基于大数据的城市土地利用分类研究
——以西宁市为例*

戴一华[1)] 刘志锋[1,2)†] 王一航[1,2)] 杨志鹏[3)]

(1)北京师范大学地理科学学部, 100875, 北京;
2)北京师范大学地表过程与资源生态国家重点实验室, 人与环境系统可持续研究中心, 100875, 北京;
3)国家自然科学基金委员会, 100085, 北京)

摘要 以西宁市为例, 基于宜出行和兴趣点(points of interest, POI)2 类常用大数据以及最大似然、支持向量机和神经网络 3 种常用分类方法, 开展了城市土地利用分类研究. 通过对比不同数据与方法组合下的城市土地利用分类精度, 确定了提取城市土地利用信息的最优数据组合方式和分类方法. 并基于分类结果对西宁市的城市土地利用格局进行了分析. 结果显示, 基于 POI 和宜出行数据的神经网络分类方法获取的研究区城市土地利用信息精度最高, 总体精度为 71.25%, Kappa 系数为 0.62. 主要原因在于综合 POI 和宜出行可以更加充分地反映不同土地利用类型的特征, 而神经网络可以有效综合多源大数据的信息. 因此, 基于多源大数据和神经网络为快速有效地获取城市土地利用信息提供了有效途径, 具有较大的应用潜力.

关键词 城市土地利用; 大数据; 兴趣点; 宜出行; 机器学习

中图分类号 K909 **DOI**: 10.12202/j.0476-0301.2020224

0 引言

城市是具有一定人口规模, 并且以非农业活动为主要生产方式的高级聚落形式[1]. 城市土地利用是城市内部与外部社会、政治、经济、技术等多种因素综合作用于城市土地的结果[2]. 城市土地利用的空间格局及其变化过程, 表征了一定时间和空间内人-地相互作用的方式与强度[3]. 城市土地是城市社会、经济发展的基础, 因此城市土地的利用方式和结构直接影响着城市发展的速度与规模[4]. 目前中国的城市化综合水平正在持续提高[5], 且这一趋势还将维持较快速度继续发展. 在此过程中, 以农业为主的乡村型社会逐渐向非农产业为主的现代城市型社会转变, 土地利用模式发生巨大的转变, 出现了工业用地多、居住用地少, 建设用地多、生态用地少等城市用地结构不合理的问题[6]. 不合理的城市空间布局不仅会导致城市土地的低效利用, 还会引发资源浪费与环境污染等许多方面的问题, 给城市的可持续发展带来不利影响[7]. 城市土地利用分类研究有助于深入了解城市土地利用格局, 分析城市发展现状以及土地利用的合理性, 对合理规划城市功能分区、提升用地效益、促进区域可持续发展具有重要的意义. 因此, 城市土地利用分类研究一直是城市规划学和城市地理学研究的核心内容之一.

传统的城市土地利用研究, 大多基于统计调查数据和各种实地测量数据获取城市土地利用信息. 例如, 李永乐等[8]以统计资料为主要数据源, 获取了 2000—2009 年全国 29 个省级行政区的城市化水平数据和城市用地面积信息, 深入探究了城市化发展与城市土地利用结构变化之间的关系. 但是实地调查和统计数据的获取往往费时费力, 难以满足快速提取和动态监测城市土地利用信息的需求. 随着卫星遥感技术日趋成熟, 各类遥感数据开始越来越多地应用于土地分类工作之中[9−10]. 匡文慧等[11]综合集成 SPOT5 影像、地形图、历史地图以及城市规划图等信息, 对长春市 1905 年以来的城市用地信息进行了恢复和提取, 从而建立了一种基于“分层分类”与“对象分割”的城市土地利用空间信息数字重建方法; 王彩艳等[12]基于资源三号卫星遥感影像, 通过多尺度分割和隶属度函数法构建合理的分类层次, 发展了一种面向对象城市土地利用信息提取方法. 目前, 遥感技术已经成

* 第二次青藏高原综合科学考察研究资助项目(2019QZKK0405); 国家自然科学基金资助项目(41871185)
† 通信作者: 刘志锋(1986—), 男, 博士, 副教授. 研究方向: 景观地理与景观可持续科学. E-mail: Zhifeng.Liu@bnu.edu.cn
收稿日期: 2020-06-07

为土地利用信息提取的主要手段. 然而,不同于传统的土地利用分类研究,城市土地利用类型主要是根据其特定的功能需求而划分[13]. 为了实现某种特定的城市功能,同一种土地利用类型内部的土地覆盖类型可以是复杂多样的,加之城市内部的建筑物在光谱特征上具有相似性,传统的遥感分类方法通常难以对城市土地利用信息进行有效的识别[14−15]. 在城市内部,人类的社会经济活动是土地利用类型发生演变的主要推动力. 而遥感数据不能直接反映人类的社会经济活动信息,因此以遥感数据为单一数据源的分类方法,很难对城市土地利用信息实现及时、准确获取.

近年来,大数据已经成为获取城市土地利用信息的一种新兴数据源. 大数据即为数据量超出传统计算机软件分析处理能力,需要特殊的信息处理模式加以分析并提取重要信息的数据集[16]. 大数据可以分为手机信令、车载GPS(global positioning system)、社交媒体大数据等行为活动数据以及兴趣点(points of interest, POI)、OpenStreetMap(OSM)、遥感数据等建成环境数据[17]. 大数据具有覆盖广、获取便捷、精度高、更新快等特点[7],并且已经在地理学研究中得到了广泛的应用[18−19]. 由于地理大数据同时带有位置信息与区域的属性——区域功能属性或社会经济属性等,大数据的出现为获取城市土地利用信息提供了新的途径. 例如:鲁国珍等[20]以深圳市为例,基于腾讯QQ用户的电子足迹数据提出了不同类型的人类时空活动指数,发展了以人类时空活动为特征的城市土地利用分类方法;宁晓刚等[7]基于POI数据,根据地块内不同类型兴趣点的面积占比设定不同的权重因子,构建了一种基于街区尺度的城市主导功能用地划分方法;陈世莉等[21]、陈泽东等[22]应用车载GPS定位数据,通过分析居民出行特征完成了城市功能区的识别;钮心毅等[23]则利用手机信令数据完成了城市空间结构识别的研究;另外,Hu等[24]、Zhang等[25]还结合POI数据与Landsat、QuickBird等遥感数据,开展了城市土地利用格局分析和城市功能区划分等工作. 然而大数据的数据量较大,内容信息复杂,在实际应用中需要选择合适的分类器加以处理. 在基于大数据的城市土地利用分类研究中,为了实现大量数据的高效处理,以及不同特征的充分融合,机器学习方法得到了广泛的应用. 常见的分类方法包括空间聚类[23,26]、支持向量机[27]、随机森林[28−29]、神经网络[30]以及元胞自动机[31]等. 例如Gong等[28]在不透水层提取的基础上,利用随机森林模型融合了传统光学数据、夜间灯光数据、POI数据和腾讯MPL(mobile-phone locating-request)数据,在全国范围内开展了城市土地利用分类研究. 该项研究综合了遥感数据与大数据在土地分类研究中的优势,并利用机器学习算法实现了特征的有效融合,在全国尺度完成了比较准确的城市土地利用分类工作. 然而,目前已有的城市土地利用分类研究所采用的大数据种类繁多,相应的分类方法多种多样,并且缺乏对不同数据源与方法的综合对比研究,导致对基于大数据的城市土地利用分类途径莫衷一是. 因此,如何选择合适的数据源和分类方法来准确获取城市土地利用信息仍需深入研究.

