基于随机森林算法的推移质输沙率计算研究

江雨润， 黄 尔
（四川大学水力学与山区河流开发保护国家重点实验室，四川 成都 610065）

摘 要：为探寻随机森林算法在预测推移质输沙率方面的效果，本文筛选了 3060 组天然河道输沙数据，将其划分为三种方案，得到了算法在单一流域、综合流域情况下的预测数据，并利用沙莫夫公式的计算结果作为对比。结果表明：水力参数与输沙率之间的相关性和参数在算法中的重要性并不完全一致；同一参数以不同形式出现在算法运算中时，会呈现不同的重要性；在足够的数据量支持下，随机森林算法相比于传统公式明显有更好的预测效果。

关键词：随机森林；机器学习；推移质；输沙率
中图分类号：C32；TV142.2；TV148+.7 **文献标识码**： A **文章编号**：1001-2184(2023)02-0001-06

Study on Bed-load Transport Rate Calculation Based on Random Forest Algorithm

JIANG Yurun, HUANG Er
(State Key Laboratory of Hydraulics and Mountain River Engineering, Sichuan University, Chengdu Sichuan 610065)

Abstract: In order to explore the effect of random forest algorithm in predicting the bed-load transport rate, 3060 sets of bed-load transport data of natural rivers were screened and divided into three schemes to obtain the prediction data of the algorithm under the condition of single watershed and comprehensive watershed. Compared with the calculation results of Shamov formula, it shows that the correlation between hydraulic parameters and bed-load transport rate is not completely consistent with the importance of the parameters in the algorithm. When the same parameter appears in different forms in the algorithm, it will show different importance. With sufficient data, the random forest algorithm has better prediction than the traditional formula.

Key words: random forest; machine learning; bed-load; transport rate

0 引 言

推移质输移，是砾石河床河流形态变化的主要驱动因素之一，对于大坝设计、流域管理、防洪、供水等水利工程非常重要。由于河床脉动、床沙组成、上游来沙条件等因素的不确定性，推移质输沙率的预测一直以来都是难点[1]。工程上应用较广的沙莫夫公式以平均流速作为决定推移质运动的主要参数[2]；刘兴年以颗粒暴露高度作为区分宽级配非均匀砂全部可动和部分可动的标准，并将此结果应用于起动概率与输沙率计算中[3]。

基于机器学习和数据挖掘技术的人工智能算法主要作为黑箱型非线性统计模型运行，能够满足较为复杂的泥沙输移预测的要求[4]。Khabat Khosravi 等采用 BA 数据挖掘算法训练了四种混合算法预测水槽推移质输沙率，Vasileios Kitsikoudis 等[5]使用了三种数据驱动技术预测了爱达荷州山区河流的推移质输沙率；国内陈雄波等[6]研究了基于神经网络和遗传算法的泥沙模型，谢世博等利用 LSTM 神经网络对卵石运动进行了轨迹预测。相较于人工神经网络，随机森林算法可以对特征变量进行评分，更好地研究变量[7]，且目前国内中文文献中暂未发现随机森林算法在推移质输沙率方面的研究，无可参考的应用实例。本文以天然河道水文泥沙数据作为研究对象，利用随机森林算法研究输沙特性及算法的应用效果。

收稿日期：2022-12-05
基金项目：第二次青藏科考，水系固体物质源－汇过程与演变（编号：2019QZKK0204）

1 研究对象及方法

1.1 研究对象

本文的数据来自 Darren Hinton 等[8]收集的 15 000 多组天然河道推移质输移数据，数据库条目齐全，包括样本描述、流量、输沙率、河道及河岸特征、粒径分布及其他属性。每一条数据均是通过已发表的文献、作者回应等途径进行汇编，并根据原始数据进行二次精度检查、转换为统一单位。

由于原数据库的输沙资料众多且精度不一，未能完全适用于本次研究，故按照以下标准对资料进行筛选：

(1)水力参数条目的完整性。本次研究筛除了大部分水力参数不齐全的河道。

(2)输沙数据的完整性。本次研究所选取的对象河道均有不少于两次洪水的数据量，且保留了输沙强度较低的数据。

基于以上标准，本文拟采用的资料包含 3 060 组天然河流输沙数据，其中圣路易斯河(St. Louis Creek)流域占 937 组。本文根据流域的单一性与综合性、水力参数是否有量纲划分了三种方案，以满足后续效果探索的需求，包含的数据组成分别为：

