引用格式：刘菁菁,刘雨思,伊迪升,等.北京市四环内街区尺度下的主题混合模式挖掘[J].地球信息科学学报,2020,22(6):1370-1382. [Liu J J, Liu Y S, Yi D S, et al. Extracting mixed topic patterns within downtown Beijing at the block level[J]. Journal of Geo-information Science, 2020,22(6):1370-1382.] **DOI:**10.12082/dqxxkx.2020.190594

北京市四环内街区尺度下的主题混合模式挖掘

刘菁菁[1,2],刘雨思[1,2],伊迪升[1,2],杨　静[1,2],张　晶[1,2*]

1. 首都师范大学 三维信息获取与应用教育部重点实验室,北京 100048; 2. 首都师范大学资源环境与旅游学院,北京 100048

Extracting Mixed Topic Patterns within Downtown Beijing at the Block Level

LIU Jingjing[1,2], LIU Yusi[1,2], YI Disheng[1,2], YANG Jing[1,2], ZHANG Jing[1,2*]

1. MOE Key Lab of 3D Information Acquisition and Application, Capital Normal University, Beijing 100048, China;

2. College of Resource Environment and Tourism, Capital Normal University, Beijing 100048, China

Abstract: Cities with different land use types influenced by rapid urbanization and urban expansion support various human activities, such as shopping, eating, living, working, and recreation. The mixed use of land can stimulate the vitality of the city, enable the city togather enough people at different points in time, thus producing more interaction, promoting diversified consumption, and improving the economic and social benefits of the city. Mixed characteristics of land use types in cities gain more popularity in many researches due to the huge practical meanings. However, previous researches on mixed characteristics calculation mainly focused on POI data,and there is a lack of consideration for detecting urban topics. Human activities usually take place in different types of points of interest, the potential relationships and spatial interactions between the different types of adjacent POIs can work together to express the potential semantics of locations. In this paper, from an urban topic perspective, a method for the consideration of the relationship between POIs was proposed, and the Hill Numbers Diversity Index was applied to calculate the mixed degree of topics at the block level. Specifically, LDA (Latent Dirichlet Allocation) topic model was firstly used to generate topic vectors of the block and the co-occurrence patterns of POIs. Secondly, the diversity index was introduced to measure the mixed degree of blocks. Then, according to the Goodness of Variance Fit (GVF) and the nature break method, the blocks wereclassified into three groups: (1) high mixed blocks, (2) medium mixed blocks, and (3) low mixed blocks. Finally, multiple linear regression was applied based on mixed degree and topics in the blocksto uncover the significant topics and mixed pattern.Results show that different mixed blocks haddifferent mixed patterns.For high mixed blocks, the topic of teahouse restaurant was significant; the topics of company, enterprise, and residence weresignificant in medium mixed blocks; and the most typical two patterns in low mixed blocks werethe existence of landscape and famous scenery topic and teahouse restaurant topic. To sum up,starting from the urban topic, this paper reveals the mixed pattern of block, and the results show thatdifferent mixed patterns reflect the characteristics of different mixed areas and present certain rules in spatial distribution, which is conducive to the deep understanding of the cityareas, so as to provide a reference for the construction of Beijing

收稿日期：2019-10-11；**修回日期：**2020-02-02.

基金项目：虚拟现实技术与系统国家重点实验室开放基金项目(01119220010011)。[**Foundation item:** The Open Project Program of the State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, No.01119220010011.]

作者简介：刘菁菁(1995—),女,山西运城人,硕士生,研究方向为空间分析与数据挖掘。E-mail: 923508685@qq.com

***通讯作者：**张　晶(1966—),女,黑龙江哈尔滨人,教授,研究方向为地理信息系统与应用。E-mail: zhangjing5946@sina.com

mixed city, and also provide suggestions for other mixed cities.

Key words: block; LDA; POI co-occurrence; topic; topic mixed pattern; topic mixed degree; TF-IDF; multiple linear regression

***Corresponding author:** ZHANG Jing, E-mail:zhangjing5946@sina.com

摘要：城市是多样性聚集的场所，且多元化和差异性日益增强，故探究土地混合利用具有一定的现实意义。现有的土地混合研究大多以POI（Point of Interest）为研究基础，着眼于城市主题的研究较少。本文采用百度POI数据，在街区尺度下考虑POI共现以提取主题，并挖掘北京市四环内的主题混合模式，其结果可以为城市规划及其建设提供参考。首先，采用LDA（Latent Dirichlet Allocation）主题模型得出街区的主题向量以及主题的POI共现模式；其次，引入多样性指数对街区的混合度进行度量，并依据自然断裂法将街区分为高混合街区、中等混合街区、低混合街区3类；最后，为了探究3类街区中的主题混合模式，先采用多元线性回归找出不同类街区中对混合度影响显著的主题，在此基础上对街区中的混合模式进行提取。结果表明：高混合街区的主题混合模式都是茶座餐厅主题与其他主题的混合；中等混合街区中的混合模式大多是以公司企业主题与住宅（商铺）主题再结合其他主题的混合；低混合街区中最典型的2种模式是茶座餐厅主题主导与风景名胜主题主导的接近单一的模式。不同的模式也体现了不同混合区的特征及其之间的差异，有助于对城市深度理解，从而为混合城市的建设提供参考。

关键词：街区；LDA；POI共现；主题；主题混合模式；混合度；TF-IDF；多元线性回归

1 引言

多样性使城市充满了魅力，混合是对多样的活动和多元的需要最好的回应[1]。城市是各种活动混合的场所，城市中的各类活动通常发生在不同的POI上，地理空间上邻近的不同类型POI之间的潜在关系以及在空间上的相互作用可共同表达位置上的潜在语义—城市主题，为深入理解和表达人类主观认知中的地理环境提供支持[2]。城市主题如住宅、大学教育、公司企业等的混合并不是毫无规律的混杂，表现为区域内各种城市主题以一定的比例混合在一起。城市土地混合对居民出行有直接影响，主题在空间上的集中可减少居民不必要的出行[3-5]、灵活应对居民的需要，且土地混合有利于城市的紧凑发展，是城市土地使用的理想状态[5]。

以往大多数城市混合研究是偏向定性的分析[6]，但大数据时代的到来使得很多学者从量化的角度去探索城市中土地混合情况，其中熵指数已经广泛应用于土地混合度的度量[7]。例如，李苗裔等[8]结合空间熵与时空熵很大程度上提升了城市功能混合度识别的准确性；Jost[9]表示熵指数衡量的是不确定性，并不完全能够度量多样性，并提出了Hill numbers多样性指数；宁晓平[10]采用此指数来揭示城市土地混合结构，并以此为基础探索土地混合利用与城市活力的关系；许思扬等[11]搭建了混合功能发展的分类框架，将框架中的混合方式与规模尺度综合分析，发现在构成城市的基本单元中，街区是混合功能发展的核心尺度。大多数土地利用研究从POI的视角在建筑物的尺度上揭示区域特征[10,12-13]，如构建POI综合赋分评价体系度量不同POI类型与不同功能区的相关性、影响力[14]，通过度量POI的显著性用于提取分层地标[15]，采用POI数据识别城市功能区[16-17]等。

