引 用 格 式：Xu Hongtao, Chen Chunbo, Zheng Hongwei, *et al*. SVR Salinization Monitoring based on Integrated Feature Subset Selection and Model Parameter Learning[J]. Remote Sensing Technology and Application, 2021, 36(1): 176-186.[徐红涛，陈春波，郑宏伟，等. 集成建模变量优选和参数学习的SVR盐渍化监测[J]. 遥感技术与应用，2021，36(1)：176-186.]
doi:10.11873/j.issn.1004-0323.2021.1.0176

集成建模变量优选和参数学习的SVR盐渍化监测

徐红涛[1,2]，陈春波[1,2]，郑宏伟[1,2]，罗格平[1,2]，杨 辽[1,2]，王伟胜[1,2]，吴世新[1,2]
（1.中国科学院新疆生态与地理研究所，荒漠与绿洲生态国家重点实验室，新疆 乌鲁木齐 830011；
2.中国科学院大学，北京 100049）

摘要：当前基于机器学习算法反演土壤盐分含量（Soil Salt Content，SSC）较少关注模型参数和建模变量的优选。基于Sentinel-1 SAR、Landsat 8 OLI、DEM数据提取8类共40个环境变量，经Pearson相关分析初步筛选出候选特征变量，分别带入格网搜索（Grid Search，GS）算法、遗传算法（Genetic Algorithm，GA）、粒子群算法（Particle Swarm Optimization，PSO）同步优选支持向量回归（Support Vector Regression，SVR）的模型参数和建模变量，然后建立盐渍化监测模型（GS-SVR、GA-SVR、PSO-SVR），选择最优模型反演玛纳斯灌区生长季SSC分布。结果表明：提取的环境变量与SSC相关性较好，植被指数和特征空间对盐渍化更为敏感；与GS-SVR相比，GA-SVR和PSO-SVR减少了建模变量，提高了模型精度，适应度值分别提高了53.87%、69.96%；生长季内，春秋季积盐，夏季脱盐，SSC均值变化趋势：整个研究区、中部和南部为降低—增加；北部为增加—降低—增加；依据生长季SSC小提琴图表明整个研究区，中部和北部SSC数值区间变化趋势为扩张—收缩—扩张，南部为扩张—收缩—平稳。

关 键 词：遗传算法；粒子群算法；土壤盐渍化；支持向量机；模型参数和建模变量优选
中图分类号：S127;TP79 **文献标志码**：A **文章编号**：1004-0323(2021)01-0176-11

1 引 言

土壤盐渍化是干旱区、半干旱区土地退化的主要表现形式之一，尤其在中国西北边陲的新疆，作为重要的农垦基地却长期遭受盐渍化的影响，弃耕地较多，农业生产效率低下[1]。因此，大尺度、精准的盐渍化监测对农业生产和生态环境保护意义重大[2]。此外，基于遥感等其他数据提取环境变量，结合机器学习算法，尤其是支持向量回归（Support Vector Regression，SVR）因擅长挖掘高维数据之间的潜在关系，已成为大尺度盐渍化定性与定量监测的主要途径[2-5]。

研究表明，植被指数、盐分指数、地形因素等环境变量可以作为盐渍化监测的有效辅助信息[3,5-6]。不同类型的环境变量虽可以从不同角度反映表层土壤的状态，但大部分均是通过波段运算得到，存在冗余信息[7]。对机器学习而言，模型参数和建模变量的优选对模型精度的提高至关重要。部分学者采用相关分析，以显著性（$p<0.05$）为标准筛选环境变量参与盐渍化模型的构建[5-6]。王飞等[8]采用特征重要性排序的方法优选建模变量，减少了不确定性，提高了盐渍化监测的准确性。遗传算法（Genetic Algorithm，GA）和粒子群算法（Particle Swarm Optimization，PSO）因具有强大的优化能

收稿日期：2019-10-08；**修订日期**：2020-12-30
基金项目：国家自然科学基金项目（41877012），中国科学院一带一路团队项目（2018-YDYLTD-002），中国科学院特色研究所项目（TSS-2015-014-FW-1-3）。
作者简介：徐红涛（1993－），男，河南驻马店人，硕士研究生，主要从事遥感与地理信息系统研究。E-mail:xuhongtao17@ mails.ucas.ac.cn
通讯作者：郑宏伟（1972－），男，山东潍坊人，博士，研究员，主要从事机器智能和模式分析，生态与地理及气候效应、遥感与GIS应用研究。E-mail:hzheng@ms.xjb.ac.cn

力,已被广泛应用于不同类型的优化问题,却鲜见于土壤盐渍化的定量评估。Nurmemet等[1]采用GA优选建模变量,使用格网搜索(Grid Search,GS)算法优化模型参数,提高了盐渍化分类的精度。谭林等[9]分别基于PSO和GA优化SVR的模型参数,发现PSO的优化能力优于GA。上述研究虽通过优选模型参数和建模变量提高了模型精度,但其两者的优选不同步,会导致算法陷入局部最优。另外,PSO和GA在盐渍化定量评估中的模型参数和建模变量的同步优选的适用性有待验证。

基于此,以玛纳斯灌区土壤盐渍化为研究对象,基于Sentinel-1 SAR、Landsat 8 OLI和DEM数据、提取8类共40个环境变量,结合Pearson相关分析初步筛选出候选特征变量(Candidate Feature Variables, CFVs),分别代入GS、GA、PSO优选SVR的建模变量与模型参数,建立研究区的盐渍化监测模型(GS-SVR、GA-SVR、PSO-SVR),选择最优模型反演灌区生长季(4~10月)土壤盐分含量(Soil Salt Content, SSC)分布并分析其时空变化,以期为盐渍化定量评估和生长季盐分含量动态变化提供一定的技术参考。

2　数据与方法

2.1　研究区概况

玛纳斯灌区位于天山北部、准格尔盆地南缘(图1)。该区年均降水110~200 mm,年均气温7.2 ℃,年均蒸发约1 600 mm,热量充足,属典型的温带大陆性干旱半干旱气候[10]。灌区内地势由东南向西北倾斜,水资源主要依靠冰雪融水和农业灌溉补给。建国之后,人口快速增长,耕地面积迅速扩大,绿洲耗水激增,粗放的农业管理方式抬升了地下水位,次生盐渍化问题严重[11]。

主要研究对象为绿洲区及附近荒漠区的盐渍化,结合玛纳斯县区划图和遥感影像,排除大片的荒漠及山区,最终确定研究区的边界,并参考吕娜娜等[12],依据地貌、水文地质和人类活动等特征将研究区划分为南、中、北部3个部分(图1)。南部主要指地下水溢出带以南的冲积平原顶部;中部和北部以S201省道为界线。

2.2　数据来源及处理

2.2.1　土壤盐分含量实测数据

2016年7~8月在玛纳斯灌区共采集105个SSC实测数据[13],用于模型的训练和验证。采样数据分布有不同盐分含量的表层土壤信息、地貌类型涵盖。采样前,首先使用GPS记录样点的地理坐标,每个样点取样3次,带回实验室,待风干去除杂质,研磨均匀后过2 mm筛,测定八大离子含量(Ca^{2+}、Mg^{2+}、K^{+}、Na^{+}、CO_3^{2-}、HCO^{3-}、CL^{-}、SO_4^{2-}),采用离子加和法计算土壤中的盐分含量,3次取样的盐分含量均值作为该样点的SSC[13]。

