文章编号：1672-8785(2023)11-0042-09

基于 Sentinel-2 遥感数据的上海市河道水质参数反演研究

季铁梅[1] 姚勇华[1] 葛婷婷[1] 杨 喆[2,3*]

(1. 上海市水文总站，上海 200232；
2. 中国科学院上海技术物理研究所，上海 200083；
3. 中国科学院大学，北京 100049)

摘 要：由于上海市河湖水网密布，科学有效地监测河湖水质，有利于巩固河湖治理成果，为新形势下水资源的保护和管理服务。研究目的是实现卫星遥感技术在城市水体水质监测中的有效应用。基于 Sentinel-2 多光谱影像，采用机器学习技术建立了城市河流水质参数反演模型，对 2019～2021 年上海市 103 条河流的溶解氧(Dissolved Oxygen，DO)、高锰酸盐指数(Permanganate index，COD_{Mn})、氨氮(Ammonia Nitrogen，NH_3-N)和总磷(Total Phosphorus，TP)4 种水质参数进行了遥感反演。分析了上海市主要河流水质参数的时空变化特征，并对上海市水环境进行了评价。结果表明，DO、COD_{Mn}和 TP 三个水质指标的反演精度优于 80%，NH_3-N 的反演精度优于 70%；4 种水质参数所属水体类别均优于 V 类；第一、第四季度水质优于第二、第三季度。

关键词：城市河流；遥感；机器学习；Sentinel-2；水质参数

中图分类号：TP79 **文献标志码**：A **DOI**：10.3969/j.issn.1672-8785.2023.11.007

Research on Retrieval of Water Quality Parameters in Rivers of Shanghai Based on Sentinel-2 Remote Sensing Data

JI Tie-mei[1], YAO Yong-hua[1], GE Ting-ting[1], YANG Zhe[2,3*]

(1. *Shanghai Hydrology Administration, Shanghai 200232, China*;
2. *Shanghai Institute of Technology and Physics, Chinese Academy of Sciences, Shanghai 200083, China*;
3. *University of Chinese Academy of Sciences, Beijing 100049, China*)

Abstract: Due to the dense network of rivers and lakes in Shanghai, scientific and effective monitoring of river and lake water quality is conducive to consolidating the achievements of river and lake management, and serving for the protection and management of water resources in the new situation. The purpose of this paper is to make satellite remote sensing technology be applied effectively in urban water quality monitoring. The re-

收稿日期：2023-05-04

基金项目：上海市水务局科研项目(沪水科 2021-10)；中国科学院上海技术物理研究所创新专项(CX-445；CX-363)

作者简介：季铁梅(1983-)，女，江苏南通人，硕士，主要从事河湖水质监测与分析评价研究。

*通讯作者：E-mail：yangzhe@mail.sitp.ac.cn

search methods are as follows: Based on Sentinel-2 multi-spectral images, an inversion model of urban river water quality parameters is established by using machine learning technology. Dissolved oxygen, permanganate index, ammonia nitrogen and total phosphorus of 103 rivers in Shanghai from 2019 to 2021 are inverted by remote sensing, and the spatio-temporal variation characteristics of water quality parameters of main rivers in Shanghai are analyzed. The water environment of Shanghai is evaluated. The results show that the inversion accuracy of DO, COD_{Mn} and TP is better than 80%, and the inversion accuracy of NH_3-N is better than 70%. The four water quality parameters are better than that of class V, and the water quality in the first and fourth quarters is better than that in the second and third quarters.

Key words: urban river; remote sensing; machine learning; Sentinel-2; water quality parameters

0 引言

河流在人类历史文明发展中发挥了重要作用。水资源对于一个城市来说至关重要，因为它提供了饮用水、交通运输动力和工业生产用水，并且可以作为景观，所以大多数著名城市都建在河流附近[1]。然而，随着城市社会和经济的快速发展，生活需求迅速增长，生活用水、农业用水和工业用水的排放超过了水体的自清洁能力，对城市河流和湖泊造成了严重污染。城市水质恶化带来饮用水安全问题并破坏生态环境，进而影响人类健康和生物多样性。

水质监测是水质评价的重要组成部分，旨在了解城市水体(特别是河流和湖泊)的水质。尽管传统的水质监测方法(包括人工现场采样、实验室测量或自动站测量)精度高，但人工方法是劳动密集型的，而且自动站的建造和维护也需要昂贵的成本[2]。虽然这些方法可能具有较高的准确性，但单个样本只能反映特定采样点的水质，在表征整个水面的水质方面受到限制。在许多情况下，水质监管部门需要对整个水面的水质情况进行全面的了解，以便进行水质管理。此外，大多数水样检测都需要试剂，而处理检测产生的废物也很昂贵。自 20 世纪 70 年代以来，遥感技术因其时间和空间特征已广泛用于水质监测[3]。