上述研究没有考虑POI之间的联系，同一种POI与不同的POI共现体现不同的城市主题，住宅类POI分布在以住宅为主的主题，也分布在以大学为主的大学主题。Gao等[2]采用POI数据与签到数据，应用基于受欢迎程度重采样的主题模型来提取不同类别的POI共现规律，发现了很多有意义的城市主题。因此，本文以城市主题为独特视角，采用Hill numbers多样性指数度量街区尺度下的主题混合度，进一步量化主题对混合度的影响程度并揭示街区主题的混合模式，意在为混合城市的建设提供参考。

2 研究区概况、数据来源及研究方法

2.1 研究区概况

北京市的核心区域主要位于四环内，作为北京商业、服务业以及公共设施聚集区域，具有很高的可达性且对人们的活动有较强的吸引力[15,18]。四环内有178 955个POI且密度较高，故本文以北京市四环内为研究区域（图1），同时采用街区（即四环内的主要道路切割城市生成的地块）为基本

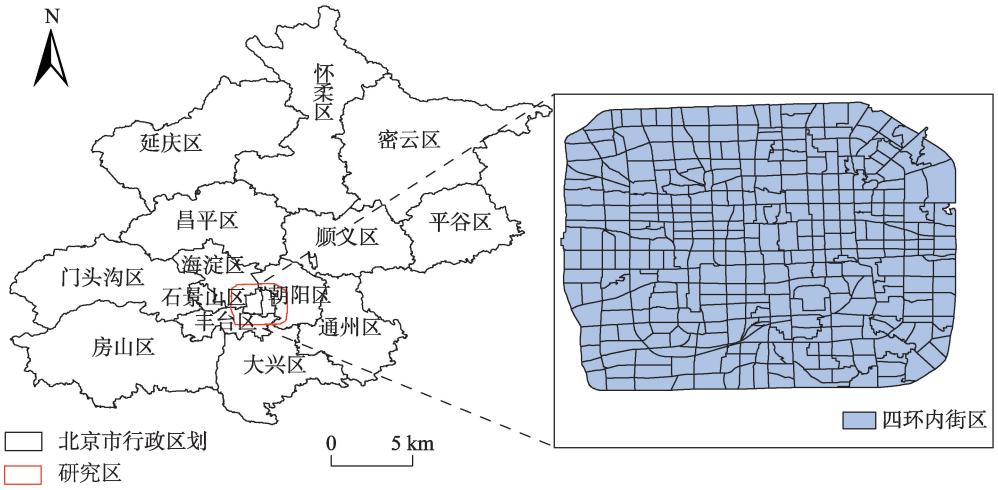


图1 北京市四环内街区示意

Fig. 1 Schematic diagram of block within the fourth ringroad of Beijing

空间单元,街区相对于格网单元,不会存在地理实体跨研究单元分布的情况,最终将四环分割为365个街区(图1)。

2.2 数据与预处理

兴趣点数据(PointOf Interest, POI)是与人们生活密切相关的地理实体的点状数据,样本量大,空间粒度精细、动态更新、可以支持人们活动的具体场所[17]。本文采用的POI采集自2016年百度地图开放平台[19],使用四环内共178 955个POI数据点,表1为本文中POI的属性信息介绍。原始POI类型中有些类之间存在共性,如茶座与咖啡厅2类中都包含咖啡厅,但是2类中并没有重复列举的情况,研究中将这种类型的POI合并为一类;还有一些POI的分类对本研究而言过于细化,如中餐厅、外国餐厅与快餐厅,重分类中将其合并为餐厅;此外,研究区中分布较为分散且被大范围包含的POI类型由于在主题提取中不具有显著性而被剔除,如公厕、报刊亭、停车场等类型。根据以上3种情况对POI数据进行重分类后得到的POI类型如表2所示。

表1 2016年北京市四环内POI属性介绍

Tab. 1 Introduction of POI Attributes within the fourth ringroad of Beijing in 2016

序号	属性字段名称	数据类型	作用描述
1	OBJECTID	Integer	唯一识别码
2	名称	String	POI点名称
3	x	Double	经度
4	y	Double	纬度
5	Type	String	POI类型

表2 2016年北京市四环内POI的类别

Tab. 2 Categories of POIs within the fourth ringroad of Beijing in 2016

类别	名称	类别	名称
1	茶座甜品	13	科研机构
2	KTV	14	培训机构
3	展览馆	15	体育场馆
4	公司企业	16	图书馆
5	大学	17	餐厅
6	风景名胜	18	文化宫
7	购物中心	19	基础教育
8	集市	20	银行
9	商铺	21	游乐园
10	酒吧	22	住宅区
11	酒店	23	医院
12	剧院		

2.3 研究方法

2.3.1 POI权重的度量方法

POI权重直接影响到主题模型对街区中潜在语义的挖掘,以POI数量作为POI的权重是不合理的,还需考虑POI的分布特征,如街区内有一个大学与很多餐厅,但是我们并不能认为数量多的餐厅的权重大于大学。本文引入文本挖掘方法中的TF-IDF

的方法度量POI的权重，但是考虑到不同POI类型的数量可能会有很大的差别，对TF取对数进行标准化，如式(1)、(2)所示。

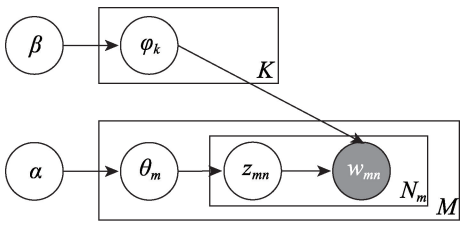
$$TFIDF_{ij}=\ln\left(TF_{ij}\right)\times IDF_i \tag{1}$$

$$IDF_i=\lg\frac{D}{\left\{j:t_i\epsilon d_j\right\}} \tag{2}$$

式中：TF_{ij} 表示j街区中i类POI的频数；D表示所有街区的数量；$\left\{j:t_i\epsilon d_j\right\}$ 表示包含i类POI的街区数量；IDF_i 表示i类POI的逆向文件频率；$TFIDF_{ij}$ 表示j街区中i类POI的权重值。

2.3.2 LDA（Latent Dirichlet Allocation）模型

本文引入机器学习和自然语言处理等领域常用的主题模型—LDA（潜在的狄利克雷分布）来挖掘隐含的主题语义[20-22]。LDA是一种文档生成方法，模型假设话题由单词的多项分布表示，文本由话题的多项分布表示，其中单词分布与话题分布的先验分布都是狄利克雷分布。LDA本质是一种概率图模型，过程如图2所示。



注：图中结点表示随机变量，其中实心结点是观测变量，空心结点表示隐变量，矩形表示重复，矩形中的数字表示重复的次数，结点α和β是模型的超参数，φ_k 表示单词分布的参数，θ_m 表示主题分布的参数，z_{mn} 表示主题，w_{mn} 表示单词[31]。