本文的研究目的在于探索一种基于大数据准确获取城市土地利用信息的途径. 以西宁市为例,基于宜出行和POI这2类常用大数据以及3种常用分类方法,对该地区城市土地利用进行分类. 系统对比了不同数据组合方式与不同分类方法下的城市土地利用分类精度,确定了提取城市土地利用信息的最优数据组合方式和分类方法. 最后基于分类结果对西宁市的城市土地利用格局进行了分析.

1 研究区与数据

1.1 研究区 研究区为青海省西宁市主城区,中心坐标为101°49′E、36°34′N,平均海拔2 261 m,年平均气温7.6 ℃,年平均降水量380 mm,属于大陆高原半干旱气候[32]. 西宁市是青藏高原上规模最大、人口最多的城市. 全市总面积7 660 km^2,市区面积约380 km^2. 2019年末,全市常住人口为238.71万人,其中城镇人口173.90万人,城镇化率达到72.85%[33].

由于区域生态环境脆弱且城市发展速度较快,及时了解西宁市城市土地利用现状以及动态变化情况,对于区域可持续发展尤为重要. 同时,西宁市土地利用方式多样,因此在西宁市开展城市土地利用分类研究,有助于验证大数据在城市土地利用分类工作中的适用性.

1.2 数据 城市土地利用信息提取用到的数据主要包括4类(表1). 其中用于划分基本分类单元的数据为OSM路网数据,数据下载自OSM网站(https://www.openstreetmap.org/). OSM道路网络的完整性好且结构精细,道路之间具有良好的拓扑关系,因此比较适用于地块单元的划分.

研究中用于提取分类特征的数据包括宜出行数据与POI数据. 其中,腾讯宜出行大数据是一种新兴的土地利用分类研究数据,具有获取成本低、时空分辨率高等特点. 宜出行数据是通过追踪腾讯公司相关在线产品的位置信息,基于腾讯产品活跃用户的街道级位置定位而产生[34],主要包含了经度、纬度、时

表 1　数据介绍

数据用途	数据名称	数据来源
划分基本分类单元	OSM路网数据	https://www.openstreetmap.org/
特征提取	腾讯宜出行数据	微信宜出行小程序
	POI数据	百度地图
标定样本真实类别	城市总体规划方案	西宁市人民政府
	Google Earth高分影像	https://www.google.com/earth/
	高德地图	https://www.amap.com/
辅助数据	1∶400万行政边界矢量数据 1∶100万行政区矢量数据	http://www.dsac.cn/

间及人口热力强度 4 种属性. 宜出行数据以空间点的形式分时段展现了人口热力值的分布, 该数值与相同位置下的人口密度成正相关, 经过分析处理后可以比较直观地表征人口的空间分布以及动态变化, 能够有效区分不同城市土地的利用特征, 从而帮助获取比较精细的城市土地利用信息. 同时, 与其他社交媒体软件相比, 微信的用户量更大且分布广泛, 因此宜出行数据相较于其他社交媒体大数据可以在更大程度上反映人口分布的真实情况[35]. 目前, 宜出行数据已经在城市的土地利用分类[28]、功能分区[36]以及空间结构分析[37]等方面得到了初步应用. 本文所使用的宜出行数据是通过网络爬虫技术获取自腾讯微信宜出行小程序. 数据获取的时间为 2019 年 11 月 6 日, 数据展现了研究区在当天 24 h 内每小时的人口热力值分布情况, 空间分辨率为 25 m. POI 数据是一种同时包含区域功能属性和位置信息的空间点数据, 可以反映城市中不同属性建筑的空间分布特征. 与其他大数据相比, POI 数据融合了电子地图与点评数据等丰富的信息, 通常包含兴趣点名称、类别、地址、地理坐标等多个字段的属性[38]. 同时, 与常用的手机信令数据相比, POI 数据还具有价格低廉、容易获取的优点[18]. 目前 POI 数据在土地利用分类等方面的应用已经趋于成熟[39−40]. 本文使用的 POI 数据来源于百度地图, 获取时间为 2014 年. 数据共分为餐饮、旅游、休闲娱乐、科研教育、交通出行等 17 种类别.

研究中对训练样本与检验样本真实地类的标定, 主要参考了西宁市城市规划图、Google Earth 高分影像以及高德地图等. 其中: 西宁市总体规划图来源于《西宁市城市总体规划》, 该方案反映了西宁市在 2001—2020 年期间的城市规划概况; Google Earth 高分影像获取自 Google Earth 官方网站(https://www.google.com/earth/), 影像获取时间为 2019 年 9 月 11 日, 影像分辨率为 8 m; 研究参考的高德地图来自于高德官方网站的在线地图(https://www.amap.com/), 包含了研究区的建筑物、道路、水体等多种信息.

此外, 研究还用到了其他辅助数据, 包括全国 1∶400 万行政边界数据与 1∶100 万行政区数据等, 均获取自基础地理国情监测数据平台(http://www.dsac.cn/).

2　方法

首先, 基于分类单元划分、城市土地利用特征提取和城市土地利用分类 3 个步骤, 提取出了西宁市城市土地利用信息. 然后, 基于目视判读结果, 对提取出的城市土地利用信息进行了精度评价, 对比了不同数据源和分类方法的精度. 最后, 基于精度最高的城市土地利用分类结果, 分析了西宁市城市土地利用空间格局(图 1).

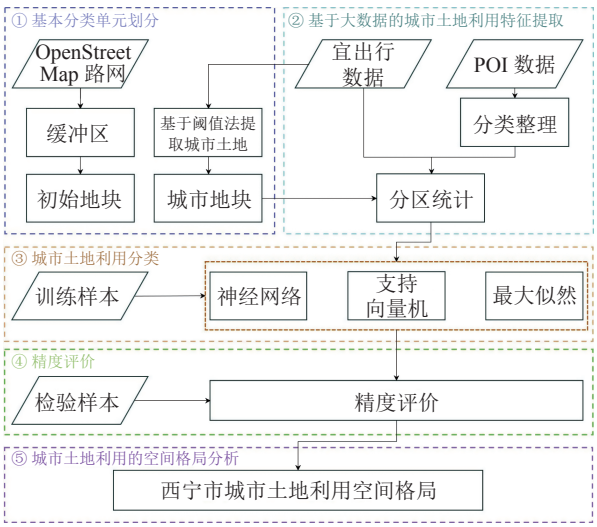


图 1　技术流程

2.1　基本分类单元划分　以城市地块为基本单元进行城市土地利用分类: 首先, 基于道路网络分割得到潜在的地块单元; 然后, 根据人口聚集情况提取得到城市地块(图 2). 参考 Gong 等[28]的研究, 根据全国城市车道宽度的平均水平, 基于 OSM 路网数据设置缓冲区, 以此作为道路的宽度. 利用道路网络对西宁市行政区的土地进行分割, 从而得到潜在的地块单元. 而后以地块为单位, 基于宜出行数据统计了各地块单元在一天内人口热力值的累积结果, 以人口热力值为主要评判依据, 参考 Dou 等[41]的研究, 通过选取最优分割阈值完成对城市地块的提取, 计算式为:

$$m_{ui}=\begin{cases}1, & H_{\mathrm{V}}\geqslant\lambda,\\ 0, & \text{otherwise},\end{cases}\tag{1}$$

式中: H_{V} 为地块的人口热力值; λ 为选取的最优阈值;

