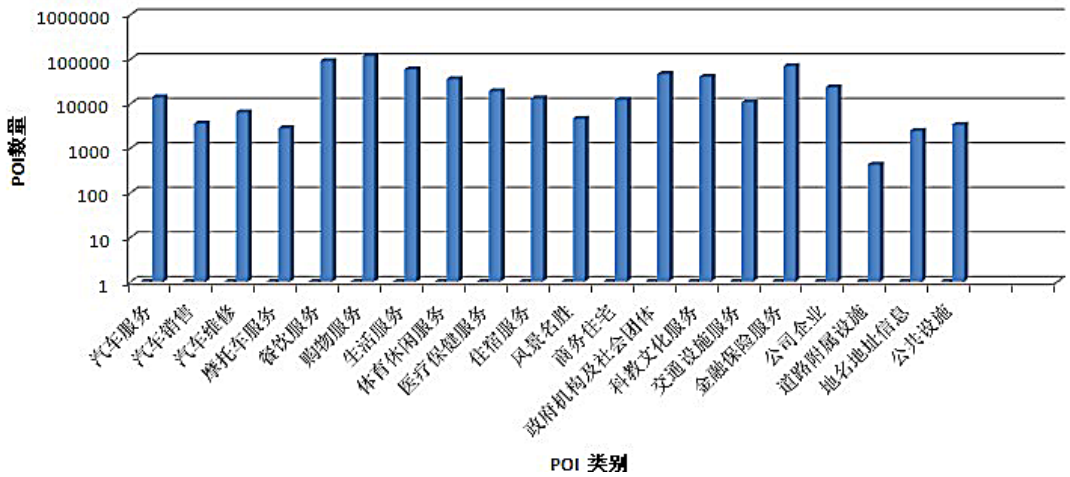

图 3　2010 年北京市兴趣点分类统计图

Fig.3 POI counts for each category in 2010

3 研究方法

本文主要通过构建 DZoF（Discovery Zones of different Functions）模型来进行城市功能区的识别。

首先，利用 SQL Sever 实现 SCD 数据和兴趣点的采集以及数据的预处理工作，构建公交站台流量数据模型（PF 模型）和兴趣点数据模型。再次，利用多维时间数据的聚类分析技术，对公交站台按照一周 7 天每天 24 小时的客流量数据进行特征构建，运用 EM 聚类算法进行聚类分析，得到簇。再次，根据兴趣点数据模型、居民出行行为特征、居民的普遍认知三个方面对聚类所得到的簇进行功能诠释，明确各个公交站台的功能。主要的功能包括公共管理及科教文化、居住、商业娱乐和风景名胜等。本研究的总体技术路线如图 4 所示。

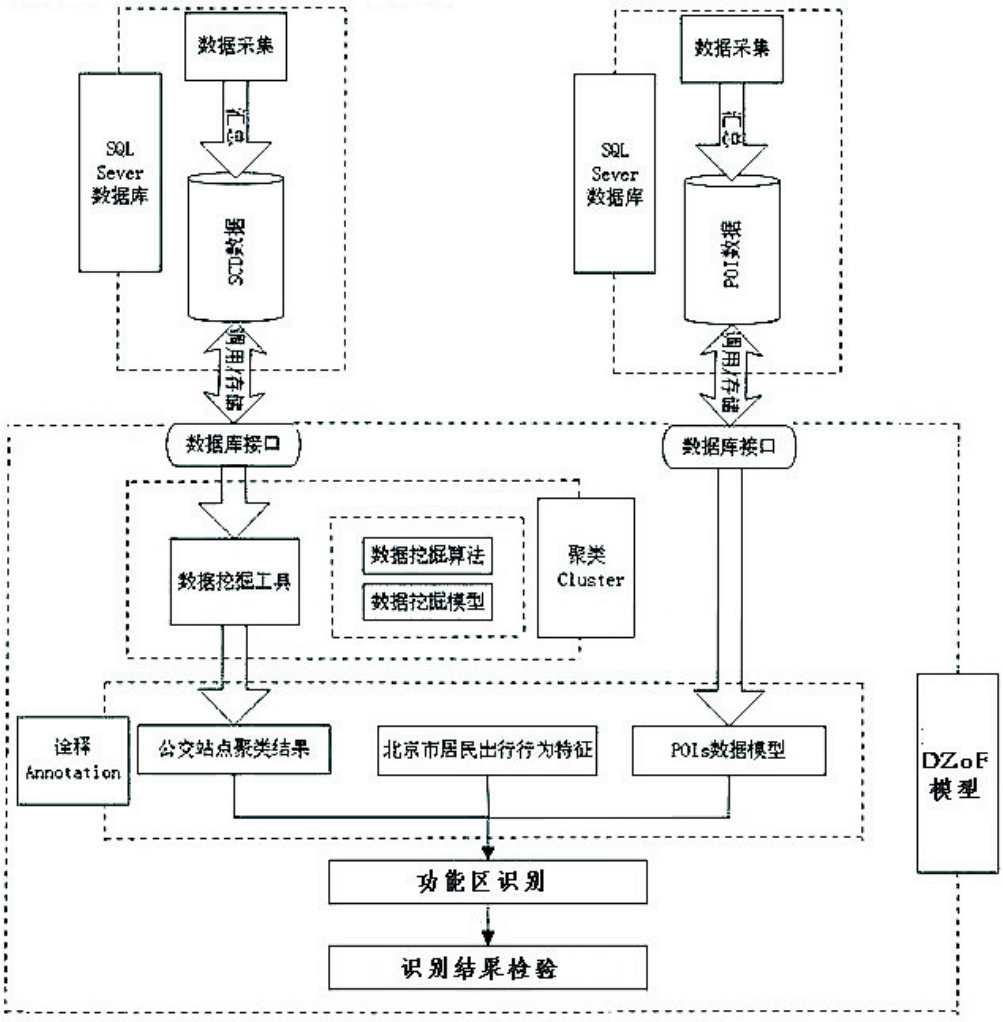

图 4 技术路线图

Fig.4 The process flow diagram of this paper

3.1 公交站台的盲聚类

3.1.1 公交 IC 卡刷卡数据预处理

原始数据是按照公交线路进行的流量统计，存在多条公交线路经过同一个站台（platform）的问题，利用 SQL Sever 软件，实现对相同公交站台的不同公交线路统计值的加和处理，统计得到每个公交站台的流量$f_{x,y,z}$，其中 x 为 Platform ID（x=1,2,…,8691），y 为日期（y=7,8,…,13），z 为时间（z=0,1…,23）。