图2　LDA模型构建示意

Fig. 2　Schematic diagram of LDA topic model construction

给定单词集合W，文本集合D，主题集合Z以及超参数α与β，根据上图，文档的生成过程为：① 生成主题的单词分布。图中β指向 φ_k，且重复K次，表示根据超参数β生成K个主题的单词分布参数 φ_k；② 生成文本的主题分布。图中α指向 θ_m，且重复M次，表示根据超参数α生成M个文本的主题分布参数 θ_m；③ 生成文本的单词序列。图中 θ_m 指向 z_{mn}，重复 N_m 次，表示根据文本的主题分布 θ_m 生成 N_m 个主题 z_{mn}；z_{mn} 指向结点 w_{mn}，同时结点 φ_k 也指向结点 w_{mn}，表示根据主题 z_{mn} 以及K个主题的单词分布 φ_k 生成单词 w_{mn} [31]。

本文将街区形成过程类比于文档形成过程，街区看作文档，主题看作城市主题，POI看作词语。LDA模型可根据给定街区的POI结构，得到街区的主题分布与主题的词分布，即每个街区的主题构成以及各个主题出现的概率大小与每个主题的POI构成以及各个POI出现的概率大小。

2.3.3 街区主题混合度的度量

本文中的混合度测度立足于主题模型提取出的城市主题，量化街区中主题的混合程度。本文引入生物多样性指数—Hill numbers多样性指数，该指数中不同的q值表示对富集种和稀疏种赋予的权重不同，公式如式(3)所示。

$$D_q=\left(\sum_{i=1}^{s}p_i^q\right)^{1/(1-q)} \tag{3}$$

式中：p_i 表示主题i的概率；q被称为阶。D_q 主要取决于q和 p_i，本文中不同的q值对应不同的主题有着不同的加权，D_q 可视为多样性指数系列，q的取值为−∞到+∞，其中q=0，1，2在多样性研究中应用较多。q=0是表示丰富度，即主题的个数，由于本文中主题模型结果中的主题个数差别不大，故不予考虑；q=1是指数型香农指数，q=2是逆辛普森指数，二者都是常见多样性指数的变形，分别用两个指数计算主题的混合度，对比两者的计算结果，选择q为1的多样性指数作为计算街区主题混合度的指标。

3　实验与结果分析

3.1 主题的提取与分析

3.1.1　POI权重结果分析

TF-IDF方法得出的结果基本可以反映不同街区的特征，与我们对街区特征的认知较为一致，同时也从侧面验证了该方法的合理性。表3中列举了4个典型的街区，且将TF-IDF方法计算的权重按照从大到小的顺序依次排列，表3中仅列出了每个街区中权重排名前六的POI类型。街区1和街区216分别对应朝阳公园所在的街区与玉渊潭公园所在的街区，二者均为公园，都承载着供居民观赏休闲的特征，但是朝阳公园相比于玉渊潭公园娱乐特性会更明显。街区345和街区360分别对应三里屯与潘家园所在的街区，从表3可看出，虽然二者权重最高的都是商铺，但是三里屯街区中商铺、餐厅与酒吧共同揭示了三里屯的购物娱乐特征，而潘家园街

表3 典型街区中前6类POI权重结果

Tab. 3 Weight results of the first six POIs in typical blocks

街区1		街区216		街区345		街区360	
POI类型	权重	POI类型	权重	POI类型	权重	POI类型	权重
游乐园	10.33	风景名胜	7.45	商铺	13.99	商铺	17.8
商铺	6.82	游乐园	1.44	餐厅	7.49	住宅区	9.30
公司企业	6.62	公司企业	0.69	酒吧	5.83	餐厅	9.28
餐厅	3.41	茶座甜品	0.36	住宅区	5.57	公司企业	6.18
风景名胜	3.37	基础教育	0.20	茶座甜品	5.35	酒店	1.99
茶座甜品	3.30	商铺	0.17	购物中心	2.20	茶座甜品	1.48

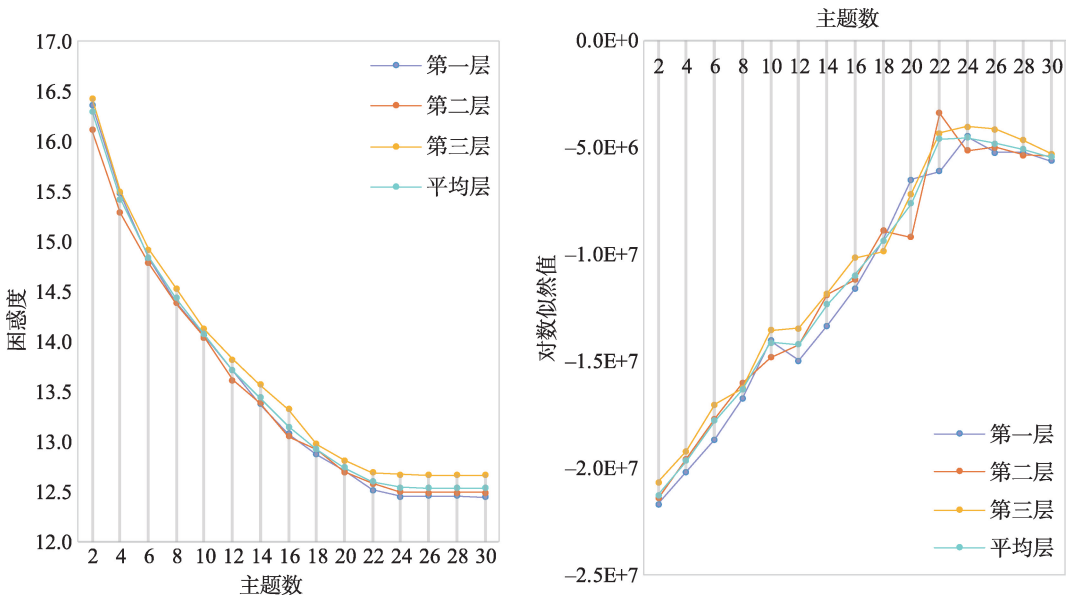
区中商铺、住宅区、餐厅与公司企业等共同体现的特征会明显区别于三里屯街区的购物娱乐特性。虽然从POI角度在一定程度上可以揭示街区的内在细粒度特征，但POI的不同共现模式表现出的不同城市主题可理解为对POI的概括，故本文采用主题而不是POI作为研究基础。

3.1.2 主题数量的确定

LDA方法中选择合适的主题数是非常重要的，主题过多会导致很多无意义的主题，过少时会使得一个主题的概括能力太大，影响结果的准确性。Blei等[21]采用困惑度（Perplexity）作为评价模型好坏的标准，通过选取困惑度最小的模型确定主题的最优数目。Griffiths等[22]根据主题的后验分布中获取样本进而计算对数似然值，选取对数似然值最大时对应的主题数为最优主题数。本文采用以上2种方法确定较为合适的主题数，考虑到一次建模结果的片面性，引入交叉检验的方法，采用3层交叉检验即随机将数据集平均分为3组，每组均作为一次测试集，剩余的2组为训练集，最后将3次的计算结果与3次结果的平均值对比选择主题数。结果如图3所示，图中的困惑度曲线随着主题数的增加而减小，在主题22~24之间达到最小且趋于平稳，而对数似然值伴随着主题数的增加不断变大，在22~24之间有最大值，2种方法的结果趋于一致，分析22~24的实验结果本文选择22作为主题数。

