杨歌, 汪维, 陶涛, 等. 城市排水管网 GIS 数据异常的检测与修复[J]. 净水技术, 2022, 41(6): 134-140.
YANG G, WANG W, TAO T, et al. Detection and recovery of abnormal GIS data in urban drainage pipelines network system[J]. Water Purification Technology, 2022, 41(6): 134-140.

扫我试试?

城市排水管网 GIS 数据异常的检测与修复

杨 歌,汪 维,陶 涛*,信昆仑,李树平,颜合想
(同济大学环境科学与工程学院,上海 200092)

摘 要 数据是城市排水管网地理信息系统(GIS)的核心和基础。只有保证高质量的数据,才能使 GIS 系统真正发挥在排水管网信息化管理、建模分析等方面的作用。文章按照"拓扑检测-文本属性项归类-管径检测-高程检测"的处理步骤提出了排水管网数据异常检测与修复的方法,并将方法应用到 ZH 市的排水管网 GIS 数据中,结果表明该方法能有效识别拓扑、文本属性、管径和高程数据的异常,并为部分异常数据提供合理的修正参考值。

关键词 排水管网 地理信息系统(GIS) 数据质量 异常数据检测 数据修复
中图分类号: TU992.2 **文献标识码**: A **文章编号**: 1009-0177(2022)06-0134-07
DOI: 10.15890/j.cnki.jsjs.2022.06.018

Detection and Recovery of Abnormal GIS Data in Urban Drainage Pipelines Network System

YANG Ge, WANG Wei, TAO Tao*, XIN Kunlun, LI Shuping, YAN Hexiang
(*College of Environmental Science and Engineering, Tongji University, Shanghai 200092, China*)

Abstract Data is the core and base of an urban drainage pipelines network geographic information system(GIS) system. Only data of high quality can guarantee effectiveness of GIS when being used for information-based management, modeling and analysis of drainage pipelines network. A set of strategies for detection and recovery of abnormal GIS data in urban drainage pipelines network system was provided, following the steps of "topology detection, text attributes classification, pipe diameters detection and elevations detection". The above method was applied in drainage pipelines network in ZH City. The research indicated that this method could well identify errors of topology, text attributes, pipe diameters and elevations, and provideed reasonable reference values for some of abnormal data.

Keywords drainage pipelines network geographic information system(GIS) data quality abnormal data detection data recovery

防治城市内涝、保障城市水环境健康、维持城市正常运行等都离不开城市排水系统良好的运营管理[1]。将地理信息系统(GIS)技术与给排水专业知识相结合,建立城市排水管网 GIS 系统,可以极大地提高管网基础信息的更新维护效率,加快相关部门对管网事故的反应速度并丰富其应对措施。排水管网 GIS 数据库建立和运行的各个阶段均可能产生数据异常,如测量、绘图或施工过程中的误差导致采集到的数据本身存在异常;人工录入、编辑处理和数据修复操作引入的数据异常;使用过程中用户的理解不当或错误使用带来的数据异常[2]。数据是贯穿于 GIS 系统各开发阶段与组成部分的重要要素,低质量的排水管网数据可能会对排水管网的建模、运行分析及管理方案的决策等多个方面造成不良影响[3]。因此,如何精准、高效地维护和管理城市排水管网 GIS 数据库,保证数据信息的完整性、准确性与系统性,是一项重要的研究课题。

国外对 GIS 用于排水系统的研究范围很广,如将 GIS 应用于计算排水管道的故障概率[4]、估算污

[收稿日期] 2021-06-21
[基金项目] 国家自然科学基金:城市雨水系统光滑粒子动力学模拟理论研究(51778452),城市排水管网运行状态微机电技术诊断理论与应用(51978493)
[作者简介] 杨歌(1998—),女,硕士,研究方向为给水排水工程设计运行最优化,E-mail: loglady997@126.com。
[通信作者] 陶涛,教授,E-mail: taotao@tongji.edu.cn。

