引文格式：闫爽,张斌,寇晓康,等. 基于随机森林模型的农村黑臭水体监测方法研究[J]. 地理空间信息, 2026, 24(4):98-102. doi:10. 3969/j.issn. 1672-4623.2026.04.020

基于随机森林模型的农村黑臭水体监测方法研究

闫　爽[1]，张　斌[2]，寇晓康[3,4*]，刘　硕[5,6]

（1. 河北省科学院地理科学研究所 河北省地理信息开发应用技术创新中心，河北 石家庄 050011；
2. 兰州大学资源环境学院，甘肃 兰州 730000；3. 石家庄铁道大学土木工程学院，河北 石家庄 050043；
4. 中核三维地理信息工程技术研究中心，河北 石家庄 050043；5. 河北省地质环境监测院，河北 石家庄 050022；
6. 河北省地质工程勘查院，河北 保定 071051）

摘　要：农村水体是农村地区生态系统必要的组成部分，全面推进乡村污水整治工作，强化黑臭水体监管体制至关重要。本研究基于GF-2影像数据，利用目前最流行的机器学习算法之一——随机森林模型，以可见光波段、改进的水体指数（BOI）、植被指数（NDVI）等数据作为输入特征，对河北省保定市徐水区崔庄镇水体进行提取，并进而对化学需氧量（OD）、氨氮（NH3-N）、总磷（TP）、溶解氧（DO）等水质指标进行遥感反演和水质分级，并对水质分级结果进行精度评估。结果显示，水质分类结果Kappa系数达到0.66，总体分类精度达到79%，证明反演结果的有效性，研究成果能够为农村地区黑臭水体的监测提供技术支撑，也可为雄安新区水环境监测治理提供范例。

关键词：GF-2；随机森林模型；黑臭水体；水质分级

中图分类号：P237　　**文献标志码：**B　　**文章编号：**1672-4623（2026）04-0098-05

Research on Monitoring Method of Black and Odorous Water Body in Rural Areas Based on Random Forest Model

YAN Shuang[1], ZHANG Bin[2], KOU Xiaokang[3,4], LIU Shuo[5,6]

(1. Institute of Geographical Sciences, Hebei Academy of Sciences, Hebei Technology Innovation Center for Geographic Information Application , Shijiazhuang 050011, China;
2. College of Earth and Environmental Sciences, Lanzhou University, Lanzhou 730000, China;
3. School of Civil Engineering, Shijiazhuang Tiedao University, Shijiazhuang 050043, China;
4. CNNC Engineering Research Center of 3D Geographic Information, Shijiazhuang 050043, China;
5. Hebei Geological Environment Monitoring Institute, Shijiazhuang 050022, China;
6. Hebei Geological Engineering Exploration Institute, Baoding 071051, China)

Abstract: Rural water bodies are an indispensable component of rural ecosystems, and it is crucial to comprehensively promote rural sewage treatment and strengthen the regulatory system for black and odorous water bodies. Based on GF-2 image data, taking the visible light band, improved water index (BOI), vegetation index (NDVI) and other data as input features, we utilized one of the most popular machine learning algorithms, the random forest model to extract water from Cuizhuang Town, Xushui District, Baoding City, Hebei Province. Furthermore, we carried out remote sensing inversion and water quality classification of chemical oxygen demand (OD), ammonia nitrogen (NH3-N), total phosphorus (TP), dissolved oxygen (DO) and other water quality indicators, and evaluated the accuracy of water quality classification results. The results show that the Kappa coefficient of water quality classification reaches 0.66, and the overall classification accuracy reaches 79%, proving the effectiveness of inversion results. The research results can provide support for the treatment of black and odorous water bodies in rural areas and an example for water environment monitoring and treatment in Xiong'an New Area, Hebei Province.

Key words: GF-2, random forest model, black and odorous water body, water quality classification

随着我国城市化进程加快，黑臭水体问题日益严重[1-2]。传统水体监测方式存在诸多弊端，多数学者利用遥感数据开展黑臭水体监测方法研究，相关研究可分为阈值法、分类法、机器学习算法三类[3-9]。雄安新区设立后，其周边地区生态环境问题受到关注。本文选取河北省保定市徐水区为研究区，基于GF-2影像数据，利用随机森林模型实现对黑臭水体的提取，并利用“遥感目视解译+现场勘察”获取的水体类别对水体提取结果进行验证，然后基于提取的水体进行水质反演及评价，以期为雄安新区周边生态环境建设和

