DOI:10.20040/j.cnki.1000-7709.2024.20230445

基于 IPSO-DBSCAN 的抽水蓄能机组状态监测数据异常检测方法

张金鹏,张孝远

(河南工业大学电气工程学院,河南 郑州 450001)

摘要:抽水蓄能机组状态监测数据受采集设备故障、通信设备异常等因素影响,数据集中存在部分异常数据,对后续机组健康状态评估及预测造成不利影响。为此,提出了一种基于改进粒子群优化算法和 DBSCAN 密度聚类算法的机组异常数据检测模型,模型针对粒子群算法易陷入局部最优解的问题对算法进行改进,之后引入轮廓系数作为适应度函数对 DBSCAN 的参数进行寻优,最后以相关系数评价异常值剔除的效果。对国内某抽水蓄能机组 2020 年 2 月初~3 月末实测导叶开度、有功功率及下机架振动数据的实例分析结果表明,所提方法能够有效检测出机组振动监测异常数据,剔除异常值后的数据相关系数得到提高,可为后续机组健康状态评估与预测奠定数据基础。

关键词:抽水蓄能;异常值检测;改进粒子群优化算法;DBSCAN

中图分类号:[TV734.1] **文献标志码**:A **文章编号**:1000-7709(2024)02-0152-05

1 引言

“双碳”目标下,我国正在构建高比例新能源电力系统。间歇性新能源的大规模接入,增加了电源侧调峰、调频压力,降低了系统稳定水平。抽水蓄能是减小新能源随机性和波动性负面影响、促进新能源灵活消纳和高效利用的最重要技术手段。为应对间歇性新能源的波动性,抽水蓄能机组需要在发电、抽水、调相、启停等多种工况之间切换[1],机组经常需要在稳定区边界运行。机组轴系在频繁启停及工况切换过程中受到水力冲击、机械失效和电磁不平衡等因素的影响可能诱发机组异常振动、结构疲劳、电气故障、运行方式破坏等各种故障与事故[2],危害性极其严重。基于状态监测数据对机组开展状态维护,对于保障抽蓄机组安全稳定性具有十分重要的意义。然而,由于存在传感器短时失效、外界干扰以及传输错误等因素[3],产生的异常值降低了原始数据的质量,给机组运行状态的准确评估带来困难,因此对抽蓄机组状态监测数据进行异常值检测,消除噪声数据,提升数据质量十分必要。万尹豪[4]在搭建水资源数据清洗系统时采用阈值判定+拉伊达准则/箱型图法[5]实现了数据异常值的检测,但基于数理统计的检测方法在样本量大、数据维度高的情况下检测精度较差;金容鑫等[6]利用 K-means 聚类方法对水电机组监测数据进行检测,有效识别出了功率一定子线圈温度的错误数据,然而 K-means 聚类中心个数 k 值对聚类效果影响较大,且算法本身仅能聚类出球状的簇,在数据分布复杂的情况下效果较差;闫亚男等[7]将 DBSCAN 聚类方法应用于水电遥测时间序列数据中,在时间序列异常值检测上取得了较好的效果,但其试验数据为下导油槽油位的一维时间序列,振动信号在机组运行过程中受功率、水头和导叶开度等工况参数变化影响较大,仅考虑单一振动指标而忽视工况参数之间的联系易导致异常识别精度下降;段然等[8]在考虑了工况参数与振动信号的关系后,构建了包含水头、有功功率和下机架振动的数据集,并用 DBSCAN 对其进行异常数据检测,采用剔除异常值后的数据绘制机组劣化曲线,所得曲线相较于原始数据及 3σ 准则筛选后

收稿日期:2023-03-23,**修回日期**:2023-05-08

基金项目:国家自然科学基金项目(51409095)

作者简介:张金鹏(1999-),男,硕士研究生,研究方向为电力设备维护自动化,E-mail:1192030766@qq.com

通讯作者:张孝远(1981-),男,博士、教授、硕导,研究方向为电力设备维护自动化、大数据分析与智能系统的应用,E-mail:freedon@haut.edu.cn

的数据绘制的曲线更为平滑,趋势更显著,但其在DBSCAN 的输入参数邻域半径(E_{Eps})和邻域内最少包含点数(M_{Minpts})的选取上仍采用人工选取的方式,主观影响较强,降低了模型筛选的精度。为此,本文将改进粒子群算法(IPSO)与DBSCAN 结合,将 DBSCAN 参数的人工选取过程转换为由 IPSO 自动寻优,以期提升模型筛选的可靠性和精度,最后,基于国内某抽水蓄能机组实测数据进行实例分析,将所提方法与 PSO-DBSCAN、K 均值聚类(Kmeans)、孤立森林(IF)和鲁棒性随机分割森林(RRCF)算法进行比较,并引入相关系数作为评价异常值检测效果的指标,验证了所提方法的优越性。

