基于潜在语义信息的城市功能区识别
——广州市浮动车GPS时空数据挖掘

陈世莉[1,2,3],陶海燕[1,2],李旭亮[1,2],卓 莉[1,2]

(1. 中山大学地理科学与规划学院 综合地理信息研究中心,广州 510275; 2. 广东省城市化与地理环境空间模拟重点实验室,广州 510275; 3. 中山大学城市化研究院,广州 510275)

摘要: 随着中国城市化进程的不断推进和深入,城市内部空间结构正发生不断的变化。城市内部形成的不同功能区标识研究,对城市结构理论以及政策制定、资源配置等方面具有非常重要的意义。这些不同的功能区包括住宅区、工业区、教育区以及办公区等。本文以大数据为依托,重点研究城市功能区的特点和分布状态,选取广州市6个区为样本,以最新道路网络为分割依据把研究样本分为439个区域。对历时一周的海量浮动车(GPS)数据以及兴趣点数据采用时空语义挖掘方法,建立潜在的狄利克雷模型(LDA)以及狄利克雷多项式回归模型(DMR);通过OPTICS聚类方法对不同模型的结果进行聚类,进而利用POI类别密度、居民出行特征等方法进行分区结果识别。同时,参考百度地图的地理信息,将研究得到的广州市功能分区结果与广州市城镇用地现状图、居民日常出行特征进行对比验证分析。研究表明,该方法基本能识别出具明显特征的城市功能区,如成熟居住区、科教文化区、商业娱乐区、开发区等。识别出的广州市不同类型的功能区呈现了以居住区和商业区为主导,其他类型功能区围绕其展开的特点。研究证明,利用大规模、高质量的个体时空数据开展人们移动行为和日常活动组织及社会空间的研究,能从一个新的视角揭示城市功能区的形成及其机制。

关键词: 主题模型;功能区;地理大数据;GPS数据;兴趣点;广州
DOI: 10.11821/dlxb201603010

1 引言

城市化是重要的全球性社会经济现象，而城市化中的功能多元化则是城市发展的基础。功能多元化形成的城市功能分工[1]，为人们的居住、工作、游憩和交通提供着各方面的便利。传统的城市功能分区研究大多是采用土地利用类型现状图、调查问卷等数据，利用各种聚类或者建立指标体系等方法来对城市功能区域进行划分[2-12]。数据的获取不仅耗时耗力效率低，且确定权重系数时主观因素太强。随着信息和通讯技术的发展，城市中大量的传感器，例如，全球定位系统（Global Position System, GPS）、全球移动通信系统（Global System for Mobile, GSM）、智能卡收费系统数据（Smart Card Data, SCD）等，可以获取大规模的、高质量的个体时空数据。这些移动定位数据获取成本低，覆盖

收稿日期: 2015-07-30; **修订日期:** 2015-11-27
基金项目: 广国家高技术发展计划(863) (2013AA122302); 广东省自然科学基金项目(S2013010012554); 国家自然科学基金项目(41371499, 41271138)[**Foundation:** Projects of National High-tech Research, No.2013AA122302; Natural Science Foundation of Guangdong Province, No.S2013010012554; National Natural Science Foundation of China, No.41371499, No.41271138].
作者简介: 陈世莉(1990-), 女, 四川达州人, 博士, 主要研究方向为空间数据挖掘、城市空间结构研究。E-mail: SLChen@126.com
通讯作者: 陶海燕(1966-), 女, 江苏扬州人, 副教授, 主要研究方向为多智能体地理模拟、空间数据挖掘。E-mail: taohy@mail.sysu.edu.cn

范围广，且具有时态特性。因此，结合大数据和数据挖掘，通过人们日常行为反映城市功能分区，可以为城市研究提供一种新的方法和途径。

近年来，大量国外学者使用手机数据、公交车数据[13-15]、GPS数据[16]以及LBS（Location Based Service）数据[17-21]开展土地利用分类、城市功能区划、活动轨迹等方面研究。同时，随着国内智慧城市建设的推广以及大数据相关研究的深入，相继有学者[16, 22-30]采用GPS数据和公交车数据，通过Google软件、数据挖掘工具、建立各种模型等来研究城市空间结构。

开展城市功能分区的研究有利于合理、健康地规划未来城市。近年来，已有学者采用大数据开展人们出行规律和特征等方面的研究[21-25, 31]，而通过挖掘大数据中的海量潜在语义信息来辅助开展城市空间结构中的城市功能分区的研究还相对较少。因此本文拟采用在文本分类领域中能快速挖掘出海量文本中潜在语义的主题模型（LDA模型和DMR模型），以广州市为例，提取浮动车轨迹数据和兴趣点数据的潜在语义信息，然后，通过OPTICS聚类方法对不同模型处理后的结果进行聚类，并对聚类后的类别进行识别。研究结果表明，该方法能有效地识别广州市不同类型的功能区，对城市规划、政策制定、资源配置等各个领域具有辅助和引导价值。

2 研究方法和思路

潜在的狄利克雷分布（Latent Dirichlet Allocation, LDA）是一种主题模型，由Blei等于2003年提出[32]。它是当前文本处理研究的范式之一，可以对文字隐含主题进行建模，不仅弥补了信息检索中传统的文本相似度计算方法的不足，并且更适合基于大规模语料库（乃至于海量互联网数据）寻找文字间的语义主题[32-33]。此模型能确定一个语料库中的每一篇文档有多个主题的概率，能充分提取出一句话或者一个单词的语义信息。

根据LDA模型理论，语料库中的每一篇文档都可以看作是由多个主题混合产生的，文档的每一个单词都是来自一个主题。当给定一个文档的所有单词的情况下，可以通过数学推导得到文档对应的主题分布。此模型能确定一个语料库中的每一篇文档属于多个主题的概率，能充分提取出一句话或者一个单词的语义信息。

LDA的计算模型[31]如图1所示。每个结点表示一个随机变量，并且根据其在生成过程中的角色予以标记，白色表示隐藏变量，灰色表示观测变量，矩形框内的变量表示需要循环计算，而箭头则表示参数和变量作用的方向。