项目来源：国家自然科学基金青年科学基金资助项目（41801277）。
第一作者简介：闫爽（1991—），硕士，主要研究方向为遥感科学应用，E-mail：imyanshuang@163.com 。
通信作者：寇晓康（1987—），博士，副教授，主要研究方向为遥感科学应用，E-mail：kou_xiaokang@163.com。

可持续发展提供数据支持[10-14]。

1　黑臭水体识别方法

本研究提出的黑臭水体识别方法包含2个步骤：①利用GF-2遥感影像，选择合适的特征组合（蓝光波段、绿光波段、红光波段、NDVI、NDWI、BOI等），利用随机森林模型粗略提取研究区内所有水体，对水体进行膨胀、腐蚀、平滑处理，消除大部分噪声得到精细水体；②基于水体提取结果，反演化学需氧量、氨氮、总磷、溶解氧等黑臭水体水质指标，并进行水体水质分级，然后结合实地调研得到的水体污染结果对反演的水质等级进行验证，评价反演精度。

1.1　基于随机森林模型的水体提取

随机森林属于集成算法，弥补了单棵决策树容易发生“过拟合”现象的不足，提高了模型的泛化能力，具有训练速度快、抗噪声能力强等优点，其工作流程为：用户输入训练数据后，随机森林模型的决策树根据数据特征进行分类，然后统计每棵树的分类结果，通过计算模型的基尼系数（式（1）），选择最优投票对应的结果作为最终预测的类别[15]。其中，数据特征反映对象的模式属性，其将模式特征定义在一个特定的空间内，进而利用不同对象间的特征差异对特征空间进行划分，达到分类的目的，合适的特征决定了分类的精度[16-17]。

$$\mathrm{Gini}(D)=1-\sum_{k=1}^{|y|}p_k^{\,2} \tag{1}$$

式中，$\mathrm{Gini}(D)$为数据集D的基尼系数，基尼系数越小，则该数据集的纯度越高，也说明决策树之间的分类结果一致性越高；p_k为抽到2个相同样本的概率；y为抽取的次数。

通过实地调研，研究区包含较多的绿藻水体、一般水体、一般水体2、裸地、植被、建筑1、建筑2，因本研究对象为水体，所以只对非水体分类进行大致分类。通过目视解译，根据GF-2影像的空间特征和光谱特征选择样点数据，对照天地图和Landsat8影像进行检验，确保训练数据集的准确性。模型输入特征选取如下：

1）植被特征。植被在近红外波段的反射值相较于其他地物偏高，同时根据现有研究，归一化差分植被指数（NDVI）常用于植被长时序监测[18]，水体和植被对近红外波段的吸收特性呈现相反的趋势，可用于区分水体与植被，因此选取NDVI作为植被提取特征。

2）水体特征。一般水体的反射主要集中在蓝绿光波段，在红波段和近红外波段的反射值较低，并且整体反射率偏低，波段间的过渡平稳。归一化差分水体指数（NDWI）对水体遥感识别较为灵敏[19]，通过利用植被和水体对近红外波段的反射特性相反，突出影像中的水体，BOI是改进后的水体指数，通过加强对植被的抑制作用，提升水体识别能力。一般水体相较于绿藻水体在红绿波段的反射值具有一定的差异，可利用红波段和绿波段的差值（$R-G$）来区分绿藻水体和一般水体。水体在全色波段（PAN）一般呈现深黑色，可以作为补充指标区分水体和其他地物，提高分类精度。综上所述，选取NDWI、BOI、$R-G$、PAN作为水体提取特征。

3）裸地特征。裸地对近红外波段的反射值偏高，并且随波长的增加平缓提升，但是裸地在近红外和绿波段的差值较大，利用这2个波段间的差值（$N-G$）能够区分裸地与其他地物。所以选取$N-G$作为裸地提取特征。

4）建筑特征。建筑物整体反射值较高，2种不同材质的建筑在可见光波段的反射值整体偏高，所以选取可见光反射值波段之和（$B+G+R$）区分建筑与其他地物。

1.2　水质反演

由于不同种类的水体中溶解含氧量以及氮磷钾等化学物质的含量不同，导致水体对可见光的反射率具有差异，因此可根据该不同波段的反射率差异对水体的化学需氧量浓度、总磷浓度、氨氮浓度、溶解氧浓度进行反演，从而据此划分水质等级。

