DOI: 10.3969/j.issn.1673-6141.2023.03.007

利用近邻因子提高二氧化氮遥感反演浓度的精度—基于随机森林算法

符 淼

(广东外语外贸大学经济贸易学院, 广东 广州 510006)

摘 要: NO_2是损害健康和破坏生态的主要大气污染物。本文基于NASA提供的Aura OMI遥感反演NO_2浓度，利用采样点8 km内的经济、人口、路网和坡度数据，以及气象、植被和高程的点值数据，采用随机森林算法、地理加权回归(GWR) 和多尺度GWR方法提高NO_2浓度的预测精度。NASA原浓度R^2为0.48，以上三种模型把交叉验证R^2分别提高到0.74、0.71和0.70，其中随机森林算法的精度最高，该算法的均方根误差 (RMSE) 和平均绝对误差 (MAE) 分别只有6.4 μg/m^{3}和4.98 μg/m^3，且其速度远快于多尺度GWR，预测精度也高于大部分现有的同等范围研究。在浓度修正方面，局部化经济人口路网因子对预测精度提高的贡献至少为11.24%。此外，基于随机森林算法还给出全国县级城市NO_2浓度估计值的分布图。

关 键 词: 二氧化氮浓度; 近邻因子; 随机森林算法; 地理加权回归; 多尺度地理加权回归

中 图 分 类 号 : X51 **文 献 标 识 码 :** A **文章编号:** 1673-6141(2023)03-258-011

Improving the accuracy of NO_2 concentrations derived from remote sensing using localized factors based on random forest algorithm

FU Miao

(School of Economics and Trade, Guangdong University of Foreign Studies, Guangzhou 510006, China)

Abstract: NO_2 is a main air pollutant that damages human health and ecological environment. Based on NASA's NO_2 concentrations retrieved from Aura OMI, the prediction accuracy of NO_2 concentration is improved in this work using the random forest algorithm, the Geographic Weighted Regression (GWR) and the Multi-scale GWR model respectively. Localized data of economy, population, road network and slope within 8 km of the sampling point, as well as the point values of meteorology, vegetation and elevation are used as correction variables in the models. It is found that the three models increase the cross validation R^2 of NASA's concentrations, from original 0.48 to 0.74, 0.71 and 0.70, respectively. Among the three models, the random forest algorithm is the most accurate one, with a low root mean square error (RMSE) of 6.4 μg/m^3 and a low mean absolute error (MAE) of 4.98 μg/m^3, and its speed is much faster

基金项目: 教育部人文社会科学研究规划基金项目 (17YJA790021), 广东省自然科学基金自由申请项目 (2017A030313439), 广州国际商贸中心研究基地专项资助 (JDZB202108)

作者简介: 符 淼 (1973-), 广东雷州人, 博士, 教授, 硕士生导师, 主要从事环境经济学研究。E-mail: cnfm@163.com.

收稿日期: 2022-01-11; **修改日期:** 2022-02-28

than multi-scale GWR. In addition, the accuracy of random forest algorithm is also higher than that of most existing studies of similar extents. In terms of the concentration correction of NO_2, it is found that the contribution of localized factors of economy, population and road network is at least 11.24%. In addition, based on the random forest algorithm, the distribution map of NO_2 estimated concentration for county-level cities in China is also presented.

Key words: NO_2 concentrations; localized factors; random forest algorithm; geographic weighted regression; multi-scale geographic weighted regression

0　引　言

氮氧化物是主要的大气污染物, 能引起支气管炎、肺气肿、哮喘和呼吸障碍等呼吸道疾病。氮氧化物还是造成酸雨以及土壤酸化的主要原因之一。人类排放的主要氮氧化物为一氧化氮 (NO), NO 在大气中被氧化为能稳定存在的二氧化氮 (NO_2)。我国 NO_2 污染比较严重, Li 等[1]的研究表明, 2010 年亚洲的氮氧化物总排放量约为 52.1 Tg, 其中中国和印度的排放量占亚洲排放量的主体。美国国家航空航天局 (NASA) 提供的 NO_2 浓度图显示, 在人口密集的京津冀、长三角和珠三角地区, NO_2 污染比较严重。NO_2 的主要来源是汽车排放、火力发电、冶炼、其他使用工业窑炉的重工业以及硝酸生产和硝化过程等。NO_2 浓度的估算是评估 NO_2 污染严重程度及其危害的主要依据。