2　构建基于 IPSO-DBSCAN 的异常数据检测模型

2.1　改进粒子群优化算法

PSO 算法的参数中,惯性权重 ω 和学习因子 c_1、c_2 对算法的收敛过程有重要作用。其中,惯性权重的意义在于保持粒子运动的惯性并扩大粒子运动的范围。ω 值大,算法全局寻优能力强,局部寻优能力弱。ω 值小,算法全局寻优能力弱,局部寻优能力强。因此,为使 PSO 算法在迭代过程初期获得较强的全局寻优能力,在后期获得较强的局部收敛能力,引入了带有随迭代次数下降的惯性权重的 PSO 算法,即标准 PSO 算法。然而,标准 PSO 算法仅考虑了惯性权重的影响,学习因子仍采用定值。本文在标准 PSO 算法的基础上引入非对称学习因子,在算法迭代初期,采用较大的 c_1 值与较小的 c_2 值,使粒子更注重个体的最优解,减少集群的影响,这样可以使粒子尽可能覆盖整个搜索空间,随着迭代次数的增加,c_1 递增,c_2 递减,可加强粒子向全局最优点的收敛能力。c_1、c_2 同样采用线性变化的方式,更新公式分别为:

$$c_1 = c_{1\mathrm{start}} + (c_{1\mathrm{start}} - c_{1\mathrm{end}})k/K \tag{1}$$

$$c_2 = c_{2\mathrm{start}} + (c_{2\mathrm{start}} - c_{2\mathrm{end}})k/K \tag{2}$$

式中,$c_{1\mathrm{start}}$、$c_{2\mathrm{start}}$ 分别为 c_1、c_2 的初值,$c_{1\mathrm{end}}$、$c_{2\mathrm{end}}$ 分别为 c_1、c_2 的终值,另 $c_{1\mathrm{start}} \neq c_{2\mathrm{end}}$、$c_{2\mathrm{start}} \neq c_{1\mathrm{end}}$,即为非对称学习因子;$k$ 为当前迭代次数;K 为迭代总数。

2.2　DBSCAN 聚类算法

DBSCAN 是一种经典空间密度聚类算法,以密度作为相似度,能够聚类出任意形状的数据簇,该算法可以用于分离噪声或异常点。DBSCAN 通过邻域半径(E_{Eps})和邻域内最少包含点数(M_{Minpts})来划分识别并划分数据簇。DBSCAN 原理示意图见图 1。

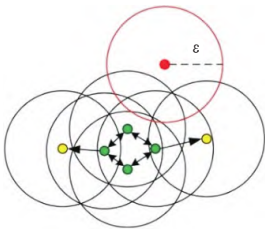

图 1　DBSCAN 原理示意图

Fig. 1　DBSCAN principle diagram

图 1 中,ε 为 DBSCAN 的邻域半径;绿点表示核心对象,即邻域内包含至少 M_{Minpts} 个样本的对象;黄点表示边界点,即该点在核心对象的邻域内而该点本身不是核心对象;红点表示离群点,该点既不是核心对象又不是边界点。

2.3　IPSO-DBSCAN 异常值检测方法

DBSCAN 算法聚类效果的优劣依赖于 E_{Eps} 和 M_{Minpts} 两个重要参数的选取。在使用中,这两个参数的值往往根据经验选取或由 K-距离图选取,之后人工观察并手动调整参数,在这个过程中,人工干预过多且主观观察存在误差。因此,本文通过 IPSO 算法寻求最优 E_{Eps} 与 M_{Minpts},以此构建 DBSCAN 聚类模型并得到最优聚类结果。这种方法剔除了参数选取时的人工误差并可以寻得最优参数,达到数据最佳聚类效果。