1）化学需氧量浓度和总磷浓度。化学需氧量浓度常用于评价水体受到有机物污染的程度，一般而言，废水和受污染水中的化学需氧量浓度较高[11]。本文参考赵倩等[11]的模型反演化学需氧量浓度和总磷浓度，计算公式如下：

$$C_{OD}=\frac{\dfrac{b_3}{b_2}-0.577\,7}{0.007} \tag{2}$$

$$C_{TP}=4.970\,3\times\left(\frac{b_3}{b_2}\right)^{11.618} \tag{3}$$

式中，C_{OD}、C_{TP}分别为化学需氧量浓度和总磷浓度，单位为mg/L；b_3为GF-2的红色波段；b_2为GF-2的绿色波段。

2）氨氮浓度。钱彬杰[20]使用水质例行监测数据构建了水体氨氮浓度模型，本文参考此模型反演氨氮浓度，计算公式为：

$$C_{NH3N}=4.519\times\left(\frac{b_2-b_3}{b_2+b_3}\right)^2-2.1\times\frac{b_2-b_3}{b_2+b_3}+0.47 \quad (4)$$

式中，C_{NH3N} 为水体氨氮浓度，单位为mg/L；b_2 为GF-2的绿波段；b_3 为GF-2的红波段。

水体溶解氧浓度。钟炜萍[21]使用水质检测数据构建了水体溶解氧模型，在钦州湾水质反演中取得较好结果，本文参考此模型反演溶解氧浓度，计算公式为：

$$C_{DO}=771.854\times\left(\frac{1}{b_3}\right)-1.476\times\frac{b_3^2}{b_2^2}+6.435 \quad (5)$$

式中，C_{DO} 为水体溶解氧浓度，单位为mg/L；b_2 为GF-2的绿波段；b_3 为GF-2的红波段。

1.3 水质等级划分

本文参考《地表水环境质量标准》(GB3838-2002)，基于化学需氧量、总磷浓度、氨氮浓度、溶解氧等指标将水质分为Ⅰ、Ⅱ、Ⅲ、Ⅳ、Ⅴ共5类，将前三类归类为一般水体，后两类分别归类为轻度黑臭水体、重度黑臭水体。

2 研究区与数据

2.1 研究区域概况

河北省保定市徐水区（38°09′~39°09′N，115°19′~115°46′E）（图1）属平原地形，高程变化较小，地理位置优越，地处保定市辖区范围中央地带。东部紧邻雄安新区，靠近水资源丰富的白洋淀；境内水资源丰富，存在3条自然河流（漕河、瀑河、萍河）并汇入白洋淀，另有人工河道（南水北调中线工程）为其供水，为当地居民生活生产提供必要保障。徐水区下设14个乡镇，从西北到东南水资源逐渐增多，东南部分布较多的小型湖泊及水稻种植地，适宜开展本文的黑臭水体监测工作。崔庄镇位于徐水区东部，拥有较多分散的湖泊并且瀑河流经此地，监测该地区水质可为黑臭水体治理提供依据，保障居民良好的生活环境，所以选择徐水区崔庄镇作为研究区。

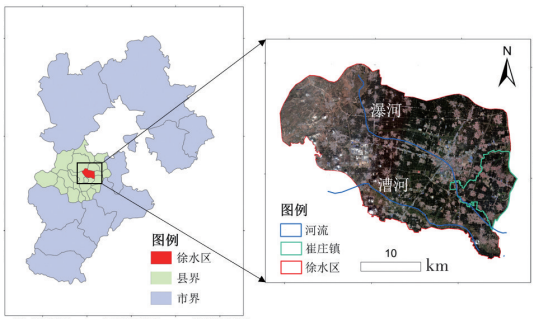


图1 徐水区GF-2影像

审图号：GS(2024)0650

2.2 研究数据

GF-2卫星搭载由2台高分辨率相机组成的PMS传感器，可以返回全色0.8 m，多光谱3.2 m的光学遥感数据。本研究所用的GF-2遥感影像数据来源于河北省高分辨率对地观测系统数据与应用中心，申请得到的GF-2数据为1级产品，采集时间为2021年4月23日。利用ENVI5.3软件和GF-2号影像自带的定标系数进行多光谱影像和全色影像的辐射定标、大气校正和正射校正，得到包含蓝光、绿光、红光、近红外、全色共5个波段的影像。

