

地理大数据中POI数据质量的评估与提升方法

薛冰^{1,2}, 赵冰玉^{1,2,3}, 李京忠^{2,4}

(1. 中国科学院沈阳应用生态研究所, 沈阳 110016; 2. 辽宁省环境计算与可持续发展重点实验室, 沈阳 110016; 3. 德国柏林工业大学规划建筑学院, 德国柏林 10623; 4. 许昌学院城乡规划与园林学院, 许昌 461000)

摘要: 地理大数据实现对区域人地系统的精细刻画, 为研究人地关系和区域发展等提供新的数据。当前, 地理大数据进入了广泛应用, 但一直缺乏对其质量的考察及相应的评估方法。兴趣点(POI)数据是地理大数据重要组成部分, 对基于位置服务和区域场景理解具有重要作用。本文提出POI类大数据评估与提升方法, 基于场地调研、GIS等方法从地物识别完整率、数据冗余率和空间位置准确率3个维度实现质量评估; 基于数据生产过程发现和总结数据质量的可能影响因素, 证明多源数据融合是提升POI数据质量的有效手段。研究发现, 基于API接口获取的高德数据量略高于百度, 空间位置准确率相当, 冗余率较低; 高德侧重识别地物入口, 适于可达性等分析; 百度侧重发现非标志性地物, 适于空间规划等分析; 发现、采集和处理阶段是降低数据质量的可能环节, 受数据保护机制影响, 数据质量与获取量及面积成反比; 多源异构地理大数据质量评估、提升与融合是提升数据“涌现价值”、促进多学科交叉融通、解决新时代地理学问题的关键途径之一。

关键词: POI数据; 地理大数据; 数据质量评估; 场地调研; GIS

DOI: 10.11821/dlxb202305014

1 引言

地理大数据实现对区域人地系统的精细刻画, 为研究人地关系和区域发展等提供新的数据语境^[1-2]。当前, 大数据管理机制及实施流程尚未完善, 易出现地理位置信息缺失等质量问题, 降低数据效率和应用深度^[3]。POI数据作为基于位置服务的底层关键数据, 记录地理实体所承载的人类活动及与地理位置的相互关联性^[4-5], 具有一定的扩展性和丰富的应用场景。随着互联网和基于位置服务的发展, POI数据信息纵深和应用场景得到长足发展^[4, 6], 从对地物基本信息记录转向于跨领域属性综合集成, 为解决空间格局、人类活动、区域综合等关键地理问题提供数据支撑^[6-8], 如甄峰等^[9]基于POI等数据实现城市内部空间结构及影响因素分析; Wang等^[10]基于POI等数据探索可持续通勤模式; 浩飞龙等^[11]基于POI数据实现城市复合功能检测。

高质量数据是有效决策和高效规划的先决条件^[12-13]。目前, POI数据采集、存储和处理能力有限, 多源数据坐标系及分类标准不同, 降低数据质量及应用效率^[14]。数据质量

收稿日期: 2022-09-13; 修订日期: 2023-03-12

基金项目: 国家自然科学基金项目(41971166); 辽宁省“兴辽英才计划”项目(XLYC2007201); 中国科学院区域发展青年学者项目(2021-003) [Foundation: National Natural Science Foundation of China, No.41971166; Liaoning Xingliao Yingcai Program, No.XLYC2007201; CAS Young Scholar of Regional Development, No.2021-003]

作者简介: 薛冰(1982-), 男, 江苏灌云人, 研究员, 博士生导师, 主要从事人地关系分析与区域可持续发展治理研究。
E-mail: xuebing@iae.ac.cn

管理通过及时追踪缺失和可疑数据,降低不确定性与未知性,发现和挖掘数据质量受损的可能原因,提出有效途径,提升数据质量和完整性^[15-16]。如薛冰等^[17]基于城市功能区识别结果反向评估POI数据质量;Wu等^[18]基于Bi-STAN模型识别用户在特定时间段内访问的位置并插入数据,减少POI数据的缺失率;Zhang等^[19]借助图像深度学习提供POI数据,减少数据采集环节的误差。

场地调研是了解场地概况和获取真实有效数据的重要手段^[20-21]。对比调研和样本数据是衡量样本有效性、精准度和覆盖量的有效方法之一^[22-23]。基于场地调研和对比分析的数据质量评估是说明数据数字化程度、采集特点及误差、遗漏情况,揭示多源数据地理及非地理属性特征的重要途径^[24-25]。Agnieszka等^[26]基于摄影测量工具获取调研数据验证土地利用数据可信度;Grant等^[27]基于现场调研数据验证无人机图像中树苗的准确性;Yang等^[28]基于实地调研验证ENVI-met数值模拟软件的可靠性。

基于此,本文提出POI类地理大数据质量评估与提升方法(图1),基于数据获取与评估、问题发现与管理、质量提升与管理等一系列措施全面评估及提升数据质量。在数据采集和预处理后,根据调研的可行性、产业发展完备度及多样性等选择研究区。基于POI特性及数据质量一般评估准则,构建质量指标体系,包括空间位置准确率、地物识别完整率和数据冗余率,综合场地调研和GIS等手段创建评估方法体系,最终基于生命周期分析偏差原因,提出数据质量提升有效途径,为自然语言、地理音频及视频等多源地理大数据质量评估及提升方案和案例参考,为地理大数据的高效利用提供基础支撑。

