高寒湿地分类的遥感特征优选研究

霍轩琳[1,2]，牛振国[2]，张波[2]，刘林崧[2]，李霞[3]

1. 长安大学 地球科学与资源学院, 西安 710054;
2. 中国科学院空天信息创新研究院 遥感科学国家重点实验室, 北京 100094;
3. 长安大学 土地工程学院, 西安 710054

摘 要：高寒湿地是青藏高原重要的地表覆盖类型之一，对于水源涵养、调节气候、维护生物多样性等起着关键作用，准确及时获知高寒湿地的时空分布对于湿地的保护和管理十分必要。遥感分类特征优选对湿地制图具有关键性的作用。虽然像光谱特征、纹理特征、地形特征等均在已有研究中有涉及，但鲜有研究聚焦光谱指数特征，深入探讨其数理统计特征和特征优选方法。本研究以甘肃首曲高寒湿地保护区为研究区，基于Sentinel-2数据得到各分类特征（光谱、植被指数、红边指数和水体指数），采用Filter和Wrapper特征选择方法包括Jeffries-Matusita距离、光谱角距离（SAD）、欧氏距离（ED）、RF-RFE算法和Relief-F算法对上述特征进行优选，并利用Filter方法的Z检验进行量化评价。研究表明：（1）所有参与分类的类别中，河流与裸地最容易区分，其次为草原与沼泽，沼泽化草甸与草甸较为难分。对沼泽、沼泽化草甸、草甸、草原邻近两类可尝试MCARI2、NDWI、DVI、EVI、EWI、IRECI、MCARI、TCARI、UGWI指数进行区分；（2）就不同指数特征对湿地信息提取贡献程度而言，水体指数特征>植被指数特征>红边指数特征；（3）从特征优选方法角度看，Filter方法中的ED距离算法与Relief-F算法表现突出；（4）最终选出适于高寒湿地信息提取的指数有RDVI、NDVI、MSR、RVI、VIgreen、RNDWI、NDWI、NDWI_B、MNDWI、EWI、CIre；（5）从不同分类特征的数理统计指标看，中值特征的分类结果最好，其次是平均值特征。本研究为湿地信息提取在特征变量优选方面提供了一种可迁移且普适性高的方法和思路。

关键词：遥感，湿地分类，高寒湿地，特征优选，青藏高原，Sentinel-2

中图分类号：P2

引用格式：霍轩琳,牛振国,张波,刘林崧,李霞.2023.高寒湿地分类的遥感特征优选研究.遥感学报,27(4): 1045-1060

Huo X L, Niu Z G, Zhang B, Liu L S and Li X. 2023. Remote sensing feature selection for alpine wetland classification. National Remote Sensing Bulletin, 27(4):1045-1060[DOI:10.11834/jrs.20222080]

1 引 言

湿地具有调节气候、保护生物多样性、蓄洪抗旱、改善环境等功能，同时为动植物提供了良好的生存条件（Moor等，2015），是地球上最重要的生态系统（森林、海洋、湿地）之一（何菊红等，2015）。青藏高原高寒湿地是青藏高原乃至西部地区最重要的生态系统，近年来受到自然和人为因素干扰，高寒湿地面积已锐减了10%，且水量和湿地的面积减少速度还在加快（王根绪 等，2007；徐新良 等，2008）。因此，及时获知高寒湿地面积、分布区域等对青藏高原高寒湿地管理与保护乃至生态系统的可持续发展至关重要。

遥感分类特征的选择是目前对湿地进行准确分类与制图的众多挑战之一。分类特征需要考虑到以下两方面，其一，湿地是陆地生态系统和水域生态系统的交界地带，具有较高的景观异质性，仅靠一种特征变量可能无法很好地进行湿地的准确提取；其二，若使用过多的特征变量参与其中将影响分类精度和效率。可见，多特征变量提取与优化以及进行有效组合将是今后湿地信息智能化提取的重点难点（张磊 等，2019）。特征选择

收稿日期：2022-03-13；**预印本：**2022-07-12

基金项目：国家重点研发计划“政府间国际科技创新合作”重点专项(编号:2021YFE0194700);第二次青藏高原综合科学考察研究项目(编号:2019QZKK0106);国家自然科学基金(编号:41971390)

第一作者简介：霍轩琳，研究方向为湿地生态与环境遥感。E-mail：h1160964174@gmail.com

通信作者简介：牛振国,研究方向为湿地生态与环境遥感。E-mail:Niuzg@aircas.ac.cn

（Feature Selection）通常情况下是将特征按照相关性准则排序，去掉冗余和不相关的特征（Guyon和Elisseeff，2003），按评价标准不同，特征选择算法可分为过滤式（Filter）、封装式（Wrapper）和嵌入式（Embedded）3种（Dash和Liu，1997；Dash等，2002；Saeys等，2007；Kira和Rendell，1992）。其中，Filter方法利用特征本身内在特性给出特征评价，特征评价越高表示该特征区分能力越强。这也是该方法最主要的特点：特征选择独立于分类学习算法。它不依赖某种分类器，因此简单，速度快，效率高。Yu和Liu（2003）以特征间的相关性为指导，用Filter方法进行特征选择，证实了该方法进行特征选择的有效性；Wrapper方法是将特定学习算法性能作为筛选子集的评估准则，每次筛选出的特征子集都需调用特定分类器进行精度验证。John等（1994）提出当满足一定条件时，将获得具有较高分类性能的识别模型。该方法准确率较高，利于关键特征的识别，在算法速度上比Filter方法慢，时间复杂度较高；Embedded方法是一种基本的归纳方法，可以说是Wrapper方法的延伸。Embedded方法将特征选择过程嵌入到分类器的建造过程中，主要的例子是套索回归的问题以及决策树如Breiman（2001）的CART算法。该方法计算效率高，但是特征中可能存在无关特征降低分类精度，在本次研究中未涉及。

针对已有的特征选择方法，有一些学者就其在湿地遥感分类方面进行了探索。如Mahdianpari等（2019）利用JM距离定量地确定不同类型湿地的可分性，结合随机森林进行分类总体精度达到88.37%；孙艳丽等（2015）利用光谱角距离（SAD）和欧氏距离（ED）双重判定提取不变特征点，提出了一种基于光谱角—欧氏距离的辐射归一化方法；郝玉峰等（2021）利用Relief-F算法计算了52个特征变量的权重，选出前20个特征变量构成最优特征集参与湿地信息提取；Han等（2012）利用Z检验方法测定区分两种植被类型的最佳纹理波段。以上提到的这些方法均属于Filter方法；解淑毓等（2021）采用Wrapper方法中典型的RFE算法进行沼泽湿地分类中的变量优选，显著减少了数据冗余。Phan等（2020）在其研究中指出GEE提供的像元级重组规则主要包括最大值、最小值、平均值、中位数和百分位数等。

总体而言，目前湿地分类特征优选的研究多集中于通过多特征变量参与、单一特征优选方法来甄选最优特征集。不同特征的统计方式和不同特征优选方法对分类的影响尚未见相关研究报道，同时不同分类特征对高寒湿地类别分类的适用性也未见相关研究。鉴于此，本文基于Sentinel-2影像数据，以首曲高寒湿地保护区为研究区域，利用随机森林分类算法，探讨数理统计特征和特征优选方法对优选的各种影响，并分析不同特征对高寒湿地类型分类的适用性。该研究将对提高高寒湿地遥感制图具有重要的参考价值。

2 研究区概况与数据源

2.1 研究区概况

甘肃黄河首曲国家级自然保护区（33°20′01″N—33°56′31″N，101°54′12″E—102°28′45″E）位于甘南藏族自治州玛曲县境内，属于内陆湿地和水域生态系统类型的自然保护区，是青藏高原典型的面积较大的高寒湿地（薛鹏飞等，2021），也是全球保存状态最为完整和原始的湿地。首曲高寒湿地属于高原大陆性气候，年均气温1.1 ℃，年平均降水量615.5 mm，全年降雨150 d左右（高斌斌，2008），给黄河贡献了黄河源区总径流量58.7%的水量，被称为“蓄水池”和“高原水塔”。

2.2 数据及预处理

2.2.1 数据源及预处理

近年来随着遥感技术的迅猛发展，越来越多的多源传感器涌现，其空间分辨率、时间分辨率、波段数量得到了巨大提升（郑阳等，2017），为湿地的遥感分类提供了更多的选择。综合考虑影像分辨率、波段、可获得性等多因素，研究采用Sentinel-2影像。Sentinel-2属于中高空间分辨率遥感影像，携带高分辨率多光谱传感器MSI，可提供可见光、近红外到短波红外的13个波段，是目前唯一在植被光谱的红边区域（670—760 m）设置3个波段的卫星。红边波段数据及其衍生指数可区分C3、C4植被（Shoko和Mutanga，2017；Korhonen等，2017；常文涛等，2020），极大促进了对植被生长信息及其健康状况的有效监测（张磊等，2019）。本研究使用的Sentinel-2数据是用GEE平台

“COPERNICUS/S2_SR”数据集中2020年1月1日至2020年12月31日的影像数据，采用已经过辐射定标和几何校正的Level-1C产品，去除云覆盖率大于10%的影像后得到34景Sentinel-2的无云影像。

2.2.2 样本数据

本研究以Global Lakes and Wetlands Database、Wetland Dataset of CAS湿地制图产品公开数据集为参考数据集，在Google Earth软件上开展样本集目视解译工作。最终取得样本点共480个，每个地类（分类体系见表1）各80个（图1）。