近年发射的 Sentinel-2 卫星生成图像的空间分辨率为 10 m，时间分辨率为 5 天，为高频率监测小型水体水质提供了机会[4]。由于 DO、COD_{Mn}、NH_3-N 和 TP 是判断水质等级的关键参数[5]，本研究使用 Sentinel-2 图像反演上海市河流的这些参数。对 Sentinel-2 图像波段和波段组合进行相关性分析，以确定反演每个水质参数的最合适波段。采用极端梯度提升树(XGBoost)机器学习方法建立一种实用、有效的非光学活性参数水质监测策略，有助于城市水资源管理。

1 研究区和数据

1.1 研究区概况

上海位于长江下游的三角洲地区(120°52′E～122°12′E，30°40′N～31°53′N)，如图 1 所示。上海江河湖泊密布，水域面积为 649.2 km^2，河流密度为 4.79 km/km^2，地表水域比率为 10.24%。主要的自然水系是长江、黄浦江和淀山湖，为居民提供了大部分生活用水。作为中国最大的城市之一，上海 2020 年人口超过 2400 万。上海依水而生，因水而兴，纵横交叉的河湖的水质与人民群众的幸福生活息息相关。

1.2 遥感影像数据

2019～2021 年的上海市 Sentinel-2 数据来自欧洲航天局(ESA)数据分发系统(https://scihub.copernicus.eu)。Sentinel-2 任务是由两颗卫星(Sentinel-2A/B)所组成的星座每 5 或 10 天对地球地表进行一次全球覆盖，旨在监测全球地表状况的变化[6]。MSI 图像包含 13 个光谱带，范围从可见光(VNIR)和近红外(NIR)到短波红外(SWIR)。波段 2(B2，492 nm)、波段 3(B3，559 nm)、波段 4(B4，665 nm)和波段 8(B8，833 nm)的空间分辨率为 10 m×10 m。波段 5(B5，703 nm)、波段 6(B6，739

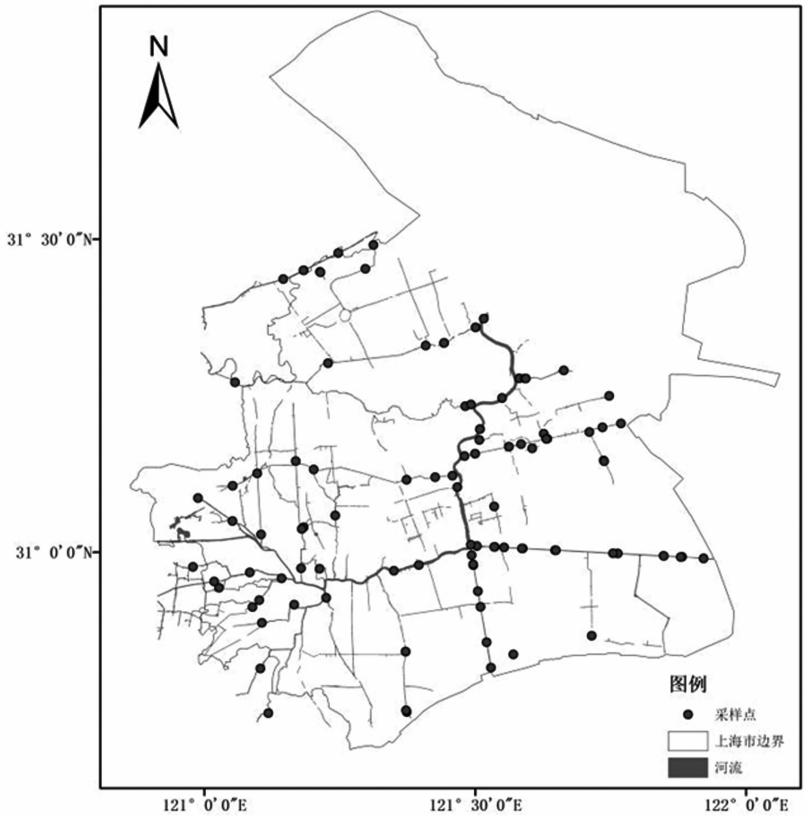


图 1 研究区、提取的河流和水质采样点

nm)、波段 7(B7，779 nm)、波段 8A(B8A，864 nm)、波段 11(B11，1610 nm)和波段 12(B12，2185 nm)的空间分辨率为 20 m×20 m。波段 1(B1，442 nm)、波段 9(B9，943 nm)和波段 10(B10，1376 nm)的空间分辨率为 60 m×60 m。B1、B9 和 B10 为三个大气波段，因此本研究未对其进行分析。

1.3 实测水质数据

根据 2019～2021 年上海市水质月度监测数据，每月共有 4000 个左右的断面实测水质浓度。将水质监测月度数据与收集到的卫星图像进行筛选，产生星地同步数据集，并将其用于水质参数反演模型的建立。高质量的输入数据是模型精度的保证。为了在保证质量的同时获取更多的数据，我们采用以下几个标准筛选数据：(1)结合实测记录和卫星影像目视去除云、太阳耀斑和藻华覆盖的样点；(2)使用±6 h 的时间窗口产生星地同步数据集；(3)每个样点 3×3 窗口的其他像元都确定为水体像元后才会被保留，以保证该样点在空间的均一性[7]。按照上述标准，共得到 139 个 Sentinel-2 星地同步数据点。