(1)圣路易斯河流域数据(水力参数由常见的有量纲及无量纲参数组成，且输沙率以有量纲形式呈现)，具体参数为：

流域面积 $F(\mathrm{km}^2)$，河道平均流量 $Q(\mathrm{m}^3/\mathrm{s})$，沉速 $\omega(\mathrm{m/s})$，$\omega=-9\frac{v}{D}+\sqrt{(9\frac{v}{D})^2+\frac{\gamma_s-\gamma}{\gamma}gD}$，$v$ 为运动粘度)，推移质中值粒径 $D_{50}(\mathrm{m})$，河道平均坡度 S，河道顶宽 $B(\mathrm{m})$，平均水深 $h(\mathrm{m})$，平均流速 $V(\mathrm{m/s})$，平均流量与平滩流量的比例关系 T(比值在 0.8 以下和 0.8 以上两种状态，用于验证部分论文中提到过的流量与平滩流量的比例关系对于输沙率的影响[9])，实测单位时间输沙率 $g_a(\mathrm{kg/s})$；

(2)圣路易斯河流域数据(水力参数均为无量纲形式)，具体参数为：

无量纲水流切应力 $\Theta=\frac{\tau}{(\rho_s-\rho)gD}$，无量纲水流功率 $W^*=\frac{\tau V}{\gamma_s D\sqrt{(\frac{\gamma_s}{\gamma}-1)gD}}$，平均坡度 S，无量纲流量 $q^*=\frac{q}{\sqrt{gD_{84}^3}}$，无量纲流速 $V^*=\frac{V}{\omega}$，无量纲粒径 $D_{gr}=D\left[g\frac{\gamma_s-\gamma}{\gamma v^2}\right]^{1/3}$，相对粗糙度 $RR=\frac{D}{h}$，雷诺数 $Re=\frac{Vh}{v}$，弗劳德数 $Fr=\frac{V}{\sqrt{gh}}$，平均流量与平滩流量的比例关系 T，爱因斯坦无量纲输沙率 $\Phi_a=\frac{\frac{g_a}{B}g}{\gamma_s D\sqrt{(\frac{\gamma_s}{\gamma}-1)gD}}$；

(3)综合流域数据，水力参数选取与方案一相同。

1.2　随机森林算法

随机森林(random forest，简称 RF)是一种高度灵活的机器学习算法，应用广泛且准确率高。其通过集成学习的思想将多棵决策树集成为一种算法。在构建决策树的过程中，均从训练数据中有放回的随机选取部分样本，并从样本中随机选取部分特征，最大限度地保证每棵树的独立性，以此找到最稳定可靠的结果。

本次研究中，训练集和测试集的划分方法采用留出法，将原数据集划分为两个互斥的集合，分别占 70%和 30%。这也意味着，上述训练步骤中的样本来源均为占总方案数据 70%的训练集；剩下 30%的测试集不参与训练，仅用于测试算法效果。单次使用留出法得到的结果往往不够稳定，故每种方案均进行三次计算(且每次的训练集与测试集都统一进行随机划分)，分别命名为序列 1、序列 2、序列 3，观察最终的计算结果。

本文中暂未对随机森林自身的参数进行过多调整，基本采用的是默认值。随机森林算法结构见图 1。

1.3　推移质输沙率公式

本次研究中，为探究随机森林算法与传统公式的效果差异，拟采用沙莫夫公式对圣路易斯河流域数据和综合流域数据进行计算，集中比较二者的性能。为保证数据来源的一致性，沙莫夫公式计算的数据均为随机森林算法中采用的测试集而非整个数据库。沙莫夫公式如下：

$$q_b=1.5D_0^{2/3}\left[\frac{v}{v_{off}}\right]^3(v-v_{off})\left(\frac{D}{h}\right)^{1/4}\quad(1)$$

式中　v_{off} 为止动流速，$v_{off}=\frac{1}{1.2}v_0=3.83\mathrm{D}^{1/3}\mathrm{h}^{1/6}$；

