武汉大学学报(信息科学版)
Geomatics and Information Science of Wuhan University
ISSN 1671-8860,CN 42-1676/TN

《武汉大学学报(信息科学版)》网络首发论文

题目：志愿者地理信息的点线面数据质量评价及其关联特征挖掘
作者：林安琪，罗文庭，吴浩
DOI：10.13203/j.whugis20230271
收稿日期：2023-07-24
网络首发日期：2023-10-19
引用格式：林安琪，罗文庭，吴浩．志愿者地理信息的点线面数据质量评价及其关联特征挖掘[J/OL]．武汉大学学报(信息科学版).
https://doi.org/10.13203/j.whugis20230271

DOI:10.13203/j.whugis20230271

引用格式：林安琪，罗文庭，吴浩. 志愿者地理信息的点线面数据质量评价及其关联特征挖掘[J].武汉大学学报（信息科学版）,2023,DOI:10.13203/J.whugis20230271

Citation: LIN Anqi，LUO Wenting，WU Hao. Data Quality Assessment and Associated Characteristic Mining of Point Line Polygon Features from Volunteered Geographic Information[J]. Geomatics and Information Science of Wuhan University,2023,DOI:10.13203/J.whugis20230271

志愿者地理信息的点线面数据质量评价及其关联特征挖掘

林安琪[1,2] 罗文庭[1,2] 吴浩[1,2*]

1 华中师范大学城市与环境科学学院，武汉 430079

2 华中师范大学地理过程分析与模拟湖北省重点实验室，武汉 430079

摘要：志愿者地理信息具有数据量大、更新频率高和采集成本低等特点，已经成为专业地理信息数据的有益补充，在众多领域发挥积极作用。然而，志愿者地理信息由非专业人士生产，缺乏严格、统一的数据生产标准和质量控制流程，导致数据质量参差不齐和空间分布不均等问题。为此，针对志愿者地理信息点线面三类要素的几何结构和应用特点，从数据完整性、重复性和精确性等不同维度设计质量评价指标，构建了由评价对象-评价元素-评价指标三层结构组成的数据质量评价框架，并深入挖掘数据质量的空间和语义关联特征。研究结果表明：（1）相比传统评价指标，所提出的指标对数据质量问题的反馈更加灵敏，使评价结果更加有区分度，有效降低了传统指标造成的评价结果不确定性。（2）志愿者地理信息数据质量的空间聚集特征差异性显著，兴趣点等点要素质量的空间聚集性最强，道路和建筑物等线、面要素质量的空间聚集性较弱，并且沿城市交通环线方向上变化明显。（3）兴趣点、道路等点和线要素质量与类别属性的关联性较为显著，而建筑物等面要素质量与其类别没有明显关联。研究结果可为志愿者地理信息的数据质量控制策略提供有益参考。

关键词：志愿者地理信息；点线面要素；空间数据质量；综合评价指标；空间和语义关联特征

中图分类号：P208　　**文献标志码**：A

Data Quality Assessment and Associated Characteristic Mining of Point Line Polygon Features from Volunteered Geographic Information

LIN Anqi[1,2] LUO Wenting[1,2] WU Hao[1,2*]

1 School of Urban and Environmental Sciences, Central China Normal University, Wuhan 430079

2 Hubei Provincial Key Laboratory for Geographical Process Analysis and Simulation, Central China Normal University, Wuhan 430079

Abstract: Objectives: With the characteristics of large amount, high update frequency and low collection cost, Volunteered Geographic Information (VGI) has become the useful supplement to classic geographic information data and plays an important role in many fields. However, due to the lack of strict and unified data production standards and quality control process, the data quality of VGI is uneven and the spatial distribution is not equal. Therefore, this study proposed the quality assessment index system compose of the evaluation object, quality element and quality index for VGI point line polygon features. **Methods:** According to different spatial data structure and application characteristics of point line polygon features, a comprehensive evaluation was conducted from different dimensions such as geometry, topology and semantic quality, and further the spatial and

收稿日期：2023-07-24

基金项目：国家自然科学基金(42201468, 42071358)；全国博士后创新人才计划（BX20220128）；全国博士后科学基金面上资助项目（2022M721283）；华中师范大学优秀青年团队项目（CCNU22QN018）。

第一作者：林安琪，博士，主要从事地理时空大数据挖掘相关研究。anqilin@ccnu.edu.cn

通讯作者：吴浩，博士，教授。haowu@ccnu.edu.cn

semantic characters of data quality were discussed. **Results:** The results show that (1) The new evaluation indexes is more sensitive than the traditional ones, and the evaluation results of each quality element are more differentiated after the index synthesis. (2) The spatial aggregation of POI semantic similarity is the strongest, while the spatial aggregation of road and building quality is weak. (3) Category attributes have significant correlation with POI interest points and road element quality, but have no significant correlation with building quality. **Conclusions：** The comprehensive quality assessment can effectively reduce the result uncertainty caused by using any single index. The spatial aggregation characteristics of VGI point, line and polygon quality are significantly different, and it changes significantly along the direction of urban ring. Category attributes have the potential to be the quality indicator of VGI data.

Key words: Volunteered geographic information; point-line-polygon features; spatial data quality; comprehensive assessment index; spatial and semantic characteristic

大数据时代的到来加快了地图学的变革进程[1-2]。志愿者地理信息（Volunteered Geographic Information，VGI）作为由大众在日常生产生活中主动或无意创建，通过互联网进行地图数据采集、传播与共享，已经成为专业制图的重要数据补充[3-5]。美国 Goodchild 教授于 2007 年对由群众生产、带有地理标签的数据进行了系统阐述，提出了 VGI 概念，并论述了其在山火监测中的应用[6, 7]。这类数据具有数据量大、更新频率高和采集成本低等优势[8]，在车辆导航、公共健康、土地管理和应急响应等领域发挥重要作用[9-12]。然而，VGI 数据通常由非专业人士产生，缺乏严格的生产标准和质量控制，导致数据质量参差不齐、空间分布不均等问题[13, 14]。特别是对基础地理信息采集与制图造成影响，若 VGI 数据几何或属性不准确，不仅会误导数据采样和验证，也会增加制图信息的不确定性[15]。因此，进行 VGI 数据质量的全方位评价和特征挖掘，对提升其应用可靠性具有重要意义。

目前，志愿者地理信息的数据质量评价主要有参考对比法和可信度指标法两类：参考对比法是通过对比志愿者地理信息与权威数据，以两者差异程度量化志愿者地理信息数据的优劣程度；可信度指标法通过分析与数据质量相关的社会经济因素，建立反映数据可信度的间接性指标[16]。相比而言，参考对比法虽然依赖于权威数据，但能够得到直观、量化的质量评价结果，是更为常用的质量评价方法[17]。利用参考对比法评价 VGI 数据质量的研究最早可追溯到 2008 年，Zulfiqar 和 Haklay 团队先后利用了英国军用地图和 Ordnance Survey 数据集，对 OpenStreetMap（OSM）平台的英国路网数据进行了完整性和位置精度评价[18]。随后的研究中提出了拓扑一致性、中位数中心、最小外接多边形等质量评价指标，对法国[19]、德国[20]、伊朗[21]、瑞士[22]和中国[23]等地的 OSM 路网进行了评价。对 VGI 点要素和面要素的质量评价研究相对较少，如文献[24]提出了基于数据密度的 VGI 建筑物完整性评价指标，应用于美国和新西兰的多个城市；文献[25]利用官方数据评价 OSM 兴趣点的主题精度和逻辑一致性。上述研究分别面向特定类型的 VGI 数据，利用单一评价指标进行数据质量评价和分析，针对点线面全要素类型的质量研究尚未见报道[26]。同时，对 VGI 数据质量的结果分析更多关注与社会经济水平的相关性[27]，较少有研究从数据本身着手，探究数据质量的空间或语义特征，从而为数据质量纠正提供更为直接的参考依据。

为此，本研究针对志愿者地理信息中点线面三类要素的几何结构和应用特点，从数据完整性、重复性和精确性等不同维度设计质量评价指标，构建了由评价对象-评价元素-评价指标三层结构组成的数据质量评价框架，并深入挖掘数据质量的空间和语义关联特征。该研究可满足不同应用场景下志愿者地理信息数据的适用性评价需求，并为科学制定质量控制策略提供重要参考。