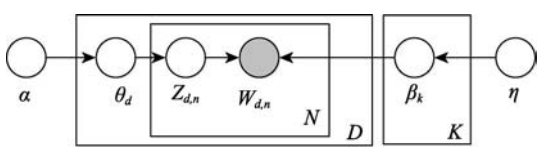

图1 LDA模型

Fig. 1 Latent Dirichlet Allocation model

这里的α和η分别为狄利克雷文档—主题分布和主题—单词分布的输入参数。假设文档有K个主题，β是一个$K\times M$的矩阵，其中M是字典中的单词数（语料库D中的所有单词）。每一个β_k是字典上的一种分布。每一个文档d的主题比例为θ_d，且$\theta_{d,k}$是主题K在d文档中的主题比例值。文档d的主题分布为Z_d，$Z_{d,n}$是第n个单词在文档d中的主题分布。文档d中的观察单词为W_d，$W_{d,n}$表示n个单词在文档d中。

显然，城市功能分区研究和文本表示研究具有一定的相似性。把一个区域映射为一篇文档，区域中的功能看作一个主题，每个区域内的居民出行活动当作单词，一个功能区域可以看作居民出行活动的集合，它包括交通设施、活动的人、相互作用等的输入[34]。

狄利克雷多项式回归（Dirichlet Multinomial Regression, DMR）模型因其参数中需要

输入先验数据，使得实验结果更贴近现实情境，与基本模型LDA相比更有优势[35]。DMR模型中的先验α对每一个区域的α_i而言，结合了每一个区域的POI特征项矢量。例如：$\alpha_{i,k}=\exp(X_i^T\lambda_i)$。因此，对于不同的POI类别分布，都会有一个$\alpha$值的分布。所以，活动类型的分布是POI特征项和移动模式的总和。最后，通过应用DMR模型，输入移动模式和POI特征项，获得了每一个区域的活动分布以及每一次活动的移动模式分布。

而本文中，移动模式M用一个元组表示：

$$M=<O_i, T_i, D_j, T_j> \tag{1}$$

式中：O_i表示起始区域；T_i表示从起始区域出发的时间；D_j表示目的区域；T_j表示到达目的区域的时间。如果一个人在T时刻从S区域出发，到达目的区域X，则用字符串“O_X_T”表示起始区域的移动模式，其中字符“O”代表起始区域S；用字符串“S_D_T”表示目的区域的移动模式，其中字符“D”代表目的区域X。

研究的实现步骤如下：

（1）提取出海量浮动车数据中的起讫点（O/D）对，并通过公式（1）实现了起讫点（O/D）对与移动模式转换的一一映射；

（2）基于Mallet平台[36]建立LDA模型和DMR模型对移动模式进行处理；

（3）LDA和DMR是非监督学习模型，本身不能用于分类，需要嵌入合适的聚类算法，所以对DMR模型处理后得到的每个功能区每种主题的概率分布，采用OPTICS聚类方法得到城市功能分区的结果；

（4）利用POI类别密度方法、居民出行特征以及问卷调查结果对分区结果进行标识；

（5）利用土地利用现状图对识别结果进行精度验证，调查问卷中某区域的研究结果与功能分区相对应的空间位置的区域特征进行对比分析，分析理论与实际的差异以及背后的原因。

3 研究数据和研究区域

研究数据由广东瑞图万方科技股份有限公司提供，其数据说明参见表1。在保证研究代表典型性的前提下，综合考虑研究对象的空间分布和时间演化，采用2014年广州行政区划调整前的范围，即传统的广州市老八区（其中芳村区已并入荔湾区，东山区则并入越秀区）[37]。同时，选取了高速公路、城市快速道、国道、省道、城市主干道、城市次干道6个等级道路把广州市老城区分割为439个区域。然而，又因华南快速干线上部的白云区部分区域主要以风景区、农田及山体组成，所以，本文将该区域独立出来，直接视为风景区，而不纳入研究范围。

表1 研究数据说明

Tab. 1 Data specification

数据	年份	数量	详细说明
广州市区划图	2014年		选取广州市白云区、海珠区、黄浦区、越秀区、荔湾区以及天河区为研究对象
道路网络数据	2014年	1028条	选取高速公路、快速路、国道、省道、城市主干道及城市次干道的道路功能等级为道路分割对象
兴趣点数据	2014年	182504个	研究对象范围内的所有兴趣点
浮动车轨迹数据	2013年	2万辆(1.315亿条)	选取研究对象范围内的所有GPS点并提取其OD点对进行分析
居民出行调查问卷	2013年	1152份	调查问卷的对象、时间、主题

兴趣点（Point Of Interest, POI）是导航、智能交通、基于位置服务等应用中一种重要的基础数据。本文所选取的POI数据的类别大类划分为15种，中类划分为65种。其中，每一个POI点，编号由大类+中类+序号组成，包括名字、经度、纬度等属性。根据本文的研究目的和需求，并经过多次试验，将POI数据合并为29个类别。需要说明的是本文中暂未考虑同一类别中不同等级兴趣点对结果可能造成的影响。

浮动车轨迹数据每隔5s左右采集一次，记录的基本信息包括出租车的车牌号码、时间、经纬度、速度、方位和载客状态等。本文采用2013年10月8日（周二）到10月14日（周一），历时一周的GPS数据。

4 基于语义信息的广州市城市功能区识别

4.1 城市功能分区结果

应用研究的实现步骤，得到9种功能分区（F0~F8）结果（图2，图3）。对比不同颜色代表的不同功能区（两图中相同的颜色可能代表不同的功能类型），可见LDA模型方法与DMR模型方法得出了大致相同的功能区域的结果，而后者在部分细节区域的区分效果更显著，例如广交会展馆和五山校区。因此本文中选择DMR模型方法结果进行聚类和识别。

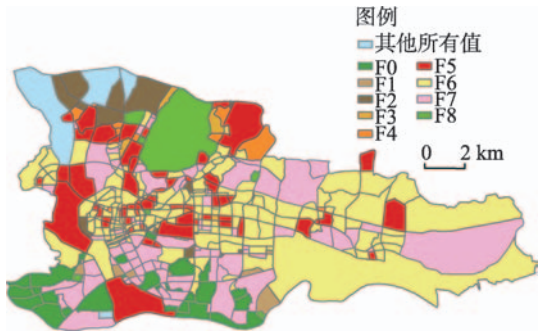

图2 LDA模型结果

Fig. 2 The result of LDA model

4.2 城市功能分区识别方法

通过建模实现相同功能的区域的聚类，之后需要根据区域的实际功能进行标识。城市功能区的分类标准很多，功能区的划分也各不相同，本文主要根据社会功能、居民需求等进行划分，主要分为：科教文化区、居住区、商业娱乐区、开发区等，并根据人类活动的密集程度对部分功能区细化为成熟区和新兴区。