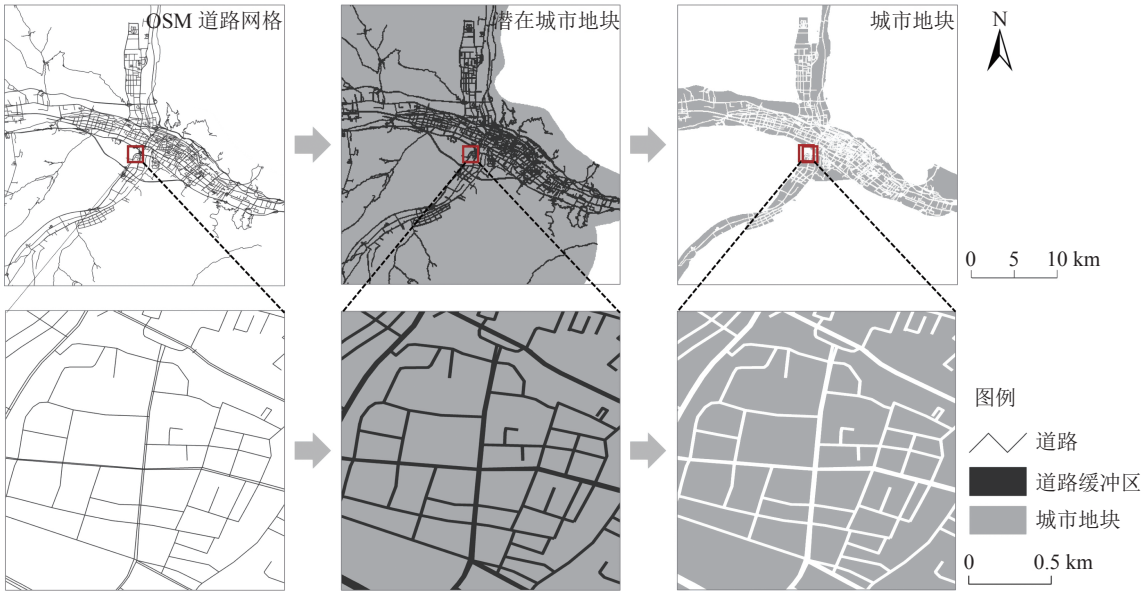


图 2 城市地块的划分与提取

m_{ui} 的值为 1,代表该地块属于城市地块,值为 0,代表该地块非城市地块.

2.2 基于大数据的城市土地利用特征提取 基于提取得到的城市地块单元,以研究区内各类兴趣点的分布密度以及人口热力值为特征进行分类. 首先参考 Gong 等[28]提出的 EULUC(essential urban land use categories)分类体系,结合研究区实际情况,将 POI 数据分为住宅、商业、公共设施、交通设施 4 类(表 2),而后基于城市地块统计各类兴趣点的分布密度. 对于宜出行数据,同样以城市地块为单位,分区统计各地块内部的人口热力累积值,从而获得逐小时人口的分布特征与动态变化特征.

表 2 城市土地利用分类体系说明

城市土地利用类型	备注
居住	公寓、住宅以及相应的服务设施用地
商业	各类商业活动用地,包括餐饮、旅馆、娱乐等;金融、媒体、保险、证券等办公用地;各类公司以及综合性商务办公楼宇用地
工业	工矿企业的生产车间、库房以及其他生产设施用地
公共	政府、军队以及其他公共安全服务用地;教育、科研用地;医疗、保健、卫生、防疫等急救设施用地;图书馆、博物馆、展览馆以及体育场馆等文化和体育服务设施用地;公园、绿地、风景名胜、旅游景区等娱乐和生态服务用地

2.3 城市土地利用分类 分类体系同样参考了 Gong 等[28]提出的 EULUC 体系. 结合研究区实际情况,本文主要基于居住用地、商业用地、工业用地以及公共用地 4 种一级用地类型展开分类. 为了探究城市土地利用信息提取的最佳数据与方法组合,本研究采用了单独基于 POI 数据、单独基于宜出行数据和 2 种数据源相结合等 3 种数据组合方式,以及最大似然法、支持向量机、人工神经网络 3 种常用的分类方法[42−44]完成分类.

其中,最大似然法分类的原理是通过一部分样本求解特征参数,而后通过计算和比较未知样本属于各类别的概率,最终将其归属于概率最大的一组类别中[45]. 其核心计算公式为

$$P_i = \frac{1}{\sqrt{(2\pi)^n |\boldsymbol{S}_i|}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\mu_i)^{\mathrm{T}} \boldsymbol{S}_i^{-1}(\boldsymbol{x}-\mu_i)\right], \quad (2)$$

式中 P_i 为特征向量 $\boldsymbol{x}$ 在第 i 个类别的概率密度,n 为波段数,μ_i 为第 i 类样本总体的均值,$\boldsymbol{S}_i$ 为第 i 类样本总体的协方差矩阵.

支持向量机分类的原理是将输入样本变换到更高维的特征空间,通过寻找最优超平面来实现样本的正确分类[41, 46]. 该方法可以有效实现非线性问题的分类,其最优分类函数表达式为

$$f(a) = \sum_{i=1}^{n} a_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j y_i y_j K(x_i, x_j), \quad (3)$$

式中 $x_i x_j$ 为训练样本,$y_i y_j$ 为样本所属类别,n 为样本总数,a 为函数解,$K(x_i, x_j)$ 为内积核函数. 核函数的选择是支持向量机应用的关键,本文选择了效果较好的径向基核函数[47]

$$K(x_i, x_j) = \exp\left(-\gamma \left\| x_i - x_j \right\|^2\right), \quad (4)$$

式中 x 为输入样本,γ 为参数,$\left\| x_i - x_j \right\|$ 为样本向量的范数.

人工神经网络则是一种模拟人脑识别图像属性的机器学习方法, 主要通过数学表达的方式模拟一系列人工神经元之间的信号传递, 而后通过各种线性和非线性运算得到输出层[48]. 其中, 反向传播(back propagation, BP)神经网络主要是采用训练样本迭代的方式, 通过使误差函数最小化来求解最优参数, 从而获得完整的运算模型以及最终分类结果. 本文构建的 BP 神经网络共包含 3 层结构, 即第 1 层输入层、第 2 层隐含层以及第 3 层输出层. 首先, 输入层可以将分类特征输入模型并传递至下一层. 输入层与隐含层之间的函数关系为

$$y_j = f\left(\sum_{k=1}^{n} w_{jk}x_k + a_j\right), \tag{5}$$

式中y_j为隐含层第 j 个神经元的输出值, x_k为隐含层第 k 个神经元的输入值, w_{jk}为输入层第 k 个神经元到隐含层第 j 个神经元之间的权值, a_j为隐含层第 j 个神经元的偏置项. $f(x)$为从输入层到隐含层的激发函数, 这里采用的是 logistic 函数

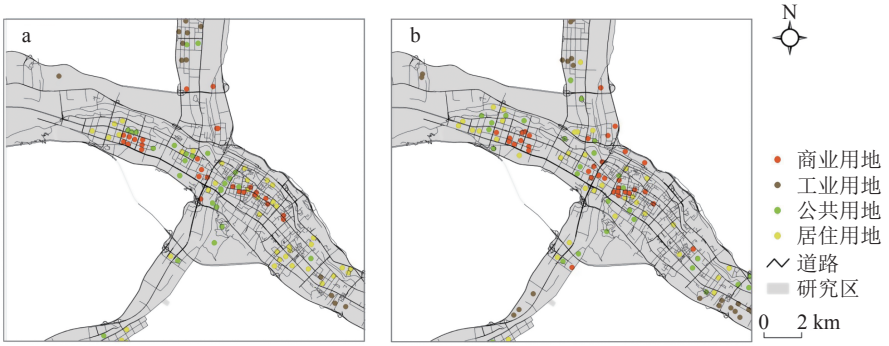
$$f(x) = \frac{1}{1+\exp(-x)}, \tag{6}$$

从隐含层到输出层的函数关系为

$$z_i = \sum_{j=1}^{m} w_{ij}y_j + b_i, \tag{7}$$

式中, z_i为输出层第 i 个神经元的输出结果, y_j为隐含层第 j 个神经元的输出结果, w_{ij}为隐含层第 j 个神经元与输出层第 i 个神经元之间的权值, b_i为输出层第 i 个神经元的偏置项.