2.3.1　适应度函数的选取

粒子的适应度就是目标函数值,是 PSO 算法寻优的目标。在本文所提模型中,DBSCAN 是一种聚类模型,其聚类效果的优劣可以用聚类效果评估指标来衡量。数据聚类效果评价指标主要有轮廓系数、CH 分数和戴维森堡丁指数(DBI)三种。在 IPSO 算法中,粒子位置朝着适应度函数的最大值寻优,而 DBI 指标越小代表聚类效果越好,极值为零,因此 DBI 并不适合作为本文的适应度函数。轮廓系数、CH 分数均是得分越高代表聚类效果越好,其中,轮廓系数范围为[−1,1],而 CH 分数得分无上限,在试验中易得到千万甚至上亿的得分,不利于结果的呈现。综上所述,本文选取轮廓系数作为改进 PSO 优化算法的适应度函数。

2.3.2　模型设计

IPSO 算法的搜索空间维度设置为二维,适应度函数为轮廓系数,模型具体框架见图 2。

算法大致流程为:①将原始监测数据经预处理后传入 DBSCAN 聚类算法;②初始化 IPSO 参数,将 DBSCAN 的两个输入参数作为 IPSO 算法

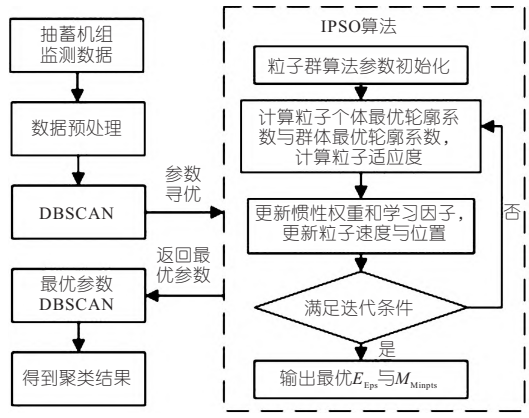

图 2 IPSO-DBSCAN 异常检测算法模型

Fig. 2 IPSO-DBSCAN anomaly detection algorithm model

的优化对象；③计算粒子群中每个粒子的轮廓系数作为适应度并更新个体最优值与群体最优值；④更新惯性权重和学习因子，更新粒子速度与位置；⑤若达到迭代次数，返回最优参数的取值，否则返回步骤③继续迭代；⑥根据最优参数构建最优 DBSCAN 聚类模型；对监测数据进行聚类，得到聚类标签。

3 实例分析

3.1 数据样本

利用某抽水蓄能机组状态监测数据验证所提方法异常值检测效果。分析数据取自 2020 年 2 月 1 日～2020 年 3 月 31 日之间的实测数据，采样频率为 1 min/次，共计 60 d 86 400 条数据。为避免机组启停时暂态数据的影响，提取出原始数据中的稳态值，共计 15 356 条数据。获取稳态数据中的有功功率(P)、导叶开度(α)、下机架振动(V)构成三维数据集(图 3)，机组实测数据中包含明显的异常值。

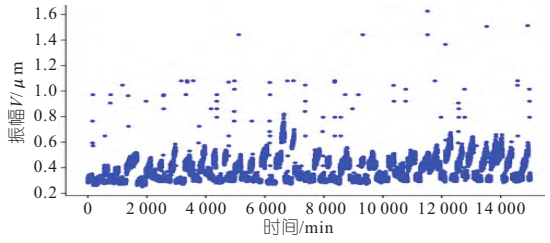

图 3 抽蓄机组实测下机架振动稳态值

Fig. 3 Steady-state vibration value of frame lower of pumped storage unit measured

图 4 为机组有功功率—导叶开度—下机架振动监测信号。由图 4 可看出，机组监测数据主要分布在三维空间的两个区域，数据中存在明显离群样本。

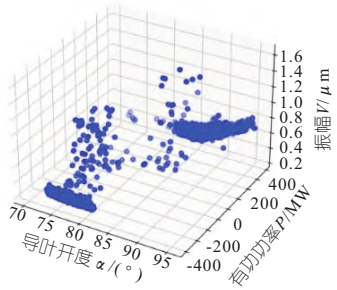

图 4 机组有功功率—导叶开度—下机架振动监测信号

Fig. 4 Unit active power-guide vane opening-frame vibration monitoring signal

3.2 DBSCAN 聚类效果

DBSCAN 聚类效果取决于 E_{Eps} 与 M_{Minpts} 的选取。试验中，按照 K 距离曲线设定 E_{Eps} 值，并根据试验结果调整，最终设置 $E_{Eps}=5$、$M_{Minpts}=6$，以此对机组导叶开度—有功功率—下机架振动进行异常数据筛选，结果见图 5。图 5(a)中，红色数据点表示检测出的异常数据。由图 5(a)可看出，共计 164 个异常点，异常值占比为 1.07%，轮廓系数得分为 0.930 4。图 5(b)中，蓝色、黄色数据点分别代表抽蓄机组抽水、发电两种工况数据。由图 5(b)可看出，人工选取参数存在一定误差，聚类簇中仍存在明显离群点未被剔除。

