

DOI:10.3876/j.issn.1000-1980.2024.03.006

考虑地形与理化性质的土壤关键水力特性 多种模型构建与比较

张赛亚^{1,2}, 张珂^{1,2,3,4,5,6}, 晁丽君¹, 李运平^{1,2}, 张兆安^{1,2}, 黄轶铭¹

(1. 河海大学水文水资源学院, 江苏 南京 210098; 2. 河海大学水灾害防御全国重点实验室, 江苏 南京 210098;
3. 河海大学长江保护与绿色发展研究院, 江苏 南京 210098; 4. 中国气象局水文气象重点开放实验室, 江苏 南京 210098;
5. 水利部水利大数据重点实验室, 江苏 南京 210098; 6. 水利部水循环与水动力系统重点实验室, 江苏 南京 210098)

摘要: 为获取我国南方典型湿润山区关键土壤水力特性(饱和导水率(K_s)与田间持水量(F_c))的精细空间数据,以南方典型湿润山区屯溪流域为样本采集区,以地形因子和土壤理化性质为输入,采用相关性分析方法建立了 3 种不同输入模式,采用多元线性回归和机器学习技术构建了多元线性回归(MLR)、遗传算法-人工神经网络(GA-BP)、支持向量机回归(SVR)和随机森林(RF)4 种表层土壤水力特性模型,并将 4 种模型同传统土壤转换函数(PTFs)进行了对比分析,探究了不同输入模式的优劣。结果表明: K_s 估算效果由优到差排序为 RF、SVR、MLR、GA-BP、PTFs, F_c 估算效果由优到差排序为 SVR、RF、GA-BP、MLR、PTFs;屯溪流域 K_s 和 F_c 的空间变化呈现一致性,整体空间分布与屯溪流域高程变化保持一致,说明湿润山区表层土壤水力特性与高程存在密切的非线性关系;SVR 与 RF 模型更适用于小样本回归问题,GA-BP 模型则需要较大的样本容量来充分捕捉特征以达到理想效果。

关键词: 土壤转换函数;多元线性回归;机器学习;饱和导水率;田间持水量

中图分类号: S152.7 **文献标志码:** A **文章编号:** 1000-1980(2024)03-0042-09

Development and comparison of multiple models for estimating key soil hydraulic properties considering terrain and soil physiochemical properties

ZHANG Saiya^{1,2}, ZHANG Ke^{1,2,3,4,5,6}, CHAO Lijun¹, LI Yunping^{1,2}, ZHANG Zhaoan^{1,2}, HUANG Yiming¹

(1. College of Hydrology and Water Resources, Hohai University, Nanjing 210098, China;
2. The National Key Laboratory of Water Disaster Prevention, Hohai University, Nanjing 210098, China;
3. Yangtze Institute for Conservation and Development, Hohai University, Nanjing 210098, China;
4. China Meteorological Administration Hydro-Meteorology Key Laboratory, Hohai University, Nanjing 210098, China;
5. Key Laboratory of Water Big Data Technology of Ministry of Water Resources, Hohai University, Nanjing 210098, China;
6. Key Laboratory of Hydrologic-Cycle and Hydrodynamic-System of Ministry of Water Resources, Hohai University, Nanjing 210098, China)

Abstract: To obtain high-precision data of key soil hydraulic properties (saturated hydraulic conductivity (K_s) and field capacity (F_c)) in typical humid mountainous areas in southern China, four models were developed for estimating key soil hydraulic properties of the topsoil, including the multiple linear regression (MLR), genetic algorithm-artificial neural network (GA-BP), support vector regression (SVR), and random forest (RF). In addition, three input-variable combination modes were also established with terrain and soil physiochemical properties as inputs that were selected using correlation analysis. Then, four estimation models are compared with the pedotransfer functions (PTFs) to estimate key soil hydraulic properties. These estimation models are selected to predict soil hydraulic properties of the Tunxi Watershed. The results show that the estimation effect of K_s ranked in descending order as RF, SVR, MLR, GA-BP and PTFs, while the results of F_c ranked as SVR, RF, GA-BP, MLR and PTFs. The spatial variations of K_s and F_c in

基金项目: 国家重点研发计划项目(2023YFC3006500);国家自然科学基金项目(52009028);中央高校基本科研业务费专项资金资助项目(B240203007)

作者简介: 张赛亚(1998—),女,硕士研究生,主要从事土壤生态水文研究。E-mail: zhangsy@hhu.edu.cn

通信作者: 张珂(1979—),男,教授,博士,主要从事水文水资源、生态水文学研究。E-mail: kzhang@hhu.edu.cn

引用本文: 张赛亚,张珂,晁丽君,等. 考虑地形与理化性质的土壤关键水力特性多种模型构建与比较[J]. 河海大学学报(自然科学版), 2024, 52(3): 42-50.

ZHANG Saiya, ZHANG Ke, CHAO Lijun, et al. Development and comparison of multiple models for estimating key soil hydraulic properties considering terrain and soil physiochemical properties[J]. Journal of Hohai University(Natural Sciences), 2024, 52(3): 42-50.

the Tunxi Watershed show a consistency with the spatial variation of elevation, which indicates that there is a close nonlinear relationship between key soil hydraulic properties and elevation in humid mountainous areas. The SVR and RF models are more suitable for the regression analysis of small samples, while the GA-BP model requires larger samples to fully capture the features to achieve good results.

