第 48 卷第 6 期 物 探 与 化 探 Vol.48,No.6
2024 年 12 月 GEOPHYSICAL & GEOCHEMICAL EXPLORATION Dec.,2024

doi: 10.11720/wtyht.2024.1554
吉子健,周志超,赵敬波,等.高放废物地质处置新场候选场址地下水位异常值识别方法[J].物探与化探,2024,48(6):1530-1538.http://doi.org/10.11720/wtyht.2024.1554
Ji Z J,Zhou Z C,Zhao J B,et al.A method for identifying anomalous values of groundwater levels at candidate sites for the geological disposal of high-level radioactive waste[J].Geophysical and Geochemical Exploration,2024,48(6):1530-1538.http://doi.org/10.11720/wtyht.2024.1554

高放废物地质处置新场候选场址地下水位异常值识别方法

吉子健[1,2],周志超[1,2],赵敬波[1,2],季瑞利[1,2],张明[1,2]
(1. 核工业北京地质研究院 环境工程研究所,北京 100029; 2. 国家原子能机构高放废物地质处置创新中心,北京 100029)

摘 要: 地下水动态监测为高放废物地质处置候选场址的安全评价提供了关键基础数据,但研究发现实际的监测数据中存在较多异常值,严重干扰了对动态过程的准确判断。因此,亟须建立一种高效的方法对异常值进行准确识别。本文基于局部加权回归的时间序列分解和最小协方差行列式方法构建了地下水位异常值检测组合模型,使最小协方差行列式方法可以在更独立的残差项中进行异常值检测。结果表明,构建的组合模型相较于最小协方差行列式方法的单一模型,其对异常数据具有更好的敏感性和检测精度;并进一步确定了组合模型的阈值应接近实际的异常值比例,以获取最佳的检测效果;此外,根据新场地段 BSQ01、BSQ25、BS35、BS26 钻孔的水位数据对组合模型的适用性进行验证,表明其能够准确识别出混淆于大量正常水位数据中的异常值,同时也适用于不同类型异常事件的检测。

关键词:时间序列异常检测;STL 分解;最小协方差行列式方法;高放废物;地质处置
中图分类号: P641 **文献标识码**: A **文章编号**: 1000-8918(2024)06-1530-09

0 引言

深地质处置作为国际上公认的处置高水平放射性废物(高放废物)的有效方法,目前已有多个国家开展了相关的研发工作。我国高放废物地质处置研究工作始于 1985 年,经过前期全国性的普查、筛选及论证,于 2019 年确定新场场址为我国首座高放废物地质处置地下实验室场址,用于开发和验证处置技术、评价处置场址的适宜性[1-2]。地下实验室开挖过程中的地下水动态响应特征,不仅是正确认识场址水文地质条件的重要因素,亦是工程建设、现场试验、场址特性评价等相关技术的重要依据,对高放废物处置库场址的最终确定具有十分重要的意义。

伴随着地下水动态监测技术的不断发展[3-4],在新场地段构建了技术先进的地下水长期动态监测网络,并由此获取了大量的地下水动态监测数据。然而研究发现,实际的监测数据中存在较多的异常值,严重干扰了对地下水动态过程的准确判断。如何从大量的监测数据中识别出水位异常值成为了地下水动态研究中亟待解决的热点问题。对此,部分学者将统计学方法或机器学习方法引入到长时间序列、大批量数据的异常值检测工作中,大幅提高了识别效率与检测结果的客观性[5-7],验证了自动化方法的可行性。目前,被广泛应用的自动识别算法可归纳为基于数理统计[8]、基于预测模型[9-10]和基于距离[11-13]3 种类型。其中,基于距离的检测算法与其他两种算法相比,更加适用于对孤立异常值的检测[14],该方法已被应用到了城市排水管网液位监测[15]、地下水环境监测[16]、网络安全监测[17]、医疗

收稿日期: 2023-12-21; **修回日期**: 2024-06-25
基金项目: 铀资源探采与核遥感全国重点实验室基金项目(NKLUR-2024-QN-004);国防科工局核设施退役治理专项科研项目(科工二司〔2022〕736 号);中核集团 2022 年基础研究项目(CNNC-JCYJ-202206)
第一作者: 吉子健(1997-),男,助理工程师,硕士,2022 年毕业于中国地质大学(北京),主要从事高放废物地质处置水文地质研究工作。Email:18332606901@163.com

数据评估[18]等领域异常事件的识别中。而最小协方差行列式方法(MCD)[19-20]因对局部异常值反映更加灵敏,适用于小样本异常事件的检测,在众多基于距离的算法中具有广泛的适用性,如:Sunderland等[18]通过对不同模型检测效果的对比,发现 MCD 方法对医疗领域错误诊断数据的识别具有更好的检测精度,能够辅助提高数据质量;孙杰[21]以 MCD 方法对教学成绩的异常值进行检测,认为其能够更快、更具有鲁棒性地识别学生成绩中的异常数据。此外,也有一些学者对检测算法进行了深入优化,提出了更为复杂的检测模型[22],如混合模型[23]、时域频域组合模型[24]和自适应模型[25]等,以提高模型的识别精度。但优化后的算法随着复杂度的提高,增加了参数的维度,难以快速获取最佳的参数组合[26],无法保障高频率、高精度监测背景下异常值检测的时效性。