1 研究方法

1.1 志愿者地理信息的数据质量评价指标体系

志愿者地理信息的数据质量评价体系由评价对象、评价元素和评价指标三层结构组成（图 1）。首先，从点、线、面要素的不同应用需求出发，设计了差异化的评价元素：点要素几何结构简单而属性信息丰富，主要根据位置和属性判断数据重复度；线要素中以道路数据应用需求最广，多用于导航、路径规划等，对几何完整性、位置精度和逻辑一致性的要求较高；对于面要素，如建筑物数据等，除了数据完整性和位置精度，轮廓的几何精确度也是考察重点。在评价指标层，根据点线面要素的数据结构特征，针对每个评价元素设计了相应的多个评价指标（具体公式见表 1），以增加评价结果可靠性。然后，利用 CRITIC-TOPSIS 指标综合方法，获得志愿者地理信息数据的综合质量评价结果，并对数据质量的空间特征和语义特征进行深入挖掘，为保障志愿者地理信息数据质量提供策略与建议。

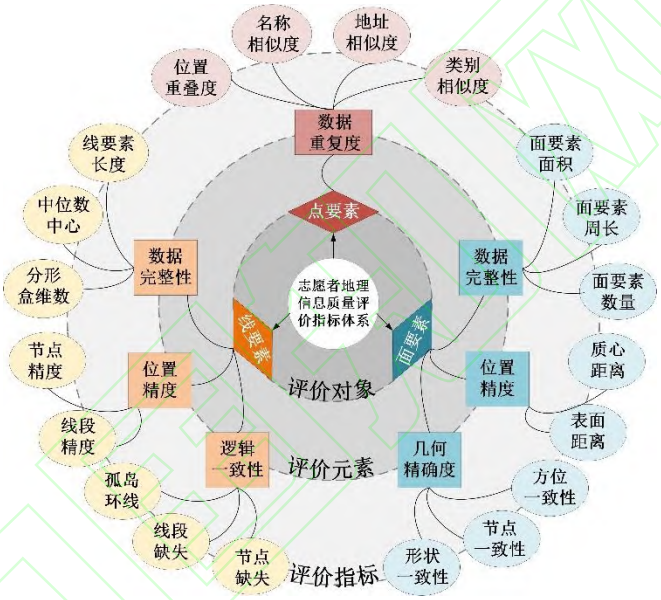

图 1 志愿者地理信息的点线面数据质量评价体系

Fig. 1 Data quality assessment system of VGI point line polygon features

1.2 点要素质量评价指标设计

点要素的属性完整度高，其中，99.99%的属性包含名称，98.89%的要素包含地址。因此，点要素主要考察数据重复度，评价指标包括位置重叠度、名称相似度、地址相似度和类型相似度。位置重叠度根据坐标空间关系判断，即当两个 POI 点数据的坐标信息完全一致，在空间上重合时，则判定为位置重叠。类型相似度可以直接对类型字段的字符串进行一致性判断。名称和地址相似度的评价更为复杂，以 VGI 兴趣点（Point of Interest，POI）为例[28]，由于缺乏标准的命名规则，大量地理实体具有多个相似名称或地址的 POI，如“湖北省群众艺术馆社会艺术考级指定考点”和“湖北省音乐家协会社会艺术考级点”，需要判断文本的模糊相似程度。为此，采用基于 TF-IDF 改进的 Simhash 算法评价同一位置下点要素的名称和地址相似度，计算流程见图 2。将本文分词处理后，通过分词向量的 hash 值与词语的 TF-IDF 权重相乘，把各分词的序列值累加形成序列串，降维后得到文本 Simhash 签名，根据海明距离判断名称或地址是否类似。

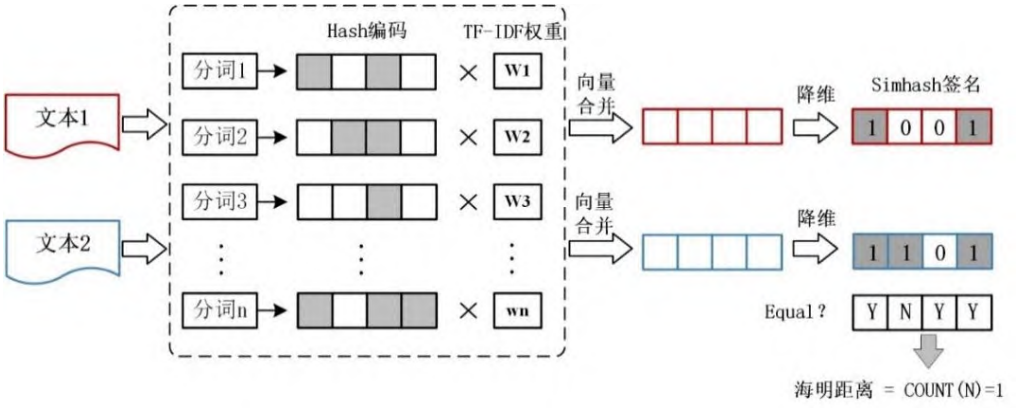

图 2 基于 TF-IDF Simhash 算法的点要素名称相似性计算原理图

Fig. 2 Schematic diagram of point name similarity based on TF-IDF Simhash algorithm

1.3 线要素质量评价指标设计

线要素主要考察数据完整性、逻辑一致性和位置精度。对于完整性，使用了经典的长度指标和中位数中心指标，其中，中位数中心测算以 1 km 格网为基本单元，分别计算参考道路与 OSM 道路在每个格网单元内的中位数中心，并通过计算这两个中位数中心之间的距离得到格网单元的中位数中心指标值；同时，引用分形盒维数方法建立了新的评价指标，度量线状要素分布的空间填充状态[29]。逻辑一致性主要检验了网络拓扑错误，包括孤岛路线、路段缺失和节点缺失。位置精度包括节点精度和线精度两个指标：节点精度是对比 VGI 与参考数据对应节点的位置误差，具体为，分别提取各参考数据和 OSM 路网的节点，采用近邻分析方法，对各 OSM 路网的节点与参考路网节点进行匹配，进而计算各匹配对节点间的距离，通过统计各格网内 OSM 节点与其匹配对距离的平均值即得到节点精度指标结果；对于线精度，通过改进经典的单缓冲区算法[30]，提出了基于双缓冲的线要素位置精度评价指标 *LA*（具体计算方法见表 1）。该方法通过将线要素转化为面要素，放大了曲线的细微几何特征，既体现了要素之间的距离，也量化了形状差异。如图 3 所示，线要素 1 到 4 分别代表了不同程度的位置精度，其中线要素 1 与参考数据的缓冲区交集最小，说明其位置精度最低。分别计算四个线要素场景的单缓冲区指标和双缓冲区指标，发现单缓冲区指标结果稳定在 0.91~0.96，说明其难以反映四个线要素位置精度的差异；双缓冲区指标结果在 0.67～0.92 之间，更为灵敏。

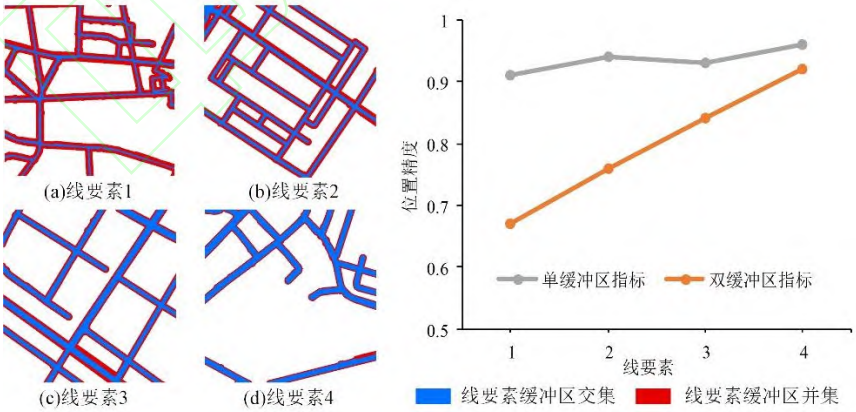

图 3 双缓冲区和单缓冲区指标对线要素位置精度的评价差异对比

Fig. 3 The comparison of location accuracy for line feature by single and double buffer method