Key words: pedotransfer functions; multiple linear regression; machine learning; saturated hydraulic conductivity; field capacity

近年来,随着人们对生态水文模型关注度的日益提高,生态水文学逐步成为研究热点。土壤水力特性作为模型基本的输入条件,是影响土壤水分运动过程的关键因素,如何快速准确地确定土壤水力特性(如饱和导水率(K_s)、田间持水量(F_c)、土壤水扩散率等)已成为生态水文学及相关学科研究的重点问题之一。土壤水力特性的估算一般分为直接测量与间接方法。直接测量方法具有精度高的优点,然而却存在样本数量有限、耗时久^[1]、无法应用于大尺度生态水文过程研究等缺陷。为了弥补直接测量的相关缺陷,间接方法应运而生。土壤转换函数(pedotransfer functions, PTFs)方法是常见间接方法之一^[2],其由较易获取的土壤理化性质作为输入,通过一定的算法推求土壤水力特性参数^[3-5];Beke 等^[6]发现田间持水量与砂粒、黏粒质量分数及土壤容重之间存在显著相关性;Wösten 等^[7]研究了土壤转换函数的准确性与可靠性,发现该函数是估算土壤水力特性参数的良好工具;Lim 等^[8]在韩国全境以林区 K_s 为研究对象,将林区海拔、地形湿度指数与坡度作为 PTFs 输入推求水力特性参数,结果良好。

21 世纪以来,随着机器学习方法的愈发成熟,越来越多的研究人员将机器学习与土壤水力特性的估算结合研究;Singh 等^[9]利用遗传算法(genetic algorithm, GA)提出了两种基于机器学习的混合 PTFs,即遗传算法与多层感知器(multilayer perceptron-genetic algorithm, MLP-GA)和支持向量机(support vector regression-genetic algorithm, SVR-GA)的结合,结果表明 SVR-GA 比 MLP-GA 算法具有更高的效率;左炳昕等^[10]基于 UNSODA 全球土壤水力性质数据库,通过 BP、SVR、K-NN 对 van Genuchten 模型参数的模拟发现 SVR 对样本预测最好,BP 次之,K-NN 稍逊。此外,众多学者也在土壤水力参数的影响因素方面做了很多工作,发现 K_s 受容重、有机质含量^[11]以及温度^[12]的影响明显,田间持水量与土壤总孔隙度、黏粒质量分数^[13]、土壤类型、土地利用类型^[14]存在密切联系,然而却很少有学者关注土壤水力特性与地形特征之间的关系。

我国南方湿润山区地形起伏较大,导致土壤构成和植被分布对生态水文过程的影响较为明显^[15]。因此,提供精细尺度下栅格化土壤水力特性对研究生态水文过程控制机理具有重要作用,也可减小生态水文模型对经验参数的依赖,为改进生态水文模型提供科学依据和数据基础^[16-17]。屯溪流域夏季强降雨频发,产汇流时间短,径流响应迅速^[18-19],生态水文过程较为典型,且该过程与土壤水力特性密切相关^[20]。本文以南方典型湿润山区屯溪流域为样本采集区,基于 K_s 和 F_c 与土壤理化性质的响应,通过建立 3 种不同输入模式和构建多元线性回归(multiple linear regression, MLR)、遗传算法-人工神经网络(GA-BP)、SVR 和随机森林(random forest, RF)4 种模型,并与 PTFs 进行对比分析,以探明 K_s 和 F_c 与地形特征之间的联系。

1 研究区域与数据

1.1 研究区概况

屯溪流域位于安徽省最南端,属于亚热带季风气候,地势分布为西高东低,高程变化较明显,山区地势条件复杂,年均降水量充沛,是湿润山区的典型代表。

研究选取屯溪流域内的典型流域为研究区(图 1),该流域地势起伏明显,由北至南高程逐渐降低,山区地形特征复杂,下垫面物理条件空间异质性明显。此外,流域多年平均气温约为 17℃,多年平均降水量约为 1 800 mm,雨量充沛,属于典型的湿润区域。

1.2 研究数据

2020 年 10 月在典型流域开展野外实地考察,进行土壤样本信息采集,分别测量有机碳、土壤孔隙率、土壤容重、土壤机械组成(砂粒、粉粒、黏粒的质量分

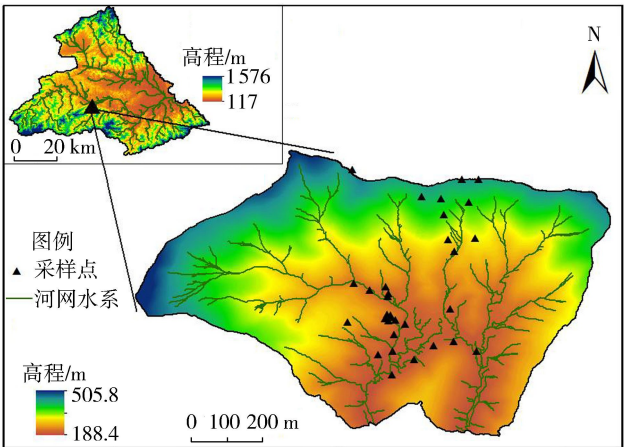


图 1 屯溪流域采样区域及采样点分布

Fig. 1 Sampling area and distribution of sampling points in Tunxi Watershed

数)、 K_s 和 F_c (表1)。此外,还采集高程、坡度、坡向以及地形湿度指数(T_{WI})等信息。在典型流域共采集 33 个样本,有效样本为 33 个,按 7 : 3 的比例随机抽取训练样本和验证样本,其中训练样本 23 个,验证样本 10 个。

表 1 采样区表层土壤理化性质统计特征

Table 1 Statistical characteristics of physiochemical properties of surface soil in the sampling area

特征统计值	$w_{SOT}/\%$	$n/\%$	$\rho/(\text{g}/\text{cm}^3)$	$w_s/\%$	$w_f/\%$	$w_c/\%$	$K_s/(\text{cm}/\text{d})$	$F_c/(\text{m}^3/\text{m}^3)$
平均值	4.195	56.444	1.067	20.880	57.665	21.539	101.197	0.546
最大值	8.620	64.590	1.420	28.268	71.910	32.600	320.256	0.964
最小值	2.327	45.760	0.600	13.989	47.817	5.100	6.048	0.347
中 值	4.276	57.370	1.080	20.597	56.369	21.600	76.752	0.537
标准值	1.380	4.362	0.153	3.243	5.501	6.155	76.920	0.118

注: w_{SOT} 为 SOT 含量, n 为孔隙率, ρ 为密度, w_s 、 w_f 、 w_c 分别为砂粒、粉粒、黏粒的质量分数。
有机碳、砂粒质量分数、黏粒质量分数和 F_c 的变异系数分别为 32.90%、15.53%、28.58%、21.68%,属于中等变异(15%~35%);孔隙率、土壤容重、粉粒质量分数的变异系数分别为 7.73%、14.33%、9.54%,为低变异系数(<15%); K_s 的变异系数为 76.01,为强变异(>35%), K_s 为强变异的重要原因是气候、地形地貌和人为活动等多因素引起土壤性质的分布不均匀^[21]。