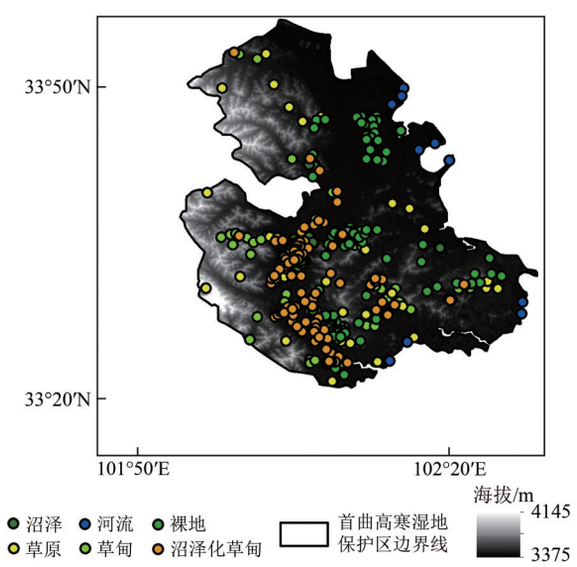


图1 黄河首曲国家级自然保护区样本点分布图

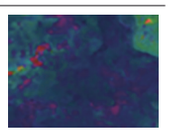
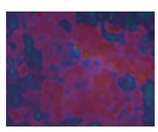
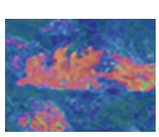
Fig.1 Distribution map of sample points in the Yellow River Shouqu National Nature Reserve

2.3 分类体系方案

White等（2020）指出，草甸沼泽对水位变化特别敏感，下垫面含水量一定程度上决定了湿地植被构成。本次不直接以湿地所在地域或湿地中细分植被为分类标准而是优先考虑造成湿地植被生长的起因——下垫面土壤含水量。随着土壤水分的增减，草原地区的草甸可能会发生演变，当水分增加时，可转变为沼泽；当水分减少时可转变为草原，沼泽化草甸是草甸与沼泽之间的过渡类型。沼泽、沼泽化草甸、草甸区域及边界的变化是监测湿地变化的重要指标，其变化可反映出当地的自然气候，生态环境的变化。因此本文土地覆被分类方案如表1所示。

表1 黄河首曲自然保护区土地覆被分类方案及影像示例

Table 1 Land cover classification scheme and wetland corresponding image of Yellow River Shouqu Nature Reserve

分类	具体说明	Sentinel-2影像举例
沼泽	地表及地表下层土壤经常过度湿润，地表生长着湿性植物和沼泽植物，有泥炭累积或虽无泥炭累积但有潜育层存在的土地	
沼泽化草甸	湿中生多年生草本植物为主的植物群落，为典型草甸向沼泽植被的过渡类型，土壤为暗色草甸土	
草甸	向草原过渡地带的典型草甸，可视为地带性植被。土壤为排水良好，湿度中等的草甸黑土	
草原	以多年生旱生型草本植物占优势的水平或垂直地带性植被类型	
水体	主要包括湖泊：陆地表面洼地积水形成的比较宽广的水域；河流：指地表上有相当大量且常年或季节性流动的天然水流	
裸地	地表有土壤覆盖且植被覆盖度小于5%的土地	

3 研究技术流程和方法

3.1 技术流程

本文的技术路线图如下图2所示。首先对Sentinel-2影像数据进行预处理。基于样本计算32种特征指数（3.2节部分）的统计特征（均值、标准差、中值、最大值、最小值）后，分别利用JM距离、ED距离、SAD距离、RF-RFE算法、Relief-F算法进行遴选，获取不同特征优选方法下的最优特征集。基于不同优选特征集利用随机森林进行湿地分类，依据特征优选方法及随机森林分类结果评价不同特征优选方案。

3.2 特征说明

本文选择了能够表征湿地植被、水文和土壤特征的指数进行分析，共选取了32种特征集，包括光谱特征、植被指数、水体指数、红边指数（表2）。

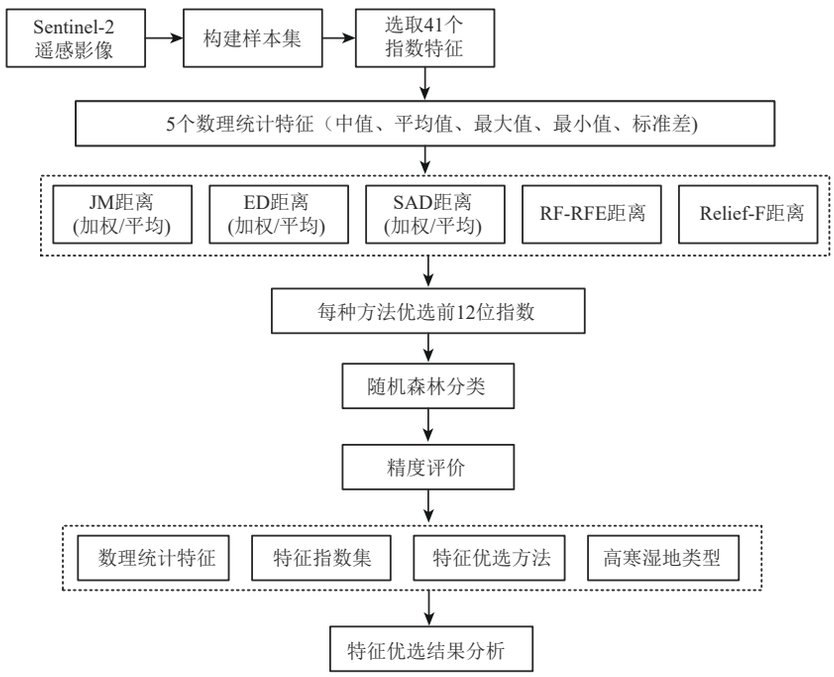


图2 特征优选技术流程图

Fig.2 Feature selection technology flow chart

表2 Sentinel-2特征集概述

Table 2 Sentinel-2 feature sets list

特征变量	指数简称	Sentinel-2A计算公式	参考文献	指数描述
光谱特征	Band	B2、B3、B4、B5、B6、B7、B8、B11、B12		
植被指数	NDVI	$(B8-B4)/(B8+B4)$	Defries和Townshend（1994）	对绿色植被的生长状况具有良好的表征能力，表征植被覆盖度和生长与健康状况
	VIgreen	$(B3-B4)/(B3+B4)$	Gitelson等（2002）	
	EVI	$2.5\times\left(\frac{B8-B4}{B8+6\times B4-7.5\times B2+1}\right)$	Huete等（2002）	通过冠层背景调节、气溶胶阻力系数和增益因子补偿土壤背景和大气效应，且对高生物量更敏感
	RDVI	$(B8-B4)/(\sqrt{B8+B4})$	Roujean和Breon（1995）	RDVI取DVI和NDVI两者之长，可用于高低不同植被覆盖的情况
	MSR	$\left(\frac{B8}{B4}-1\right)/\left(\sqrt{\frac{B8}{B4}}+1\right)$	Chen（1996）	该指数根据对两个光谱带组合得出的几种植被指数的评估而制定，包括NDVI、SR、SAVI、SAVI1、SAVI2、WDVI、GEMI、NLI和RDVI
	TCARI	$3\times\left((B8-B4)-0.2\times(B8-B3)\times\left(\frac{B8}{B4}\right)\right)$	Haboudane等（2002）	对叶绿素含量变化非常敏感
	TVI	$0.5\times(120\times(B8-B3)-200\times(B4-B3))$	Broge和Leblanc（2001）	该指数基于叶绿素吸收导致红色反射率降低，叶组织丰度导致NIR反射率增加，这两者都会增加三角形的总面积
	RVI	$B8/B4$	Birth和Mcvey（1968）	用于估算和检测植被生物量，并对高植被覆盖度反应敏锐
	DVI	$B8-B4$	Tucker（1979）	DVI对土壤背景的变化较RVI敏感，植被覆盖度高时，对植被的灵敏度有所下降
	GCVI	$\frac{B8}{B3}-1$	Gitelson等（2003）	GCVI比NDVI具有更大的动态范围，适合高密度植被覆盖区域