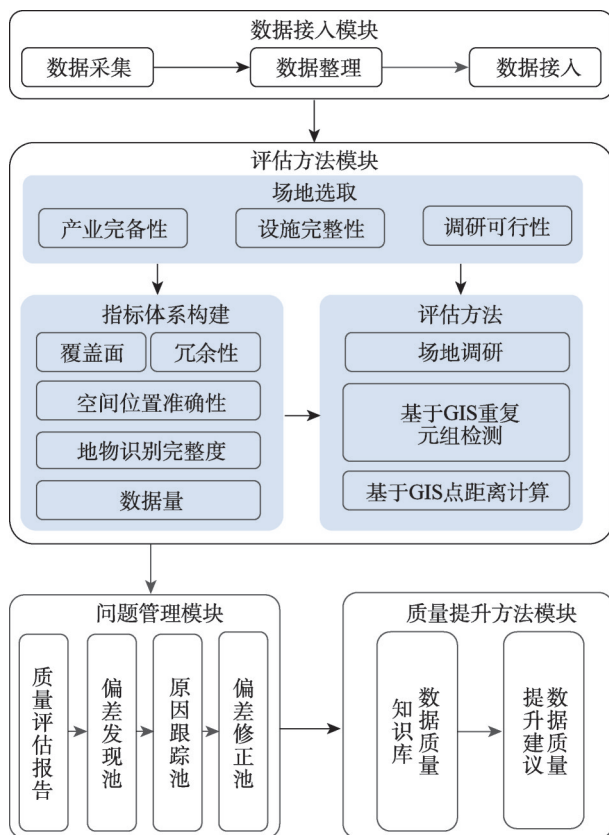


图1 POI类地理大数据质量评估与提升方法示意图

Fig. 1 Schematic diagram of the quality assessment and enhancement method for POI-type geographic big data

2 数据与方法

2.1 POI数据获取

2021年12月在中国知网以POI数据为数据源的文章数量共计32篇,其中以高德地图为数据源(简称“高德”)文章24篇,占比75%,以百度地图为数据源(简称“百度”)文章5篇,占比15.63%,两者占比90.63%。故选择百度地图和高德地图POI进行质量评估。POI数据基于WEB API接口获取,百度获取方式是“以圆形区域检索”和“以行政区检索”,高德则是“周边检索”和“以行政区检索”,获取时间是2021年12月

9日—2021年12月18日。预处理包括清洗和坐标系投影转换等。清洗是删除研究区以外数据,包括:①地物及POI采集点均位于研究区外;②地物位于研究区外,POI采集点位于研究区内。坐标系投影转换包括地理坐标系转化投影。本文分别借助百度数据源及高德数据源API接口,将坐标系进行解密,转换为WGS1984,并基于沈阳地理位置,将WGS 1984投影为WGS 1984 UTM ZONE 51N。

2.2 场地选取

城市选择辽宁省沈阳市。沈阳作为东北地区和沈阳都市圈的中心城市,经济产业较为发达,具备沈阳故宫等多个名胜古迹,北陵公园等多个公园绿地以及东北大学等多所高校,城市要素种类齐全,具备评估全样本、全行业POI数据的条件^[29-30]。因考虑社区发展完整性、公众关注度及场地调研可行性等因素,选择青年公园、沈阳体育学院(简称“沈体”)和中国医科大学第一附属医院(简称“医大一”)作为评估场地。

2.3 指标体系构建

评估维度包括地物识别完整率、数据冗余率和空间位置准确率。地物识别完整率是指地物的数字化程度,反映数据获取的覆盖面和精准度,基于场地调研了解地物的地理位置及基本社会属性,明确数据采集的特点及偏向性^[31];数据冗余率是检测数据库元组的重复率,反映数据的有效性和可用性,通过检测一定地理范围内重复元组,降低因地物重名等产生的重复率,提升数据评估准确性^[32];空间位置准确率表示数据经纬度坐标与真实位置的接近程度,一定程度上决定分析和决策层面的可信度,通过统计经纬度坐标的偏差量、偏差区间以及偏差率,评估数据的误差程度^[33]。

2.4 评估方法

(1) 场地调研

场地调研是了解地物实际空间位置及本质信息,掌握第一手真实资料的有效手段^[34]。本文基于观察法和询问法获取地物的数量、地理及非地理属性等,为精准评估数据质量提供支撑。流程为:①研究区格网化处理。因研究区内建筑物数量较多,为防止调研期间出现遗漏及重复等问题,进行格网化处理,依据研究区面积、地物丰富度和调研执行度等因素,沈体为100 m×100 m,青年公园为50 m×50 m,医大一为25 m×25 m。②调研信息标记及预处理。调研主要借助GPS等工具记录地物的空间信息,询问法了解地物更新情况。预处理是指数据接入、清洗及数字化。本文基于转换器,将.gpx格式转换为.shp格式,糅合询问所得信息,数字化调研信息。调研时间是2021年12月20—27日。

(2) 基于GIS重复元组检测模型

该模型用于测算冗余率。因城市存在相似地物,若仅对比非地理属性相似度,可能提高数据冗余率,因此本文在对比非地理属性基础上,借助GIS邻域分析进行点距离计算,认为一定空间距离内相似POI数据即为重复元组,即“名称”“类别”字段相同,空间位置接近的样点。流程:对样本进行循环遍历,寻找相似元组;计算重复元组间的空间距离;结合场地调研确定重复元组可能性。根据重复元组占比计算冗余率。

$$P_r = \frac{\sum_{i=0}^n (C_i - 1)}{C_{sum}} \quad (1)$$

式中: P_r 指冗余率; C_{sum} 指元组总数; i 指元组索引; n 指总量; $(C_i - 1)$ 指第 i 个元组重复数。

(3) 基于GIS点距离计算模型

该模型用于计算偏差点的偏差距离。本文通过计算调研与样本数据间的欧氏距离确

定POI数据的偏离距离、区间及特征,为挖掘数据采集阶段产生误差的可能因素提供基础支撑。公式为:

$$Ds = \sqrt{(s_x - p_x)^2 + (s_y - p_y)^2} \quad (2)$$

式中: Ds 表示偏差距离; (s_x, s_y) 表示场地调研获取的地物经纬度坐标; (p_x, p_y) 表示POI数据的经纬度坐标。

(4) 比值法

本文借助比值法 (Count Ratio, CR) 辅助计算数据地理及非地理属性准确率:

$$A_i = \frac{a_i}{S_i} \times 100\% \quad (3)$$

$$D_j = \frac{d_j - 1}{S_j} \times 100\% \quad (4)$$

式中: A_i 表示第 i 项指标的空间位置准确率和数据冗余度; D_j 表示第 i 项指标的地物识别完整率; a_i 或 d_j 表示基于第 i 或 j 项指标内容, 获取的正确元组个数; S_i 表示基于第 i 项指标内容, 获取的样本总数。