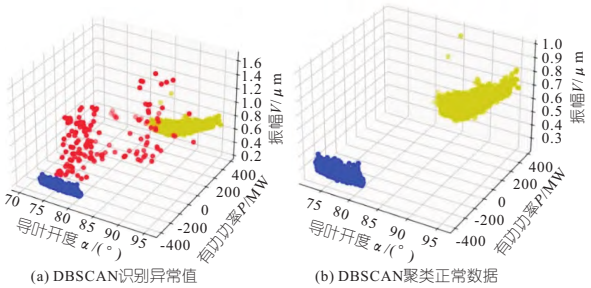

图 5 DBSCAN 算法结果

Fig. 5 DBSCAN algorithm result

3.3 基于 IPSO-DBSCAN 的抽水蓄能机组状态监测数据异常值检测

为降低 DBSCAN 算法参数选取对聚类结果造成的不利影响，采用 IPSO 算法对 DBSCAN 的参数进行自动寻优，以求得到最优聚类效果，从而提高异常检测的可靠性。

IPSO 算法的初始参数设置为粒子维度为 2；粒子数量为 30；最大迭代次数为 50；惯性权重 ω 初始值为 1.6，终值为 0.8；学习因子 c_1、c_2 初始值分别为 2.5、1，终值分别为 0.5、2.25；随机数 r_1、r_2 在区间[0,1]内均匀配置。

最终选取最优 E_{Eps} 值为 2.01，最优 M_{Minpts} 值为 8.25。图 6(a)为 IPSO-DBSCAN 对试验数据的异常值检测结果。由图 6(a)可看出，共计

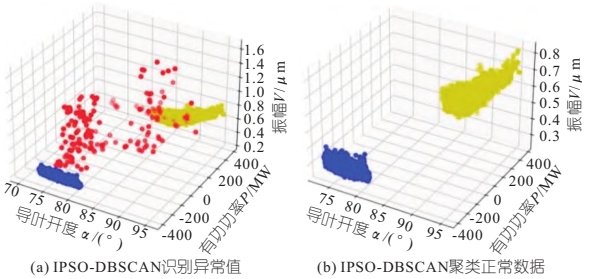

图 6　IPSO-DBSCAN 检测结果

Fig. 6　IPSO-DBSCAN detection results

178 个异常点,异常值占比为 1.16%,最大轮廓系数为 0.977 6。可见 IPSO-DBSCAN 可以有效适应全工况下的抽水蓄能机组数据并识别出样本中的异常点。

相比于图 5(b),图 6(b)所示 IPSO-DBSCAN 筛选后的正常聚类簇中未包含明显离群样本,同时,检测结果的异常值占比及轮廓系数得分均有所提升。

3.4　算法对比分析

3.4.1　对比试验

为验证 IPSO-DBSCAN 算法在异常检测方面的优势,采用 PSO-DBSCAN、Kmeans、IF 和 RRCF 作为对比。对比试验结果见图 7。

PSO-DBSCAN 算法参数设置为 $\omega=1.5$,c_1、c_2 均为 2,其他参数与 IPSO-DBSCAN 设置一致,算法识别异常值 167 个,轮廓系数得分为 0.949 2,异常值占比为 1.08%,结果见图 7(a)、(b)。Kmeans 算法设置类簇中心为 7 时效果最好,算法识别异常值个数为 149 个,轮廓系数得分为 0.933 7,异常值占比为 0.97%,结果见图 7(c)、(d)。IF 识别异常值个数 285,异常值占比为 1.86%,结果见图 7(e)、(f)。RRCF 算法识别异常数据个数为 326,异常值占比为 2.12%,结果见图 7(g)、(h)。可见,不同算法识别出的正常数据中均包含部分明显离群样本。

3.4.2　评价指标

本文所使用的数据为机组正常运行时采集数据,没有人工添加的异常值标签,轮廓系数虽然能够在缺乏标签的情况下对聚类结果进行度量,但其适用范围仅限于聚类算法,在异常值检测结果评估方面不具通用性。为进一步定量分析异常值检测效果的优劣,本文引入相关系数作为评价指标进行对比。

相关系数是变量之间相关程度的定量表达。抽蓄机组的振动信号与其导叶开度和有功功率之间有很强的相关性,异常值的存在会影响参数间的相关关系,因此采用相关系数作为评价指标判