续表

特征变量	指数简称	Sentinel-2A计算公式	参考文献	指数描述
	SAVI	$1.5\times(B8-B4)/(B8+B4+0.5)$	Huete（1988）	消除不同背景的土壤所反射的不同光谱对植被特性的影响，SAVI包含了土壤调节系数，更适用于低植被覆盖地区
	gNDVI	$(B8-B3)/(B8+B3)$	Gitelson和Merzlyak（1996）	gNDVI与叶绿素含量和叶面积指数有显著相关性
	MCARI	$\left((B8-B4)-0.2\times(B8-B3)\right)\times(B8/B4)$	Daughtry等（2000）	对叶片叶绿素浓度和背景反射率都有响应
	GI	$B3/B4$		
	NDWI	$(B3-B8)/(B3+B8)$	McFeeters（1996）	能够抑制植被信息，突出水体；对建筑物和土壤的分离有一定影响；受冰雪、薄云和山体阴影影响较大，适用于地势平坦地区
	NDWI_B	$(B2-B4)/(B2+B4)$	张帅旗 等（2020）	可以在一定程度上抑制与水体无关的背景信息，增强水体信息
	MNDWI	$(B3-B11)/(B3+B11)$	徐涵秋（2005）	能够较好去除居民地和土壤等影响，突出水体；受阴影影响大
	RNDWI	$(B12-B4)/(B12+B4)$		能削弱混合像元和山体阴影的影响，较好地提取水陆边界；一般适用于山区等地形起伏较大地区
水体指数	LSWI	$(B8-B11)/(B8+B11)$	Xiao等（2005）	对植被冠层含水量和土壤含水量敏感
	EWI	$(B3-B8-B12)/(B3+B8+B12)$	闫霈 等（2007）	能够抑制居民地、土壤和植被等噪声；易受到阴影及浅滩的影响；适合半干旱地区的水体提取
	UGWI	$\frac{B3^3-B2\times B4\times B8}{B3^3+B2\times B4\times B8}$	段纪维 等（2021）	能较完整地提取不同泥沙含量的洪水、坑塘水面及阴影下各种水体，不易受阴影和暗色地物等干扰
	SWI	$B2+B4-B8$	陈文倩 等（2015）	较好地区分水体和阴影，能削弱积雪和山体裸地的影响，适用于山区的水体提取
	NDVIre1	$(B8-B5)/(B8+B5)$	Gitelson和Merzlyak（1994）	用波段5替代NDVI中的红波段
	NDVIre2	$(B8-B6)/(B8+B6)$	Gitelson和Merzlyak（1994）	用波段6替代NDVI中的红波段
	NDVIre3	$(B8-B7)/(B8+B7)$	Gitelson和Merzlyak（1994）	用波段7替代NDVI中的红波段
	NDre1	$(B6-B5)/(B6+B5)$	Gitelson和Merzlyak（1994）；Barnes等（2000）	用红边的峰和谷来代替传统NDVI中的红光和近红外波段，可用于估算植物叶面积指数和叶绿素含量
	NDre2	$(B7-B5)/(B7+B5)$	Gitelson和Merzlyak（1994）；Barnes等（2000）	将NDVI中的近红外和红波段替换为波段7和波段5，它可用于精细农业、森林监测、植被胁迫性探测等
红边指数	CIre	$\frac{B7}{B5}-1$	Gitelson等（2006）	该指数与植物叶绿素，氮素含量具有显著的线性关系
	MTCI	$(B6-B5)/(B5-B4)$	Dash和Curran（2004）	对植物叶片叶绿素含量较敏感，其值越大代表叶绿素含量越高
	MCARI2	$\left((B6-B5)-0.2\times(B6-B3)\right)\times\left(\frac{B6}{B5}\right)$	Wu等（2008）	该指数对植物中的叶绿素含量较为敏感，其值越大表示叶绿素含量越高
	REP	$705+35\times\frac{0.5\times(B4+B7)-B5}{B6-B5}$	Guyot和Baret（1988）	当植物叶片的叶绿素含量增加时，REP向长波方向移动，反之则向短波方向移动
	IRECI	$(B7-B4)/(B5/B6)$	Frampton等（2013）	该指数与植物冠层叶绿素含量和叶面积指数具有相关关系，可定量表征植物的叶绿素含量

3.3 特征优选方法

本文选择包括JM距离、ED距离、SAD距离、Relief-F算法、Z检验在内的5种Filter方法，以及Wrapper方法的RFE算法进行本次实验。

3.3.1 Jeffries-Matusita距离

Jeffries-Matusita（JM）距离基于数据正态分布的假设得到不同类别的分离度，在模式识别和特征选择领域中较为广泛的使用（Dabboor等，2014）。对训练样本集$\boldsymbol{C}(i, j=1, 2, \cdots, C, i \neq j)$中两个待分地类$w_i$和$w_j$之间的JM可分性准则定义如下

$$J_{ij}=2(1-\mathrm{e}^{-d_{ij}}) \tag{1}$$

式中，d_{ij}是两个待分地类w_i和w_j之间的Bhattacharyya距离，定义为

$$d_{ij}=-\mathrm{In}\left(\int\sqrt{P\left(x/w_i\right)P(x/w_j)}\,\mathrm{d}x\right) \tag{2}$$

式中，$P\left(x/w_i\right)$和$P(x/w_j)$是地类w_i和w_j的随机变量x的条件概率密度函数，通常假设多元正态分布，Bhattacharyya距离公式表达为

$$d_{ij}=\frac{1}{8}\left(\boldsymbol{m}_j-\boldsymbol{m}_i\right)^{\mathrm{T}}\left(\frac{\boldsymbol{\Sigma}_i+\boldsymbol{\Sigma}_j}{2}\right)^{-1}\left(\boldsymbol{m}_j-\boldsymbol{m}_i\right)+\frac{1}{2}\mathrm{In}\frac{\left|\frac{\boldsymbol{\Sigma}_i+\boldsymbol{\Sigma}_j}{2}\right|}{\sqrt{\left|\boldsymbol{\Sigma}_i\right|\left|\boldsymbol{\Sigma}_j\right|}} \tag{3}$$

式中，$\boldsymbol{m}_i$和$\boldsymbol{m}_j$分别表示均值；$\boldsymbol{\Sigma}_i$和$\boldsymbol{\Sigma}_j$分别表示w_i和w_j的协方差矩阵，上标T表示矩阵的转置。

3.3.2 欧氏距离

欧氏距离ED（Euclidian Distance）是常见的相似性度量方法，其实质是通过一定的准则函数，求两个不同地类的像元对应的光谱向量之间的距离，此距离代表两像元的差异程度。两种地类的欧氏距离越大，代表两种待分地类间的可分性越强，反之，则表示可分性越弱（Carvalho Júnior等，2011）。由于欧氏距离算法默认每一个维度是相同权重，因此如果不同维度取值范围差别较大时需要先对其进行归一化，ED值计算公式为

$$\mathrm{ED}=\sqrt{\sum_{i=1}^{N}(X_i-Y_i)^2} \tag{4}$$

式中，i表示波段，N表示波段总数，X_i和Y_i分别表示两种待分地类样本集所对应的像元亮度值。

3.3.3 光谱角距离

光谱角距离SAD（Spectral Angle Distance）是常用的光谱分类方法。在光谱空间中，每个像元对应一个多维光谱向量，将两个向量之间的夹角定义为光谱角。光谱角越小，两光谱越相似，属于同类地物的可能性越大。由于光谱角距离不受光照、阴影等条件的影响，因此可以突出目标光谱形状特征。两种待分地类光谱的相似度越高，SAD值越大，最大取值为1（Kruse等，1993）。计算公式为

$$\theta=\cos^{-1}\frac{\sum_{i=1}^{N}X_iY_i}{\sqrt{\sum_{i=1}^{N}X_i{}^2\sum_{i=1}^{N}Y_i{}^2}} \tag{5}$$

$$\mathrm{SAD}=\cos\theta \tag{6}$$

式中，i表示波段，N表示波段总数，X_i和Y_i分别表示两种待分地类样本集所对应的像元亮度值。

3.3.4 Z检验方法

Z检验方法能够用来测定两种地物类型在不同特征变量间统计显著性差异。具体步骤为首先将特征变量分为植被指数、水体指数、红边指数3组，再分别计算两种湿地类型在不同变量的Z值。Z统计表达式如下（Han等，2012）：

$$Z=\frac{(u_1-u_2)}{\sqrt{\frac{\left(s_1^2\right)}{n_1}+\frac{\left(s_2^2\right)}{n_2}}} \tag{7}$$

式中，u_1和u_2指的是两种待分地类的平均像素值；n_1和n_2指两种待分地类样本个数；s_1和s_2指两种待分地类像素值的标准差。Z值越大，待分地类在此特征上的差异就越显著。

以上4种方法均用于判别两地类之间的可分性，为直观地表达指数对所有地类区分的能力，进行以下步骤：根据在样本解译时发现的沼泽、沼泽化草甸、草甸、草原4种类型越相邻越难区分的认识，采用加权（表3）对类型组合对应的JM值等进行处理，紧密相邻的两类权重赋4，次相邻的两类权重赋3，以此类推。在本次实验中同时采用了平均方法做补充。

表3 各类型组合权重分配

Table 3 Weight distribution of various types of combinations

类型组合	权重分配	类型组合	权重分配	类型组合	权重分配
河流、裸地	1	沼泽化草甸、沼泽	4	草原、裸地	1
沼泽、裸地	1	草甸、裸地	1	草原、河流	1
沼泽、河流	1	草甸、河流	1	草原、沼泽	2
沼泽化草甸、裸地	1	草甸、沼泽	3	草原、沼泽化草甸	3
沼泽化草甸、河流	1	草甸、沼泽化草甸	4	草原、草甸	4

3.3.5 Relief-F算法

Relief-F算法是基于分析邻近样本对类别的区分能力继而确定特征的权重，核心思想是一个优秀的特征应该使得同类的样本更加靠近，而使得不同类的样本更加分散，它是Relief算法的拓展（刘吉超和王锋，2021）。其原理为假设数据集$\boldsymbol{D}$中有N个类别的样本，对属于第n类中样本$\boldsymbol{R}$，首先在同类即第n类的样本中寻找$\boldsymbol{R}$的k个最近邻样本$\boldsymbol{H}$，作为猜中近邻；在第n类之外的每个类中均找到$\boldsymbol{R}$的k个最近邻样本$\boldsymbol{M}$作为猜错近邻，最后定义的权重为

$$W(\boldsymbol{A}) = W(\boldsymbol{A}) - \sum_{j=1}^{k}\frac{diff\left(\boldsymbol{A},\boldsymbol{R},H_j\right)}{mk} + \sum_{C\neq class(R)}\left(\frac{p(\boldsymbol{C})}{1-p\left(class(\boldsymbol{R})\right)}\sum_{j=1}^{k}diff\left(\boldsymbol{A},\boldsymbol{R},\boldsymbol{M}_j(\boldsymbol{C})\right)\right)/(mk) \tag{8}$$

