陈晓楠, 顾起豪, 张召, 等. 南水北调中线总干渠水情数据智能清洗[J]. 南水北调与水利科技(中英文), 2024, 22(3): 436-444. CHEN X N, GU Q H, ZHANG Z, et al. Intelligent cleaning of hydrological data in the main canal of the Middle Route of the South-to-North Water Transfers Project[J]. South-to-North Water Transfers and Water Science & Technology, 2024, 22(3): 436-444. (in Chinese)

南水北调中线总干渠水情数据智能清洗

陈晓楠[1], 顾起豪[1], 张召[2], 靳燕国[1], 顾沁扬[3]

(1. 中国南水北调集团中线有限公司, 北京 100038; 2. 中国水利水电科学研究院, 北京 100038;
3. 水利部水利水电规划设计总院, 北京 100120)

摘要: 南水北调中线总干渠水位、流量等实时水情数据受外界扰动、测量系统误差等因素影响而产生的病态水情数据将造成调度模型计算失真, 甚至导致计算失败。为此, 针对上下游流量数据空间上的逻辑错误和水位数据时间序列的跳变, 分别建立基于粒子群优化的水量平衡模型和指数加权滑动平均模型, 对病态水情数据在空间、时间上实施横向、纵向清洗处理。以穿黄节制闸至漳河节制闸间的渠段为典型研究区间, 利用模型自动识别流量倒挂点, 并对该渠段涉及的 12 座节制闸、26 处分水点的流量数据进行统一修正, 实现了上下游逻辑上的合理性。同时, 选取研究渠段内的闫河节制闸为代表, 在 48 h 内运行基本稳定状态下, 对每 2 h 的闸前水位数据序列进行分析, 自动识别出跳变数据并进行合理修正。结果表明: 建立的模型可自动识别病态水情数据并进行智能清洗, 处理后的数据能够较好地满足输水调度分析决策的需要, 因此该模型具有推广应用的价值。

关键词: 南水北调中线; 数据清洗; 输水调度; 粒子群优化算法; 指数加权滑动平均模型

中图分类号: TV68 **文献标志码:** A **DOI:** 10.13476/j.cnki.nsbdqk.2024.0045

南水北调中线工程的输水调度是工程运行的核心业务, 而及时、准确的水情数据是实时调度决策最重要的依据及基础。此外, 水情数据也是智慧水利发展阶段中各种模型计算的基础。但是, 病态水情数据会使模型计算结果失真、不具备实用意义, 甚至导致模型计算失败。因此, 水情数据清洗是十分必要的。数据清洗研究最早出现在美国, 用于保险号码的修正[1]。随着大数据技术的普及和应用、各种模型的产生与发展, 数据的精准度越来越重要, 数据清洗也因此越来越受到重视。近年来, 数据清洗在各行各业[2-8]都有着长足的发展, 应用愈加广泛。

水情数据清洗是数据清洗领域在水利行业的应用与拓展[9]。侯峰等[10]针对水质异常数据, 采用孤立森林算法对异常数据实现识别和剔除, 随后基于 AdaBoost 算法对剔除后缺失数据进行插补, 该方法比随机森林算法插补效果更好, 精度更高; 薛萍[11]对调水工程水位数据清洗中异常值检测和数据填充等进行研究, 结果表明滤波+3σ 模型在异常值检测中更占优势, 而插值算法简单、合理, 适用于数据填充; 陈锃[12]针对水质采集数据中各类噪声, 提出了基于经验小波变换和多尺度模糊熵的自适应去噪方法, 该方法与小波变换、相似模态分解、完整相似模态分解和经验小波变换对比, 去噪效果更佳, 而针对数据缺失问题, 提出了一种迁移学习和长短期记忆模型相结合的方法, 可使数据填补准确率得到较大的提升; 张佳鸿等[13]针对深圳市南山区智慧水务系统大数据中存在的脏数据问题, 构建了“数据预处理-异常值检测-空缺值插补”三阶段的清洗模型, 可使脏数据平均清洗率达到 94%, 清洗效果良好; 付贵[14]充分考虑了水文数据异常识别的动态环境, 提出了基于随机森林算法的水文监测数据异常识别方法, 主要是通过改进随机森林进行异常值特征提取, 采用语义相似性度量方法实现异常数据的识别和聚类, 测试表明效果良好。

南水北调中线工程沿线布设了 64 个节制闸, 将总干渠划分为 60 多个渠段, 形成一个高度水力协同

收稿日期: 2023-11-14 **修回日期:** 2024-04-19 **网络出版时间:** 2024-05-27
网络出版地址: https://link.cnki.net/urlid/13.1430.TV.20240523.1604.008
基金项目: 国家自然科学基金项目 (52209046); 水利青年科技英才资助项目
作者简介: 陈晓楠(1979—), 男, 河北沙河人, 正高级工程师, 博士, 主要从事输水调度运行管理研究。E-mail: chenxiaonan@nsbd.cn
通信作者: 顾起豪(1995—), 男, 江苏南通人, 助理政工师, 主要从事输水调度运行管理研究。E-mail: guqihao@nsbd.cn

的“串联水库群”，工程在节制闸等处设置了水位计、流量计，实时采集水位和流量等水情数据。由于明渠输水受到外界风浪干扰、自身设备采集数据跳变等因素，水情要素在时间序列维度上有时会产生异常值。此外，由于测量系统误差等原因，可能出现上下游水情数据空间序列维度上的异常值问题。例如，稳定状态下渠段上游端实测流量比实际值偏小 2%，而下游端测流偏大 2%，单独看每个测流点精度满足要求，但产生了下游流量大于上游流量的逻辑错误，导致后续调度分析计算的失效等。针对南水北调中线水情数据中出现的流量倒挂问题，现已有部分研究成果[9]，该成果采用了区间流量最长序列法，该方法对单一闸门流量进行清洗的结果十分良好，但在长距离大范围闸门流量数据清洗过程中，清洗适用条件相对严苛，且区间流量最长序列法中的距离公式部分存在一定改进空间。因此，本文利用现代人工智能技术，对水位要素进行自身时间序列上的清洗，对流量要素进行空间上逻辑关系处理，形成清洗条件较宽泛的“纵向”与“横向”相结合的中线水情数据智能清洗方法。