3 数据质量评估

本文基于场地调研获取地物共 126 个, 包括青年公园 35 个, 沈体 48 个, 医大一 43 个, 结合 GPS 定位器、观察及询问法等记录地物的空间位置及地物建造、翻新、营业及开放等非地理属性。同时基于 GIS 平台等实现 POI 与调研数据的对比分析, 包括检测元组重复率、计算偏差距离等, 发现高德及百度数据源在冗余率、地物识别完整率及空间位置准确率 3 个维度的特征:

冗余率特征 (图 2、表 1): ① 冗余率均较低, 高德数据源略低于百度数据源。共获取高德数据 83 条, 重复数据 1 条, 占数据总量 1.2%, 占所在地物数据总量的 3.4%; 百度数据源数据 75 条, 重复数据 1 条, 占数据总量的 1.3%, 占所在地物数据总量的 4.3%。② 重复数据易出现在边界不清晰或名称不明晰的地物。如青年公园休息区具有面积较大且边界不清晰的特征, 易导致采集人员多次采集; 医大一 4 号楼病房, 可称为病房或 4 号楼病房, 在融合或更新阶段, 未能实现多个元组融合或新旧数据的更替。

地物识别完整率特征 (图 2、表 1): ① 高德地物识别完整率高于百度。高德数据量大于百度, 最大数据差存在于医大一, 共 3 个, 最小数据差存在于沈体, 共 1 个。② 高德对地物的入口识别度更高, 识别青年公园入口 12 个, 占比 85.71%; 百度数据源识别青年公园入口 4 个, 占比 28.57%。③ 百度数据源对同一地物多种属性识别率更高, 百度数据源识别医大一急诊楼 4 种属性, 包括急诊、卒中中心、住院部和消毒供应中心生活区。④ 地物识别完整率与地物对外开放度成正比。3 个场地的社会属性与开放程度不同, 青年公园、医大一和沈体分别为城市绿地, 医疗服务和教育服务, 对社会公众开放程度依次从高到低, 数据采集难度依次上升, 地物识别完整率下降。

空间位置信息不准确共有 3 种场景 (图 3): ① 位于地物周边道路或空地等, 距离不超过 50 m (图 3a~3c), 如青年公园东 4 门 (高德) 和医大一 1 号楼 (高德)、沈体游泳馆 (百度)、医大一卒中中心 (百度)。② 位于地物周边的其他地物上, 距离不超过 50 m, 可能导致地物功能错乱 (图 3d~3f)。如医大一体检中心、3 号楼和 3 号楼病房 (高德)。③ 位于地物周边, 距离区间是 100~200 m (图 3g~3i), 属于错误位置信息, 如沈体

表1 网络地图与调研数据对比
Tab. 1 Comparison of e-map and survey data

地点	来源	数据量(条)	地物识别完整率(%)	重复数量	冗余率(%)	误差数	位置准确率(%)
青年公园	高德	26	71.43	0	0	1	96.15
	百度	23	62.86	1	4.30	0	100.00
沈体	高德	27	54.17	0	0	2	92.60
	百度	26	52.08	0	0	3	88.46
医大一	高德	29	65.17	1	3.40	4	86.21
	百度	26	58.14	0	0	4	84.62

注：青年公园实地调研的数据量为35条；沈阳体育学院实地调研的数据量为48条；中国医科大学第一附属医院实地调研的数据量为43条。



图3 空间位置偏差示意

Fig. 3 Schematic diagram of spatial position deviation

图书馆（高德和百度），分别距离实际地物165 m和105 m，沈体北门（高德），距离实际地物102 m，无法正确表达地物的真实信息。

数据量特征（图2、表1）：① 基于宏观角度，高德数据量远大于百度，是其7.82倍。根据“以行政区划区域检索”，获取高德POI数据411501条，百度POI数据52591条。② 对比“以圆形区域检索或周边搜索”数据获取方式，发现返回的数据量和数据精准度相差较大，以青年公园为例，百度数据“以圆形区域检索”获取数据是以“以行政区划区域检索”获取数据的3.83倍，“以行政区划区域检索”无重复和存在误差数据；高

德数据“以周边搜索”获取数据是“以行政区划区域检索”获取数据的1.63倍,“以行政区划区域检索”返回空间位置误差数据2条。此外,对比两种数据源面积较大地物的采集点一般位于地物中心;面积较大场地内商家门店识别率较低,如沈体内存在瑞星咖啡、理发店和菜鸟驿站等商业服务业设施,未能成功识别。

4 数据质量影响因素

基于质量评估结果,基于API接口返回的数据均存在不同程度的质量问题,如存在重复数据、地物及其属性未能完全识别、空间位置获取有误、数据量受数据获取方式影响等。本文基于数据生命周期,发现并总结影响数据质量环节及因素,为提高数据质量及应用效率提供技术支撑和可行性建议。

生产过程包括发现、采集、处理和发布4个阶段^[35]。发现和采集主要有3种方式:采集员实地调查、众包和基于遥感影像等数据的自动识别^[36]。实地调查是采集员通过驾驶车辆或步行形式发现地物,借助采集设备对获取地物详细信息并传回数据中心,采集员可能未发现全部地物及其属性,降低地物识别率,亦因采集员记录地物空间位置习惯差异,降低空间位置准确性^[37]。众包是通过浮动车或用户发现并向数据中心反馈数据,地物发现率及用户对地物的定位习惯影响地物识别率及空间位置准确性^[38]。基于遥感影像等数据的自动识别在一定程度上提高空间位置准确性,却无法识别地物的多种属性^[39],如沈体同一栋教学楼设立成人教育部、社会体育学院等多个部门(图2b)。

处理阶段包括数据接入、标准化、判重、融合4个程序^[40](图4)。因来源及内容多样性,须先进行规范化处理,将其转化为可处理的格式。在接入后,进行标准化处理,包括标准化属性字段及验证数据行政区划正确性^[41]。判重是将新接入数据与原有数据进行对比,构建模型判断相似度,当相似度达到阈值则为重复,进行数据的融合与更新^[42];若