式中，$diff(\boldsymbol{A},\ \boldsymbol{R}_1,\ \boldsymbol{R}_2)$表示样本$\boldsymbol{R}_1$和样本$\boldsymbol{R}_2$在特征$\boldsymbol{A}$上的差，其计算公式$\boldsymbol{M}_j(\boldsymbol{C})$表示类别$\boldsymbol{C}$中的第$j$个最近邻样本，$p(\boldsymbol{C})$为该类别的比例。

3.3.6 递归特征消除法

相比递归特征消除法RFE（Recursive feature elimination）（Elavarasan等，2020），随机森林和RFE相结合形成RF-RFE，其能够更加合理的决定最终特征子集的大小，避免了人为因素造成的影响。RF-RFE算法用于特征选择（Wu等，2017），是采用随机森林算法得到的重要性排序进行后向迭代删除特征重要度最小的特征，再将其余特征用随机森林算法重新评估后得到新的特征重要性排序，重复步骤，每次删掉特征重要性小的特征，最终得到分类的最优特征集。

3.4 分类方法

随机森林算法是由Breiman（2001）提出的一种统计学习理论，研究表明随机森林算法具有速度快，准确度高，稳定性好的优势。因此论文采用随机森林分类方法进行湿地信息分类。

$$H(x) = \arg\max_{y}\sum_{k} I(h_k(x) = y) \tag{9}$$

3.5 精度分析

混淆矩阵是一种特定的矩阵用来呈现算法性能的可视化效果，主要通过比较分类结果与实际测量值之间的混淆程度进行精度评价。本文利用混淆矩阵，分别计算总体精度OA（Overall Accuracy）、Kappa系数、生产者精度PA（Producer's Accuracy）、用户精度UA（User's Accuracy）。其中总体精度和Kappa系数作为评价总体分类精度的指标，生产者精度和用户精度作为衡量各类的漏分和错分误差的指标。

4 结 果

为了方便叙述，论文用到的特征提取方法、数理统计方法和数据处理方式分别按表4编码进行论述。

表4 编号对照表

Table 4 Number comparison table

优选方法	代码1	数理特征	代码2	处理方式	代码3
JM距离	A	中值	a	加权	1
ED距离	B	平均值	b	平均	2
SAD距离	C	最大值	c		
RF-RFE算法	D	最小值	d		
Relief-F算法	E	标准差	e		

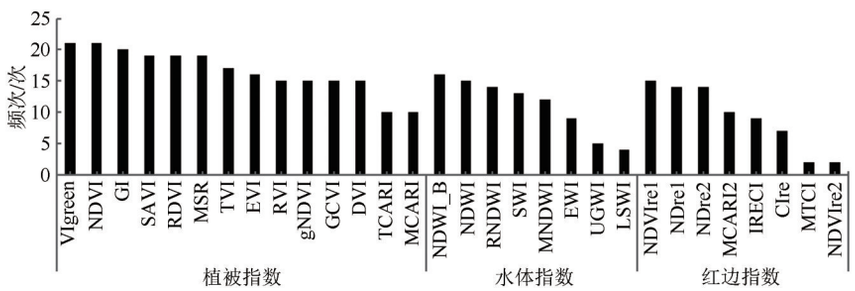
4.1 指数的优选

根据JM距离、SAD距离、ED距离、Relief-F算法和RF-RFE算法计算结果，逐一得出在水体指数、植被指数、红边指数中表现较好的指数。

4.1.1 JM距离

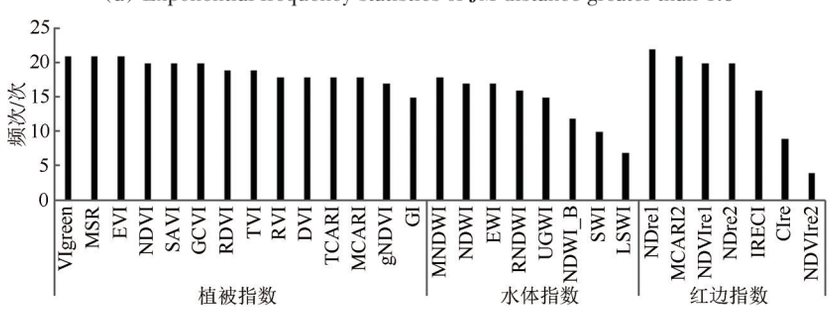
JM距离取值范围为［0，2］，JM距离大于1.8表示样本间可分性较好，据此选出每类组合中JM值大于1.8的指数，并对这些指数进行频次统计（图3(a)）。植被指数在频次上大于水体指数和红边指数，其中最突出的是NDVI、VIgreen这两个指数；水体指数中频次数在前两位的指数为NDWI_B、NDWI；红边指数中NDVIre1、NDre1、NDre2指数出现频次较多。

采取加权和平均两种计算方式对数据进行处理，按重要性从大到小排序后，根据前12位指数。发现加权方式优选的指数TVI、DVI、SAVI和以平均方式优选指数RDVI、SAVI，不论在哪一个数理统计特征中都有。当将两种方式所选的指数放在一起统计，优选至少出现5次的指数作为JM距离方法所选最优指数集，包括：SAVI、RDVI、TVI、DVI、MSR、EWI、NDWI、RNDWI、SWI、EVI、NDVI、gNDVI、RVI。基本包含了JM距离大于1.8所选频次较高的植被指数和水体指数，但是红边特征指数没能入选。



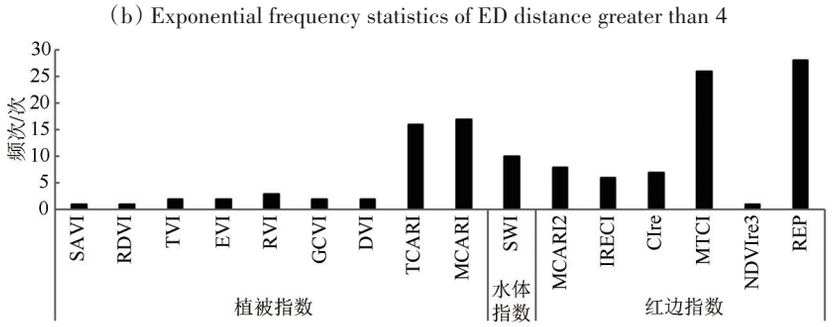
（a）JM距离大于1.8的指数频次统计

（a）Exponential frequency statistics of JM distance greater than 1.8



（b）ED距离大于4的指数频次统计

（b）Exponential frequency statistics of ED distance greater than 4



（c）SAD距离小于0.6的指数频次统计

（c）Exponential frequency statistics of sad distance less than 0.6

图3 JM距离、ED距离、SAD距离指数频次统计

Fig.3 Frequency statistics of JM distance, ED distance and SAD distance index

4.1.2 ED距离

ED值越大说明光谱距离越大，地类可分性越强。ED值大于4表示样本间可分性好，依此标准对优选的指数进行频次统计，指数出现频次如图3（b）所示。红边指数NDre1出现次数最多。另外在红边指数中MCARI2、NDre2、NDVIre1均

表现出较好的区分能力。植被指数VIgreen、MSR、EVI指数出现21次，其余指数与之相比相差不大，植被指数在分类中指数数量最多。水体指数MNDWI、NDWI、EWI表现优异。

同样对15种类型组合的ED值进行加权与平均两种方式计算，得出加权方式优选的指数出现4次以上的有：EVI、RVI、SAVI、GCVI、RDVI；平均方式优选的指数出现4次以上的是GCVI、RDVI、NDVI、SAVI、EVI、MSR、NDre1、NDre2、NDVIre1。综合两种方式所选的指数，同样以出现5次以上的指数作为ED距离方法优选指数集，包括GCVI、EVI、RDVI、SAVI、MSR、NDre1、NDre2、NDVIre1、gNDVI、RVI、NDVI、NDWI。所选水体指数只有NDWI一个。

4.1.3 SAD距离

光谱角检测是比较两类数据相似程度的光谱对比方法，当SAD值越趋近于1，两类的光谱就越相似。以SAD值小于0.6的指数出现的频次进行统计，其指数出现频次统计如图3（c）所示，指数频次显示出较大差距，植被指数MCARI、TCARI不同于其他方法所选指数。进一步研究发现，TCARI值是MCARI值的3倍，相比其他植被指数，对叶绿素浓度的变化十分敏感；红边指数REP出现最多，其次是MTCI指数。REP指数可定量表征植物的叶绿素含量，而MTCI指数同样对叶绿素含量敏感。SAD距离所选植被指数和红边指数均对植物叶片含的叶绿素有较好的表征。水体指数仅选出SWI，该指数能较好区分水体和阴影。

对SAD值分别进行加权和平均处理后发现，加权方式优选指数中出现4次以上的指数为：TCARI、MCARI、MCARI2、IRECI、EWI、GCVI、NDWI、MNDWI；平均方式优选指数出现次数4次以上的为：TCARI、MCARI、MCARI2、IRECI、CIre、RVI、GCVI、LSWI、EVI。将两种方式所选指数共同考虑得到SAD距离方法优选的指数集为TCARI、MCARI、MCARI2、IRECI、CIre、RVI、GCVI、EWI、MNDWI、EVI、LSWI、NDWI。