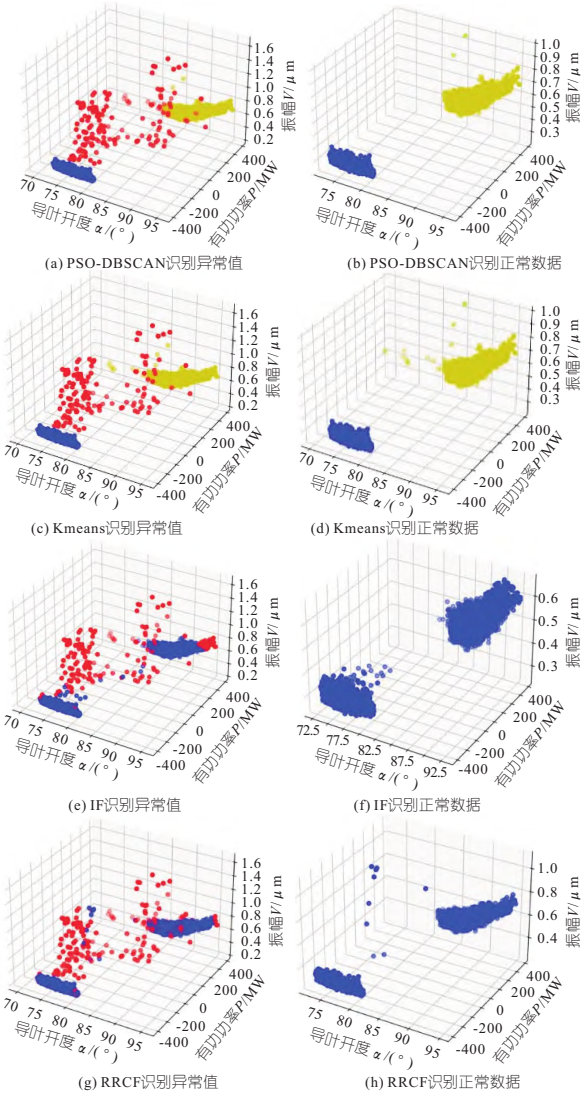

图 7　对比试验结果

Fig. 7　Compare experimental results

定异常值检测的效果,剔除异常值后相关性越强代表检测的效果越好,本文采用 pearson、kendall 两种相关系数作为评价指标。

根据评价指标对五种异常值检测方法结果进行评价,结果见表 1。

表 1　5 种方法检测结果及评价指标

Tab. 1　Detection results and evaluation indexes of five methods

检测方法	检测结果		pearson		kendall	
	异常值个数	占比/%	导叶开度	有功功率	导叶开度	有功功率
原始数据			0.792 8	0.743 9	0.600 3	0.454 2
DBSCAN	164	1.07	0.899 3	0.840 8	0.608 3	0.455 6
PSO-DBSCAN	167	1.08	0.899 4	0.840 8	0.609 5	0.456 0
Kmeans	149	0.97	0.897 3	0.840 8	0.608 3	0.455 6
IF	285	1.86	0.891 2	0.833 3	0.601 9	0.453 8
RRCF	326	2.12	0.890 1	0.834 3	0.608 7	0.456 0
IPSO-DBSCAN	178	1.16	**0.901 0**	**0.842 0**	**0.609 7**	**0.456 8**

注:加粗数据代表最大相关系数值。

由表 1 可知,以相关系数作为评价指标可以有效反映异常值检测的效果,在评价结果中,使用 IPSO-DBSCAN 方法剔除异常值后的数据在两种相关系数的评价中相较于对比试验中效果最好的

PSO-DBSCAN 分别提升了 1.6‰、1.2‰、0.2‰、0.8‰，进一步说明所提出的 IPSO-DBSCAN 算法优于其他几种常用异常值检测方法。

4 结论

a. 针对抽水蓄能机组状态监测数据中蕴含的异常值，提出一种基于 IPSO-DBSCAN 的异常值检测方法，该方法避开了 DBSCAN 算法人工调节参数导致检测精度下降的问题，并采用某抽水蓄能机组实测有功功率－导叶开度－下机架振动数据进行验证，结果表明所提方法相较于传统 DBSCAN 算法具有更好的检测性能。

b. 为验证模型的异常值检测性能，将 IPSO-DBSCAN 模型与 PSO-DBSCAN、Kmeans、IF 和 RRCF 模型对比，以数据间的相关系数作为评价指标，证明了 IPSO-DBSCAN 模型在抽蓄机组异常值检测上具有更好的识别效果。

参考文献：

[1] YAHIA Z, PRADHAN A. Simultaneous and sequential stochastic optimization approaches for pumped storage plant scheduling with random breakdowns[J]. Energy, 2020, 204: 117896.