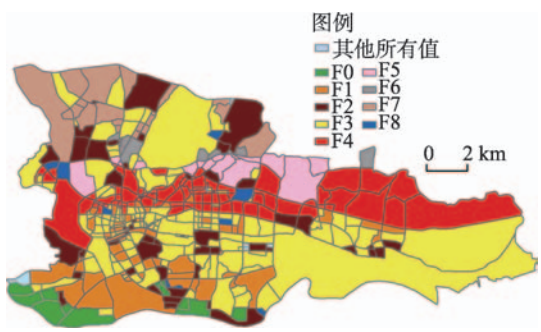

图3 DMR模型结果

Fig. 3 The result of DMR model

在这里主要是根据已有数据识别功能区，通过以下三个方面来识别城市的一个功能区：

（1）计算得出每种功能区的FD（Frequency Density），即通过计算每一类POI在每类功能区中的密度，并对其进行内部索引排序即IR（Internal Ranking），从而得到该类型区域内POI的分布特征，推测该类型区域能实现的功能。其中第 j 类的POI类别在各个区域中的密度 p_j 计算公式如下：

$$p_j = \frac{\text{每类}POI\text{类别的数目}}{\text{每种功能区的各个区域面积总和}} \tag{2}$$

其结果如表2所示。F0至F8表示不同类型的功能区，FD为POI密度，IR为单一功能区中不同POI密度的排序索引，越靠前密度值越大。颜色深浅则表示单一POI种类在不同功能区中的分布情况，较深则代表集中分布，反之亦然。

（2）通过每个功能区周末和工作日的移动模式频率，即离开和到达该功能区的频次，探索该类型区域人群移动规律，从而推测功能区类型，其结果如图4所示。

另外，利用Matlab（2014a）对浮动车的起讫点进行计算，得到从早上6:00开始到晚上24:00结束的时间段内，工作日和周末到达以及离开某种功能区的热点图。横坐标表示一天内时间变化，纵坐标代表不同的功能区，颜色越深表示出现的频次越高，反之亦然（图5，图6，图7）。

（3）经验标注。对于一个长久居住在一个城市中的经验者，他们非常清楚地知道城市的地标建筑和最能体现城市特色的区域，例如，区域中包含的琶洲国际会展中心，大家肯定会认同这个区域是一个会展区。在实验中完成聚类，得到城市功能区后，有经验的人们可以帮助我们更好地标注、识别功能区，可以得到更详细、准确的城市功能分区结果。

4.3 识别结果

利用以上城市功能区的识别方法，结合POI密度排序（表2）、周末/工作日每个小时的流量特征（图4）以及工作日/周末的到达/离开每个功能区的热点图（图5，图6，图7）的数据，发现F1~F5功能覆盖齐全，进出人流量持续较大，说明F1~F5建设比较完善，从而得出其为成熟建成区；相应地，F0、F6~F7为欠发达或正待开发地区。而F8因其功能区域内部的POI数目和GPS数据过少，无法正确地识别功能区的类别，因此，本文未对F8区域进行识别。下面对聚类所得结果进行功能识别：

（1）花卉园林[F0]。这个功能区大都由栽培花卉园林的种植地、花卉园林的公司、湿地公园以及其他景区（例如，宗祠）组成。从F0的FD值可以看出（表2），园艺销售、林场生产在这个功能区中占据较高比例。所以，本文把F0归为以花卉园林生产为主要特色的欠发达区域。虽然其地方偏僻，但仍有少量人前往，同时城中村中的居住人口相对拥挤，这也导致这个功能区具有早晚高峰出行量较多的现象。

（2）成熟专业批发区[F1]。这个功能区包含了广州市具有特色的批发市场和销售市场，例如，广州国际轻纺城即专业布匹及辅料批发市场，特别地，这个功能区中还包括了作为汽车销售、建材销售的赛马场。但围绕批发区含有居民居住以及生活的各种完善的配套设施，这个功能区属于城市建设、规划比较早的区域，其本身有多种功能（表2，图5），但区域中的居民仍然有对外活动以及出行的需求，所以，其对外活动相对较多。

（3）风景名胜地[F2]。这个功能区与F1有相似之处，都是比较成熟的城市规划建成区（表2）。同时，这个功能区有很多风景名胜和地标建筑，例如，广州塔、海心沙、奥林匹克体育中心、世界大观、华南植物园等，进出口交易中心（广交会展馆）。这个区域工作日相对F1而言，早上基本没有出行高峰期，周末10:00-14:00之间却有较高的到达峰值（图4a、4b），这与表2中会展中心、旅游业等POI类别排序比较靠前的结果相一致。

（4）成熟居住、文化区[F3]。这个功能区不仅具有大量成熟、高级的居住区并且还含有广州市传统文化景区，例如，白云公园、烈士陵园等，同时还有中山大学的南校区、北校区以及一些综合性医院和大量生活服务设施。文化教育在POI密度排序中的比值较大（表2）；工作日的早上7:00-10:00是居民出行的高峰期，晚上22:00是居民搭乘出租车回到住宿的高峰期；周末的早上的10:00-13:00是居民出行的高峰期，而23:00是居民回程的峰值（图4a、4b）。

（5）成熟商业娱乐区[F4]。该功能区具有以大厦、写字楼、批发销售市场、餐饮以及购物等为主的兴趣点数据分布特征，同时其中包含较多的居民区分布。广州典型商业

表2 POI密度和功能区排序

Tab. 2 Overall POI density vector and ranking of functional zones (FD: Frequency Density, IR: Internal Ranking, F: Functional Zones)