4.1.4 RF-RFE特征重要性排序

利用RF-RFE算法优选得到的前12位指数按重要性排序（表5），不同的数理统计特征出现很多重复的指数，一定程度上说明这些指数在重要性排序中表现较好。

表5 RF-RFE算法特征重要性前12位特征指数集排序

Table 5 RF-RFE algorithm feature importance top twelve feature index set

数理统计特征	1	2	3	4	5	6	7	8	9	10	11	12
中值	RDVI	DVI	gNDVI	GI	NDWI	NDWI_B	MNDWI	RNDWI	EWI	CIre	VIgreen	TVI
平均值	VIgreen	NDVI	TVI	GCVI	GI	NDWI_B	MNDWI	RNDWI	EWI	CIre	RDVI	DVI
最大值	VIgreen	NDVI	RDVI	MSR	TVI	DVI	GI	MNDWI	UGWI	LSWI	RVI	NDWI_B
最小值	TCARI	RVI	DVI	SAVI	NDWI	NDWI_B	RNDWI	EWI	UGWI	SWI	EVI	CIre
标准差	VIgreen	TCARI	TVI	DVI	GI	RNDWI	EWI	LSWI	MCARI	MCARI2	MTCI	IRECI

注：从1到12，重要性程度逐渐降低。

其中，植被指数有RDVI、DVI、GI、VIgreen、TVI，水体指数有NDWI_B、MNDWI、RNDWI、EWI，红边指数有CIre。

4.1.5 Relief-F特征重要性排序

依据Relief-F算法优选的指数按重要性排序（表6），其中，中值、平均值和标准差特征除了指数重要性排序不同外，选择的指数均相同，并且部分指数也出现在最大值和最小值特征中。指数特征中表现优异的指数分别为：水体指数中的RNDWI、MNDWI、EWI，植被指数有VIgreen、GI、MSR、GCVI、RVI，红边指数为NDre1、NDVIre1、NDre2，反映出这些指数对于区分地类较为重要。另外还发现重要性排序第一位的总是水体指数。

4.2 数理统计特征的优选

样本的空间分布具有随机性特征，虽然在湿地分类中通常采用平均值的方式对样本进行处理，但哪一种统计特征更能代表样本的属性特征，目前为止没有相关的研究。为此本文研究分别根据平均值、中值、最大值、最小值和标准差进行各类特征的计算，在此基础上评价对属性特征优选的影响。

表6 Relief-F算法特征重要性前12位特征指数集排序

Table 6 Relief-F algorithm feature importance top twelve feature index set

数理统计特征	1	2	3	4	5	6	7	8	9	10	11	12
中值	NDWI_B	VIgreen	GI	RNDWI	CIre	MNDWI	LSWI	NDVI	NDre2	NDre1	EWI	MSR
平均值	NDWI_B	VIgreen	GI	NDVI	CIre	RNDWI	MSR	MNDWI	LSWI	NDVIre1	NDre2	GCVI
最大值	MNDWI	LSWI	NDVI	EWI	VIgreen	gNDVI	MSR	NDWI	GI	UGWI	GCVI	NDre2
最小值	RNDWI	NDWI	SWI	RDVI	UGWI	SAVI	NDVIre1	TVI	MCARI2	EVI	RVI	MSR
标准差	RNDWI	MNDWI	VIgreen	GI	EWI	NDre1	MSR	NDVIre1	NDre2	GCVI	RVI	LSWI

注：从1到12，重要性程度逐渐降低。

4.2.1 JM距离

对表5基于JM距离优选的指数集分别采用随机森林分类器进行分类，得出中值和平均值特征总体精度和Kappa系数最高，均为86.70%、0.840，标准差特征的最低。依据JM距离计算结果，指数数理统计特征的可分性能力可排序为平均值特征=中值特征>最小值特征>最大值特征>标准差特征。

4.2.2 ED距离

对ED距离优选得到的指数集进行随机森林分类，分类精度结果如表7。

表7 ED算法分类精度统计

Table 7 ED algorithm classification accuracy statistics

类别	B-a-2		B-b-2		B-c-1		B-d-1		B-e-1	
	PA/%	UA/%	PA/%	UA/%	PA/%	UA/%	PA/%	UA/%	PA/%	UA/%
草原	88.50	95.80	88.50	92.00	88.50	95.80	88.50	88.50	88.50	88.50
草甸	79.20	76.00	79.20	73.10	79.20	63.30	79.20	70.40	75.00	60.00
沼泽化草甸	76.00	70.40	72.00	72.00	64.00	69.60	72.00	75.00	56.00	73.70
沼泽	77.80	82.40	77.80	87.50	77.80	93.30	83.30	88.20	83.30	88.20
河流	100.00	100.00	100.00	100.00	95.80	88.50	100.00	100.00	95.80	100.00
裸地	100.00	100.00	100.00	96.30	92.30	96.00	96.20	100.00	92.30	85.70
总体精度/%	**87.40**		86.70		83.20		86.70		81.80	
Kappa系数	**0.849**		0.840		0.798		0.840		0.781	

注：加粗数字为最大总体精度、Kappa系数。

中值特征的总体精度与Kappa系数最高，达到87.40%、0.849；标准差特征的最小；平均值与最小值特征的相同。从ED方法看，数理特征的可分性能力排序为：中值特征>平均值特征=最小值特征> 最大值特征>标准差特征。

4.2.3 SAD距离

对SAD距离计算得到的指数集逐一进行随机森林分类，得出分类精度结果。将SAD距离方法优选的指数组合利用随机森林分类方法，比较其精度后发现，虽然所选出的指数与前两种方法选出的指数有较大不同，但是平均值特征的总体精度依然能达到87.40%，Kappa系数达0.848，SAD距离结果显示不同统计特征可分性能力大小排序为：中值特征>平均值特征>最大值特征=最小值特征>标准差特征。

4.2.4 RF-RFE特征重要性排序

根据RF-RFE特征重要性排序选出指数集（表5）的对应分类结果：平均值特征的总体精度与Kappa系数最高，达87.40%、0.848，而标准差特征的最低，中值和最小值的相同。依据总体精度和Kappa系数的大小，可将其排序为平均值特征>中值特征=最小值特征>最大值特征>标准差特征。

4.2.5 Relief-F特征重要性排序

对Relief-F算法计算结果取前12位指数（表6）进行随机森林分类得到结果见表8。

中值和平均值特征的总体精度和Kappa系数均为87.40%、0.849，最大值特征的最低，比中值和平均值低了4.90%、0.059。以数理统计特征的维度，依据总体特征和Kappa系数的大小，可将其排序为中值特征=平均值特征>最小值特征>标准差特征>最大值特征。

表8 Relief-F算法分类精度统计

Table 8 Relief-F algorithm classification accuracy statistics

类别	E-a		E-b		E-c		E-d		E-e	
	PA/%	UA/%	PA/%	UA/%	PA/%	UA/%	PA/%	UA/%	PA/%	UA/%
草原	88.50	95.80	88.50	88.50	88.50	95.80	88.50	85.10	80.80	87.50
草甸	87.50	75.00	87.50	77.80	79.20	61.30	75.00	69.20	75.00	62.10
沼泽化草甸	68.00	73.90	68.00	73.90	64.00	64.00	72.00	75.00	64.00	72.70
沼泽	77.80	82.40	77.80	82.40	72.20	86.60	83.30	88.20	88.90	84.20
河流	100.00	100.00	100.00	100.00	95.80	95.80	100.00	100.00	100.00	100.00
裸地	100.00	96.30	100.00	100.00	92.30	100.00	96.20	100.00	96.20	100.00
总体精度/%	**87.40**		**87.40**		82.50		86.00		83.90	
Kappa系数	**0.849**		**0.849**		0.790		0.832		0.807	

注：加粗数字为最大总体精度、Kappa系数。

总体而言，湿地信息提取最适合的数理统计特征为中值和平均值。

4.3 高寒湿地类型的区分能力

本文将沼泽、沼泽化草甸、草甸、河流、草原和裸地等土地覆被类型分别两两组合分析其可分性，共有15组类型组合。

4.3.1 河流、裸地

河流与裸地因其截然不同的形状特征、下垫面环境，是所有类别中分类精度最高的，其制图精度和用户精度可达到100%。另外能用于区分河流或裸地与其他类别的指数数目很多。

4.3.2 沼泽、沼泽化草甸、草甸、草原

草原的制图精度和用户精度的中位数是88.5%，沼泽化草甸与草甸的制图与用户精度主要集中在72%，相比较低。沼泽相比于沼泽化草甸与草甸两类湿地过渡类型来说，制图精度和用户精度均要高，大部分情况集中在83.3%。沼泽化草甸与沼泽、草原与草甸、草甸与沼泽化草甸这3种类型组合难以区分，因此，结合JM距离、ED距离、SAD距离计算结果，尝试找出易于区分这3种组合类型的指数。具体做法如下：依据JM距离、ED距离、SAD距离指数的中值特征计算结果，选出沼泽化草甸与沼泽、草原与草甸、草甸与沼泽化草甸类型组合前10%的指数，共计25个，根据指数出现的频次排序，选出前10%的指数为：MCARI2、NDWI、DVI、EVI、EWI、IRECI、MCARI、TCARI、UGWI。

5 讨 论

5.1 分类指数集的构建

由于每种方法优选的指数不完全一致，为确定对于湿地分类最优的特征，我们分别取Filter方法（JM距离、ED距离、SAD距离、Relief-F算法）和Wrapper方法（RF-RFE算法）计算结果中精度最高的指数集合，然后以这5个指数集合为基础，统计各个指数在5个集合中出现的次数，以众数为指标作为最终优选的指数，结果包括了11个指数。分别为植被指数RDVI、NDVI、MSR、RVI、VIgreen，水体指数RNDWI、NDWI、NDWI_B、MNDWI、EWI和红边指数CIre。

5.2 数理统计特征的选择

对优选出的11个指数，分别对5个统计特征进行随机森林分类，得到的总体精度和Kappa系数如表9。

表9 不同数理统计特征下分类精度统计

Table 9 Classification accuracy statistics under different mathematical statistics characteristics