1　典型研究渠段概况

选择穿黄节制闸至漳河节制闸间的渠段为典型研究区间，该段工程位于河南省郑州市荥阳市以北至河北省邯郸市磁县。从长序列历史水情数据来看，该段除穿黄、漳河以外节制闸的水情数据出现病态数据的频率较高。穿黄节制闸、漳河节制闸分别作为总干渠黄河南北分界点和河南与河北分界点，属于重要断面，长期以来重点对其流量开展率定，水情数据的可信度高。故本研究选取穿黄节制闸（桩号 483+471）作为研究段起点，漳河节制闸（桩号 731+366）作为研究段终点，研究渠段见图 1。

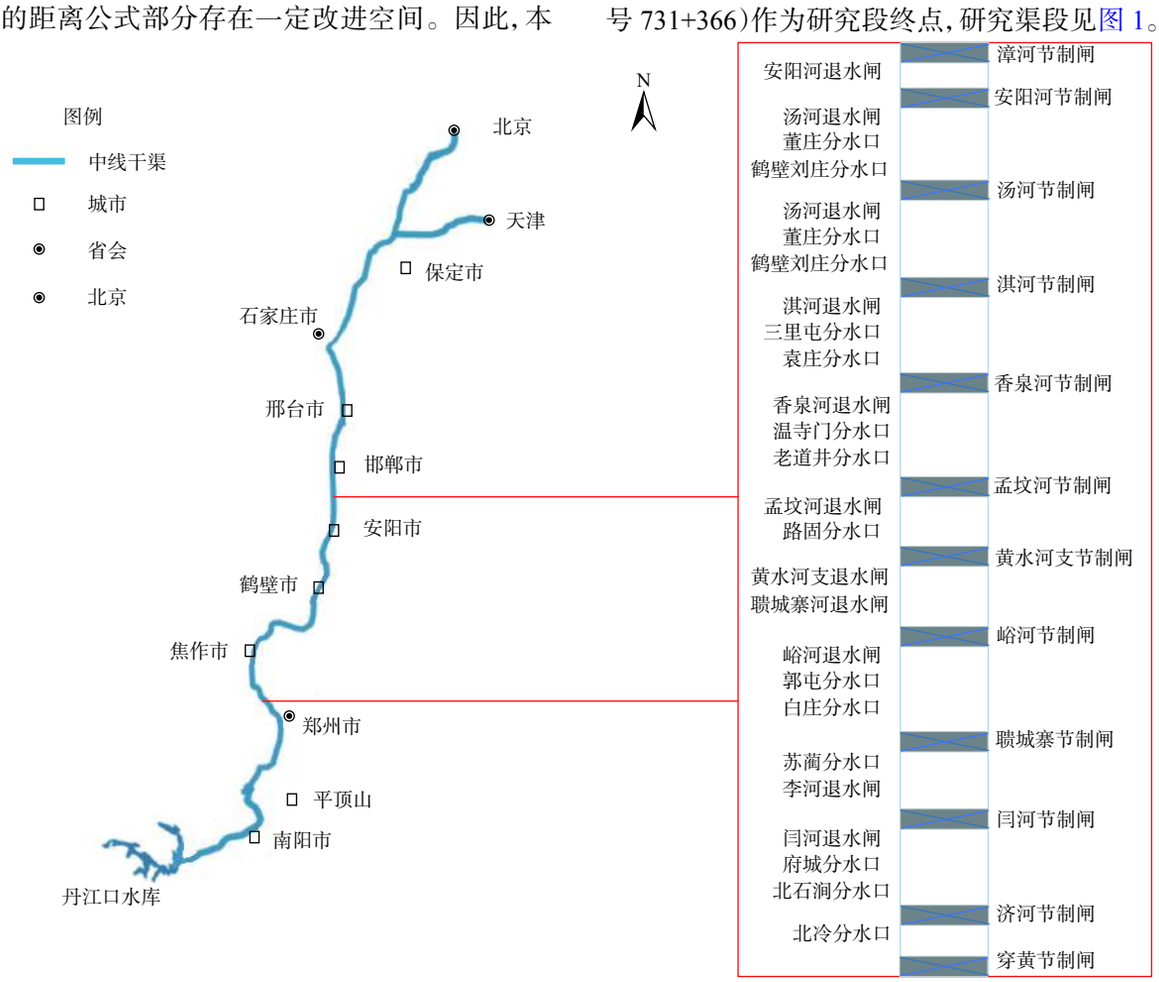

图 1　研究渠段

Fig. 1　Research channel

2　病态数据分析及清洗思路

2.1　病态数据分析

为支撑南水北调中线工程的调度运行管理，工程沿线在节制闸前后、控制闸闸后、退水闸闸前设置了水位计，节制闸处、分水口处设置了流量计，所有闸门处设置了开度仪[15]，共计 668 个水位计、163 个流量计、909 个开度仪，用以监测全线的水位、流