当前, 大范围 NO_2 浓度估算方法多基于卫星遥感数据。针对美国的大城市, Bechle 等[2]从 Aura 卫星的臭氧监测仪 (OMI) 数据反演得到 NO_2 浓度, 发现源自 OMI 的估计与现场测量值之间的空间相关性很强。Chan 等[3]使用神经网络模型, 基于对流层 NO_2 垂直柱密度 (VCD) 和其他气象卫星数据, 发现 NO_2 估算结果与现场监测浓度数据的相关系数为 0.80。 Cooper 等[4]发现基于哨兵 5P 卫星对流层观测仪 (TROPOMI) 或 OMI VCD 方法容易低估 NO_2 的浓度, 主要原因在于卫星产品的分辨率较低, 以及从 VCD 到地表浓度的算法精确度有待提高。为提高估算精度, 有的研究加入了土地利用状况, Knibbs 等[5]基于土地利用回归 (LUR) 模型, 利用对流层 NO_2 VCD 和土地利用状况解释了 NO_2 81% 的空间变化, 均方根误差 (RMSE) 为 2.63 μg/m^3。Novotny 等[6]采用 OMI 的 NO_2 垂直柱密度数据, 辅以人口密度、土地利用、气象以及到主次要道路的距离, 发现模型拟合 R^2 等于 0.78, 能够较好地捕获城市和近路地区 NO_2 浓度。拟合 R^2 又称拟合优度, 为样本回归模型回归平方和 (ESS) 和总平方和 (TSS) 之比, 代表模型已解释的部分, 取值范围在 0 和 1 之间, 拟合 R^2 不含样本外的检验, 区别于后面提到的交叉验证 R^2, 拟合 R^2 一般简称 R^2。Larkin 等[7]在全球范围将遥感数据与土地利用、人口和道路信息相结合, 其模型解释了全球 NO_2 54% 的变化。Harper 等[8]利用 LUR 模型预测了重庆市 NO_2 浓度的变化。施媛媛等[9]利用 LUR 分析了武汉市 NO_2 浓度的空间分布。Anand 和 Monks[10]利用 LUR 模型模拟香港特别行政区 NO_2 日浓度从 2005 年至 2015 年的变化, 交叉验证结果表明基于 OMI 数据的预测能力较高。He 等[11]利用机器学习和自适应加权方法拟合 GOME-2B 的 NO_2 测量值和地表观测值, 发现两者相关系数较高。Chen 等[12]比较了 16 种 NO_2 浓度的预测算法, 包括线性逐步回归、正则化技术和机器学习方法等, 发现以上方法的预测能力相差不大, 交叉验证 R^2 在 0.57 到 0.62 之间。交叉验证 R^2 采用模型预测样本外检验点的因变量值, 而后把预测值与观察到的相应因变量值比较, 计算 R^2, 交叉验证 R^2 体现基于样本得到的模型能否适用于样本外的数据, 故难度更大, 它的值一般低于同模型的拟合 R^2。Wang 等[13]利用 OMI 和 TROCOMI 对流层

NO_2 VCD 和 GEOS-Chem 模型, 发现 COVID-19 疫情封控措施导致中国 2020 年 2 月和 3 月地表 NO_2 浓度分别降低了 42% ± 8% 和 26% ± 9%。

与已有研究相比, 本研究采用随机森林算法, 同时与常用的地理加权回归 (GWR) 方法比较, 以及与多尺度地理加权回归 (MGWR) 比较, 以找出精度较高的模型。同时, 在县一级统计数据和地理信息数据基础上, 采用所研究点附近 8 km 范围内的经济、人口、交通、坡度和土地利用等统计数据作为预测因子, 并结合气象、高程和植被等因子的点值, 提高预测的准确性。