2 研究方法

2.1 4 种模型

2.1.1 MLR 模型

MLR 被广泛应用于含有多个解释变量的预测因子。然而在土壤特性估算方法中回归函数通常是较为复杂的非线性函数,为解决这类问题,本文将预测因子与解释变量之间的关系处理为线性关系:

$$y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_i x_i \tag{1}$$

式中: y 为因变量; x_1 、 x_2 、 x_3 、 x_i 为解释变量; α_0 为偏差; α_1 、 α_2 、 α_3 、 α_i 为各解释变量对应的系数。

2.1.2 GA-BP 模型

GA-BP 模型的泛化能力及非线性映射能力较强^[22]。研究^[23]表明,3 层 BP 网络具有较强的非线性映射能力,能够以任意精度逼近一个连续函数。因此,本文选用 3 层网络:输入层、隐含层及输出层各为一个。其他参数设置如下:输入层和隐含层的激励函数为 tansing 函数,输出函数为 purelin 函数,训练函数为 trainlm。学习速率为 0.01,训练目标为 0.001。遗传算法中,进化代数为 100,种群规模为 50。由于本文样本数量有限,神经元个数过多容易引起模型训练的过拟合,故本文设为 1 个神经元。

2.1.3 SVR 模型

SVR 基于有限样本信息在模型的复杂性和模型的学习能力中寻求最优方案,即寻找一条最优拟合曲线,尽可能描述更多数据点。其构建的回归模型为

$$f(x) = w^T X + b \tag{2}$$

式中: w 为权重; X 为样本数据矩阵; b 为偏置项。与一般线性回归相比,SVR 在 $f(x)$ 两侧人为设定了一个间隔带“ ε ”,即当数据落在间隔带以内不计损失(图 2)。

利用优化目标(式(3))和约束条件(式(4)),得到最终的回归函数(式(5))。

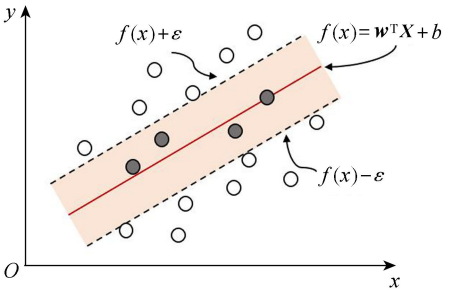


图 2 支持向量机结构示意图
Fig.2 Diagram of support vector regression structure

$$\min_{w,b,\xi_i,\hat{\xi}_i} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^l (\xi_i + \hat{\xi}_i) \tag{3}$$

$$\text{s. t. } \begin{cases} f(x_i) - y_i \leq \varepsilon + \xi_i \\ y_i - f(x_i) \leq \varepsilon + \hat{\xi}_i \end{cases} \quad (i = 1, 2, \cdots, l) \tag{4}$$

$$f(x) = \sum_{i=1}^l (\alpha_i^* + \alpha_i) K(x_i, x) + b \tag{5}$$

式中: ξ_i 和 $\hat{\xi}_i$ 为松弛变量; C 为 C-SVM 惩罚参数; I 为训练样本总量; y_i 为第 i 个训练样本目标; α_i 、 α_i^* 为拉格朗日乘子; $K(x_i, x)$ 为核函数。

由式(5)可知, $f(x)$ 与核函数密切相关,核函数的选择对模型性能有较大影响。高斯径向基核函数具有超参数少,非线性强的优点^[24],因此本文选取高斯径向基核函数。

2.1.4 RF 模型

RF 属于一种有监督学习算法,能够处理离散型与连续性数据,且不需要对数据进行归一化处理,对存在部分缺失数据的模拟比较友好。本文采用基于 Bootstrap 自主抽样和 CART 算法的 RF 模型。RF 通常在运行过程中能够自主保持误差平衡,不需要过度的人工干预,具有较快的运行速度。

2.2 PTFs

选取 Wösten 等^[7,25-26]提出的 3 种 PTFs (Wösten 函数、Weynants 函数、Jabro 函数) 估算 K_s , 选择 Beke 等^[6,27-28]提出 3 种 PTFs (BM 函数、AP 函数、GL 函数) 估算 F_c , PTFs 详见表 2。

表 2 本文选取的 PTFs

Table 2 Selected pedotransfer functions in this paper			
预测变量	函数	输入变量	PTFs
K_s	Weynants	$w_{SOT}、\gamma、w_s$	$e^{8.2927-0.0794w_s-1.9639\gamma+0.0192w_{SOT}}$
	Wösten	$w_{SOM}、\gamma、w_f、w_c$	$0.4327e^{-0.6381+1.0518\gamma^2-0.1206w_{SOM}+0.0545w_{SOM}^2+0.6033\ln(w_c+w_f)}$
	Jabro	$\gamma、w_f、w_c$	$e^{-2.6611+0.6871w_f+2.1362\lnw_c-2.2498\gamma}$
F_c	AP	$\gamma、w_s$	$0.5920+0.0290w_s-0.0294w_s\gamma$
	GL	$w_{SOM}、\gamma、w_s、w_f、w_c$	$0.0173w_s+0.0142w_f+0.0187w_c-0.0272w_{SOM}-0.8597\gamma$
	BM	$\gamma、w_s、w_f$	$1.1907-0.0019w_s-0.0098w_f\gamma$

注: γ 为容重, w_{SOM} 为 SOM 含量, w_{SOM} 与 w_{SOT} 之间的关系为: $w_{SOM}=1.724w_{SOT}$ 。

2.3 数据预处理

本文构建 MLR、GA-BP、SVR、RF 模型,所用的输入变量包括 w_{SOT} 、 n 、 γ 、 w_s 、 w_f 、 w_c 、高程、坡度、坡向、 T_{WI} 。因土壤水力特性受多种理化性质的复杂影响且并非属于简单的线性关系^[29],故而将输出变量与输入变量做变形处理。基本步骤如下:①将原始输入数据分别做 x 、 x^2 、 $\ln x$ 、 e^x 处理,输出变量 K_s 和 F_c 分别处理为 $\ln x$ 与 x ;②将 K_s 和 F_c 与所有输入变量做 Pearson 相关性检验,得到 4 种变形对应的相关性系数矩阵;③选取各个变量相关性最高的形式,从而确定输入变量的形式。综上所述,得到 4 种变形对应的相关性系数矩阵(表 3)。