POI类别	F0		F1		F2		F3		F4		F5		F6		F7		F8	
	FD	IR	FD	IR	FD	IR	FD	IR	FD	IR	FD	IR	FD	IR	FD	IR	FD	IR
国际国外组织机构	0.0000	23	0.0008	26	0.0007	26	0.0005	28	0.0026	24	0.0000	27	0.0000	22	0.0000	25	0.0000	25
技术开发	0.0000	24	0.0035	23	0.0032	21	0.0025	21	0.0085	18	0.0022	22	0.0047	19	0.0002	23	0.0393	16
综合性大学	0.0000	25	0.0049	20	0.0054	19	0.0033	20	0.0060	21	0.0090	15	0.0000	23	0.0007	21	0.0187	19
综合医院	0.0000	26	0.0037	22	0.0004	28	0.0023	23	0.0012	25	0.0004	25	0.0000	24	0.0000	26	0.0075	23
广播电台	0.0005	22	0.0031	25	0.0015	23	0.0012	25	0.0028	23	0.0015	23	0.0000	25	0.0002	24	0.0019	24
会展中心	0.0000	27	0.0002	28	0.0011	25	0.0006	27	0.0002	29	0.0000	28	0.0000	26	0.0000	27	0.0000	26
博物馆	0.0000	28	0.0008	27	0.0007	27	0.0008	26	0.0005	27	0.0000	29	0.0000	27	0.0000	28	0.0000	27
体育机构各类场馆	0.0038	18	0.0167	17	0.0116	17	0.0113	16	0.0120	15	0.0071	16	0.0156	16	0.0035	16	0.0187	20
银行和保险服务	0.0213	9	0.1024	6	0.0827	7	0.0678	5	0.0857	5	0.0373	8	0.0779	6	0.0233	8	0.1758	5
农场林场、果场生产	0.0066	15	0.0035	24	0.0015	24	0.0025	22	0.0039	22	0.0052	17	0.0000	28	0.0009	19	0.0206	18
园艺销售	0.3623	1	0.0218	16	0.0187	16	0.0064	18	0.0068	19	0.0052	18	0.0125	17	0.0023	18	0.0094	21
电脑网络高新技术	0.0027	20	0.0599	8	0.0718	9	0.0601	7	0.2128	2	0.0523	6	0.0405	12	0.0042	15	1.2590	1
纺织服装生产厂	0.0776	5	0.4281	1	0.1597	3	0.0967	3	0.0772	7	0.1318	3	0.4798	1	0.2385	1	0.2469	4
化工原料及制品	0.0169	10	0.0586	9	0.0748	8	0.0292	13	0.0429	10	0.0358	9	0.0374	13	0.0138	11	0.0804	8
宾馆住宿	0.0098	12	0.0523	11	0.0600	10	0.0408	9	0.0538	9	0.0314	10	0.0561	8	0.0203	9	0.0692	10
酒吧、咖啡	0.1290	3	0.3013	2	0.2379	2	0.2183	1	0.2221	1	0.1736	1	0.2430	3	0.1342	3	0.3349	3
娱乐业	0.0098	13	0.0468	12	0.0506	11	0.0337	12	0.0386	12	0.0288	11	0.0607	7	0.0119	13	0.0561	13
旅游业	0.0087	14	0.0430	14	0.0444	13	0.0340	11	0.0319	13	0.0105	14	0.0421	11	0.0170	10	0.0486	14
自行车修理	0.0066	16	0.0041	21	0.0021	22	0.0017	24	0.0006	26	0.0015	24	0.0031	20	0.0007	22	0.0000	28
开发区	0.0000	29	0.0000	29	0.0000	29	0.0001	29	0.0003	28	0.0004	26	0.0000	29	0.0000	29	0.0000	29
商务办公楼	0.0038	19	0.0444	13	0.0491	12	0.0351	10	0.0423	11	0.0250	12	0.0436	10	0.0068	14	0.0823	7
大厦/写字楼	0.0011	21	0.0077	19	0.0066	18	0.0053	19	0.0092	17	0.0052	19	0.0171	15	0.0009	20	0.0094	22
住宅区	0.0317	7	0.2254	3	0.2521	1	0.1869	2	0.1646	4	0.1374	2	0.2196	4	0.0299	6	0.4284	2
布匹、服装辅料销售	0.0109	11	0.0985	7	0.0277	15	0.0272	15	0.0105	16	0.0045	21	0.0078	18	0.0128	12	0.0318	17
批发部	0.1366	2	0.1071	4	0.0842	6	0.0630	6	0.0833	6	0.0478	7	0.0467	9	0.0306	5	0.0786	9
商场商厦、购物中心	0.0240	8	0.0423	15	0.0334	14	0.0281	14	0.0253	14	0.0228	13	0.0327	14	0.0247	7	0.0599	11
物流仓储、速递运输	0.0322	6	0.0568	10	0.0864	5	0.0483	8	0.0558	8	0.0975	5	0.0872	5	0.1594	2	0.0861	6
车船修理、销售租赁	0.0853	4	0.1046	5	0.1080	4	0.0854	4	0.1992	3	0.1090	4	0.2664	2	0.0866	4	0.0599	12
道路附属	0.0060	17	0.0092	18	0.0049	20	0.0070	17	0.0068	20	0.0052	20	0.0016	21	0.0026	17	0.0412	15