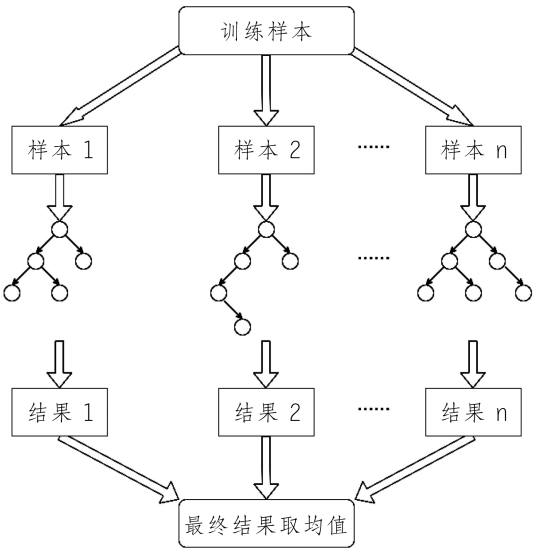

图 1　随机森林算法结构图

v_0 为起动流速；D_0 为非均匀砂中最粗一组的平均粒径。如这一组占沙样的 40%～70%，则公式系数为 3；如占 20%～40%或 70%～80%，系数等于 2.5；如占 10%～20%或 80%～90%，系数等于 1.5。公式中单位使用千克(kg)、米(m)、秒(s)。

1.4　模型评估

本次研究中，使用了 R^2(决定系数 Coefficient of Determination)、RMSE(均方根误差 Root Mean Square Error，亦称标准误差)、MAE(平均绝对误差 Mean Absolute Error)三种模型性能评估指标，各指标公式如下：

$$R^2=\left[\frac{\sum_{i=1}^{n}(X_o-\overline{X}_o)(X_e-\overline{X}_e)}{\sqrt{\sum_{i=1}^{n}(X_o-\overline{X}_o)^2\sum_{i=1}^{n}(X_e-\overline{X}_e)^2}}\right]^2 \tag{2}$$

$$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_e-X_o)^2} \tag{3}$$

$$MAE=\frac{1}{n}\sum_{i=1}^{n}|X_e-X_o| \tag{4}$$

式中　X_o 和 X_e 分别为实测值和预测值；$\overline{X}_o$ 和 $\overline{X}_e$ 分别为实测值和预测值的平均值；n 为数据点的数量。各项指标中，R^2、$RMSE$、MAE 的评估标准分别为：$R^2\geqslant 0.5$ 时，为较合适的结果，越接近 1 则表示效果越好；$RMSE$、MAE 均为越小越好。

2　计算结果及分析

2.1　水力参数相关性分析

本文将首先展示每种方案下水力参数与输沙率之间的相关系数 R(Correlation coefficient)，这并不需要划分数据集，因此，后面的相关性分析数据来源为三种方案下各自全部的数据。各方案下水力参数与推移质输沙率之间的相关性热力图见图 2。