表 3 输入变量各变形形式与 $\ln K_s$ 和 F_c 的相关系数

Table 3 Correlation coefficient matrix of various deformation forms, $\ln K_s$ and F_c								
输入变量	$\ln K_s$				F_c			
	x	x^2	$\ln x$	e^x	x	x^2	$\ln x$	e^x
w_{SOT}	0.3024	0.24	0.3428	-0.001784	0.307	0.2466	0.3533	0.05972
n	0.4776	0.4838	0.47	0.2354	0.681	0.6816	0.6784	0.2716
γ	-0.4266	-0.4537	-0.382	-0.4508	-0.9526	-0.9263	-0.9624	-0.926
w_s	-0.2374	-0.24	-0.2311	-0.1732	-0.04901	-0.03221	-0.06425	0.1274
w_f	-0.4803	-0.49	-0.4689	-0.3002	-0.229	-0.2316	-0.225	-0.1066
w_c	0.5095	0.4662	0.524	0.03235	0.1913	0.1473	0.2218	-0.07315
高程	0.4628	0.4213	0.5041	0.014	0.6927	0.6896	0.6915	0.06306
坡度	-0.2230	-0.2017	-0.2345	-0.01313	-0.2849	-0.2998	-0.1785	-0.0004124
坡向	0.07325	0.2126	-0.1214	0.1522	-0.06837	0.01765	-0.1507	0.1229
T_{WI}	-0.03888	-0.07553	0.003617	-0.1748	-0.032	0.00548	0.02599	0.05666

依据 4 种变形对应的相关性系数(表 3),取相关系数最大对应的形式为最佳变形形式。对于 $\ln K_s$, n 、 γ 、 w_s 、 w_f 、坡向的最佳变形形式为 x^2 ; w_{SOT} 、 w_c 、高程、坡度为 $\ln x$; T_{WI} 为 e^x 。对于 F_c , 高程的最佳变形形式为 x , n 、 w_f 、坡度为 x^2 , w_{SOT} 、 γ 、 w_c 、坡向为 $\ln x$; w_s 、 T_{WI} 为 e^x 。

2.4 输入模式及评价指标

为确定不同模型下各因子的最优组合,本文基于研究对象的特点以及各因子之间的复杂关系,根据变量

间的显著性将输入模式分为 3 种:①输入全部变量,②输入相关性表现在 0.3 以上的变量,③输入相关性表现在 0.5 以上的变量。分别记为 First、Second、Third(表 4)。

本文选取相关系数(R)、均方根误差(RMSE)和偏差(Bias)作为评价指标。

3 结果与分析

3.1 不同模型的结果对比

为探究典型湿润山区关键土壤水力特性与地形和土壤理化性质之间的复杂关系,获取典型湿润山区关键土壤水力特性的精细空间数据,本文基于 MLR、GA-BP、SVR、RF 模型及 PTFs,得到 $\ln K_s$ 和 F_c 的散点图(图 3)及估算精度(表 5)。

表 4 3 种模式对应的输入变量

Table 4 Input variables corresponding to three modes		
输入模式	预测 $\ln K_s$ 输入变量	预测 F_c 输入变量
First	w_{SOT} 、 n 、 γ 、 w_s 、 w_f 、 w_e 、高程、坡度、坡向、 T_{WI}	w_{SOC} 、 n 、 γ 、 w_s 、 w_f 、 w_e 、高程、坡度、坡向、 T_{WI}
Second	w_{SOT} 、 n 、 γ 、 w_f 、 w_e 、高程	w_{SOC} 、 n 、 γ 、高程、坡度
Third	w_f 、 w_e 、高程	n 、 γ 、高程

