基于 CNN+SCatBoost 模型的中国 2010—2020 年 1km×1km 分辨率 XCO_2 年度数据重建

刘宇莘

重庆交通大学 重庆 南岸 400074

【摘　要】：长时间序列、高分辨率的二氧化碳柱平均干空气摩尔分数（XCO_2 ）数据集是评估碳排放与实施低碳循环的重要基础。针对主流卫星 XCO_2 观测数据空间分辨率较粗的问题，本文提出融合卷积神经网络（CNN）与空间 CatBoost（SCatBoost）的混合模型，以全球陆地 1° 分辨率 GLM-XCO_2 数据集为基础，重建 2010—2020 年中国 1km×1km 分辨率 XCO_2 年度数据集。交叉验证结果显示，CNN+SCatBoost 模型 R^2 达 0.93，RMSE 为 0.17ppm，性能显著优于 SCatBoost（R^2=0.91、RMSE=0.20ppm）与 CatBoost（R^2=0.77、RMSE=0.32ppm）。2010—2020 年中国 XCO_2 年均浓度从 388.74ppm 增至 412.96ppm，空间上东部沿海高于西部内陆，高浓度区域呈由东向西渐进式扩张。本研究生成的高分辨率数据集为碳减排政策制定提供了科学支撑。

【关键词】：XCO_2 ；CNN+SCatBoost；高分辨率；时空特征；中国

DOI:10.12417/3041-0630.26.06.071

1 引言

CO_2 浓度升高引发的全球气候变暖对生态系统和人类社会发展构成重大挑战。中国作为全球最大发展中国家，正全力推进“碳达峰、碳中和”目标，精准掌握 XCO_2 时空变化特征，是开展碳排放评估、制定减排策略的关键前提。当前 CO_2 监测主要依赖地面观测与卫星遥感。地面观测精度高但站点稀疏；卫星遥感（GOSAT、OCO-2 等）覆盖广，但 XCO_2 数据空间分辨率多为 1° -2° ，难以满足中小尺度研究需求。现有高分辨率重建方法中，CatBoost 模型处理分类特征能力强，但忽略空间信息；SCatBoost 融入空间距离，改善了空间异质性处理，但缺乏深层空间特征自动学习能力。为此，本文提出 CNN+SCatBoost 混合模型，通过 CNN 提取深层空间特征，由 SCatBoost 建模非线性关系并处理空间异质性，实现优势互补，重建中国 1km×1km 分辨率 XCO_2 数据集。

2 材料与方法

2.1 数据来源

基础 XCO_2 数据：采用全球陆地 1° 分辨率 GLM-XCO_2 数据集（https://dataverse.harvard.edu/），基于 GOSAT 和 OCO-2 卫星观测数据通过地质统计学方法生成，时间跨度 2010—2020 年。

影响因素变量：选取人类活动（夜间灯光 NTL、人口密度 POP）、植被条件（总初级生产力 GPP、归一化植被指数 NDVI）、气象条件（气温 TMP、降水量 PRE）3 大类 6 类变量。所有数据重采样至 1km×1km，经归一化处理消除量纲影响，随后通过空间统计方法聚合至 1° 分辨率，与基础 XCO_2 数据匹配，形成 2010—2020 年共 10604 个训练样本。

2.1.1 数据处理

所有数据均采用双线性插值法重采样至 1km×1km 分辨率，经归一化处理消除量纲影响，归一化公式如下：

$$x^* = \frac{x - x_{min}}{x_{max} - x_{min}}$$

其中，x_{max}和x_{min}分别为样本数据的最大值和最小值。随后通过空间统计方法将影响因素变量聚合至 1° 分辨率，与基础 XCO_2 数据匹配，形成 2010—2020 年共 10604 个训练样本。