本文利用Landsat8 Collection 2 Tier 1 TOA Reflectance（陆地卫星8号集合2一级TOA反射率，简称Landsat8）影像作为辅助数据来采集样本。Landsat8数据分辨率较低，但波段较多，易于区分不同地物。在采集地物样本时，通过对比GF-2和Landsat8数据，保证采集的样本点具有相似的光谱特征，从而确保样本数据集的真实性。本研究将训练样本分为4类，分别为水体、裸地、植被、建筑，共采集样本18 093条数据，训练指标详见1.1节。本文以“遥感目视解译+现场勘察”获取的水体类别数据作为检验数据，用于评价结果精度，该数据集包括徐水区19处水体的现场照片及水体类型，分布在徐水区中部及东部等含较多闭塞坑的区域。为了更好地分析可能的污染源，本文采用ESRI开发的全球10 m分辨率土地覆盖原始数据作为辅助数据，该数据是基于哨兵2号卫星数据结合机器学习开发，将全球覆盖的土地利用类型分为九类，分别是水体、树木、淹没植被、农作物、建成区、裸地、冰雪、云层、牧场。

3 结果分析

3.1 水体提取

通过样本训练，调节分类器的各项参数，使随机森林模型获得较强的抗干扰能力，提升分类精度，最终选取参数如下：决策树数量为290，决策树最大深度为18，分割内部节点所需的最小样本数为9，分割叶子节点的最小样本数选定为9，最大特征数为3。将数据集按7∶3分为训练数据集和验证数据集，基于上述参数，使用随机森林模型进行水体提取。但是受到样本数量和精度的影响，分类结果仍然会产生一些椒盐噪声和误判，需要对初步提取后的水体像元进行形态学运算，剔除噪声，这些噪声大多是单独的像元，所以可设定面积阈值将其剔除。本文首先采用广度优先搜索算法，寻找像元集合的最大连通区域，从影像

的第一个像元开始搜索，若其8个方向上有与其相邻的水体像元即可并入该水体集合，更新该集合的面积，以此类推直至将所有的像元搜索完毕，利用200 m^2作为阈值可剔除绝大部分噪声，同时保留了小面积的水体集合，考虑到人工道路的形态学特征，同时引入了形态学参数（长宽比 > 5），剔除了研究区内细长且狭窄的区域。得到最终水体像元空间分布图。由图2a可以看出，粗提取得到的水体在研究区北部存在零碎像元。通过对比天地图发现，这些零碎像元是模型错分的非水体地物，集中分布在北部居民区。并且，在崔庄镇南部分布着一条公路被错误识别为水体，则根据形态学特征将其剔除，将这些不连续像元去除掉之后，能够更加清晰地展示水体的空间分布。

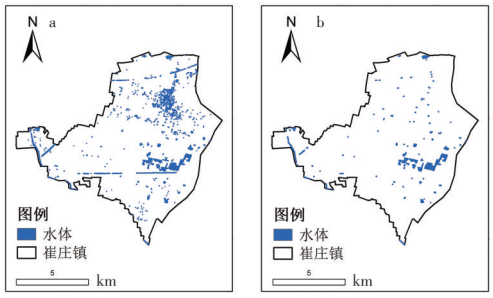


图2　剔除噪声前后水体分布对比

审图号：GS(2024)0650

3.2　水质反演

基于模型反演得到的水体数据，利用式（2）~（5）反演水体的化学需氧量浓度、总磷浓度、氨氮浓度、溶解氧浓度，再根据《地表水环境质量标准》（GB3838-2002）中对以上指标的综合划分方法，绘制了研究区内的水质等级空间分布图（图3）。化学需氧量浓度整体保持平衡，化学需氧量浓度范围主要在Ⅰ-Ⅲ类，水体受有机物污染较少。总磷浓度较少，范围主要分布在Ⅰ-Ⅲ类，局部地区存在化学需氧量和总磷浓度超标的现象，主要集中在瀑河和农田区，这些地区可能存在“向河道内排污”和“农药使用不当”等行为，造成化学物质积聚；整体的氨氮浓度基本达标，说明水体中含有较少的氨氮物质，该区域受氨氮污染较少，可初步推测，该区域排放工业废水较少或没有工业废水排放，农作物对氮化合物等肥料实现了饱和利用；水体溶解氧量整体偏高，说明水中含有较多的氧气，能够满足生物生存条件，说明该区域水中有机物含量较少，即微生物消耗氧气较少，使得水体中溶解氧含量较高。