水处理厂服务人口[5]、计算城市污废水的热量[6]等,但目前尚未有针对排水管网 GIS 数据异常检测与修复的研究。国内学者则在排水管网 GIS 数据质量方面进行了一定的有益探索,如秦立为[3]梳理了排水管网 GIS 系统的常见数据异常形式和原因等,据此提出了相应的数据质量评价指标及方法。孔彦虎[7]提出了对拓扑异常、属性缺失、图层合并等问题的检测和修复方法,但仅针对单个案例提出该方法,未进行系统性研究。王腾龙[8]结合案例分析了管网中的拓扑异常,在空间数据库中编写规则来检测和修复数据,对人工判断的依赖性较强。以上研究方法一般仅能识别和修正管网拓扑错误和部分属性错误,而针对文本属性、管径和高程等重要数值属性的研究甚少,尚缺乏高效可靠的检测与修复策略。本文提出了系统性解决排水管网各类主要数据异常问题的方法,具有一定的普适性,并采用人机结合的处理方式大大降低了人工成本,为排水管网 GIS 数据的智能化维护与管理提供了有效的技术支撑。

1 拓扑检测与修复

排水管网的拓扑异常包括排水节点与排水管渠的缺失和重复、管段端点或节点的位置不准确、排水管渠方向错误、管网始端节点与终端节点标注错误等。针对管段和节点的重复、孤立、缺失或位置错误等问题,可借助 ArcGIS 软件的“拓扑”工具箱,为排水管网中的点、线、面要素建立合理的空间规则,并在错误检查器中进行相应的拓扑修复[9]。而针对部分特殊的拓扑问题,可进一步采用结构化查询语言(structured query language, SQL)在 SQL Server 或其他软件中编写查询与修正的规则来进行检测与修复:判断排放口或出水口,即查询某个管段的下游节点不为任意管段的上游节点;判断反向管段,即判断某个管段的起点只为其他一根或多根管段的起点,不是任意管段的终点,且它的终点只为其他一根或多根管段的终点,不是任意管段的起点。

2 文本属性项归类与修复

排水管网数据库的文本属性项主要包括排水体制、管渠材料、管理归属部门 3 类。属性项的表述不应出现存在空值、同种类型表述方式不同、表述方式存在包含关系等错误。本研究基于 Jaccard 相似系数(Jaccard similarity coefficient)矩阵进行排水管网文本属性项的归类与修复。Jaccard 相似系数的数学定义是集合 A 与集合 B 的交、并集大小的比值,Jaccard 相似系数越大,说明两文本间的相似度越高,从而实现文本分类[10]。基于 Jaccard 相似系数矩阵的排水管网属性项半自动分类方法包括如下 5 个步骤。

(1) 剔除某一属性项中的重复值,形成一组新的数据 $TA = \{da_1, da_2, da_3, \dots, da_N\}$,计算 da_i 在原始数据 TA^{raw} 中出现的概率 p_i,按照文本出现的概率从大到小排序,得到两组相互对应的数据 $\{p_1, p_2, p_3, \dots, p_N\}$ 和 $\{da_1, da_2, da_3, \dots, da_N\}$。

(2) 根据 Jaccard 相似系数的相反数计算 TA 不同文本之间的 1-gram 距离 $J_{i,j}$,如式(1)。

$$J_{i,j} = 1 - \frac{2 \times \bar{L}}{L_i + L_j} \tag{1}$$

其中:L_i——文本 da_i 的字符个数;

L_j——文本 da_j 的字符个数;

$\bar{L}$——文本 da_i 与 da_j 中相同的字符个数。

(3) $J_{i,j}$ 形成距离矩阵 J,如图 1 所示,$J_{i,j}$ 表示 da_i 跟 da_j 的 1-gram 距离。为避免重复计算,只需计算距离矩阵 J 的上三角数据,距离矩阵 J 的对角线数值全部为 0。

$$J = \begin{array}{c} \\ da_1 \\ da_2 \\ \cdots \\ \cdots \\ da_{N-1} \\ da_N \end{array} \begin{array}{c} \begin{array}{cccccc} da_1 & da_2 & \cdots & \cdots & da_{N-1} & da_N \end{array} \\ \begin{bmatrix} 0 & J_{1,2} & J_{1,3} & \cdots & J_{1,N-1} & J_{1,N} \\ 0 & 0 & J_{2,3} & \cdots & \cdots & J_{2,N} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & J_{N-3,N-1} & J_{N-3,N} \\ 0 & \cdots & \cdots & \cdots & 0 & J_{N-1,N} \\ 0 & 0 & \cdots & \cdots & 0 & 0 \end{bmatrix} \end{array}$$