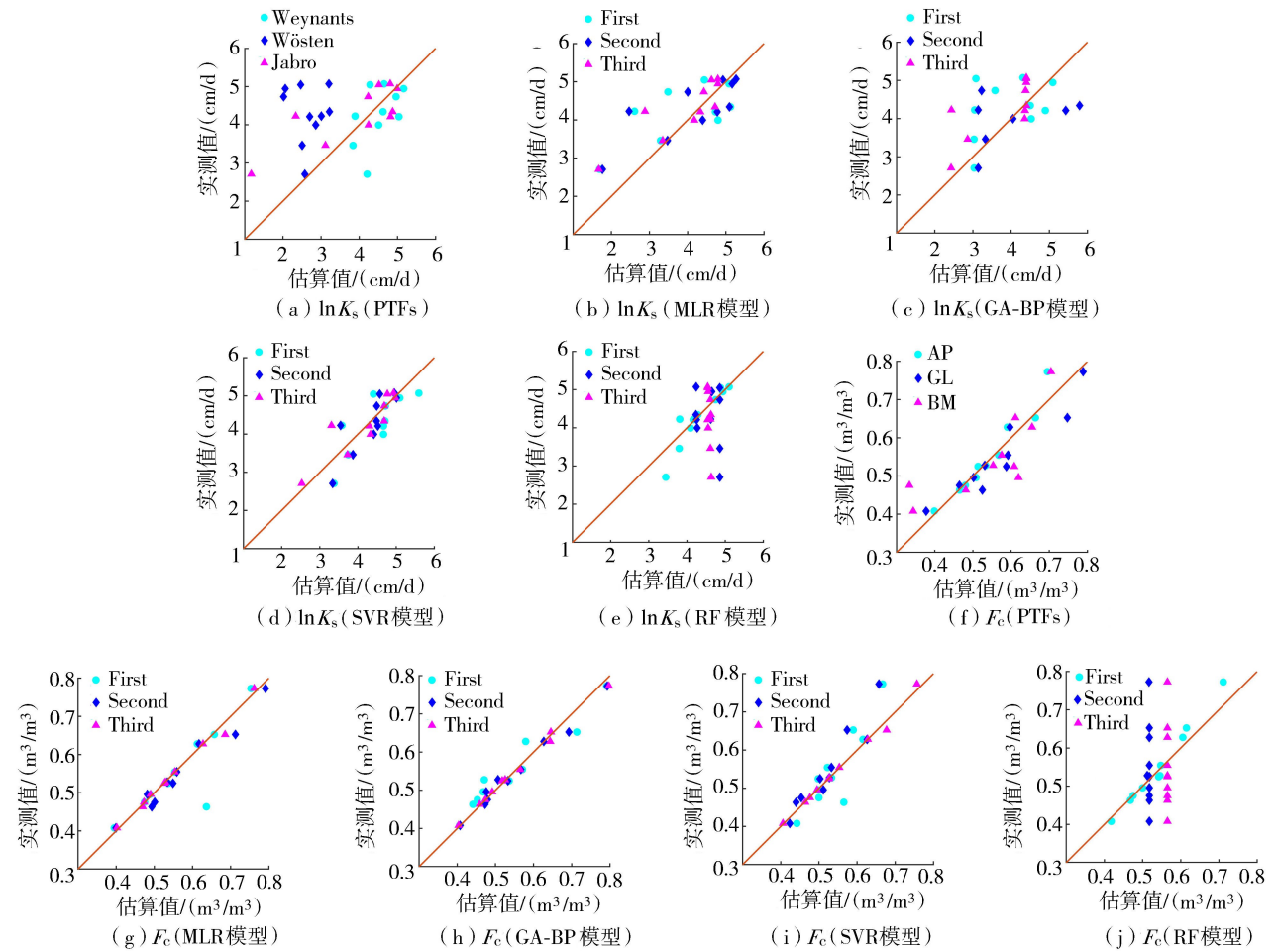


图 3 5 种模型 $\ln K_s$ 和 F_c 散点图对比

Fig. 3 Scatter plot comparison of F_c and $\ln K_s$ for five models

从估算 $\ln K_s$ 来看,各模型估算效果参差不齐。PTFs 估算效果不理想,整体分布偏离射线 $y=x$,平均相关系数为 0.4170,平均均方根误差为 3.5205 cm/d,平均偏差为 0.5773 cm/d,说明选择的 PTFs 模型在湿润山区出现估算偏高的现象,估算效果较差,不适合应用于湿润山区土壤水力特性的估算。MLR 模型结果总体分布在射线 $y=x$ 周围,平均相关性系数为 0.7919,平均均方根误差为 2.0332 cm/d,平均偏差为-1.8961 cm/d,与 PTFs 模型不同的是 MLR 模型出现了估算值远低于实测值的情况。GA-BP 模型的结果虽然整体分布在 $y=x$ 周围,但其离散程度偏大,平均相关性系数为 0.6048,平均均方根误差为 0.8819 cm/d,平均偏差为-0.1475 cm/d,估算效果同样不理想。除 RF 模型的 Second 与 Third 模式外,SVR 与 RF 模型的估算效果明显优于上述的

3 种模型,平均相关系数达到 0.83 以上,平均均方根误差介于 0.17~0.3 cm/d 之间,平均偏差达到 0.05 cm/d 左右,与 MLR 与 GA-BP 模型相差了一个数量级。

表 5 5 种模型的估算精度
Table 5 Estimation accuracy of five methods

模型	函数/输入模式	$\ln K_s$			F_c		
		R	RMSE/(cm/d)	Bias/(cm/d)	R	RMSE/(m ³ /m ³)	Bias/(m ³ /m ³)
PTFs	Weynants/AP	0.5148	2.0950	-0.2428	0.9724	0.0903	0.0087
	Wösten/GL	-0.0647	5.7507	1.6121	0.9426	0.1428	-0.0209
	Jabro/BM	0.8009	2.7159	0.3625	0.7854	0.2344	0.0017
MLR	First	0.7263	2.0484	-1.8426	0.8543	0.0558	0.0137
	Second	0.7830	1.9866	-1.8440	0.9818	0.0256	0.0135
	Third	0.8665	2.0647	-2.0016	0.9940	0.0116	0.0016
GA-BP	First	0.3964	0.9093	-0.3639	0.9654	0.0343	-0.0073
	Second	0.6755	1.0231	0.3527	0.9918	0.0178	0.0051
	Third	0.7424	0.7134	-0.4313	0.9978	0.0106	0.0019
SVR	First	0.7785	0.2456	0.1827	0.8921	0.0029	-0.0071
	Second	0.8335	0.1619	0.0459	0.9504	0.0022	-0.0253
	Third	0.8858	0.1286	-0.0487	0.9957	8.87×10 ⁻⁵	6.98×10 ⁻⁴
RF	First	0.9307	0.2955	0.0529	0.9908	0.0254	-0.0133
	Second	-0.2350	0.8755	0.2971	-0.0581	0.1071	-0.0601
	Third	-0.6112	0.8022	0.3110	0.0023	0.1029	0.0272

从估算 F_c 来看,除 RF 模型中 Second 和 Third 模式外,各模型估算效果均较为理想。PTFs 估算结果集中在射线 $y=x$ 周围,平均相关系数为 0.9001,平均均方根误差为0.1558 m³/m³,平均偏差为 0.5773 m³/m³,说明估算值比实测值偏高。MLR 模型比 PTFs 模型的估算效果更好,平均偏差为 0.0096 m³/m³ 远小于 PTFs 模型,然而 MLR 模型的结果中出现了异常点,说明该模型对异常值的处理能力一般。GA-BP、SVR 与 RF 模型表现出较高的一致性,平均相关系数分别为 0.9850、0.9461、0.9908;平均偏差分别为-0.001 m³/m³、-0.0106 m³/m³、-0.0133 m³/m³,说明机器学习在典型湿润山区土壤水力特性的估算具有更好的适用性。

3.2 不同输入模式对比

对比发现,机器学习模型整体表现优于 PTFs 与 MLR 模型,同时也发现在 RF 模型中,随着输入模式的改变,估算效果发生显著变化,为此,本文为准确找到各模型的最佳输入模式,根据预测因子与输入变量之间相关性强弱选取了 3 种输入模式,具体精度见表 5。