3.1.3 主题模型结果分析

采用LDA建模的结果中包含描述街区的主题分布以及描述主题的POI分布，其中主题的POI分布揭示了不同的POI共现模式能够体现怎样的主题。表4中列举了22个主题中的部分主题及其前5



(a) 困惑度与主题数 (b) 对数似然值与主题数

图3 LDA模型中2种确定主题数方法折线图

Fig. 3 Line graph of two methods to determine topic number in LDA model

表4　LDA建模结果中部分主题的前5类POI

Tab. 4　Top 5 POIs of some topics in LDA modeling results

主题1	主题2	主题4	主题7	主题12	主题14
茶座甜品	风景名胜	酒吧	图书馆	购物中心	科研机构
餐厅	餐厅	餐厅	住宅区	商铺	商铺
商铺	酒店	茶座甜品	餐厅	餐厅	茶座甜品
培训机构	剧院	住宅区	商铺	茶座甜品	餐厅
公司企业	公司企业	商铺	银行	酒吧	基础教育
主题15	主题16	主题17	主题19	主题20	主题21
酒店	游乐园	公司企业	住宅区	体育场馆	大学
餐厅	茶座甜品	餐厅	商铺	餐厅	餐厅
商铺	公司企业	培训机构	餐厅	商铺	商铺
体育场馆	商铺	基础教育	公司企业	培训机构	茶座甜品
文化宫	购物中心	文化宫	酒店	公司企业	公司企业

类POI的组合（前5类的POI的概率之和大于95%），每一类POI隶属于每个主题的概率都不同，按其概率从大到小排序，以便突出每个主题的特色。

从表4可知，餐厅在上述大多主题中都有体现，表示餐厅与不同的POI共现可以体现不同的主题，上述主题中，主题2中风景名胜与周围经常会出现的餐厅、酒店等POI共同体现了风景名胜主题；主题4中的酒吧、餐厅与茶座甜品等POI共同体现的酒吧主题可以体现后海、三里屯等街区的酒吧文化；一个购物中心往往只表现为一个POI，其中分布着很多商铺、餐厅与茶座甜品类型的POI，它们的共现体现了主题12的购物商铺主题；主题7为图书馆、住宅区与餐厅等的共现，主题21为大学主题，表现为大学、餐厅、商铺、茶座甜品等POI的共现，但是由于POI分类中将大学的宿舍划分为住宅类，校图书馆划分为图书馆类，并没有将其统一划分为大学类，故在街区尺度，描述一个大学主题主导的街区可能会涉及主题21和主题7等的共现，但具体情况还要根据实际情况确定。

3.2　主题混合度的度量及其分布

街区是由多个主题共现的，每个街区都可以表示为22维主题的多项式分布，即每个街区的主题混合情况都不同。首先采用Hill numbers多样性指数度量街区的主题混合度，其次为了了解主题混合度的空间分布特征，采用自然断裂法对街区分类，自然断裂法通过计算各种分类的方差和，选取值最小的作为最优的分类结果。但是此方法的类数需要自己指定，本文采用GVF即方差拟合优度确定最优分类数，一定范围内，GVF越大，分类效果越好，GVF随着分类数k变化的折线图如图4所示，分类数3是一个拐点，之后曲线趋近于水平，因此选择3类作为最优分类数。

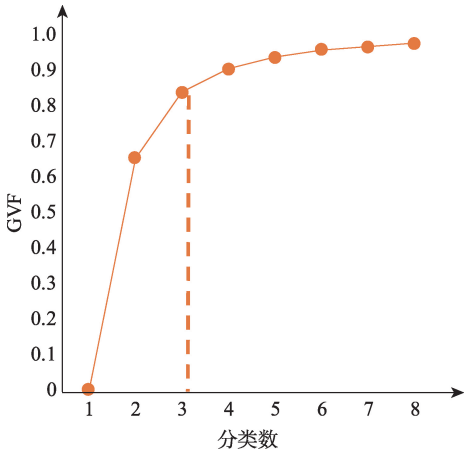


图4　不同分类数下的GVF值

Fig. 4　GVF in different classification numbers

根据自然断裂法将其分为3类，空间分布如图5所示。将其按照环状结构以及行政区划分进行统计分析（图6），环状结构中从内到外高混合度街区比例逐渐降低，中等混合街区与低混合街区比例逐渐增高；以行政区划分的统计结果表明，东城区与西城区高混合街区所占比例较高，丰台区低混合街区占比较高，高混合街区占比最小，结果与实际认知相吻合。因此，进一步探究区域间的混合度差异以及不同混合程度的街区与主题之间的关系具有实际意义。

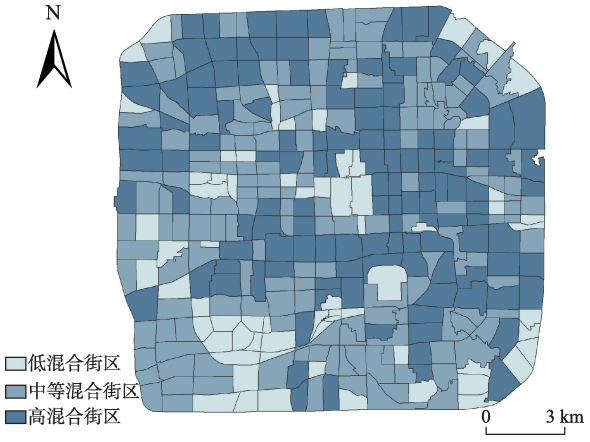


图5 四环区域内自然断裂法划分的街区分布

Fig. 5 Block distribution within the fourth ringroad of Beijing by natural break method

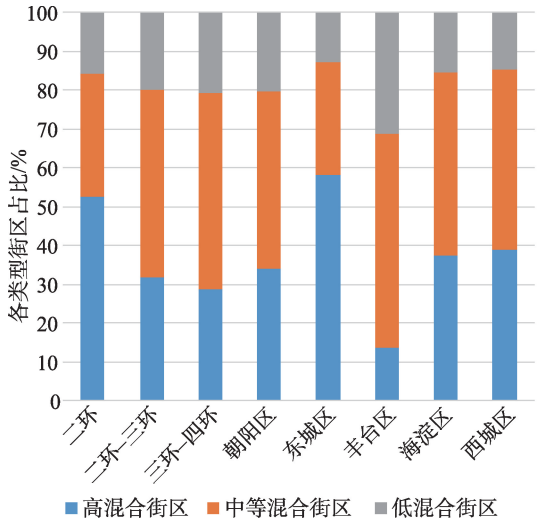


图6 四环区域内3类街区在环状与行政区结构中的占比

Fig. 6 The proportion of three types of blocks in the structure of ring and administrative district within the fourth ringroad of Beijing