3.1.2 PF（Platform flows）数据模型

针对每一个公交站台（platform），构建一个上车流量统计向量（Inflows vector），$<X_{7,0},X_{7,1},\ldots,X_{i,j},\ldots,X_{13,23}>$，其中 $X_{i,j}$ 为该站点（platform）在 2008 年 4 月 i 日，第 j 个小时内的上车人数（i=7,8,…,13；j=0,1…,23）。同时，构建一个下车流量统计向量（Outflows vector），$<Y_{7,0},Y_{7,1},\ldots,Y_{i,j},\ldots,Y_{13,23}>$，其中 $Y_{i,j}$ 为该站点（platform）在 2008 年 4 月 i 日，第 j 个小时内的下车人数（i=7,8,…,13;j=5,6…,23）。

本研究数据通过转换为二维时间序列数据，并进行维度归约，构造线性函数，利用不同时间上下车人数的比值作为公交站台流量相似性比较的指标数据：

$$Z_{ij} = \frac{X_{i,j}}{Y_{i,j}} \qquad (1)$$

($X_{i,j}$ 为 4 月 i 日，第 j 小时内的上车人数，$Y_{i,j}$ 为 4 月 i 日，第 j 小时内的下车人数)

因此，针对每一个公交站台（platform）构建 PF 模型，即 $<Z_{7,0},Z_{7,1},\ldots,Z_{i,j},\ldots,Z_{13,23}>$，其中 $Z_{i,j}$ 为该站点（platform）在第 i 天，第 j 个小时内的上车人数（i=7,8,…,13；j=0,1…,23）。

3.1.3 数据降维

根据北京市一星期内 24 小时平均公交上下车流量数据统计（如图 5），可以看出出行时间集中在 5:00-23:00 时间段（图 5 中红线以上部分）。对原数据进行

冗余特征的约简，PF 数据由 168 维降为 126 维。

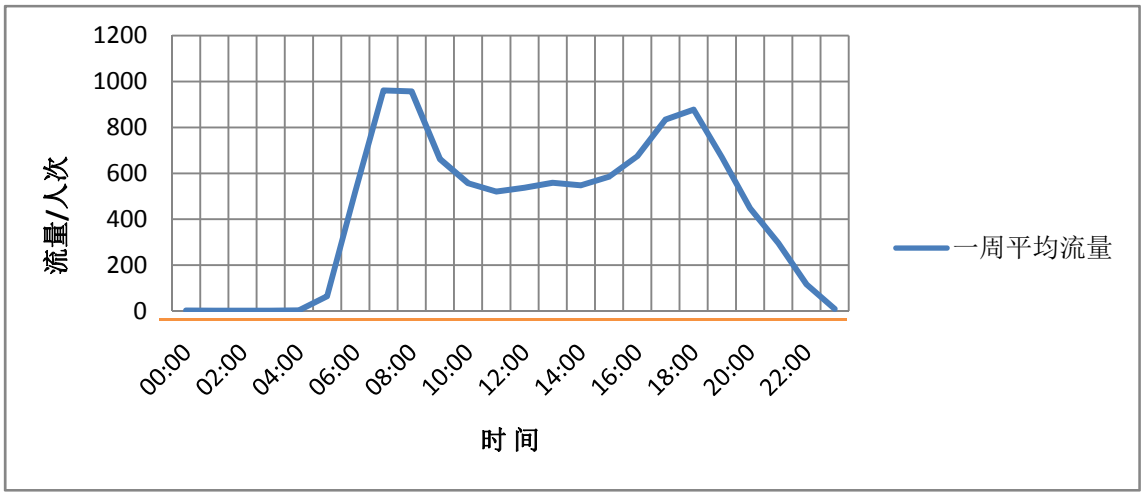

图 5 一星期 24 小时平均公交上下车流量数据

Fig.5 The average flows in various departure hours of a week

接下来，对 Platform ID 为 1934 的公交站台一星期内 24 小时上车流量数据进行绘图（图 6，图 7），通过对图像的观察，发现工作日内 weekdays 的流量数据具有较强的一致性性，而周末两天的（weekends）流量数据则具有较强的一致性。