量、闸门开度、水温、流速等关键的水情数据信息。监测设备以每秒采集相关数据、监测平台每 0.5 h 推送至综合管理平台为原则,形成庞大的且不断扩展的数据库。

从多年的调度运行实际来看,在调度人员决策使用的综合管理平台上,时常会出现水情的病态数据。例如,流量数据一般会出现以下问题:在平稳状态下流量发生非趋势性跳变;流量数值未出现异常但上下游流量出现倒挂现象。其成因有 4 种:一是流量计发生故障,导致读数有误,甚至无法读取数据;二是数据传输过程中因丢包等因素导致数据在输水调度管理平台显示有误;三是外界扰动较强时容易造成测量数据发生突变,单次异常的持续时间往往较短,分布离散[16],产生随机误差;四是流量计本身工作和数据传输均正常,但由于系统误差的存在,上下游流量数据出现倒挂现象,不符合物理机理。水位数据常出现的问题有:水位值缺失或有误,或因外界干扰造成的数据突变现象。其成因与流量数据问题成因的前 3 类相似。

因此,南水北调中线工程的水情病态数据主要可以分为两类:一类是自身时空问题的纵向病态数据,该类病态数据主要特征是从单点数据分析就直观表达为突变数据、空值等明显错误值;另一类是存在上下游逻辑问题的横向病态数据,此类病态数据的特征为从单点数据分析来看是合理的,但不能通过上下游间数据的合理性分析(病态数据分类及成因见图 2)。鉴于流量数据的上下游关联较强、上游对下游影响明显,故流量病态数据认为均可采用横向清洗方式进行清洗,水位数据则考虑用纵向清洗方法进行清洗。

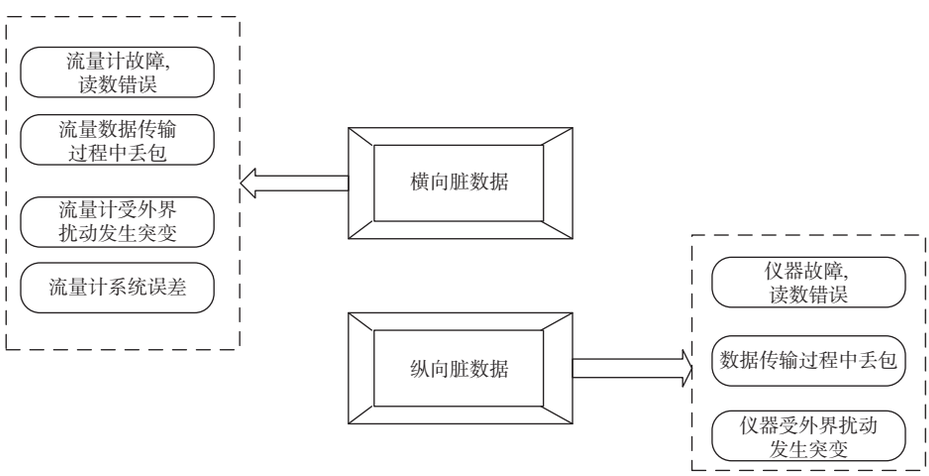

图 2　病态数据分类及成因

Fig. 2　Ill-conditioned data classification and factors

2.2　清洗思路

南水北调中线工程在调度运行过程中,闸前水位及过闸流量数据是进行调度分析、指令决策的基础,也是最核心的两种水情数据。本文主要是在平稳状态下对水位、流量数据进行清洗研究。

对于存在上下游逻辑问题的横向病态数据,由图 2 可知主要存在于流量病态数据之中。统筹考虑节制闸的过闸流量及分水口、退水闸的分、退水流量,默认节制闸数据相较于分水口、退水闸的流量数据更为准确,且首、末节制闸的流量数据准确、合理。故以流量平衡为硬约束条件,宽浅式破坏原则为目标函数(适应度函数),利用标准粒子群算法对选择时段内正在分(退)水的的分水口、退水闸的流量数据进行迭代更新(正在分(退)水的最后一个的分水口(退水闸)的流量不参与算法更新),通过控制分水口、退水闸流量数据的更新区间,加以考虑沿线的输水损失率,推求节制闸过闸流量数据,末端分水口(退水闸)的流量利用更新后的上游节制闸与下游节制闸、流量损失作差可得。详细过程如下:

第一步:建立标准粒子群算法模型,设置相应参数;

第二步:根据研究区实测数据,进行输入;

第三步:由于根据经验认为研究段首、末节制闸的流量数据是准确的,因此,通过算法更新研究段除了所选时刻正在分水(退水)的最后一个分水口

（退水闸）的其余分水口（退水闸）流量，考虑渠段损失率，相应地计算各节制闸的流量。最后一个正在分水（退水）的分水口（退水闸）由该口门所在渠段的上游节制闸流量和下游节制闸流量、流量损失相减所得，更新完毕。

第四步：验证未参与算法更新和约束的研究段所选时刻内正在分水（退水）的最后一个分水口（退水闸）的计算值是否满足更新区间，如不满足，返回第三步；如满足，继续。

第五步：输出计算后的节制闸过闸流量和更新后的分水口和退水闸的流量。

对于存在自身时空问题的纵向病态数据，究其成因，流量数据和水位数据都会出现此类问题，但考虑到流量数据本身存在很强的上下游关联性，故流量病态数据均可以通过利用横向病态数据清洗方法进行清洗，且更为有效和符合物理机理。对于水位病态数据的纵向清洗，建立指数加权滑动平均模型，自动识别超出阈值范围的病态数据，再进行修正。过程如下：