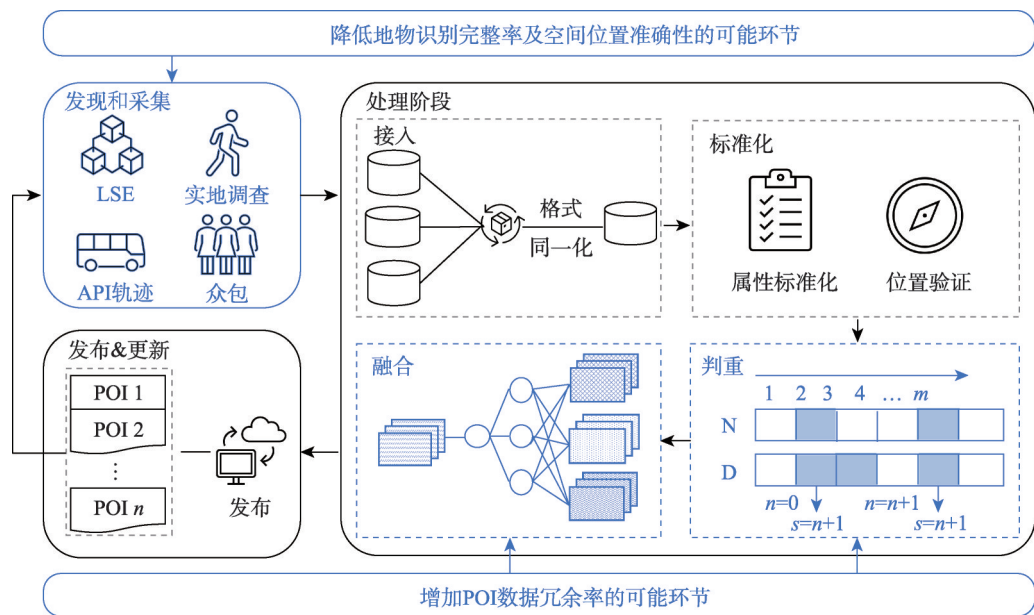


图4 影响POI数据质量的可能环节示意
Fig. 4 Schematic diagram of possible links that affect POI data quality

无重复数据,将新增数据添加至数据库。此过程可能因同一地物多次采集信息有异,无法将其判定为重复数据,提高冗余率。融合是将多源数据合并满足不同业务需求,如与其他平台对接获取扩展属性等^[43]。此阶段因多源数据描述不同,提升数据冗余率,如沈体(高德)中的住院部和4号楼住院部(图2a)。数据更新和发布是一个长久且持续过程。电子地图商会根据地物更新进行更新和融合。

基于此,冗余数据可能产生在处理阶段中的判重和融合程序(图4),原因是同一地物的多次采集内容相差较大,相似程度降低;地物识别完整率下降可能发生在发现和采集阶段(图4),原因是采集人员或用户未能发现地物及其全部属性;空间位置获取有误的可能发生在采集阶段(图4),原因是采集员定位习惯有异。受数据保护机制影响,用户无法通过API获取全部数据,数据质量受损,与搜索面积成反比^[44]。

5 数据质量提升方法

多源数据丰富度和信息完善度不同。在数据质量一定的情况下,获取更丰富与完整的数据信息,提升数据质量是研究者的迫切需求之一。基于多种数据源,抽取精准度较高的数据进行融合处理,是提高数据质量的有效途径之一。

本文以高德和百度数据源为例进行数据融合,尝试说明多源数据融合在提升数据质量和丰富数据信息完整性的可行性。因考虑到多源数据存在重复元组、空间位置准确率的不确定性及属性组织结构的差异性,本文对获取的数据进行初步处理:①删除重复数据,如医大一4号楼病房与病房(高德、图5a);②选择空间位置准确率较高的数据,若多源数据均识别某地物,选取空间位置偏差小的数据,如保留医大一卒中中心(高德),删除医大一卒中中心(百度、图5a);③属性信息合并,将多源数据扩展属性合并,提高属性完整度,如高德数据的商家电话等。

多源数据融合后,质量得到一定提升(图5、表2)。表现在:①地物识别完整率显著提高,融合后数据地物识别完整率提升14.58%~32.56%,百度提升16.67%~32.56%,高德提升14.58%~25.53%,说明两种数据源采集的地物存在一定差异性;②有效避免重复数据,通过判别多源数据相似度并建立合集,有效消除原数据的重复数据;③空间位置准确率得以提升,空间位置准确率提升0%~8.6%,百度提升0%~8.6%,高德提升0.99%~7.4%,说明百度数据的空间位置准确率浮动较大。

基于此,证明多源数据融合方法是降低重复率,提升地物识别完整率和空间位置准确率的有效途径。目前已有研究基于POI数据的空间位置和属性信息进行大体量数据的融合,如张巍等^[45]在空间位置属性的基础上借助非空间属性相似度进行多源数据融合;吴张峰等^[46]通过构建母库融合多源POI数据形成内容规整、信息量丰富的融合库等。但如何匹配拟链接对象,如何设置容差以及如何验证融合后数据质量等是当下及未来很长一段时间需要关注的问题^[47]。

6 总结与展望

本文基于微观视角构建POI大数据质量评估与提升方法,包括“数据获取与评估→问题发现与总结→质量提升及管理”等多个逐层递进、相辅相成的模块。在数据获取与评估模块,借助场地调研、GIS等方法从地物识别完整率、数据冗余率和空间位置准确



图5 融合数据与实际调研数据对比

Fig. 5 Comparison of fusion data and survey data

表2 网络地图与融合后数据对比
Tab. 2 Comparison of e-map and fusion data

地点	来源	数据量(条)	地物识别完整率(%)	重复数	冗余率(%)	误差数	位置准确率(%)
青年公园	融合	34	94.29	0	0	1	97.14
	高德	26	71.43	0	0	1	96.15
	百度	23	62.86	1	4.30	0	100.00
沈体	融合	34	68.75	0	0	1	97.06
	高德	27	54.17	0	0	2	92.60
	百度	26	52.08	0	0	3	88.46
医大一	融合	40	90.70	0	0	5	87.50
	高德	29	65.17	1	3.40	4	86.21
	百度	26	58.14	0	0	4	84.62