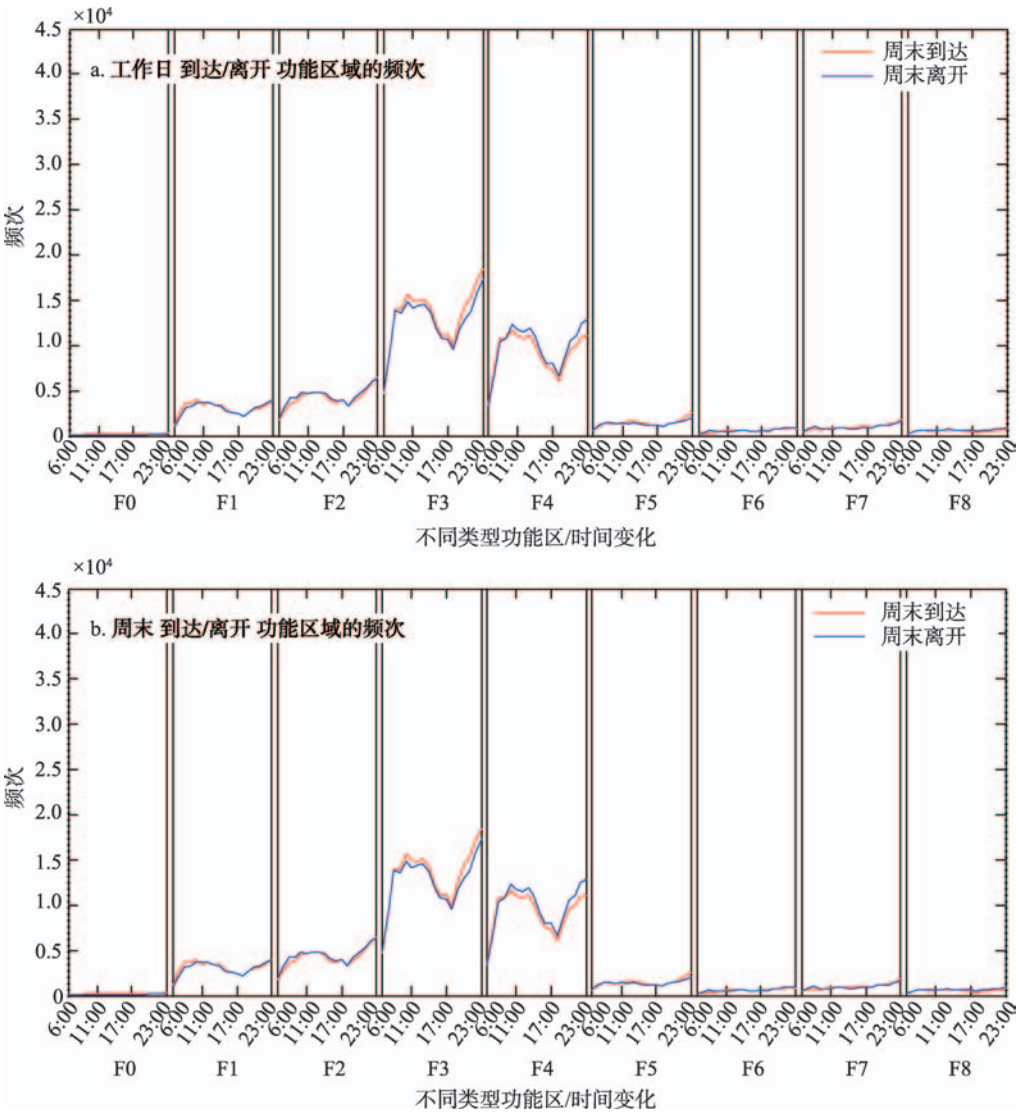

图4　工作日和周末到达/离开功能区域的频次

Fig. 4　Frequency of arrival/departure of functional regions on weekdays and weekends

购物、娱乐地，例如北京路和上下九也在其中。北京路是由大量大型商场构成，上下九主要是由服务业组成。这也说明，这个功能区以服务业为主。同时，通过具有流量特征的图4以及图6，可以看出该功能区工作日下班时段（20:00-23:00）会出现到达流量高蜂，说明很多居民在该功能区内的区域中活动，不仅表现在功能区中区域内部大量的活动，且含有功能区外的居民希望进入这一地区参与活动，其功能区主要以商业（CBD）、开发区为主。

（6）科教文化区[F5]。这个功能区包含大量的科研机构和教育学校等POI信息（例如华南师范大学、华南理工大学、华南农业大学、暨南大学等）。这个功能区中的区域人口流动表现在：周末前往F3区域，而在工作日前往F4的人流较多。由于大学生的活动时间相对比较自由，所以其出行存在任意性。