2 方法

2.1 遥感影像预处理

下载到的 Sentinel-2 影像为 L1C 级产品，需要进行大气校正才能实现后续处理。使用 ESA 开发的独立大气校正模块 Sen2cor 进行大气校正，获得 L2A 级地表反射率产品。通过重复上述预处理步骤获得图像的遥感反射率。使用 ENVI 软件中的 Band Math 工具将影像像元值除以 10000，得到 Sentinel-2 图像的归一化遥感反射率。之后采用 SNAP 软件将不同空间分辨率的图像重采样至 10 m，合成图像波段并使用归一化水体指数(NDWI)提取水体信息。接着通过对数据进行质量检查共获得了 25 幅图像。

2.2 机器学习算法

在本研究中，利用目前机器学习领域使用最广泛的高效算法之一的 XGBoost 来训练 4 个

水质参数反演模型。XGBoost 是一种通过融合许多回归树来构建强分类器的提升树模型，具有良好的预测精度，广泛应用于机器学习。该算法通过不断增加树来实现特征分裂：每棵树有若干个叶子结点，每个叶子结点对应一个分数，最后每棵树中对应分数的和就是预测值[8]。

模型的输入是 10 维的特征，除了 B2、B3、B4、B5、B6、B7、B8 和 B8A 以外，还使用了找出的 8 个与 DO、COD_{Mn}、NH_3-N 或 TP 浓度相关的波段比值。这些指数与不同水体中的水质浓度存在关联，能够加快模型收敛。随机将星地同步数据分为两类：97 个用来建模，42 个用来验证。

2.3　统计分析指标

主要采用均方根误差(Root-Mean Squared Error，RMSE)、平均相对误差百分比(Mean Absolute Percentage Error，MAPE)、平均绝对误差(Mean Absolute Error，MAE)作为统计分析指标。它们已被证实能够较为完整地评估水体水质遥感的算法性能[9]。

$$RMSE = \left[\frac{\sum_{i=1}^{n}(E_i - M_i)^2}{n}\right]^{1/2} \tag{1}$$

$$MAPE = 100\% \times Mean\left[\sum_{i=1}^{n}|E_i - M_i|/M_i\right] \tag{2}$$

$$MAE = 10^Y\ Y = \left[\sum_{i=1}^{n}|\log_{10}(E_i) - \log_{10}(M_i)|/n\right] \tag{3}$$

式中，n 是样本数，E 是估算值，M 是实测值。在计算时对 MAE 进行 log10 变换是为了更好地评估非高斯分布的数据集。MAE 计算的是绝对值，但可以转换为百分比(MAE 是相对差异)。比如，1.50 的 MAE 代表相对差异为 50%。

3　结果

3.1 水质参数反演模型精度

表 1 列出了各水质参数的反演模型在验证集数据上的精度结果。由于 NH_3-N 和 TP 为非正态分布，我们筛选了部分实测值，并调整了建模数据和验证数据的比例。从建模角度来看，DO、COD_{Mn}和 TP 三个水质指标的反演精度优于 80%，NH_3-N 的反演精度优于 70%。

3.2　水质参数反演结果及时空变化分析

3.2.1　水质年际变化

DO 的年平均分布呈现出越宽的河道浓度越高的空间特征。黄浦江、拦路港-泖河-斜塘、太浦河、北横港和大治河等较宽河道的年平均 DO 浓度基本在 7 mg/L 左右，明显优于较窄河道(见图 2)。COD_{Mn}指数在三年中的总体浓度没有明显的变化，大河道浓度基本为 3～4.5 mg/L，其空间分布表现为东部河道浓度低于西部河道(见图 3)。三年中 NH_3-N 年平均浓度最高出现在 2020 年，而三年中浓度较高的地区都出现在青浦区的北横港、拦路港-泖河-斜塘、淀浦河附近(见图 4)。TP 的年际变化呈现出逐年降低的趋势，2019 年平均浓度为 0.15～0.2 mg/L 左右，2020～2021 年平均浓度为 0.1～0.15 mg/L 左右(见图 5)。值得注意的是，三年中 TP 浓度较低的区域也在青浦区的北横港、拦路港-泖河-斜塘、淀浦河附近，这与 NH_3-N 的情况相反。

3.2.2　水质季度变化

DO 浓度最高的是第一季度(1～3 月)，最低的是第三季度(7～9 月)，如图 6 所示。COD_{Mn}浓度最低的是第四季度(10～12 月)，最高的是第二季度(4～6 月)。COD_{Mn}的季度空间分布也呈现出明显的东部低于西部的特征。