通过随机抽样的方法选取训练样本, 用于分类模型的训练(图 3-a). 研究区共包含 1 231 个城市地块单元, 取 10% 数量的样本用于训练, 并且保证训练样本在 4 种地类中平均分布. 因此, 采用随机抽样的方法分别在 4 种地类中随机选取 30 个样本, 用于分类模型中参数的训练. 样本真实类别的标定参考了西宁市城市规划图、Google Earth 高分影像以及高德地图. 基于训练后的分类方法, 得到了 9 种不同数据与方法组合的分类结果.



a. 训练样本; b. 检验样本.

图 3　样本的空间分布

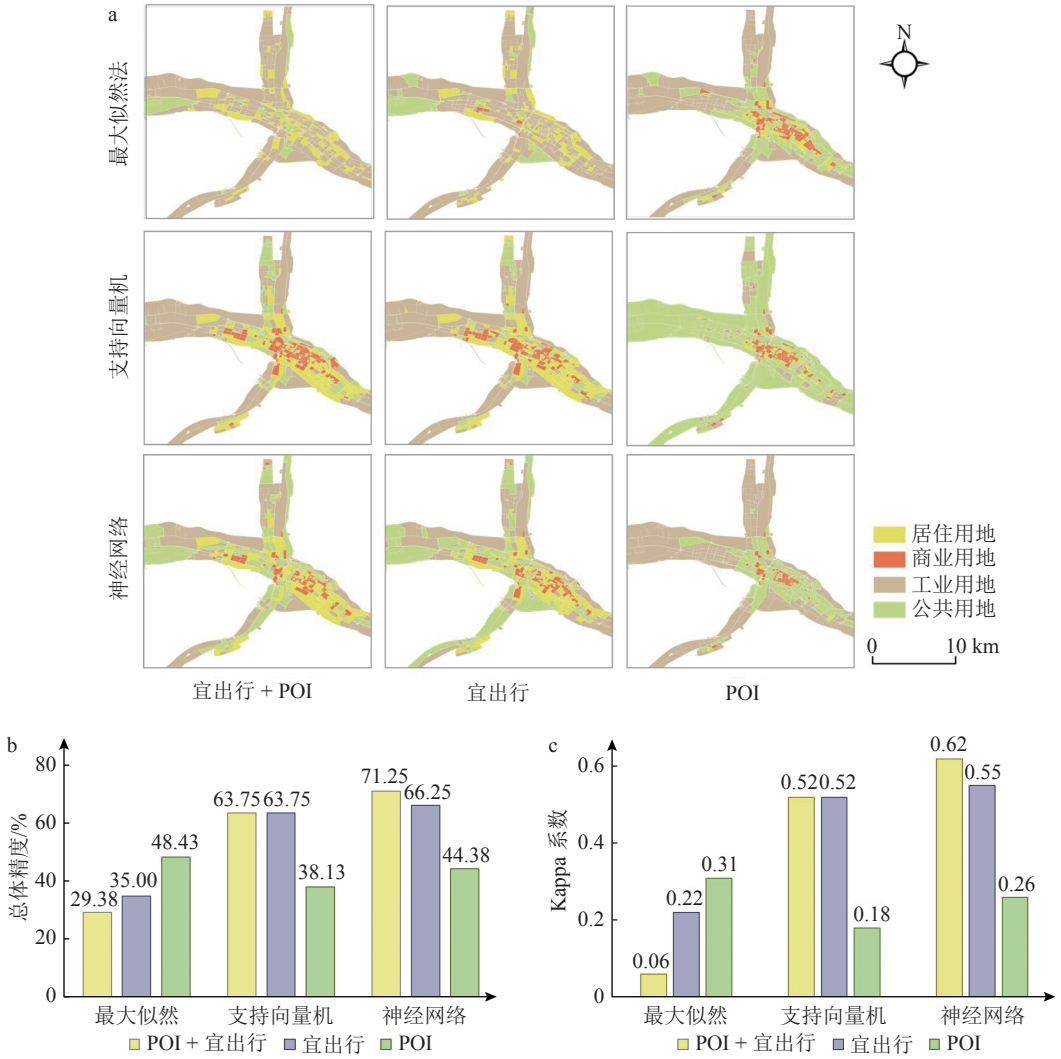
2.4　精度评价　参考翟天林等[10]的研究, 分别对不同数据和不同方法的分类结果进行了精度评价. 采取随机取样的方法, 在每种类别中随机选取 40 个检验样本, 共选取了 160 个样本点(图 3-b). 借助西宁市城市规划图、Google Earth 高分影像以及高德地图, 基于目视判读标定其真实类别. 通过构建混淆矩阵的方法, 分别得到各个数据与方法组合分类结果的总体精度与 Kappa 系数, 以及各地类的用户精度与制图精度.

2.5　城市土地利用的空间格局分析　基于分类结果, 对研究区的土地利用格局进行了总体分析与分区分析. 1)分别计算研究区全区和各地类的占地面积, 在此基础上比较不同用地类型在研究区内的面积占比. 2)参考渠爱雪等[49]的方法, 分别对西宁市城市土地利用格局进行了方位分区与圈层划分, 其中: 方位分区主要是按照东南西北 4 个方位, 在现有道路网络的基础上进行划分; 圈层的划分则以西宁市中心为中点, 分别在 1、3、5、8、15 km 处设置环状缓冲区, 依此对研究区进行分割. 3)分别对各分区内不同用地类型的占地面积与比例进行了统计, 一方面在单个分区内部对比不同用地类型的面积比例与分布位置, 另一方面则探究各用地类型在不同分区内占地面积和比例的变化规律.

3　结果

3.1　分类结果与精度　结合 POI 与宜出行 2 种数据源以及神经网络分类方法的城市土地利用分类总体效果最好(图 4). 精度评价结果显示: 基于 2 种数据源和神经网络分类的总体精度达到 71.25%, Kappa 系

数达到 0.62, 在不同数据源与方法的组合之中精度最高; 基于宜出行数据与神经网络方法的城市土地利用分类, 其总体精度为 66.25%, Kappa 系数为 0.55; 支持向量机处理宜出行数据或结合多源数据同样可以比较准确地实现 4 种用地类型的划分, 分类的总体精度均为 63.75%, Kappa 系数均为 0.52; 使用 POI 单一数据源的分类和基于最大似然法的分类效果较差, 其总体精度均<50%, 不能有效区分不同地类的特征.



a. 城市土地利用分类结果; b. 总体精度; c. Kappa 系数.

图 4 基于不同数据源与不同方法的分类结果对比

从不同用地类型的角度对比总体分类精度较高的 4 种数据方法组合发现(表 3~6), 在 4 种用地类型中: 工业用地的分类效果最好, 在总体精度>50.00%的分类组合之中, 基于 POI 和宜出行的神经网络分类对工业用地的错分和漏分都相对较少, 制图精度为 80.00%, 用户精度为 71.11%; 居住用地的分类效果仅次于工业用地, 在 4 种数据与方法的组合之中, 基于 POI 与宜出行的神经网络分类对居住用地的划分同样最为准确, 其制图精度为 77.50%, 用户精度为 75.61%; 公共用地与商业用地在 4 种地类中分类效果相对较差, 对于公共用地, 基于神经网络的方法分类效果较好, 但错分现象比较明显, 对于商业用地, 结合 2 种数据源的分类方法效果好于基于单一数据源的分类. 相对而言, 基于 POI 与宜出行的神经网络分类综合表现最好, 其中: 公共用地的制图精度为 70.00%, 用户精度为 60.87%; 商业用地的制图精度为 57.50%, 用户精度为 82.14%. 综上, 基于 POI 与宜出行的神经网络分类, 既能够达到较高的总体精度, 又可以对单一用地类型取得较好的分类效果, 是目前最佳的城市土地利用分类方式.