图 1 距离矩阵

Fig. 1 Distance Matrix

(4) 根据距离矩阵 J,从 da_1 遍历到 da_N,将 1-gram 距离小于某一阈值的两文本划分为同一类别,形成分类矩阵 C。对分类矩阵 C 进行人工调整形成 C',以确保分类准确。如图 2 所示,处于矩阵 C' 同一行的文本为同一类别,排在前面的文本出现的频率较大,分类矩阵的行数等于类别数目。

(5) 根据调整后的分类矩阵 C',当原始属性数据 $da_t^{raw} = C'_{i,j}$,且 $j \neq 1$ 时,将 da_t^{raw} 修改为 $C'_{i,1}$。属性项的修正可以利用 ArcGIS 软件自带的功能通过编写规则高效实现。

文本属性项问题的处理流程如图 3 所示,异常

铸铁管	铸铁	Z	NULL	[]
钢筋混凝土管	钢筋混凝土	钢混	钢筋砼管	钢筋砼
混凝土管	混凝土	混	砼	[]
PE 管	PE	PEE	聚乙烯管	聚乙烯
HDPE 管	高密度聚乙烯	[]	[]	[]

注:NULL 为原数据中的空值;[ ]为矩阵长度不够的占位空值

图 2　人工调整后的分类矩阵

Fig. 2　Classification Matrix after Manual Adjustment

的属性值将被修改为其可能正确的表述中与其距离“最近”的修正值,而距离“最近”通过比较 1-gram 距离确定。被归为同一类的所有属性值,将被修改为“最常见”的正确属性值,而“最常见”通过统计各属性项在数据集中出现的次数确定。这种方法具有无需挑选训练集对其进行标记、人工操作简单、维护成本低等优点。

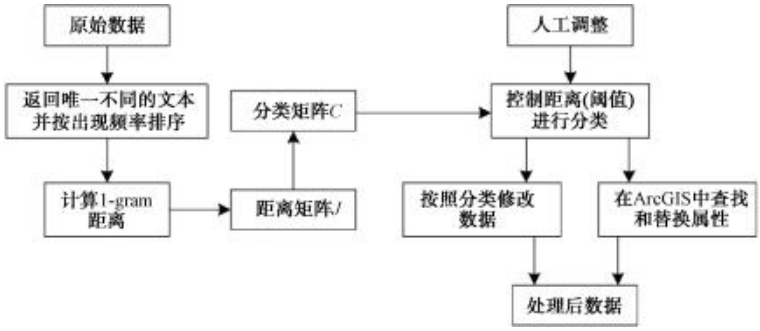

图 3　属性项半自动分类方法流程图

Fig. 3　Flow Chart of Semi-Automatic Classification Method of Attribute Items

3　管径检测与修复

本文提出基于规则的异常管径修复方法,排水管网的管径问题包括错误值和缺失值,处理流程如图 4 所示。将明显不在取值范围内的错误管径值修改为 0,针对管径值为 0 的管段进行逐个判断填充,根据用户自定义的一系列参数取值,结合上下游管段的实际存在情况填充缺失管径。在程序编写中,函数如式(2)。

$$[H,\ \text{ifup},\ \text{tick}] = \text{GetH}(t,\ \text{H0set},\ \text{FL\_Q},\ \text{ifupprior},\ \text{ifdel},\ \text{Nmax}) \quad (2)$$

其中:t——临时变量,表示第 t 根缺失管径的管段;

H0set——缺失管径的管段集合;

FL_Q——管段数据;

ifupprior——确定非管网始端管段管径按上游或下游管段管径填充的参数;

ifdel——确定管网始端管段管径按下游管段管径填充或直接删除的参数;

Nmax——单根孤立管段管径的填充值;