综上所述,为降低地下水位异常值对场址水文地质特征评价工作的干扰,本文选取了能够更加快速、准确地识别局部异常值的 MCD 方法,同时考虑到地下水位数据的时变特征,又将基于局部加权回归进行时间序列数据分解的统计学方法[27](STL 分解)和 MCD 方法[28]结合,提出了一种适用于高放废物地质处置新场候选场址高频率、高精度地下水位异常值检测的研究方法,并通过与单一 MCD 方法的异常检测模型进行对比,验证了其在实际应用中的准确性和适用性。

1　研究区概况

新场候选场址位于甘肃省酒泉市肃北县。场址地形多为低山丘陵,海拔一般在 1 650~1 800 m。所在区域气候条件属半沙漠大陆性气候,夏季酷热,冬季严寒。多年平均降雨量为 60~80 mm[29],多年平均蒸发量大于 3 000 mm。

区域范围内地下水主要类型为松散岩类孔隙水、碎屑岩类裂隙孔隙水、变质岩类裂隙水及火成岩类裂隙水(图 1)。其中,新场场址范围内以火成岩类裂隙水为主,钻孔揭露岩体渗透系数普遍小于 10^{-8} m·s^{-1},属于典型的低渗透性花岗岩裂隙岩体。区内地下水在断裂构造和地形的控制作用下,以新场场址中部南北分水岭为界向 NE 和 SE 方向径流。地下水系统的主要补给来源为大气降水的垂向入渗补给,并以侧向径流的方式向下游地区排泄。

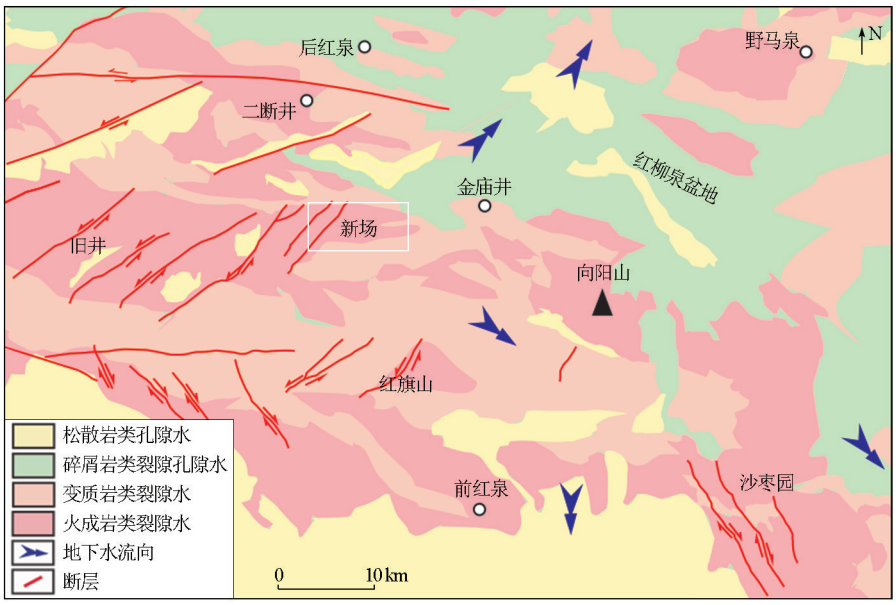

图 1　新场地段周边水文地质

Fig.1　Hydrogeological map of the Xinchang site

2　研究方法

2.1　数据准备

水位数据取自于研究区内的浅部(<100 m)和深部(约 600 m)监测钻孔(图 2),采用 Solinst Levelogger 3001 或 Seametrics LevelScout 2X 型水位计自动获取监测数据,相应的传感器测量精度为 0.1%FS。为保证数据分析结果的一致性,将水位数据样本间隔统一调整为 1 d。考虑到监测数据中还同时包含固体潮、蒸散发等因素的影响,且地下水位在受上述因素影响下的微动态变化幅度一般小于 10 cm。因此,通

过比较相邻时刻的水位样本值来标注异常数据样本，将水位差大于 15 cm 的样本作为异常事件的起始时刻，并将其恢复至正常变化水平的时刻作为异常事件的结束时刻，以降低水位微动态效应的影响。起始和结束时刻内包含的超出正常水位变化范围的水位值则被标注为待检测的异常值。

结合已有地下水位数据中异常值的标注结果，选取 BS35、BSQ01、BS26 和 BSQ25 共 4 个钻孔（图 2 中绿色钻孔），验证模型算法在研究区地下水位异常值检测工作中的适用性。同时结合 4 个钻孔中呈现的水位异常波动特征，将其划分为由地表洪流或降雨后沿钻孔井壁通道快速入渗、抽水取样等因素影响下的几种响应类型，以分析模型对不同类型异常事件的检测能力。其中，对于沿井壁快速入渗水量引起的水位