（7）交通连接区[F6]。这个功能区中的区域是广州重点区域及交通热点，用于连接

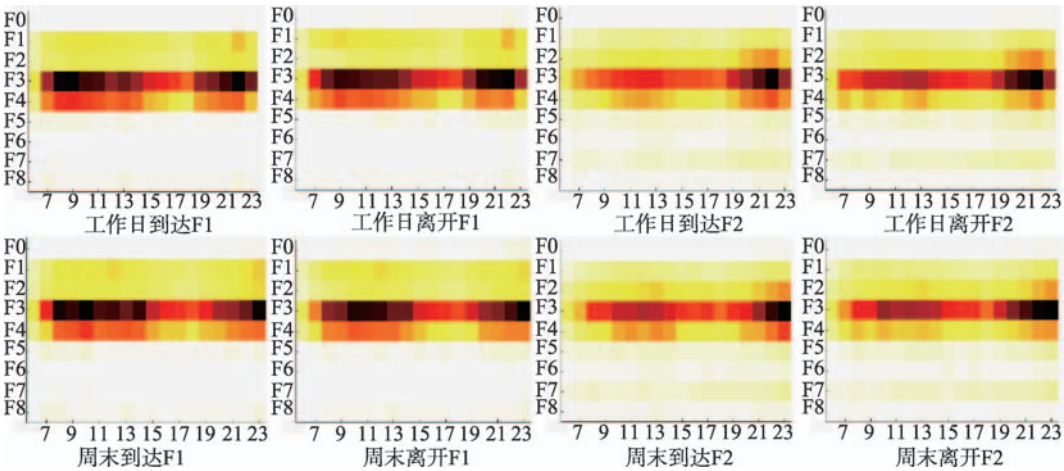

图5 工作日和周末到达与离开功能区域F1、F2

Fig. 5 Weekday and weekend arrival/departure regions of F1 and F2

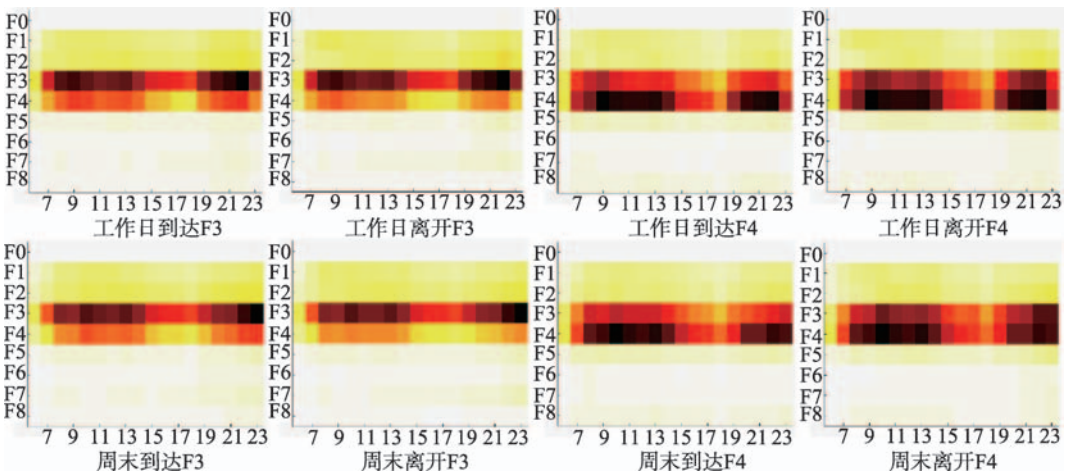

图6 工作日和周末到达与离开区域F3、F4

Fig. 6 Weekday and weekend arrival/departure regions of F3 and F4

市区内外的其他区域，所以包含一些娱乐生活设施，但更主要的还是与居住相关的设施（表2）。此外，全时间段有一定量前往其他区域的车流量且呈均匀分布（图7）。

（8）待开发工业区[F7]。此功能区中的区域位于广州北部郊外，是功能非常局限的待开发区域，由于这些区域正处于建设、发展阶段，所以其物流运输的占比相对较大（表2）。同时，这个区域因其房价成本相对低廉，也包含一些布匹和服装等生产工厂。此郊区以产业为主，前往F1、F2功能区的人流要少于更具吸引力、更成熟的F3、F4区域（图6）。

（9）未识别区域[F8]。因其功能区中区域内部的POI数目和GPS数据过少，导致OPTICS聚类未能识别出功能区。所以，本文未对这个功能区中的区域进行分析。

5 精度验证

为了检验DMR模型应用于城市功能区识别的效果，参考百度地图的地理信息，将研

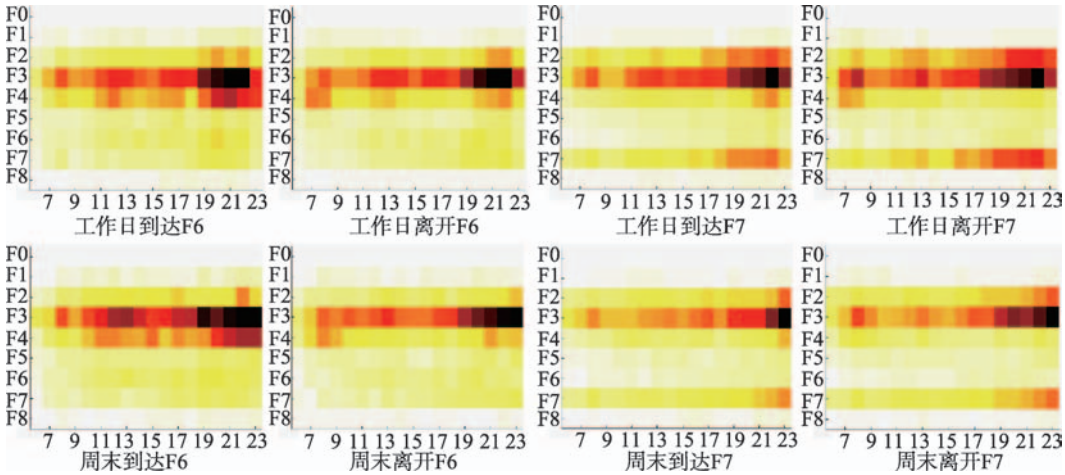

图7　工作日和周末到达与离开区域F6、F7

Fig. 7　Weekday and weekend arrival/departure regions of F6 and F7

究得到的广州市功能区域划分结果与广州市2007年的城镇用地现状图、居民日常出行特征对比分析。综合考虑广州市商住以及居住高度混合的用地现状与研究结果对比分析，DMR模型能对广州市主要的功能区有效地进行识别，具有一定的准确度。其中，若干典型地区的对比结果如表3所示。

本文使用的调查问卷包括了3种预设类型的社区，每种类别的社区体现出与本文研究中识别的城市功能区有相类似的空间分布特征和居民行为规律。特别地，以越秀区的旧城社区（洪庆坊社区、三眼井社区）为例，体现出居民的分布以及出行活动特征符合本文判定的城市功能区的类别。

在城市功能区中的F1功能区中与洪庆坊社区和三眼井社区位置重合的区域具有城市建设、规划比较早的区域结论（表2，图5）。而在调查问卷中，越秀区的旧城社区体现出具备各类社会功能，就业机会充裕，机关、单位大院型用地的复合程度并不高的特征；而功能区F1也具有较为完善齐全的生活配套和公共服务设施。同时，F1功能区中与两个社区相对应的区域（表2），在上午和下午的工作时段维持活跃度高值，午间时分出现一个活跃度谷值；下班（18:30）后进入活动低活跃期并渐次收敛，至晚上23:00后才能基本恢复平静，与图4a规律类似。

总体而言，调查问卷发放区域与本文划分的不同类别的功能区中重合的部分，都具有相类似的居民日常出行的时空特征。旧城区的居民以上午7:00开始活动，至晚上23:00后才基本恢复平静（图5，图6，图7）。特别地，外来人口集中的城市商业集聚区的居民，一天活动开始的时间相对更早；下班后活动密度稍有下降但仍保持一定水平，没有出现急剧回落的情况。

6　结论与讨论

结合文本分类思想和传统城市功能分区的理念，基于人们日常移动行为的大数据，运用潜在语义的主题模型（LDA模型和DMR模型）开展城市功能分区研究，在传统方法的基础上拓展了通过居民行为探究城市空间结构的研究思路，为城市研究提供一种新的方法和途径，为发展和验证城市理论提供了一种重要的分析手段。