H——管径填充值;

ifup——记录返回管径值是上游还是下游管径的参数;

tick——记录管段的上游或下游管段的总数。

具体输入参数说明如表 1 所示,输出参数说明如表 2 所示。

表 1　输入参数说明

Tab. 1　Description of Input Arguments

参数	参数值	参数设置说明
参数 1	ifdel=0	管网始端管段管径填充为-1 表示删除
	ifdel=1	不考虑 ifupprior,管网始端管段管径按下游管段管径填充
参数 2	ifupprior=0	非管网始端管段管径按上游管段管径填充
	ifupprior=1	非管网始端管段管径按下游管段管径填充
标记值	Nmax	单根孤立管段管径按 Nmax 填充,一般取较大的数

表 2　输出参数说明

Tab. 2　Description of Output Arguments

序号	H	ifup	tick	描述
1	上游最大管径	0	>0	上游有管段,上游管段数为 tick
2	下游最小管径	1	>0	下游有管段,下游管段数为 tick
3	Nmax	-1	=0	上下游都没有管段,孤立管段
4	0	-1	>0	上下游管径都是空值
5	-1	-1	>0	标记优先被删除的管网始端空管段

若返回的管径值为 0,说明此时管网中至少存

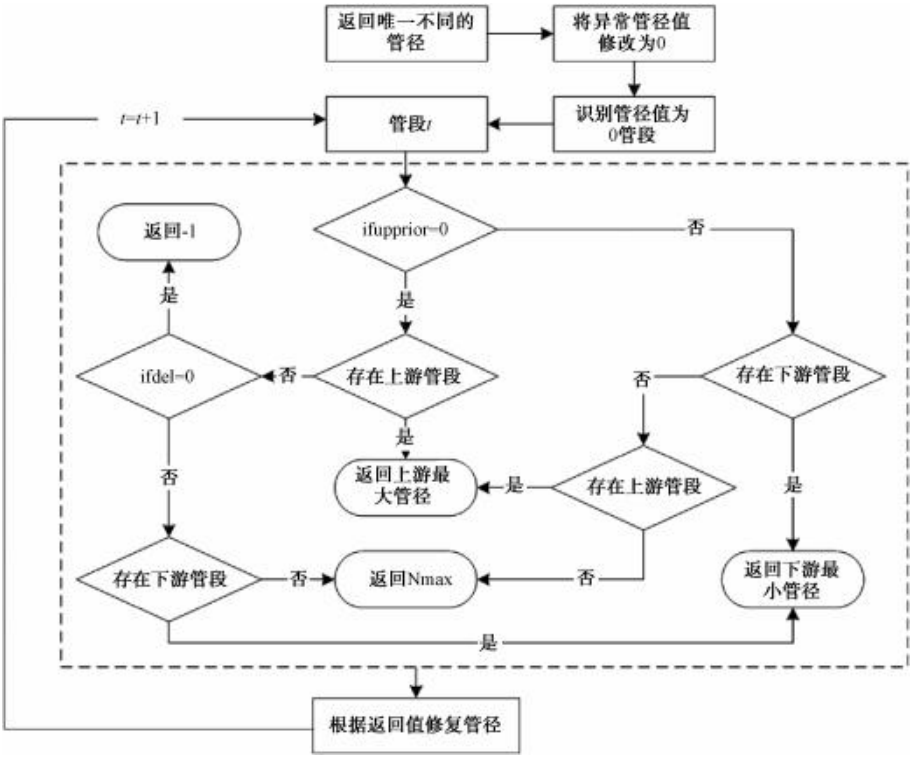

图 4 管径问题的处理流程图

Fig. 4 Flow Chart of Solution to Pipeline Diameter Problem

在两条连续的缺失管径管道,这种情况可以通过两次或多次填充解决。但考虑到多次填充可能带来较大误差,重复填充的次数不宜过多,关于管段孤立或连续管径缺失的情况还需结合外部信息才能更好地进行填充。

4 高程检测与修复

在一个管段分支中,排水管道的高程一般遵循从上游到下游逐渐下降且无突变的规律,基于这种规律可以识别出异常的高程值。首先对管段高程进行预处理,将明显异常的高程值以“埋深 0.7 m”为标准进行修正,以降低后续不合理检测或修复情况的发生概率。然后整理排水管网 GIS 数据,构造节点索引的管段分支集合。每个分支用一个数组来表示,按从下游往上游追溯排列,分支 branch[i]=[分支终端节点编号, 管段 1 编号, 管段 2 编号, …, 管段 $n-1$ 编号, 管段 n 编号, 分支始端节点编号],分支集合 topology = {branch[1], branch[2], …, branch[$m-1$], branch[m]}。构造这个管段分支集合的函数流程如图 5 和图 6 所示。