数理统计特征	总体精度/%	Kappa 系数
中值	88.10	0.857
平均值	86.70	0.84
最大值	85.30	0.823
最小值	83.20	0.798
标准差	85.30	0.824

可以看出，中值特征的总体精度和Kappa系数均是最高的，其次是平均值特征，和不同方法分

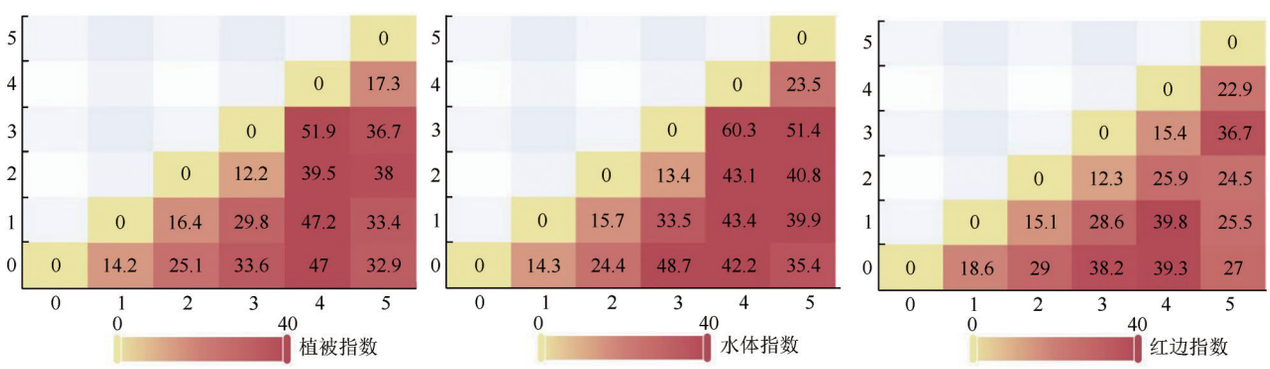
别进行评价的结果一致（4.2节部分）。说明基于样本的中值或平均值可以代表样本的属性。

5.3 特征优选方法的比较

由于中值特征的精度以及分类的效果最好，因此用中值特征计算的精度及混淆矩阵来比较5种特优选的方法。可以发现Filter方法的Relief-F算法与ED距离算法的总体精度和Kappa系数均为87.40%、0.849，JM距离与SAD距离得到的总体精度和用户精度均为86.70%、0.840，Wrapper方法的RFE-RF算法的总体精度和用户精度为86.00%、0.832，由此可知，Filter方法的Relief-F与ED距离算法在本次高寒湿地分类指数特征研究中略胜一筹。

5.4 指数特征的重要性排序

为了能够更好地查看水体指数特征、植被指数特征与红边指数特征对湿地分类的贡献，利用Filter方法的Z检验方法对样本点的指数特征进行定量分析。依据5.2节的结果，最终选择指数的中值特征进行计算。



（a）两两类之间植被指数显著性差异 （b）两两类之间水体指数显著性差异 （c）两两类之间红边指数显著性差异

（a）Significant difference in vegetation index between two classes （b）Significant difference of water body index between two classes （c）Significant difference in red edge index between two classes

图4 两两类之间的植被指数、水体指数和红边指数的显著性差异（0-草原、1-草甸、2-沼泽化草甸、3-沼泽、4-河流、5-裸地）

Fig.4 Significant differences in vegetation index, water body index and red edge index between the two classes (0-grassland, 1-meadow, 2-swamped meadow, 3-swamp, 4-river, 5-bare land)

图4表明了不同类型的指数特征对湿地的可分离度，Z值越大，两种湿地类型在此指数上的差异度越明显。通过观察决定以Z值大于37为界限进行统计，可得水体指数特征Z值最大为60.3，大于37的类型组合有8个；植被指数特征Z值最大为51.9，大于37的类型组合有5个；红边指数特征Z值最大为39.8，大于37的类型组合有3个。由此可知，不同指数特征的特征重要性程度排序为：水体指数特征>植被指数特征>红边指数特征。

5.5 分类特征的不确定性

沼泽化草甸与沼泽、草甸与沼泽化草甸、草原与草甸这3种类型组合的可分性较差，而这3类组合的分类情况决定着湿地分类结果的好坏。从类型角度来看，草原、草甸、沼泽化草甸、沼泽邻近两类型无论在植被长势还是下垫面水分含量都存在渐变的过程，邻近两类型之间没有明显的地缘间隙。因此，在用到的分类特征时是易混淆的。JM距离、ED距离、SAD距离计算结果也证明了这一点，后续需要深入研究上述类型的分类技术。

此外，SAD距离方法所得指数集的随机森林分类精度与JM距离方法一样能达到86.70%，但是两种方法所选指数集却大相径庭，接下来可以从指数的物理机理、波段组合方式方面进一步研究。

6 结 论

本文基于Sentinel-2遥感影像，以首曲高寒湿地为研究区，通过Filter方法（JM距离、ED距离、SAD距离、Relief-F算法）、Wrapper方法（RF-RFE算法）共同对植被指数特征、水体指数特征、红边指数特征进行分析，并借助随机森林分类方法计算了所选指数集的精度和混淆矩阵，最后利用了Z检验方法对3种指数特征进行了定量比较，主要得到以下结论：

（1）特征优选方法：在所选的Filter和Wrapper

方法中，Filter方法的ED距离与Relief-F算法得出的指数集其分类精度高于其他方法，精度最大相差1.4%。说明ED距离与Relief-F算法在本次湿地分类研究中具有最好结果。其原因是Wrapper方法在特征选择中过分依赖聚类参数，缺乏合适的评价准则评估不同特征子空间的特征子集，所以采用Filter方法相对取得了较好的结果。

（2）湿地分类优选指数：植被指数RDVI、NDVI、MSR、RVI、VIgreen，水体指数RNDWI、NDWI、NDWI_B、MNDWI、EWI，红边指数CIre。具体来说，水体指数在湿地分类中占重要地位，尤其是NDWI、MNDWI、NDWI_B，植被指数在所选总指数中所占数量最多，其中NDVI指数尤为重要，红边指数虽没有前两类指数表现突出但是也不可或缺，表现好的指数有NDVIre1、NDre1、NDre2、CIre。通过Z检验对指数特征的定量分析，可得出水体指数特征的重要性程度大于植被指数特征，也大于红边指数特征。

（3）统计特征评价结果：基于中值的总体精度最高，达到88.10%。

（4）最易区分的是河流与裸地，其次为沼泽与草原，沼泽化草甸与草甸较为难分。沼泽化草甸与沼泽、草原与草甸、草甸与沼泽化草甸这3种类型组合较难区分，可尝试如下指数进行分类：MCARI2、NDWI、DVI、EVI、EWI、IRECI、MCARI、TCARI、UGWI。

参考文献（References）

Barnes E M, Clarke T R, Richards S R, Colaizzi P D, Haberland J, Kostrzewski M, Waller P, Choi C, Riley E, Thompson T, Lascano R J, Li H and Moran M S. 2000. Coincident detection of crop water stress, nitrogen status and canopy density using ground-based multispectral data//Proceedings of the 5th International Conference on Precision Agriculture. Bloomington, Minnesota, USA: [s.n.]: 1-15

Birth G S and Mcvey G R. 1968. Measuring the color of growing turf with a reflectance Spectrophotometer. Agronomy Journal, 60(6): 640-643 [DOI: 10.2134/agronj1968.00021962006000060016x]

Breiman L. 2001. Random forest. Machine Learning, 45(1): 5-32 [DOI: 10.1023/A:1010933404324]

Broge N H and Leblanc E. 2001. Comparing prediction power and stability of broadband and hyperspectral vegetation indices for estimation of green leaf area index and canopy chlorophyll density. Remote Sensing of Environment, 76(2): 156-172 [DOI: 10.1016/s0034-4257(00)00197-8]

Carvalho Júnior O A, Guimarães R F, Gillespie A R, Silva N C and Gomes R A T. 2011. A new approach to change vector analysis using distance and similarity measures. Remote Sensing, 3(11): 2473-2493 [DOI: 10.3390/rs3112473]

Chang W T, Wang H, Ning X G and Zhang H C. 2020. Extraction of Zhalong wetlands information based on images of Sentinel-2 red-edge bands and Sentinel-1 radar bands. Wetland Science, 18(1): 10-19 (常文涛, 王浩, 宁晓刚, 张翰超. 2020. 融合Sentinel-2红边波段和Sentinel-1雷达波段影像的扎龙湿地信息提取. 湿地科学, 18(1): 10-19) [DOI: 10.13248/j.cnki.wetlandsci.2020.01.002]

Chen J M. 1996. Evaluation of vegetation indices and a modified simple ratio for boreal applications. Canadian Journal of Remote Sensing, 22(3): 229-242 [DOI: 10.1080/07038992.1996.10855178]

Chen W Q, Ding J L, Li Y H and Niu Z Y. 2015. Extraction of water information based on China made-GF-1 remote sense image. Resources Science, 37(6): 1166-1172 (陈文倩, 丁建丽, 李艳华, 牛增懿. 2015. 基于国产GF-1遥感影像的水体提取方法. 资源科学, 37(6): 1166-1172)

Dabboor M, Howell S, Shokr M and Yackel J. 2014. The Jeffries-Matusita distance for the case of complex Wishart distribution as a separability criterion for fully polarimetric SAR data. International Journal of Remote Sensing, 35(19/20): 6859-6873 [DOI: 10.1080/01431161.2014.960614]