第一步：建立指数加权滑动平均法模型，设置相应参数；

第二步：根据研究区实测数据，进行输入；

第三步：人为给定误差区间，识别超出误差阈值的水位数据，将超出误差区间的水位病态数据进行清洗；

第四步：输出更新后满足误差区间的水位数据。

3　模型建立

3.1　横向病态数据清洗模型

3.1.1　标准粒子群模型

粒子群优化（partical swarm optimization, PSO）算法是一种群智能优化算法，源于对鸟群捕食行为的研究，由 Kennedy 等[17] 提出。PSO 算法存在容易陷入局部最优、收敛速度受惯性权值影响较大的缺点[18]，但是这些缺点可以通过多次重复实验和调整 PSO 算法参数的方法尽量避免[19]。PSO 算法在水库调度、水资源配置、水力模型参数研究等多方面得到了广泛的应用且效果显著[19-24]，目前发展较为成熟。

PSO 算法采用常数惯性权重和学习因子，其速度及位置公式为

$$v_{id}(t+1)=\omega v_{id}(t)+c_1r_1(P_{id}-x_{id}(t))+c_2r_2(P_{gd}-x_{id}(t)) \tag{1}$$

$$x_{id}(t+1)=x_{id(t)}+v_{id}(t+1) \tag{2}$$

式（1）和（2）中：$v_{id}(t)$表示第i个粒子的第d维度第t代的速度；$x_{id}(t)$表示第i个粒子的第d维度第t代的位置；ω为惯性因子，常数，大小程度决定了全局搜索能力；c_1为自身学习因子，决定着局部搜索能力；c_2为社会学习因子，决定着全局搜索能力；常数r_1、r_2一般为[0,1] 的随机数；P_{id}代表个体最优解；P_{gd}代表全局最优解。

3.1.2　目标函数（适应度函数）

选取的目标函数参考宽浅式破坏原则。宽浅式破坏原则一般多用于水资源配置方面的研究[25]，其概念主要是在来水不足的情况下，为防止各用水户出现大范围集中缺水，在各用水户之间均匀地分配缺水量[26-27]。将目标函数定义为各节制闸、分水口、退水闸的变化量占原流量和的比值最小，具体公式为

$$R=\sum|(Q_{\mathrm{f}}-Q_{\mathrm{fs}})|+\sum|(Q_{\mathrm{t}}-Q_{\mathrm{ts}})|+\sum\left|(Q_{\mathrm{j}}-Q_{\mathrm{js}})\right|/\left(\sum Q_{\mathrm{fs}}+\sum Q_{\mathrm{ts}}+\sum Q_{\mathrm{js}}\right) \tag{3}$$

式中：R代表总变化率；$\sum|Q_{\mathrm{f}}-Q_{\mathrm{fs}}|$代表分水口计算值和实测值之差的绝对值的和；同理，$\sum|Q_{\mathrm{t}}-Q_{\mathrm{ts}}|$、$\sum\left|Q_{\mathrm{j}}-Q_{\mathrm{js}}\right|$分别代表退水闸和节制闸计算值和实测值之差的绝对值的和；$\sum Q_{\mathrm{fs}}$、$\sum Q_{\mathrm{ts}}$、$\sum Q_{\mathrm{js}}$分别代表分水口、退水闸、节制闸实测流量的和。

3.1.3　约束条件

（1）流量平衡约束

$$Q_{\mathrm{in}}-\sum Q_{\mathrm{f}}-\sum Q_{\mathrm{t}}-\sum Q_{\mathrm{s}}=Q_{\mathrm{out}} \tag{4}$$

式中：Q_{in}表示渠段入流；$\sum Q_{\mathrm{f}}$表示渠段所有分水口的总分水流量；$\sum Q_{\mathrm{t}}$表示渠段所有退水闸的总退水流量；Q_{s}表示由研究段的输水损失率求得的各节制闸段输水损失流量的和；Q_{out}表示渠段出流。

（2）分水口、退水闸的误差约束

鉴于本文选取时段的分水口、退水闸流量均不大，因此需设置合理的误差系数（W）来约束其变化值。

$$\begin{cases} W=0.05 & 0<Q_{\mathrm{f}}<1 \\ W=0.10 & 1\leqslant Q_{\mathrm{f}}<2 \\ W=0.15 & Q_{\mathrm{f}}\geqslant 2 \end{cases} \tag{5}$$

（3）速度和位置约束

$$\begin{cases} v>V_{\max} & v=V_{\max} \\ v<V_{\min} & v=V_{\min} \end{cases} \tag{6}$$

$$\begin{cases} x>X_{\max} & x=X_{\max} \\ x<X_{\min} & x=X_{\min} \end{cases}$$

3.2　纵向病态数据清洗模型

纵向病态数据采用指数加权滑动平均模型。指数加权滑动平均(exponentially weighted moving averages)是深度学习中众多算法的一项基础方法,它是指对观察值给予不同的权重,并以最后的观察值(与计算值邻近时刻)为基础、按照历史观察值的不同权重计算当前的值。其特点在于计算时各数值的权重会随时间呈指数形式递减,越靠近当前的观察值权重越大,也就是代表越靠近当前的观察值对计算结果的影响越大。计算公式为

$$V_t = \beta^n V_{t-n} + (1-\beta)\left(\beta^{n-1}\theta_{t-n+1} + \cdots + \beta^0\theta_t\right) \tag{7}$$

$$\beta = \frac{n-1}{n}$$

式中:$\beta \in [0,1)$,代表衰减系数;θ_t代表变量V在t时刻的取值;n代表历史值数量。

4　结果分析

4.1　横向病态数据清洗

4.1.1　数据来源

考虑到平衡状态的条件,需要选取研究段分水口和退水闸流量每日变化不大,全段及穿黄节制闸往南、漳河节制闸往北若干个闸门未动作的情况。选取 2023 年 5 月 10 日上午 8 时数据作为横向病态数据清洗的基础数据,各节制闸过闸流量值见表 1,横向病态数据示图见图 3,分水口和退水闸的分退水流量见表 2。