3.2 西宁市城市土地利用空间格局 西宁市主城区中: 公共用地占地面积最大, 约为 123.80 km^2, 是主城区总面积的 48.07%; 其次是工业用地, 为 93.01 km^2, 占总面积的 36.11%; 居住用地占地面积为 33.24 km^2,

表 3　基于宜出行数据的支持向量机分类混淆矩阵　　%

分类类别	居住用地	商业用地	公共用地	工业用地	总和	用户精度
居住用地	67.50	22.50	12.50	5.00	26.88	62.79
商业用地	15.00	50.00	15.00	0	20.00	62.50
公共用地	2.50	10.00	42.50	0	13.75	77.27
工业用地	15.00	17.50	30.00	95.00	39.38	60.32
总和	100.00	100.00	100.00	100.00	100.00	
制图精度	67.50	50.00	42.50	95.00		

注: 总体精度为63.75% , Kappa系数为0.52.

表 4　基于 POI 与宜出行的支持向量机分类混淆矩阵　　%

分类类别	居住用地	商业用地	公共用地	工业用地	总和	用户精度
居住用地	65.00	12.50	10.00	0	21.88	74.29
商业用地	22.50	62.50	17.50	0	25.63	60.98
公共用地	0	10.00	35.00	7.50	13.13	66.67
工业用地	12.50	15.00	37.50	92.50	39.38	58.73
总和	100.00	100.00	100.00	100.00	100.00	
制图精度	65.00	62.50	35.00	92.50		

注: 总体精度为63.75% , Kappa系数为0.52.

表 5　基于宜出行数据的神经网络分类混淆矩阵　　%

分类类别	居住用地	商业用地	公共用地	工业用地	总和	用户精度
居住用地	77.50	22.50	5.00	0	26.25	73.81
商业用地	7.50	40.00	5.00	0	13.13	76.19
公共用地	7.50	22.50	80.00	32.50	35.63	56.14
工业用地	7.50	15.00	10.00	67.50	25.00	67.50
总和	100.00	100.00	100.00	100.00	100.00	
制图精度	77.50	40.00	80.00	67.50		

注: 总体精度为66.25% , Kappa系数为0.55.

表 6　基于 POI 与宜出行的神经网络分类混淆矩阵　　%

分类类别	居住用地	商业用地	公共用地	工业用地	总和	用户精度
居住用地	77.50	15.00	10.00	0	25.63	75.61
商业用地	10.00	57.50	2.50	0	17.50	82.14
公共用地	2.50	22.50	70.00	20.00	28.75	60.87
工业用地	10.00	5.00	17.50	80.00	28.13	71.11
总和	100.00	100.00	100.00	100.00	100.00	
制图精度	77.50	57.50	70.00	80.00		

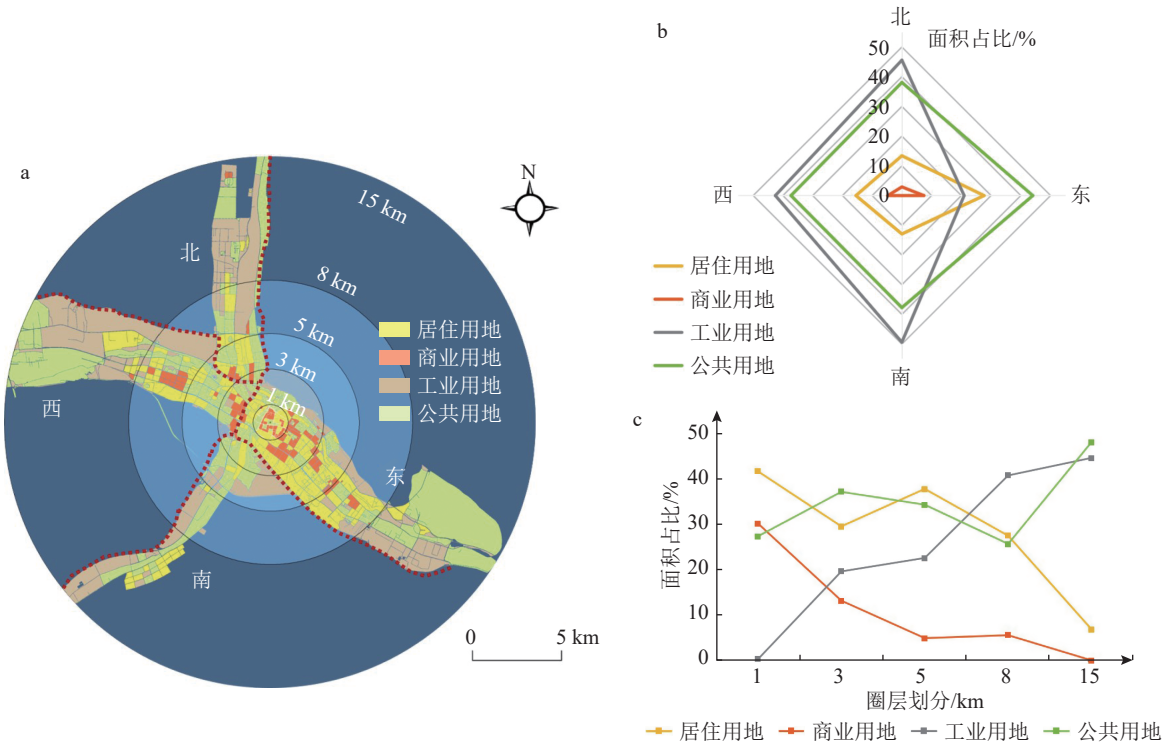
注: 总体精度为71.25% , Kappa系数为0.62.

占总面积的 12.91%; 商业用地占地面积最少, 为 7.49 km^2, 占总面积的 2.91%.

西宁市主城区的范围可以被大致划分为东西南北 4 个区域, 不同用地类型在各区域的分布存在显著的特征(图 5). 其中, 东部区域内土地利用方式多样, 商业与居住用地大多分布于这片区域. 在该区域内, 商业用地主要围绕一个明显的核心聚集分布, 居住用地则分布于商业区外围, 相对比较连续. 西部区域的

用地类型同样比较丰富. 其中, 商业用地在该区域内集中分布于东、西 2 片商业区, 居住用地则分布比较零散. 工业用地与公共用地在西区的面积占比远高于商业和居住用地. 南、北分区的土地利用格局比较相似, 工业用地与公共用地为主要的用地类型, 居住用地的分布面积相对较小, 并且基本没有商业用地在该区域分布.

由上分析可知: 公共用地在各方位的分区内均有广泛分布; 工业用地主要分布于北、西、南 3 个方位, 在东部区域的面积占比最低; 居住用地与商业用地在东、西 2 个方位分布最广, 而在南、北方位的分布面积则极小.



a. 方位分区与圈层划分; b. 各地类在不同方位占比; c. 各地类在不同圈层占比.