2.2 模型构建

本研究构建 CatBoost、SCatBoost、CNN+SCatBoost 三种模型进行对比，均以 XCO_2 浓度为输出变量，输入变量根据模型特征调整。

（1）CatBoost 模型

CatBoost：以 NTL、POP、GPP、NDVI、TMP、PRE 为输入，XCO_2 为输出，构建非线性预测模型。

（2）SCatBoost 模型

在 CatBoost 基础上融入空间信息，采用 Haversine 方法计算样本间的球面距离，将空间关系量化后作为额外输入特征，以处理 XCO_2 的空间异质性。Haversine 方法计算公式如下：

$$\text{haversin}(\theta) = \sin^2\left(\frac{\theta}{2}\right) = \frac{[1-\cos(\theta)]}{2} \quad (2)$$

$$h = \text{haversin}(u_i - u_j) + \cos(u_1)\cos(u_2)\text{haversin}(v_i - v_j) \quad (3)$$

$$\text{Distance} = 2 \times r \times \text{asin}(\sqrt{h}) \quad (4)$$

其中，u 和 v 分别表示纬度和经度，(u_i, v_i)和(u_j, v_j)为两个

样本的地理位置，r 为地球半径（约 6371 km）。模型输入为 NTL、POP、GPP、NDVI、TMP、PRE 及空间距离，输出为 XCO_2 浓度。

（3）CNN+SCatBoost 模型

CNN+SCatBoost：模型分为 CNN 特征提取和 SCatBoost 预测两个阶段。①CNN 特征提取：采用 3 层一维卷积网络（Conv1D），每层 32 个卷积核（核大小=3），通过最大池化层降维，最后经全连接层输出 64 维深层特征向量；②SCatBoost 预测：将 CNN 提取的深层特征与空间距离信息结合作为输入，建模与 XCO_2 的非线性关系。

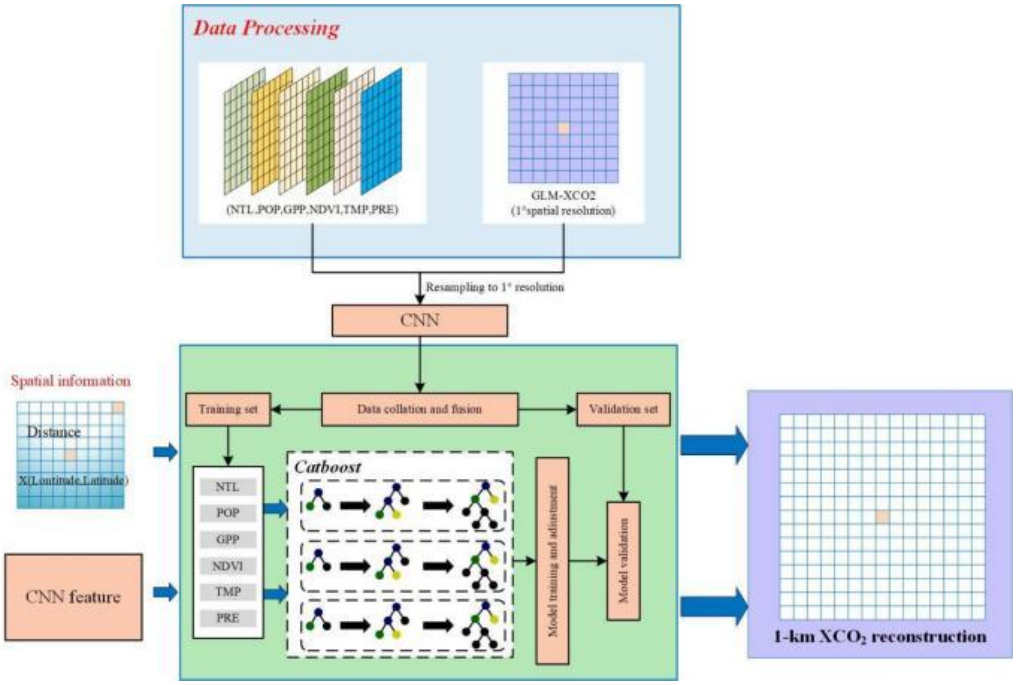

图 1 CNN-SCatBoost 模型图

Fig.1 CNN-SCatBoost model diagram

2.3 模型训练与评估

（1）训练策略：三种模型均采用 10 折交叉验证（90%训练集、10%验证集）优化超参数：CatBoost 迭代次数=2000、学习率=0.01；SCatBoost 迭代次数=2000、学习率=0.01；CNN 学习率=0.001、迭代次数=100，SCatBoost 部分参数与单独训练时一致。