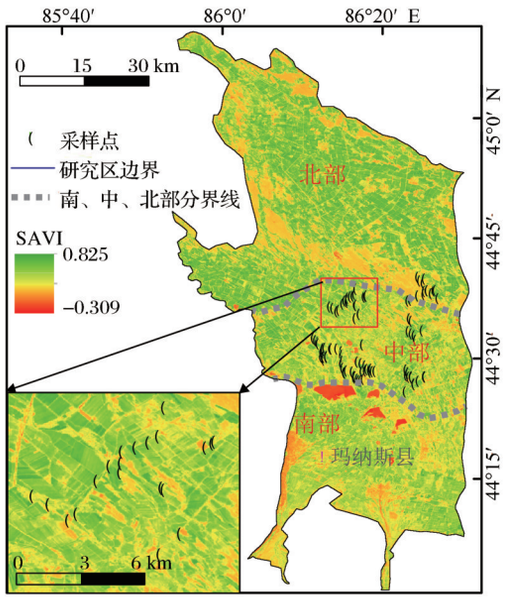


图1　研究区位置及采样点示意图

Fig.1　The location of study area and the distribution of sampling sites

2.2.2　辅助数据

研究采用的辅助数据包括Sentinel-1 SAR、Landsat 8 OLI(144/29)和DEM数据以及SSC实测数据。数据空间分辨率、用途、数据来源及数据获取时间如表1所示。基于Landsat 8 OLI和Sentinel-1 SAR影像获取时间尽可能接近和研究区无云或少云覆盖的原则,选取每月月末影像作为反演生长季SSC的辅助数据。研究区7~8月的Landsat影像云覆盖较多,仅8月11日云覆盖较少,故7~8月仅一期数据。

另外,选取6月24日Landsat 8 OLI和6月28日Sentinel-1 SAR数据以及DEM数据提取环境变量,结合采样数据建立盐渍化监测模型,应用于其他时间的辅助数据,反演研究区生长季SSC分布。为了便于分析,以Landsat 8 OLI影像获取时间为生长季的时间节点。

2.2.3　环境变量提取

环境变量与表层土壤的理化性质息息相关,可作为盐渍化监测的有效辅助信息。通过整合前人的

表 1 实验数据

Table1 Experimental data

数据	空间分辨率/m	用途	来源	影像获取时间
Sentinel-1 SAR影像	10	提取环境变量	欧洲空间局	2016年0417/0522/0628/0815/1002/1026
Landsat 8 OLI影像	30		美国地质调查局	2016年0421/0523/0624/0811/0928/1030
DEM	30		美国国家航空航天局	2000年
SSC实测数据	-	模型的训练与验证	野外实测	2016年7～8月

研究，本研究共提取8类（微波物理量、植被指数、盐分指数、下垫面反射因素、特征空间、缨帽变换因子、地表反射率、地形因素）共40个环境变量作为盐渍化监测的辅助信息（表2）。本研究中，实验数据的获取、处理及环境变量的提取均通过JavaScript编程在谷歌地球引擎（Google Earth Engine，GEE）云平台实现。将提取的环境变量导出至本地，微波物理量采用三次卷积内插法重采样至30 m，以便后续的处理和分析。GEE是专门服务于对地观测数据研究的云平台，已存档了Landsat、Sentinel、Modis等常用遥感数据集，且均以经过预处理可直接用于后续分析，并提供JavaScript和Python语言的API接口，便于研究者进行海量的复杂分析及可视化[14]。

3 研究方法

本研究的总体路线如图2所示，首先基于多源数据提取环境变量，结合相关分析，以显著性（$p<0.05$）为标准初步筛选出候选特征变量（Candidate Feature Variables，CFVs），分别代入PSO、GA、GS 3种算法同步优选建模变量与模型参数，建立研究区的3种盐渍化监测模型（PSO-SVR、GA-SVR、GS-SVR），选择最优的模型反演玛纳斯灌区的生长季SSC分布并分析其时空变化。

3.1 支持向量机与优化算法

3.1.1 支持向量机

支撑向量机（Support Vector Machine，SVM）是以VC维和结构风险最小化为理论基础的机器学习算法。其主要思想是将低维空间线性不可分的数据通过核函数（线性、多项式和径向基函数）映射到高维空间以寻找线性可分的分类面。通过引入不敏感函数ε，SVM在曲线拟合中得以应用并发展为支持向量回归（Support Vector Regression，SVR）。因径向基函数需要优化的参数少且实际应用效果较好[1,7,20]，故本研究选取径向基函数（公式（1））作为SVR的核函数。此外，惩罚参数C和核函数参数γ对模型精度影响较大，需对其进行优选。

$$K(x,x') = \exp\left(-\gamma\|x-x\|^2\right) \tag{1}$$

其中：x'代表支持向量；x代表特征空间；γ表示核函数的宽度，γ越小，支持向量越多。

3.1.2 遗传算法

遗传算法是由霍兰德教授提出，以自然选择和遗传变异为理论基础的概率优化算法[20]。在迭代优化时，需要将问题的每一个候选解进行编码，通常采用二进制的0、1表示，为1时表示该候选解被选中，所有候选解组合在一起构成染色体（也称为个体）。通过对个体结构的随机初始化，依据构造的适应度函数，结合遗传操作（自然选择、交叉、变异）实现问题的优化[7]。其中，个体的适应度值越大，该个体被保留的概率越大，参与交叉和变异的概率越小。

3.1.3 粒子群算法

粒子群算法（Particle Swarm Optimization，PSO）源于对鸟群觅食行为的研究，最早由Kennedy和Eberhart提出[9]。与遗传算法使用相似，需要将问题的所有候选解进行编码，经随机初始化，结合适应度函数，使算法在迭代过程中向最优解方向聚合[21]。在迭代过程中，区别于遗传算法，粒子群算法主要更新每个粒子（也称为个体）的当前位置、速度和历史最优位置以及粒子群的最优位置。如在一个n维空间的寻优问题，第k次迭代，粒子群的位置为$X_k=(X_{k1}, X_{k2}\cdots, X_{kn})^T$，速度为$V_k=(V_{k1}, V_{k2}\cdots, V_{kn})^T$，第k+1次迭代，个体的速度与位置更新如公式（2）～（3）所示。

$$V_{id}(k+1)=w\times V_{id}(k)+c_1\times r_1\times\left(p_{best}-X_{id}(k+1)\right)+ c_2\times r_2\times(g_{best}-X_{id}(k)) \tag{2}$$

$$X_{id}(k+1)=X_{id}(k)+V_{id}(k+1) \tag{3}$$

其中：p_{best}为该个体的历史最优位置；g_{best}为粒子群的最优位置；id为粒子群中的第id个个体；w为惯性变量，调整收敛方向；c1和c2分别为自身和社会学习率，调节迭代的步长；$r1$和$r2$为0～1的互相独立的随机数。

使用粒子群算法优选模型参数和建模变量时，个体的速度更新公式不变，位置更新需要将当前个体的速度采用sigmoid函数（公式（4））转换至0～1

表 2　基于 Sentinel-1 SAR、Landsat 8 OLI、DEM 衍生的环境变量

Table2　Environmental factors derived from Sentinel-1 SAR, Landsat 8 OLI images and DEM data