表 1 各水质参数反演模型误差

水质参数	验证点数量	MAPE	MAE	RMSE
DO	42	17.81%	1.18	1.50 mg/L
COD_{Mn}	42	15.04%	1.15	0.63 mg/L
NH_3-N	20	28.24%	1.29	0.15 mg/L
TP	30	19.89%	1.21	0.03 mg/L

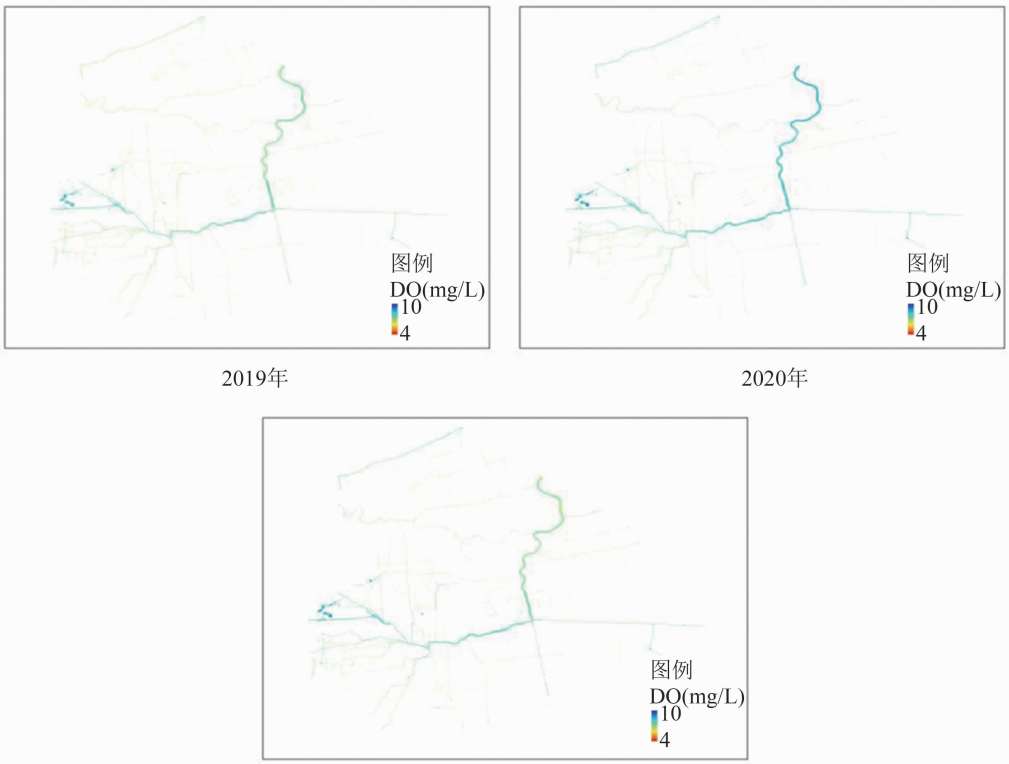


图 2 DO 在 2019～2021 年的平均结果

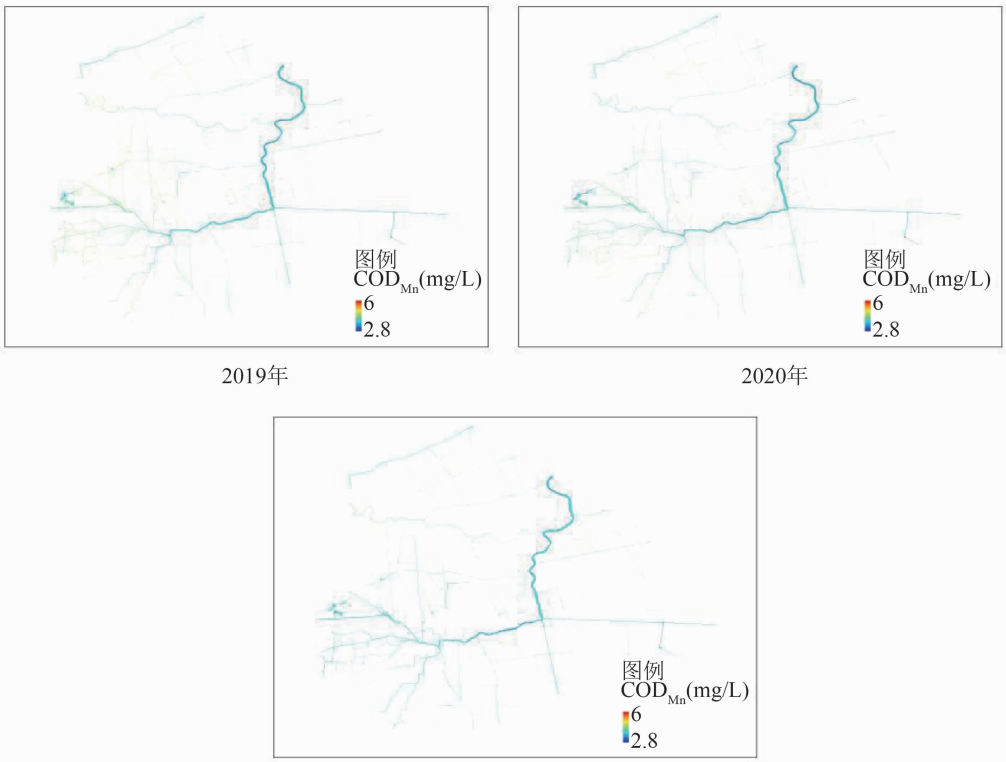


图 3 COD_{Mn}在 2019～2021 年的平均结果