在估算 $\ln K_s$ 时,MLR、GA-BP 与 SVR 模型均表现出 Third 模式要优于 First 与 Second 模式,其中,SVR 模型中的 Third 模式是三者表现最优的,相关系数为 0.8858、均方根误差为 0.1286 cm/d、偏差为-0.0487 cm/d,说明 SVR 模型在小样本回归问题上具有较强的泛化能力。GA-BP 模型表现为三者最差的原因:①数据样本较小,在训练、评估、超参数调整等过程中无法达到输出要求;②侧面说明 GA-BP 模型适合大容量样本集开发,对于小样本的训练、学习与预测无法发挥其优势。然而在 RF 模型中却出现与之相反的现象,其中 First 模式的相关系数为 0.9307、均方根误差为 0.2955 cm/d、偏差为 0.0529 cm/d,且随着输入模式的改变,估算效果显著下降,甚至出现了多值相同的异常情况,也因此出现相关系数小于 0,估算失真,说明了 RF 模型需要多特征的输入来估算预测因子。RF 模型的 First 模式均方根误差与偏差表现不如 SVR 模型是因为 RF 模型出现个别偏离程度大的点,而 SVR 模型则是整体偏离程度不大。除个别数据点,综合来看表现最优的是 RF 模型中的 First 模式。

F_c 与 $\ln K_s$ 估算既有相同又有不同。相同的是 MLR、GA-BP 与 SVR 模型同样也是 Third 模式优于 First 与 Second 模式,在 RF 模型中也出现随着输入模式的改变,估算效果显著下降甚至失真。不同的是 MLR、GA-BP 与 SVR 模型的 Third 模式与 RF 模型中的 First 模式表现十分相似(表 5),由此可以看出表现最优的是 SVR 模型的 Third 模式。同时,除 RF 模型外,其他模型的估算值要高于实测值,说明将 SVR 模型应用于屯溪流域土壤水力特性的估算会造成估算值要稍高于实测值。

综上,对于估算 $\ln K_s$ 、 F_c ,MLR、GA-BP、SVR 模型的最佳输入模式为 Third 模式,RF 模型的最佳输入模式为 First 模式。

3.3 屯流域表层土壤水力特性时空分布

5 种模型在估算土壤表层水力特性的表现优劣明显:估算 K_s 时,与 PTFs 与 MLR 模型相比,RF 和 SVR 模型的泛化能力最强,这样的情况也出现在估算 F_c 的过程,但两者在对 K_s 和 F_c 的估算精度并非完全一样,相比较而言,RF 模型在估算 K_s 时表现最优,而 SVR 模型估算 F_c 表现最优。因此,本文选择 RF 模型预测屯流域 K_s ,SVR 模型预测 F_c 。

基于栅格化土壤表层理化性质数据^[30]和屯流域地形特征得到屯流域 K_s 和 F_c 的空间分布(图 4),分辨率均为 250 m。发现屯流域 K_s 介于 36.08 ~ 162.01cm/d 之间,纵向格局呈现中间低,南北高,横向则表现为东低西高,整体格局与该地区的高程变化呈现一致性,说明南方湿润山区的 K_s 与高程存在密切的非线性关系; F_c 介于 0.42 ~ 0.83 m³/m³ 之间,纵向上的变化为先减小后增大,横向格局与 K_s 保持一致,呈现东低西高,是地形和人为活动共同导致的。

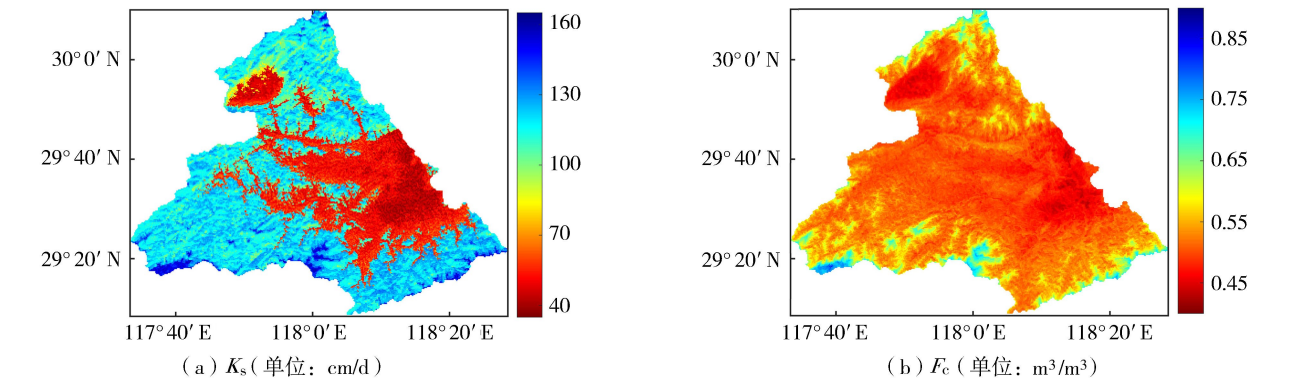


图 4 屯流域 K_s 和 F_c 空间分布
Fig.4 Spatial distribution of K_s and F_c in Tunxi Watershed

3.4 机器学习在小样本回归问题上的泛化能力

由表 6、表 7 可知,模型估算 F_c 的效果较好,而估算 K_s 相对较差,可能是其数值变化较大,样本容量较小,捕捉特征较为困难。

表 6 机器学习模型在小样本容量中 $\ln K_s$ 的泛化能力

Table 6 $\ln K_s$ generalization ability of machine models in small sample

模型	R		RMSE/(cm/d)		Bias/(cm/d)	
	训练集	验证集	训练集	验证集	训练集	验证集
GA-BP	0.5708	0.7424	0.8625	0.7134	-0.2224	-0.4313
SVR	0.5872	0.8858	0.8311	0.1286	0.1238	-0.0487
RF	0.7882	0.9307	0.6745	0.2955	0.0543	0.0529

表 7 机器学习模型在小样本容量中 F_c 的泛化能力

Table 7 F_c generalization ability of machine models in small sample

模型	R		RMSE/(m ³ /m ³)		Bias/(m ³ /m ³)	
	训练集	验证集	训练集	验证集	训练集	验证集
GA-BP	0.9399	0.9978	0.2931	0.0106	-0.2780	0.0019
SVR	0.9968	0.9957	0.0117	8.87×10 ⁻⁵	-0.0021	6.98×10 ⁻⁴
RF	0.9236	0.9908	0.0564	0.0254	0.0032	-0.0133