图 5 西宁市城市土地利用空间格局分析

西宁市各种土地利用类型的空间分布, 不仅在各个方位上具有不同的特征, 而且随着研究区域与城市中心距离的增加而呈现出明显的梯度分布规律(图 5): 在距离市中心 1 km 范围内, 城市土地利用类型主要为居住用地、商业用地和公共用地, 基本没有工业用地在此分布; 距离市中心 1~3 km 范围内, 城市土地利用类型比较多样, 且土地利用方式以公共用地、居住用地为主, 商业用地的面积比例虽然显著下降, 但分布比较密集, 主要形成东、西 2 片对称分布的商业聚集区; 在 3~5 km 范围内, 用地类型依旧以居住用地和公共用地为主, 商业用地面积占比继续下降, 工业用地的面积占比不断上升; 在 5~8 km 范围内, 工业用地面积占比显著上升, 并成为主要的用地类型; 同时在东西两侧对称位置出现了小型的商业聚集区, 商业用地面积占比略有回升; 距离市中心 8 km以外, 城市土地利用类型以公共用地和工业用地为主, 二者均呈片状连续分布, 仅存一些小型的居民点零散分布, 商业用地基本消失. 通过分析不同地类在各圈层内的分布规律可知, 随着研究区域与城市中心之间距离的增加, 商业用地的分布面积呈现显著的下降趋势, 而工业用地的分布比例则不断上升. 公共用地与居住用地在城市中心均为主要的土地利用方式, 随着研究区域逐渐远离城市中心, 二者的分布比例出现分异, 居住用地的分布面积逐渐减少, 而公共用地在 4 种土地利用方式中始终占据比较重要的地位.

4 讨论

4.1 结合多源大数据能够更加准确地提取城市土地利用信息 通过精度评价与实地调查验证得知(图 6), 结合兴趣点与宜出行数据有利于准确提取城市土地利用信息. 对于不同的城市土地利用类型, 单一数据源并不能全面展现其土地利用特征. 如果仅使用 POI 作为单一数据源, 则只能刻画城市中不同功能建筑的分布特征. 但是, 与工业、公共用地相比, 城市居住用

地及商业用地范围内通常建筑物分布比较密集、建筑的属性构成复杂, 同时还可能存在居住点、商业点混杂分布的情况. 因此, 2 种用地类型可能在建筑物属性及分布方面呈现出相似的特征. 除此之外, 公共用地内部的建筑特征差异较大, 医院、学校等公共用地附近常有居民点、商业点等混杂分布, 建筑物属性与居住用地、商业用地类似; 而公园、绿地等区域内 POI 分布密度相对较小、种类属性单一, 与工业用地存在相似之处. 将 POI 作为单一数据源参与城市土地利用分类的实际效果不够理想. 仅选择宜出行数据进行分类, 虽然能通过不同时间段人口分布的特征及其差异划分多数城市土地的利用类型, 但分类结果仍然存在较大误差. 例如, 学生、老师、工人、公司职员等人群在一天内的行为动态具有一定相似性, 在夜晚大多聚集于住宅区, 而工作时间则分别聚集于学校、工厂以及写字楼等区域, 这一现象在某种程度上模糊了公共用地、工业用地以及商业用地之间的差异. 因此, 将宜出行数据作为单一数据源进行分类, 也同样存在不足之处.

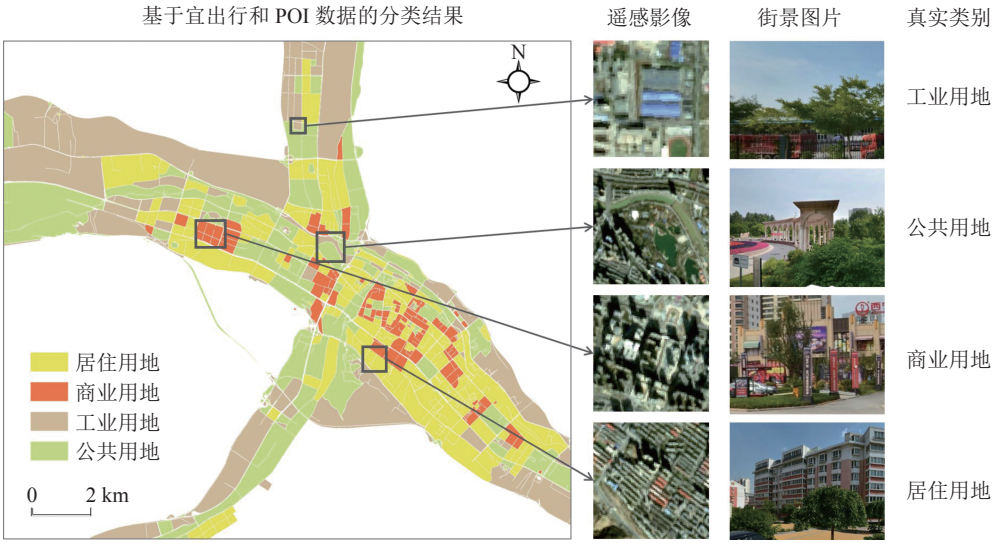
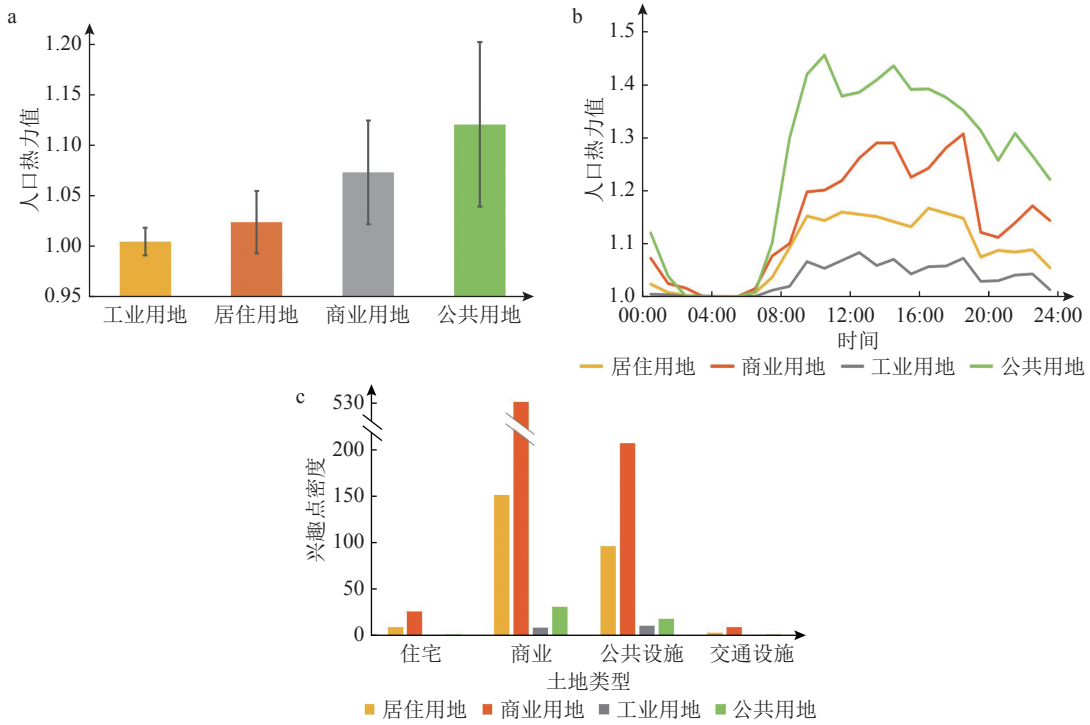


图 6　城市土地利用分类结果与真实地类对比

为了展示宜出行与 POI 数据所反映的城市土地利用特征, 以不同用地类型为单元, 分别统计了人口热力值在不同城市土地内部的平均值以及不同属性 POI 在各类城市用地中的分布密度(图 7). 结果显示, 在 4 类城市土地中, 公共用地在 24 h 内的人口热力平均值最高, 人口分布最为密集, 其次是商业和居住用地, 工业用地的人口分布相对比较稀疏. 与之类似, 4 类城市土地内部的人口热力平均值在大多数时段也具有相同的特征, 并且其时间变化曲线均具有一定的区分度. 总体来说, 在 06:00—24:00 之间, 4 种用地类型范围内的人口热力值普遍较高, 并且不同城市土地人口分布特征之间的区分度也比较明显. 其中: 公共用地的平均人口密度最高; 商业用地与居住用地的人口密度变化趋势相似程度较高, 但是在 09:00—20:00 之间, 商业用地范围内的人口明显更为密集; 相比之下, 工业用地范围内的人口密度最小, 变化趋势也比较简单, 但是在 00:00—06:00 之间, 4 种用地类型范围内人口热力平均值普遍较低, 并且变化趋势比较相似, 尤其是在 02:00—06:00 之间, 公共用地与商业用地的变化特征相似程度最高, 这会导致明显的混分现象. 在此基础上, 进一步结合不同属性 POI 的分布特征可以有效弥补宜出行数据的缺点, 突出公共用地、商业用地以及其他用地类型之间的差异. 在 4 种城市用地类型中, 居住用地与商业用地较之工业用地与公共用地的 POI 分布密度明显更高, 同时商业用地范围内的商业点分布密度显著高于其他用地类型. 因此, POI 数据同样能够为城市土地利用信息提取提供有效的分类特征, 并且在宜出行数据的基础上结合 POI 能够提高信息的丰富程度, 从而有效提高城市土地利用分类精度.