1.4 面要素质量评价指标设计

面要素主要考察数据完整性、位置精度和几何精确度。完整性主要通过对比 VGI 面要素与参考数据的面积、周长和数量差异。位置精度从全局和局部角度考察，设计质心距离和表面距离两个指

标。几何精确度考察了要素轮廓的形状、方位和节点指标，其中形状指标是利用 Boyce-Clark 半径指数（BCI）测度 VGI 与参考数据的差异，该算法以面要素质心为起点生成一系列角度相等的射线，将射线与面要素边界相交得到一系列半径，计算 BCI 指数（具体计算公式见表 1）。将 BCI 指标与面积、周长和圆形度 3 个传统指标用于对比四个面要素的差异（图 4），结果显示，4 个面要素的 BCI 指数差异最大，而其他 3 个指数差异非常小，无法反映面要素的形状区别。这说明 BCI 对形状差异的检测能力更强，更适用于考察面要素轮廓的几何精确度。

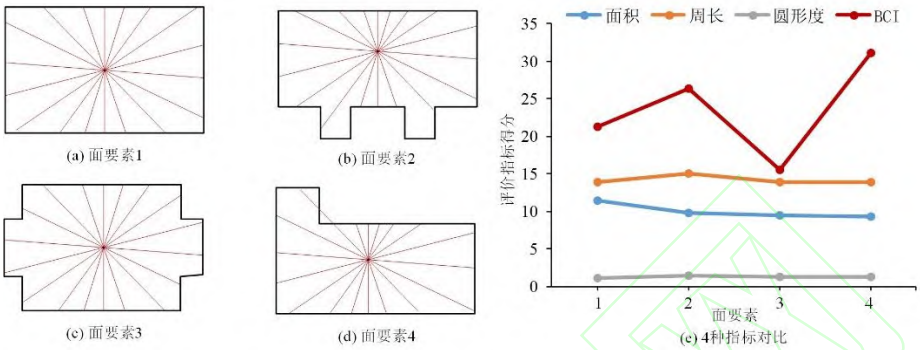

图 4 BCI 半径形状指数与其他指标的效果对比

Fig. 4 The effect comparison between BCI Radius shape index and other indexes

表 1 志愿者地理信息的点线面数据质量评价指标计算公式

Tab. 1 Equations of quality assessment indexes for VGI point line polygon features

要素类型	评价指标	计算公式	指标描述
点要素	位置重叠度	$\sum I(dis(P_i,P_j)=0)$	统计位置重合的 POI 数量，$I()$为指示函数，当任 POI 点 i 和 j 的位置重合则为 $I()$ =1，否则为 0
	类型重复度	$\sum I(str(T_i)=str(T_j))$	当 POI 点位置重合，统计其中类型一致的 POI 数量
	名称相似度	$\mathrm{tf}(t_j,d_k)*\log\frac{\lvert D\rvert}{\lvert\{k:t_j\in d_k\}\rvert+1}$	$\mathrm{tf}(t_j,d_k)$为特征词t_j在文本d_k中出现的频率，$\log\frac{\lvert D\rvert}{\lvert\{k:t_j\in d_k\}\rvert+1}$表示特征词在语料库中的逆频率，|D|为文本总数量，$\lvert\{k:t_j\in d_k\}\rvert$表示含有特征词的文本数
	地址相似度	$\mathrm{tf}(t_j,d_k)*\log\frac{\lvert D\rvert}{\lvert\{k:t_j\in d_k\}\rvert+1}$	该指标计算原理与名称相似度指标一致
线要素	线要素长度	$\lvert\sum L_{CGI}-\sum L_{REF}\rvert$	$\sum L_{CGI}$ 和$\sum L_{REF}$分别代表 VGI 和参考线状要素的总长度。
	中位数中心	$dis(N_{CGI\_i},N_{REF\_i})$	N_{CGI_i},N_{REF_i}分别代表 VGI 和参考路网的中位数中心。
	分形盒维数	$\lvert D_{CGI}-D_{REF}\rvert$，$D=\lim_{s\to 0}-\frac{\log N(s)}{\log s}$	D_{CGI}和D_{REF}分别为 VGI 和参考数据的盒维数，S为网格尺寸，N为覆盖格子数量
	节点精度	$\frac{1}{n}\sum_{i=1}^{n} dis(N_{CGI\_i},N_{REF\_i})$	$dis(N_{CGI\_i},N_{REF\_i})$ 为 VGI 数据与参考数据中第 i 对匹配节点的距离
	线段精度	$1-\frac{S(CGI\cap REF)}{S(CGI\cup REF)}$	$S(CGI\cap REF)$是 VGI 与参考数据缓冲区的交集面积，$S(CGI\cup REF)$则代表了并集面积
	孤岛环线	$\sum R_{FI}$	路网中与其他任何道路没有连接的路段
	线段缺失	$\sum R_{AJ}$	路网中应该相连但却没有相连的路段
	节点缺失	$\sum R_{IWJ}$	路网中相交路段所缺失的节点
面	面要素面积	$\sum A$	建筑物总面积A

要素	面要素周长	$\sum P$	建筑物总周长P
	面要素数量	$\sum N$	建筑物总数量N
	质心距离	$\frac{1}{n}\sum_{i=1}^{n} dis(C_{CGI\_i}, C_{REF\_i})$	$dis(C_{CGI_i}, C_{REF_i})$为 VGI 与参考数据质心的距离，$n$ 为总节点数
	表面距离	$(A_{union} - A_{inter})/A_{union}$	A_{union}代表 VGI 和参考建筑物的并集面积，A_{inter}代表二者的交集面积
	方位一致性	$\lvert\sum D_{CGI} - \sum D_{REF}\rvert$	VGI 与参考建筑物的外接矩形方向差值
	节点一致性	$\lvert\sum N_{CGI} - \sum N_{REF}\rvert$	VGI 与参考建筑物的节点数差值
	形状一致性	$SS = \lvert D_{CGI} - D_{REF}\rvert$， $BCI=\sum_{i=1}^{n}\left\lvert\frac{r_i}{\sum_{i=1}^{n} r_i}\times 100\% - \frac{100}{n}\right\rvert$	通过 BCI 指数计算，n 为等角的辐射线数量，r_i为质心到边界的半径

2 志愿者地理信息数据质量的关联特征挖掘方法

2.1 基于 TOPSIS-CTRTIC 的志愿者地理信息数据质量综合得分计算方法

单一评价指标难以准确反映志愿者地理信息的质量水平，有必要综合考虑不同指标的影响，提升评价结果的稳定性，从而保障数据质量关联特征挖掘的可靠性。为此，引用 TOPSIS-CTRTIC 方法对单一评价指标进行综合处理：首先，利用 CRITIC 赋权方法计算各指标权重ω_j，与标准化的质量评价得分矩阵相乘；然后，采用欧氏距离函数计算待评价单元与正负理想解之间的差异 D_i^+和D_i^-，并计算样本与理想解之间的相对逼近度T_i：

$$T_i = \frac{D_i^-}{D_i^+ + D_i^-} \tag{1}$$

式中，T_i的取值范围为$[0, 1]$，T_i越接近于 0，则数据质量越好。通过该方法，形成对志愿者地理信息点线面要素多个质量元素的综合评价结果。

2.2 志愿者地理信息数据质量的空间关联特征挖掘方法

志愿者地理信息数据质量存在一定的空间异质性，不同质量元素表现出的空间特征值得进一步分析。为此，利用全局莫兰指数和局部莫兰指数，考察 VGI 数据质量的空间聚集特征，挖掘数据质量的空间关联性。首先，采用全局莫兰指数判断 VGI 数据质量是否存在空间自相关性，若数据质量具有全局聚集特征，进一步采用局部莫兰指数，探明哪些区位出现了数据质量的高值或低值聚集区，以及聚集区的范围大小。局部莫兰指数I_i的计算公式为：

$$I_i = \frac{Z_i}{S^2}\sum_{j\neq i}^{n} w_{ij} Z_j \tag{2}$$

式中，$Z_i = y_i - \bar{y}$，$Z_j = y_j - \bar{y}$，$S^2 = \frac{1}{n}\sum(y_i - \bar{y})^2$，$w_{ij}$为空间权重值，$n$ 为研究单元的总数，y_i和y_j分别代表了第 i 和第 j 个研究单位的数据质量评价得分。