类别	名称	公式	参考文献
微波物理量植被指数	后向散射系数(BC)		
	归一化植被指数(NDVI)	(NIR－R)/(NIR＋R)	[6]
	扩展的归一化植被指数(ENDVI)	(NIR＋$SWIR_{b2}$－R)/(NIR＋$SWIR_{b2}$＋R	[15]
	增强植被指数(EVI)	2.5×(NIR－R)/(NIR＋6×R－7.5×B＋1)	[6]
	扩展的增强植被指数(EEVI)	2.5×(NIR＋$SWIR_{b1}$)/(NIR＋$SWIR_{b1}$＋6×R－7.5×B＋1)	[15]
	土壤调节植被指数(SAVI)	(1＋L)×(NIR－R)/(NIR＋R＋L)	[6]
	修改型土壤调节植被指数(MSAVI)		[16]
	差值植被指数(DVI)	NIR－R	[16]
	比值植被指数(RVI)	NIR/R	[16]
	大气阻抗植被指数(ARVI)	(NIR－(2×R－B))/(NIR＋(2×R－B))	[16]
	广义差分植被指数(GDVI)	(NIR^2－R^2)/(NIR^2＋R^2)	[6]
	非线性植被指数(NLI)	(NIR^2－R)/(NIR^2＋R)	[6]
	绿色大气阻抗指数(GARI)	(NIR－(G＋γ×(B－R)))/(NIR＋(G＋γ×(B－R)))	[6]
盐分指数	盐分指数(SI)	$\sqrt{B\times R}$	[6]
	盐分指数1(SI1)	$\sqrt{G\times R}$	[6]
	盐分指数2(SI2)	$\sqrt{R^2+G^2+NIR^2}$	[6]
	盐分指数3(SI3)	$\sqrt{G^2+R^2}$	[6]
	盐分指数(S1)	B/R	[6]
	盐分指数(S2)	(B－R)/(B＋R)	[6]
	盐分指数(S3)	G×R/B	[6]
	盐分指数(S5)	B×R/G	[6]
	盐分指数(S6)	NIR×R/G	[6]
	冠层响应盐分指数(CRSI)	$\sqrt{(NIR\times R-G\times R)/(NIR\times R+G\times R)}$	[6]
下垫面反射因素	短波红外地表反照度(α_{short})	0.36×B＋0.13×R＋0.37×NIR＋0.09×$SWIR_{b1}$＋0.07×$SWIR_{b2}$－0.002	[17]
	可见光地表反照度(α_{vis})	0.44×B＋0.17×G＋0.24×R	[17]
特征空间	植被指数-盐分指数特征空间(NSI)	$\sqrt{(MSAVI-1)^2+SI^2}$	[18]
	植被指数-湿度指数特征空间(NWI)	$\sqrt{(MSAVI-1)^2+(WI-1)^2}$	[18]
	湿度指数-盐分指数特征空间(WSI)	$\sqrt{(1-WI)^2+SI^2}$	[18]
缨帽变换因子	绿度指数(GVI)	－0.16×B－0.28×G－0.49×R＋0.79×NIR＋0.0002×$SWIR_{b1}$－0.14×$SWIR_{b2}$	[19]
	湿度指数(WI)	0.03×B＋0.2×G＋0.31×R＋0.16×NIR－0.68×$SWIR_{b1}$－0.61×$SWIR_{b2}$	[19]
	亮度指数(BI)	0.20×B＋0.42×G＋0.55×R＋0.57×NIR＋0.31×$SWIR_{b1}$＋0.23×$SWIR_{b2}$	[19]
波段反射率	B2/B3/B4/B5/B6/B7	B/G/R/NIR/$SWIR_b$/$SWIR_{b2}$	
地形因素	高程/坡度/地表粗糙度	Elevation/Slope/Roughness	[4]

注:B、G、R、NIR、$SWIR_{b1}$、$SWIR_{b2}$分别为蓝、绿、红、近红外、短波红外1、短波红外2波段的反射率。短波红外1、2的波长范围分别为1.56～1.66 μm、2.10～2.30 μm;L=0.5和γ=0.9是气溶胶和大气相关参数

之间,并对该个体的位置更新(公式(5))。

$$S\left(V_{id}^{k+1}\right)=\frac{1}{1+e^{V_{id}^{k+1}}} \tag{4}$$

$$X_{id}^{k+1}=\begin{cases}1 & rand<S(V_{id}^{k+1})\\ 0 & rand\geqslant S(V_{id}^{k+1})\end{cases} \tag{5}$$

其中:rand表示0～1的随机数。

3.2　SVR的模型参数与建模变量的优选

3.2.1　Pearson 相关分析

本研究采用Pearson相关分析初步筛选出与SSC显著相关($p<0.05$)的环境变量作为CFVs,为了消除CFVs的量纲差异对建模精度的影响,每个候选特征变量均进行离差标准化(公式(6))。

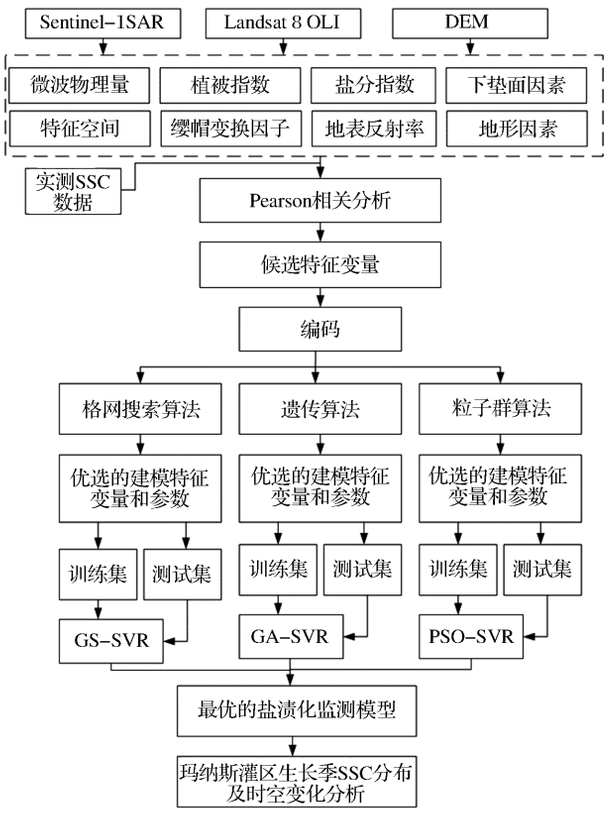


图 2 实验和算法计算流程图

Fig.2 Flow chart of experiment and algorithm calculation

$$CFV_i = \frac{CFV_{i,original} - CFV_{i,min}}{CFV_{i,max} - CFV_{i,min}} \quad (6)$$

其中：CFV_i为第 i 个标准化后的 CFV；$CFV_{i,original}$为第 i 个从 GEE 中计算得到的 CFV；$CFV_{i,min}$，$CFV_{i,max}$分别为第 i 个 CFV 的最小值与最大值。

3.2.2 个体设计及优化算法参数设置

每个个体由模型参数与候选特征变量组成（图 3），其中参数 C 的长度为 20，值域 0～200，间隔为 10；参数 γ 的长度为 10，值域为 0～20，间隔为 2；CFVs 为与 SSC 显著相关的环境变量。遗传算法中，交叉概率 P_c=0.5；变异概率 P_m=0.1。粒子群算法中，惯性变量 w=0.5，自身和社会学习率分别为 2、1。最大迭代次数设置为 50，种群（粒子群）数量设置为 60。