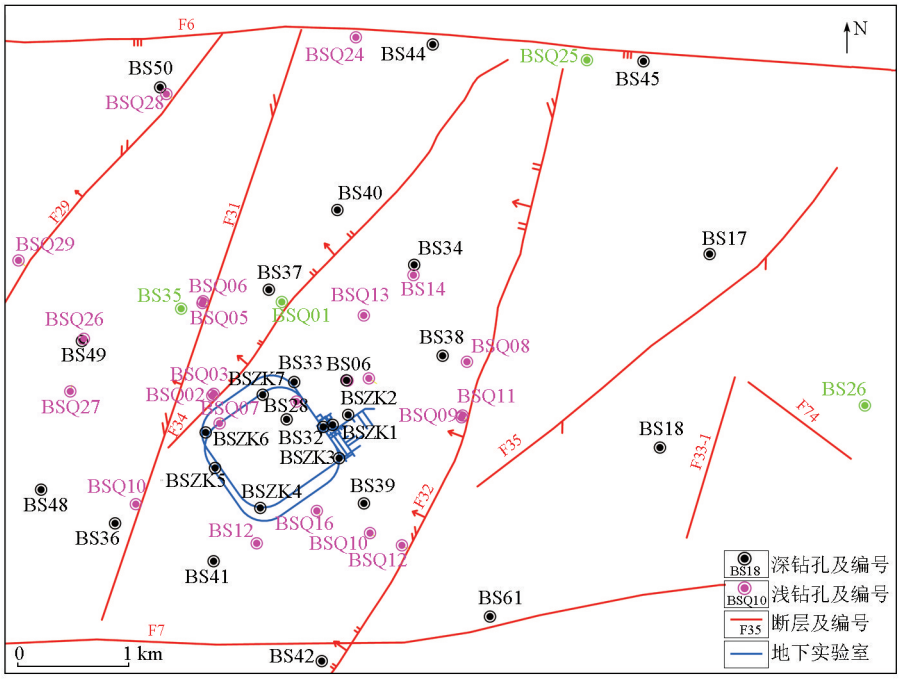

图 2 新场地段内地下水位监测钻孔位置

Fig.2 Location map of monitoring boreholes around the Xinchang site

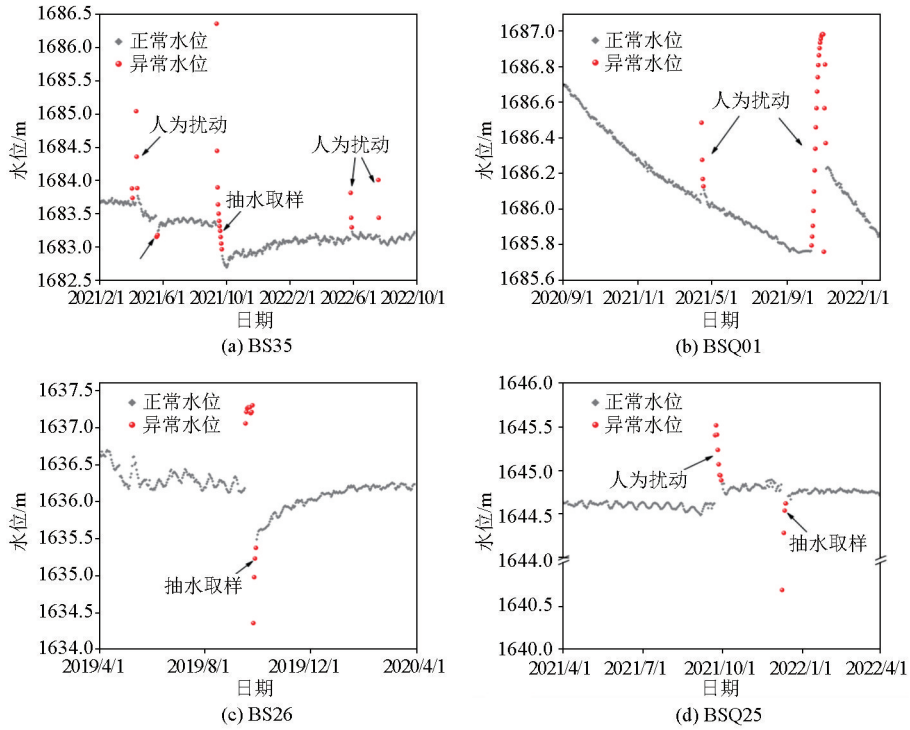

图 3 钻孔地下水位监测数据及异常值分布

Fig.3 Distribution of groundwater level monitoring data and anomaly values in boreholes

异常变化,受限于入渗量大小的差异,呈现出先骤然上升后又恢复至初始时刻水位或高于初始时刻水位的两种不同特征(图 3a、b);而因抽水取样造成的水位异常波动,则呈现出水位骤然下降后逐渐缓慢恢复的异常特征(图 3c、d)。

2.2　基于 MCD 方法的异常值检测单一模型

最小协方差行列式方法(MCD)依托数据序列中的异常值会影响整体分布的特点,通过迭代方法计算每个样本子集的协方差矩阵,并获取具有最小协方差行列式的样本子集[28,30]。在迭代计算的过程中该模型能够平衡剔除异常点和最小化协方差行列式对样本子集选择的影响,确保选择的样本子集能够保持正常数据的基本结构,削弱异常值影响,提高 MCD 方法在异常值检测中的鲁棒性和准确性。子集协方差矩阵计算公式为:

$$\hat{\sum}_{MCD}=\frac{k_{MCD}(h,n,p)}{h-1}\sum_{i\in\zeta_{MCD}}(x_i-\hat{\mu}_{MCD})(x_i-\hat{\mu}_{MCD})^{-1},\tag{1}$$

式中:$\hat{\sum}_{MCD}$ 为 MCD 方法计算的样本子集 ζ_{MCD} 的协方差矩阵;h 为样本子集的样本数量;n 为样本数据集中的样本数量;p 为每个样本的维数;$k_{MCD}(h,n,p)$ 为保证协方差估计量一致性和无偏性的比例常数;$\hat{\mu}_{MCD}$ 为计算的样本子集的均值;x_i 为第 i 个数据样本。