2.3 志愿者地理信息数据质量的语义关联特征挖掘方法

数据质量不仅与地理位置相关，与其社会属性也存在一定关联性。经济水平、受教育程度、贡献者信誉度等外部因素对 VGI 数据质量的影响已经得到证实[31-33]，有必要进一步探索数据质量的内部语义特征。类别属性是 VGI 数据最重要的语义信息之一，各类型要素都有对应的分类体系，如高德 POI 数据分为公共管理、商业服务、居住等 8 大类，OSM 道路可分为高速公路、一级公路、二级公路等 8 类，OSM 建筑物可与城市土地利用类型相对应[34]，分为居住、商业、工业、交通和公共管理 5 类。本研究以 1km 格网为单元，分别统计点要素数量、线要素长度和面要素面积，以及对应格网的点线面质量得分，计算数据质量和类别属性的相关系数R：

$$R = \frac{\sum_{i=1}^{n}(a_i - \bar{a})(b_i - \bar{b})}{\sqrt{\sum_{i=1}^{n}(a_i - \bar{a})^2 (b_i - \bar{b})^2}} \tag{3}$$

式中，a_i和b_i分别代表第 i 个格网单元的质量评价得分和所统计的点线面数据量，$\bar{a}$和$\bar{b}$代表二者对应的平均值，$R \in [-1,1]$， 正负和绝对值分别代表相关方向与程度。

3 案例分析与讨论

3.1 案例数据

以武汉市 2020 年 POI 数据、OSM 道路和 OSM 建筑物数据为例（图 5），进行针对点线面要素的志愿者地理信息数据质量评价。POI 数据通过编写 python 网络爬虫程序调用高德地图 API 获取，OSM 道路和建筑物数据通过 OSM 官方平台下载，共收集到 POI 947 419 个，OSM 道路 29 979 km 和建筑物 18.98 km^2。同时，收集了四维图新的道路数据和天地图建筑物轮廓数据作为高质量的参考数据，通过将其与 VGI 数据进行多方位对比，评价数据质量。完成数据收集后，对所有原始数据进行了格式转换和投影变化等预处理。

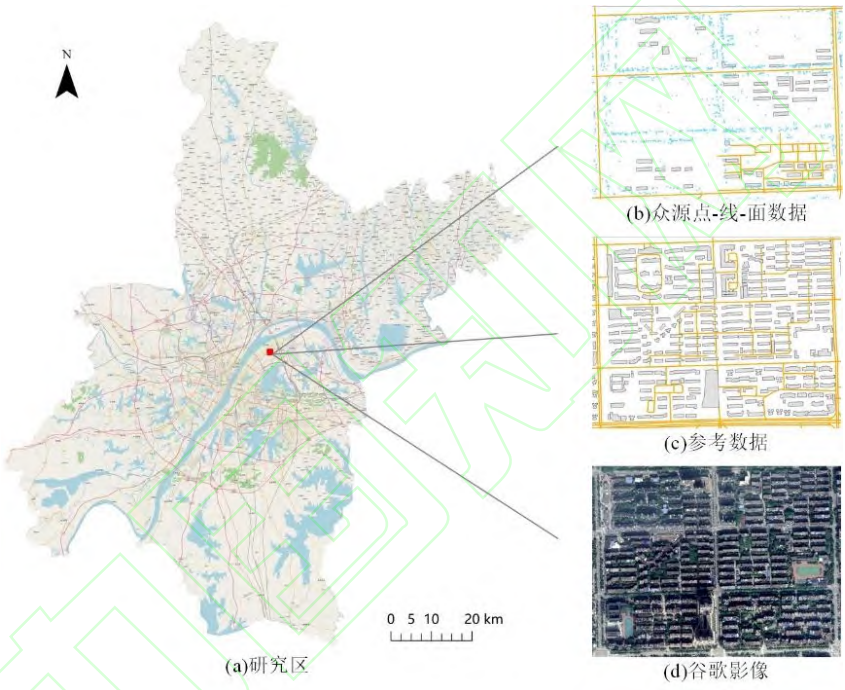

图 5 研究区 VGI 数据与参考数据

Fig. 5 VGI data and reference data in the Study Area

3.2 志愿者地理信息数据质量的单一指标评价结果

以 1 km×1 km 格网为评价单元，去除数据缺失的格网后，分别有 4 547、7 462 和 645 个格网参与 VGI 点、线、面的质量评价。按照所设计的评价指标，计算所有格网 VGI 点、线、面要素的 20 种评价指标结果。图 6 为质量评价结果的箱线图，发现属于同一质量元素（同色系）的指标评价结果数值分布趋势总体类似，说明该类指标结果趋势相同，不同质量元素之间差异较大，体现了 VGI 数据质量的多面性。POI 数据重复度、道路逻辑一致性和建筑物几何精度的评价指标得分整体偏低，平均值接近于 0，由于所有指标均为负向指 标，说明这三方面质量普遍较好；道路和建筑物的位置精度评价结果平均值在 0.5 以上，而且指标的数值分布范围大，说明道路和建筑物的位置精度问题严重。同时发现，道路和建筑物的完整性指标箱型图分布比较类似，说明同一数据源（OSM）的不同类型要素，完整度和位置误差水平保持一致。

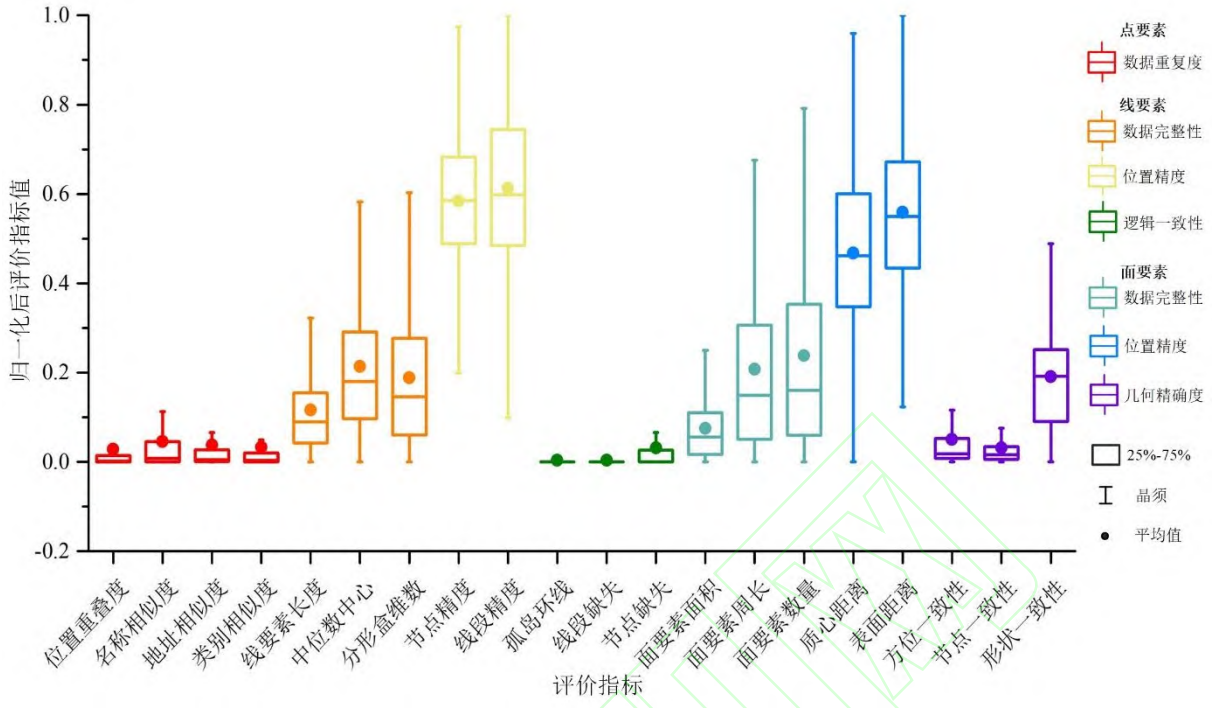

图 6 武汉市志愿者地理信息数据质量的单一指标评价结果箱线图

Fig. 6 Boxplot of single quality index results for VGI data in Wuhan

如图 7 所示，所有单一指标的质量评价结果按照优（前 25%）、良（25%～50%）、中（50%～75%）、差（75%～100%）4 个等级划分，并与对应格网关联。总体来看，点、线、面数据质量存在明显的空间异质特征，POI 数据 4 个指标（图 7（a）-7（d））的高值区主要集中在城市中心，说明中心地区的 POI 数据冗余度高于周边区域。道路数据的完整性和位置精度指标（图 7（e）-7（i））空间分布的随机性更强，质量问题严重；对比来看，道路逻辑一致性的 3 个指标（图 7（j）-7（l））质量较好，特别是远城区地带。建筑物数据由于数量较少，整体覆盖全市 7.15%，并未观察到明显的数据质量空间分布特征。