表3 识别结果与现状的对照分析

Tab. 3 The comparisons between recognized results and realities

对照区域	中国进出口商品交易会展馆（广交会展馆）
对照图	
识别结果	识别图中A区域（深红色）为文化区与现状图A区域相一致
对照区域	五山区是高校（老教学区）云集地，包含华南理工、华南农业、华南师范、暨南大学等均在此有校区，是广州市的科教文化区
对照图	
识别结果	B区域在百度地图中为科教区与识别图中的B区域（粉色）相互对应
对照区域	亚洲单体建筑面积最大的现代化纺织品批发市场——广州国际轻纺城
对照图	
识别结果	百度地图中的F区域即用红色圈标出的区域和识别图中的F区域相一致

以广州市为例，采用连续一周的浮动车GPS数据以及29类POI数据，建立潜在的狄利克雷模型（LDA）以及狄利克雷多项式回归模型（DMR），得到研究区域属于不同功能的概率；其次，通过OPTICS聚类方法对不同模型的结果进行聚类；然后，利用POI类别密度、居民出行特征等对分区结果进行识别；最后，利用土地利用现状图、百度地图及问卷调查验证其功能分区识别结果。

研究发现，广州市不同类型的功能区呈现出不同的功能特征，如：专业批发区主要以批发市场和销售市场为主，尤其以轻纺城为代表；风景名胜地主要以广州市的风景区为主，包括广州塔、海心沙、华南植物园等地标性建筑物和景区；商业娱乐区主要以大厦、写字楼为主，餐饮、购物其次，同时具有服装生产和运输业发达的特点。而各个行政区也具有不同的功能特征，例如，白云区以白云山的风景区为主，辅以轻工业以及交通要道功能区；天河区主要以商业娱乐区为主，居住休闲等设施完善的生活区为辅；越

秀区，总体上以居住区和商业区为主导，其它类型功能区围绕其展开的特点。该结果采用百度地图、现状图以及调查问卷对识别结果进行验证，与广州市的实际情况基本符合。

由于本文选择浮动车GPS数据和兴趣点数据作为主要的数据源，尽管反映出的客观规律值得借鉴，但是仅仅以出租车GPS数据代表人们的日常出行行为存在一定的偏差，今后拟结合公交数据、地铁数据等公共交通数据开展进一步研究；同时会进一步研究同一类别中不同等级兴趣点对结果可能造成的影响。此外，使用地理空间大数据进行功能分区的研究存在一些不足[31]，例如活动频率与人口密度密切相关，频率的特征既可能是规律性的也存在特殊事件引起的噪音等，这里并未做过细的区分，在后续研究中将结合大数据和人口统计数据等多源数据进行深入分析和研究。

参考文献(References)

[1] Li Dehua. Principles of Urban Planning. Beijing: China Architecture & Building Press, 2001. [李德华. 城市规划原理. 北京: 中国建筑工业出版社, 2001: 12]

[2] Tian G, Wu J, Yang Z. Spatial pattern of urban functions in the Beijing metropolitan region. Habitat International, 2010, 34(2): 249-255.

[3] Dou Zhi. Research on the spatial clustering algorithm for urban function zoning [D]. Chengdu: Sichuan Normal University, 2010. [窦智. 城市功能区划分空间聚类算法研究[D]. 成都: 四川师范大学, 2010.]

[4] Li Xinyun. Research on methods and application for urban spatial data mining [D]. Taian: Shandong University of Science and Technology, 2004. [李新运. 城市空间数据挖掘方法与应用研究[D]. 泰安: 山东科技大学, 2004.]

[5] Qu Guoqing, Jiang Yuchun. Cluster analysis and its application in land utilization classification. Resource Development & Market, 1999, 15(4): 4-7. [曲国庆, 姜玉春. 聚类分析及其在土地利用分类中的应用. 资源开发与市场, 1999, 15(4): 4-7.]

[6] Shi Yufeng, Wang Yan. Study on urban function partition based on self- organizing neural network. Computer Engineering, 2006, 32(18): 206-207. [史玉峰, 王艳. 基于自组织神经网络的城市功能分区研究. 计算机工程, 2006, 32(18): 206-207.]

[7] Wang Hui. Spatial impacts of new economies and the implications for city planning and decision-making. Geographical Research, 2007, 26(3): 577-589. [王慧. 城市“新经济”发展的空间效应及其启示: 以西安市为例. 地理研究, 2007, 26(3): 577-589.]

[8] Wang Hui. Rise of new special development zones and polarization of socio-economic space in Xi'an. Acta Geographica Sinica, 2006, 61(10): 1011-1024. [王慧. 开发区发展与西安城市经济社会空间极化分异. 地理学报, 2006, 61(10): 1011-1024.]

[9] Wang Hui, Tian Pingping, Liu Hong. Spatial structuring of the 'new economies' in Xi'an and its mechanisms. Geographical Research, 2006, 25(3): 539-550. [王慧, 田萍萍, 刘红. 西安城市“新经济”发展的空间特征及其机制. 地理研究, 2006, 25(3): 539-550.]

[10] Wang Yan, Song Zhenbai, Wu Peilin. A study on spatial clustering of urban function partition. Areal Research and Development, 2009, 28(1): 27-31. [王艳, 宋振柏, 吴佩林. 城市功能分区的空间聚类方法 研究及其应用: 以济南市为例. 地域研究与开发, 2009, 28(1): 27-31.]

[11] Wu Wenheng, Xu Zewei, Yang Xinjun. Quantitative research of spatial development differentiation in Xi'an from the perspective of urban functional zoning. Geographical Research, 2012, 31(12): 2173-2184. [吴文恒, 徐泽伟, 杨新军. 功能分区视角下的西安市发展空间分异. 地理研究, 2012, 31(12): 2173-2184.]

[12] Zhu Zhilin. Research on the urban function zoning divide based on fuzzy clustering. Dam and Safety, 2006(S): 28-31. [朱枝琳. 基于模糊聚类的城市功能区划分研究. 大坝与安全, 2006(S): 28-31.]

[13] Joh C H, 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. Washington, DC: AAG Annual Meeting, 2010.

[14] Sun L, Lee D, Erath A, et al. Using smart card data to extract passenger's spatio-temporal density and train's trajectory of MRT system. Urban Computing, 2012: 142-148.

[15] Zhong C, Huang X, Arisona S M, et al. Inferring building functions from a probabilistic model using public transportation data. Computers, Environment and Urban Systems, 2014, 48(6): 124-137.

