DOI:10.20040/j.cnki.1000-7709.2023.20230577

基于 IF-GEP 的河湾最大冲刷深度预测方法

陈骏峰,肖丽蓉,周晓泉,黄宇航

(四川大学水力学及山区河流开发保护国家重点实验室,四川 成都 610065)

摘要: 为解决传统河流流经弯道的最大冲刷深度预测过程中存在的不足,将孤立森林(IF)和基因表达式编程(GEP)方法相结合,建立了一个基于 IF 的 GEP 河湾最大冲刷深度预测模型(IF-GEP),并将该模型与传统 GS-SVR 和 RF 模型及现有经验公式进行对比。结果表明,IF-GEP 预测模型在测试集上取得了较好的预测效果,且预测精度明显高于现有公式及传统的 GS-SVR 和 RF 模型。最后将该预测模型应用于多条不同河流的预测中,IF-GEP 预测模型的预测结果与实际测量值较吻合,说明该预测模型具有良好的预测能力和较高的泛化性能。

关键词: 河湾最大冲刷深度;孤立森林;基因表达式编程;GS-SVR;RF

中图分类号: [TV91] **文献标志码:** A **文章编号:** 1000-7709(2023)09-0019-04

1 引言

当河流流经弯道时,由于离心力等作用,水流会产生螺旋运动和水面横比降,对凹岸河床的冲刷作用可能引起河岸不稳定、滑坡等,甚至会导致河堤倒塌,造成巨大的损失,因此研究河流弯道最大冲刷深度,提前做好保护工作具有十分重大的意义[1]。目前关于河湾最大冲刷深度方面的研究已较多,如王木兰等[2]根据上游来水为清水在弯道启动流速公式的基础上求得急流下弯道冲刷深度公式;CHATLEY H 等[3-5]分别提出了三个最大冲刷深度公式,且其选择的特征均为曲率半径、河床宽度及平均冲刷深度。但由于其复杂的作用机理和众多的影响因素,各公式均基于经验,精度和适用性有待进一步探究。对于这种复杂的非线性"黑箱"关系式,使用机器学习方法进行研究是一种普遍而可行的途径。孤立森林(IF)[6]为随机森林算法的一种改良算法,是一种具有高精准度、无监督等优点的离群点检测算法。与其他离群点检测算法相比,孤立森林具有简单高效、无需依赖样本分布及能够检测多维度的离群点等优点。基因表达式编程(GEP)是一种结合了遗传编程(GP)和遗传算法(GA)优点的算法,其运行速度比 GA 和 GP 提高了数倍,可应用于逻辑回归、函数发现和时间序列等多个领域。IF 和 GEP 算法已在土木水利领域已有了一些应用,并取得了较好的效果[7,8]。为此,本文将孤立森林(IF)与基因表达式编程(GEP)相结合,建立了一个基于 IF 的 GEP 河湾最大冲刷深度预测模型(IF-GEP),并将该模型应用于河湾最大冲刷深度的预测中,取得了良好的预测效果,可为河湾最大冲刷深度的预测及凹岸的防护设计提供一定的参考。

2 基于 IF 的 GEP 河湾最大冲刷深度预测模型

河湾最大冲刷深度受河湾曲率半径 R_c、水面宽度 B 等因素影响,每个因素对最大冲刷深度的影响权重也不同,根据实测数据所提出的经验公式作为参考(表 1),这些公式包含了曲率半径 R_c、水面宽度 B 和平均冲刷深度 $\bar{d}$ 三个因素,对这种机理尚不明确的非线性关系式,运用 GEP 方法预测是一种较好的方式,其预测精度往往比人工拟合的经验关系式更加可靠和适用。

表 1 不同的预测公式

Tab. 1 Different prediction formulas

模型	预测公式
文献[3]	$(d_{max}/\bar{d}) = 1 + 2B/R_c$
文献[4]	$(d_{max}/\bar{d}) = 2.07 - 0.19\ln(R_c/B - 2)$
文献[5]	$(d_{max}/\bar{d}) = 3.37 - 0.66\ln(R_c/B)$