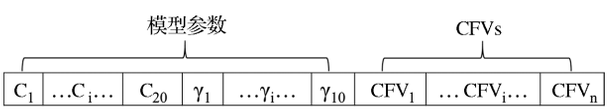


图 3 个体结构组成

Fig.3 The composition of individual structure

3.2.3 模型精度评估及适应度函数设置

基于优选模型参数和建模变量的 SVR 监测玛纳斯灌区生长季 SSC 分布，模型评估采用决定系数（Coefficient of Determination，R^2）和均方根误差（Root Mean Square Error，RMSE）。将采样数据随机分成两部分，75% 的采样数据用于训练模型，25% 用于验证模型精度。以建立的模型在验证数据上的性能（R^2，RMSE）作为优化算法的适应度函数（公式（9））。适应度值越大，R^2 越大，RMSE 越小，模型越优。

$$R^2 = \left(\frac{\sum_{i=1}^{n}(p_i - \bar{p}) \times (o_i - \bar{o})}{\sqrt{\sum_{i=1}^{n}(p_i - \bar{p})^2 + \sum_{i=1}^{n}(o_i - \bar{o})^2}}\right)^2 \quad (7)$$

$$RMSE = \sqrt{\sum_{i=1}^{n}\frac{(p_i - o_i)^2}{n}} \quad (8)$$

$$Fitness = 1\,000 \times \frac{R^2}{RMSE} \quad (9)$$

其中：o_i和 p_i分别为实测和预测的 SSC；$\bar{o}$和 $\bar{p}$分别为实测和预测的 SSC 均值；n 为样点个数。

4 结果与分析

4.1 SSC 与环境变量之间的 Pearson 相关分析及 SSC 的统计特征

为初步筛选出对盐渍化监测有效的环境变量，本研究计算了每个环境变量与 SSC 之间的相关系数及显著性（图 4）。整体而言，除 S6、BI、B5、Elevation、Slope、Roughness 外，其他环境变量均与 SSC 显著相关（$p<0.05$），但 $|r|$ 均小于 0.4 且相差甚微。就不同类型的环境变量而言，SSC 与微波物理量、绝大部分植被指数、缨帽变换因子、地形因素均呈负相关，与盐分指数、下垫面反射因素、特征空间、波段反射率的相关性正负皆有。其中，植被指数和特征空间对 SSC 的敏感性优于其他类型的环境变量，均与 SSC 显著相关（$p<0.05$）。剔除不显著相关的 6 个环境变量，将剩余的 34 个环境变量按 $|r|$ 降序排列组成 CFVs。

由图 5 可知，全部采样数据（2.65～99.50 g·kg^{-1}）均值为 27.83 g·kg^{-1}，变异系数为 81.42%，研究区盐渍化较为严重。全部采样数据、训练集和验证集的 SSC 分布均不均匀但三者分布相似，均呈在 0～40 g·kg^{-1} 分布较为集中，高于 40 g·kg^{-1} 分布相对离散的趋势。训练集（6.62～99.50 g·kg^{-1}）和验证集（2.65～83.65 g·kg^{-1}）均值分别为 28.25 g·kg^{-1}、24.62 g·kg^{-1}，变异系数分别为 80.60%、80.98%。训练集和验证集与全部采样数据的 SSC 分布基本一致，说明基于训练集和验证集的模型训练和验证是可行的[4]。

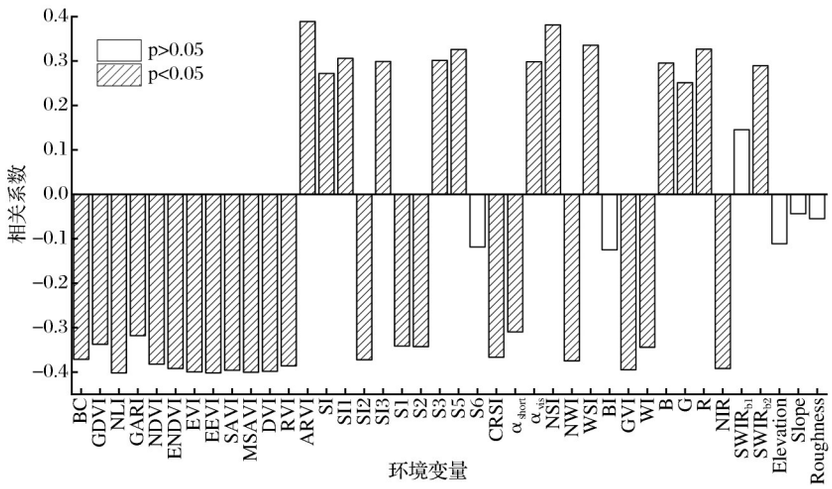
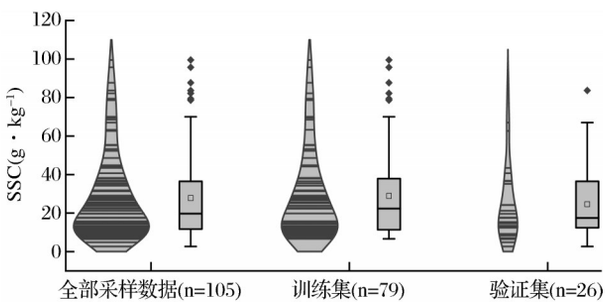


图 4　环境变量与 SSC 的相关系数及显著性

Fig.4　Correlation coefficient and significance between environmental factors and SSC



(图中小提琴图的宽度表示数据分布密度,越宽表示数据分布密度越大,高度表示数据的分布范围)

图 5　土壤样品 SSC 统计特征

Fig.5　Statistical characteristics of SSC

4.2　不同 SSC 估算模型

为了对比不同优化算法在盐渍化建模过程中模型参数和建模变量优选的能力,将 CFVs 分别带入格网搜索(Grid Search, GS)算法、GA、PSO,同步优选 SVR 的模型参数和建模变量,分别建立研究区的盐渍化监测模型(GS-SVR、GA-SVR、PSO-SVR)。其中 GS 算法将全部候选特征变量作为模型的输入,只优化模型参数。不同模型的精度、模型参数和建模变量个数如表 4 所示。不同模型优选的建模变量如图 6 所示,GA-SVR 和 PSO-SVR 的平均、最佳适应度值随迭代次数的变化如图 7 所示。

表 4　不同 SSC 估算模型精度对比

Table 4　Accuracy comparison of different SSC estimation models

方法	训练		验证			模型参数	建模变量个数
	R^2	RMSE/g·kg^{-1}	R^2	RMSE/g·kg^{-1}	Fitness/kg·g^{-1}	C, γ	
GS-SVR	0.78	11.00	0.63	12.30	51.22	90,4	34
GA-SVR	0.74	13.16	0.77	9.77	78.81	110,8	18
PSO-SVR	0.72	12.56	0.80	9.19	87.05	100,10	14