由图3可知，水体主要集中在崔庄镇中东部的农业种植区，其中分布较多的轻度黑臭水体及重度黑臭

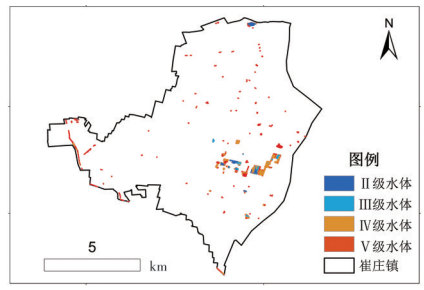


图3　崔庄镇水体水质分级图

审图号：GS(2024)0650

水体。该区域也分布着池塘，被识别为一般水体，即水质达标；瀑河河道内的水体某些指标超限，被识别为重度黑臭水体，可初步推测水体受到严重污染；北部分布一所湖泊，水质优良，说明该区域受到的污染较少。为了探究黑臭水体污染物的潜在来源，本研究分别以轻度黑臭水体和重度黑臭水体为中心，构建30 m缓冲区，并与ESRI土地利用数据进行叠加分析。随后，计算缓冲区内农作物、建成区、牧场3类地物的面积，以量化不同土地利用类型对黑臭水体的潜在影响。其中黑臭水体缓冲区内的作物区占据了较大比例，其次是建成区，最后是牧场。由于当地生产条件的制约，大面积区域为农田。而随着经济的不断发展，建成区的面积逐渐扩大，生活垃圾也随之累积，成为了新的主要污染源。下面根据土地利用来分析黑臭水体的形成原因。①作物区通常是农业用地，可能使用大量的化肥和农药。若这些化肥和农药没有得到有效的管理和回收，部分会被雨水冲刷进入水体，增加水中的氮、磷含量，导致富营养化，从而引发水质恶化。尤其在雨季，表面径流和灌溉水可能带入大量的养分，导致水体的藻类繁殖过度，从而形成“黑臭”现象。②建成区通常指城市或乡村建设区，包括住宅、道路、工业设施等，受到人类活动影响较大。这些区域的主要污染源包括生活污水、工业废水、垃圾废弃物等，尤其是未经处理的生活污水会直接流入水体，带入大量有机污染物（如氮、磷、油脂）和病原体，造成水体有机物浓度超标，微生物需要消耗水中大量氧气才能降解这些有机物，造成水体溶解氧含量降低，严重时甚至会导致河流内生物死亡，进而引起水体颜色加深、气味加剧。堆积在河道内的垃圾会减缓河水流速，水面漂浮物阻碍阳光射入水体，破坏水生态系统的平衡，造成水体水质恶化。建成区的排水系统如果不健全，雨水径流也可能带入垃圾、重金属等污染物，进一步加剧水质的恶化。实地调研发现，瀑河两岸分布着较多村庄，村民生活产生的生活垃圾成为主要污染源。为了日常生活的便利

或者村内缺少统一的垃圾处理站，一部分居民选择向河道排放污水和堆积垃圾。③牧场主要是养殖动物的区域，动物的排泄物（如粪便、尿液）可能带入水体，尤其是牲畜数量较多的牧场。在没有良好管理的情况下，牲畜排泄物中的氮、磷等营养物质会进入水体，导致其富营养化。此外，牧场地面覆盖物较少，雨水容易将污染物冲刷入水体。

3.3 水质反演精度评价

本文基于混淆矩阵分析模型精度，包括整体正确识别率、黑臭水体错分率、黑臭水体漏分率、正常水体错分率、正常水体漏分率和Kappa系数6个精度指标。从表1中可以发现：水质分类结果Kappa系数达到0.66，总体分类精度达到79%，证明反演结果的有效性。分类别看：对于一般水体，生产者精度为87.5%，漏分率为12.5%，有1个轻度黑臭水体和1个重度黑臭水体被误判，用户精度为77.8%，错分率为22.2%；对于轻度黑臭水体，生产者精度为50%，即漏分率为50%，无其他类型水体误判为轻度黑臭水体，用户精度达到100%，错分率为0；对于重度黑臭水体，生产者精度为85.7%，漏分率为14.3%，有1个一般水体和1个轻度黑臭水体被误判为重度黑臭水体，用户精度达75%，错分率达25%。

表1 水体分级混淆矩阵

实测值预测值	一般水体	轻度黑臭水体	重度黑臭水体	漏分率/%
一般水体	7	0	1	12.5
轻度黑臭水体	1	2	1	50
重度黑臭水体	1	0	6	14.3
错分率/%	22.2	0	25	
Kappa系数		0.66		
整体正确识别率/%		79.0		