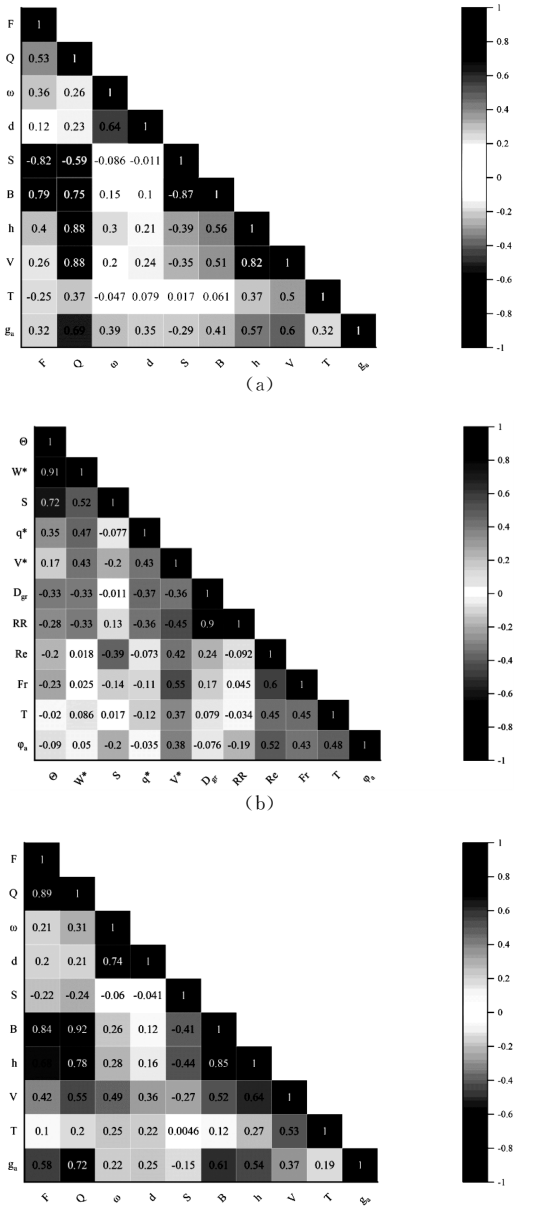

图 2　各方案下水力参数与推移质输沙率之间的相关性热力图

对比图 2(a)和图 2(c)可以看出，无论是综合流域还是单一流域，流量、流速、水深、河宽等参数与输沙率之间的相关性都较好，与前人的研究保

持一致，且均有其他参数呈正相关，唯独坡度呈负相关的现象。

针对圣路易斯河流域，图 2(b)中无量纲水力参数中的雷诺数是与爱因斯坦无量纲输沙率相关性最高的参数。相比于单一流域，在综合流域中，流域面积、流量、河道宽度等参数与输沙率的相关性显著增强，而沉速、粒径、坡度、流速与输沙率的相关性有所减弱。

2.2　随机森林算法的运行结果

后续的预测结果均为对应方案随机划分的测试集预测结果，不包含训练集，但输入参数重要性来自训练集，其原理在于未纳入计算的数据(袋外数据)中的某个特征(在此处为水力参数)加入随机噪声后的准确率会受到影响，影响越大则重要性越高。

2.2.1　方案一的运行结果

方案一中序列 1、序列 2、序列 3 各自的测试集运行结果与参数重要性排序见图 3，左侧图为运行结果，右侧图为参数重要性排序。决定系数见图中，其他评估指标见表 1。

表 1　随机森林算法与沙莫夫公式的性能评估表

评估指标	R^2	*RMSE* /kg·s^{-1}	*MAE* /kg·s^{-1}
方案一序列 1	0.780	0.033	0.020
方案一序列 2	0.800	0.042	0.022
方案一序列 3	0.780	0.034	0.020
方案二序列 1	0.527	0.002	0.002
方案二序列 2	0.577	0.002	0.002
方案二序列 3	0.476	0.002	0.002
方案三序列 1	0.668	1.170	0.263
方案三序列 2	0.684	1.987	0.330
方案三序列 3	0.678	2.156	0.364
沙莫夫方案一	0.502	1.977	1.504
沙莫夫方案三	0.502	21.561	4.658

方案一中三次随机取样之后的计算结果差距不大，因此，可以排除样本的特殊性带来的误差。从整体来看，随着推移质输沙率的增大，图中点位呈收缩状，预测精度有明显的提高，这是因为在临近起动的低输沙状态下还有其他具有重要影响力的因素未能考虑进来。同时，尽管预测效果整体差距不大，但依旧可以看出序列 2 的预测效果相对较好，这意味着训练集样本的选择确实对