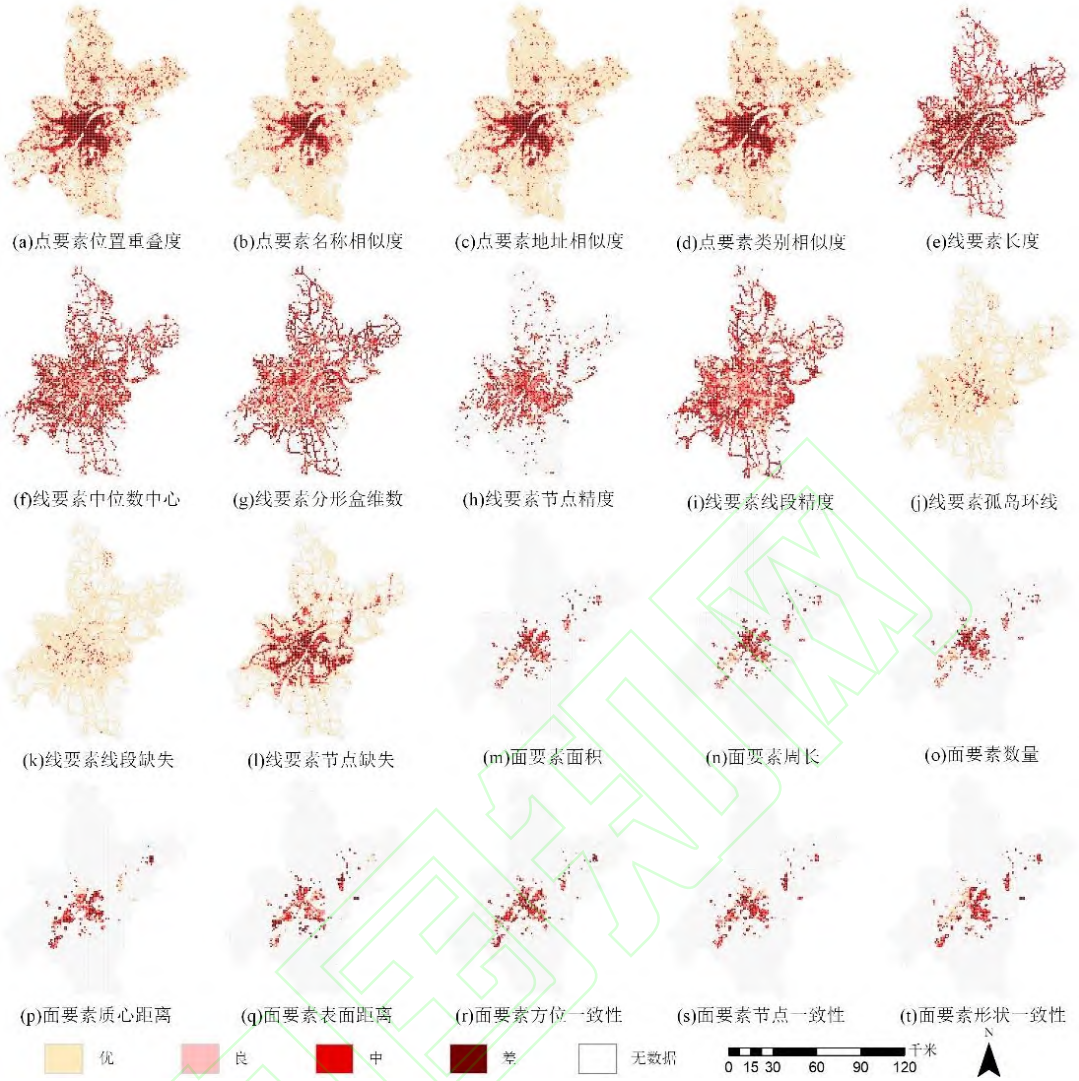

图 7 武汉市志愿者地理信息数据质量的单一指标评价结果的空间分布图

Fig. 7 Spatial distribution of single quality index results for VGI data in Wuhan

3.3 志愿者地理信息数据质量的综合评价结果

图 8 为武汉市 VGI 数据质量的综合质量评价结果，包括 POI 数据相似度、道路完整性、道路位置精度、道路逻辑一致性，以及建筑物完整性、位置精度和几何精确度。通过指标综合，VGI 数据质量的空间分布随机性减弱，空间特征更为明显，如图 7（h）和（i）所示线要素位置精度的两个评价指标，构成了一个综合指标（图 8（c）），更容易观察到位置精度的聚集趋势。这一现象反映出质量元素的综合评价结果对比单一指标更加具有区分度，有效降低了评价结果的不确定性。具体来看，POI 数据重复度的高值最集中，超过 68.09%的差值（深蓝色）聚集在城市三环线以内；对于道路质量，数据完整性的空间分布随机性较强，位置精度在城市中心区域的优秀率更高(49.54%)，逻辑一致性呈现从城市中心向四周逐渐变好的趋势；建筑物完整性和位置精度的空间分布趋势表现出一定的互补性，城市中心的建筑物虽然完整性不高，但是位置精度较高，而左下角建筑物完整性高的区域位置精度较差。