3.3 街区的主题混合模式发现

3.3.1 提取与混合街区显著相关的主题

本文从主题的独特视角出发量化四环内街区的混合程度，其混合程度的高低是否受特定主题的影响是一个值得探讨的问题，即挖掘与混合街区显著相关的主题。本文分别对上文中划分的3类街区构建多元线性回归模型，其中，因变量（主题混合度）为连续型变量且服从正态分布，为了找到与不同混合度街区显著相关的主题，采用逐步进入的方法，此方法会对所有可能的影响因子组合建立模型，选出最优拟合模型，可保证最终留在模型中的自变量既显著又没有严重的共线性。构建的回归模型如式(4)所示。

$$Y=\mathrm{b}+b_1x_1+b_2x_2+\cdots+b_nx_n+\epsilon \quad (4)$$

式中：因变量 Y 为街区的主题混合度；自变量 x_1~x_n 表示主题的概率；b为常数；b_1至b_n 为模型的系数，表示每一个自变量对因变量的影响程度；ϵ 表示残差。

上述实验构建了3个回归模型，分别探究高混合街区、中等混合街区与低混合街区与22个主题的关系。表5中构造F检验统计量检验回归方程的显著性，计算结果中*Sig.*值均为0.00，表明3个回归方程中自变量和因变量之间有显著的线性关系。得到拟合结果如表6所示，显著值*Sig.*值均小于0.05，表示回归系数达到95%的置信度，即模型中的自变量均与因变量显著线性相关。3类不同混合度街区的回归模型调整后的拟合优度分别为0.497、0.361与0.425，这可能是由于在同一种混合度的街区中，影响其混合度的主题组合有很多，回归结果中提取出的组合是对这一类街区都有显著影响的一组主题，一定程度上可以揭示一些现实规律。

高混合街区的拟合结果中，住宅（商铺）主题、体育场馆主导的主题、大学主题以及医院主题与街区混合度呈现负相关，相较于高混合街区中的其他街区，存在这些主题的街区混合度较低，这是由于在街区尺度下，住宅、体育场馆、大学、医院自身的占地面积较大，且街区中与这些主题共现的其他主题类型相对固定，故在高混合区中起着降低混合度的作用。购物中心主题、酒吧主题、商业娱乐主题与混合度呈现正相关，可以显著提高街区的混合度，说明这类主题的出现对其他主题具有很大的吸引力且吸引的主题类型较丰富、不确定性高，从而提高街区主题的多样性，增加街区的活力。且实验结果与实际认知相符，购物中心、酒吧与商业娱乐场所一般分布在城市中混合度较高的街区，如三里屯、北海、西单等。

表5 3类混合街区多元线性回归方程的显著性

Tab. 5 Significance of multivariate linear regression equation for three kinds of mixed block

回归模型	*F*	*Sig.*
高混合街区	16.706	0.000
中等混合街区	11.399	0.000
低混合街区	9.467	0.000

表6 街区混合度与22维主题的多元线性回归结果

Tab. 6 Results of multiple linear regression for block mixing degree and 22 topics

	变量	非标准化系数	标准误差	标准系数	*t*	*Sig.*
高混和街区(R^2=0.497)	（常量）	17.763	0.549	—	32.336	0.000
	主题19	−17.463	2.849	−0.464	−6.131	0.000
	主题20	−6.754	3.210	−0.151	−2.104	0.037
	主题12	8.440	2.665	0.214	3.168	0.002
	主题4	9.335	2.683	0.241	3.479	0.001
	主题1	−22.574	4.687	−0.433	−4.816	0.000
	主题21	−7.635	2.113	−0.304	−3.614	0.000
	主题18	−5.585	2.114	−0.188	−2.641	0.009
	主题13	7.611	3.118	0.160	2.440	0.016
中等混合街区(R^2=0.361)	（常量）	12.132	0.403	—	30.069	0.000
	主题19	−8.329	1.575	−0.366	−5.289	0.000
	主题20	−4.761	1.914	−0.181	−2.488	0.014
	主题17	−5.142	1.869	−0.186	−2.750	0.007
	主题10	6.550	2.290	0.196	2.861	0.005
	主题22	5.275	2.331	0.160	2.263	0.025
	主题13	4.525	2.265	0.141	1.998	0.047
低混合街区(R^2=0.425)	（常量）	4.508	0.265	—	17.016	0.000
	主题5	20.417	4.915	0.396	4.154	0.000
	主题1	5.628	1.857	0.288	3.031	0.004
	主题3	11.469	4.556	0.243	2.517	0.014
	主题6	12.533	4.640	0.260	2.701	0.009
	主题2	−2.460	1.014	−0.236	−2.426	0.018

中等混和街区的回归结果中，对其混合度提升有显著影响的主题为基础教育主题、银行商铺主题以及公司企业周围的商业娱乐主题，有降低作用的主题为住宅（商铺）主题、体育场馆以及公司企业主题。综上来看，相对于与街区混合度呈正相关的主题，住宅（商铺）主题、体育场馆以及公司企业主题中起主导作用的POI—住宅、体育场馆、公司企业占地面积相对较大，即与更多类型主题共现的可能性就越小，且住宅、公司企业一般成片分布，也是导致其降低街区混合的原因。

低混合街区中对混合度有显著提升作用的主题为附近为培训机构的住宅主题、附近分布集市的住宅主题以及展览馆主题，与之相比，由于以风景名胜主题主导的街区边界大多与风景名胜边界吻合如天坛、陶然亭、故宫等，且主题相对单一，街区中主题多样性较差，能显著降低街区的混合度。

每个主题在不同类型的街区中扮演着不同的角色，发挥着不同的作用，多元线性回归结果揭示了不同类型街区的混合度受哪些主题的影响以及这些主题怎么影响街区混合度。街区由一个或多个主题共同构成，每个主题对街区的影响不同，故不同街区中主题的混合方式是一个值得探究的问题。

3.3.2 挖掘街区的主题混合模式

上述研究中表明不同街区的混合度不同，但是街区中的主题是如何混合的仍然是一个待解决的问题。为了探究不同街区的主题混合模式，以上述提取出的与混合街区显著相关的主题为变量，对3种混合度的街区分别进行层次聚类，进而计算每一类中与混合度显著相关的主题在街区中占比的平均值，挖掘3种混合街区中的主题混合模式。

（1）高混合街区的主题混合模式

依据层次聚类中的树状图将高混合街区分为5类，图7中的柱状图显示了5类中主题占比均值的大小，将值大于8种主题均值的主题组合定义为每一类的主题混合模式（表7中已将主题数根据LDA转化为具体主题），不同的模式体现高混合街区中的不同特征，如模式4为大学区的主题混合模式，体现了高混合街区的教育特征，模式1与模式5为表明高混合区的娱乐特征。观察表7中的混合模式发现，茶座餐厅主题出现在每一种模式中，与其他不