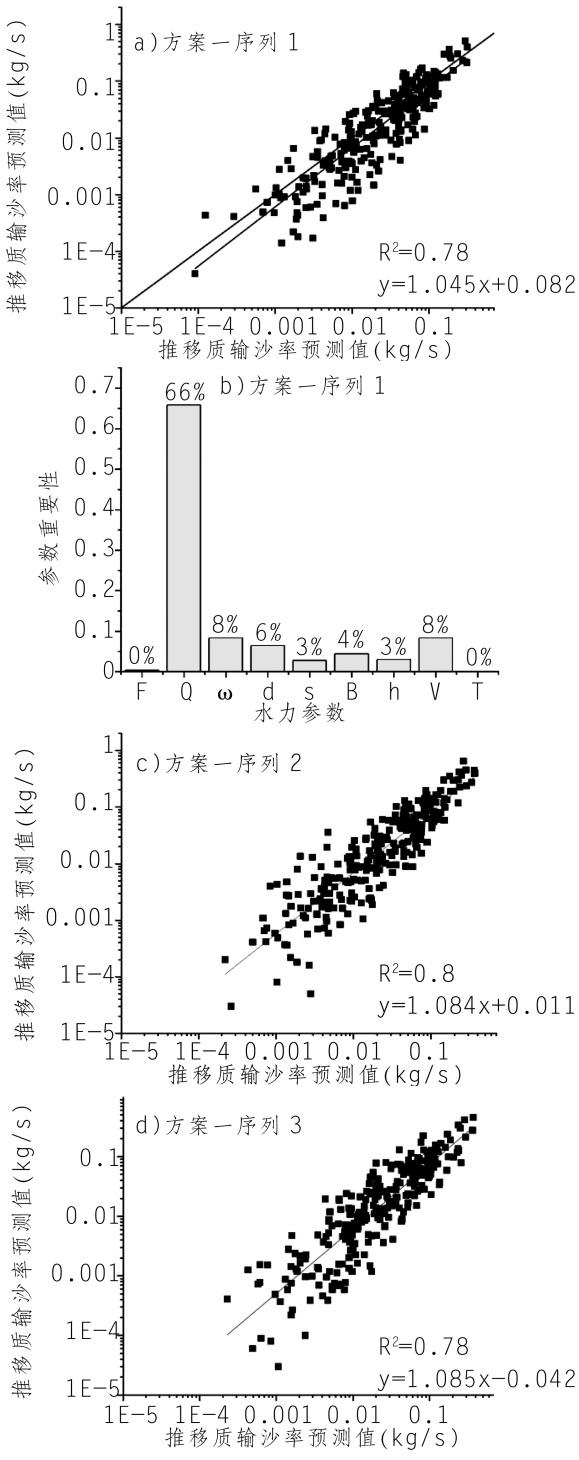

图 3　方案一中序列 1、序列 2、序列 3 各自的测试集运行结果与参数重要性排序

预测能力有影响。

对比前文中的相关性热力图可以看出，在选定的水力参数中，流量 Q 无论是在相关性方面还是算法的重要性方面均有极佳的表现，其他常规参数彼此之间差距不大，但与流量 Q 有了

断层式的差距。另外，可以察觉到相关性与参数在算法中的重要性并不对等，这是由于对于单一流域而言，流量 Q 不仅具有相对最高的相关性，同时极为敏感，加入随机噪声后，袋外准确率大幅度下降。

2.2.2　方案二的运行结果

与方案一相似，三次运行结果与参数重要性见图 4，评估指标见后面表 1。方案二、三的效果与方案一接近，此处仅放置具有代表性的图片。

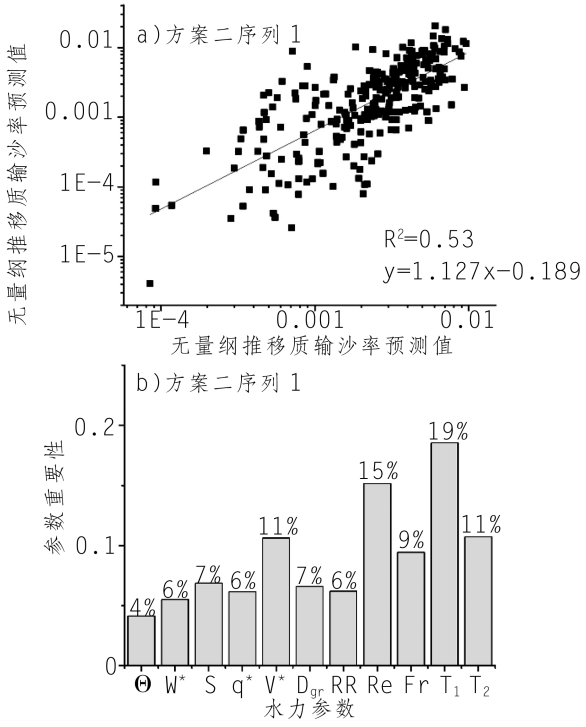

图 4　随机森林方案二中三次运行结果与参数重要性

方案二与方案一的数据来源均为圣路易斯河流域，只是参与算法运行的参数组合不同。在方案二中，各序列下的预测效果仍处于一定的范围内，因此，结果较为可靠。方案二中同样存在随着输沙率增大预测精度相应提高的现象，原因与方案一相同。

各水力参数的重要性均不高，但相对而言较为均衡。图中 T_1 与 T_2 是 T 的两种表现形式，同时出现在运算中（T_1 表示流量与平滩流量的比值在 0.8 以下时为类别 1，在 0.8 以上时为类别 2，而 T_2 则刚好相反）。这是为了验证同一个参数以不同的形式出现在随机森林的运算中，会呈现不同的重要性。其原因是不同的排列顺序与其他参数有着不同的关联性。

同时，图 4 也显示了纯粹以无量纲参数运行的算法效果并不理想。相比于方案一，方案二的参数组合在低输沙率的情况下表现尤其不佳。这是因为参数在无量纲化的过程中，加强了相互之间的关联性，这也导致生长的决策树相似性高于方案一的情况，而更高的相似性意味着较低的预测能力。