从训练集和验证集来看,各模型的训练集预测效果大部分要比验证集差,原因是数据集样本容量小,随机抽样后使得某些异常值保留在训练集中,造成了训练集的预测效果反而降低。从 3 种机器学习模型看,在训练集和验证集上的表现保持一致,其中 $\ln K_s$ 的预测效果最好的是 RF 模型,SVR 模型次之,GA-BP 模型较差;同样,GA-BP 模型对 F_c 的预测效果也比 SVR 和 RF 模型稍差,然而,SVR 模型在小样本回归问题上的泛化能力最优,RF 模型次之。

本文实地勘测数据来源于二级流域的嵌套小流域,具有典型湿润山区特征,采样点分布的设置遵循代表性、随机性等原则,覆盖流域的高程范围,且试验测定取多组土样的均值,很大程度上避免了数据极值。因

此,基于采样实测数据估算土壤表层水力特性具有较高的可靠性。土壤水力特性是土壤经过外因与内因驱动共同形成,与动态变化的土壤湿度存在明显区别,因此土壤水力特性与土壤理化性质和地形特征之间存在的时滞后性响应较为隐秘且复杂。此外,本文采用的多元线性分析与机器学习模型虽然是众多学者验证后切实可行的方法,然而其局限性也较为明显,如 GA-BP 模型对小样本回归问题的泛化能力较弱,RF 模型需要捕捉多特征的输入优化估算结果。

4 结 论

- a. 对数据进行预处理(x 、 x^2 、 $\ln x$ 、 e^x)后,得到 K_s 与粉粒(x^2)、黏粒($\ln x$)、高程($\ln x$)存在显著的非线性关系,而 F_c 与土壤孔隙度(x^2)、容重($\ln x$)、高程(x)存在显著的非线性关系。
- b. 在 5 种模型中, K_s 估算精度由高到低依次为 RF、SVR、MLR、GA-BP、PTFs, F_c 则为:SVR、RF、GA-BP、MLR、PTFs。与 PTFs 模型和 MLR 模型相比,GA-BP、SVR、RF 模型具有更简便、高效、低误差的优势。
- c. 依据评价指标择优选取 RF 模型预测屯溪流域 K_s ,SVR 模型预测 F_c ,结果发现屯溪流域 K_s 和 F_c 的空间变化基本一致,纵向格局为中间低,南北高,横向则是东低西高,整体空间分布与屯溪流域高程变化保持一致,说明湿润山区表层土壤水力特性与高程存在密切的非线性关系。
- d. K_s 和 F_c 的验证集的表现都要远优于训练集,原因是数据样本容量较小,随机抽样后使得某些异常值保留在训练集中;在对小样本数据进行回归训练中,SVR 模型的泛化能力最优,RF 模型次之,而 GA-BP 模型则更合适大容量数据集。

参考文献:

[1] 丁建丽,穆艾塔尔·赛地. 基于土壤转换函数的土壤水力性质估算研究进展[J]. 新疆大学学报(自然科学版),2017,34(1):96-101. (DING Jianli,MUATTAR Saydi. Assessment of soil hydraulic properties using pedotransfer functions-a review[J]. Journal of Xinjiang University(Natural Science Edition),2017,34(1):96-101. (in Chinese))

[2] 徐绍辉,刘建立. 土壤水力性质确定方法研究进展[J]. 水科学进展,2003,14(4):394-401. (XU Shaohui,LIU Jianli. Advances in approaches for determining unsaturated soil hydraulic properties[J]. Advances in Water Science,2003,14(4):394-401. (in Chinese))

[3] 李慧霞,刘建立,朱安宁,等. 预测天然文岩流域土壤饱和导水率的土壤转换函数方法比较研究[J]. 土壤,2010,42(3):438-445. (LI Huixia,LIU Jianli,ZHU An'ning,et al. Comparison study of soil pedo-transfer functions in estimating saturated soil hydraulic conductivity at Tianranwenyanqu Basin[J]. Soils,2010,42(3):438-445. (in Chinese))

[4] ROSSETI R D C,AMORIM R S S,DI RAIMO L A D,et al. Pedotransfer functions for predicting soil-water retention under Brazilian Cerrado[J]. Pesquisa Agropecuaria Brasileira,2022,57:90-95.

[5] GHANBARIAN B,TASLIMITEHRANI V,PACHEPSKY Y A. Accuracy of sample dimension-dependent pedotransfer functions in estimation of soil saturated hydraulic conductivity[J]. Catena,2017,149:374-380.

[6] BEKE G J,MACCORMICK M I. Predicting volumetric water retentions for subsoil materials from Colchester County,Nova Scotia [J]. Canadian Journal of Soil Science,1985,65:233-236.

[7] WÖSTEN J H M,LILLY A,NEMES A,et al. Development and use of a database of hydraulic properties of European soils[J]. Geoderma,1999,90(3/4):169-185.

[8] LIM H,YANG H,CHUN K W,et al. Development of pedo-transfer functions for the saturated hydraulic conductivity of forest soil in south korea considering forest stand and site characteristics[J]. Water,2020,12(8):2217-2235.

[9] SINGH V K,PANDA K C,SAGAR A,et al. Novel genetic algorithm (GA) based hybrid machine learning-pedotransfer function (ML-PTF) for prediction of spatial pattern of saturated hydraulic conductivity[J]. Engineering Applications of Computational Fluid Mechanics,2022,16(1):1082-1099.

[10] 左炳昕,查元源. 基于机器学习方法的土壤转换函数模型比较[J]. 灌溉排水学报,2021,40(5):81-87. (ZUO Bingxin,ZHA Yuanyuan. Using machine learning methods as a pedotransfer function to estimate soil hydraulic parameters[J]. Journal of Irrigation and Drainage,2021,40(5):81-87. (in Chinese))

[11] 石文豪. 基于天津地区土壤水力参数的传递函数构建研究[D]. 天津:天津大学,2021.