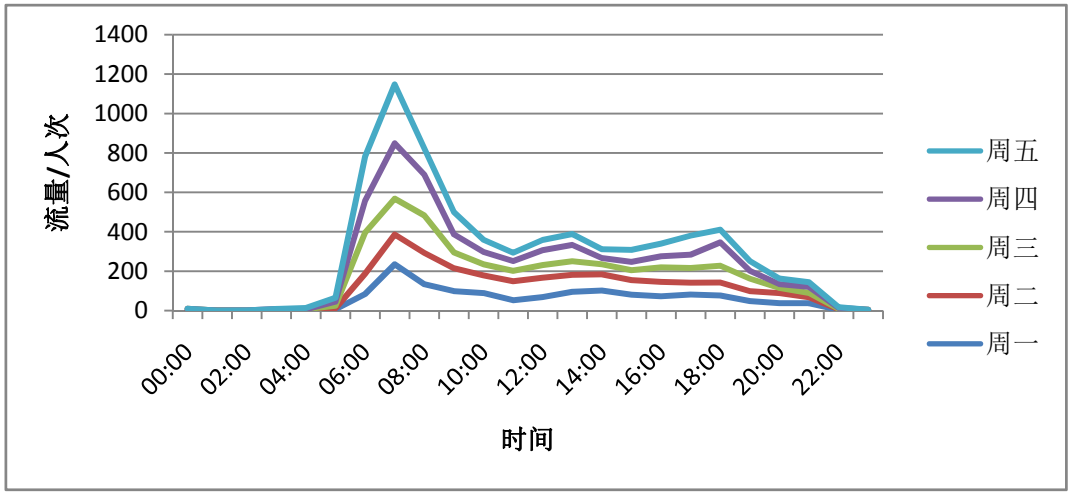

图 6 Platform_ID 为 1934 的公交站台工作日 24 小时上车流量

Fig.6 The inflows in various departure hours on weekdays of the 1934th platform

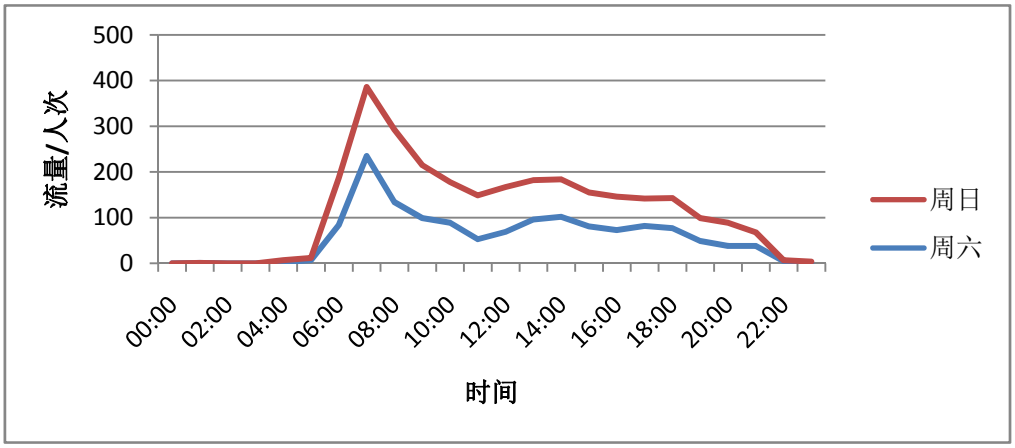

图 7　Platform_ID 为 1934 的公交站台休息日 24 小时上车流量

Fig.7　The inflows in various departure hours i On the weekend of the 1934th platform

进而，对工作日和周末的公交站台的流量数据进行了相关性分析，计算各属性间皮尔孙（Pearson）相关系数。结果显示，不同工作日内相同时间段的流量数据在 0.01（双侧）水平上呈显著相关。因此，对数据进行特征创建，对一周工作日内和周末两天的数据分别进行计算其算数平均数。这样原数据维度就由 126 维降至 36 维，在保留原数据特征的基础上，避免了维度灾难，使聚类算法的效果更好，同时降低了数据挖掘的时间和内存需求。

3.1.4 期望最大化(expectation-maximization, EM)算法

本文选用期望最大化(expectation-maximization, EM)算法作为公交站台的聚类方法。该算法对于每个对象，计算其属于每个分布的概率，就相当于 K 均值算法中将每个对象指派到一个簇中的步骤；算法中进行最大化似然估计模型的参数相当于均值算法中计算簇的质心。而 EM 算法相较于 K 均值算法更具有一般性，可以适用于各种不同的类，也可以发现不同大小的簇。同时，EM 算法是基于模型的算法，可以消除与数据相关联的复杂性。

EM 算法聚类是一种是用混合模型的聚类方法。基于模型的聚类的原理是假设数据是由一个统计过程得到的，可以按照统计模型对数据进行分类。因此，可以试图找出最佳拟合数据的统计模型，并由数据来估计该模型的参数。该算法的基本过程大致可归结为：首先，对参数做初始猜想；然后，迭代地改进这些估计。算法中对参数的估计是利用最大似然法进行的。一维高斯分布所产生点的概率密度为：

$$\text{prob}(\chi|\Theta) = \Pi\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \qquad (2)$$

如果 σ 和 μ 的值是未知的，就需要一个过程来估计他们，即选择使得上式最大化的 σ 和 μ。这种估计模型参数的方法在统计学上称作最大似然估计法。

3.2．城市功能区的识别

3.2.1 公交服务区 POI 数据的采集

本研究以 500 米为半径界定公交站服务区的空间尺度[36]，建立以公交站台为中心，以 500 米为半径的城市公交站服务区。针对每个公交站台，统计不同类别的 POI 数据个数，表示为$P_{i,j}$，其中 i 为公交站台的 ID 号(i=1,2,…,8691)，j 为 POI 一级编码（j=1,2,…,20）。

3.2.2 数据标准化处理

2010 年北京市的 POI 数据中，餐饮服务 POI 共有 90819 个，购物服务 POI 共有 116499 个，而风景名胜 POI 只有 4575 个。在统计分析的过程中，这种 POI 数量级的不同会影响对于功能的识别。

因此，需要对原始 POI 数据按照如下公式进行 Z-Score 标准化。

$$x_{ij}^{*} = \begin{cases} \dfrac{x_{i,j} - \overline{x_j}}{S_j} & (S_j \neq 0) \\ 0 & (S_j = 0) \end{cases} \quad (i = 1,2,\dots,n; j = 1,2,\dots,m) \qquad (3)$$

x_{ij}^{*}为标准值，x_{ij}为原始值，S_j为标准差，$\overline{x_j}$为均值，m 和 n 为数据矩阵行数、列数

3.2.3 POI 数据模型

对于每一个公交站台，构建一个 POI 数据的特征向量，FD（Frequency Density）数据模型，表示为<fd_1，fd_2，…,fd_{20}>。

其中 fd_i 表示在公交站台服务区 r 内的第 i 类兴趣点的频数密度（frequency density），即：

$$fd_i = \frac{\text{标准化后该公交站台服务区内第 i 类 POI 的个数}}{\text{公交站台服务区 r 的面积}} \quad (4)$$

同样，对于每一个公交站台，构建了另外一个 POI 数据的特征向量，CR (Category Ratio)数据模型，表示为<cr_1，cr_2，…,cr_{20}>。

其中 cr_i 表示第 i 类兴趣点在该区域中所有兴趣点的百分比，即：

$$cr_i = \frac{\text{标准化后该公交站台服务区内第 i 类 POI 的个数}}{|\text{标注化后该公交站台服务区内 POI 的总数}|} \quad (5)$$

3.2.4 城市功能的识别

本研究利用传统的数据获得规则，再将这些规则用于大数据的模式识别。即利用从已有的居民在出行时间和出行目的相关性特征，居民的普遍认知，以及 POI 数据模型等，对公交站台盲聚类得到的簇（cluster）进行功能识别判断。

功能的识别主要是以下三种方法的综合应用：

（1）通过计算得出每个簇（cluster）的 FD(Frequency Density)模型，并对该数据进行排序（得到内部排名）；其次计算得出每个簇（cluster）中的 CR(Category Ratio)模型，进行排序（得到外部排名）。