对获取的样本子集,计算其均值和协方差矩阵,并依据公式(2)[31]计算各个数据点到样本子集整体的距离,将偏离数据中心的样本标记为异常值;此外,通过不断调整模型阈值来优化检测结果,挑选出最优的阈值大小作为判别标准,完成最终的异常值检测工作。本研究选取 pycaret 模型库中的 anomaly 模型来实现此算法。

$$RD(x)=\sqrt{(x-\hat{\mu}_{MCD})^{t}\hat{\sum}_{MCD}^{-1}(x-\hat{\mu}_{MCD})},\tag{2}$$

式中:$RD(x)$ 为样本 x 距离样本子集整体的距离;$\hat{\mu}_{MCD}$ 为 MCD 方法计算的样本子集的均值;$\hat{\sum}_{MCD}$ 为 MCD 方法计算的样本子集协方差矩阵。

2.3　结合 STL 分解与 MCD 方法的异常值检测组合模型

基于 MCD 方法的单一模型仅是对协方差行列式的估计,忽略了时间序列数据的时间依赖性和序列相关性,导致无法对异常值进行准确识别。因此,在传统 MCD 方法的基础上引入 STL 分解方法,用于处理数据的时间序列相关性,降低趋势性和微动态效应对检测结果的干扰,使 MCD 方法可以在更独立、相对无关的残差部分进行异常检测,增加模型对异常值的敏感程度,提高识别结果的准确性和可靠性。

STL 分解方法作为异常值检测的预处理过程,其主要原理为基于局部加权回归的方法,将给定的时间序列数据分解为趋势项、周期项和残差项(式(3))[27]。在预处理过程中可分为内循环和外循环两个独立的迭代过程,其中,内循环主要处理数据序列的趋势性和周期性,外循环则是通过添加稳健性权重项,来抑制数据异常值对时间序列分解的影响。

$$Y_t=T_t+S_t+R_t\ ,\tag{3}$$

式中:Y_t 为 t 时刻的数据的实际值;T_t 为时间序列分解后的趋势项;S_t 为时间序列分解后的周期项;R_t 为时间序列分解后的残差项;t 为数据样本的时间,$t=1,2,\cdots,N$。

在完成 STL 分解后,依托单一 MCD 方法对拆分后的残差项序列执行异常值检测,其操作流程如图 4 所示。其中 STL 分解方法的实现选取了 statsmodels 统计分析库中的 seasonal.STL 模型算法,而 MCD 异常值检测算法则采取与单一模型相同的实现形式。

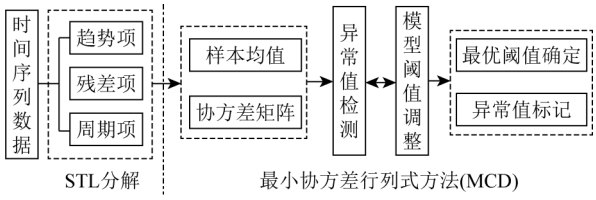

图 4　STL 分解和最小协方差行列式方法组合的异常检测流程

Fig.4　Anomaly detection process for the combined STL decomposition and minimum covariance determinant method

3　研究结果

3.1　评价指标

为定量评价时间序列异常检测算法的精度,引入了精确率、召回率和 F_1 值 3 个评价指标作为模型性能的判别标准,更全面地评价检测结果的可靠性。其中,精确率反映的是模型检测到的异常数据中实际异常值的占比;召回率是指模型正确识别的异常值占全部异常值的比例,反映模型对于异常值的查全率;而 F_1 值则是精确率和召回率的加权平均值。求取的 3 个评价指标值越大,表明模型检测的精度越高。计算方法见式(4)~(6)[32]。

$$P=\frac{T_P}{T_P+F_P}\ ,\tag{4}$$

$$R = \frac{T_P}{T_P + F_N}, \tag{5}$$

$$F_1 = \frac{2 \cdot P \cdot R}{P + R}, \tag{6}$$

式中：P 为精确率；R 为召回率；T_P 为模型将实际的异常值预测为异常值的样本数量；F_N 为模型将实际的异常值预测为正常值的样本数量；F_P 为模型将实际的正常值预测为异常值的样本数量。

3.2 模型阈值确定

MCD 方法中模型阈值的大小决定了数据序列中被判定为异常值的比例，合理的设置阈值可以帮助正确捕捉数据中的全部异常事件，提高模型的检测精度。考虑到选取的 4 个钻孔中异常值比例在 3.4%～4.6%不等，为评价不同模型阈值情况下的检测效果，确定最优的模型阈值大小，将模型阈值设定在 2.0%～8.0%区间范围内，并以精确率、召回率和 F_1 值 3 个评价指标作为不同阈值条件下模型性能的判别标准。