Dash M, Choi K, Scheuermann P and Liu H. 2002. Feature selection for clustering-a filter solution//2002 IEEE International Conference on Data Mining. Maebashi City, Japan: IEEE: 115-122 [DOI: 10.1109/ICDM.2002.1183893]

Dash J and Curran P J. 2004. The MERIS terrestrial chlorophyll index. International Journal of Remote Sensing, 25(23): 5403-5413 [DOI: 10.1080/0143116042000274015]

Dash M and Liu H. 1997. Feature selection for classification. Intelligent Data Analysis, 1(1/4): 131-156 [DOI: 10.1016/S1088-467X(97)00008-5]

Daughtry C S T, Walthall C L, Kim M S, de Colstoun E B and McMurtrey J E III. 2000. Estimating corn leaf chlorophyll concentration from leaf and canopy reflectance. Remote sensing of Environment, 74(2): 229-239 [DOI: 10.1016/s0034-4257(00)00113-9]

Defries R S and Townshend J R G. 1994. NDVI-derived land cover classifications at a global scale. International Journal of Remote Sensing, 15(17): 3567-3586 [DOI: 10.1080/01431169408954345]

Duan J W, Zhong J S, Jiang L and Zhong M M. 2021. Extraction method of ultra-green water index for flood area after rain based on GF-2 image. Geography and Geo-Information Science, 37(3): 35-41 (段纪维, 钟九生, 江丽, 钟森淼. 2021. 基于GF-2影像的雨后洪涝区超绿水体指数提取方法. 地理与地理信息科学, 37(3): 35-41) [DOI: 10.3969/j.issn.1672-0504.2021.03.006]

Elavarasan D, Vincent P M D R, Srinivasan K and Chang C Y. 2020. A hybrid CFS filter and RF-RFE wrapper-based feature extraction for enhanced agricultural crop yield prediction modeling. Agriculture, 10(9): 400 [DOI: 10.3390/agriculture10090400]

Frampton W J, Dash J, Watmough G and Milton E J. 2013. Evaluating

the capabilities of Sentinel-2 for quantitative estimation of biophysical variables in vegetation. ISPRS Journal of Photogrammetry and Remote Sensing, 82: 83-92 [DOI: 10.1016/j.isprsjprs.2013.04.007]

Gao B B. 2008. Research on the Status Quo of Wetland in Shouqu Nature Reserve of the Yellow River in Gansu Province and its protection Countermeasures. Gansu Science and Technology, 24(12): 3-5 (高斌斌. 2008. 甘肃黄河首曲自然保护区湿地现状及其保护对策研究. 甘肃科技, 24(12): 3-5) [DOI: 10.3969/j.issn.1000-0952.2008.12.002]

Gitelson A A, Kaufman Y J, Stark R and Rundquist D. 2002. Novel algorithms for remote estimation of vegetation fraction. Remote Sensing of Environment, 80(1): 76-87 [DOI: 10.1016/s0034-4257(01)00289-9]

Gitelson A A, Keydan G P and Merzlyak M N. 2006. Three-band model for noninvasive estimation of chlorophyll, carotenoids, and anthocyanin contents in higher plant leaves. Geophysical Research Letters, 33(11): L11402 [DOI: 10.1029/2006gl026457]

Gitelson A and Merzlyak M N. 1994. Spectral reflectance changes associated with autumn senescence of *Aesculus hippocastanum* L. and *Acer platanoides* L. leaves. Spectral features and relation to chlorophyll estimation. Journal of Plant Physiology, 143(3): 286-292 [DOI: 10.1016/s0176-1617(11)81633-0]

Gitelson A A and Merzlyak M N. 1996. Signature analysis of leaf reflectance spectra: algorithm development for remote sensing of chlorophyll. Journal of Plant Physiology, 148(3/4): 494-500 [DOI: 10.1016/s0176-1617(96)80284-7]

Gitelson A A, Viña A, Arkebauer T J, Rundquist D C, Keydan G and Leavitt B. 2003. Remote estimation of leaf area index and green leaf biomass in maize canopies. Geophysical Research Letters, 30(5): 1248 [DOI: 10.1029/2002GL016450]

Guyon I and Elisseeff A. 2003. An introduction to variable and feature selection. Journal of Machine Learning Research, 3: 1157-1182

Guyot G and Baret F. 1988. Utilisation de la haute resolution spectrale pour suivre l'etat des couverts vegetaux//Proceedings of the 4th International Conference on Spectral Signatures of Objects in Remote Sensing. Aussois, France: [s.n.]: 279-286

Haboudane D, Miller J R, Tremblay N, Zarco-Tejada P J and Dextraze L. 2002. Integrated narrow-band vegetation indices for prediction of crop chlorophyll content for application to precision agriculture. Remote Sensing of Environment, 81(2/3): 416-426 [DOI: 10.1016/s0034-4257(02)00018-4]

Han N, Wang K, Yu L and Zhang X Y. 2012. Integration of texture and landscape features into object-based classification for delineating-Torreyausing IKONOS imagery. International Journal of Remote Sensing, 33(7): 2003-2033 [DOI: 10.1080/01431161.2011.605084]

Hao Y F, Man W D, Wang J H, Liu M Y and Zhang K. 2021. Wetland information extraction based on Relief-F algorithm and decision tree method. Journal of Liaoning Technical University (Natural Science), 40(3): 225-233 (郝玉峰, 满卫东, 汪金花, 刘明月, 张阔. 2021. Relief-F算法及决策树方法下的湿地信息提取. 辽宁工程技术大学学报(自然科学版), 40(3): 225-233)

He J H, Zhang T B, Yi G H, Bie X J, Luo N, Wang Q and Fan W W. 2015. Wetland information extraction and change detection in Zoige Plateau area based on the EOS/MODIS. Geomatics and Spatial Information Technology, 38(9): 38-41 (何菊红, 张廷斌, 易桂花, 别小娟, 罗娜, 王强, 范微维. 2015. 基于EOS/MODIS若尔盖高原地区湿地信息提取及变化监测. 测绘与空间地理信息, 38(9): 38-41)

Huete A, Didan K, Miura T, Rodriguez E P, Gao X and Ferreira L G. 2002. Overview of the radiometric and biophysical performance of the MODIS vegetation indices. Remote Sensing of Environment, 83(1/2): 195-213 [DOI: 10.1016/S0034-4257(02)00096-2]

Huete A R. 1988. A soil-adjusted vegetation index (SAVI). Remote Sensing of Environment, 25(3): 295-309 [DOI: 10.1016/0034-4257(88)90106-X]

John G H, Kohavi R and Pfleger K. 1994. Irrelevant features and the subset selection problem//Proceedings of the Eleventh International Conference on Machine Learning. New Brunswick, NJ: Rutgers University, 121-129 [DOI: 10.1016/B978-1-55860-335-6.50023-4]

Kira K and Rendell L A. 1992. The feature selection problem: traditional methods and a new algorithm//Proceedings of the Tenth National Conference on Artificial Intelligence. San Jose, California: AAAI Press: 129-134

Korhonen L, Hadi, Packalen P and Rautiainen M. 2017. Comparison of Sentinel-2 and Landsat 8 in the estimation of boreal forest canopy cover and leaf area index. Remote Sensing of Environment, 195: 259-274 [DOI: 10.1016/j.rse.2017.03.021]

Kruse F A, Lefkoff A B and Dietz J B. 1993. Expert system-based mineral mapping in northern Death Valley, California/Nevada, using the airborne visible/infrared imaging spectrometer (AVIRIS). Remote Sensing of Environment, 44(2/3): 309-336 [DOI: 10.1016/0034-4257(93)90024-r]

Liu J H and Wang F. 2021. A semi-supervised feature selection algorithm based on Relief-F. Journal of Zhengzhou University (Natural Science Edition), 53(1): 42-46, 53 (刘吉超, 王锋. 2021. 基于Relief-F的半监督特征选择算法. 郑州大学学报(理学版), 53(1): 42-46, 53) [DOI: 10.13705/j.issn.1671-6841.2020196]

Mahdianpari M, Salehi B, Mohammadimanesh F, Homayouni S and Gill E. 2019. The first wetland inventory map of newfoundland at a spatial resolution of 10 m using Sentinel-1 and Sentinel-2 data on the Google earth engine cloud computing platform. Remote Sensing, 11(1): 43 [DOI: 10.3390/rs11010043]

McFeeters S K. 1996. The use of the normalized difference water index (NDWI) in the delineation of open water features. International Journal of Remote Sensing, 17(7): 1425-1432 [DOI: 10.1080/01431169608948714]

Moor H, Hylander K and Norberg J. 2015. Predicting climate change effects on wetland ecosystem services using species distribution modeling and plant functional traits. Ambio, 44(S1): 113-126 [DOI: 10.1007/s13280-014-0593-9]

Phan T N, Kuch V and Lehnert L W. 2020. Land cover classification using Google earth engine and random forest classifier—The role

of image composition. Remote Sensing, 12(15): 2411 [DOI: 10.3390/rs12152411]

Roujean J L and Breon F M. 1995. Estimating PAR absorbed by vegetation from bidirectional reflectance measurements. Remote Sensing of Environment, 51(3): 375-384 [DOI: 10.1016/0034-4257(94)00114-3]

Saeys Y, Inza I and Larrañaga P. 2007. A review of feature selection techniques in bioinformatics. Bioinformatics, 23(19): 2507-2517 [DOI: 10.1093/bioinformatics/btm344]