（2）根据每个簇（cluster）的出行时间流量特征进行判断。

（3）居民的普遍认知。人们通常了解一些知名区域的功能，比如故宫、中关村、北海公园等

4. 结果

4.1 公交站台聚类结果及 TAZ 尺度汇总

利用 EM 算法对公交站点依据流量数据进行聚类，分别得到 6 个不同的簇（每一个公交站点唯一地属于一个簇，C0-C5）。之后，利用公交站台和交通分析小区（TAZ）的空间从属关系，对每一个交通分析小区进行统计，选取分布最多的簇别作为该交通分析小区的类别，将聚类结果在交通分析小区尺度上汇总（Sparse 为未分类区域），如图 8 所示。

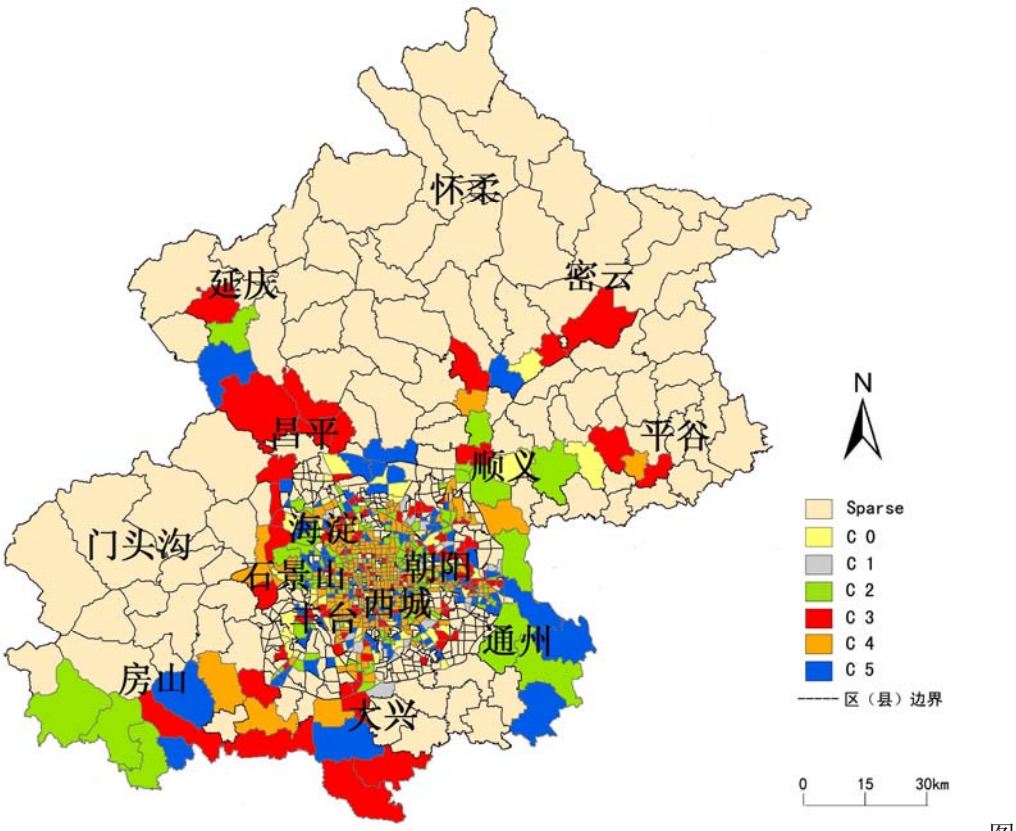

图 8 北京市功能区域图

Fig.8 Functional regions of the Beijing

4.2 功能识别

4.2.1 POI 模型建立

依照 4.1 节公交站台盲聚类结果，对各聚类所得簇（C0—C5）分别建立 POI 数据模型，计算各功能区的 FD（Frequency Density）值和 RCR(Rank of Category ratio)值，如表 2 所示。

表 2 EM 聚类所得功能区的兴趣点特征值

（FD: Frequency Density，RCR: Rank of Category Ratio）

Tab.2 Overall POI feature vector and ranking of functional regions by DZoF:FD: Frequency Density，RCR: Rank of Category Ratio

	C0		C1		C2		C3		C4		C5	
POI	FD	RCR	FD	RCR	FD	RCR	FD	RCR	FD	RCR	FD	RCR
汽车服务	-0.077	**7**	-0.025	**6**	0.073	**9**	0.03	**19**	0.021	**18**	-0.02	**6**
汽车销售	-0.075	**6**	0.034	**2**	-0.006	**19**	0.089	**14**	0.073	**13**	-0.063	**12**
汽车维修	-0.005	**3**	0.032	**3**	0	**18**	0.119	**9**	0.084	**12**	-0.012	**4**
摩托车服务	0.063	**1**	0.006	**5**	0.057	**13**	0.085	**15**	0.041	**16**	0.117	**1**

餐饮服务	-0.186	18	-0.109	13	0.142	1	0.149	7	0.205	5	-0.095	18
购物服务	-0.173	16	-0.141	16	0.039	15	0.214	3	0.107	11	-0.051	10
生活服务	-0.156	13	-0.157	18	0.099	5	0.216	2	0.114	10	-0.026	7
体育休闲服务	-0.16	14	-0.114	14	0.124	2	0.095	10	0.307	1	-0.057	11
医疗保健服务	-0.106	9	0.013	4	0.06	12	0.187	5	0.056	15	-0.004	3
住宿服务	-0.183	17	-0.14	15	0.075	8	0.187	4	0.18	6	-0.034	8
风景名胜	-0.129	12	-0.076	11	0.042	14	-0.033	20	0.167	8	-0.075	15
商务住宅	-0.073	5	-0.094	12	0.072	10	0.152	6	0.07	14	-0.018	5
政府机构及社会团体	-0.124	11	-0.18	20	0.082	6	0.135	8	0.224	2	-0.11	20
科教文化服务	-0.202	19	-0.173	19	0.068	11	0.067	18	0.22	3	-0.095	17
交通设施服务	-0.173	15	-0.076	10	0.111	4	0.089	13	0.17	7	-0.066	13
金融保险服务	-0.214	20	-0.057	8	0.114	3	0.094	11	0.216	4	-0.105	19
公司企业	-0.12	10	-0.144	17	0.075	7	0.069	17	0.117	9	0.017	2
道路附属设施	-0.008	4	-0.039	7	0.031	16	0.092	12	-0.036	20	-0.068	14
地名地址信息	0.015	2	0.039	1	0.016	17	0.084	16	0.021	17	-0.079	16
公共设施	-0.102	8	-0.068	9	-0.048	20	0.214	1	-0.036	19	-0.044	9