根据构造出的全部管段的分支集合,基于最小

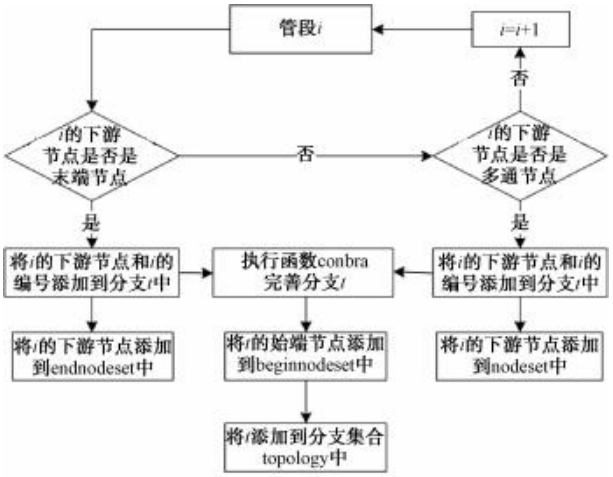

图 5 构造管段分支集合的主函数流程图

Fig. 5 Flow Chart of Main Function of Topology Structure

二乘法进行分支管段高程异常值的检测。对每个分支上各根管段的下游(或上游)高程做直线拟合(分支管段数为两根)或二次曲线拟合(分支管段数≥3根),并做出曲线的 50%置信区间,超出置信区间的管段上游和下游高程被认为可能是错误的,需要进