2.2.3　方案三的运行结果

随机森林方案三中三次运行结果与参数重要性见图 5，其他评估指标见后面表 1。方案二、三的效果与方案一接近，此处仅放置具有代表性的图片。

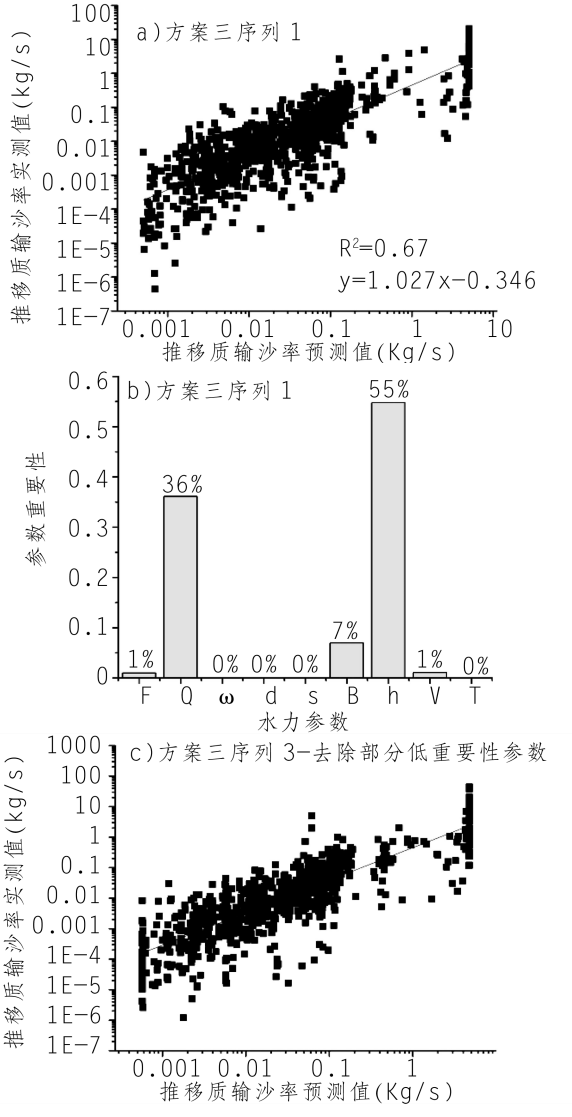

图 5　随机森林方案三中三次运行结果与参数重要性

根据图 5 所示，随机森林算法对于综合流域数据同样有较为良好的预测能力。从中可以看

出，推移质输沙率在低输沙率和较高的输沙率情况下，预测结果离散程度较大，效果相对较差。前者的原因已在前面解释过，而后者是因为较大的输沙率来自于山体滑坡或者泥石流等现象，超出了预测的范围。

对比相关性分析（图 2）部分，同样可以得到与方案一部分类似的结论，即相关性与参数重要性并不完全一致。在综合流域中，流量与水深对于算法来说是尤为重要的两个参数，其他参数的重要性可以忽略不计。在图 5 中，去掉了部分重要性近乎为 0 的水力参数后（如沉速、推移质中值粒径等），与原来的计算结果差距极小。这是由于综合流域中，流速与水深这两个参数不仅与输沙率的相关性较高，且在算法中表现得更为敏感，加入随机噪声后，袋外准确率大幅度下降。

2.3　沙莫夫公式计算结果

现在使用沙莫夫公式在两种方案下进行计算，分别对应方案一和方案三。从前文的结果来看，每种方案三次随机的数据并未存在划分过程的额外偏差，因此，仅需对比每种方案的序列 1 即可。沙莫夫公式方案一和方案三序列 1 的计算结果见图 6。