由表 4 可知,相对 GS-SV(R^2=0.63,RMSE=12.30 g·kg^{-1})而言,GA-SVR(R^2=0.77,RMSE=9.77 g·kg^{-1})和 PSO-SVR(R^2=0.80,RMSE=9.19 g·kg^{-1})在减少建模变量的同时,模型精度均有不同程度的提高,适应度值分别提高了 53.87%、69.96%。结合图 6 可知,不同模型的建模参数和建模变量均不同,说明模型参数和建模变量的依赖关系。此外,部分环境变量与 SSC 相关性较好,并未参与模型的构建,这主要是因为 Pearson 相关分析仅能体现 SSC 对单个候选特征变量之间的依赖关系,忽略了候选特征变量对 SSC 监测的协同或抑制作用;而 GA-SVR 和 PSO-SVR 考虑了候选特征变量的组合效应,使部分候选特征变量不参与模型的建立,并且在优选建模变量的同时又优化了模型参数。而相对于 GA-SVR 而言,PSO-SVR 在每次迭代时,都会记录每个个体的当前最优解,并确定是否更新粒子群的历史最优解,且迭代优化均在历史最优解的基础上进行,提高了优化效率,使得 PSO-SVR 收敛较早且效果优于 GA-SVR(图 7);而 GA 迭代优化过程复杂,不具备'记忆'能力,可能会丢失当前最优解的信息。

为了对比 3 种方法反演的 SSC 的局部特征,在

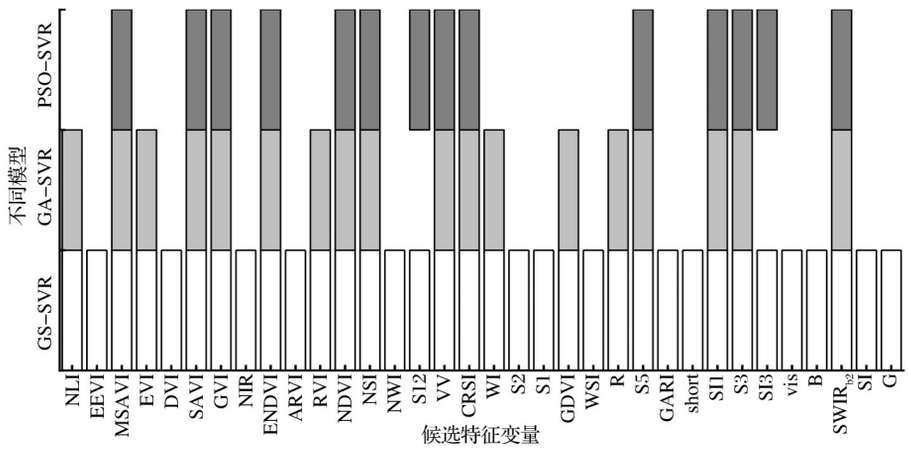


图 6 不同模型选择的建模特征变量

Fig.6 The feature subset selected by different models

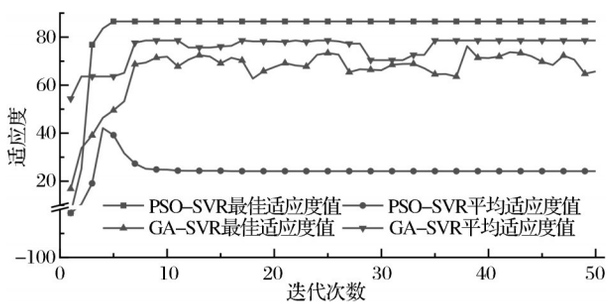


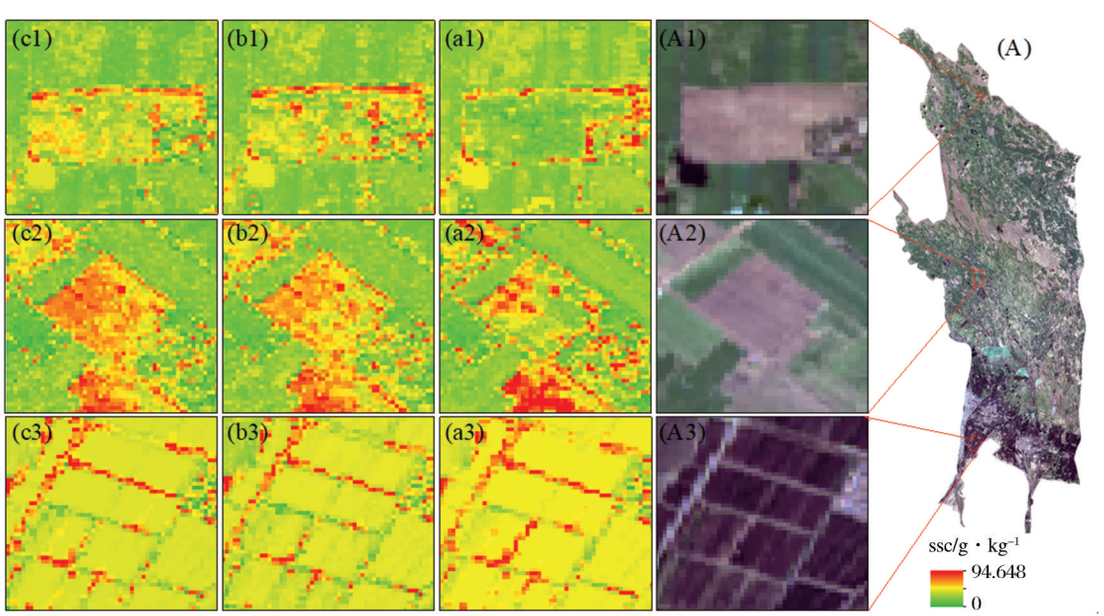
图 7 适应度随迭代优化的变化

Fig.7 Fitness changes with iterative optimization

南、中、北部各选一小子区(图 8)。对比可知,基于 PSO-SVR 和 GA-SVR 反演的 SSC 分布地物轮廓更为清晰,且同地物 SSC 分布的均质性更好。说明 PSO 和 GA 在盐渍化监测中的模型参数和建模变量优选是可行的,又因 PSO-SVR 的精度较优,故选取 PSO-SVR 模型反演玛纳斯灌区 2016 年生长季 SSC 分布。

4.3 玛纳斯灌区生长季 SSC 分布

随时间推移,SSC 分布变化明显(图 9(a)~(f);4 月末,SSC 较高且分布较为集中,至 8 月中旬,绿洲区 SSC 趋于下降,绿洲边缘 SSC 趋于增加,而后绿洲区 SSC 趋于增加,绿洲区边缘 SSC 趋于减少,且 SSC 分布较为离散。就不同区域而言,玛纳斯灌区整个生长季 SSC 均值基本呈南部>中部>北部的规律(图 10),这主要是因为研究区为典型的绿洲经济,农业灌溉需水量大。为了满足灌溉需求,南部兴修水库;长期受土壤母质和水库截流的影响,该



(1、2、3 分别代表北、中和南部的小子区;A、a、b、c 分别代表小子区的真彩色合成、GS-SVR、GA-SVR、PSO-SVR 反演的 SSC 分布)

图 8 子区的位置及 SSC 分布

Fig.8 Sub-region location and SSC distribution

区地下水位偏高,盐分表聚普遍[10],而中部处于地下水溢出带和水库下方,地下径流缓慢,加之水库下渗,抬升了地下水位,强烈的蒸发导致土壤盐分积累[12]。因枯水期河流经常发生断流,北部水资源主要来源于农业灌溉,地下水位较低,盐分表聚相对不明显。