（2）评估指标：采用决定系数（R^2）、均方根误差（RMSE）和平均绝对误差（MAE）衡量模型性能，

此外，采用瓦里关地面监测站（WLG）2010—2020 年的实测 XCO_2 数据进行精度验证，进一步验证重建数据集的可靠性。

3 结果与讨论

3.1 模型性能对比（以 2018 年为例）

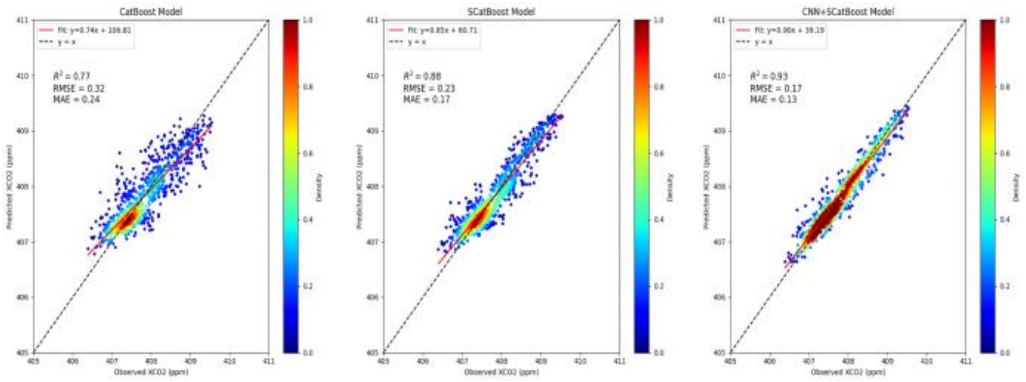

图 2 2018 年三种模型十折交叉验证结果

交叉验证中，CNN+SCatBoost 模型表现最优，R^2=0.93、RMSE=0.17ppm、MAE=0.13ppm；SCatBoost 次之（R^2=0.88、RMSE=0.23ppm）；CatBoost 最差（R^2=0.77、RMSE=0.32ppm）。瓦里关地面站独立验证结果一致，CNN+SCatBoost 模型重建数据与实测数据一致性最高（MAPE=0.37%），进一步验证了其高精度优势。

表 3 2018 年基于瓦里关地面站数据的验证结果

模型	RMSE(ppm)	MAE(ppm)	MAPE(%)
CatBoost	2.05	2.01	0.50
SCatBoost	1.81	1.78	0.44
CNN+SCatBoost	1.52	1.50	0.37

3.2 2018 年 XCO_2 空间分布对比

原始 1° 分辨率数据难以体现区域内部差异；CatBoost 预测结果空间平滑度高，忽略小尺度异质性；SCatBoost 能体现部分空间细节，但复杂地形区域存在局部偏差；CNN+SCatBoost 对细节刻画最精准，可清晰地呈现东部城市群核心城区与郊区的浓度梯度，准确捕捉西部低浓度中心的空间范围。