4 结 语

本文以河北省保定市徐水区崔庄镇为研究区，利用随机森林模型构建分类器，以可见光波段、改进的水体指数（BOI）、植被指数（NDVI）等指标作为输入特征，基于GF-2影像结合形态学运算完成细小水体提取；基于水体提取结果对化学需氧量（OD）、氨氮（NH3-N）、总磷（TP）、溶解氧（DO）等水质指标进行遥感反演和水质分级，并利用混淆矩阵进行精度评估，Kappa系数达到0.66，整体正确识别率达到79%。研究成果获取了徐水区崔庄镇不同水质水体的空间分布，可为政府部门进行黑臭水体监管治理提供依据，并为雄安新区水环境监测治理提供了范例。

参考文献

[1] 付丽,郭雯雯,宋开山,等. 2020年长春市黑臭水体遥感识别研究[J]. 湿地科学,2022,20(4):537-547

[2] 张宝东,王彪,吴艳兰,等. 安徽省农村黑臭水体特征分析及识别[J]. 生态环境学报,2024,33(8):1257-1268

[3] 陈凯,肖雪,李冠达,等. 西南多源遥感影像水资源监测方法研究[J]. 地理空间信息,2024,22(10):14-18

[4] 张永红. 基于高分卫星的辽宁省城市黑臭水体遥感监测[J]. 皮革制作与环保科技,2023,22(4):112-114

[5] 李玲玲,李云梅,吕恒,等. 基于决策树的城市黑臭水体遥感分级[J]. 环境科学,2020,41(11):5060-5072

[6] 曹云,杭鑫,高艺,等. 利用高分二号影像对城市黑臭水体遥感监测:以南京市主城区为例[J]. 四川环境,2023,42(1):208-217

[7] 张颖. 基于遥感数据对某农村区域黑臭水体解译识别[J]. 绿色科技,2024,26(12):71-77

[8] 王妍,姚杰,杨朴,等. 北京市黑臭水体治理的动态遥感监测及影响因素分析[J]. 环境工程学报,2022,16(9):3092-3101

[9] 解明权,闵松寒,杨辉,等. 基于贝叶斯卷积网络的农村黑臭水体遥感识别算法研究[J]. 环境科学学报,2024,44(3):215-226

[10] 顾佳艳,何国富,占玲骅,等. 上海市黑臭水体光谱特征分析及遥感识别方法探究[J]. 环境科学研究,2022,35(1):70-79

[11] 赵倩. 广州市黑臭水体评价模型构建及污染溯源研究[D]. 淄博:山东理工大学,2021

[12] 张宁宁. 基于GF-2的哈尔滨市黑臭水体动态监测与评价[D]. 哈尔滨:哈尔滨师范大学,2022

[13] 史佳睿,申茜,彭红春,等. 基于三种高分辨率影像样本迁移学习的城市水体提取研究[J]. 遥感技术与应用,2022,37(3):731-738

[14] 陈紫怡,牛犇牛,杨海燕,等. 我国黑臭水体现状、治理及展望[J].中国水运,2025,46(2):74-77

[15] 王奕森,夏树涛. 集成学习之随机森林算法综述[J]. 信息通信技术,2018,12(1):49-55

[16] 韩帅兵,苏成,寇晓康,等. 基于随机森林模型的山区铁路线暴雨洪灾风险评估:以朔黄铁路原平段为例[J]. 地理与地理信息科学,2024,40(2):44-50

[17] 徐聪,于丽君,柳泽,等. 基于特色村镇的多源影像最优融合分类方法探讨[J]. 地理空间信息,2024,22(3):66-71

[18] 龙昶宇. 利用时序NDVI分析安徽省植被参数时空变化[J]. 地理空间信息,2024,22(5):88-92

[19] 罗福正,杨柳,卢彬,等. 基于多时相Sentinel-1影像的江南地区水稻信息提取[J]. 地理空间信息,2024,22(11):73-77

[20] 钱彬杰,黄迪,谭晓慧,等. 基于Landsat8 OLI影像的沭河临沂段氮磷污染物反演[J]. 山西建筑,2019,45(1):159-161

[21] 钟炜萍,姚焕玫,陈华权. 基于自动监测和Sentinel-2影像的钦州湾溶解氧反演模型研究[J]. 广西科学院学报,2020,36(4):392-398