收稿日期: 2023-03-12,**修回日期:** 2023-04-28

作者简介: 陈骏峰(1999-),男,硕士研究生,研究方向为水力学及河流动力学,E-mail:1044128575@qq.com

通讯作者: 周晓泉(1966-),男,博士、副研究员,研究方向为水力学及河流动力学,E-mail:xiaoquan_zhou@126.com

2.1 孤立森林(IF)算法

样本数据源自文献[5]中的 223 组实测数据。由于在测量过程中不可避免的会产生一些离群点,为减少这些离群点对后续建模的影响,在使用这些数据前需对其进行处理,使用孤立森林的方法去除离群点。

孤立森林算法不借助密度等指标去描述样本点之间的差异,而是利用二叉树随机分割数据集来判断离群点,其具体工作原理为:①随机选择样本子集作为根节点;②将样本不断地进行二分形成叶子节点,直到每个叶子节点只有一个样本或达到了预设的树高度,重复此过程,形成随机森林;③计算每个样本的平均路径长度及异常分数;④根据异常分数判断每个样本是否为离群点。路径长度和异常分数计算公式为:

$$c(m)=2H(m-1)-2(m-1)/m \tag{1}$$

$$s(x,m)=2^{-\frac{E(h(x))}{c(m)}} \tag{2}$$

式中,$c(m)$ 为搜索路径函数平均值;$H(k)$ 为调和数,表示从 1～k 的倒数之和;$s(x,m)$ 为计算异常分数;$E(h(x))$ 为样本 x 的路径长度的期望值。

基于孤立森林方法,将 223 组实测数据减少为 180 组不包含离群点或离群点影响较小的数据。

2.2 基因表达式编程(GEP)

在建立模型前需先选择适当的适应度函数以确定个体的选择概率,并据此计算个体的适应度值。经典的 GEP 算法有三种适应度计算函数,即绝对误差的适应函数、基于相对误差的适应函数和用于逻辑合成问题的适应度函数。其中,均方根误差(R_{RMSE})是一种具有较高的稳健性的适应度函数,适用于连续变量的优化问题。采用 R_{RMSE} 函数作为适应度函数。

先对数据进行量纲处理,参照已有公式的处理方法,将 R_c/B、$B/\bar{d}$ 和 $\sqrt{v^2/(gB)}$ 作为输入量,$d_{\max}/\bar{d}$ 作为输出量代入 GEP 模型,其中 144 组数据作为训练集,36 组数据作为测试集。接着进行参数的设置,在 GEP 中需设置的参数包括染色体个数、基因个数、基因头部长度、基因间连接函数等,不同的参数对预测结果的精度有直接影响,因此选择合适的参数至关重要。设置不同的参数组合后进行迭代计算,最终确定最优参数。其计算方法就是先随机生成一定数量的种群个体,计算个体中染色体上基因的适应值,保留最优个体,对其他个体进行选择,原则是适应值大的个体相比适应值小的个体被选中的概率大,接着对被选中的个体进行复制、变异、重组、转座等一系列操作由此形成新的个体;评估新个体的适应值,重复上述过程,进行多次进化迭代,从而找到较为优良的种群。最优参数见表 2。

表 2 遗传参数设置

Tab. 2 Genetic parameter setting

参数	取值	参数	取值
染色体个数	1 000	RIS 转座率	0.005
基因个数	2	基因转座率	0.003
基因头部长度	5	单点重组率	0.003
连接函数	Addition	两点重组率	0.003
变异率	0.002	基因重组率	0.003
IS 转座率	0.005	进化次数	500

通过计算可得到最大冲刷深度的预测函数:

$$\frac{d_{\max}}{\bar{d}}=2\sqrt{\frac{v^2}{gB}}+1.34+\frac{B/\bar{d}}{13.23R_c/B+56.27} \tag{3}$$

其表达式树见图 1,其中 d_0、d_1、d_2 为输入值,即 R_c/B、$B/\bar{d}$ 和 $\sqrt{v^2/(gB)}$,c_0、c_1、c_2、c_3、c_4、c_5 均为常数,其值分别为 3.141 5、−4.615 8、6.711 1、2.034 1、−3.637 0、−4.254 0。