就生长季而言,SSC呈明显的春季(3~5月)、秋季(9~11月)积盐,夏季(6~8月)脱盐的趋势,中、南部整个生长季SSC均值变化趋势与整个研究区类似:即降低—增加;北部整个生长季SSC均值变化趋势为增加—降低—增加(图10)。这主要与灌溉脱盐和蒸发积盐的博弈强弱相关。南部农作物主要以春玉米和棉花为主,中、北部主要以棉花为主。春季4月中下旬作物开始播种,至春季5月末,作物需水较少,除播前灌外,灌溉次数少且量小,而中、南部受冰雪融水的持续供给,脱盐效果较北部显著,而北部地形相对平坦开阔且与沙漠接壤,蒸发强于中部和南部,盐分积累较多,SSC均值在5月末高于中、南部;在夏季6月至8月中旬,春玉米拔节、抽雄,棉花现蕾、开花、结玲,需水均较多,灌溉淋洗量高于蒸发返盐量,SSC均值呈明显的下降趋势,而北部砂质土壤含量较大,灌溉对土壤盐分的淋溶作用优于中、南部,使得SSC均值在6月末骤降且低于中、南部,直到8月中旬,灌溉淋洗量基本与蒸发返盐量持平;秋季灌溉较少,蒸发返盐量多于灌溉淋洗量,SSC均值呈逐渐增加的趋势,而北部蒸发较强,盐分表聚的累积效应使得SSC均值在10月末再次超越中部。

图11反映了SSC的分布密度和变异程度。整个研究区,南、中、北部SSC数值分布均存在明显的离散值。整个研究区与中、北部的SSC数值分布相似,SSC数值区间呈扩张—收缩—扩张的趋势且SSC分布呈高值—低值—高值聚集。南部SSC数值区间呈扩张—收缩—平稳的趋势,不同阶段SSC分布均主要为20~30 g·kg^{-1},但其他区间的SSC分布随季节变化较大。这主要是因为研究区包含部分非绿洲区,春夏季降水稀少,蒸散发强烈,盐分持续表聚,提高了SSC的上限,而灌溉洗盐使得绿洲区SSC降低且趋于低值聚集;而8月中旬之后,蒸发强烈,灌溉较少,聚盐效果明显,使得SSC趋于高值聚集,而该时段降水相对较多,淋溶作用降低了非绿洲区的SSC,缩小了SSC的数值区间;进入10月,作物基本成熟,地表裸露,蒸发提高了SSC数值的上限和聚集程度。而南部地下水位偏高,盐渍化较为严重,虽然生长季内SSC变化明显,SSC在20~30 g·kg^{-1}仍分布较多;步入10月,蒸发的积盐效应显著,抬升了SSC数值的上下限,使得SSC的数值区间变化不大但SSC分布更为聚集。

5　结果与讨论

5.1　结　果

本文基于多源数据提取环境变量,经相关分析,分别代入GS、GA、PSO 3种算法,优选了SVR的模型参数和建模变量并分别建立了盐渍化监测模型,选择精度最优的模型反演了玛纳斯灌区2016年生长季SSC并分析了其时空变化。结果表明:

(1)提取的环境变量除S6、BI、B5、Elevation、Slope、Roughness外,均与SSC显著相关($p<0.05$),且植被指数和特征空间对SSC更为敏感。

(2)相对GS-SVR(R^2=0.63,RMSE=12.30 g·kg^{-1})而言,GA-SVR(R^2=0.77,RMSE=9.77 g·kg^{-1})和PSO-SVR(R^2=0.80,RMSE=9.19 g·kg^{-1})

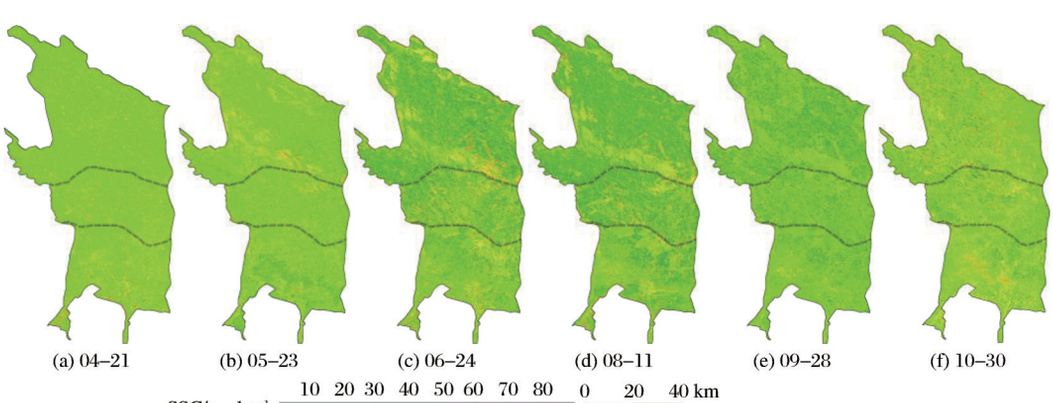


图9　玛纳斯灌区2016年生长季SSC空间分布

Fig.9　Spatial distribution of SSC in the growing season of Manasi Irrigation District in 2016

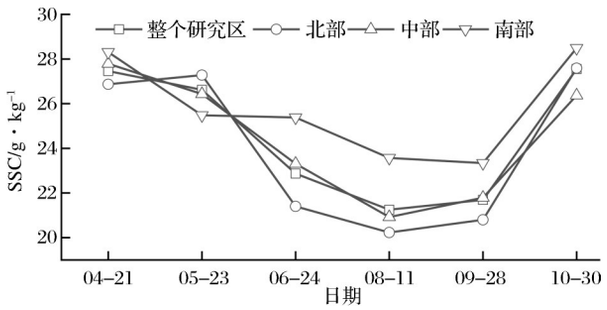


图 10 玛纳斯灌区 2016 年生长季 SSC 均值变化

Fig.10 Change of the average SSC during the growing season in the Manasi Irrigation District, 2016

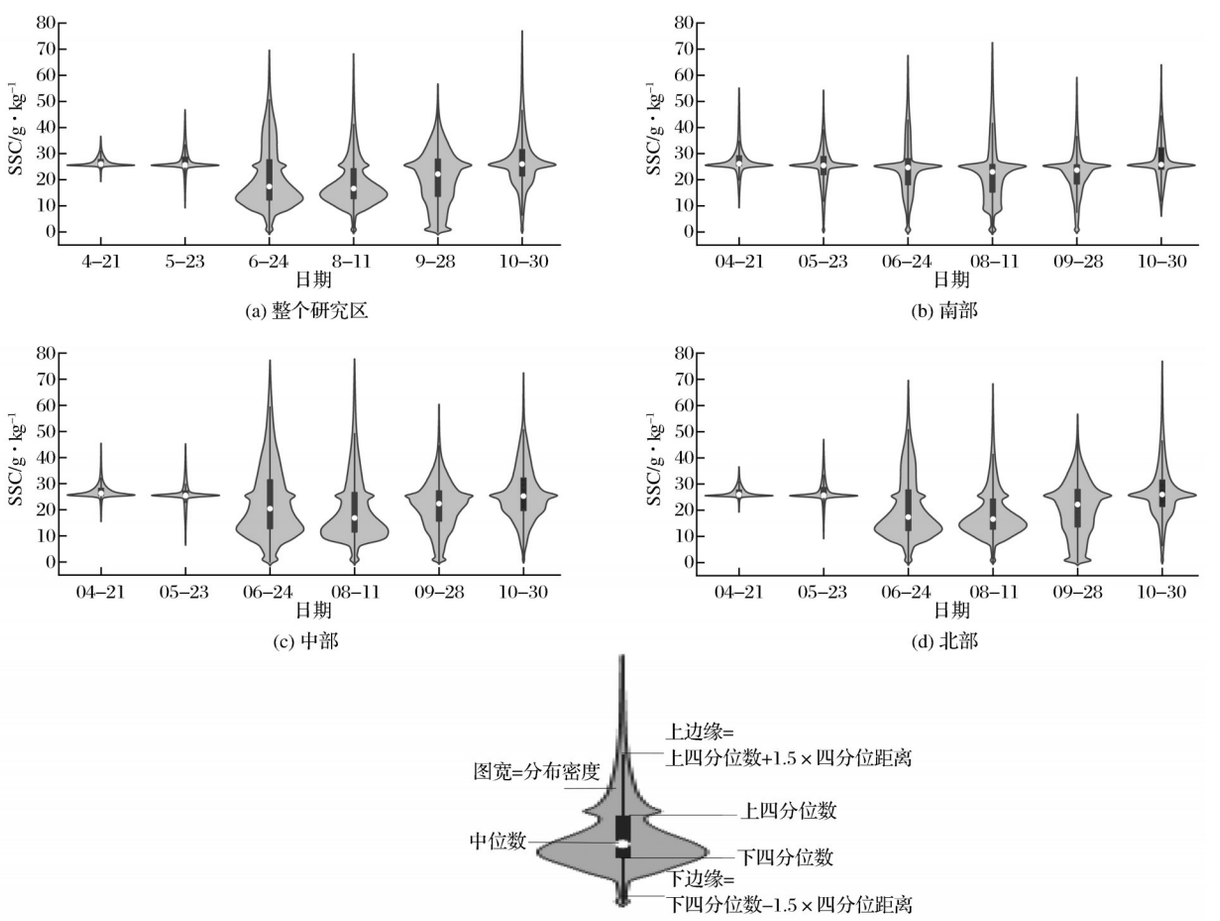
在减少建模变量的同时，提高了模型精度，适应度值分别提高了 53.87%、69.96%。