图7 高混合街区层次聚类结果分布及各类中主题均值

Fig. 7 The distribution of hierarchical clustering results and the histogram of mean value of various subjects in high mixed block

表7 高混合街区主题混合模式

Tab. 7 Mixed pattern of high mixed blocks

模式类型	主题混合模式
模式1(高1类)	购物中心主题+住宅(商铺)主题+茶座餐厅主题+其他
模式2(高2类)	医院主题+茶座餐厅主题+其他
模式3(高3类)	住宅(商铺)主题+茶座餐厅主题+其他
模式4(高4类)	大学主题+茶座餐厅主题+住宅(商铺)主题+其他
模式5(高5类)	酒吧主题+茶座餐厅主题+其他

同主题共同分布表现为不同的高混合模式,说明茶座餐厅主题分布街区较广,且与其他主题共同分布的街区一般表现为高混合。

(2)中等混合街区的主题混合模式

根据层次聚类的树状图将中等混合街区分为5类,得到的聚类结果以及每一类的均值统计图如图8,中等混合街区的主题混合模式如表8。由表8可知除模式4之外,其余模式都以公司企业主题与住宅(商铺)主题混合再结合其他主题构成,可见公司企业主题与住宅主题混合是中等混合街区中一种经典的关联模式。现如今,很多住宅区分布在公司企业附近以减少通勤时间,也从侧面验证了这种关联模式的正确性。

(3)低混合街区的主题混合模式

采用层次聚类将该区域街区划分为3类,并按照主题占比均值对其统计,得到结果如图9,进而得出主题混合模式(表9),分析可知模式2与模式3主题相对比较单一,模式2所在街区为主要是茶座餐厅主导的街区,对比街区本身的实际情况,该模式街区中有一部分为火车站主题主导,但是由于数据中并没有火车站这种类型的POI,又因为火车站附近分布大量的餐厅与茶座甜品店,故结果中将这类街区归属于模式2,模式2中除火车站所在的街区存在误分之外,其他街区以茶座餐饮主题为主,大

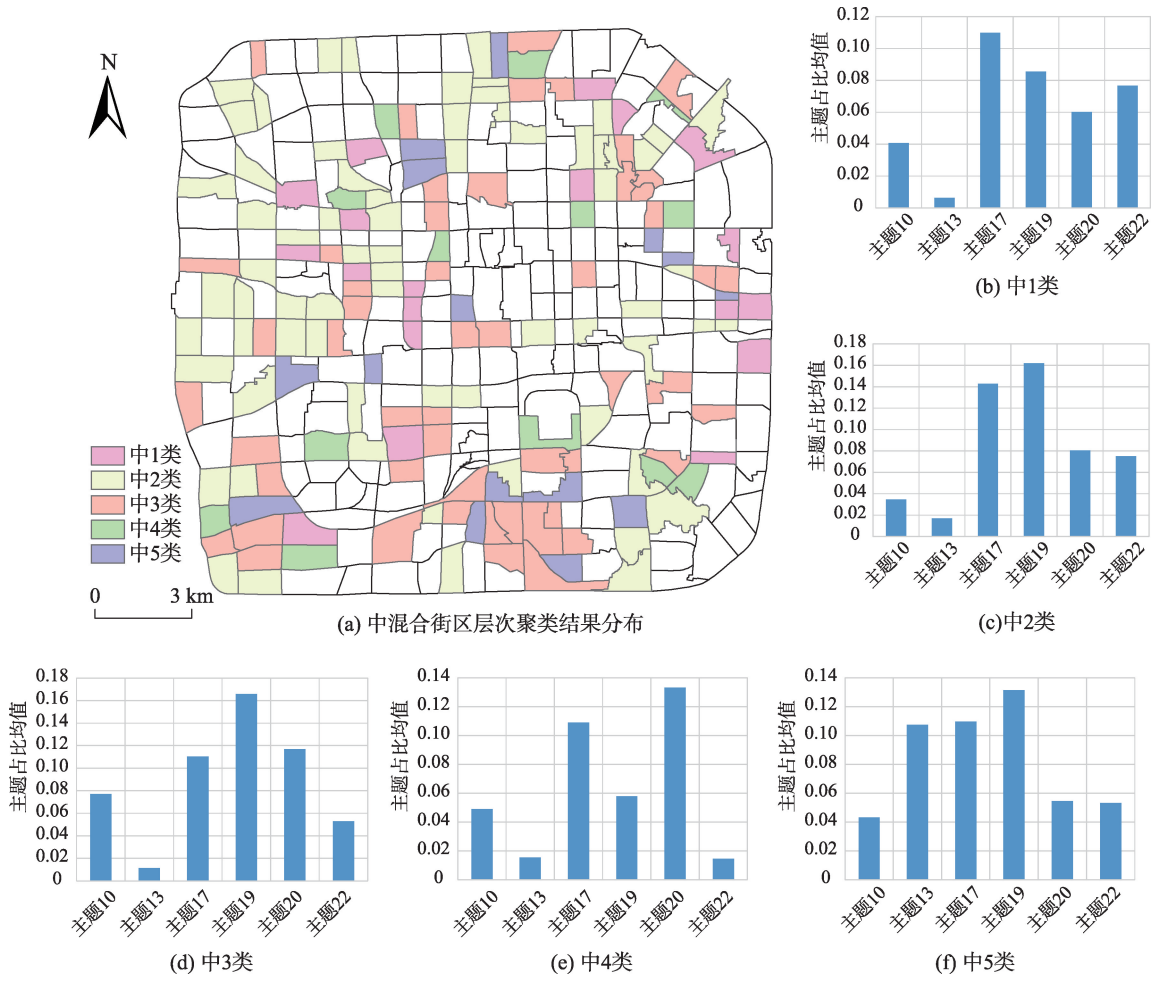


图8　中混合街区层次聚类结果分布及各类中主题均值

Fig. 8　The distribution of hierarchical clustering results and the histogram of mean value of various subjects in middle mixed block

表8　中混合街区主题混合模式

Tab. 8　Mixed pattern of medium mixed blocks

模式类型	主题混合模式
模式1(中1类)	公司企业主题+住宅(商铺)主题+银行主题+其他
模式2(中2类)	公司企业主题+住宅(商铺)主题+其他
模式3(中3类)	住宅(商铺)主题+公司企业主题+体育场馆主题+其他
模式4(中4类)	公司企业主题+体育场馆主题+其他
模式5(中5类)	休闲娱乐主题+公司企业主题+住宅(商铺)主题+其他

多为面积较小的美食聚集区；模式3主要是以风景名胜主题主导，这是由于风景名胜的边界大多与街区范围重合，故这一类街区的主题多样性较低。除此之外，模式1所在街区主要包含住宅(集市)主题以及展览馆主题，此类中的住宅(集市)主题不同于高混合街区中的住宅(商铺)主题，住宅(集市)主题主导的街区住宅多与集市分布在一起。相较于住宅(商铺)主题与其他主题主导的街区，住宅(集市)主题与其他主题主导的街区混合程度较低。