表 1　研究段节制闸流量表

Tab. 1　The flow of the controlling gates in the study area

节制闸序号	节制闸名称	流量/($m^3 \cdot s^{-1}$)
1	穿黄隧洞出口节制闸	176.71
2	济河倒虹吸出口节制闸	174.91
3	闫河倒虹吸出口节制闸	169.71
4	聩城寨河倒虹吸出口节制闸	171.57
5	峪河暗渠进口节制闸	169.19
6	黄水河支倒虹吸出口节制闸	170.99
7	孟坟河倒虹吸出口节制闸	166.34
8	香泉河倒虹吸出口节制闸	159.30
9	淇河倒虹吸出口节制闸	156.89
10	汤河涵洞式渡槽进口节制闸	152.07
11	安阳河倒虹吸出口节制闸	152.88
12	漳河倒虹吸出口节制闸	150.33

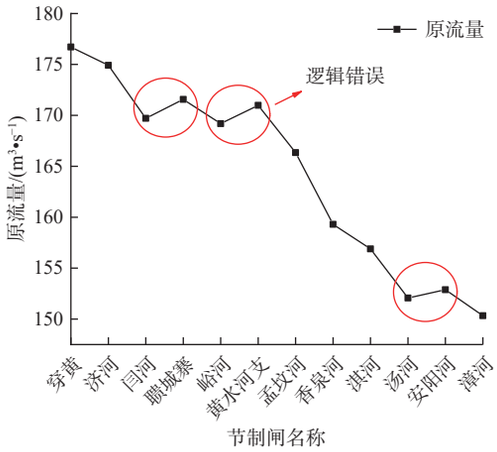

图 3　横向病态数据

Fig. 3　Horizontal pathological data

表 2　分水口、退水闸流量

Tab. 2　The flow of the diversion outlets and release gates

名称	流量/($m^3 \cdot s^{-1}$)	名称	流量/($m^3 \cdot s^{-1}$)
北冷分水口	0.22	老道井分水口	4.17
北石涧分水口	0.60	温寺门分水口	0.85
府城分水口	1.36	香泉河退水闸	5.00
闫河退水闸	0	袁庄分水口	0.50
李河退水闸	0	三里屯分水口	6.82
苏蔺分水口	1.34	淇河退水闸	0
聩城寨河退水闸	0	鹤壁刘庄分水口	0.61
白庄分水口	0	董庄分水口	1.02
郭屯分水口	0.32	汤河退水闸	0
峪河退水闸	0	小营分水口	1.33
黄水河支退水闸	0	南流寺分水口	1.92
路固分水口	0.44	安阳河退水闸	0
孟坟河退水闸	0	漳河退水闸	0

4.1.2　研究段损失率计算

利用现有水体数据以及分、退水情况计算研究段的损失率,计算公式为

$$\iota_z = \frac{w_{sta} - w_{end} - w_{fen}}{w_{sta}} \tag{8}$$

式中:ι_z代表研究段总损失率;w_{sta}代表初时刻水体,m^3;w_{end}代表末时刻水体,m^3;w_{fen}代表计算时段内分、退水总水体,m^3。

研究渠段不同运行时段的损失率结果见表 3。

研究段全长 247.90 km,共有 12 个节制闸,通过距离平均分配损失率,计算公式为

$$\iota_n = 1 - (1-\iota_z)^{\left(\frac{d_n}{d_z}\right)} \tag{9}$$

式中:ι_n代表研究段第 n 个节制闸和第 n+1 个节制闸之间渠段的损失率,n = 11;d_n研究段第 n 个节制

闸和第 n+1 个节制闸之间渠段的距离，n = 11; d_z代表研究段的总长度。研究段各节制闸间渠段损失率结果见表 4。

表 3　研究段损失率结果

Tab. 3　The results of loss rate in the study area

运行时段	损失率/%
2014年12月—2015年10月	5.70
2015年11月—2016年10月	2.50
2016年11月—2017年10月	1.53
2017年11月—2018年10月	0.44
2018年11月—2019年10月	0.39
2019年11月—2020年10月	1.31
2020年11月—2021年10月	2.44
2021年11月—2022年10月	0.05
2022年11月—2023年05月	1.29
平均值	1.74

表 4　研究段各节制闸间渠段损失率结果

Tab. 4　The results of loss rate of canals between controlling gates in the study area

节制闸序号	节制闸名称	损失率/%
1	穿黄隧洞出口节制闸	0.13
2	济河倒虹吸出口节制闸	0.20
3	闫河倒虹吸出口节制闸	0.15
4	聩城寨河倒虹吸出口节制闸	0.10
5	峪河暗渠进口节制闸	0.19
6	黄水河支倒虹吸出口节制闸	0.13
7	孟坟河倒虹吸出口节制闸	0.17
8	香泉河倒虹吸出口节制闸	0.21
9	淇河倒虹吸出口节制闸	0.17
10	汤河涵洞式渡槽进口节制闸	0.20
11	安阳河倒虹吸出口节制闸	0.10
12	漳河倒虹吸出口节制闸	

4.1.3　数据清洗结果及分析

在 2023 年 5 月 10 日上午 8 时的研究段流量数据中，有 3 处出现明显的下游流量大于上游流量的横向逻辑错误的现象。以流量平衡为基础，通过标准粒子群模型迭代计算，经多次重复实验调整，迭代参数见表 5，计算结果如表 6。