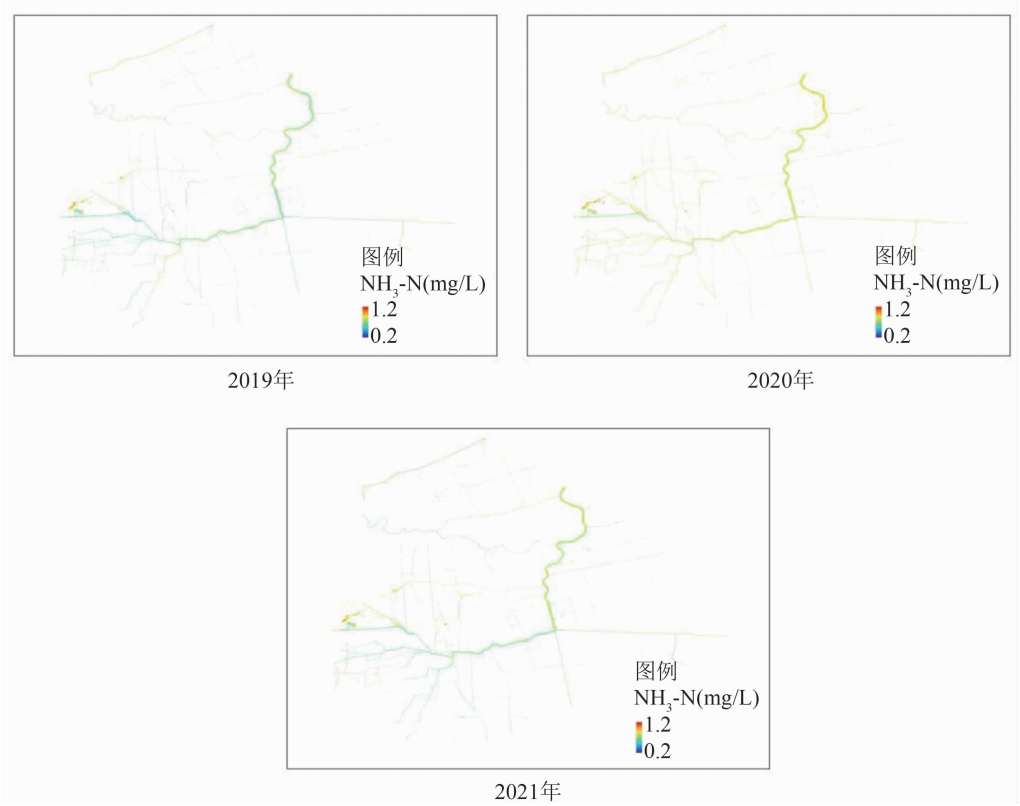


图 4 NH_3-N 在 2019～2021 年的平均结果

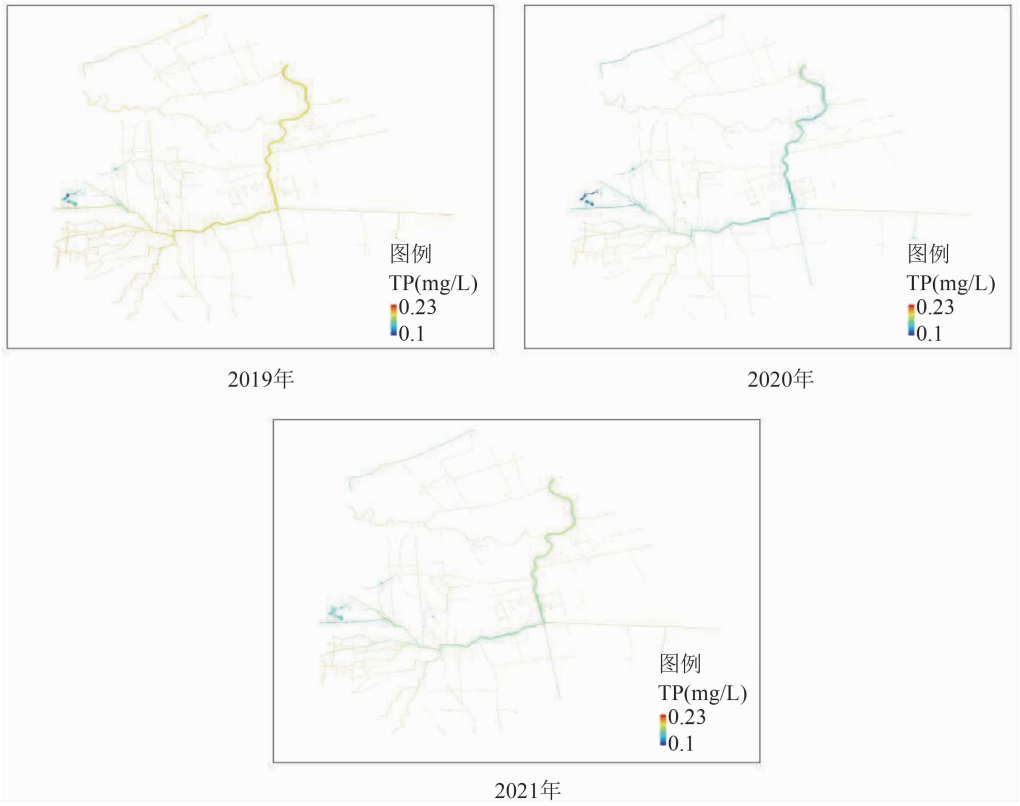


图 5 TP 在 2019～2021 年的平均结果