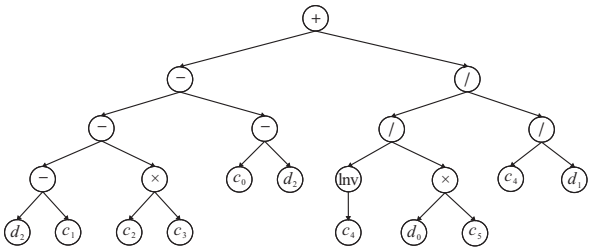

图 1 GEP 模型表达式树

Fig. 1 GEP model expression tree

2.3 IF-GEP 计算流程

采用孤立森林的方法去除离群点,然后将经过孤立森林得到的数据代入 GEP 算法中建立 GEP 模型,从而预测河湾最大冲刷深度,其计算流程见图 2。

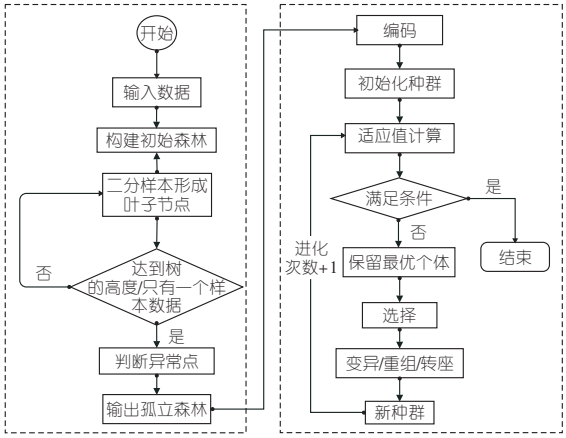

图 2 IF-GEP 计算流程

Fig. 2 Flow chart of IF-GEP calculation

3 结果分析

3.1　模型结果分析

采用均方根误差、平均方差(M_{MSE})、相对平方根误差(R_{RRSE})及决定系数 R^2 4 个指标对模型的训练结果进行评估见表 3。由表 3 可知,IF-GEP 模型训练集的 R^2 达到了 0.913 0,均方根误差、平均方差、相对平方根误差的值分别为 1.155 7、1.335 7、0.295 0,均较小。为分析该预测模型的优劣性,建立了基于网格搜索法寻优的支持向量回归(SVR)模型和随机森林(RF)模型与之进行对比,并与表 1 所列 3 个现有经验公式进行比较,利用测试集数据对不同模型和公式的预测结果进行评估。由于文献[4]公式的局限性,36 组测试集数据中有一组数据不适用于文献[4]公式,因此使用剩下 35 组数据进行分析,结果见表 4。

表 3　IF-GEP 模型训练结果分析

Tab. 3　Analysis of IF-GEP model training results

项目	R_{RMSE}	M_{MSE}	R_{RRSE}	R^2
训练集	1.155 7	1.335 7	0.295 0	0.913 0
测试集	0.908 0	0.824 4	0.287 0	0.917 6

表 4　IF-GEP 模型与现有公式及其他模型对比分析

Tab. 4　Comparative analysis of IF-GEP model with existing formulas and other models

模型	R_{RMSE}	M_{MSE}	R_{RRSE}	R^2
文献[3]	2.850 3	8.124 1	0.961 9	0.074 8
文献[4]	1.390 4	1.933 1	0.469 2	0.779 8
文献[5]	2.175 7	4.733 5	0.734 2	0.460 9
GS-SVR 模型	0.993 5	0.987 1	0.335 3	0.887 6
RF 模型	1.112 2	1.236 9	0.375 3	0.859 1
IF-GEP 模型	0.919 0	0.844 6	0.310 1	0.903 8

由表 4 可看出,现有公式中,文献[4]公式的预测精度最高,文献[3]公式最低,三个公式的精度均低于 IF-GEP 模型,且文献[4]公式具有一定的局限性,需满足 $R_c/B>2$ 的条件;RF 模型和 GS-SVR 模型的预测精度均较高,说明用机器学习的方法来预测河湾最大冲刷深度这种复杂的非线性关系具有较好的效果。IF-GEP 模型的 R_{RMSE}、M_{MSE}、R_{RRSE} 值在与现有公式及 RF、GS-SVR 模型的对比中均最低,R^2 值最高,达到了 0.903 8,而其他三个公式及 RF、GS-SVR 模型的 R^2,能够比较明显地看出 IF-GEP 模型的预测结果优于其他三个公式及 RF 和 GS-SVR 模型,说明 IF-GEP 模型在训练集和测试集包含的这些河流中对河湾最大冲刷深度的预测结果较合理可靠,与现有经验公式及 GS-SVR 模型比较优势较大,略优于 RF 模型。