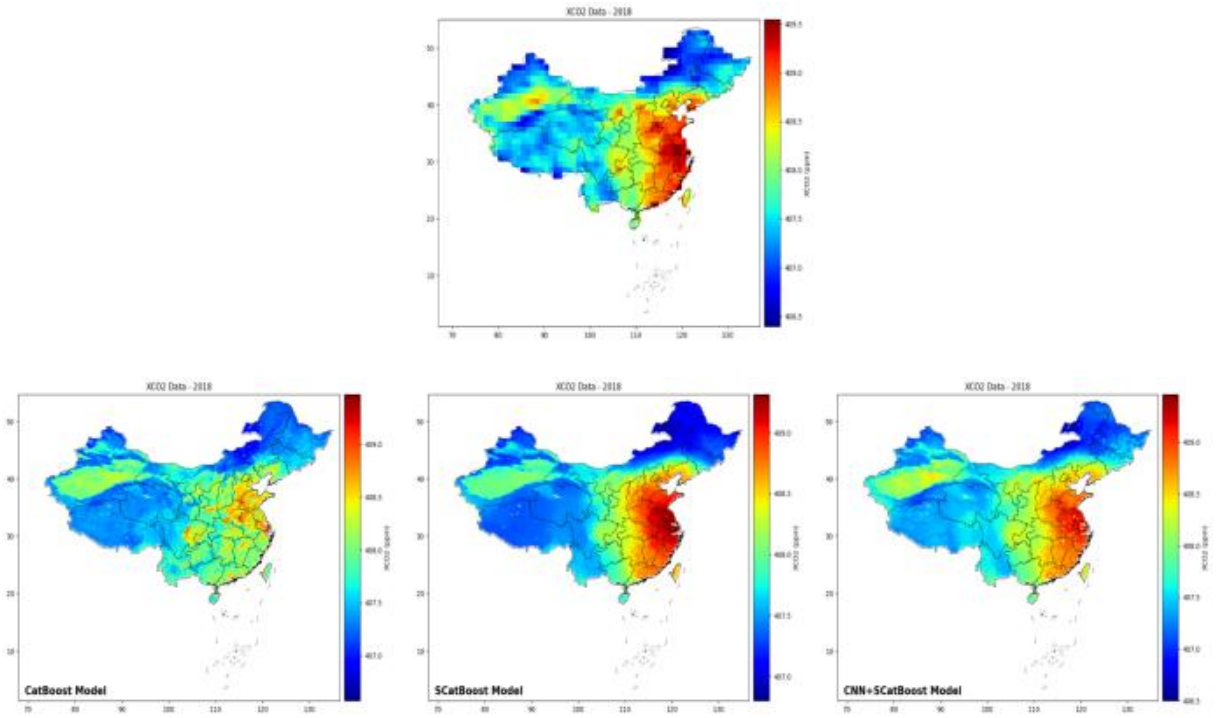

图 2 原始 1° 分辨率 XCO2 空间分布和三种模型得到的 1km 分辨率 XCO2 空间分布

3.3 2010-2020 年 CNN+SCatBoost 模型预测效果（密度散点图）

2010—2020 年 CNN+SCatBoost 模型预测值与观测值的密度散点图显示，各年份数据点均紧密聚集在 1:1 线附近，所有年份 R^2 均高于 0.90，RMSE 介于 0.17-0.20ppm 之间。表明该模型在长时间序列上具有稳定、优异的预测性能。