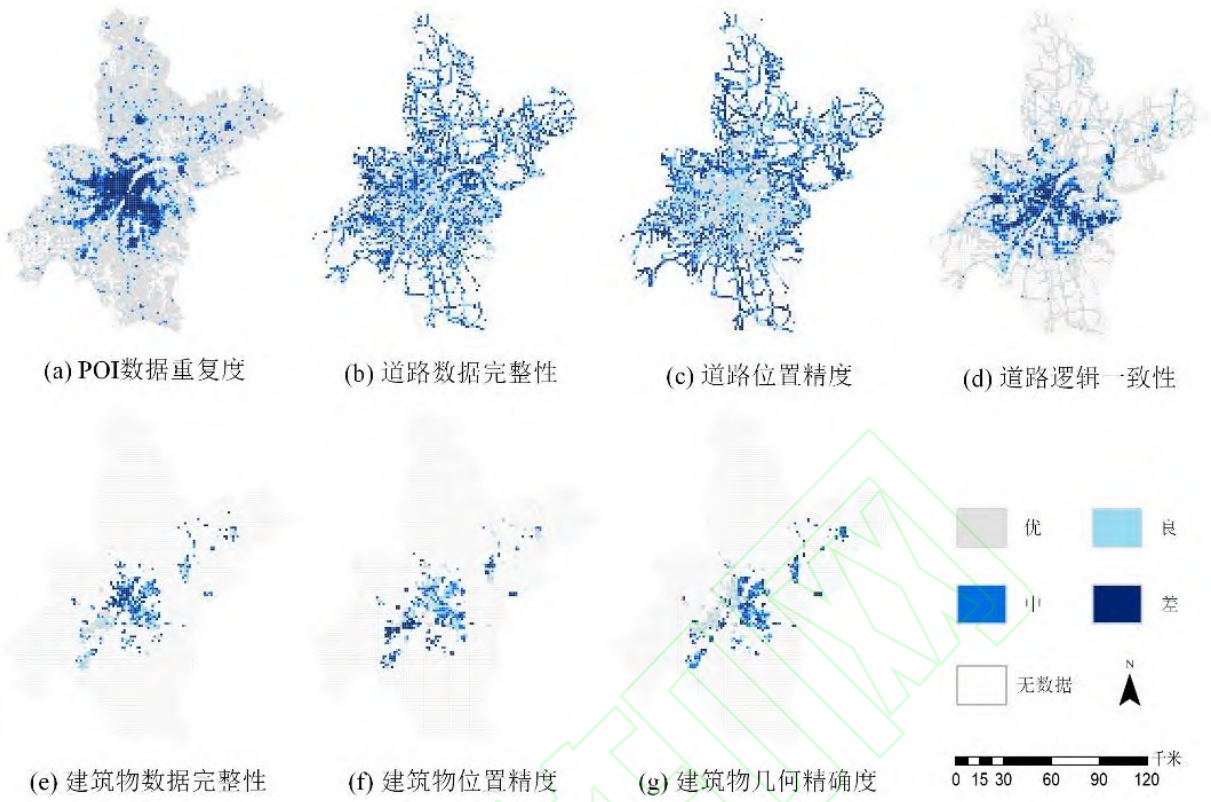

图 8 武汉市志愿者地理信息数据质量的综合质量评价空间分布图

Fig. 8 Spatial distribution of comprehensive quality index for VGI data in Wuhan

3.4 志愿者地理信息数据质量的空间和语义关联特征分析结果

3.4.1 空间关联特征分析结果

从全局空间特征来看，POI 数据重复度的莫兰指数为 0.589，空间自相关性最强；OSM 路网逻辑一致性、位置精度的莫兰指数分别为 0.376 和 0.329，表明空间自相关性较弱，而数据完整性随机分布，无空间聚集特征；建筑物完整性、位置精度和几何精确度的莫兰指数分别为 0.451、0.461 和 0.451，空间聚集程度类似。

图 9 展示了综合指标的局部空间聚集特征，进一步明确数据质量的空间聚集位置和范围特点。POI 数据重复度的高值聚集区斑块较大，且完全集中在三环内，可能是城市核心区域 POI 数据增长更快，产生冗余的风险更高。对于道路数据，道路完整性的聚集斑块最小、最分散，而位置精度和逻辑一致性的空间聚类特征相反：逻辑一致性的高值聚类分布在五环以内，而且聚类斑块较大，低值聚类主要分布在五环外，且聚类斑块较小，可能是由于中心城区更新频率更快，在数据融合中更可能发生拓扑错误；位置精度的低值聚类集中在五环线以内，说明城市中心道路位置精度优于郊区。对于 OSM 建筑物，完整性较差（红色）的数据聚集在三环以内，而位置精度较差的聚集在三环以外，几何精度呈现左低右高的趋势。

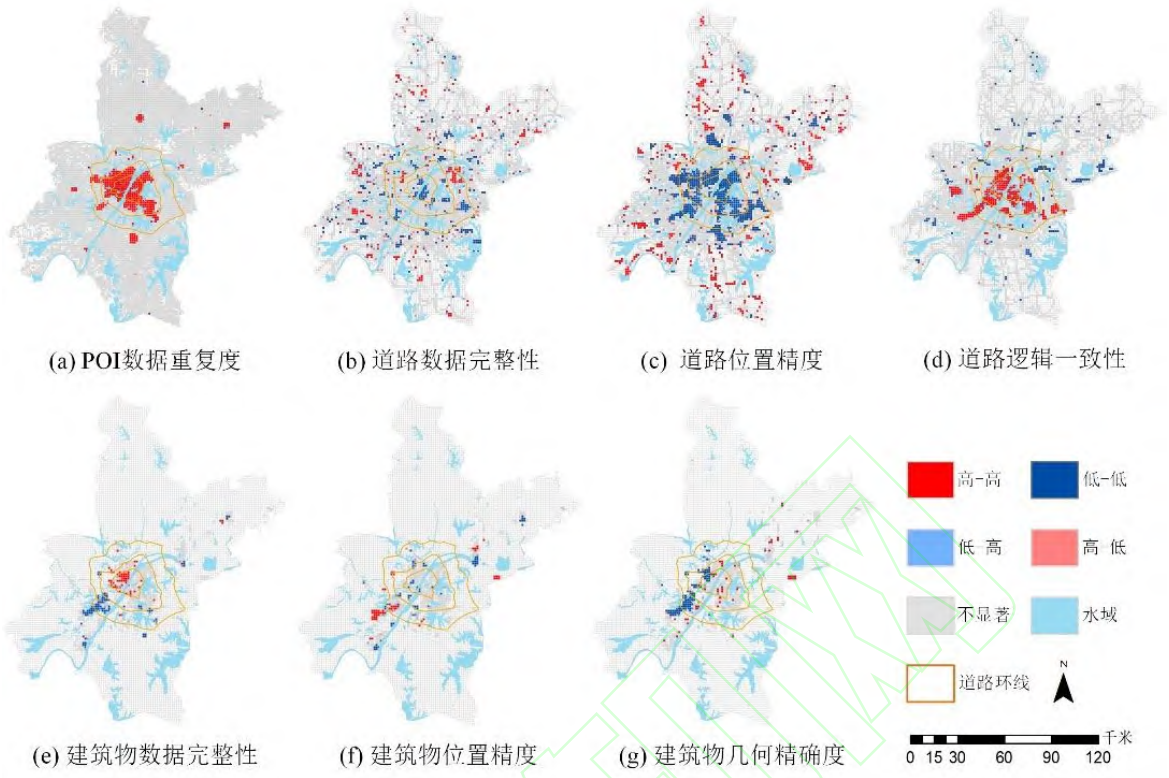

图 9 武汉市志愿者地理信息数据质量的空间关联特征

Fig. 9 Spatial associated characteristic of VGI data quality in Wuhan

从一环内到五环外共 6 个区域，以各环区域为横轴，各环内数据质量为高值/低值聚集的格网与环内格网总数的比值为纵轴，绘制直方图可清晰地展示空间关联特征，结果如图 10 所示。POI 数据重复度、道路逻辑一致性和建筑物几何精确度 3 个质量元素，在沿内环向外环的方向上，高值聚集性逐渐减弱；相反，道路数据完整性和位置精度在沿内环向外环的方向上，低值聚集性逐渐减弱。

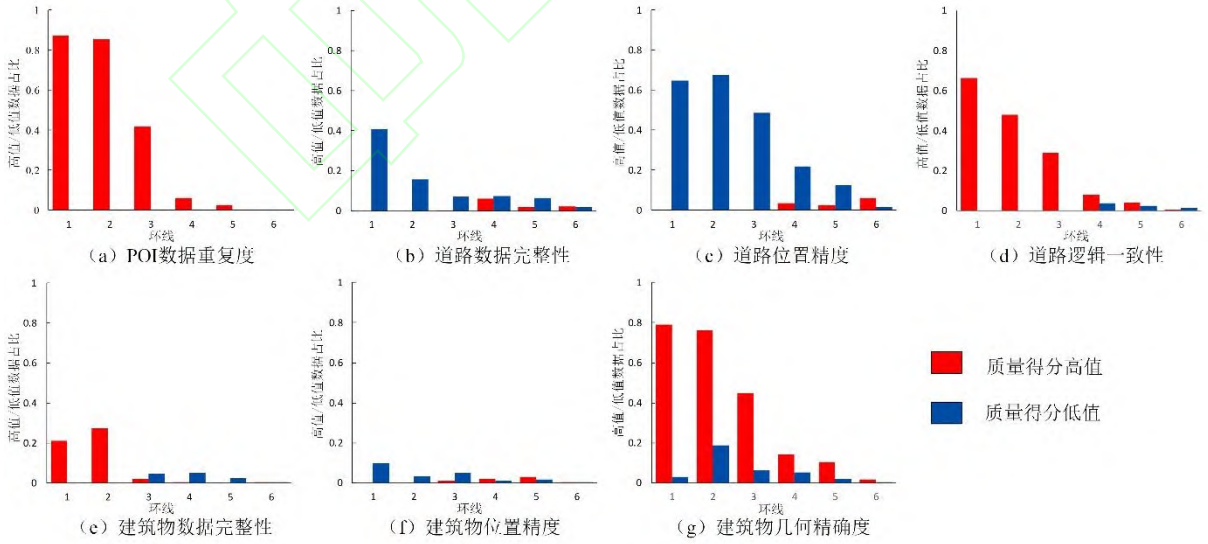

图 10 武汉市志愿者地理信息数据质量沿城市交通环线的变化趋势

Fig. 10 Change trend along the urban loop of VGI data quality in Wuhan

3.4.2 语义关联特征分析结果

对 VGI 数据质量的语义特征分析，主要验证了数据质量和类别属性的相关性，结果见图 11。在点线面要素中，点要素数据质量与类别属性的相关性最强，有 6 种类型都与数据质量呈现显著正向相关，其中商业服务、公共管理、道路设施和居住类的相关系数大于 0.5，说明这 4 类 POI 越多的区域，POI 数据重复度越高。道路的 3 个质量评价元素与类别属性的相关性差异较大：逻辑一致性与道路类型正相关，且相关性随着道路等级降低而减弱，说明等级越高的道路，产生拓扑错误可能性越大；道路完整性与道路类型呈负相关，高速公路与数据完整性的负相关性较强，可能与高速公路等级高且在城市区域占比小有关；位置精度和住宅区道路的负相关系数最高（0.39），说明这类道路位置精度较高，可能是因为低等级道路较窄，与权威数据重合的概率更高。建筑物质量与用地类型的相关性较弱，说明建筑物数据质量的类别差异性并不明显。