[12] 秦文静. 黄土水力运动参数经验模型参数的传递函数研究[D]. 太原:太原理工大学,2020.

[13] 李向富,刘目兴,易军,等. 三峡山地不同垂直带土壤层的水文功能及其影响因子[J]. 长江流域资源与环境,2018,27(8):1809-1818. (LI Xiangfu,LIU Muxing,YI Jun,et al. Soil hydrological function of different altitudinal hillslopes of the Three Gorges Mountain and its impact factors[J]. Resources and Environment in the Yangtze Basin,2018,27(8):1809-1818. (in

Chinese))

- [14] 黄兴成,熊仕娟,李渝,等. 贵州赤水河流域土壤持水特征的影响因素分析[J]. 西南大学学报(自然科学版),2022,44(5): 58-64. (HUANG Xingcheng,XIONG Shijuan,LI Yu,et al. The influencing factors of soil water holding capacity in the Chishui River Basin of Guizhou[J]. Journal of Southwest University(Natural Science Edition),2022,44(5):58-64. (in Chinese))
- [15] ZHANG K,ALI A,ANTONARAKIS A,et al. The sensitivity of north american terrestrial carbon fluxes to spatial and temporal variation in soil moisture: an analysis using radar-derived estimates of root-zone soil moisture[J]. Journal of Geophysical Research-Biogeosciences,2019,124(11):3208-3231.
- [16] 张海阔,姜翠玲,李亮,等. 基于 HYDRUS-1D 模拟的变水头入渗条件下 VG 模型参数敏感性分析[J]. 河海大学学报(自然科学版),2019,47(1):32-40. (ZHANG Haikuo,JIANG Cuiling,LI Liang,et al. Parameter sensitivity analysis of VG model in the varying-head infiltration based on HYDRUS-1D simulation[J]. Journal of Hohai University(Natural Sciences),2019,47(1): 32-40. (in Chinese))
- [17] 刘家宏,王东,李泽锦,等. 厦门海绵化改造区入渗性能测定及特征[J]. 水资源保护,2019,35(6):9-13. (LIU Jiahong, WANG Dong,LI Zejin,et al. Measurement and characteristic analysis of infiltration performance in Xiamen spongification reconstruction area[J]. Water Resources Protection,2019,35(6):9-13. (in Chinese))
- [18] 严登明,翁白莎,宋新山,等. 那曲流域草地覆盖变化对降雨产流过程的影响[J]. 水资源保护,2019,35(6):44-51. (YAN Dengming,WENG Baisha,SONG Xinshan,et al. Effect of grassland cover change on rainfall and runoff yield in Naqu Watershed [J]. Water Resources Protection,2019,35(6):44-51. (in Chinese))
- [19] 牛欣怡,鲁程鹏,卢佳赞,等. 机器学习模型在地下水埋深模拟中的适应性分析[J]. 河海大学学报(自然科学版),2022, 50(4):74-82. (NIU Xinyi,LU Chengpeng,LU Jiayun,et al. Adaptability analysis of machine learning model in groundwater depth simulation[J]. Journal of Hohai University(Natural Sciences),2022,50(4):74-82. (in Chinese))
- [20] 赵宇寒,曹建生,朱春雨,等. 壤中流形成机制及其生态水文效应研究进展[J]. 中国生态农业学报(中英文),2022,30 (1):38-46. (ZHAO Yuhan,CAO Jiansheng,ZHU Chunyu,et al. Research progress on the formation mechanism of subsurface flow and its eco-hydrological effects[J]. Chinese Journal of Eco-Agriculture,2022,30(1):38-46. (in Chinese))
- [21] 管孝艳,王少丽,高占义,等. 盐渍化灌区土壤盐分的时空变异特征及其与地下水埋深的关系[J]. 生态学报,2012,32 (4):198-206. (GUAN Xiaoyan,WANG Shaoli,GAO Zhanyi,et al. Spatio-temporal variability of soil salinity and its relationship with the depth to groundwater in salinization irrigation district[J]. Acta Ecologica Sinica,2012,32(4):198-206. (in Chinese))
- [22] 刘佩瑶,郝振纯,王国庆,等. 新安江模型和改进 BP 神经网络模型在闽江水文预报中的应用[J]. 水资源与水工程学报, 2017,28(1):40-44. (LIU Peiyao,HAO Zhenchun,WANG Guoqing,et al. Application of Xin'anjiang model and the improved BP neural network model in hydrological forecasting of the Min River[J]. Journal of Water Resources & Water Engineering, 2017,28(1):40-44. (in Chinese))
- [23] 孙娜,周建中,朱双,等. 基于小波分析的两种神经网络耦合模型在月径流预测中的应用[J]. 水电能源科学,2018,36 (4):14-17. (SUN Na,ZHOU Jianzhong,ZHU Shuang,et al. Application of hybrid models based on wavelet analysis and two different neural networks in prediction of monthly runoff[J]. Water Resources and Power,2018,36(4):14-17. (in Chinese))
- [24] 姚望. 基于支持向量回归的定向能量沉积熔道尺寸预测[D]. 广州:华南理工大学,2020.
- [25] WEYNANTS M,VERECKEN H,JAVAUX M. Revisiting vereecken pedotransfer functions: introducing a closed-form hydraulic model[J]. Vadose Zone Journal,2009,8(1):86-95.
- [26] JABRO J D. Estimation of saturated hydraulic conductivity of soils from particle size distribution and bulk density data[J]. 1992, 35(2):557-560.
- [27] AINA P O,PERIASWAMY S P. Estimating available water-holding capacity of western Nigerian soils from soil texture and bulk density, using core and sieved samples[J]. Soil Science,1985,140:55-58.
- [28] GUPTA S C,LARSON W E. Estimating soil water retention characteristics from particle size distribution,organic matter percent, and bulk density[J]. Water Resources Research,1979,15:1633-1635.
- [29] 王志强. 科尔沁沙地土壤水力特性的推算[D]. 呼和浩特:内蒙古农业大学,2003.
- [30] LIU F,WU H,ZHAO Y,et al. Mapping high resolution national soil information grids of China[J]. Science Bulletin,2022,67 (3):328-340.

(收稿日期:2023-07-21 编辑:刘晓艳)