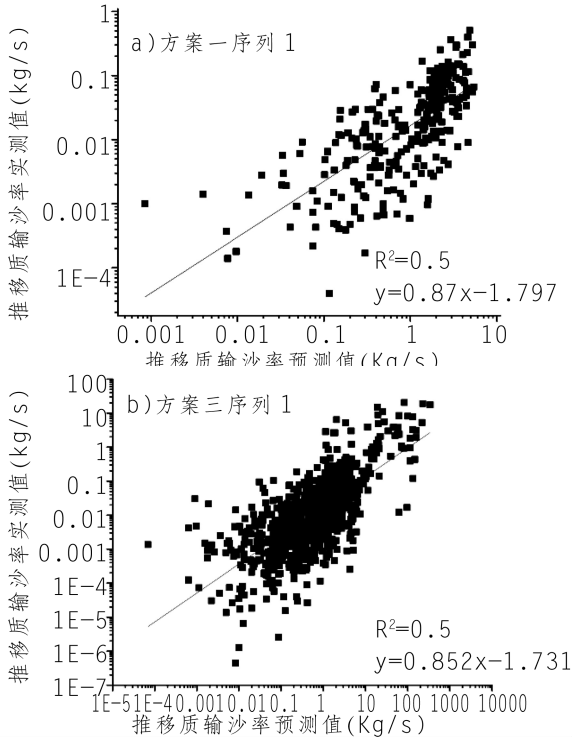

图 6　沙莫夫公式方案一和方案三序列 1 的计算结果

从图 6 及表 1 可知，沙莫夫公式无论应用在单一流域（方案一）还是综合流域（方案三），其预测效果均劣于使用随机森林算法得到的结果。由此可见，在足够的数据量支持下，随机森林算法相比于传统公式有着更好的预测效果；但其缺点也极为明显，需要大量数据来训练算法，远不如公式灵活。

3　结　论

本文从 Darren Hinton 等收集的 15 000 多组天然河道推移质输移数据中筛选出 3 060 组作为研究对象，对随机森林算法在预测推移质输沙率方面的效果进行了研究，可以为以后该算法在此领域的研究提供参考。结论如下：

（1）水力参数与输沙率之间的相关性与参数在算法中的重要性并不完全一致。原因在于，这不仅与相关性有关，还与其在算法中的敏感性有关。

（2）相比于常规有量纲参数，纯粹以无量纲参数运行的算法效果并不理想，且同一参数以不同形式出现在算法运算中时，会呈现不同的重要性。这是因为参数在无量纲化的过程中，加强了参数之间的关联性，导致生长的决策树相似度更高而预测效果更差。

（3）在足够的数据量支持下，随机森林算法相比于传统公式明显有着更好的预测效果。但其缺点就在于需要预先提供大量的已有数据作为支撑，远不如推移质输沙率公式灵活。

参考文献：

[1]　谢世博. 基于 LSTM 神经网络的河流推移质卵石运动研究[D]. 华北水利水电大学，2021.

[2]　范中海. 川江推移质计算方法比选与防沙安全设计[D]. 四川大学，2005.

[3]　刘兴年，陈远信. 非均匀推移质输沙率[J]. 成都科技大学学报，1987，(2)：29-36.

[4]　KHOSRAVI K，COOPER J R，DAGGUPATI P，et al. Bedload transport rate prediction：Application of novel hybrid data mining techniques[J]. J Hydrol，2020，585.

[5]　KITSIKOUDIS V，SIDIROPOULOS E，HRISSANTHOU V. Machine Learning Utilization for Bed Load Transport in Gravel－Bed Rivers[J]. Water Resources Management，2014，28(11)：3727-43.

[6]　陈雄波，王俊昀，龚钰婷，等. 基于神经网络和遗传算法的泥沙模型研发及应用[J]. 人民黄河，2020，42(12)：18-22.

[7]　陆龙妹，赵明松，卢宏亮，等. 人工神经网络和随机森林在回归问题中的应用比较[J]. 科技创新与应用，2019，(10)：31-2+6.

（下转第 17 页）

[5] Gong Fengqiang, Wang Yunliang, Luo Song. Rockburst proneness criteria for rock materials: Review and new insights [J]. Journal of Central South University, 2020, 27 (10): 2793-2821.

[6] 蔡美峰,冀东,郭奇峰.基于地应力现场实测与开采扰动能量积聚理论的岩爆预测研究[J]. 岩石力学与工程学报,2013,32(10):1973-1980...

[7] 魏新江,陈涛涛,王霄,等.岩爆灾害研究与进展[J].现代隧道技术,2020,57(2):1-12..

[8] 何满潮,苗金丽,李德建,等.深部花岗岩试样岩爆过程实验研究[J]. 岩石力学与工程学报,2007,26(5):865-876..

[9] 黄润秋,王贤能.深埋隧道工程主要灾害地质问题分析[J].水文地质工程地质, 1998,(4):23-26..

[10] 谢杰辉.岩石破裂过程的自组织临界特性及岩爆倾向性[D].南华大学,2018:18-24..