4.2.2 居民出行时间流量特征

由 EM 算法聚类得出的簇（C0-C5）在一周内的工作日、休息日出行时间流量特征（上下车人数），如图 9-12 所示：

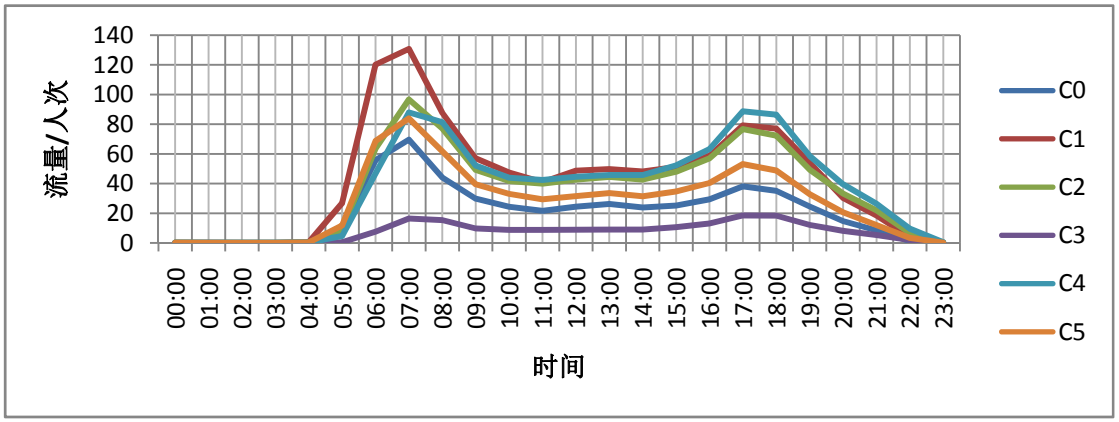

图 9　EM 聚类所得各簇的工作日上车流量

Fig.9 The inflows on weekdays of clusters clustered by EM

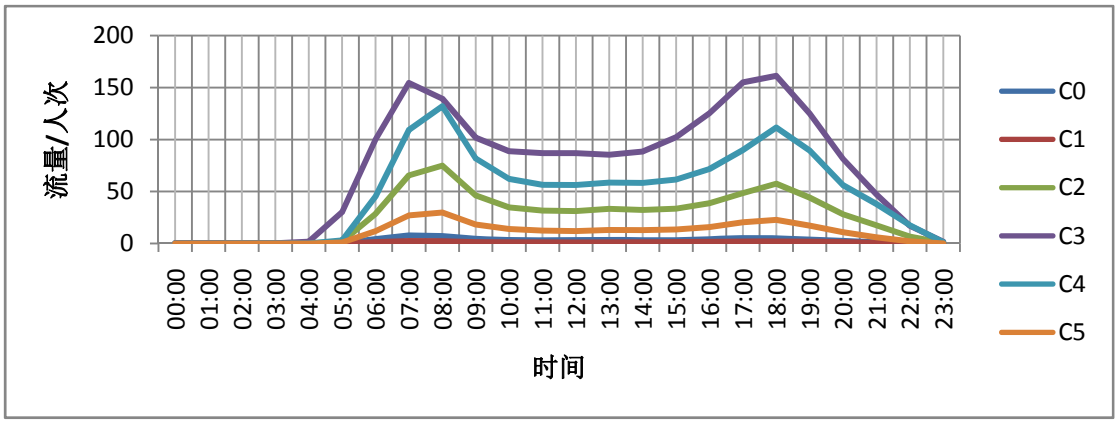

图 10　EM 聚类所得各簇的工作日下车流量

Fig.10 The outflows on weekdays of clusters clustered by EM

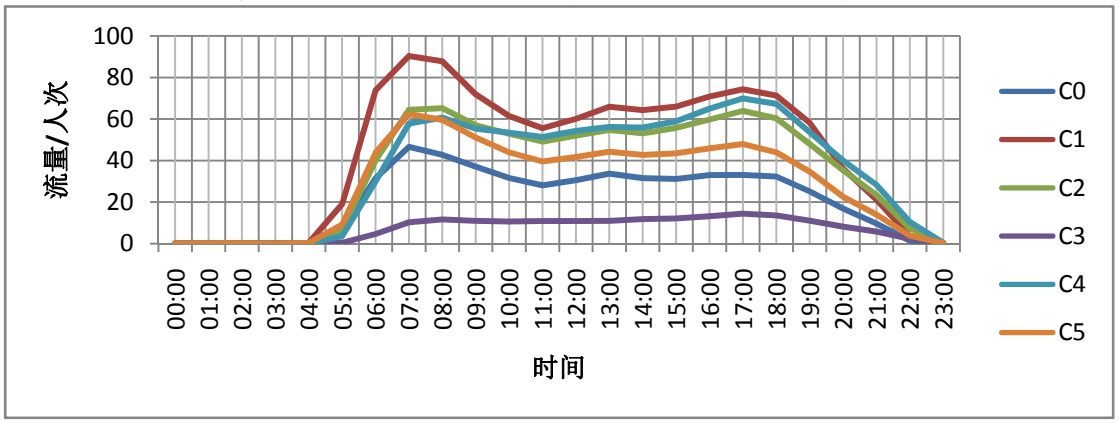

图 11　EM 聚类所得功能区的休息日上车流量

Fig.11 The inflows on weekends of clusters clustered by EM

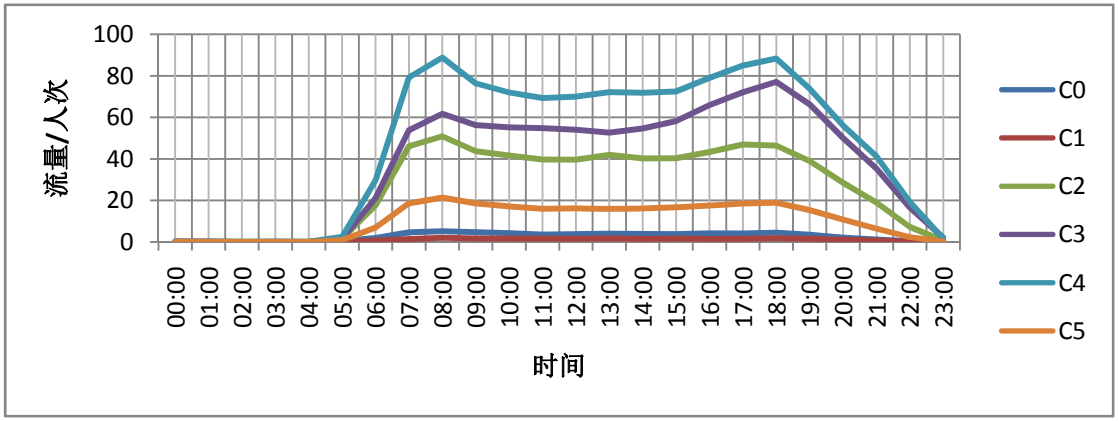

图 12　EM 聚类所得功能区的休息日下车流量

Fig.12 The outflows on weekends of clusters clustered by EM

4.2.3 识别结果

针对 EM 算法聚类结果进行功能识别：

a.成熟居住区(C0)

该区域内住宅兴趣点分布较广，商务住宅比例较高（FD 值为-0.073），且为

居民生活服务的医疗卫生服务、住宿服务和教育服务等兴趣点配套均衡，是典型的居住区兴趣点分布。

同时通过一周流量数据分析，我们可以发现该区域工作日最大的出发流量高峰是在清晨（7-8 点的上班时段），返程流量高峰出现在傍晚（17-19 点的下班时段），是典型的居住区出行模型。

b.待开发区（C1）

该区域兴趣点主要为摩托车、汽车服务，分布较多的汽车 4S 店、摩托车销售、汽车摩托车维修点等，周边基础设施建设尚不健全。

c.风景名胜区(C2)

该区域分布比例最高的兴趣点是风景名胜点，在该类别相较于其他区域具有较高 FD（Frequency Density）值（FD 值为 0.042）。同时为游客服务的餐饮服务、住宿服务等在外部排名中也较靠前，而且该区域工作日、休息日出行流量差距不大，每天不同时段出行相对平均，休息日出行流量高于工作日流量。

d.商业娱乐区(C3)

该区域兴趣点数据分布特征可以看出：餐饮服务、购物服务、生活服务的 FD 值较高，在所有簇中分别排名第 2、第 1 和第 1。与此同时，餐饮购物信息点在区域内部所有信息点中占比例较高（CR 值较大）。例如，餐饮服务类的簇内 CR 值排名第 7，购物服务类的簇内 CR 排名第 3。同时，通过流量特征图，可以看出该区域工作日下班时段（17:00-19:00）会出现下车流量高峰，说明很多居民在该区域消费购物，以及参加休闲娱乐活动。

e.公共管理及科教文化区(C4)

该区域分布比例最高的兴趣点是政府机构及社会团体，相较于其他区域具有较高 FD（Frequency Density）值(FD 值为 0.22)，该类型兴趣点占该区域内中兴趣点数的 9.7%，RCR 值排名为第二。并且该区域内科教文化兴趣点较多。同时交通服务设施、体育休闲服务、住宿服务等在外部排名中也较靠前。