表 5　迭代参数

Tab. 5　Iteration parameters

迭代参数	取值	迭代次数
ω	0.8	200
c_1=c_2	2.0	500
r_1 = r_2	[0,1)的随机数	1 200

表 6　标准粒子群算法结果

Tab. 6　The results of PSO algorithm

迭代次数	最佳适应度值	最佳适应度值代数
200	0.014 906	118
500	0.014 837	200
1 200	0.014 794	954

从图 4 可以清晰地看出，当迭代 1 200 次时，在第 954 代的时候达到最小适应度值，收敛效果良好，认为未陷入早熟。鉴于目标函数的特性，取适应度值最小为最优解，更新后的研究段分水口、退水闸流量清洗结果及占原流量比值情况见表 7，研究段节制闸流量清洗结果见表 8，节制闸过闸流量清洗前后对比见图 5。

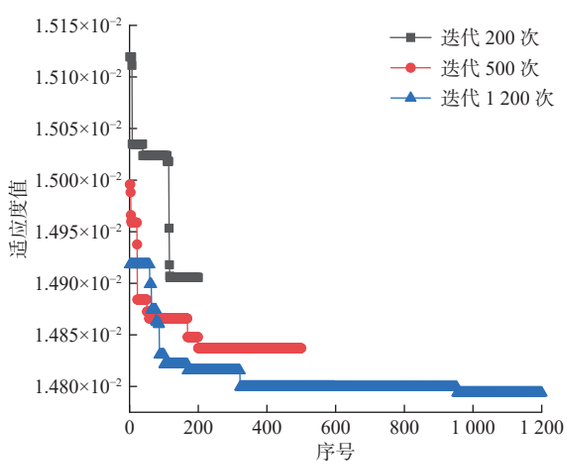

图 4　适应度值计算图

Fig. 4　Fitness value diagram

通过实例验证，标准粒子群算法对横向病态数据清洗有着明显的效果，原实际测量的闫河节制闸—聩城寨节制闸、峪河节制闸—黄水河支节制闸、汤

河节制闸—安阳河节制闸这 3 组具有横向逻辑错误特征的数据得到了清洗修正,更新后流量均在原测量值的流量计误差范围内,分水口、退水闸修正值都在更新控制范围内,且清洗后研究段符合流量平衡原理,清洗效果良好。

表 7　研究段分水口、退水闸流量清洗结果及占比

Tab. 7　The results and proportion of diversion outlets and release gates in the study area after cleaning

名称	流量/($m^3 \cdot s^{-1}$)	占原流量百分比/%
北冷分水口	0.231	105.0
北石涧分水口	0.630	105.0
府城分水口	1.496	110.0
苏蔺分水口	1.215	90.7
郭屯分水口	0.336	105.0
路固分水口	0.418	95.0
老道井分水口	3.545	85.0
温寺门分水口	0.808	95.0
香泉河退水闸	4.250	85.0
袁庄分水口	0.475	95.0
三里屯分水口	5.797	85.0
鹤壁刘庄分水口	0.560	95.0
董庄分水口	0.918	90.0
小营分水口	1.197	90.0
南流寺分水口	1.877	97.8

表 8　研究段节制闸流量清洗结果

Tab. 8　The flow cleaning results of controlling gates in the study area

节制闸序号	节制闸名称	流量/($m^3 \cdot s^{-1}$)
1	穿黄隧洞出口节制闸	176.71
2	济河倒虹吸出口节制闸	176.25
3	闫河倒虹吸出口节制闸	173.77
4	聩城寨河倒虹吸出口节制闸	172.30
5	峪河暗渠进口节制闸	171.80
6	黄水河支倒虹吸出口节制闸	171.47
7	孟坟河倒虹吸出口节制闸	170.84
8	香泉河倒虹吸出口节制闸	161.94
9	淇河倒虹吸出口节制闸	155.32
10	汤河涵洞式渡槽进口节制闸	153.56
11	安阳河倒虹吸出口节制闸	150.48
12	漳河倒虹吸出口节制闸	150.33

4.2　纵向病态数据清洗

4.2.1　数据来源

考虑平衡状态下节制闸闸前水位较为平稳或处于缓变状态,在日常调度过程中,调度人员一般认为在调度平稳阶段同一节制闸相邻时刻水位差值不应大于 0.03 m,否则需要修正。因此,在平稳状态下对同一节制闸相邻时刻变化幅度较大(水位差值大于 0.03 m)的节制闸闸前水位需要进行纵向病态数据的清洗。

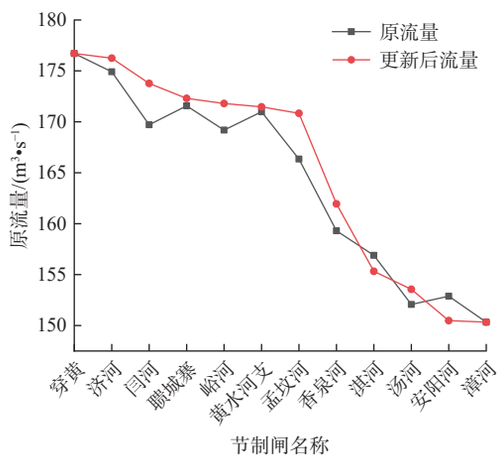

图 5　节制闸过闸流量清洗前后对比

Fig. 5　Controlling gate flow cleaning results comparison chart

选择 2023 年 7 月 21 日 8 时至 2023 年 7 月 23 日 8 时闫河节制闸闸前水位数据作为纵向清洗的基础。鉴于当时南水北调中线工程已进入大流量输水阶段,全线流量大、水位高,上游济河节制闸、闫河节制闸和下游聩城寨节制闸均已提离水面退出调度,受调度影响较小。

4.2.2　分析计算

运用指数加权滑动平均模型对闫河节制闸闸前水位进行识别与清洗。由图 6 可得:第一个病态数据出现在 n=11(7 月 24 日 04:00),因此,式(7)中参数 β 的取值为 $\beta_1 = 0.91$;第二个病态数据出现在 n=22(7 月 25 日 02:00),因此,式(7)中参数 β 的取值 $\beta_2 = 0.95$。将参数代入公式中计算,结果见图 6。