3.2　应用验证

为验证 IF-GEP 模型在其他河流中的适用性,用文献[5]中的 15 组训练集和测试集数据,包括威尔士 Neath、英国德文郡 Yscir 等多条不同河流,具体统计参数见表 5。采用 IF-GEP 模型及 GS-SVR、RF 模型和其他三种现有经验公式对这 15 组河流的最大冲刷深度数据进行预测,并将预测值与实际值进行对比,结果见图 3。其中 1 组数据由于不适用文献[4]公式使用实测值代替。由表 5、图 3 可看出,在这 15 组实测数据中 IF-GEP 模型的预测结果与实测值相近,表现最好;其他公式和模型中表现最好的是 RF 模型,略差于 IF-GEP 模型。GS-SVR 模型虽然测试集中表现良好,但在不同河流的适用性上与 IF-GEP 模型还有着一定的差距。IF-GEP 模型几乎每个样本的预测值相比其他公式和模型均更接近实测值,这说明 IF-GEP 模型在这些河流的河湾最大冲刷深度预测中适用性较好。

表 5　河流实测数据统计

Tab. 5　Statistics of river measured data

河流名称	数量	R_c/m	B/m	$\bar{d}$/m	v	d_{max}/m
Neath	1	91.33	30.9	2.46	2.26	3.41
Yscir	1	74.96	18.6	1.81	1.34	2.89
Ccidog	2	72.43	18.3	1.35	1.94	2.18
		72.43	14.8	1.17	2.77	1.89
Tweed(Peebles)	1	179.55	31.6	2.26	2.15	2.99
Mint	1	196.56	18.4	1.41	0.88	2.18
Glendaramackin	2	94.50	14.4	1.27	2.46	2.78
		110.07	17.0	1.12	2.36	2.15
Roading	1	21.00	12.0	1.30	1.13	2.22
South Esk	1	67.01	23.0	1.22	0.48	1.89
Seven	1	44.00	9.1	0.87	1.35	1.30
Wylie	2	66.94	9.7	0.87	0.84	1.35
		66.94	9.8	0.93	0.78	1.25
Croasedale Beck	1	116.24	14.6	1.04	3.49	1.83
Exe	1	97.55	32.4	1.64	2.90	2.88

注:表中 v 单位为 m/s。

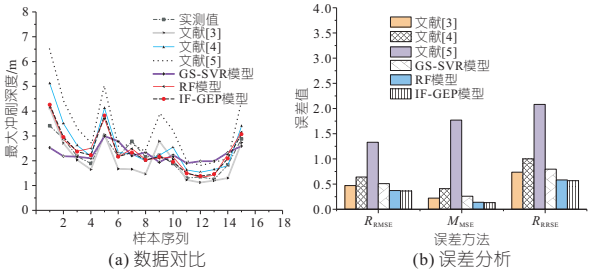

图 3　河流最大冲刷深度公式对比图

Fig. 3　Comparison of formulas for maximum river scour depth

4　结论

a. 应用孤立森林算法处理原始数据,将得到的数据经过量纲处理后代入 GEP 模型中进行训练和优化,建立了一个基于 IF 的 GEP 河湾最大冲刷深度的预测模型,并将该模型与 GS-SVR、RF 模型及现有经验公式进行对比。结果表明,

GS-SVR、RF 和 IF-GEP 模型的表现明显优于现有公式，IF-GEP 模型的表现明显优于 RF 模型，略优于 GS-SVR 模型，说明机器学习的方法在处理这种复杂的非线性问题中有着较好的效果，且 IF-GEP 模型的预测效果相比 GS-SVR 和 RF 模型在这些河流中更适用且预测效果更好，精度更高。

b. 将 IF-GEP 模型应用于训练集和测试集以外的多条不同河流，IF-GEP 模型对于河湾最大冲刷深度的预测值与实测值相差较小，且表现明显优于 GS-SVR 和现有公式，略优于 RF 模型。说明 IF-GEP 模型在这些河流中也表现出了良好的预测效果，具有一定的推广价值，可为河湾最大冲刷深度的预测及凹岸的防护设计提供一定的参考。

参考文献：

[1] 张良然. 河渠缓流弯道冲刷深度计算的探讨[J]. 南昌大学学报(工科版)，2001，23(4)：84-86，106.

[2] 王木兰，汪德爟. 明渠弯道水流与冲刷问题[J]. 河海大学学报(自然科学版)，1978(1)：117-129.