4　结论与讨论

POI在空间上不是绝对孤立的存在，由于人的活动大多发生在POI上，人的活动与POI的分布是相互影响相互制约的，不同类型的POI会共同出现在某一区域体现同一个主题，不同主题的混合体现街区不同的特性，且挖掘街区的主题混合模式有助于宏观上了解四环内的主题混合情况。本文以POI为数据基础，采用LDA主题模型以及空间分析方法度量街区尺度上主题混合的空间特征，并以此为基础度量主题对混合度的影响程度，探究北京市四环内街区的主题混合模式。实验结果中，不同混合度街区中有不同的主题混合模式，且反映了不同街区

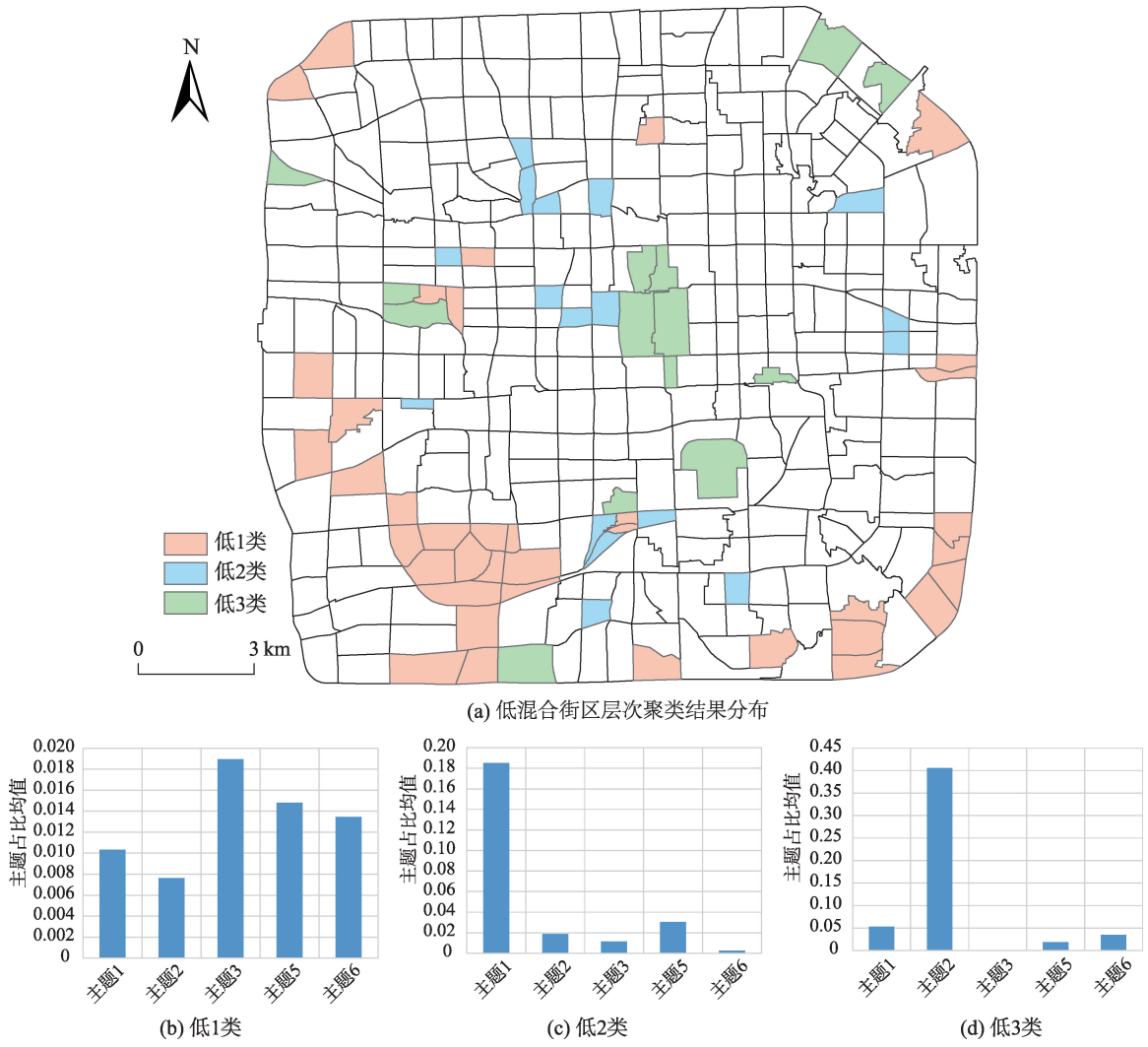


图9 低混合街区层次聚类结果分布及各类中主题均值

Fig. 9 The distribution of hierarchical clustering results and the histogram of mean value of various subjects in low mixed block

表9 低混合街区主题混合模式

Tab. 9 Mixed pattern of low mixed blocks

模式类型	主题混合模式
模式1(低1类)	住宅(集市)主题+展览馆主题+其他
模式2(低2类)	茶座餐厅主题+其他
模式3(低3类)	风景名胜主题+其他

的特征。具体表现为:

(1)高混合街区中发现了5种混合模式,是对人们认知中繁华地段城市主题组合的表达,主要分布在中关村、海淀大学区、三里屯等街区。除此之外,发现5种混合模式均为茶座餐厅主题与其他主题的混合,可见茶座餐厅主题与其他主题的混合一般分布在高混合街区。

(2)中等混合街区中的混合模式有3种,表达为以公司企业主题与住宅(商铺)主题再结合其他主题的混合,主要分布在住宅与公司企业均有分布的街区,丰台区分布较多。

(3)低混合街区中最典型的两种模式是茶座餐厅主题主导与风景名胜主题主导的接近单一的模式,相较于上述2类街区,该类街区空间上分布较为分散,主要分布在故宫、天安门、天坛等景区所在的街区、一些以餐饮为主的小吃街以及主题占比较少的街区。

结果既可为北京混合城市的建设提供参考,同时也可为其他混合城市提供建议,城市主题的发现是我们从微观尺度了解街区的一个新的思路,也为宏观上探究街区中主题的混合度提供了新的视角。

但是研究仍存在一定的局限性:首先,对功能区的识别依赖POI类型,例如该实验中的POI中没

有有关火车站的POI信息，故并没有得出火车站主题以及以火车站主题主导的街区。其次，主题的混合模式研究是基于街区尺度开展的，没有考虑到不用尺度对空间分异的影响，后续研究会加入尺度因素继续研究。

参考文献(References)：

[1] 黄毅.城市混合功能建设研究[D].上海:同济大学,2008. [Huang Y.A study of urban mixed-use development in theory and practice: The case of Shanghai[D]. Shanghai: Tongji University, 2008.]

[2] Gao S, Janowicz K, Couclelis H. Extracting urban functional regions from points of interest and human activities on location-based social networks[J]. Transactions in GIS, 2017,21(3):446-467.

[3] Bordoloi R, Mote A, Sarkar P P. Quantification of land use diversity in the context of mixed land use[J]. Procedia-Social and Behavioral Sciences, 2013,104:563-572.