综上, 宜出行数据与 POI 数据分别反映了城市中人口与建筑的分布特征. 结合 2 种数据源, 可以使人口信息与建筑信息相互补充, 从而更加准确地提取城市土地利用信息. 此外, 通过对比不同分类方法处理多源大数据的效果可知: 最大似然法虽然简单易行, 但要求训练样本必须超过输入图像的波段数, 难以有效融合具有逐小时信息的宜出行数据和多类别的 POI 数据, 因此分类效果最差; 支持向量机算法可以基于



a. 24 h 内宜出行数据平均值和标准差；b. 不同时段宜出行数据平均值；c. 不同属性 POI 数量.

图 7　不同城市土地利用类型的数据特征对比

少量训练样本进行分类，分类精度明显优于最大似然法，但其基于大规模训练样本进行求解时效率较低，而且经典支持向量机算法只可用于 2 类的划分，在多类信息的提取过程中仍面临困难；人工神经网络具有良好的非线性映射能力、自学习和自适应能力以及容错能力，能够充分融合 POI与宜出行数据，从而具有最高的分类精度. 因此，基于 POI 和宜出行数据的神经网络分类方法是获取研究区高精度城市土地利用信息的一种有效技术手段，具有较好的应用推广潜力.

4.2　不足和展望　本文基于 POI 与宜出行 2 类常用的大数据以及 3 种常用的分类方法，探索了利用大数据快速、准确获取城市土地利用信息的途径. 其中，宜出行数据作为一种新兴的数据源，具有较高的时空精度，与 POI 数据结合可以有效提高城市土地利用分类的精度.

但是研究仍存在不足之处. 首先，西宁市作为青藏高原上规模最大的城市，目前正处于快速的城市化进程中. 在研究过程中发现，西宁市主城区周边存在大量正在建设之中的土地，这部分土地存在用地类型难以判定和土地利用特征不规律的问题，因此会对分类精度产生不利影响. 其次，研究区内不同属性 POI 的分布总量存在差异，一个住宅小区往往有大量商铺和公共设施，不同城市用地内 POI 的分布密度普遍呈现出商业点和公共设施分布密集、住宅和交通设施分布稀疏的特点. 同时考虑到地块内部具有明显的空间异质性，比如存在商住两用的地块，居住与公共用地内部分布有较大比例的商业点. 这些特点导致 4 类土地利用仍存在混分现象. 再次，城市地块划分所用到的道路数据不够精细也会影响分类的准确性：一方面，城区边缘的路网通常不够精细，当本文用人口密度提取城市地块时，容易因为城市边缘的地块划分过于粗糙而导致城市范围提取不够精确；另一方面，部分地块内部存在多种用地类型而未能被精准分割，这也给分类工作带来了不利的影响. 除此之外，分类所使用的宜出行数据作为一种社交媒体大数据存在一定程度的偏性问题；但是该数据的获取基于腾讯平台，考虑到腾讯公司的用户数量及其服务群体的普遍性，宜出行数据尚能够表征大部分人群的行为特征.

在未来的研究中，首先需要提升路网的精细程度，并借助面向对象图像分割技术，以提高单个地块内部土地利用方式的均质性. 同时，可以采用更加详细的城市土地利用分类体系，引入软分类的思想，对单个地块内不同土地利用类型的占比分别进行评定，以解决混合地块的用地类型难以判定的问题. 此外，在下一步研究中将引入手机信令数据，以进一步提高分类精度. 目前，本文仅对西宁市的城市土地进行了比较粗略的划分. 通过结合深度学习以及多源遥感数据，还可以进一步提升分类结果的精度和地类的精细程度.

5　结论

通过对比不同数据组合方式与不同分类方法下的城市土地利用分类精度, 探索了一条利用大数据快速、准确获取城市土地利用信息的途径. 结果显示, 结合宜出行与 POI 数据可以有效提升城市土地利用分类的精度. 而人工神经网络可以充分融合多源大数据信息, 从而实现最优的分类效果. 目前研究在基本分类单元以及分类体系的精细程度上仍存在不足, 未来可以借助更精细的道路数据来应对混合地块的问题, 同时可以结合深度学习与多源遥感数据, 实现更精细和更高精度的城市土地利用分类.

6　参考文献

[1] 许学强, 周一星, 宁越敏. 城市地理学[M]. 2版.北京: 高等教育出版社, 2009

[2] 王一航, 夏沛, 刘志锋, 等. 中国绿洲城市土地利用/覆盖变化研究进展[J]. 干旱区地理, 2019, 42(2): 341

[3] 刘纪远, 张增祥, 徐新良, 等. 21世纪初中国土地利用变化的空间格局与驱动力分析[J]. 地理学报, 2009, 64(12): 1411

[4] 赵涛, 郑新奇, 邓祥征. 城市土地利用优化配置分析应用: 以济南市为例[J]. 地球信息科学, 2004. 6(2): 53

[5] 简新华, 黄锟. 中国城镇化水平和速度的实证分析与前景预测[J]. 经济研究, 2010, 45(3): 28

[6] LEE T L. Action strategies for strengthening industrial clusters in southern Taiwan[J]. Technology in Society, 2006, 28(2): 533

[7] 宁晓刚, 刘娅菲, 王浩, 等. 基于众源数据的北京市主城区功能用地划分研究[J]. 地理与地理信息科学, 2018, 34(6): 42

[8] 李永乐, 吴群, 舒帮荣. 城市化与城市土地利用结构的相关研究[J]. 中国人口·资源与环境, 2013, 23(4): 104

[9] 赵萍, 傅云飞, 郑刘根, 等. 基于分类回归树分析的遥感影像土地利用/覆被分类研究[J]. 遥感学报, 2005, 9(6): 708

[10] 翟天林, 金贵, 邓祥征, 等. 基于多源遥感影像融合的武汉市土地利用分类方法研究[J]. 长江流域资源与环境, 2016, 25(10): 1594

[11] 匡文慧, 张树文, 刘纪远, 等. 城市用地空间信息分类与数字重建: 以长春百年城市内部用地变化为例[J]. 遥感学报, 2010, 14(2): 345

[12] 王彩艳, 王瑗玲, 王介勇, 等. 基于面向对象的北京市区城市内部用地信息提取[J]. 自然资源学报, 2015, 30(4): 705

[13] 李伟峰, 欧阳志云, 肖燚. 景观生态学原理在城市土地利用分类中的应用[J]. 生态学报, 2011, 31(3): 593

[14] DEWAN A M , YAMAGUCHI Y. Land use and land cover change in Greater Dhaka, Bangladesh: using remote sensing to promote sustainable urbanization[J]. Applied Geography, 2009, 29(3): 390