f.新兴居住区（C5）

该区域的兴趣点数据结构和 C0 类似，按照各类兴趣点数量占该区域内中兴趣点总数的比例进行排名，住宅类位列第 5，同时该区域内医疗保健第 3、生活服务类第 7，是典型的居住区兴趣点分布结构。

另一方面，从工作日、休息日站点流量数据可以看出，该区域工作日最大的出发流量高峰是在清晨（7-8 点的上班时段），返程流量高峰出现在傍晚（17-19 点的下班时段），是典型的居住区出行模型。但是该区域流量相对于 C0 区域较少，日流量为 C0 的 1/4 左右，这说明该区域人流量不大，尚处在发展阶段。

g.未分类区域（Sparse）

由于山地、森林、河流等原因，部分区域无公交流量数据，本文将该类区域归为一类。

根据功能区识别结果，对各功能区内的面积、人口进行了统计，如表 3 所示。

表 3 各功能区信息统计

Tab.3 Information of each cluster

功能区	TAZ 个数	面积（km^2）	人口（人）
未分类区域（Sparse）	357	11061.13	2561345
成熟居住区(C0)	63	326.2918	572609
待开发区（C1）	25	82.10567	162647
风景名胜区(C2)	155	973.0869	1878303
商业娱乐区(C3)	129	2068.882	1911066
公共管理及科教文化区(C4)	267	918.993	3551985
新兴居住区（C5）	122	974.2913	1271898

4.3 识别结果的检验

为了检验 DZoF 模型识别结果的准确性，本文将实验得到的北京市不同功能区域图与北京市城市总体规划（2004-2020）中的用地现状图[5]以及谷歌地图进行对比。其中，若干典型地区的对比结果如表 4 所示。

表 4 识别结果与现状图对照分析

Tab.4 Comparison and analysis between Recognition results and the status quo of figure

对照区域	北京市著名的风景名胜区——十渡风景区

[5]数据来源于北京市规划设计研究院

对照图	
识别结果	识别图中 A 区域为风景名胜（绿色）与现状图 A 区域相一致
对照区域	北京市文物保护单位——岔道城遗址（八达岭关城西北 3 华里）
对照图	
识别结果	识别图 B 区域为风景名胜区，按照地理位置刚好对应岔道口遗址风景区
对照区域	永定河北京市段流域形成了一个三角洲，以林地为主。
对照图	
识别结果	C 区域在识别图中为未分类区域，主要为河流、山地、森林等，与现状图中 C 区域所表示的永定河三角洲相符合。
对照区域	海淀区高校云集，著名的北京大学、清华大学、中国人民大学等均位于海淀区。同时又有中关村等，是北京的科教文化区。

对照图	
识别结果	D 区域在现状图中为科教文化区（紫色），与识别图中的 D 区域相互对应。
对照区域	东城区部分
对照图	
识别结果	识别图中标注的 E 区域主要为风景名胜区，F 为商业娱乐区，通过和该区域的谷歌地图对照，E 区域实际是北京日坛公园，而 F 区域内则汇集了国贸商城、国贸饭店、嘉里商场、万达广场等多个购物娱乐中心，是北京著名的商业娱乐区。