图 5 为以 3 个不同评价指标为标准反映的模型性能随阈值的变化情况。BSQ01、BSQ25、BS26 和 BS35 钻孔的最优阈值分别为 5.2%、2.6%、3.6%和 3.0%，且模型性能的变化以最优阈值为界可分为两种不同的变化趋势。首先，当模型阈值小于最优阈值时，3 个评价指标变化一致，均随着设定阈值的增大而增加，当达到最优阈值时模型性能最佳（图 5 灰色区域）；其次，当模型阈值大于最优阈值时，精确率和 F_1 值呈现降低趋势，召回率则呈现平稳变化趋势，表明当阈值增大到一定程度后，模型的宽容度也随之增大，将存在更多的正常样本被误判，造成模型的检测精度逐渐下降（图 5 蓝色区域）。此外，研究结果还表明，4 个钻孔获取的最优阈值均小于或等于模型实际的异常值比例（图 5），表明构建的模型能够较好地适应实际的数据特征，具有较高的稳定性，能够准确识别出混淆的异常值。

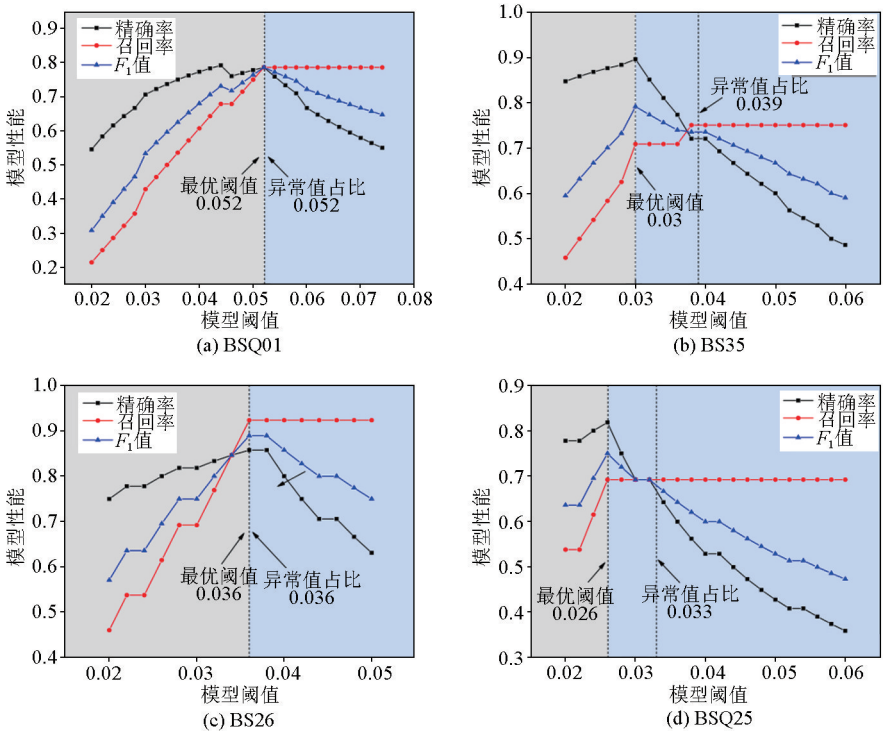

图 5 模型性能随模型阈值变化

Fig.5 Map of model performance with model thresholds

3.3 组合模型与单一模型检测效果对比

依托确定的最优阈值，将构建的组合模型与仅采用 MCD 方法的单一模型进行对比，评价不同模型的检测效果。整体上，组合模型能够有效地识别大部分水位异常值（图 6b、7b、8b、9b），而单一模型仅可识别出两次水位大幅度上升的异常事件，模型漏报率较高（图 6a、7a、8a、9a）；且单一模型与组合模型相比，其识别的异常事件中包含了较多的正常值，模型误报率更高。分析原因，认为其检测精度较差可归因于模型仅采用了基于距离的处理方法，无法准确识别出地下水位数据中的微动态效应，在标注异常样本时错误处理了部分周期性变化特征，因而模型误报率较高。而组合模型在进行异常检测之前，首先应用STL分解剔除了数据中的趋势性和微