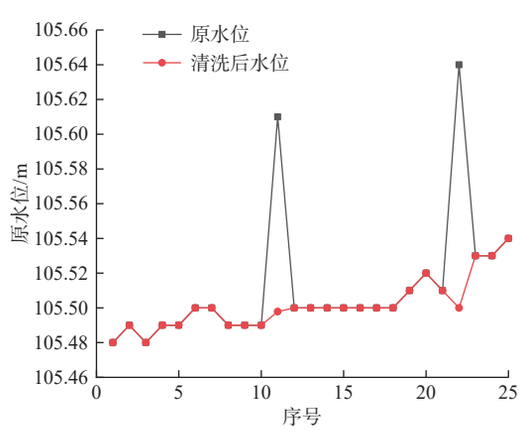

图 6　闫河节制闸水位清洗前后对比

Fig. 6　Comparison of water level of Yanhe controlling gate after cleaning

结果显示，经过指数加权滑动平均模型的清洗，病态水位数据得到很好的修正，相邻时刻的水位变化值回到正常范围内，且与相邻（上一时刻）水位值的关联较大，符合输水调度规律，清洗效果良好。

5　结 论

南水北调中线工程通水运行以来积累了大量的水情数据。在运行时间增长、设备设施老化和干渠沿线气候条件变化等多方面因素影响下，水情数据在采集、传输、展示等各方面不可避免地会出现错误，产生病态数据。本文将水情病态数据原因归纳为存在自身时空问题的纵向病态数据和存在上下游逻辑问题的横向病态数据。

本研究的主要目的是将水情数据中最为常用的流量、水位的病态数据优化至合理区间范围内，以便于数据能够较好地满足输水调度分析决策的需要，为后期中线输水调度数字孪生中模型建立提供部分数据基础。因此，针对横向病态数据，选用了目前运用广泛、成熟度高，且建模计算相对简易、效果良好的标准粒子群算法（PSO 算法），以宽浅式破坏原则为目标函数建立模型，结果表明：具有横向逻辑错误的节制闸流量数据得到了很好的修正，更新后流量均在原测量值的流量计误差范围内，分水口、退水闸修正值都在控制范围内，研究段符合流量平衡原理，清洗效果良好，病态数据回归至合理区间范围内。针对存在自身时空问题的纵向病态数据问题，选取经典有效的指数加权滑动平均模型，结果表明：病态水位数据回到正常范围内，且与相邻（上一时刻）水位值的关联较大，符合输水调度规律，清洗效果良好。因此，该在相似的大型调水工程中具有一定的推广应用的价值。

参考文献：

[1] GALHARD H, FLORESCU D, SHASHA D, et al. An extensible framework for data cleaning[C]//Data Engineering, 2000 Proceedings of 16th International Conference Data Engineering. San Diego, CA, USA. IEEE, 2000: 312. DOI: 10.1109/ICDE.2000.839429

[2] 亓星, 修德皓, 程关文, 等. 滑坡变形监测数据的实时过滤方法及应用[J]. 水利水电技术(中英文), 2022, 53(7)：129-138. DOI: 10.13928/j.cnki.wrahe.2022.07.012.

[3] 李莉, 梁袁, 林娜, 等. 考虑时空相关性的风电机组风速清洗方法[J/OL]. 太阳能学报, 1-9.[2024-05-22]. https://doi.org/10.19912/j.0254-0096.tynxb.2023-0201.. DOI: 10.19912/j.0254-0096.tynxb.2023-0201

[4] 王婧怡, 陈胤佳, 袁野, 等. 面向*K*-近邻学习模型的高效数据清洗框架[J]. 计算机科学与探索, 2023, 17(9)：2241-2251. DOI: 10.3778/j.issn.1673-9418.2207105.

[5] 李晨阳. 基于相关性的时序数据清洗方法研究[D]. 沈阳: 沈阳航空航天大学, 2022. DOI:10.27324/d.cnki.gshkc.2022.000239. DOI: 10.27324/d.cnki.gshkc.2022.000239

[6] 琚佳彬, 赵健, 杨克新. 考虑数据潮流模型的配电网关键节点辨识[J]. 水利水电技术(中英文), 2023, 54(9)：26-36. DOI: 10. 13928 / j. cnki. wrahe. 2023. 09.003.

[7] LI S, HE H, ZHAO P, et al. Data cleaning and restoring method for vehicle battery big data platform[J]. Applied Energy, 2022, 320：119292. DOI: 10.1016/J.APENERGY.2022.119292.

[8] 万文华, 梁雪容, 郑航, 等. 中国水电站数据库构建、验证与定量分析[J]. 水利水电技术(中英文), 2022, 53(12)：185-195. DOI: 10.13928/j.cnki.wrahe.2022.12.019.

[9] 位文涛, 靳燕国, 张召, 等. 南水北调中线工程流量监测站点倒挂数据清洗模型及应用[J]. 南水北调与水利科技 (中英文), 2022, 20(6)：1158-1167. DOI: 10.13476/j.cnki.nsbdqk.2022.0114.

[10] 侯锋, 李朋, 庞洪涛, 等. 基于AdaBoost算法的水质监测数据清洗方法[J]. 水电站机电技术, 2023, 46(5)：109-111, 126. DOI: 10.13599/j.cnki.11-5130.2023.05.031.