1 数 据

NO_2 的卫星反演浓度来源于 NASA Aura 卫星验证数据中心 (AVDC) 的 OMSNO2 (OMI surface NO_2) 产品。这一反演浓度的数据基础是 Aura 卫星上的 OMI 所获取的 NO_2 VCD。OMI 主要检测大气分子对反向散射的可见光和紫外光的吸收程度, 在天底点其视野 (FoVs) 的尺寸为 13 km × 24 km, 刈幅宽度为 2600 km, 每两秒完成一次测量。反演算法采用 Lamsal 等[14]提出的方法, 该方法先将 OMI 的 NO_2 柱观测值网格化, 并利用 GMI-Replay 化学传输模型得到的垂直结构来估算地表 NO_2 浓度。Auro 卫星的当地赤道穿越时间 (LECT) 为 13:45 ± 00:15, 故得到的浓度为当地时间 13:45 左右的浓度, 分辨率为 0.1° × 0.1°。2015 年该浓度的年平均值和地面监测站观测到的 NO_2 浓度年平均值的直方图见图 1。两者的相关系数为 0.6937, R^2 为 0.4813, 总的来说精度不高, 且从分布上看以上算法明显低估了 NO_2 浓度值。遥感反演浓度偏差的主要原因之一是卫星遥感的时空尺度与地面站点的观测尺度有所不同。因此本研究将通过考虑局部化的经济、人口、地理、气象、土地利用和道路网络数据来提高浓度估计的精度。在模型中各采样点的卫星反演 NO_2 浓度采用 2015 年均值, 以与县一级经济和人口统计数据保持一致。

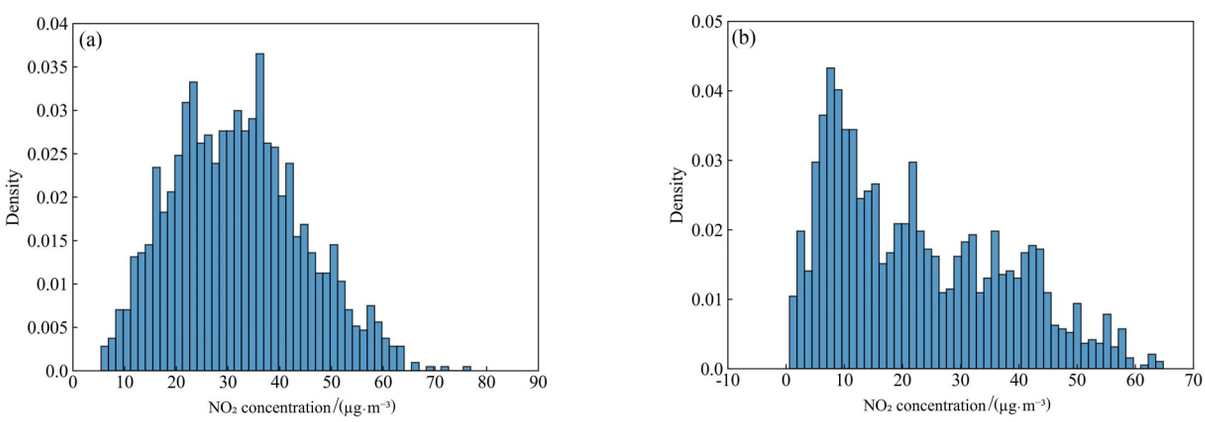


图 1 NO_2 地面观测浓度和卫星遥感反演浓度的直方图比较。(a) 地面观测浓度; (b) NASA 卫星反演浓度

Fig. 1 The comparison of the distributions of NO_2 ground-level observed and satellite-derived concentrations. (a) Ground-level observed NO_2 concentrations; (b) NASA satellite-derived NO_2 concentrations

环境监测站 NO_2 地面观测浓度来自中国环境监测中心 (CEMC)。所用的浓度值为各个监测站 2015 年 NO_2 浓度的年平均值。在删除缺失值后, 剩余 1494 个站点的数据。NO_2 地面监测浓度为模型的因变量。模型的核心自变量为卫星反演 NO_2 浓度, 为进一步提高模型的预测精度, 在模型的自变量一侧引进了降水量和风力等气象变量以及经济、人口、道路和地形等近邻因子。郑凯莉等[15]发现气象因子和 NO_2 污染紧密相关。