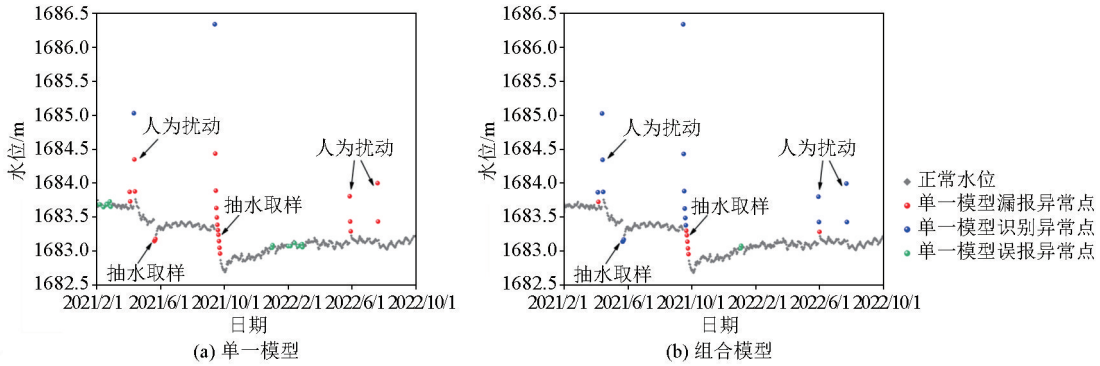

图 6　BS35 号孔组合模型与单一模型检测效果

Fig.6　Combined model and single model detection results in borehole BS35

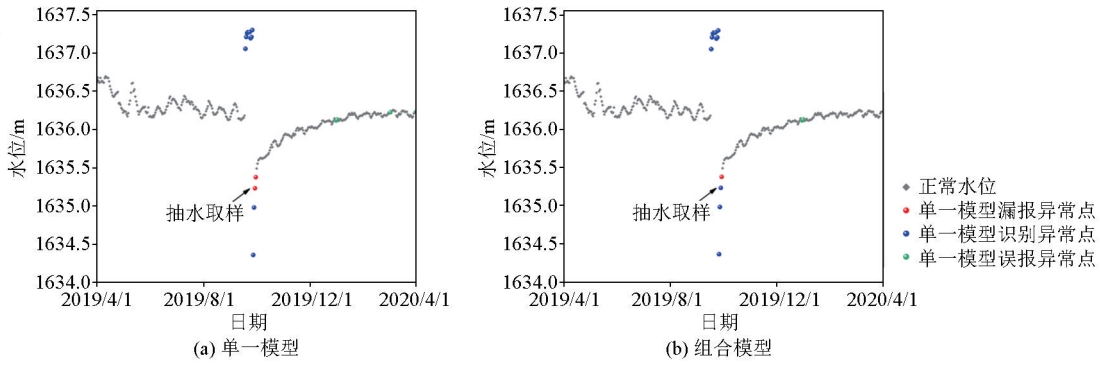

图 7　BS26 号孔组合模型与单一模型检测效果

Fig.7　Combined model and single model detection results in borehole BS26

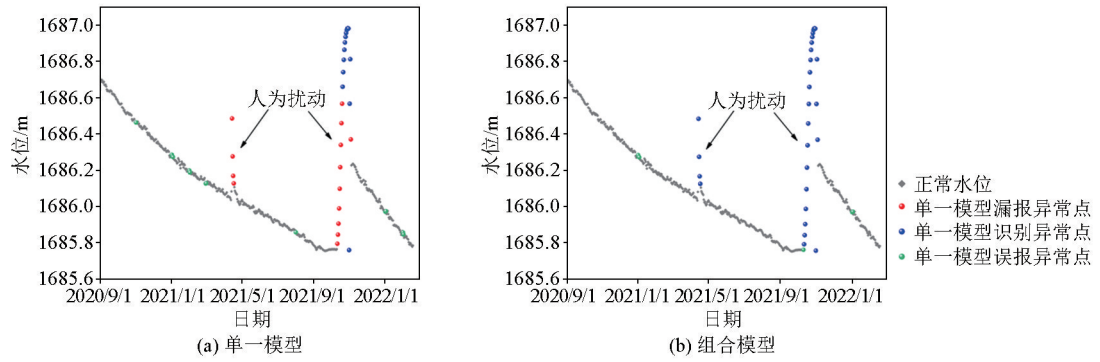

图 8　BSQ01 号孔组合模型与单一模型检测效果

Fig.8　Combined model and single model detection results in borehole BSQ01

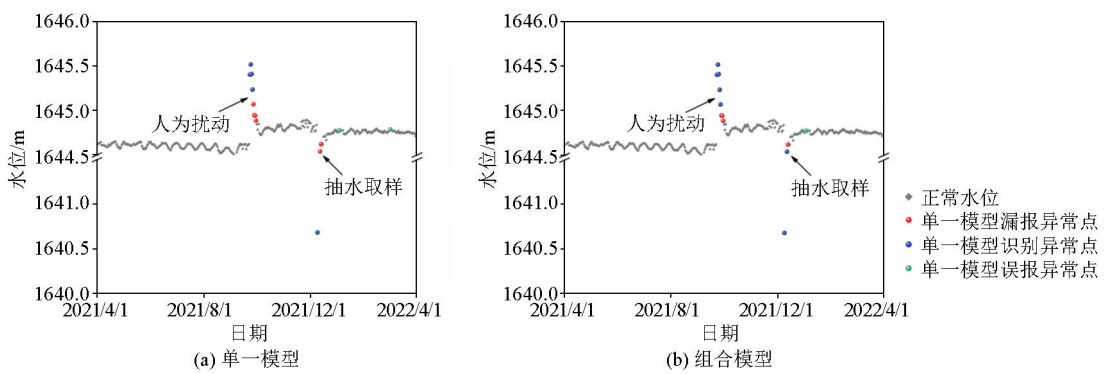

图 9　BSQ25 号孔组合模型与单一模型检测效果

Fig.9　Combined model and single model detection results in borehole BSQ25

动态效应,仅处理更独立、相对无关的残差部分,提高了模型对于一些小幅度水位异常事件的敏感性和检测精度。

此外,对于 4 个钻孔中几种不同类型异常事件的检测结果,组合模型基本能够对各种不同波动幅度的异常事件进行准确识别,仅在 BS35 号孔中对因抽水取样造成的水位骤然下降事件检测结果较差(图 6b)。在对 BS35 号孔检测流程中的每个步骤进行单独分析时发现,产生这种偏差的原因是由于 STL 分解方法将未能识别的异常值错误处理为了具有下降趋势的正常水位,使得分解后的残差项中该部分异常值的差异特征与正常点相差不大,导致 MCD 方法无法对其进行准确识别。而单一模型因受到水位微观波动等因素的影响,仅能识别出部分水位波动在 1 m 以上的异常事件类型(图 7a、8a),对于 1 m 以内的小幅度水位异常波动则均存在较大程度的漏报,模型的适用性较差。