[11] 薛萍. 调水工程水位数据清洗及预测模型研究[D]. 济南: 济南大学, 2022. DOI: 10.27166/d.cnki.gsdcc.2022.000736

[12] 陈锃. 钱塘江流域水质时间序列数据清洗及预警研究[D]. 杭州: 杭州电子科技大学, 2022. DOI: 10.27075/d.cnki.ghzdc.2022.000287

[13] 张佳鸿, 陈兴晖. 南山区智慧水务系统及大数据清洗模型的构建与应用[J]. 水利技术监督, 2021, 29(12)：32-35,121. DOI: 10.3969/j.issn.1008-1305.2021.12.011.

[14] 付贵. 基于改进随机森林算法的水文监测数据异常识别研究[J]. 水利科技与经济, 2022, 28(8)：76-80. DOI: 10.3969/j.issn.1006-7175.2022.08.017.

[15] 陈晓楠, 靳燕国, 许新勇, 等. 南水北调中线干线智慧输水调度的思考[J]. 河海大学学报 (自然科学版), 2023, 51(5)：46-55. DOI: 10.3876/j.issn.1000-1980.2023.05.007.

[16] 李阳, 沈小军, 张扬帆, 等. 基于速度-关联约束的风电机组风速感知异常数据识别方法[J]. 电工技术学报, 2023, 38(7)：1793-1807. DOI: 10.19595/j.cnki.1000-6753.tces.211893.

[17] KENNEDY J, EBERHART. R. Particle swarm optimization[J]. Proc of 1995 IEEE Int Conf Neural Networks, 2011, 4(8): 1942-1948. DOI: 10.1007/978-0-398-30164-8_630.
[18] 吴昌友, 王福林, 马力. 一种新的改进粒子群优化算法[J]. 控制工程, 2010, 17(3): 359-362. DOI: 10.14107/j.cnki.kzgc.2010.03.010.
[19] 刘欣蔚, 王浩, 雷晓辉, 等. 粒子群算法参数设置对新安江模型模拟结果的影响研究[J]. 南水北调与水利科技, 2018, 16(1): 69-74,208. DOI: 10.13476/j.cnki.nsbdqk.20180011.
[20] 杜佰林. 基于模拟退火粒子群算法的大荔县水资源优化配置研究[D]. 西安: 西安理工大学, 2021. DOI: 10.27398/d.cnki.gxalu.2021.000819.
[21] 贾本有, 吴时强, 范子武, 等. 粒子群算法在河道水动力模型参数校正中的应用[J]. 南水北调与水利科技, 2018, 16(3): 143-148. DOI: 10.13476/j.cnki.nsbdqk.2018.0080.
[22] 李彤姝, 黄睿, 孙志鹏, 等. 基于多目标粒子群算法的渠系优化配水研究[J]. 灌溉排水学报, 2020, 39(9): 95-100,125. DOI: 10.13522/j.cnki.ggps.20190294.
[23] 宋健蛟, 赵红莉, 蒋云钟. 粒子群算法在密云水库供水优化配置中的应用[J]. 南水北调与水利科技, 2015, 13(2): 378-381. DOI: 10.13476/j.cnki.nsbdqk.2015.02.041.
[24] 张宏洋, 韩鹏举, 马聪, 等. 基于改进粒子群算法的土石坝动力参数反演研究[J]. 水利水电技术 (中英文), 2023, 54(6): 110-123. DOI: 10.13928/j.cnki.wrahe.2023.06.010.
[25] 李景海. 基于规则的水资源配置模型研究[D]. 北京: 中国水利水电科学研究院, 2005.
[26] 马兴华, 左其亭. 区域尺度初始水权分配模型及应用研究[C]//环境变化与水安全: 第五届中国水论坛论文集, 北京: 中国水利水电出版社, 2007: 681-385.
[27] 韦瑞深, 严子奇, 周祖昊, 等. 基于宽浅式破坏原则的水库旱限水位优化方法[J]. 水资源保护, 2023, 39(4): 152-158,166. DOI: 10.3880/j.issn.1004-6933.2023.04.019.

Intelligent cleaning of hydrological data in the main canal of the Middle Route of the South-to-North Water Transfers Project

CHEN Xiaonan[1], GU Qihao[1], ZHANG Zhao[2], JIN Yanguo[1], GU Qinyang[3]

(1. *China South-to-North Water Diversion Middle Route Co., Ltd, Beijing* 100038, *China*;
2. *China Institute of Water Resources and Hydropower Research, Beijing* 100038, *China*;
3. *General Institute of Water Resources and Hydropower Planning and Design, Ministry of Water Resources, Beijing* 100120, *China*)

Abstract: The real-time hydrological data such as water level and discharge of the main canal of the Middle Route of the South-to-North Water Transfers Project are affected by external disturbances, measurement system errors and other factors. The ill-conditioned hydrological data will cause the calculation distortion of the scheduling model, and even lead to the failure of the calculation.

Aimed at the logical errors in the upstream and downstream flow data space and the jump of the time series of water level data, the water balance model based on particle swarm optimization and the exponential weighted moving average model were established respectively, and the pathological water regime data was cleaned horizontally and vertically in space and time. Taken the channel section between the Yellow River controlling gate and the Zhanghe River controlling gate as a typical research interval, the flow inversion point was automatically identified by the model. The flow data of 12 controlling gates and 26 water diversion points involved in the channel section were uniformly corrected to realize the rationality of upstream and downstream logic. At the same time, the Yanhe controlling gate in the research section was selected as the representative. Under the basic stable state of operation within 48 hours, the water level data sequence in front of the gate every 2 hours was analyzed, and the jump data was automatically identified and reasonably corrected.

The results showed that the model established could automatically identify the pathological water regime data and carried out intelligent cleaning. The processed data was able to better meet the needs of water transfer scheduling analysis and decision-making. So the model has the value of popularization and application.

Key words: Middle Route of the South-to-North Water Transfers Project; data cleaning; water dispatching; particle swarm optimization algorithm; exponential weighted moving average model