此外，我们还将研究结果与详细的北京市各交通分析小区土地利用数据进行分析对比，以检验识别的总体准确率。根据公共用地面积（具体包含公共设施用地和市政用地）的大小对 1118 个交通分析小区进行排序，选取前 50 个公共交通分析小区，除去无公交 IC 卡刷卡信息的分析小区，共有 44 个分析小区。其中有 22 个被模型识别为公共管理及科教文化区，准确率达 50%。采用同样的方法，对居住用地进行分析对比，准确率为 58.06%，对比结果如表 5 所示。

表 5 识别结果与交通分析小区用地情况对比分析

Tab.5 Comparison and analysis between Recognition results and land use of TAZ

		有效对照数据	识别结果	准确率
公共管理及科教文化区	交通分析小区数目	44	22	50.00%
	总面积（m^2）	59596348	19824488	
居住区	交通分析小区数目	31	18	58.06%

	面积（m^2）	49319636	15423667	

综合考虑北京市商住、产住高度混合的用地现状与研究对比分析，DZoF 模型对于北京市主要的功能区能有效地加以识别，具有一定的准确度。

5.结论和讨论

本研究基于北京市 2008 年 4 月连续一周的 77976010 条公交 IC 卡刷卡数据和北京市 2010 年 113810 条兴趣点数据，通过构建 DZoF 模型，进行了北京市城市功能区的识别，共得到 7 个类别的功能区，分别为公共管理及科教文化区、风景名胜区、商业娱乐区、成熟居住区、新兴居住区和尚未分类区域。其中公共管理及科教文化涵盖交通分析小区（TAZ）267 个，总面积 918.993 平方公里；商业娱乐区涵盖交通分析小区（TAZ）129 个，总面积 2068.882 平方公里；风景名胜区涵盖交通分析小区（TAZ）155 个，总面积 973.0869 平方公里。

研究结果显示，DZoF 模型对于北京市城市功能区特征具有一定的识别能力。本研究能够更好的帮助人们轻松地理解一个复杂的城市的空间功能结构，辅助城市规划者基于人类活动（human mobility）和兴趣点数据开展不同城市功能区的规划，对城市规划具有指导和参照价值，同时也可以为房地产开发的选址提供重要的决策支持。

本研究主要有三方面的潜在创新：第一，基于海量的公交 IC 卡刷卡数据，通过人类出行活动来研究城市空间结构。第二，将城市研究的传统方法与大数据挖掘相结合，即从已有的城市居民出行调查数据识别出居民的出行行为特征，再将这些特征作为规则应用到城市区域的功能识别上；第三，构建了 DZoF（识别城市功能区）模型，对城市不同区域的功能进行了识别研究。总体上，本文提出的基于城市兴趣点和公交 IC 卡刷卡数据，运用数据挖掘的方法进行城市空间结构动态研究，为大都市空间结构研究提供了一种新的分析思路和研究方法。

本研究仍有一些不足之处需要在未来的进一步研究中加以改进：（1）本文研究在数据模型构造时选择了上下车人数作为指标，由于一票制公交车无法得到乘客下车数据，所以研究中忽略了一票制公交车乘客刷卡数据，未来，可以通过构建新的数据模型，有效地利用一票制公交车乘客刷卡数据，提高识别结果的准确性。(2) 2010 年北京市公共汽（电）车出行比率为 28.2%，轨道交通出行比率占

11.5%，出租车出行比率为 6.6%，小汽车出行比率为 34.2%[6]，公交数据具有一定的局限性和片面性。在未来研究中，将进一步通过增加轨道交通和出租车数据，完善人类活动信息，得到更为可靠准确的结果。（3）现实情况中，商住、产住等混合用地是广泛存在的。本研究将实验结果在 TAZ 尺度上汇总时，忽略了混合用地的因素，选取比率最大的功能作为该区的功能表征。在未来的研究中，也可考虑加入混合用地分类。

参考文献

[1] Lai S-K, Han H. Complex: the new ideas of urban planning. Beijing: china-building Press,2009,9.10-13 [赖世刚, 韩昊英. 复杂：城市规划的新观点. 北京：中国建筑工业出版社, 2009,9 .10-13.]

[2] Batty M. Cities as Complex Systems: Scaling, Interactions, Networks, Dynamics and Urban Morphologies Springer. Berlin, DE, 2009.

[3] Batty M. Invisible cities. Environment and Planning B: Planning and Design, 1990(17):127-130.

[4] Neuhaus F, "Urban diary-a tracking project" UCL working paper series. Paper 151. Available on line: http://discovery.ucl.ac.uk/19245/

[5] Calabrese F, Ratti C. Real time Rome. Networks and Communication Studies, 2006, 20:247-258.

[6] Ratti C, Pulselli R M, Williams S, Frenchman D. Mobile landscapes: Using location data from cell phones for urban analysis. Environment and Planning B: Planning and Design 2006, 33(5):727-748.

[7] Gonzalez M C, Hidalgo C A, Barabasi A L. Understanding individual human mobility patterns. Nature, 2008, 453: 779-782.

[8] Goodchild M F, Janelle D. The city around the clock: Space-time patterns of urban ecological structure. Environment and Planning A, 1984, 16:807-820.

[9] Goodchild M F, Klinkenberg B, Janelle D G. A factorial model of aggregate spatio-temporal behavior: Application to the diurnal cycle. Geographical Analysis, 1993, 5:277-294.

[10] Jiang B, Yin J, Zhao S. Characterizing human mobility patterns in a large street network. Physical Review E, 2009, 80(2):1136-1146.

[11] Hamilton B W. Wasteful Commuting. The Journal of Political Economy, 1982,90(5):1035-1053.