[4] Maleki M Z, Zain M F M, Ismail A. Variables communalities and dependence to factors of street system, density, and mixed land use in sustainable site design[J]. Sustainable Cities and Society, 2012,3:46-53.

[5] 龚咏喜,李贵才,林姚宇.土地利用对城市居民出行碳排放的影响研究[J].城市发展研究,2013,20(9):112-118. [Gong Y X, Li G C, et al. Impact of land use on urban household travel carbon emissions[J]. Urban Development Studies, 2013,20(9):112-118.]

[6] 包宇.城市土地混合利用测度研究——以深圳市为例[J].湖北农业科学,2016,55(22):5794-5797. [BaoY. Measure of mixed urban land use: Case of Shenzhen city[J]. Hubei Agricultural Sciences, 2016,55(22):5794-5797.]

[7] Cervero R, KockelmanK. Travel demand and the 3Ds: density, diversity, and design[J]. Transportation Research Part D: Transport and Environment, 1997,2(3):199-219.

[8] 李苗裔,马妍,孙小明,等.基于多源数据时空熵的城市功能混合度识别评价[J].城市规划,2018,42(2): 97-103. [Li M Y, Ma Y, Sun X M. Application of spatial and temporal entropy based on multi-source data for measuring the mix degree of urban functions[J]. City Planning Review, 2018,42(2):97-103.]

[9] Lou J. Entropy and Diversity[J]. Oikos, 2010,113(2):363-375.

[10] 宁晓平.土地利用结构与城市活力的影响分析[D].深圳:深圳大学,2016. [Ning X P. Analysis of the influence of land use structure and urban vitality[D]. Shenzhen: Shenzhen University, 2016.]

[11] 许思扬,陈振光.混合功能发展概念解读与分类探讨[J].规划师,2012,28(7):105-109. [Xu S Y, Chen Z G. Interpretation and classification of mixed function development concept[J]. Planners, 2012,28(7):105-109.]

[12] Yue Y, Zhuang Y, Yeh A G O, et al. Measurements of POI-based mixed use and their relationships with neighbourhood vibrancy[J]. International Journal of Geographical Information Systems, 2016,31(4):1-18.

[13] 康朝贵,刘瑜,邬伦.城市手机用户移动轨迹时空熵特征分析[J].武汉大学学报·信息科学版,2017,42(1):63-69. [Kang C G, Liu Y, Wu L. An analysis of entropy of human mobility from mobile phone data[J]. Geomatics and Information Science of Wuhan University, 2017,42(1):63-69.]

[14] 康雨豪,王玥瑶,夏竹君.利用POI数据的武汉城市功能区划分与识别[J].测绘地理信息,2018,43(1):81-85. [Kang Y H, Wang YY, Xia Z Z. Identification and classification of wuhan urban districts based on POI[J]. Journal of Geomatics, 2018,43(1):81-85.]

[15] Zhao W F, Li Q Q and Li B J. Extracting hierarchical landmarks from urban POI data[J]. Journal of Remote Sensing, 2011,15(5):973-988.

[16] Chi J, Jiao L, Dong T, et al. Quantitative identification and visualization of urban functional area based on POI data[J]. Journal of Geomatics, 2016,41(2):68-73.

[17] Han H, Xiang Y U, Long Y. Identifying urban functional zones using bus smart card data and points of interest in Beijing[J]. City Planning Review, 2016,40(6):52-60.

[18] Lynch K. Good city form[M]. Massachusetts: MIT press, 1984.

[19] 百度地图开放平台[EB/OL]:http://lbsyun.baidu.com/. [Baidu Map Open Platform: http://lbsyun.baidu.com/.]

[20] David M. Probabilistic topic models[J]. IEEE Signal Processing Magazine, 2010,27(6):55-65.

[21] Blei D M, Ng A Y, Jordan M I. Latent dirichletallocation [J]. Journal of Machine Learning Research, 2003,3(1): 993-1022.

[22] Griffiths T L, Steyvers M. Finding scientific topics[J]. Proceedings of the National Academy of Sciences, 2004, 101(1):5228-5235.

[23] 刘瑜,詹朝晖,朱递,等.集成多源地理大数据感知城市空间分异格局[J].武汉大学学报·信息科学版,2018,43(3): 327-335. [Liu Y, Zhan Z H, Zhu D, et al. Incorporating multi-source big geo-data to sense spatial heterogeneity patterns in an urban space[J]. Geomatics and Information Science of Wuhan University, 2018,43(3):327-335.]

[24] 刘瑜.社会感知视角下的若干人文地理学基本问题再思考[J].地理学报,2016,71(4):564-575. [Liu Y. Revisiting several basic geographical concepts: A social sensing perspective[J]. Acta Geographica Sinica, 2016,71(4):564-575.]

[25] Huang J, Levinson D, Wang J, et al. Tracking job and housing dynamics with smartcard data[J]. Proceedings of the National Academy of Sciences, 2018,115(50):12710-12715.

[26] Wang B, Zhen F, Wei Z, et al. A theoretical framework and methodology for urban activity spatial structure in e-society: Empirical evidence for Nanjing city, China[J]. Chinese Geographical Science, 2015,25(6):672-683.

[27] Zhen F, Tang J, Chen Y. Spatial distribution characteristics of residents' emotions based on Sina Weibo big data: A case study of Nanjing[M]//Big Data Support of Urban Planning and Management. Springer, Cham, 2018:43-62.

[28] 秦萧,甄峰.大数据与小数据结合:信息时代城市研究方法探讨[J].地理科学,2017,37(3):321-330. [Qin X, Zhen F. Combination between big data and small data: New methods of urban studies in the information era[J]. Scientia Geographica Sinica, 2017,37(3):321-330.]

[29] 刘瑜,康朝贵,王法辉.大数据驱动的人类移动模式和模型研究[J].武汉大学学报·信息科学版,2014,39(6):660-666. [Liu Y, Kang C G, Wang F H. Towards big data-driven human mobility patterns and models[J]. Geomatics and Information Science of Wuhan University, 2014, 39(6):660-666.]

[30] 陈瑗瑗,高勇.利用社交媒体的位置潜语义特征提取与分析[J].地球信息科学学报,2017,19(11):1405-1414. [Chen Y Y, Gao Y.Extracting and analyzing latent semantic characteristics of locations using social media data. Journal of Geo-information Science, 2017,19(11):1405-1414.]

[31] 李航.统计学习方法—第二版[M].北京:清华大学出版社,2019. [Li H. Statistical learning methods, second edition[M]. Beijing: Tsinghua University Press, 2019.]

[32] 陈泽东,谯博文,张晶.基于居民出行特征的北京城市功能区识别与空间交互研究[J].地球信息科学学报,2018, 20(3):291-301. [Chen Z D, Qiao B W, Zhang J. Identification and spatial interaction of urban functional regions in Beijing based on the characteristics of residents' traveling[J]. Journal of Geo- information Science, 2018,20(3): 291-301.]