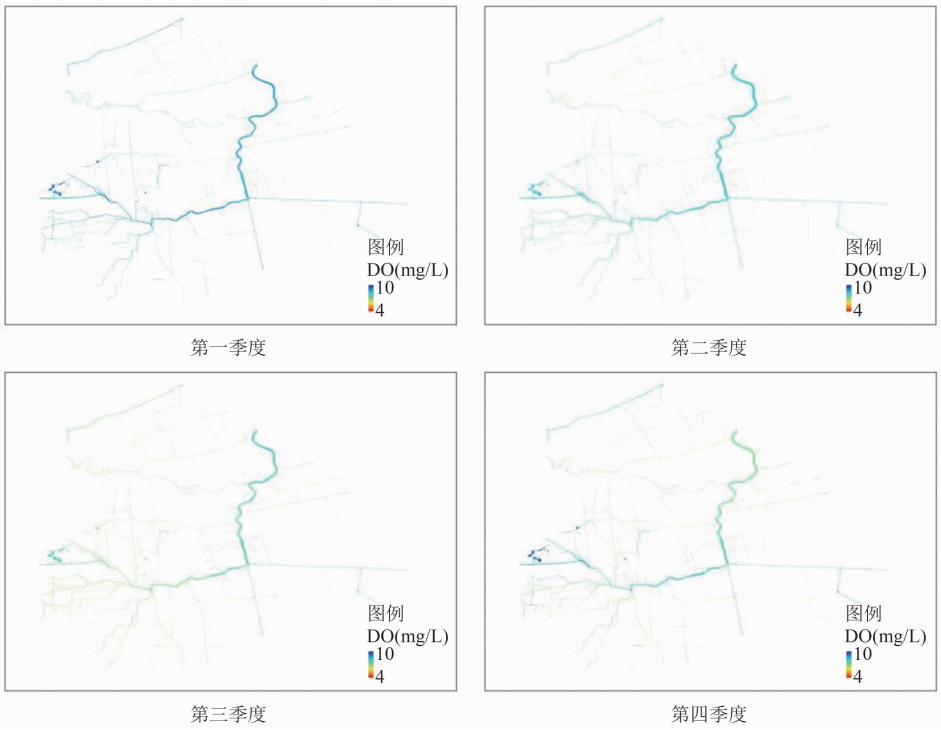


图 6 DO 的季度平均结果

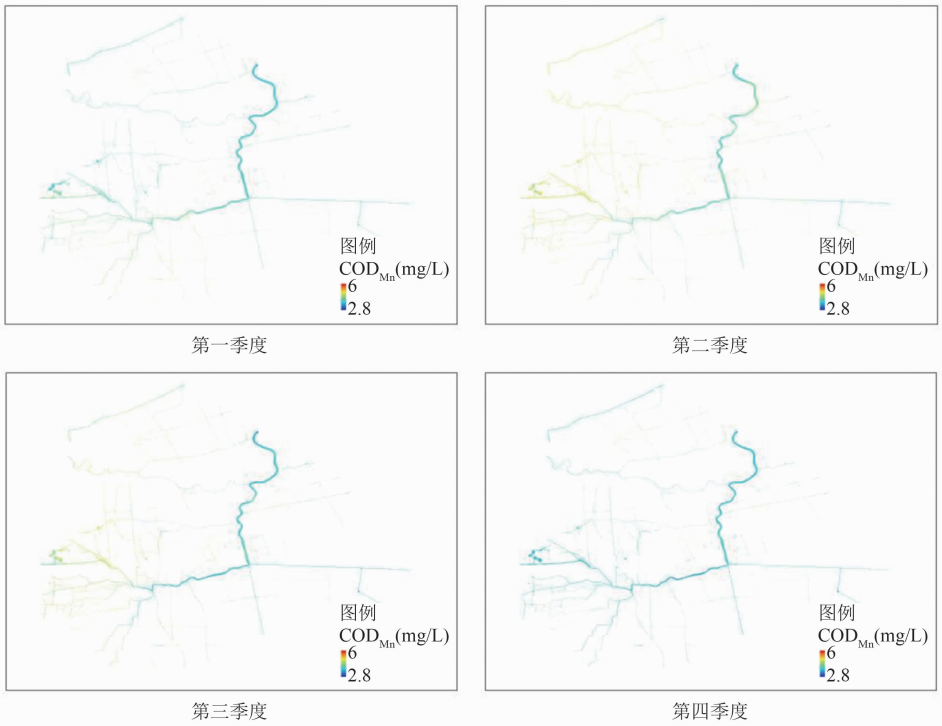
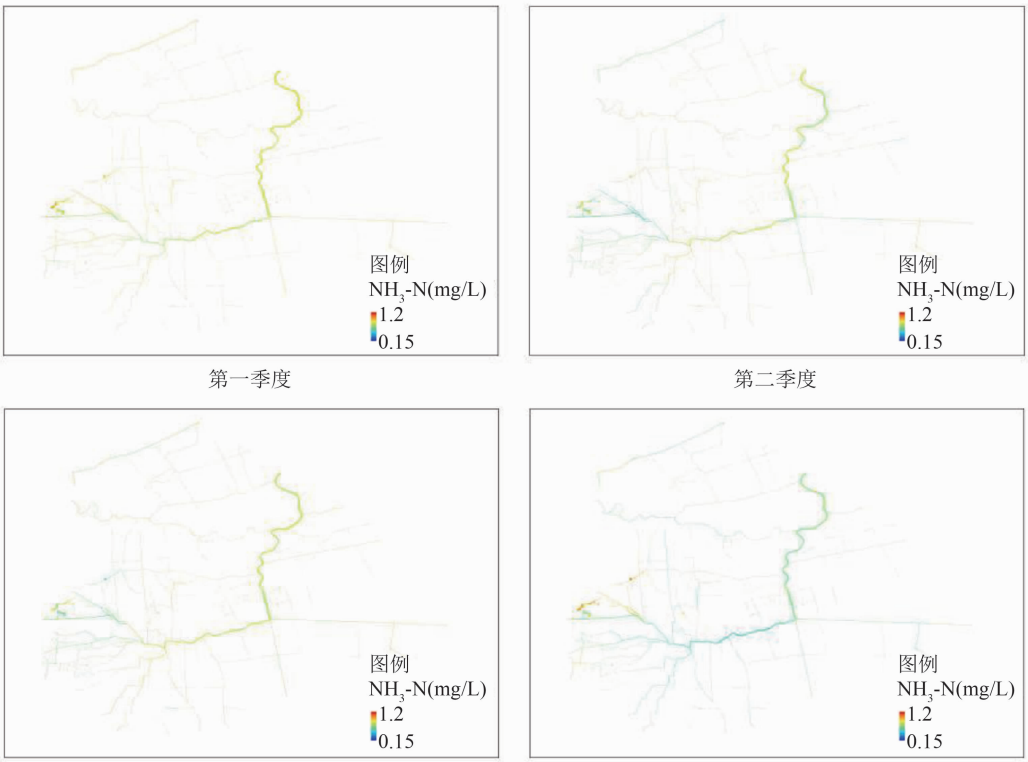