本研究的气象数据来自ERA5, ERA5原始数据为月度数据, 其分辨率为0.25°。采集的气象数据包括10 m高度风速 (10SI)、地面2 m高度温度 (2T)、边界层高度 (BLH)、相对湿度 (RH)、地表气压 (SP)、地表太阳辐射 (SSR)和总降水量 (TP) 等。此外, 还采用了来自MODIS的增强植被指数 (EVI), EVI分辨率为1 km, 数据更新间隔为16天。以上气象和植被变量均采用2015年各采样点的年平均值。

县级数据如第一产业、第二产业、第三产业产值和劳动力, 以及各类污染物的工业排放量, 均来自相应地级市统计年鉴的县级数据。人口密度分布来源于WorldPop, 并根据县级人口普查和统计数据进行调整, 以保证在县级总量上与统计数据一致。以上经济和人口数据均为2015年数据。道路网络来自OpenStreetMap, 由于OpenStreetMap的数据更新有一定的滞后性, 尤其是用户不活跃的地区更新比较慢, 故采用2018年初下载的数据, 以保证其与污染监测数据在时间上更为接近。地形数据方面, 数字高程模型 (DEM) 来自NASA航天飞机雷达地形测绘任务 (SRTM), 坡度基于DEM计算。土地利用数据来自全球30 m分辨率的地表覆盖数据 (GlobeLand30), 该数据由国家基础地理信息中心于2014年发布, 与2015年比较接近。

以上数据的采样点集之一是全国的环境监测站点, 共1494个点, 其数据包含NO_2的地面观测浓度, 作为估计模型的基础数据。采样点集之二是全国各县级辖区的城区中心点, 去掉一些缺少数据的县, 共2858个点, 这些点对应的地理坐标上一般没有监测站, 只能采集到自变量的值, 无NO_2浓度观测值, NO_2浓度需用模型预测得到。

2　方　法

随机森林算法由Breiman[16]提出。随机森林算法首先用自举法 (Bootstrap) 得到数据的K个自举样本, 每个自举样本将构建一个独立的决策树 (也称为Estimator), 故得到K个决策树。在建立决策树时, 在树的每个分裂节点, 随机选择M个候选特征, 而后从中选出最佳的分裂特征进行树的分裂, 每次分裂得到两个分支, 且每棵树均在没有修剪的情况下尽可能地分裂生长。最后的预测值是每个决策树输出值的平均值, 从而避免使用单个决策树导致预测偏误。本模型中, 用来预测的特征就是前文提到的经济、人口、地理、气象、交通和土地利用等变量。M和K可通过袋外 (Out of bag) 检验误差确定, 也可以通过交叉验证确定。通过交叉验证, 确定适用本模型的最优K和M分别是250和8。预测质量的判断公式为

$$g(\boldsymbol{X},\boldsymbol{Y})=a_k I\left[h_k(\boldsymbol{X})=\boldsymbol{Y}\right]-\max_{Y'\neq Y}\left\{a_k I\left[h_k(\boldsymbol{X})=\boldsymbol{Y}'\right]\right\}, \tag{1}$$

$$Q^*=P_{X,Y}\left[g(\boldsymbol{X},\boldsymbol{Y})<0\right], \tag{2}$$

式中$\boldsymbol{X}$为预测因子, $\boldsymbol{Y}$为浓度观察值, a为预测值的平均投票数, k取值从1到K, $I(\bullet)$为指示函数, $h_k(\bullet)$即为第k个树预测器, P表示概率, 加下标 ($\boldsymbol{X}$, $\boldsymbol{Y}$) 表示在 ($\boldsymbol{X}$, $\boldsymbol{Y}$) 空间上的概率。

式 (1) 为边际函数。式中前后两项差额衡量的是在 ($\boldsymbol{X}$, $\boldsymbol{Y}$) 空间上, 正确的预测值平均投票数超过错误预测值平均投票数的程度。共K个决策树进行投票, 边际值越大, 预测的可信度越高。式 (2) 计算边际函数小于零的概率, 即泛化误差Q^*。关于随机森林用于回归时的泛化误差性质, Breiman[16]证明了以下结论

$$Q^*(f)=E_{X,Y}E_{\boldsymbol{\Theta}}\left[\boldsymbol{Y}-h(\boldsymbol{X},\boldsymbol{\Theta})\right]^2\leq\bar{\rho}Q^*(t), \tag{3}$$