[11] 冯涛,谢学斌,王文星,等.岩石脆性及描述岩爆倾向的脆性系数[J].矿冶工程,2000,20(4):18-19..

[12] 李庶林,冯夏庭,王泳嘉,等.深井硬岩岩爆倾向性评价[J]. 东北大学学报(自然科学版),2001,22(1):60-64.

[13] ZHOU Jian, SHI Xiuzhi, HUANG Rendong, et al. Feasibility of stochastic gradient boosting approach for predicting rockburst damage in burst-prone mines[J]. Transactions of Nonferrous Metals Society of China, 2016, 26 (7): 1938-1945.

[14] Kidybiński A. Bursting liability indices of coal[J]. International Journal of Rock Mechanics and Mining Sciences and Geomechanics Abstracts, 1981, 18 (4): 295-304.

[15] Zhou Jian, Li Xibing, Mitri Hani S, et al. Classification of rockburst in underground projects: comparison of ten supervised learning methods[J]. Journal of Computing in Civil Engineering, 2016, 30(5).

[16] 谢和平,陈至达.岩石断裂的微观机理分析[J]. 煤炭学报,1989,(2):57-67.

[17] 李庶林,冯夏庭,王泳嘉,等.深井硬岩岩爆倾向性评价[J]. 东北大学学报(自然科学版),2001,(1):60-64..

[18] 邓林,武君,吕燕.基于岩石应力应变过程曲线的岩爆能量指数法[J]. 铁道标准设计,2012,(7):108-111.

[19] Yin Xiangchu, Wang Yucang, Peng Keyin, et al. Development of a New Approach to Earthquake Prediction: Load/Unload Response Ratio (LURR) Theory[J]. Pure and Applied Geophysics, 2000, 157 (11-12): 2365-2383.

[20] Gong Fengqiang, Wu Chen, Luo Song, et al. Load - unload response ratio characteristics of rock materials and their application in prediction of rockburst proneness[J]. Bulletin of Engineering Geology & the Environment, 2019, 78 (7): 5445-5466.

[21] 中华人民共和国住房和城乡建设部.工程岩体分级标准 GB/T 50218-2014[S].北京:中华人民共和国住房和城乡建设部,2014.

[22] 国家铁路局.铁路隧道设计规范. TB 10003-2016[S].北京:国家铁路局,2016.

作者简介:

唐登志(1977-),男,四川华蓥人,本科,高级工程师,主要从事高速公路项目管理工作;

白根铭(1992-),男,云南昆明人,硕士,工程师,主要从事高速公路项目管理工作;

陈　爽(1993-),男,湖北荆州人,硕士,工程师,主要从事高速公路项目管理工作;

曲宏略(1984-),男,山东青州人,博士,副教授,主要从事岩土与地下工程方面的教学和研究工作;

李博文(1990-),男,四川达州人,硕士,主要从事岩土方面的研究工作;

蔡永灵(1998-),男,四川广安人,硕士,主要从事岩土方面的研究工作.

(责任编辑:卓政昌)

(上接第 6 页)

[8] HINTON D, HOTCHKISS R, AMES D P. Comprehensive and Quality－Controlled Bedload Transport Database[J]. Journal of Hydraulic Engineering, 2017, 143(2), 89-93.

[9] RYAN S E, PORTH L S, TROENDLE C A. Defining phases of bedload transport using piecewise regression[J]. Earth Surface Processes and Landforms, 2002, 27(9): 71-75.

作者简介:

江雨润(1995-),男,四川眉山人,硕士研究生,主要从事泥沙输移方面的研究;

黄　尔(1972-),男,四川营山人,研究员、博导,主要从事泥沙输移及河道演变规律等方面的研究.

(责任编辑:卓政昌)

(上接第 10 页)

姚　强(1987-),男,陕西宝鸡人,副教授,博导,主要从事工程爆破、水电水利工程施工技术方面的研究;

吴　钰(1996-)男,四川成都人,硕士研究生,主要从事三维激光技术在堆石坝施工中的应用研究;

王　千(1992-)男,四川达州人,本科工学学士,工程师,主要从事水利水电工程施工;

陈星艮(2000-)男,四川乐山人,硕士研究生,主要从事水工结构工程研究;

李洪涛(1979-),男,湖北仙桃人,教授,博导,主要从事工程爆破、水电水利工程施工技术方面的研究.

(责任编辑:卓政昌)