[15] SHALABY A, TATEISHI R. Remote sensing and GIS for mapping and monitoring land cover and land-use changes in the Northwestern coastal zone of Egypt[J]. Applied Geography, 2006, 27(1): 28

[16] DOUGLAS L. The Importance of 'big data': a definition [EB/OL]. Gartner, 2012-06-21

[17] 龙瀛, 毛其智. 城市规划大数据理论与方法[M]. 北京: 中国建筑工业出版社, 2019

[18] 薛冰, 李京忠, 肖骁, 等. 基于兴趣点(POI)大数据的人地关系研究综述: 理论、方法与应用[J]. 地理与地理信息科学, 2019, 35(6): 51

[19] LUM C, KIM K J, MAGLIO P P. Smart cities with big data: reference models, challenges, and considerations[J]. Cities, 2018, 82: 86

[20] 鲁国珍, 常晓猛, 李清泉, 等. 基于人类时空活动的城市土地利用分类研究[J]. 地球信息科学学报, 2015, 17(12): 1497

[21] 陈世莉, 陶海燕, 李旭亮, 等. 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[J]. 地理学报, 2016, 71(3): 471

[22] 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

[23] 钮心毅, 丁亮, 宋小冬. 基于手机数据识别上海中心城的城市空间结构[J]. 城市规划学刊, 2014(6): 61

[24] HU T Y, YANG J, LI X C, et al. Mapping urban land use by using Landsat images and open social data[J]. Remote Sensing, 2016, 8(2): 151

[25] ZHANG X Y, DU S H, WANG Q. Hierarchical semantic cognition for urban functional zones with VHR satellite images and POI data[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2017, 132: 170

[26] 韩昊英, 于翔, 龙瀛. 基于北京公交刷卡数据和兴趣点的功能区识别[J]. 城市规划, 2016, 40(6): 52

[27] 赵恒谦, 贾梁, 尹政然, 等. 基于多源遥感数据的北京市通州区土地利用/覆盖与生态环境变化监测研究[J]. 地理与地理信息科学, 2019, 35(1): 38

[28] GONG P, CHEN B, LI X, et al. Mapping essential urban land use categories in China (EULUC-China): preliminary results for 2018[J]. Science Bulletin, 2020, 65(3): 182

[29] 孙士杰, 孙群, 陆川伟, 等. 基于出租车上下客数据的城市功能区提取方法[J]. 测绘科学技术学报, 2019, 36(6): 91

[30] 季顺平, 田思琦, 张驰. 利用全空洞卷积神经元网络进行城市土地覆盖分类与变化检测[J]. 武汉大学学报(信息科学版), 2020, 45(2): 233

[31] 黎华, 王道飘, 刘博源, 等. 元胞自动机和粗集支持下的土地利用变化研究[J]. 华中师范大学学报(自然科学版), 2018, 52(6): 910

[32] 陈克龙, 苏茂新, 李双成, 等. 西宁市城市生态系统健康评价[J]. 地理研究, 2010, 29(2): 214

[33] 西宁市统计局. 西宁市2019年国民经济和社会发展统计公报[A]. 西宁: 西宁统计年鉴, 2020
[34] 刘云舒, 赵鹏军, 梁进社. 基于位置服务数据的城市活力研究: 以北京市六环内区域为例[J]. 地域研究与开发, 2018, 37(6): 66
[35] 申犁帆, 王烨, 张纯, 等. 轨道站点合理步行可达范围建成环境与轨道通勤的关系研究: 以北京市44个轨道站点为例[J]. 地理学报, 2018, 73(12): 2423
[36] CHEN Y, LIU X, LI X, et al. Delineating urban functional areas with building-level social media data: a dynamic time warping (DTW) distance based k-medoids method[J]. Landscape & Urban Planning, 2017, 160: 48
[37] 段亚明, 刘勇, 刘秀华, 等. 基于宜出行大数据的多中心空间结构分析: 以重庆主城区为例[J]. 地理科学进展, 2019, 38(12): 1957
[38] ZHU X Y, ZHOU C H. POI inquiries and data update based on LBS[C]. Information Engineering and Electronic Commerce. Washington: IEEE Computer Society, 2009
[39] 文聪聪, 彭玲, 杨丽娜, 等. 主题模型与SVM组合的小尺度街区用地分类方法[J]. 地球信息科学学报, 2018, 20(2): 167
[40] 武新宇, 孙立双, 谢志伟, 等. 遥感影像与POI数据相结合的城市建成区提取适用性研究: 以沈阳市为例[J]. 测绘通报, 2019, 512(11): 142
[41] DOU Y Y, LIU Z F, HE C Y, et al. Urban land extraction using VIIRS nighttime light data: an evaluation of three popular methods[J]. Remote Sensing, 2017, 9(2): 175
[42] 张兵. 高光谱图像处理与信息提取前沿[J]. 遥感学报, 2016, 20(5): 1062
[43] 杜培军, 夏俊士, 薛朝辉, 等. 高光谱遥感影像分类研究进展[J]. 遥感学报, 2016, 20(2): 236
[44] 杨立民, 朱智良. 全球及区域尺度土地覆盖土地利用遥感研究的现状和展望[J]. 自然资源学报, 1999, 14(4): 340
[45] 骆剑承, 王钦敏, 马江洪, 等. 遥感图像最大似然分类方法的EM改进算法[J]. 测绘学报, 2002, 31(3): 234
[46] 田源, 塔西甫拉提·特依拜, 丁建丽, 等. 基于支持向量机的土地覆被遥感分类[J]. 资源科学, 2008, 30(8): 1268
[47] 王新明, 梁维泰, 周方, 等. 基于支持向量机和Getis因子的高分辨率遥感图像分类[J]. 地理与地理信息科学, 2008, 24(4): 16
[48] 卢文路, 刘志锋, 何春阳, 等. 基于Sentinel-1A合成孔径雷达数据和全卷积网络的城市建设用地监测方法研究[J]. 干旱区地理, 2020, 43(3): 750
[49] 渠爱雪, 卞正富. 徐州城市建设用地空间格局特征及其演化[J]. 地理研究, 2011, 30(10): 1783

Urban land use classification based on big data: case of Xining

DAI Yihua[1)] LIU Zhifeng[1,2)†] WANG Yihang[1,2)] YANG Zhipeng[3)]

(1)Faculty of Geographical Science, Beijing Normal University,100875, Beijing , China;
2)Center for Human-Environment System Sustainability (CHESS), State Key Laboratory of Earth Surface Processes and Resource Ecology (ESPRE), Beijing Normal University, 100875, Beijing , China;
3)National Natural Science Foundation of China, 100085, Beijing, China)

Abstract Urban land use is the result of interactions among social, political, economic, technological and other factors within and without cities. Urban land use classification not only helps to analyze land use pattern, but also has great significance for rational urban zoning and promotion of sustainable development. Urban land use classification in Xining is done based on two types of commonly used big data (Easygo, points of interest or POI) and three common classification methods (Maximum Likelihood, Support Vector Machine, Artificial Neural Networks). By comparing the accuracy of results under different data and methods, optimal data combination and classification method for extracting urban land use information are determined. The classification results are used to analyze urban land use patterns in Xining. Urban land use information obtained by neural network classification method based on Easygo and POI was found to have the highest accuracy, with overall accuracy at 71.25% and a Kappa coefficient at 0.62. Easygo and POI can reflect more information about characteristics of different land use. Artificial Neural Networks can fully integrate information of multi-source big data. Therefore, it provides a potential way to timely and accurately obtain urban land use information with multi-source big data and Artificial Neural Networks.

Keywords urban land use; big data; point of interest; Easygo; machine learning

【责任编辑: 刘先勤】