(3)生长季内，灌区 SSC 呈春、秋季积盐，夏季脱盐的季节变化；SSC 均值变化趋势：中、南部和整个研究区为降低—增加；北部为增加—降低—增加。

(4)整个研究区与中、北部的 SSC 数值分布相似，SSC 数值区间呈扩张—收缩—扩张的趋势且 SSC 分布呈高值—低值—高值聚集；南部 SSC 数值区间呈扩张—收缩—平稳的趋势，不同阶段 SSC 分布均主要为 20～30g·kg^{-1}，但其他区间的 SSC 值分布呈明显的季节变化。

5.2 讨 论

环境变量对盐渍化的敏感性受采样点分布及时间、植被覆盖、土壤母质、影像获取时间等众多因素的影响[6]。本研究的采样点大多分布于中部，地形起伏小，地形因素均没有表现出很好的盐渍化监测能力。雷达穿透能力强，辐射传输过程中信息损失较少，可以较好地反映真实的地表状况，后向散射系数(BC)表现出较好的盐渍化监测能力，说明雷达数据在盐渍化监测方面的应用潜力[22-23]。波段反射率除 $SWIR_{b1}$ 外，均与 SSC 显著相关；可见光(B、G、R)和近红外(NIR)波段对土壤盐渍化的响应较为敏感，SSC 与可见光波段的反射率呈负相关，与近红外波段呈正相关；这主要是因为植被光合作用



(四分位距离为上下四分位数之差，上下边缘的距离表示数据的分布区间，距离越大表示数值分布越离散，越小表示数值分布越集中)

图 11 2016 年玛纳斯灌区生长季 SSC 小提琴图

Fig.11 SSC violin plots during the growing season in Manasi Irrigation District, 2016

吸收叶绿素,在可见光波段的反射率低,而在近红外波段的反射率高;受盐分胁迫的植被,光合作用减弱,在可见光波段的反射能力增强,而在近红外波段的反射能力减弱[6]。其他类型的环境变量均由波段运算得到,相对而言,植被指数对盐渍化的监测能力优于其他类型的环境变量,这与王飞等[15]的研究一致。采样数据大多分布于植被长势较好的中部,裸露地表少,水分状况良好,植被指数、盐分指数、GVI、WI均表现出较好的盐渍化监测能力,而*BI*对盐渍化的响应较弱。可见光地表反照度(α_{vis})随盐渍化的加重呈增加趋势,这与Wang等[24]的研究一致。而复合型指数NSI、NWI、WSI通过整合不同类型的环境变量对盐渍化的响应,间接提供了盐渍化地表的信息[2,18]。

参考文献(References):

[1] Nurmemet I, Sagan V, Ding J L, *et al*. A WFS-SVM Model for Soil Salinity Mapping in Keriya Oasis, Northwestern China Using Polarimetric Decomposition and Fully PolSAR Data [J]. Remote Sensing, 2018, 10(4): 598. doi: 10.3390/rs10040598.

[2] Guo B, Yang F, Fan Y W, *et al*. Dynamic Monitoring of Soil Salinization in Yellow River Delta Utilizing MSAVI - SI Feature Space Models with Landsat Images[J]. Environmental Earth Sciences, 2019, 78(10): 308. doi: 10.1007/s12665-019-8319-8.

[3] Adisa A, Hassani A, Shokri N. Predicting Long-term Dynamics of Soil Salinity and Sodicity on a Global Scale[J]. Proceedings of the National Academy of Sciences, 2020, (Accepted/In press). doi: 10.1073/pnas.2013771117.

[4] Vermeulen D, Niekerk A V. Machine Learning Performance for Predicting Soil Salinity Using Different Combinations of Geomorphometric Covariates[J]. Geoderma, 2017, 299: 1-12. doi: 10.1016/j.geoderma.2017.03.013.

[5] Peng J, Biswas A, Jiang Q S, *et al*. Estimating Soil Salinity from Remote Sensing and Terrain Data in Southern Xinjiang Province, China[J]. Geoderma, 2019, 337: 1309-1319. doi: 10.1016/j.geoderma.2018.08.006.

[6] Xu H T, Chen C B, Zheng H W, *et al*. AGA-SVR-based Selection of Feature Subsets and Optimization of Parameter in Regional Soil Salinization Monitoring[J]. International of Remote Sensing, 2020, 41(12): 4470-4495. doi: 10.1080/01431161.2020.1718239.

[7] Zhou T, Lu H L, Wang W W, *et al*. GA-SVM based Feature Selection and Parameter Optimization in Hospitalization Expense Modeling [J]. Applied Soft Computing Journal, 2019, 75: 323-332. doi: 10.1016/j.asoc.2018.11.001.

[8] Wang Fei, Yang Shengtian, Ding Jianli, *et al*. Environmental Sensitive Variable Optimization and Machine Learning Algorithm Using in Soil Salt Prediction at Oasis[J]. Transactions of the Chinese Society of Agricultural Engineering, 2018, 34(22): 102-110.[王飞, 杨胜天, 丁建丽, 等. 环境敏感变量优选及机器学习算法预测绿洲土壤盐分[J]. 农业工程学报, 2018, 34(22): 102-110.]

[9] Tan Lin, He Bingyu, Liu Weiguo, *et al*. Estimation of Chlorophyll Content Eremurus Chinensis based on Optimization Support Vector Machine[J]. Chinese Journal of Ecology, 2017, 36(2): 555-562.[谭林, 何秉宇, 刘卫国, 等. 基于优化SVR高光谱指数的独尾草叶绿素含量估算[J]. 生态学杂志, 2017, 36(2): 555-562.]