[16] Qi G, Li X, Li S, et al. Measuring social functions of city regions from large-scale taxi behaviors. Pervasive Computing and Communications Workshops, 2011: 384-388.

[17] Cranshaw J, Schwartz R, Hong J I. The livehoods project: utilizing social media to understand the dynamics of a city. ICWSM, 2012.

[18] Gaubatz P. Changing Beijing. Geographical Review, 1995: 79-96.

[19] Kling F, Pozdnoukhov A. When a city tells a story: Urban topic analysis. Advances in Geographic Information Systems, 2012: 482-485.

[20] Pozdnoukhov A, Kaiser C. Space-time dynamics of topics in streaming text. Location-based Social Networks, 2011: 1-8.

[21] Yin Z, Cao L, Han J, et al. Geographical topic discovery and comparison. World Wide Web, 2011: 247-256.

[22] Liu Y, Wang F, 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(1): 73-87.

[23] Pulliam H R. Sources, sinks, and population regulation. The American Naturalist, 1988, 132(5): 652-661.

[24] Yuan J, Zheng Y, Xie X. Discovering regions of different functions in a city using human mobility and POIs. Knowledge Discovery and Data Mining, 2012: 186-194.

[25] Dong Xiaojing, Yu Zhiwei, Fu Weiwei. Data processing and analyzing system for bus IC card based on GIS. Geospatial Information, 2009, 7(5): 124-126. [董晓晶, 余志伟, 伏伟伟. 基于GIS的公交IC卡数据处理及分析系统. 地理空间信息, 2009, 7(5): 124-126.]

[26] Gao Lianxiong, Wu Jianping. An algorithm for mining passenger flow information from smart card data. Journal of Beijing University of Posts and Telecommunications, 2011, 34(3): 94-97. [高联雄, 吴建平. 从智能卡数据挖掘客流信息的算法. 北京邮电大学学报, 2011, 34(3): 94-97.]

[27] Long Ying, Zhang Yu, Cui Chengyin. Identifying commuting pattern of Beijing using bus smart card data. Acta Geographica Sinica, 2012, 67(10): 1339-1352. [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): 1339-1352.]

[28] Ni Zhongyun, Lei Fanggui, Yang Wunian. Application of google earth in Chengdu functional zoning. Scientific and Technological Management of Land and Resources, 2007, 24(4): 121-124. [倪忠云, 雷方贵, 杨武年. Google Earth在成都市功能分区研究中的应用. 国土资源科技管理, 2007, 24(4): 121-124.]

[29] Yu Xiang. Discovering zones of different functions using bus smart card data and points of interest: A case study of Beijing [D]. Hangzhou: Zhejiang University, 2014. [于翔. 基于城市公交刷卡数据和兴趣点的城市功能区识别研究[D]. 杭州: 浙江大学, 2014.]

[30] Zhang Yu, Hu Xinhua. A detection method of expressway traffic congestion with probe car data. Journal of Transport Information & Safety, 2012, 30(6): 87-89. [章玉, 胡兴华. 基于IC卡数据的居民公交乘车距离研究. 交通信息与安全, 2012, 30(6): 87-89.]

[31] Liu Y, Liu X, Gao S. Social sensing: A new approach to understanding our socioeconomic environments. Annals of the Association of American Geographers, 2015, 105(3): 1-19.

[32] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation. The Journal of Machine Learning Research, 2003, 3: 993-1022.

[33] Wei X, Croft W B. LDA-based document models for ad-hoc retrieval. Research and Development in Information Retrieval, 2006: 178-185.

[34] Li Wenbo, Sun Le, Zhang Dakun. Text classification based on labeled-LDA model. Chinese Journal of Computers, 2008, 31(4): 620-627. [李文波, 孙乐, 张大鲲. 基于Labeled-LDA模型的文本分类新算法. 计算机学报, 2008, 31(4): 620-627.]

[35] Karlsson C. Clusters, functional regions and cluster policies. IBS and CESIS Electronic Working Paper Series, 2007, 84: 1-24.

[36] Mimno D M A. Topic models conditioned on arbitrary features with dirichlet-multinomial regression. ArXiv Preprint ArXiv,1206.3278, 2012.

[37] http://www.gz.gov.cn/gzgov/s2294/zjgzcon.shtml.

Discovering urban functional regions using latent semantic information:
Spatiotemporal data mining of floating cars GPS data of Guangzhou

CHEN Shili[1, 2, 3], TAO Haiyan[1, 2], LI Xuliang[1, 2], ZHUO Li[1, 2]
(1. Center of Integrated Geographic Information Analysis, School of Geography and Planning, Sun Yat-sen University, Guangzhou 510275, China; 2. Guangdong Provincial Key Laboratory of Urbanization and Geo-simulation, Guangzhou 510275, China; 3. Urbanization Institute of Sun Yat-sen University, Guangzhou 510275, China)

Abstract: China has been experiencing rapid urbanization at an unprecedented rate and as a result, urban internal space structure has evolved significantly. It is of great significance to label different functional regions (DFR) inside a city for urban structure analysis, policy making, and resource allocation. These DFRs include residential district, industrial district, education district, and the administration district. This paper explored the characteristics and distribution of urban functional regions based on big geographic data. With the latest road network data, the study area (i.e., 6 districts of Guangzhou city in Guangdong Province, China) was partitioned into 439 segments. By applying the employment of spatial and temporal semantic mining method to the one-week massive floating cars GPS data and the point of interest data, we developed a Latent Dirichlet Allocation (LDA) and Dirichlet Multinomial Regression (DMR) model. Moreover, OPTICS clustering method was employed to process the results of LDA and DMR to identify different functional zones. Meanwhile, status map of Guangzhou urban planning, and resident travel characteristics were used to verify the verification of mentioned results. The results show that this method can identify the obvious characteristics of urban functional areas, such as mature residential area, science and education culture area, commercial area, and development zone. The results also show that residential and commercial areas are dominant DFRs in Guangzhou city, which are surrounded by other types of functional regions. This paper brings a new perspective on using large-scale and high quality individual space-time data to study human migration and daily activities, as well as to explore social space to unveil the formation and mechanism of urban functional zones.

Keywords: latent dirichlet allocation; functional regions; big geographic data; GPS data; point of interest; Guangzhou