基于此,计算了 BSQ01、BSQ25、BS26、BS35 这 4 个钻孔地下水位异常值检测结果的精确率、召回率和 F_1 值(表 1),以对组合模型的实际应用效果进行定量评价。可以看出,构建的组合模型与单一模型相比,在具有较高召回率的同时也能够兼顾良好的精确率。由此可见,构建的组合模型能够准确识别出混淆于大量正常水位数据中的异常值,在实际应用中取得了较好的检测效果,对于地下实验室场址周边地下水位监测数据中的异常值检测具有较好的适用性,并能够适用于不同类型地下水位异常事件的识别。

表 1 场址周边钻孔地下水位异常值检测评价指标得分

Table 1 Evaluation index score for groundwater level anomaly detection in boreholes around the site

钻孔号	单一模型			组合模型		
	精确率	召回率	F_1 值	精确率	召回率	F_1 值
BS26	0.71	0.76	0.74	0.85	0.92	0.88
BS35	0.11	0.08	0.09	0.89	0.71	0.79
BSQ01	0.50	0.50	0.50	0.78	0.64	0.71
BSQ25	0.64	0.54	0.58	0.82	0.69	0.75

4 结论与展望

1)依托 STL 时间序列分解和基于距离的 MCD 方法构建了地下水位异常检测组合模型,使 MCD 方法可以在更独立、相对无关的残差部分进行异常值检测,提高了模型对异常数据的敏感性和检测精度;以精确率、召回率和 F_1 值 3 个评价指标为标准,分析了模型性能随阈值的变化情况,表明模型阈值在接近实际异常值比例时,检测效果最佳。

2)将构建的组合模型应用到新场地段 BSQ01、BSQ25、BS26、BS35 等 4 口钻孔地下水位异常值的检测中,并通过与单一模型进行对比,证明其能够准确识别出混淆于大量正常水位数据中的异常值;且相比于单一模型,能够更好地适用于不同类型和不同水位波动幅度下地下水位异常事件的检测。

尽管现阶段已构建了适用于高放废物处置场址地下水位异常值的检测方法,并取得了较好的应用效果,但地下水动态监测是一个长期性工作,在未来的实际工作中仍存在当前研究未考虑的地下水位异常事件类型,仍需要进一步获取更新的水位监测数据,完善构建的异常值检测方法,提高模型的适用性。此外,还需要进一步构建适用于高放废物处置场址地下水位时间序列数据的缺失值处理方法,以保障地下水位监测数据的连续性。

参考文献(References):

[1] 郭永海,王驹,金远新.世界高放废物地质处置库选址研究概况及国内进展[J].地学前缘,2001,8(2):327-332.
Guo Y H,Wang J,Jin Y X.The general situation of geological disposal repository siting in the world and research progress in China [J].Earth Science Frontiers,2001,8(2):327-332.

[2] Wang J,Chen L,Su R,et al.The Beishan underground research laboratory for geological disposal of high-level radioactive waste in China:Planning,site selection,site characterization and in situ tests[J].Journal of Rock Mechanics and Geotechnical Engineering,2018,10(3):411-435.

[3] Calderwood A J,Pauloo R A,Yoder A M,et al.Low-cost,open source wireless sensor network for real-time,scalable groundwater monitoring[J].Water,2020,12(4):1066.

[4] Drage J,Kennedy G.Building a low-cost,internet-of-things,real-time groundwater level monitoring network[J].Groundwater Monitoring & Remediation,2020,40(4):67-73.

[5] Muharemi F,Logofătu D,Leon F.Machine learning approaches for anomaly detection of water quality on a real-world data set[J].Journal of Information and Telecommunication,2019,3(3):294-307.

[6] Pang G S,Shen C H,Cao L B,et al.Deep learning for anomaly detection:A review[J].ACM Computing Surveys,2021,54(2):1-38.

[7] Schmidl S,Wenig P,Papenbrock T.Anomaly detection in time series:A comprehensive evaluation[J].Proceedings of the VLDB Endowment,2022,15(9):1779-1797.

[8] Rousseeuw P J,Hubert M.Anomaly detection by robust statistics [J].WIREs Data Mining and Knowledge Discovery,2018,8(2):e1236.

[9] Yu Y,Zhu Y L,Li S J,et al.Time series outlier detection based on sliding window prediction[J].Mathematical Problems in Engineering,2014:1-14.

[10] Kulanuwat L, Chantrapornchai C, Maleewong M, et al. Anomaly detection using a sliding window technique and data imputation with machine learning for hydrological time series[J]. Water, 2021, 13(13): 1862.

[11] Cabana E, Lillo R E, Laniado H. Multivariate outlier detection based on a robust Mahalanobis distance with shrinkage estimators[J]. Statistical Papers, 2021, 62(4): 1583-1609.

[12] Sripriya T P, Srinivasan M R, Gallo M. Robust distance measure to detect outliers for categorical data[J]. Soft Computing, 2020, 24(18): 13557-13564.

[13] Li J B, Izakian H, Pedrycz W, et al. Clustering-based anomaly detection in multivariate time series data[J]. Applied Soft Computing, 2021, 100: 106919.

[14] Smiti A. A critical overview of outlier detection methods[J]. Computer Science Review, 2020, 38: 100306.

[15] 何黎,陈磊,纪莎莎,等.基于 K-shape 聚类的连续液位监测数据异常检测方法[J].中国给水排水,2023,39(11):56-61.
He L, Chen L, Ji S S, et al. Abnormal detection of continuous water level monitoring data based on K-shape clustering[J]. China Water & Wastewater, 2023, 39(11): 56-61.