图 7 COD$_{Mn}$ 的季度平均结果

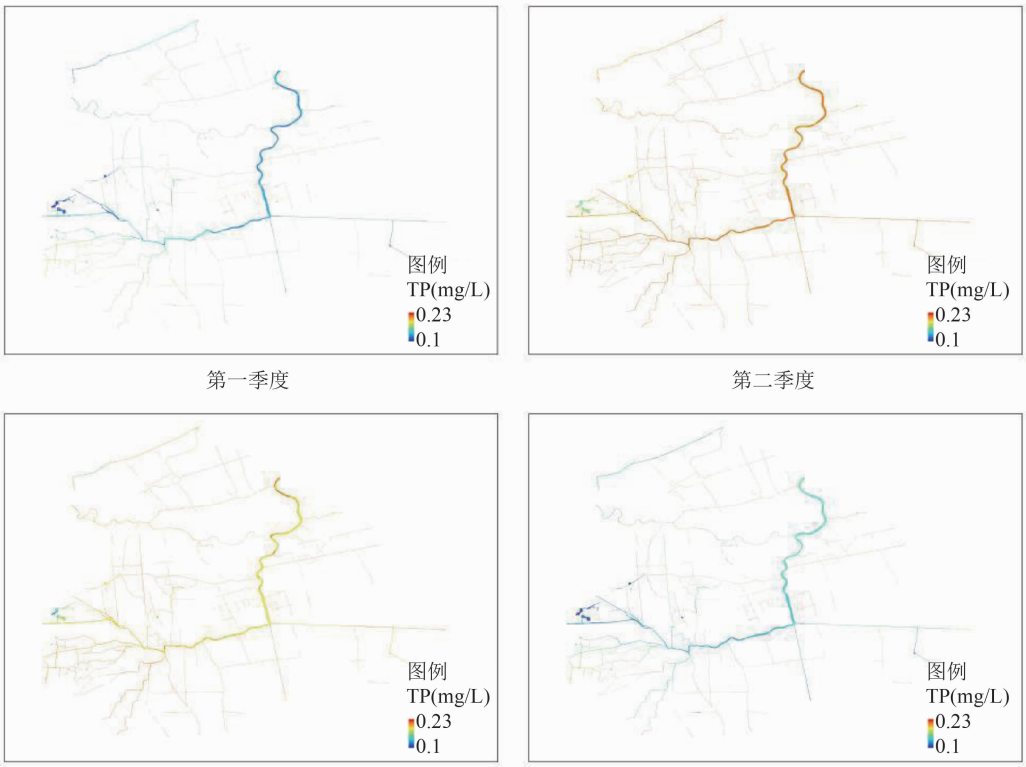
NH_3-N 浓度整体最低的是第四季度(10～12 月)，但在北横港、拦路港-泖河-斜塘、淀浦河区域浓度明显高于其他河道，整体浓度最高的是第一季度(1～3 月)，如图 7 所示。TP 浓度呈现出明显的季度差异，第二季度(4～6 月)、第三季度(7～9 月)的浓度明显高于第一季度(1～3 月)和第四季度(10～12 月)，如图 8 所示。



第一季度　　第二季度

第三季度　　第四季度

图 8 NH_3-N 的季度平均结果



第一季度　　第二季度

第三季度　　第四季度

图 9 TP 的季度平均结果

4 结束语

本文开发了一种实用有效的机器学习方法，采用 Sentinel-2 遥感图像监测上海市河水中的水质参数浓度。用 70%的实测水质数据和同步卫星影像训练的机器学习算法精度较高。该算法能够反演浑浊河流中的 DO、COD_{Mn}、NH_3-N 和 TP。采用 XGBoost 算法反演了 2019～2021 年上海市 103 条河流中水质参数的时空分布。研究结果表明：采用 Sentinel-2 多光谱遥感数据可以对河宽 30 m 以上水体进行水质监测；DO、COD_{Mn}和 TP 三个水质指标的反演精度优于 80%，NH_3-N 的反演精度优于 70%；4 种水质参数所属水体类别均优于 V 类水；各水质参数的季节变化较大，第一、第四季度水质普遍优于第二、第三季度水质；不同水质参数在不同年份的空间分布变化无明显规律。

机器学习算法的模型精度对数据量要求较高，在后续的研究中将积累更多的测量数据对模型进行优化，从而提高反演精度。为了实现对更多尺度较小的河流的水质监测，后续需要研究高空间分辨率和高光谱/多光谱数据的融合技术，采用多源遥感数据提高河湖水质监测的频率和精度。

参考文献

[1] Miao S, Liu C, Qian B, et al. Remote sensing-based water quality assessment for urban rivers: a study in Linyi development area [J]. *Environmental Science and Pollution Research*, 2020, **27**: 34586–34595.

[2] 张克，张凯，牛鹏涛，等. 遥感水质监测技术研究进展 [J]. **现代矿业**，2018，**34**(11)：171–174.

[3] 张兵，李俊生，申茜，等. 长时序大范围内陆水体光学遥感研究进展 [J]. **遥感学报**，2021，**25**(1)：37–52.

[4] Dey J, Vijay R. A critical and intensive review on assessment of water quality parameters through geospatial techniques [J]. *Environmental Science and Pollution Research*, 2021, **28**: 41612–41626.

[5] 王桥. 中国环境遥感监测技术进展及若干前沿问题 [J]. **遥感学报**，2021，**25**(1)：25–36.

[6] 王歆晖，田华，季铁梅，等. 哨兵 2 卫星综合水质指标的河流水质遥感监测方法 [J]. **上海航天**，2020，**37**(5)：92–97.

[7] Niroumand-Jadidi M, Bovolo F, Bruzzone L, et al. Inter-comparison of methods for chlorophyll-a retrieval: Sentinel-2 time-series analysis in Italian lakes [J]. *Remote Sensing*, 2021, **13**(12): 2381.

[8] Cao Z, Ma R, Duan H, et al. A machine learning approach to estimate chlorophyll-a from Landsat-8 measurements in inland lakes [J]. *Remote Sensing of Environment*, 2020, **248**: 111974–111988.

[9] Seegers B N, Stumpf R P, Schaeffer B A, et al. Performance metrics for the assessment of satellite data products: an ocean color case study [J]. *Optics Express*, 2018, **26**(6): 7404–7422.