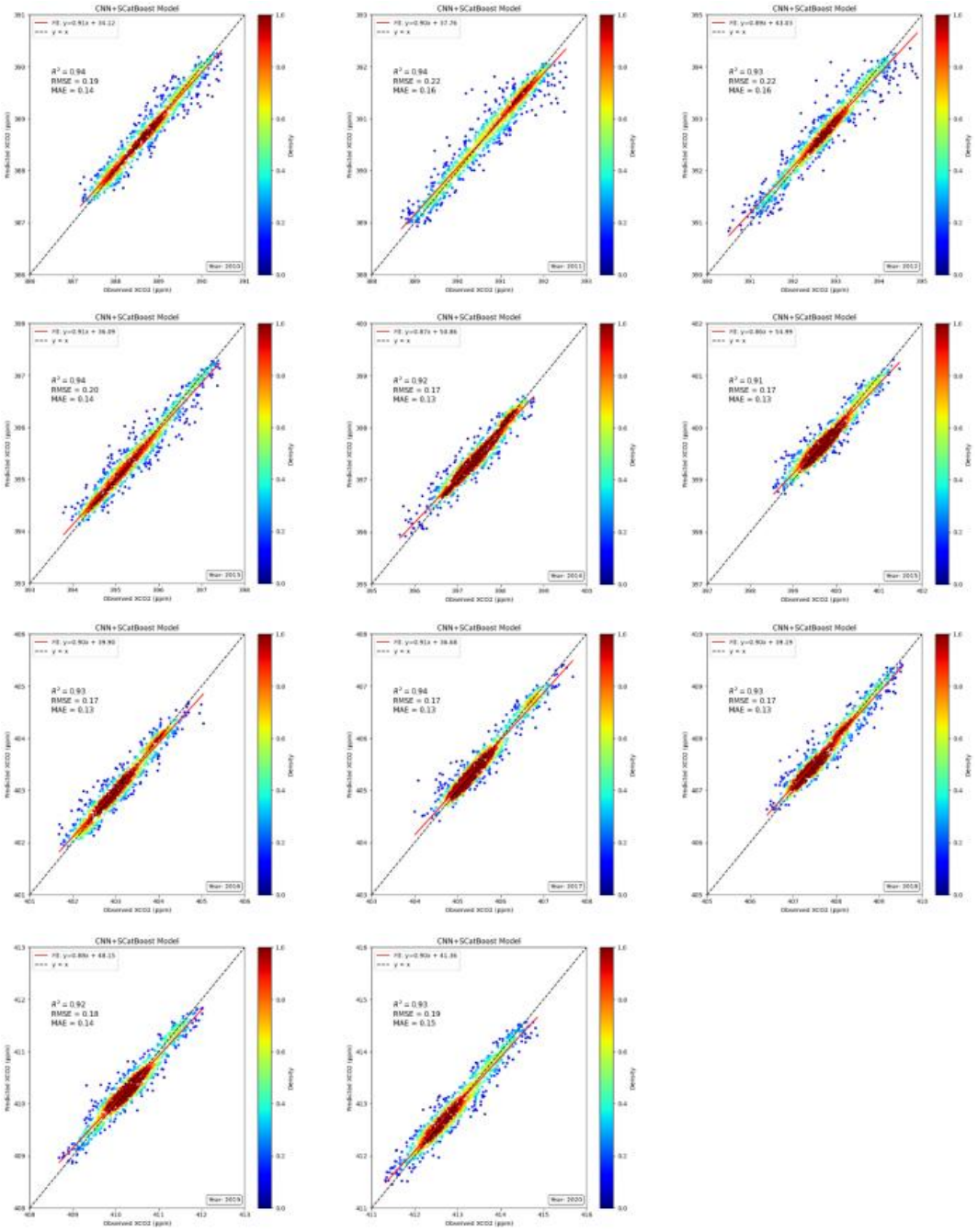

图 2 2010—2020 年使用 CNN+SCatBoost 模型十折交叉验证结果

注：黑色虚线为 1:1 线，红色实线为线性回归拟合线

3.4 2010-2020 年 XCO_2 时空演变特征

时间趋势：2010—2020 年中国 XCO_2 年均浓度从 389.72ppm 增至 412.58ppm，累计增长 22.86ppm，年均增长率 2.29ppm/年（$p<0.001$）。其中 2010—2015 年为平缓增长阶段（年均增长 2.15ppm），2016—2020 年为加速增长阶段（年均增长 2.43ppm），城市化与工业化进程加快是主因。

空间分布：始终保持“东高西低”格局，呈现东部高值核心区—中部过渡区—西部低值区三级特征。高浓度区域集中在东部沿海及华北平原人类活动密集区，低浓度区域分布在西部青藏高原、西北荒漠及西南高植被覆盖区。2010—2020 年，高浓度区域呈由东向西渐进式扩张特征，东部高浓度区向中部延伸，中部整体浓度提升，东西部浓度梯度局部差异小幅缩小。人类活动强度的空间分异与区域发展进程是核心驱动因素，自然条件为辅助约束。