[10] Yang H C, Chen Y, Zhang F H. Evaluation of Comprehensive Improvement for Mild and Moderate Soil Salinization in Arid Zone[J]. PloS One, 2019, 14(11): e0224790. doi: 10.1371/journal.pone.0224790.

[11] Zhang T Y, Wang L, and Han Y. Evaluating the Sensitivity of Ecosystems to Soil Salinization in the Manasi River Basin [J]. Polish Journal of Environmental Studies, 2017, 26(2): 917-924. doi: 10.15244/pjoes/65836.

[12] Lü Nana, Luo Geping, Ding Jianli, *et al*. Spatio-temporal Variation of Soil Salinity in Wastelands inside and Outside of Oasis in Manas River Watershed in the Context of Dripping Irrigation[J]. Journal of Natural Resources, 2017, 32(9): 1542-1553.[吕娜娜, 罗格平, 丁建丽, 等. 滴灌背景下玛纳斯流域绿洲内外荒地土壤盐分时空变化趋势分析[J]. 自然资源学报, 2017, 32(9): 1542-1553.]

[13] Han Yan. Dynamic Monitoring and Risk Assessment of Soil Salinization in Manasi River Basin, Xinjiang[D]. Shihezi: Shihezi University, 2018.[韩燕, 新疆玛纳斯河流域土壤盐渍化动态监测及风险性评价[D]. 石河子: 石河子大学, 2018.]

[14] Gorelick N, Hancher M, Dixon M, *et al*. Google Earth Engine: Planetary-scale Geospatial Analysis for Everyone[J]. Remote Sensing of Environment, 2017, 202: 18-27. doi: 10.1016/j.rse.2017.06.031.

[15] Wang Fei, Ding Jianli, Wei Yang, *et al*. Sensitivity Analysis of Soil Salinity and Vegetation Indices to Detect Soil Salinity Variation by Using Landsat Series Images: Applications in Different Oases in Xinjiang, China[J]. Acta Ecoligica Sinica, 2017, 37(15): 5007-5022.[王飞, 丁建丽, 魏阳, 等. 基于Landsat系列数据的盐分指数和植被指数对土壤盐度变异性的响应分析——以新疆天山南北典型绿洲为例[J]. 生态学报, 2017, 37(15): 5007-5022.]

[16] Guo S S, Ruan B Q, Chen H R, *et al*. Characterizing the Spatiotemporal Evolution of Soil Salinization in Hetao Irrigation District (China) Using a Remote Sensing Approach[J]. International Journal of Remote Sensing, 2018, 39(20): 6805-6825. doi: 10.1080/01431161.2018.1466076.

[17] Liang S L. Narrowband to Broadband Conversions of Land Surface Albedo Algorithms[J]. Remote Sensing of Environment, 2001, 76(2): 213-238. doi: 10.1016/S0034-4257(00)00205-4.

[18] Li Yanhua, Ding Jianli, Sun Yongmeng, *et al*, Remote Sensing Monitoring Models of Soil Salinization based on the Three-Dimensional Feature Space of MSAVI-WI-SI[J]. Research of Soil and Water Conservation, 2015, 22(4): 113-117.[李艳华, 丁建丽, 孙永猛, 等. 基于三维特征空间的土壤盐渍化遥感模型[J]. 水土保持学报, 2015, 22(4): 113-117.]

[19] Crist E P. A TM Tasseled Cap Equivalent Transformation for Reflectance Factor Data [J]. Remote Sensing of Environment, 1985, 17(3): 301-306. doi: 10.1016/0034-4257(85)90102-6.

[20] Sukawattanavijit C, Chen J, Zhang H S. GA-SVM Algorithm for Improving Land-Cover Classification Using SAR and Optical Remote Sensing Data[J]. IEEE Geoscience and Remote Sensing Letters, 2017, 14 (3): 284-288. doi: 10.1109/LGRS.2016.2628406.

[21] Elahe A, Ali D B, Najmeh N S, *et al*. Crop Mapping Using Random Forest and Particle Swarm Optimization based on Multi-Temporal Sentinel-2[J]. Remote Sensing, 2020, 12(9): 1449. doi: 10.3390/rs12091449.

[22] Mahdi T M, Mahdi H, Kamran E. Soil Salinity Mapping Using Dual-polarized SAR Sentinel-1 Imagery[J]. International Journal of Remote Sensing, 2018, 40(1): 237-252. doi: 10.1080/01431161.2018.1512767.

[23] Yang R M, Guo W W. Using Sentinel-1 Imagery for Soil Salinity Prediction Under the Condition of Coastal Restoration [J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2019, (99): 1-7. doi: 10.1109/JSTARS.2019.2906064.

[24] Wang H S, Jia G S. Satellite-based Monitoring of Decadal Soil Salinization and Climate Effects in a Semi-arid Region of China[J]. Advances in Atmospheric Sciences, 2012, 29(5): 1089-1099. doi: 10.1007/s00376-012-1150-8.

SVR Salinization Monitoring based on Integrated Feature Subset Selection and Model Parameter Learning

Xu Hongtao[1,2], Chen Chunbo[1,2], Zheng Hongwei[1,2], Luo Geping[1,2], Yang Liao[1,2], Wang Weisheng[1,2], Wu Shixin[1,2]

(1.*State Key Laboratory of Desert and Oasis Ecology, Xinjiang Institute of Ecology and Geography, Chinese Academy of Sciences, Urumqi* 830011, *China*;
2.*University of Chinese Academy of Sciences, Beijing* 100049, *China*)

Abstract: Salinization is one of the main forms of land degradation which leads to fragile ecological environment and low efficiency of agricultural production. Remote sensing combined with machine learning algorithm is one the most popular methods in salinization monitoring. In terms of machine learning algorithm, the model feature subset and parameters is vital to modeling accuracy. Therefore, accurate identification and optimization of model parameters and feature subset is crucial for machine learning based inversion and prediction of Soil Salt Content (SSC). Based on Sentinel-1 SAR, Landsat 8 OLI images and DEM data, a total of 40 environmental factors of 8 categories were extracted. In conjunction with Pearson correlation analysis, the Candidate Feature Variables (CFVs) were initially selected. The CFVs were introduced into the Grid Search (GS) algorithm, Genetic Algorithm (GA) and the Particle Swarm Optimization (PSO) to simultaneously identify the model parameter and feature subset of Support Vector Regression (SVR). Salinization monitoring models (GS-SVR, GA-SVR, PSO-SVR) were established, respectively. The optimal model was applied into the SSC prediction of Manasi Irrigation District in growing season, 2016. The results show that the extracted environmental factors showed good correlations with SSC, and the vegetation indices and feature spaces were more sensitive to salinization than other types of environmental factors. Compared with GS-SVR, the GA-SVR and PSO-SVR methods improved the accuracy of the salinization monitoring while reducing the number of feature subset, and the fitness value increased by 53.87% and 69.96%, respectively. During the growing season, salt accumulates in spring and autumn and fades in summer. The trend of average SSC of the whole study area and the central part and the southern part was decreasing-increasing, while the northern part was increasing-decreasing-increasing. According to the SSC violin plots in the growing season, it was found that the trend of SSC range of the whole study area and the central part and the northern part was expansion-contraction-expansion, while it was expansion-contraction-stability in southern part of study area. This study provided the technical support for accurate salinization monitoring and dynamic change of SSC in growing season.

Key words: Genetic Algorithm; Particle Swarm Optimization; Soil Salinization; Support Vector Machine; Model parameters and feature subset selection