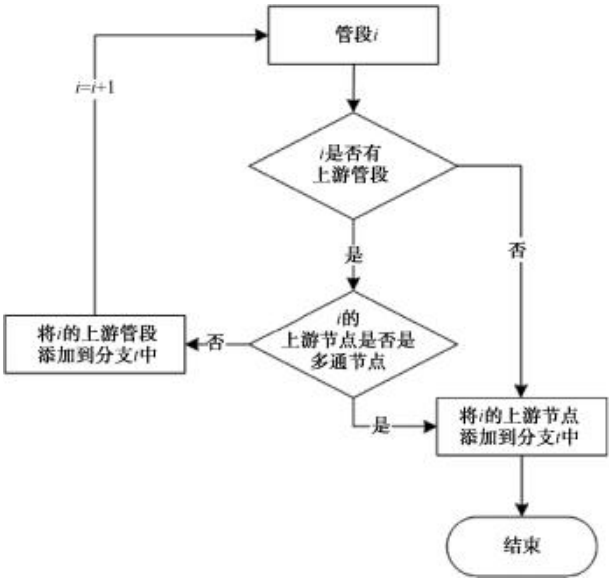

图 6　构造管段分支的迭代函数流程图

Fig. 6　Flow Chart of Iterated Function of Structural Branch Pipelines

行标记并结合外部信息具体判断。分支管段高程处理方法实现的流程如图 7 所示。

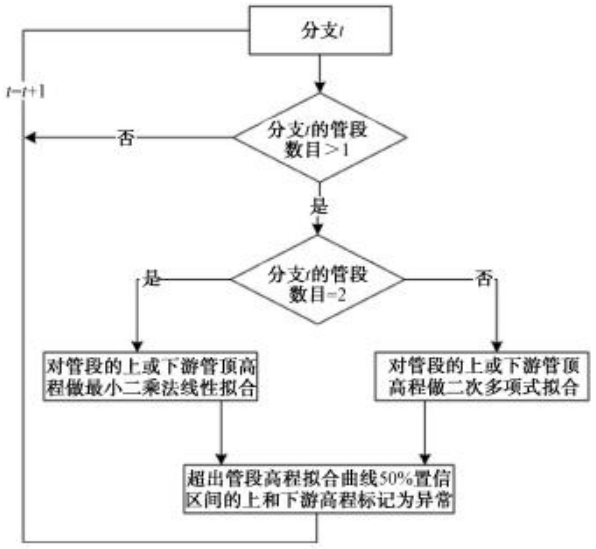

图 7　分支高程异常检测流程图

Fig. 7　Flow Chart of Abnormal Detection of Branch Elevation

分支高程修复结束后,还需要对各分支连接处的高程进行检测。若多通节点处有上游管段下游管底标高≤(下游管段上游管底最小标高-上游管段直径),则认为管段高程可能有误。若检查井处不满足井底标高≤下游管段上游管底标高或井底标高≤上游管段下游管底标高,则认为检查井井底标高可能有误。

5　程序开发

Matlab 具有强大的数据处理能力,适应于排水管网数据量大且复杂的特点,因此,本研究根据上述方法,在 Matlab 的 GUIDE 中编写交互界面,开发排水管网异常数据检测与修复程序,从而实现人机结合的数据异常处理策略。该程序包括 4 个界面,如图 8 所示。界面一是主界面,用于读取 GIS 数据、保存 GIS 数据中的列信息、将修改后的 GIS 数据写入新的 GIS 文件中,由用户输入管段编号等数据的所在列,方便程序读取。界面二是排水管网属性项归类与分析界面,由用户输入属性数据表名及待分析属性所在列,程序可计算并展示出该属性所有独立文本出现的次数及分类矩阵,用户可根据经验进行人工调整并据此进行修复。界面三是管径异常值检测与修复界面,由用户输入管径所在列,程序可计算并展示出管径所有值出现的次数,用户可将明显异常的管径修改为 0,程序可根据用户设定的参数返回相应的管径修正参考值。界面四是检测和修复管段高程界面,由用户输入相应高程所在列,将高出地面的管段高程和埋深过大的高程修改为埋深 0.7 m,接下来可以对所有分支进行曲线拟合,识别异常高程,并将曲线上的点作为异常高程的推荐修改值。程序可以将异常的上下游高程及推荐值在界面的表格中展示,用户可根据经验判断修正表格中的高程推荐值,据此修复 GIS 数据。

6　工程案例分析

以 ZH 市部分区域的排水管网 GIS 数据作为工程案例的研究对象。评估范围内有排水管道共 1 969.77 km,排水沟渠共 505.01 km。对原始数据进行预处理,用节点索引的分支管段拓扑结构表示法构造管网的管段分支集合和特殊节点集合,简化后得到排水管道共 6 805 个分支。异常数据检测与修复结果如表 3 所示。

将所有异常管段和节点写入新的 GIS 文件,检测与修复的结果可在 ArcGIS 中与原始 GIS 数据对比展示。如图 9 所示,左侧是不同类型异常的图层列表,将异常数据分为“拓扑”“管渠尺寸”“高程”和“其他”4 大类,此外可能正确的修正值也被写入 GIS 文件的属性数据库中以供参考。