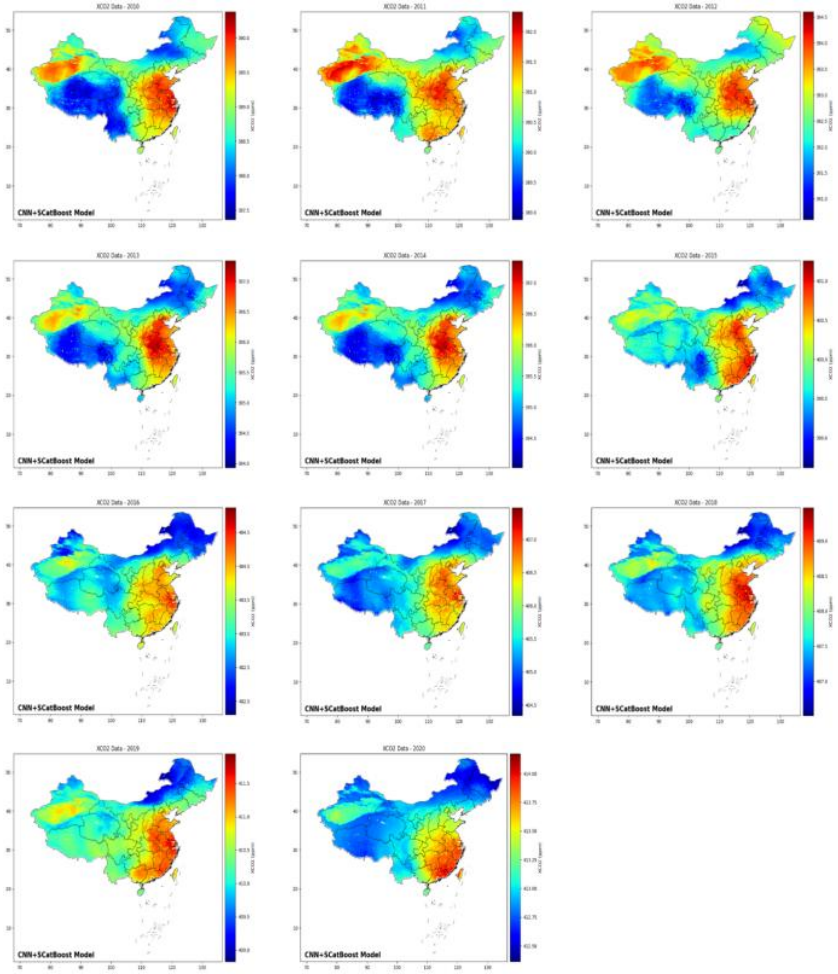

图 4 2010—2020 年 1km 分辨率 XCO2 空间分布

3.5 模型优势分析

CNN+SCatBoost 模型实现深度学习与空间机器学习的优势互补：一是通过卷积操作捕捉影响因素的局部与全局空间关联，提供深层特征支撑；二是结合 SCatBoost 的空间异质性处理能力，精准捕捉中小尺度空间差异；三是时序预测稳定，有效降低过拟合风险，适用于长时间序列高分辨率数据集重建。

4 结论

本文构建CNN+SCatBoost混合模型，成功重建2010—2020年中国 1km×1km 分辨率 XCO_2 年度数据集。主要结论：①模型性能优异，2018 年交叉验证 R^2=0.93、RMSE=0.17ppm，瓦里关站验证 MAPE=0.37%，2010-2020 年各年份 R^2 均超 0.90，空间细节刻画精准、时序预测稳定；②时间上，全国 XCO_2 年均浓度从 389.72ppm 增至 412.58ppm，年均增长 2.29ppm，2016 年后加速增长；③空间上，始终呈东高西低格局，高浓度区集中于东部城市群，2010-2020 年由东向西渐进式扩张，人类活动强度与区域发展进程是核心驱动因素。

本研究生成的高分辨率 XCO_2 数据集为中小尺度碳排放监测、碳源碳汇评估提供了数据支撑，CNN+SCatBoost 的建模思路也为其他地理环境变量的高分辨率重建提供了参考。

参考文献：

[1] Wu,C.Ju,Y.Yang,S.Zhang,Z.Chen,Y.Reconstructing annual XCO(2)at a 1 kmx1 km spatial resolution across China from 2012 to 2019 based on a spatial CatBoost method.Environ Res 2023,236,116866.[CrossRef].

[2] 48.He,S.Yuan,Y.Wang,Z.Luo,L.Zhang,Z.Dong,H.Zhang,C.Machine Learning Model-Based Estimation of XCO2 with High Spatiotemporal Resolution in China.Atmosphere 2023,14,436.[CrossRef].

[3] 54.Li,T.Wu,J.Wang,T.Generating daily high-resolution and full-coverage XCO(2)across China from 2015 to 2020 based on OCO-2 and CAMS data.Sci.Total Environ.2023,893,164921.[CrossRef].