[12] Liu Zhilin, Wang Maojun. Job accessibility and its impacts on commuting time of urban residents in Beijing: From a spatial mismatch perspective. Acta Geographica Sinica, 2011, 66(4): 457-467. [刘志林, 王茂军. 北京市职住空间错位对居民通勤行为的影响分析: 基于就业可达性与通勤时间的讨论. 地理

[6]数据来源于《2011 北京市交通发展年度报告》

学报, 2011, 66(4): 457-467.]
[13] Wang D, Chai Y. The jobs-housing relationship and commuting in Beijing, China: the legacy of Danwei. Journal of Transport Geography, 2009, 17: 30-38.
[14] Anas A, Arnott R, Small K A. Urban Spatial Structure. Journal of Economic Literature, 1998, 36(3) 1426-1464.
[15] McMillen D P, McDonald J F. A Nonparametric Analysis of Employment Density in a Polycentric City. Journal of Regional Science, 1997, 37(4): 591-612.
[16] Lu D, Weng Q. (2005). Urban land-use and land-cover mapping using the full spectral information of Landsat ETM+ data in Indianapolis, Indiana. Photogrammetric Engineering & Remote Sensing, 2005, 71(11):1275-1284.
[17] Xiao J, Shen Y, Ge J, et al. Evaluating urban expansion and land use change in Shijiazhuang, China, by using GIS and remote sensing. Landscape and Urban Planning, 2006, 75(1-2), 69-80.
[18] Qi G, Li X, Li S, et al. Measuring Social Functions of City Regions from Large-scale Taxi Behaviors. The 9th IEEE International Conference on Pervasive Computing and Communications (PerCom'11), Work in Progress, Seattle, USA: March 21-25, 2011:384–388.
[19] Liu Y, Wang F H, Xiao Y, et al. Urban land uses and traffic ‘source-sink areas’: Evidence from GPS-enabled taxi data in Shanghai. Landscape and Urban Planning, 2012, 106:73-87.
[20] Pulliam H R. Sources, sinks, and population regulation. American Naturalist, 1988, 132:652-661.
[21] Yuan J, Zheng Y, Xie X. Discovering Regions of Different Functions in a city Using Human Mobility and POIs. The 18th ACM SigKdd Conference on Knowledge Discovery and Data Mining, Beijing, China: August 12-16, 2012.
[22] Sun L, Lee D, Erath A. Using Smart Card Data to Extract Passenger’s Spatio-temporal Density and Train’s Trajectory of MRT System. ACM SIGKDD International Workshop on Urban Computing, Beijing, China : August 12, 2012.
[23] CH J, Hwang C. A Time-geographic analysis of trip trajectories and land use characteristics in Seoul metropolitan area by using multidimensional sequence alignment and spatial analysis. 2010 AAG Annual Meeting, Washington, DC: 2010.
[24] Long Ying, Zhang Yu, Cui chenyin. Identifying Communing Pattern of Beijing using Bus Smart Card Data. Acta Geographica Sinica,2012,67(10):1-12.[龙瀛,张宇,崔承印.利用公交刷卡数据分析北京职住关系和通勤出行.地理学报，2012, 67(10):1-12.]
[25] Zhao Weifeng, Li Qingquan, Li Bijun. Using urban POIs data extraction hierarchical landmarks.Journal Of Remote Sensing,2011(05):981-989.[赵卫锋,李清泉,李必军.利用城市POIs数据提取分层地标.遥感学报,2011(05):981-989.]
[26] Witten I H, Frank E. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann, 1999.2-3.
[27] Tan P-N, Steinbach M, Kumar V. Introduction to Data Mining. Post&Telecom Press,2006,2-7.
[28] Jiang S, Ferreira J, Gonzalez M C. Clustering daily patterns of human activities

in the city. Data Mining and Knowledge Discovery, April 23,2012,1-33.
[29] Sun J.B, Yuan J, Wang Y, et al. Exploring space-time structure of human mobility in urban space. Physical A,2011,390:929-942.
[30] Jiang S, Ferreira J, Gonzalez M. Discovering Urban Spatial-Temporal Structure from Human Activity Patterns. UrbComp'12, August 12, 2012. Beijing, China.
[31] Pan G, Qi G, Wu Z, et al. Land-use Classification Using Taxi GPS Traces. IEEE Transactions on Intelligent Transportation Systems (T-ITS), 2013, 14(1):113-123.
[32] Agrawal R, Raloutsos C, Swami A. Efficient Similarity Search in Sequence Databases. In Proc of the 4th International Conference on Foundations of Data Organization and Algorithms, Berlin, Germany, Springer-Verlag, 1993:69~84.
[33] Sun Wenshuang, Chen Lanxiang. Multivariate statistical analysis. Beijing: Higher Education Press,1994.[孙文爽, 陈兰祥. 多元统计分析. 北京: 高等教育出版社, 1994.]
[34] Korn F, Jagadish H V, Faloutos C. Efficiently Supporting Ad Hoc Queries in Large Datasets of Time Sequences. In Proc. of the ACM SIGMOD International Conference on Management of Data, Tucson, USA, ACM, 1997:289~300.
[35] Jiang Yuan, Zhang Zhaoyang, Qiu Peiliang. Clustering Algorithms Used in Data Mining. Journal of Electronics & Information Technology, 2005.27(4):655-659.[姜园,张朝阳,仇佩亮等.用于数据挖的聚类算法.电子与信息学报,2005.27(4):655-659.]
[36] Huang Libin. City rail transit sites influence evaluation research of regional development—In Shanghai xujiahui rail hub. Tongji University,2006.03.[[黄丽彬. 大城市的轨道交通站点对地区发展的影响评价研究——以上海徐家汇轨道枢纽为例. 同济大学, 2006.03].

Discovering Functional Zones Using Bus Smart Card Data and Points of Interest in Beijing

HAN Haoying[1], YU Xiang[1], LONG Ying[2]
（1.College of Public Administration, Zhejiang University, Hangzhou 310029,China；
2.Beijing Institute of City Planning, Beijing 100045,China）

Abstract: Cities comprise various functional zones, including residential, educational, commercial zones etc... It is important for urban planners to identify different functional zones and understand their spatial structure within the city in order to make better urban plans. In this research, we used 77976010 bus smart card records of Beijing City in one week in April 2008 and converted them into two-dimensional time series data of each bus platform, Then, through data mining in the big database system and previous studies on citizens' trip behavior, we established the DZoF (Discovering Zones of different Functions) model based

on SCD (Smart Card Data) and POIs (Points of Interest), and pooled the results at the TAZ (traffic analysis zone) level. The results suggested that DzoF model and cluster analysis based on dimensionality reduction and EM (expectation-maximization) algorithm can identify functional zones that well match the actual land uses in Beijing. The methodology in the present research can help urban planners and the public understand the complex urban spatial structure and contribute to the academia of urban geography and urban planning.

Key Words: bus smart card data (SCD); POIs; functional zones; human mobility; Beijing