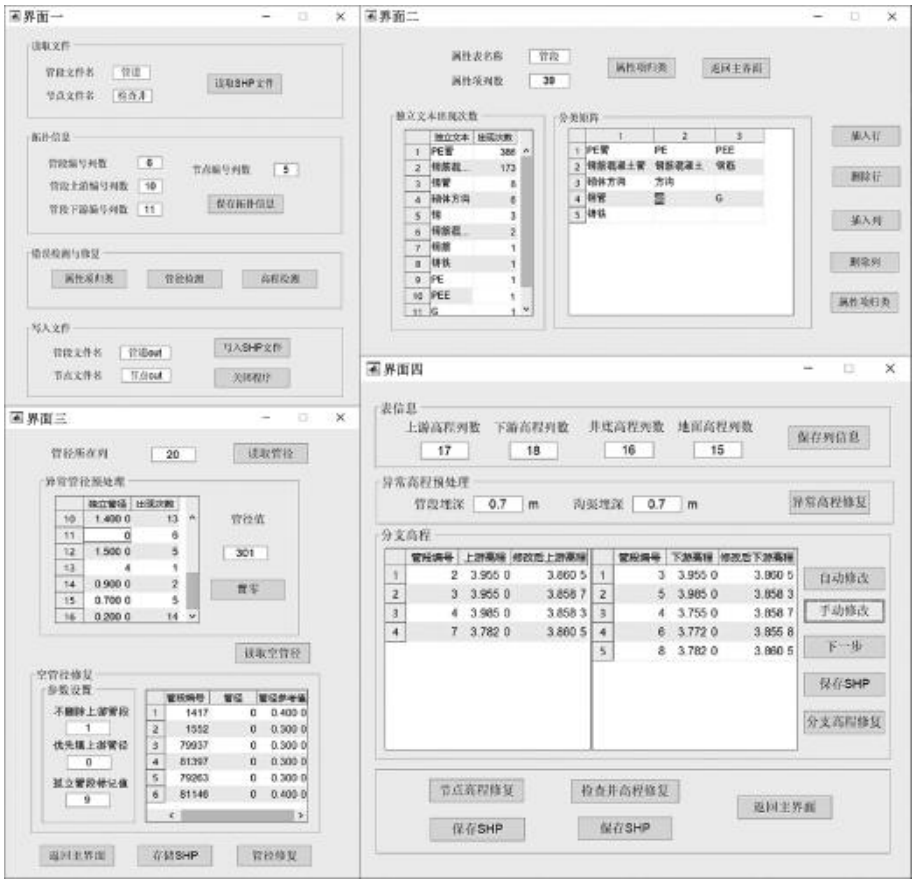

图 8 排水管网异常数据检测与修复的程序界面

Fig. 8 Detection and Recovery of Drainage Pipelines Network Abnormal Data for User Interface

表 3 ZH 市排水管网 GIS 数据异常数据检测与修复结果

Tab. 3 Results of Detection and Recovery of Drainage Pipelines Network Abnormal GIS Data of ZH City

问题类型	检测与修复情况
拓扑	孤立节点有雨水口 31 个、管线点 19 个、窨井 2 897 个；节点缺失的排水管渠 18 个；管段末梢节点共 8 686 个，其中未标注为排放口的节点数为 8 476 个；可能存在管段反向的管段数为 27 根
文本属性	存在排水管渠的管材标注为 NULL、同种材料表述方式不同、表述方式存在包含关系等问题，应将“聚氯乙烯”修改为“PVC”，“塑料”和“NULL”等表述进行人工判断后予以修正；在排水体制方面，同时存在“雨污”和“雨污合流”两种方式表达同种类型的排水体制，建议更新为数量更多的“雨污合流”；管理归属部门存在表述不清、表述错误、相同属性表述不同等问题，应结合外部信息进一步核实
管径	管径异常的管段共 146 条，其中 DN300000 共 74 条，DN0 共 69 条，DN110 共 2 条，DN301 共 1 条；根据上下游管径填充法，参数设置为 ifupprior=0，ifdel=1，Nmax=9，即不删除上游始端空管段，按照上游管径进行填充；存在管道与沟渠相连和多条管径值为 0 的管道相连的情况，填充 3 次后仅有 21 根管径被修复
高程	主干管分支管道中上游高程异常共 3 301 个，管道下游高程异常共 1 770 个；多通节点的上游 875 根管道中的下游高程不符合排水管道的连接规则；共 2 320 个检查井井底高程不符合设计规则

7 结论

目前，我国绝大部分城市都已经建成或正在建立城市排水管网 GIS 数据库，但这些数据库往往存在着不同程度的数据质量问题。只有保证高质量的数据，才能使 GIS 系统真正发挥出在排水管网信息化管理和建模分析等方面的作用。本文着重研究和实现了具有一定普适性的解决排水管网各类主要数据异常问题的方法：总结了拓扑检测的两种实现；提出了文本属性项的半自动分类与修复方法；实现了基于规则的上下游管径填充法；构造了一种节点索