注：青年公园实地调研的数据量为35条；沈阳体育学院实地调研的数据量为48条；中国医科大学第一附属医院实地调研的数据量为43条。

率3个维度实现POI数据质量评估，为评价数据质量状态、考察数据在应用层面满足程度提供支撑。在问题发现与总结模块，基于数据生产过程发现和总结数据质量的可能影响因素，为整改生产流程和形成智慧生产提供基础。在质量评估及管理模块，多源数据融合是提升POI数据质量的有效手段，是提高数据高效应用和有效决策的重要支撑。

大体量数据质量评估是提升POI数据应用性能的关键。POI数据主要应用于城市(群)、国家及全球等大尺度研究，揭示地理要素空间发展的区域性、综合性和复杂性等特征，如基于城市群尺度分析东北地区城市空间结构特征^[48]或基于全国尺度分析地方小吃空间扩散格局及模式^[49]等。面向大体量数据，如何评估数据质量^[50-51]（构建数字与物质空间映射通道，设置评估指标阈值范围等）、实现多源数据融合^[52-53]（检测关键字符空值率与相似率、匹配拟链接对象等）是提升数据质量与物质空间数字化程度的核心问题^[54-55]，亦是精细刻画陆地表层状态演化和地理学服务于决策的重要支撑^[3-4]。

地理大数据质量评估与提升需要“结合特征，因材施教”。地理大数据实现对“人”“地”的精细化刻画，具有种类多且属性各异的特点。兼顾数据特征构建标准数据库与指标体系是面向多源异构地理大数据质量评估与提升的关键一步^[28, 40, 56]，多源地理大数据融合是挖掘数据“涌现价值”、发挥数据潜力的核心步骤^[50-51]，融合GIS、人工智能和机器学习等跨学科技术构建的自动化管理平台^[22, 38, 57]是实时获取与处理数据、构建数字孪生城市和实现城镇可持续发展的重要技术与装备。地理大数据及计算与可视化技术的发展是实现多学科交叉融通、解决新时代地理学问题趋于复杂化的重要突破口与增长点^[58-60]。

参考文献(References)

[1] Cheng Changxiu, Shi Peijun, Song Changqing, et al. Geographic big-data: A new opportunity for geography complexity study. *Acta Geographica Sinica*, 2018, 73(8): 1397-1406. [程昌秀, 史培军, 宋长青, 等. 地理大数据为地理复杂性研究提供新机遇. *地理学报*, 2018, 73(8): 1397-1406.]

[2] Tsai W P, Feng D P, Pan M, et al. From calibration to parameter learning: Harnessing the scaling effects of big data in geoscientific modeling. *Nature Communications*, 2021, 12: 5988. DOI: 10.1038/s41467-021-26107-z.

[3] Xue Bing, Zhao Bingyu, Li Jingzhong. Urban complexity studies from the perspective of geography: A review based on the literature in the past 20 years. *Progress in Geography*, 2022, 41(1): 157-172. [薛冰, 赵冰玉, 李京忠. 地理学视角下城市复杂性研究综述: 基于近20年文献回顾. *地理科学进展*, 2022, 41(1): 157-172.]