通过语义关联特征分析，可以发现 VGI 不同类型数据质量问题的替代性指标，从而对数据质量评估与纠正提供有针对性的意见。结果显示，道路的等级与位置精度、逻辑一致性等要素都具有显著相关性。例如，道路等级越高，产生拓扑错误可能性越大。在 VGI 道路数据应用中，可以根据道路的等级属性，判断是否需要进行拓扑纠正、精度纠正等。具体来说，对于高等级的主干道更应该重视拓扑纠正，对于低等级的辅路等更应该重视其位置精度。

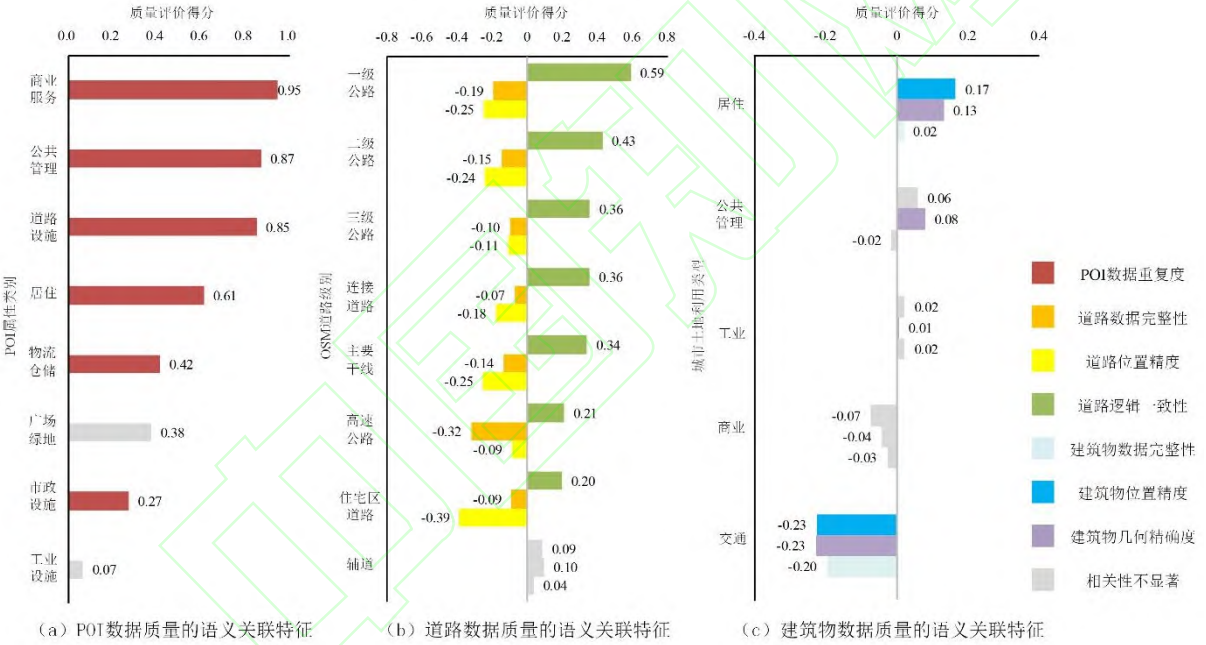

图 11 武汉市志愿者地理信息数据质量的语义关联特征

Fig. 11 Semantic associated characteristic of VGI data quality in Wuhan

4 总结与展望

本研究针对点线面三类要素，构建了由评价对象-评价元素-评价指标三层结构组成的志愿者地理信息数据质量评价指标体系，以武汉市 POI 兴趣点、OSM 道路和 OSM 建筑物为例，多维度评价志愿者地理信息数据质量，并对数据质量的空间关联特征和语义关联特征进行了深入挖掘，为科学制定质量控制策略提供了依据。主要结论如下：

1）新型质量评价指标的引入改善了 VGI 数据质量评价的可靠性。一方面，新型指标对数据质量问题的反馈更加灵敏；另一方面，与传统指标形成优势互补，使综合评价结果区分度增强，并有效降低了质量评价的不确定性。

2）对于 VGI 点线面三种几何要素的质量，它们的空间关联特征存在显著的差异性。从全局看，点状要素质量的空间聚集特征高于线要素和面要素质量；从局部看，POI 数据重复度、道路逻辑一致性和建筑物几何精确度，从内环向外环延伸方向上数据质量趋优，而道路数据完整性和位置精度的趋势恰巧相反。

3）点、线两种几何要素质量的语义关联特征显著，而面要素缺乏语义关联特征。POI 点要素重复度与其类别具有强烈正相关关系；线要素的完整性、位置精度与类别呈现负相关关系，而逻辑一致性与类型的正相关关系随着道路等级降低而减弱。

所提出的 VGI 数据质量评价指标体系主要包括经典指标和改进的新指标组成，主要利用与参考数据对比的方式，利用 ArcMap 和 Python 语言进行批量化处理，研究方法具有很好的可行性。结合数据统计与空间分析方法，研究结果体现了新型质量评价指标的优势，以及点、线、面三种类型数据质量的空间关联特征和语义关联特征，结果分析具有科学性。本研究的应用潜力主要体现在 VGI 数据质量控制领域：一方面，根据点、线、面要素质量评价元素的空间关联性特征，有助于快速确定特定类型数据质量问题集中的区域，从而有针对性的进行数据纠正与问题排查，提高质量控制效率。另一方面，根据所发现的数据质量语义关联特征，可以在缺乏参考数据而无法直接进行质量评价的情况下，根据数据本身的属性来判断 VGI 数据的可靠性。

本文所提出的数据质量评价方法主要针对点、线、面三类数据，研究局限体现在未将更多形式的志愿者地理数据纳入评价范畴。随着志愿者地理信息的发展，涌现出更丰富的数据形式，如社交媒体数据、Flickr 等平台上带有地理标签的图片数据、大众点评等生活服务平台的位置打卡数据等。在后续研究中，有必要改进现有评价体系，适应不同数据类型的质量评价需求。

参考文献

[1] Wang Jiayao. Cartography: From Digital to Intelligent[J]. *Geomatics and Information Science of Wuhan University*, 2022,47(12):1963-1977. [王家耀.地图科学技术:由数字化到智能化[J].武汉大学学报(信息科学版),2022,47(12):1963-1977.]

[2] Li Deren, Zhang Hongyun, Jin Wenjie. The Mission of Geo-spatial Information Science in New Infrastructure Era[J]. *Geomatics and Information Science of Wuhan University*, 2022,47(10): 1515-1522.[李德仁,张洪云,金文杰.新基建时代地球空间信息学的使命[J].武汉大学学报(信息科学版), 2022,47(10): 1515-1522.]

[3] Yan Y W, Feng C C, Huang W, et al. Volunteered geographic information research in the first decade: a narrative review of selected journal articles in GIScience[J]. *International Journal of Geographical Information Science*, 2020, 34 (9):1765–1791.

[4] Heipke C. Crowdsourcing geospatial data[J]. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2010, 65(6): 550-557.

[5] Shan Jie, Qin Kun, Huang Changqing, et al. Methods of Crowd Sourcing Geographic Data Processing and Analysis[J]. *Geomatics and Information Science of Wuhan University*, 2014,39(4):390-396. [单杰,秦昆,黄长青,等.众源地理数据处理与分析方法探讨[J].武汉大学学报(信息科学版),2014,39(4):390-396.]