[3] CHATLEY H. Curvature effects in open channels [J]. Engineering，London，England，1931，131.

[4] THORNE C R，ABT S R. Velocity and scour prediction in river bends[R]. Army Engineer Waterways Experiment Station Vicksburg MS Hydraulics Lab，1993.

[5] US Army Corps of Engineers. Hydraulic design of flood control channels[M]. Washington，DC：Government Printing Office，1994.

[6] FEI T L，KAIM T，ZHOU Z H. Isolation forest [C]// IEEE International Conference on Data Mining IEEE，2008.

[7] 赵新华，范振东，何宇，等. 基于数据重构与孤立森林法的大坝自动化监测数据异常检测方法[J]. 中国农村水利水电，2021(9)：174-178.

[8] 李超群，魏清顺. 采用基因表达式编程的潜水泵性能预测[J]. 水电能源科学，2020，38(4)：150-153，180.

Maximum Erosion Depth Prediction of River Bend Based on IF-GEP

CHEN Jun-feng，XIAO Li-rong，ZHOU Xiao-quan，HUANG Yu-hang

(State Key Laboratory of Hydraulics and Mountain River Engineering，Sichuan University，Chengdu 610065，China)

Abstract：In order to address the limitations in forecasting the maximum scour depth of conventional river bends，this study amalgamated the methodologies of isolated forest (IF) and gene expression programming (GEP). An IF-GEP model for predicting the maximum scour depth of river bends was established. The validation results demonstrate that the IF-GEP prediction model surpasses existing formulations in terms of its accuracy on the test set. Moreover，it exhibits enhanced predictive performance compared to the traditional GS-SVR and RF models. Application of the prediction model to various rivers yielded remarkably close results to the actual measured values，affirming its strong predictive capability and robust generalization performance.

Key words：maximum scour depth of river bend；isolated forest；gene expression programming；GS-SVR；RF

(上接第 43 页)

[10] WHIPPLE A A，VIERS J H. Coupling landscapes and river flows to restore highly modified rivers [J]. Water resources research，2019，55(6)：4512-4532.

[11] BIRON P M，CARVER R B，CARRÉ D M. Sediment transport and flow dynamics around a restored pool in a fish habitat rehabilitation project：Field and 3D numerical modelling experiments[J]. River research and applications，2012，28(7)：926-939.

Numerical Simulation of Effect of Channel Morphology Reconfiguration on Migration and Diffusion of Pollutants in Urban Stream

DENG Lin-yue[1a]，TANG Jie[1a]，CHEN Yao[1a,1b]，LIU Fei[1a,1b]，HOU Yi-zhi[1a]，GAN Chun-juan[2]，TAN Yu-qing[1a]

(1a. School of River and Ocean Engineering；1b. Engineering Laboratory of Environmental Hydraulic Engineering of Chongqing Municipal Development and Reform Commission，Chongqing Jiaotong University，Chongqing 400074，China；2. Chongqing Municipal Research Institute of Design Ltd. Co.，Chongqing 400012，China)

Abstract：A typical urban channelized (Urban) channel in Liangtan River，Chongqing，was reshaped into six types of channel by a numerical simulation tool named RiverBuilder. And then a two-dimensional hydrodynamic convection-diffusion model was constructed to study the effect of channel morphology reconfiguration on migration and diffusion index，such as water turn-over time (T_{TOT})，pollutant concentration curve (C_{CC})，pollutant reaching maximum time (M_{MT}) and pollutant arriving time (A_{AT}). The results show that the channel morphology reconfigured by changing the width (W_{bf})，depth (D_{bf})，and meandering (M_d) of the Urban channel can inhibit the migration and diffusion of pollutants to a certain extent，but the influencing effect is not as good as that of the composite channel based on the variable D_{bf}. Meanwhile，the near-natural (Natural) channel has the strongest anti-pollute capacity and inhibiting ability of pollutants diffusion，indicating it is more suitable for the self-purification process of pollutants. It is confirmed that the channel morphology based on disordered and complex changes in W_{bf}，D_{bf}，and M_d is closer to the Natural channel，which can provide good eco-hydraulic conditions for the improvement of river water quality.

Key words：channel morphology；hydrodynamic model；pollutants；migration and diffusion；geometric variables；eco-hydraulics