[4] Xue Bing, Li Jingzhong, Xiao Xiao, et al. Overview of man-land relationship research based on POI data: Theory, method and application. *Geography and Geo-Information Science*, 2019, 35(6): 51-60. [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. 地理与地理信息科学, 2019, 35(6): 51-60.]
- [5] Zheng Minrui, Zheng Xinqi, Li Tianle, et al. Big data driven functional interaction patterns and governance strategy for Beijing-Tianjin-Hebei region. *Acta Geographica Sinica*, 2022, 77(6): 1374-1390. [郑敏睿, 郑新奇, 李天乐, 等. 京津冀城市群城市功能互动格局与治理策略. 地理学报, 2022, 77(6): 1374-1390.]
- [6] Xue Bing, Xu Yaotian, Zhao Bingyu. Application and reflection of POI big data from the perspective of geography. *Journal of Guizhou Normal University (Natural Sciences)*, 2022, 40(4): 1-6, 14, 128. [薛冰, 许耀天, 赵冰玉. 地理学视角下POI大数据的应用研究及反思. 贵州师范大学学报(自然科学版), 2022, 40(4): 1-6, 14, 128.]
- [7] Xue Bing, Xiao Xiao, Li Jingzhong, et al. POI-based spatial correlation of the residences and retail industry in Shenyang city. *Scientia Geographica Sinica*, 2019, 39(3): 442-449. [薛冰, 肖骁, 李京忠, 等. 基于POI大数据的沈阳市住宅与零售业空间关联分析. 地理科学, 2019, 39(3): 442-449.]
- [8] Liu Yu, Guo Hao, Li Haifeng, et al. A note on GeoAI from the perspective of geographical laws. *Acta Geodaetica et Cartographica Sinica*, 2022, 51(6): 1062-1069. [刘瑜, 郭浩, 李海峰, 等. 从地理规律到地理空间人工智能. 测绘学报, 2022, 51(6): 1062-1069.]
- [9] Zhen Feng, Li Zherui, Xie Zhimin. Analysis of urban internal spatial structure characteristics and its influencing factors based on population flow: A case study of Nanjing. *Geographical Research*, 2022, 41(6): 1525-1539. [甄峰, 李哲睿, 谢智敏. 基于人口流动的城市内部空间结构特征及其影响因素分析: 以南京市为例. 地理研究, 2022, 41(6): 1525-1539.]
- [10] Wang R X, Wu J P, Qi G Q. Exploring regional sustainable commuting patterns based on dockless bike-sharing data and POI data. *Journal of Transport Geography*, 2022, 102: 103395. DOI: 10.1016/j.jtrangeo.2022.103395.
- [11] Hao Feilong, Shi Xiang, Bai Xue, et al. Geographic detection and multifunctional land use from the perspective of urban diversity: A case study of Changchun. *Geographical Research*, 2019, 38(2): 247-258. [浩飞龙, 施响, 白雪, 等. 多样性视角下的城市复合功能特征及成因探测: 以长春市为例. 地理研究, 2019, 38(2): 247-258.]
- [12] Danshkohan A, Alimoradi M, Ahmadi M, et al. Data quality and data use in primary health care: A case study from Iran. *Informatics in Medicine Unlocked*, 2022, 28: 100855. DOI: 10.1016/j.imu.2022.100855.
- [13] Li Deren, Zhang Guo, Jiang Yonghua, et al. Opportunities and challenges of geo-spatial information science from the perspective of big data. *Big Data Research*, 2022, 8(2): 3-14. [李德仁, 张过, 蒋永华, 等. 论大数据视角下的地球空间信息学的机遇与挑战. 大数据, 2022, 8(2): 3-14.]
- [14] Fu B, Xiao X, Li J Z. Big data-driven measurement of the service capacity of public toilet facilities in China. *Applied Sciences*, 2022, 12(9): 4659. DOI: 10.3390/app12094659.
- [15] Gu R, Qi Y, Wu T Y, et al. SparkDQ: Efficient generic big data quality management on distributed data-parallel computation. *Journal of Parallel and Distributed Computing*, 2021, 156: 132-147.
- [16] Saleem R, Saint R, Clark W, et al. A pragmatic and industry-oriented framework for data quality assessment of environmental footprint tools. *Resources, Environment and Sustainability*, 2021, 3: 100019. DOI: 10.1016/j.resenv.2021.100019.
- [17] Xue Bing, Zhao Bingyu, Xiao Xiao, et al. A POI data-based study on urban functional areas of the resources-based city: A case study of Benxi, Liaoning. *Human Geography*, 2020, 35(4): 81-90. [薛冰, 赵冰玉, 肖骁, 等. 基于POI大数据的资源型城市功能区识别方法与实证: 以辽宁省本溪市为例. 人文地理, 2020, 35(4): 81-90.]
- [18] Wu J H, Hu R M, Li D S, et al. Where have you been: Dual spatiotemporal-aware user mobility modeling for missing check-in POI identification. *Information Processing & Management*, 2022, 59(5): 103030. DOI: 10.1016/j.ipm.2022.103030.
- [19] Zhang J B, Liu X J, Liao W L, et al. Deep-learning generation of POI data with scene images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2022, 188: 201-219.
- [20] Zhang Jia, Wang Chen. Research on rural e-commerce and influential factors of product diversification: Based on the field investigation and analysis of Taobao villages in Zhejiang province. *Progress in Geography*, 2020, 39(8): 1260-1269. [张佳, 王琛. 农村电子商务与产品多样化影响因素探究: 基于浙江淘宝村的实地调研分析. 地理科学进展, 2020, 39(8): 1260-1269.]
- [21] Chen Li, Xu Jingxue, Zhang Wenzhong, et al. Residents' sense of urban public security and community environment: Analysis based on a large-scale questionnaire survey of Beijing. *Acta Geographica Sinica*, 2021, 76(8): 1939-1950. [谌丽, 许婧雪, 张文忠, 等. 居民城市公共安全感知与社区环境: 基于北京大规模调查问卷的分析. 地理学报, 2021, 76(8): 1939-1950.]
- [22] Wang Junsong, Yan Yan. Complexity, relatedness and urban technology evolutionary path: A comparative study between