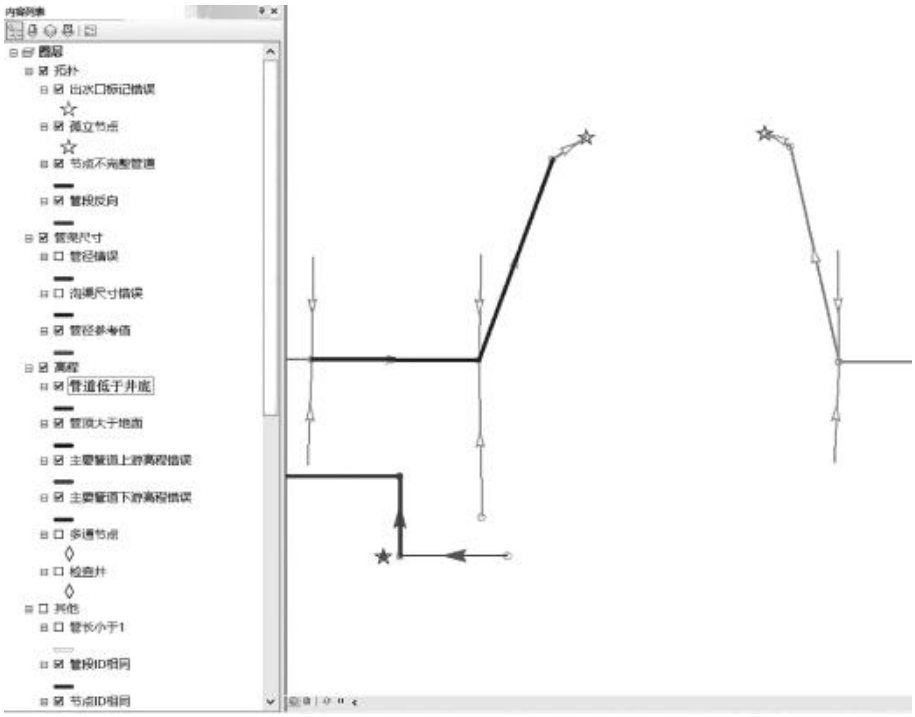

图 9　在 ArcGIS 中查看异常数据

Fig. 9　Check of Abnormal Data in ArcGIS

引的分支管段拓扑结构表示法,在此基础上实现了基于最小二乘法的分支管段高程异常检测与修复;基于上述策略开发了具有属性项归类、管径修复和高程修复功能的排水管网异常数据检测与修复程序,实现了人机交互式的异常数据检测与修复;最后将该方法应用到 ZH 市某区域排水管网 GIS 数据的异常检测与修复中,验证了方法的可行性与实用性。

参考文献

[1] 谷俊鹏, 何维华. 城市排水管网运营综合评估方法的探讨[J]. 给水排水, 2018, 54(s2): 244-251.

[2] 聂小波. 基础地理信息数据质量检查系统设计与实现[D]. 北京: 中国地质大学, 2006.

[3] 秦立为. 排水管网 GIS 系统数据质量评价与控制[D]. 上海: 同济大学, 2008.

[4] GHAVAMI S M, BORZOOEI Z, MALEKI J. An effective approach for assessing risk of failure in urban sewer pipelines using a combination of GIS and AHP-DEA[J]. Process Safety and Environmental Protection, 2020, 133: 275 - 285. DOI: 10.1016/j.psep.2019.10.036.

[5] KELLER V, FOX K, REES H G, et al. Estimating population served by sewage treatment works from readily available GIS data [J]. Science of the Total Environment, 2006, 360(1/2/3): 319-327. DOI: 10.1016/j.scitotenv.2005.08.043.

[6] PELDA J, HOLLER S. Spatial distribution of the theoretical potential of waste heat from sewage: A statistical approach[J]. Energy, 2019, 180: 751 - 762. DOI: 10.1016/j.energy.2019.05.133.

[7] 孔彦虎. 基于 GIS 的城市排水管网数据处理与校验[D]. 昆明: 云南大学, 2012.

[8] 王腾龙. 基于 GIS 的城市排水管网数据标准研究[D]. 昆明: 云南大学, 2013.

[9] 王宝利. Geodatabase 中基于规则的拓扑关系[J]. 测绘与空间地理信息, 2004, 27(3): 17-19.

[10] MANNING C D, RAGHAVAN P, SCHÜTZE H. 信息检索导论(修订版)[M]. 王斌,李鹏, 译. 北京: 人民邮电出版社, 2019.