Shoko C and Mutanga O. 2017. Examining the strength of the newly-launched Sentinel 2 MSI sensor in detecting and discriminating subtle differences between C3 and C4 grass species. ISPRS Journal of Photogrammetry and Remote Sensing, 129: 32-40 [DOI: 10.1016/j.isprsjprs.2017.04.016]

Sun Y L, Zhang X, Shuai T, Shang K and Feng S N. 2015. Radiometric normalization of hyperspectral satellite images with spectral angle distance and Euclidean distance. Journal of Remote Sensing, 19(4): 618-626 (孙艳丽, 张霞, 帅通, 尚坤, 冯淑娜. 2015. 光谱角—欧氏距离的高光谱图像辐射归一化. 遥感学报, 19(4): 618-626) [DOI: 10.11834 /jrs.20154176]

Tucker C J. 1979. Red and photographic infrared linear combinations for monitoring vegetation. Remote Sensing of Environment, 8(2): 127-150 [DOI: 10.1016/00344257(79)90013-0]

Wang G X, Li Y S, Wang Y B and Chen L. 2007. Typical alpine wetland system changes on the Qinghai-Tibet Plateau in recent 40 years. Acta Geographica Sinica, 62(5): 481-491 (王根绪, 李元寿, 王一博, 陈玲. 2007. 近40年来青藏高原典型高寒湿地系统的动态变化. 地理学报, 62(5): 481-491) [DOI: 10.3321/j.issn:0375-5444.2007.05.004]

White L, Ryerson R A, Pasher J and Duffe J. 2020. State of science assessment of remote sensing of great lakes coastal wetlands: responding to an operational requirement. Remote Sensing, 12(18): 3024 [DOI: 10.3390/rs12183024]

Wu C Y, Niu Z, Tang Q and Huang W J. 2008. Estimating chlorophyll content from hyperspectral vegetation indices: modeling and validation. Agricultural and Forest Meteorology, 148(8/9): 1230-1241 [DOI: 10.1016/j.agrformet.2008.03.005]

Xiao X M, Boles S, Liu J Y, Zhuang D F, Frolking S, Li C S, Salas W and Moore B. 2005. Mapping paddy rice agriculture in southern China using multi-temporal MODIS images. Remote Sensing of Environment, 95(4): 480-492 [DOI: 10.1016/j.rse.2004.12.009]

Xie S Y, Fu B L, Li Y, Liu Z L, Zuo P P, Lan F W, He H C and Fan D L. 2021. Classification method on marsh wetlands in Honghe National Nature Reserve based on multi-dimensional remote sensing images. Wetland Science, 19(1): 1-16 (解淑毓, 付波霖, 李颖, 刘兆礼, 左萍萍, 蓝斐芫, 何宏昌, 范冬林. 2021. 基于多维度遥感影像的洪河国家级自然保护区沼泽湿地分类方法研究. 湿地科学, 19(1): 1-16) [DOI: 10.13248/j.cnki.wetlandsci.2021.01.001]

Xu H Q. 2005. A study on information extraction of water body with the modified normalized difference water index (MNDWI). Journal of Remote Sensing, 9(5): 589-595 (徐涵秋. 2005. 利用改进的归一化差异水体指数(MNDWI)提取水体信息的研究. 遥感学报, 9(5): 589-595) [DOI: 10.11834/jrs.20050586]

Xu X L, Liu J Y, Shao Q Q and Fan J W. 2008. The dynamic changes of ecosystem spatial pattern and structure in the Three-River Headwaters region in Qinghai Province during recent 30 years. Geographical Research, 27(4): 829-838 (徐新良, 刘纪远, 邵全琴, 樊江文. 2008. 30年来青海三江源生态系统格局和空间结构动态变化. 地理研究, 27(4): 829-838) [DOI: 10.11821/yj2008040011]

Xue P F, Li W L, Zhu G F, Zhou H K, Liu C L and Yan H P. 2021. Changes in the pattern of an alpine wetland landscape in Maqu County in the first meander of the Yellow River. Chinese Journal of Plant Ecology, 45(5): 467-475 (薛鹏飞, 李文龙, 朱高峰, 周华坤, 刘陈立, 晏和飘. 2021. 黄河首曲玛曲县高寒湿地景观格局演变. 植物生态学报, 45(5): 467-475) [DOI: 10.17521/cjpe.2020.0288]

Yan P, Zhang Y J and Zhang Y. 2007. A study on information extraction of water system in semi- arid regions with the enhanced water index (EWI) and GIS based noise remove techniques. Remote Sensing Information, (6): 62-67 (闫霈, 张友静, 张元. 2007. 利用增强型水体指数(EWI)和GIS去噪音技术提取半干旱地区水系信息的研究. 遥感信息, (6): 62-67) [DOI: 10.3969/j.issn.1000-3177.2007.06.015]

Yu L and Liu H. 2003. Feature selection for high-dimensional data: a fast correlation-based filter solution//Proceedings of the Twentieth International Conference on Machine Learning. Washington, DC, USA: AAAI Press: 856-863

Zhang L, Gong Z N, Wang Q W, Jin D D and Wang X. 2019. Wetland mapping of Yellow River Delta wetlands based on multi-feature optimization of Sentinel-2 images. Journal of Remote Sensing, 23(2): 313-326 (张磊, 宫兆宁, 王启为, 金点点, 汪星. 2019. Sentinel-2影像多特征优选的黄河三角洲湿地信息提取. 遥感学报, 23(2): 313-326) [DOI: 10.11834/jrs.20198083]

Zhang S Q, Zhou B R, Shi F F, Chen Q and Su S L. 2020. Study on information extraction method of alpine wetland in Qinghai-Xizang Plateau based on remote sensing data of GF-1 Satellite. Plateau Meteorology, 39(6): 1309-1317 (张帅旗, 周秉荣, 史飞飞, 陈奇, 苏淑兰. 2020. 基于高分一号卫星遥感数据的青藏高原高寒湿地信息提取方法研究. 高原气象, 39(6): 1309-1317) [DOI: 10.7522/j.issn.1000-0534.2019.00131]

Zheng Y, Wu B F and Zhang M. 2017. Estimating the above ground biomass of winter wheat using the Sentinel-2 data. Journal of Remote Sensing, 21(2): 318-328 (郑阳, 吴炳方, 张淼. 2017. Sentinel-2数据的冬小麦地上干生物量估算及评价. 遥感学报, 21(2): 318-328) [DOI: 10.11834/jrs.20176269]

Remote sensing feature selection for alpine wetland classification

HUO Xuanlin[1,2], NIU Zhenguo[2], ZHANG Bo[2], LIU Linsong[2], LI Xia[3]

1.School of Earth Science and Resources, Chang'an University, Xi'an 710054, China;
2.State Key Laboratory of Remote Sensing Science, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China;
3.School of Land Engineering, Chang'an University, Xi'an 710054, China

Abstract: Alpine wetlands are an important surface cover type on the Qinghai—Tibet Plateau because they play a key role in water conservation, climate regulation, and biodiversity maintenance. Accurate and timely knowledge of the temporal and spatial distribution of alpine wetlands is necessary for wetland protection and management. The selection of remote sensing classification features is crucial in wetland mapping. Although spectral, texture, and topographic features have been investigated, studies focusing on spectral index features and their mathematical statistical features and feature selection methods are limited. This study aims to classify alpine wetlands from the aspects of mathematical statistical features, alpine wetland types, feature selection methods, and selected feature sets combined with random forest classification algorithm using Sentinel-2 image data and taking the Shouqu Alpine Wetland Reserve as the research site. An in-depth and comprehensive analysis on the spectral index characteristics of alpine wetlands is performed to optimize the classification characteristics of alpine wetlands.

The Gansu Shouqu Alpine Wetland Reserve was used as the research area, and classification characteristics (spectrum, vegetation index, red edge index, and water body index) were obtained on the basis of Sentinel-2 data. Filter and wrapper feature selection methods, including Jeffries–Matusita distance, Spectral Angular Distance (SAD), Euclidean Distance (ED), RF-RFE algorithm, and Relief-F algorithm are utilized to optimize these features. Meanwhile, Z test is applied for quantitative evaluation.

The following conclusions can be drawn from this study. (1) Among the categories of alpine wetlands involved in the classification, rivers and bare land are the easiest to distinguish, followed by grasslands and swamps and then swampy meadows and meadows. MCARI2, NDWI, DVI, EVI, EWI, IRECI, MCARI, TCARI, and UGWI indices can be used to differentiate among adjacent swamps, swampy meadows, meadows, and grasslands. (2) The order of contribution of different index characteristics to wetland information extraction in terms of degree is water body index characteristics > vegetation index characteristics > red edge index characteristics. (3) ED and Relief-F algorithms in the filter method demonstrate excellent performance from the perspective of feature optimization methods. (4) A suitable alpine wetland information extraction method is selected using the indices RDVI, NDVI, MSR, RVI, VIgreen, RNDWI, NDWI, NDWI_B, MNDWI, EWI, and CIre. (5) The mathematical statistics of different classification features indicated that the median feature obtains the best classification result, followed by the average value feature.

We provide detailed results from feature optimization methods, wetland classification optimization index, statistical feature evaluation, and categories involved in alpine wetland classification using multi-dimensional analysis. To the best of our knowledge, this study provides a novel transferable and universal method for the selection of characteristic variables for wetland information extraction.

Key words: remote sensing, wetland classification, alpine wetland, feature selection, Qinghai-Tibet Plateau, Sentinel-2

Supported by National Key Research and Development Program of China (No. 2021YFE0194700); Second Comprehensive Scientific Research Project on Qinghai-Tibet Plateau (No. 2019QZKK0106); National Natural Science Foundation of China (No. 41971390)