- Beijing, Shanghai and Shenzhen in China. *Progress in Geography*, 2022, 41(4): 554-566. [王俊松, 颜燕. 复杂度、关联度与城市技术演化路径: 基于北京、上海、深圳的对比分析. *地理科学进展*, 2022, 41(4): 554-566.]
- [23] Li Zhixuan, Zhen Feng, Zhang Shanqi, et al. Characteristics of elderly activity space by public transport and influencing factors: Based on the comparative analysis of daily and occasional activities. *Progress in Geography*, 2022, 41(4): 648-659. [李智轩, 甄峰, 张珊琪, 等. 老年人公交活动空间特征及影响因素研究: 基于日常与偶发活动的对比分析. *地理科学进展*, 2022, 41(4): 648-659.]
- [24] Long Y. Redefining Chinese city system with emerging new data. *Applied Geography*, 2016, 75: 36-48.
- [25] Martinez-Morata I, Bostick B C, Conroy-Ben O, et al. Nationwide geospatial analysis of county racial and ethnic composition and public drinking water arsenic and uranium. *Nature Communications*, 2022, 13: 7461. DOI: 10.1038/s41467-022-35185-6.
- [26] Cienciala A, Sobolewska-Mikulska K, Sobura S. Credibility of the cadastral data on land use and the methodology for their verification and update. *Land Use Policy*, 2021, 102: 105204. DOI: 10.1016/j.landusepol.2020.105204.
- [27] Pearse G D, Tan A Y S, Watt M S, et al. Detecting and mapping tree seedlings in UAV imagery using convolutional neural networks and field-verified data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2020, 168: 156-169.
- [28] Yang J Y, Hu X Y, Feng H Y, et al. Verifying an ENVI-met simulation of the thermal environment of Yanzhong Square Park in Shanghai. *Urban Forestry & Urban Greening*, 2021, 66: 127384. DOI: 10.1016/j.ufug.2021.127384.
- [29] Xue Bing, Xiao Xiao, Li Jingzhong, et al. POI-based analysis on the affecting factors of property prices' spatial distribution in the traditional industrial area. *Human Geography*, 2019, 34(4): 106-114. [薛冰, 肖骁, 李京忠, 等. 基于POI大数据的老工业区房价影响因素空间分异与实证. *人文地理*, 2019, 34(4): 106-114.]
- [30] Xue Bing, Xiao Xiao, Li Jingzhong, et al. POI-based analysis on retail's spatial hot blocks at a city level: A case study of Shenyang, China. *Economic Geography*, 2018, 38(5): 36-43. [薛冰, 肖骁, 李京忠, 等. 基于POI大数据的城市零售业空间热点分析: 以辽宁省沈阳市为例. *经济地理*, 2018, 38(5): 36-43.]
- [31] Liu Yali, Wang Yanfen, Du Jianqing, et al. Big earth data promotes assessment of even development. *Bulletin of Chinese Academy of Sciences*, 2021, 36(8): 963-972. [刘雅莉, 王艳芬, 杜剑卿, 等. 地球大数据助力均衡发展评估. *中国科学院院刊*, 2021, 36(8): 963-972.]
- [32] Yang Jianhao, Song Chao, Zhou Gofu, et al. A government data quality management architecture based on multi-classification sub-chain. *Journal of Information Security Research*, 2022, 8(4): 374-385. [杨建豪, 宋超, 周国富, 等. 一种基于多分类子链的政务数据质量管理架构. *信息安全研究*, 2022, 8(4): 374-385.]
- [33] Zhang L, Sun Z, Zhang J, et al. Modeling hierarchical category transition for next POI recommendation with uncertain check-ins. *Information Sciences*, 2020, 515: 169-190.
- [34] Tang Jie, Ye Xiaoqi. Survey on historic districts of National Historical and Cultural City Research Center in Nanlang Ancient Town, Zhongshan City, Guangdong Province. *City Planning Review*, 2022, 46(7): 90-91. [唐劼, 叶孝奇. 广东省中山市南朗古镇国家历史文化名城研究中心历史街区调研. *城市规划*, 2022, 46(7): 90-91.]
- [35] Tan Haining, Yao Di, Bi Jingping, et al. A next POI recommendation method for data-poor cities. *Chinese High Technology Letters*, 2021, 31(12): 1248-1260. [谭海宁, 姚迪, 毕经平, 等. 面向数据匮乏城市的下一个POI推荐方法. *高技术通讯*, 2021, 31(12): 1248-1260.]
- [36] Zhou Shiyang, Hu Junzhi, Ji Chenghui. POI processing method, device, electronic equipment and computer storage medium. Guangdong: CN110795515B, 2022-04-12. [周世洋, 卢俊之, 季成晖. 兴趣点POI的处理方法、装置、电子设备及计算机存储介质. 广东: CN110795515B, 2022-04-12.]
- [37] Yu Dan. Method, device and storage medium for obtaining quality freshness of map data. Beijing: CN111986552B, 2022-04-15. [俞丹. 地图数据质量鲜度获取方法、装置及存储介质. 北京: CN111986552B, 2022-04-15.]
- [38] Hu Zhihui, Cai Bo, Zhu Yungao. National tax and local tax joint data acquisition system and its operation method. Zhejiang: CN108229921B, 2022-02-18. [胡志怀, 蔡博, 朱贻高. 国税地税联合数据采集系统及其操作方法. 浙江: CN108229921B, 2022-02-18.]
- [39] Xie Kun. Application of basic geographic data in digital city data collection. *Geomatics & Spatial Information Technology*, 2021, 44(4): 162-163, 167. [解琨. 基础地理数据在数字城市数据采集中的应用. *测绘与空间地理信息*, 2021, 44(4): 162-163, 167.]
- [40] Du Jun, Xu Ruifeng, Cao Xiaohang, et al. Feedback method, terminal and server of geographic element information of navigation electronic map. Beijing: CN101608925B, 2013-07-10. [杜钧, 徐瑞峰, 曹晓航, 等. 导航电子地图地理要素信息的反馈方法、终端及服务器. 北京: CN101608925B, 2013-07-10.]
- [41] Jia Zhibin, Feng Chengping, Liu Zhaohui. The invention relates to a data processing method and device. Guangdong:

- CN114329236A, 2022-04-12. [贾志宾, 丰成平, 刘朝辉. 一种数据处理方法及装置. 广东: CN114329236A, 2022-04-12.]
- [42] Lin Zhipeng. Location method, device, device and storage medium based on POI spatial distance. Beijing: CN111726860B, 2022-04-08. [林志鹏. 基于 POI 空间距离的定位方法、装置、设备和存储介质. 北京: CN111726860B, 2022-04-08.]
- [43] Shen Lei, Li Naiqiang. Research on the integration and update technologies of electronic map based on multi-source vector data. *Geospatial Information*, 2021, 19(7): 119-122, 8. [沈蕾, 李乃强. 多源矢量数据的电子地图整合更新技术研究. *地理空间信息*, 2021, 19(7): 119-122, 8.]
- [44] Cheng Peng, Luo Lijun. The invention relates to a POI information supplement method and device. Beijing: CN103218375B, 2016-08-17. [程鹏, 罗丽俊. 一种 POI 信息补充方法及装置. 北京: CN103218375B, 2016-08-17.]
- [45] Zhang Wei, Gao Xinyuan, Li Ruishan. Multi-source POI data fusion based on the spatial location information. *Periodical of Ocean University of China*, 2014, 44(7): 111-116. [张巍, 高新院, 李瑞珊. 空间位置信息的多源 POI 数据融合. *中国海洋大学学报(自然科学版)*, 2014, 44(7): 111-116.]
- [46] Wu Zhangfeng, Xia Lanfang. Study on the method of matching and fusion of multi-source POI. *Bulletin of Surveying and Mapping*, 2018(3): 143-146. [吴张峰, 夏兰芳. 多源异构 POI 融合方法及应用. *测绘通报*, 2018(3): 143-146.]
- [47] Wang Zhiguang. Methods and devices for measuring the quality of map POI data. Beijing: CN105608112A, 2016-05-25. [王智广. 衡量地图 POI 数据的质量的方法和装置. 北京: CN105608112A, 2016-05-25.]
- [48] Xue Bing, Xiao Xiao, Li Jingzhong, et al. Analysis of spatial economic structure of Northeast China cities based on points of interest big data. *Scientia Geographica Sinica*, 2020, 40(5): 691-700. [薛冰, 肖晓, 李京忠, 等. 基于兴趣点 (POI) 大数据的东北城市空间结构分析. *地理科学*, 2020, 40(5): 691-700.]
- [49] Zhu Bangyao, Wu Yuanyuan. Spatial diffusion pattern and mode of local snacks: Based on POI data of four famous local snacks in China. *Scientia Geographica Sinica*, 2021, 41(12): 2179-2185. [朱邦耀, 吴媛媛. 地方小吃空间扩散格局与模式: 基于中国四大知名地方小吃 POI 数据的研究. *地理科学*, 2021, 41(12): 2179-2185.]
- [50] Du Yunyan, Yi Jiawei, Xue Cunjin, et al. Modeling and analysis of geographic events supported by multi-source geographic big data. *Acta Geographica Sinica*, 2021, 76(11): 2853-2866. [杜云艳, 易嘉伟, 薛存金, 等. 多源地理大数据支撑下的地理事件建模与分析. *地理学报*, 2021, 76(11): 2853-2866.]
- [51] Li Pengfei, Zhang Ya, Sun Qinke. Points of interest synthetic similarity calculation method and its application. *Science of Surveying and Mapping*, 2021, 46(9): 178-183. [李鹏飞, 张亚, 孙钦珂. 兴趣点综合相似度计算方法及应用研究. *测绘科学*, 2021, 46(9): 178-183.]
- [52] Li J X, Hong D F, Gao L R, et al. Deep learning in multimodal remote sensing data fusion: A comprehensive review. *International Journal of Applied Earth Observation and Geoinformation*, 2022, 112: 102926. DOI: 10.1016/j.jag.2022.102926.
- [53] Guo Mingjun, Chen Qin, An Xiaomi, et al. Construction and practical application of big data development index in China: From the government data and social data fusion perspective. *Big Data Research*, 2022, 8(2): 182-192. [郭明军, 陈沁, 安小米, 等. 我国大数据发展指数构建及实践应用: 从政务数据与社会数据融合的视角. *大数据*, 2022, 8(2): 182-192.]
- [54] Pei Tao, Huang Qiang, Wang Xi, et al. Big data aggregation: Connotation, classification, and framework. *National Remote Sensing Bulletin*, 2021, 25(11): 2153-2162. [裴韬, 黄强, 王席, 等. 地理大数据聚合的内涵、分类与框架. *遥感学报*, 2021, 25(11): 2153-2162.]
- [55] Li Jianhui, Li Yuepeng, Wang Huajin, et al. Scientific big data management technology and system. *Bulletin of Chinese Academy of Sciences*, 2018, 33(8): 796-803. [黎建辉, 李跃鹏, 王华进, 等. 科学大数据管理技术与系统. *中国科学院院刊*, 2018, 33(8): 796-803.]
- [56] Pei Tao, Liu Yaxi, Guo Sihui, et al. Principle of big data mining. *Acta Geographica Sinica*, 2019, 74(3): 586-598. [裴韬, 刘亚溪, 郭思慧, 等. 地理大数据挖掘的本质. *地理学报*, 2019, 74(3): 586-598.]
- [57] Liu Yu, Yao Xin, Gong Yongxi, et al. Analytical methods and applications of spatial interactions in the era of big data. *Acta Geographica Sinica*, 2020, 75(7): 1523-1538. [刘瑜, 姚欣, 龚咏喜, 等. 大数据时代的空间交互分析方法和应用再论. *地理学报*, 2020, 75(7): 1523-1538.]
- [58] Yang Jun, You Haolin, Zhang Yuqing, et al. Research process on human settlements: From traditional data to big data+. *Progress in Geography*, 2020, 39(1): 166-176. [杨俊, 由浩琳, 张育庆, 等. 从传统数据到大数据+的人居环境研究进展. *地理科学进展*, 2020, 39(1): 166-176.]
- [69] Chen Min, Lv Guonian, Zhou Chenghu, et al. Geographic modeling and simulation systems for geographic research in

the new era: Some thoughts on their development and construction. *Scientia Sinica (Terrae)*, 2021, 51(10): 1664-1680. [陈旻, 闫国年, 周成虎, 等. 面向新时代地理学特征研究的地理建模与模拟系统发展及构建思考. *中国科学: 地球科学*, 2021, 51(10): 1664-1680.]

- [60] Li Xin, Yuan Linwang, Pei Tao, et al. Disciplinary structure and development strategy of information geography in China. *Acta Geographica Sinica*, 2021, 76(9): 2094-2103. [李新, 袁林旺, 裴韬, 等. 信息地理学学科体系与发展战略要点. *地理学报*, 2021, 76(9): 2094-2103.]

Evaluation and enhancement methods of POI data quality in the context of geographic big data

XUE Bing^{1,2}, ZHAO Bingyu^{1,2,3}, LI Jingzhong^{2,4}

(1. Institute of Applied Ecology, CAS, Shenyang 110016, China; 2. Key Lab for Environmental Computation and Sustainability of Liaoning Province, Shenyang 110016, China; 3. Planning Building Environment, Technical University of Berlin, Berlin 10623, Germany; 4. College of Urban Planning and Architecture, Xuchang University, Xuchang 461000, Henan, China)

Abstract: Geographic big data enables a fine-grained depiction of regional human-terrestrial systems and provides new data for the study of human-terrestrial relations and regional development. At present, geographic big data research has entered the stage of widespread application, but the examination of its quality and the corresponding evaluation methods have been lacking to guarantee the widespread and efficient application of the data. POI is an important part of geographic big data and plays an important role in location-based services and an understanding of regional scenarios. This paper proposes a method to assess and enhance POI-type big data, and realize quality evaluation based on site research, GIS and other methods from three dimensions: feature identification completeness, data redundancy rate and spatial location accuracy; discover and summarize possible influencing factors of data quality based on data production process, and prove that multi-source data fusion is an effective means to enhance POI data quality. We found that: the volume of Amap data acquired based on API interface is slightly higher than that of Baidu, the accuracy rate of spatial location is comparable and the redundancy rate is lower; Amap focuses on identifying the entrance of features, which is suitable for analysis such as accessibility; Baidu focuses on discovering non-significant features, which is suitable for analysis such as spatial planning; the discovery, acquisition and processing stages are possible links to reduce data quality, which is influenced by data protection mechanism, and the data quality is inversely proportional to the acquisition volume and area. The quality assessment, enhancement and integration of multi-source heterogeneous geographic data is one of the key ways to enhance the "emergent value" of data, promote trans- and cross-multidisciplinary and solve geographic problems in the new era.

Keywords: POI data; geographic big data; data quality assessment; site research; GIS