[2] 朱溪，李超顺，邹雯，等. 一种水电机组振动趋势预测方法和系统：CN112651290A[P]. 2021-04-13.

[3] 徐搏超. 基于参数关联性的电站参数异常点清洗方法[J]. 电力系统自化，2020，44(20)：142-147.

[4] 万尹豪. 水资源数据清洗整编算法的研究与实现[D]. 南昌：南昌大学，2022.

[5] 蔡思宇，刘庆涛，孙龙，等. 一种数据异常动态识别与多模式自匹配的数据清洗技术：CN112286924A[P]. 2021-01-29.

[6] 金容鑫，娄岱松，黄华德，等. 水电机组状态监测数据清洗方法[J]. 中国农村水利水电，2022(7)：187-192.

[7] 闫亚男，韩长霖，陈小松，等. 融合时间序列特性的水电遥测数据清洗优化框架[J]. 水力发电，2021，47(11)：79-83，95.

[8] 段然，周建中，蔡银辉，等. 基于低质量数据的水电机组变工况状态指标构建方法[J]. 水电能源科学，2022，40(6)：183-187.

Abnormal Detection Method of Condition Monitoring Data of Pumped Storage Unit Based on IPSO-DBSCAN

ZHANG Jin-peng, ZHANG Xiao-yuan

(College of Electrical Engineering, Henan University of Technology, Zhengzhou 450001, China)

Abstract: The status monitoring data of pumped storage unit is affected by factors such as collection equipment failures and communication equipment abnormalities, and there are some abnormal data in the dataset, which has a negative impact on the subsequent assessment and prediction of the health status of the units. Therefore, a unit anomaly data detection model based on improved particle swarm optimization algorithm and DBSCAN density clustering algorithm was proposed. The model improves the particle swarm optimization algorithm to address the problem of easily falling into local optima, and then introduced contour coefficient as fitness function to optimize the parameters of DBSCAN. Finally, the correlation coefficient was used to evaluate the effect of eliminating outliers. The measured guide vane opening, active power and lower frame vibration data of a domestic pumped storage unit from early February to late March 2020 were used for example analysis. The results show that the proposed method can effectively detect the abnormal data of vibration monitoring, and the correlation coefficient between the data after removing the outlier is improved, which lays a data foundation for the subsequent unit health status assessment and prediction.

Key words: pumped storage; outlier detection; improved particle swarm optimization algorithm; DBSCAN

(上接第 107 页)

Statistics Characteristics and Correlation Distribution Model of Material Parameters of Low Elastic Modulus Concrete Cut-off Wall

HE Jin-wen[1a,1b], JIA Shu[1a,1b], PAN Chun-ling[2], CHEN Zhao[1a,1b], ZHANG Shi-yao[1a,1b]

(1a. Hubei Key Laboratory of Hydropower Engineering Construction and Management;
1b. College of Hydraulic&Environmental Engineering, China Three Gorges University, Yichang 443002, China;
2. Fujian Research Institute of Water Conservancy and Hydropower, Fuzhou 350001, China)

Abstract: In order to obtain the statistics characteristics and correlation distribution models of permeability coefficient, elastic modulus (E-modulus) and compressive strength of concrete cutoff wall with low E-modulus, based on the test data of 2 earth-rockfill dams, AIC method was used to identify the optimum distribution of material parameters and best Copula functions for correlation distribution model. The statistical uncertainty was simulated by Bootstrap method. The results show that compressive strength, permeability coefficient and elastic modulus obey Extreme type I distribution, Lognormal distribution and Weibull distribution or Extreme type I distribution, respectively, with coefficient of variation being 0.17, 0.40 and 0.11. The average Pearson coefficient of correlation between compressive strength and permeability coefficient, E-modulus and compressive strength, E-modulus and permeability coefficient are −0.79, 0.67 and −0.55, and the best Copula are Plackett Copula, Gaussian Copula and Gaussian Copula, respectively. The results can provide reference for seepage and strength reliability calculation of low E-modulus concrete cutoff wall.

Key words: low E-modulus concrete; material parameter; statistics characteristics; correlation distribution; compressive strength; permeability coefficient; E-modulus