[6] Goodchild M F. Citizens as sensors: the world of volunteered geography[J]. *GeoJournal*, 2007, 69(4): 211-221.

[7] Goodchild M F, Glennon J A. Crowdsourcing geographic information for disaster response: a research frontier[J]. *International Journal of Digital Earth*, 2010, 3(3): 231-241.

[8] Du Yunyan, Yi Jiawei, Xue Cunjin, et al. Modeling and analysis of geographic events supported by multi-source geographic big data[J]. *Acta Geographica Sinice*, 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[J].地理学报,2021,76(11):2853-2866.]

[9] Yang Wei, Ai Tinghua. Extracting Arterial Road Polygon from OpenStreetMap Data Based on Delaunay Triangulation[J]. *Geomatics and Information Science of Wuhan University*, 2018,43(11):1725-1731.[杨伟,艾廷华.运用 Delaunay 三角网提取 OpenStreetMap 主干道多边形[J].武汉大学学报(信息科学版),2018,43(11):1725-1731.]

[10] Li Hanqi, Jia Peng, Fei Teng. Geographical association between dietary tastes and chronic diseases in China:An exploratory study using crowdsourcing data mining techniques[J]. *Acta Geographica Sinice*, 2019, 74(8): 1637-1649. [李瀚祺,贾鹏,费腾.基于众源数据挖掘的中国饮食口味与慢性病的空间关联[J].地理学报,2019,74(8):1637-1649.]

[11] Zhou De, Zhong Wenyu, Zhou Ting, Qi Jialing. Assessment on Urban Mixed Land Use and Analysis of Its Influencing Factors Based on POI Data: A Case of the Main Districts of Hangzhou City. *China Land Science*[J], 2021, 35(8):96-106. [周德, 钟文钰, 周婷等. 基于 POI 数据的城市土地混合利用评价及影响因素分析——以杭州市主城区为例[J]. 中国土地科学, 2021, 35(8):96-106.]

[12] Yi Jiawei, Wang Nan, Qian Jiale, et al. Spatio-temporal responses of urban road traffic and human activities in an extreme rainfall event using big data. *Acta Geographica Sinice*[J], 2020, 75(3): 497-508. [易嘉伟, 王楠, 千家乐, 等. 基于大数据的极端暴雨事件下城市道路交通及人群活动时空响应. 地理学报[J], 2020, 75(3): 497-508.]

[13] Ma Chao, Sun Qun, Xu Qing, et al. Accuracy evaluation and improvement of volunteer geographic information based on image matching[J]. *Bulletin of Surveying and Mapping*. 2017(3):22-25. [马超,孙群,徐青,等.基于影像匹配的自发地理信息道路精度评价与改善[J].测绘通报,2017(3):22-25.]

[14] Zhu Jianjun, Song Yingchun, Hu Jun, et al. Challenges and Development of Data Processing Theory in the Era of Surveying and Mapping Big Data[J]. *Geomatics and Information Science of Wuhan University*, 2021,46(7): 1025-1031.[朱建军,宋迎春,胡俊,等.测绘大数据时代数据处理理论面临的挑战与发展[J].武汉大学学报(信息科学版),2021,46(7):1025-1031.]

[15] Foody, G., See, L., Fritz, S., et al. Accurate Attribute Mapping from Volunteered Geographic Information: Issues of Volunteer Quantity and Quality[J]. *Cartographic Journal*, 2015,52: 336-344.

[16] Zhao Yijiang, Zhou Xiaoguang. Version similarity-Based model for volunteers' reputation of volunteered geographic information: a case study of polygon[J]. *Acta Geodaetica et Cartographica Sinica*, 2015,44(5):578-584. [赵肄江,周晓光.地理信息志愿者信誉度评估的版本相似度模型——以面目标为例[J].测绘学报,2015,44(5):578-584.]

[17] Zhu Fuxiao, Wang Yanhui. On the comprehensive evaluation of the data quality for OSM road network from the perspectives of multi-level and multi-granularity[J]. *Journal of Geo-information Science*, 2017,19(11):1422-1432. [朱富晓,王艳慧.多层次多粒度 OSM 路网目标数据质量综合评估方法研究[J].地球信息科学学报,2017,19(11):1422-1432.]

[18] Haklay M. How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets[J]. *Environment and planning B: Planning and design*, 2010, 37(4): 682-703.

[19] Girres J F, Touya G. Quality assessment of the French OpenStreetMap dataset[J]. *Transactions in GIS*, 2010, 14(4): 435-459.

[20] Dorn H, Törnros T, Zipf A. Quality evaluation of VGI using authoritative data—A comparison with land use data in Southern Germany[J]. *ISPRS International Journal of Geo-Information*, 2015, 4(3): 1657-1671.

[21] Forghani M, Delavar M R. A quality study of the OpenStreetMap dataset for Tehran[J]. *ISPRS International Journal of Geo-Information*, 2014, 3(2): 750-763.

[22] Fan H, Yang B, Zipf A, et al. A polygon-based approach for matching OpenStreetMap road networks with regional transit authority data[J]. *International Journal of Geographical Information Science*, 2016, 30(4): 748-764.

[23] Wang Ming, Li Qingquan, Hu Qingwu, et al. Quality analysis on crowd sourcing geographic data with open street map data[J]. *Geomatics and Information Science of Wuhan University*, 2013,38(12):1490-1494. [王明,李清泉,胡庆武,等.面向众源开放街道地图空间数据的质量评价方法[J].武汉大学学报(信息科学版),2013,38(12):1490-1494.]

[24] Zhou Q. Exploring the relationship between density and completeness of urban building data in OpenStreetMap for quality estimation. *International Journal of Geographical Information Science*, 2018, 32 (2): 257–281.

[25] Balducci F. Is OpenStreetMap a good source of information for cultural statistics ? The case of Italian museums[J].*Environment and Planning B:Urban Analytics and City Science*，2021，48(3): 503-520.

[26] Fan Hongchao，Kong Gefei，Yang Anran. Current status and prospects of research for voluntered geographic information[J]. *Acta Geodaetica et Cartographica Sinica* .2022, 51(7);1653-1668. [范红超,孔格菲,杨岸然.众源地理信息研究现状与展望[J].测绘学报.2022,51(7):1653-1668.]

[27] Antoniou V. and Skopeliti A. Measures and indicators of vgi quality: an overview[C]//*ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, La Grande France, 2015: 345–351.

[28] Yang Min, Ai Tinghua , Lu Wei, et al. A Real-time Generalization and Mult-scale Visualization Method for POI Data in Volunteered Geographic Information[J]. *Acta Geodaetica et Cartographica Sinica*, 2015.44(2):228-234. [杨敏,艾廷华,卢威,等.自发地理信息兴趣点数据在线综合与多尺度可视化方法[J].测绘学报,2015,44(2):228-234.]

[29] Wu H, Lin A Q, Clarke K C, et al. A comprehensive quality assessment framework for linear features from Volunteered Geographic Information[J]. *International Journal of Geographical Information Science*, 2021, 35(9): 1826-1847.

[30] Goodchild M F, Hunter G J. A simple positional accuracy measure for linear features[J]. *International journal of geographical information science*, 1997, 11(3): 299-306.

[31] Huang Mengni, Zhou Xiaoguang, Zhao Yijiang. Cleaning Model of OSM Data which Considering the Trustworthiness[J]. *Geomatics &Spatial Information Technology*, 2017,40(1):177-181. [黄梦妮, 周晓光, 赵肄江. 顾及可信度的OpenStreetMap数据清理[J]. 测绘与空间地理信息, 2017, 40(1):177-181.]

[32] Ma Chao, Sun Qun, Xu Qing, Wen Bowe. The research status and tendency of the volunteer geographic information data quality[J]. *Science of Surveying and Mapping*, 2017,42(3):93-97. [马超,孙群,徐青,温伯威.志愿者地理信息数据质量研究现状与趋势[J].测绘科学,2017,42(3):93-97.]

[33] Camboim S, Bravo J, and Sluter C, 2015. An investigation into the completeness of, and the updates to, OpenStreetMap data in a heterogeneous area in Brazil[J]. *ISPRS International Journal of Geo-Information*, 4 (3), 1366–1388.

[34] 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): 182-187.