[16] Shi H X, Guo J, Deng Y D, et al. Machine learning-based anomaly detection of groundwater microdynamics: Case study of Chengdu, China[J]. Scientific Reports, 2023, 13(1): 14718.

[17] Ayadi A, Ghorbel O, Obeid A M, et al. Outlier detection approaches for wireless sensor networks: A survey[J]. Computer Networks, 2017, 129(1): 319-333.

[18] Sunderland K M, Beaton D, Fraser J, et al. The utility of multivariate outlier detection techniques for data quality evaluation in large studies: An application within the ONDRI project[J]. BMC Medical Research Methodology, 2019, 19: 102.

[19] Hardin J, Rocke D M. Outlier detection in the multiple cluster setting using the minimum covariance determinant estimator[J]. Computational Statistics & Data Analysis, 2004, 44(4): 625-638.

[20] Hubert M, Debruyne M, Rousseeuw P J. Minimum covariance determinant and extensions[J]. WIREs Computational Statistics, 2018, 10(3): e1421.

[21] 孙杰.基于 FAST-MCD 算法的异常成绩检测研究[J].现代计算机,2021,27(29):59-62.
Sun J. Research on the abnormal grade detection based on the FAST-MCD algorithm[J]. Modern Computer, 2021, 27(29): 59-62.

[22] Zhou Y J, Ren H R, Li Z W, et al. Anomaly detection via a combination model in time series data[J]. Applied Intelligence, 2021, 51(7): 4874-4887.

[23] Lin S, Clark R, Birke R, et al. Anomaly detection for time series using VAE-LSTM hybrid model[C]//ICASSP 2020 ~ 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020: 4322-4326.

[24] Yokkampon U, Chumkamon S, Mowshowitz A, et al. Anomaly detection using variational autoencoder with spectrum analysis for time series data[C]//2020 Joint 9th International Conference on Informatics, Electronics & Vision (ICIEV) and 2020 4th International Conference on Imaging, Vision & Pattern Recognition (icIVPR), 2020: 1-6.

[25] Lyu J M, Wang Y Q, Chen S J. Adaptive multivariate time-series anomaly detection[J]. Information Processing & Management, 2023, 60(4): 103383.

[26] Samariya D, Thakkar A. A comprehensive survey of anomaly detection algorithms[J]. Annals of Data Science, 2023, 10(3): 829-850.

[27] Cleveland R B, Cleveland W S. STL: A seasonal-trend decomposition procedure based on Loess[J]. Journal of official statistics, 1990, 6(1): 3-73.

[28] Rousseeuw P J, Driessen K V. A fast algorithm for the minimum covariance determinant estimator[J]. Technometrics, 1999, 41(3): 212-223.

[29] Li J B, Zhang Y K, Zhou Z C, et al. Using multiple isotopes to determine groundwater source, age, and renewal rate in the Beishan preselected area for geological disposal of high-level radioactive waste in China[J]. Journal of Hydrology, 2024, 629: 130592.

[30] Hubert M, Debruyne M. Minimum covariance determinant[J]. WIREs Computational Statistics, 2010, 2(1): 36-43.

[31] Rousseeuw P J, Hubert M. Robust statistics for outlier detection[J]. WIREs Data Mining and Knowledge Discovery, 2011, 1(1): 73-79.

[32] 李航.统计学习方法[M].北京:清华大学出版社,2012.
Li H. Statistical learning methodology[M]. Beijing: Tsinghua University Press, 2012.

A method for identifying anomalous values of groundwater levels at candidate sites for the geological disposal of high-level radioactive waste

JI Zi-Jian[1,2], Zhou Zhi-Chao[1,2], Zhao Jing-Bo[1,2], JI Rui-Li[1,2], ZHANG Ming[1,2]

(1. *Division of Environmental Engineering, Beijing Research Institute of Uranium Geology, Beijing* 100029, *China*; 2. *CAEA Innovation Center for Geological Disposal of High-Level Radioactive Waste, Beijing* 100029, *China*)

Abstract: Dynamic groundwater monitoring provides critical foundational data for the safety assessment of candidate sites for the geological disposal of high-level radioactive waste. However, research has revealed that actual monitoring data frequently contain numerous anomalous values, severely interfering with the accurate assessment of the dynamic monitoring process. Therefore, there is an urgent need to develop an efficient method to accurately identify these anomalous values. This study built a combined model for anomalous value detection of the groundwater level using local weighted regression-based time series decomposition and the minimum covariance de-

terminant (MCD) method. This combined model allowed the MCD method to achieve anomaly detection in more independent residuals. Results indicate that the combined model exhibited higher sensitivity and detection accuracy for anomalous data than the single MCD model. Furthermore, this study established that the threshold of the combined model should be close to the actual proportion of anomalous values to achieve optimal detection results. Besides, this study validated the applicability of the combined model using groundwater level data from boreholes BSQ01, BSQ25, BS35, and BS26 at the new site. The validation results demonstrate that the combined model can accurately identify anomalous values amidst a large volume of data on the normal groundwater level and is applicable to the detection of different types of anomalous events.

Key words: time-series anomaly detection; STL decomposition; minimum covariance determinant (MCD) method; high-level radioactive waste; geological disposal

(本文编辑:蒋实)