式中f表示随机森林; $\boldsymbol{\Theta}$为决策树生长所依赖的随机向量, 它所生成的树预测器为$h(\boldsymbol{X},\boldsymbol{\Theta})$; $\bar{\rho}$为$\boldsymbol{Y}$–$h(\boldsymbol{X},\boldsymbol{\Theta})$和$\boldsymbol{Y}$–$h(\boldsymbol{X},\boldsymbol{\Theta}')$的加权相关系数; t表示决策树。

式 (3) 表示左侧随机森林泛化误差小于等于右侧决策树泛化误差均值和$\bar{\rho}$的乘积。可见决策树之间的独立性越强和决策树的预测误差越低, 随机森林回归的预测越准确。随机森林算法也存在不足, 比如在做回归预测时, 它得不到连续的预测值, 且预测结果不会超出训练数据集的范围。

为提高预测精度, 除了NO_2反演浓度、气象、高程和植被指数使用采样点的点值外, 人口、GDP、工业产值、路网长度、坡度均值、不同土地覆盖类型面积均为距采样点 8 km 范围内的局部化数据。局部化的采样、求和与求均值基于GIS编程实现。GDP、工业产值和就业人口等统计数据通常只提供到县级, 因此需要利用 8 km 范围内人口占该县总人口的比率将县级数据分配到各个采样点上。在亚洲, 使用人口密度来拆分数据更为合理, Larkin 等[7]发现人口密度仅解释全球NO_2变化的1%, 但它解释了亚洲NO_2变化的3%。

本研究使用 8 km 缓冲区, 原因是 Zhai 等[17]发现 8 km 范围内获取的土地利用数据对大气污染物有显著影响。Meng 等[18]发现监测点周围 2 km 缓冲区内主要道路长度、10 km 缓冲区内的工业源数量、5 km 缓冲区内的农业用地面积以及人口数对NO_2浓度有显著的影响。Knibbs 等[5]发现除污染物的卫星遥感数据外, 对大气污染解释能力较强的解释变量是 8 km 内的道路长度和 10 km 内的工业用地面积。Gilliland 等[19]认为, 空气污染物的城市扩散范围为 4～50 km。综合以上研究, 且经过校验, 采用 8 km 缓冲区较为合适。解释变量包含 8 km 内路网长度之和, 是因为交通在城市空气污染中起关键作用, 特别是在城市已迁出大型点污染源的情况[20]。

为比较不同模型的预测精度, 还采用GWR进行对比研究, GWR 的计量模型为

$$y_i=\beta_{0i}+\sum_{k=1}^{m}\beta_{ki}x_{ik}+\varepsilon_i, \tag{4}$$

$$\hat{\boldsymbol{\beta}}_i=\left[\boldsymbol{X}'\boldsymbol{W}(i)\boldsymbol{X}\right]^{-1}\boldsymbol{X}'\boldsymbol{W}(i)y, \tag{5}$$

$$w_j(i)=\begin{cases}\left(1-\dfrac{d_{ij}^2}{b^2}\right)^2, & d_{ij}\le b\\ 0, & d_{ij}>b\end{cases}, \tag{6}$$

式中y_i是NO_2的浓度, β_{0i}和β_{ki}是随位置 i 地理坐标变化而变化的截距和系数, x_{ik}是前面描述过的解释变量, ε_i为随机干扰项, 系数向量$\boldsymbol{\beta}_i$可通过式 (5) 估计, 其中$\boldsymbol{X}$是解释变量矩阵, $'$表示矩阵转置操作, $\boldsymbol{W}(i)$ 是权重的对角线矩阵, 对角线中的权重通过式 (6) 中所示的双平方 (Bisquare) 函数计算, d_{ij}是位置i和j之间的距离, b表示带宽。

GWR是常见的大气污染物浓度预测方法。与GWR相比, MGWR为不同的预测变量分配不同的带宽, 即各变量的权重在不同的空间尺度上变化[21]。MGWR的计量模型为

$$y_i=\beta_{b0i}+\sum_{k=1}^{m}\beta_{bki}x_{ik}+\varepsilon_i, \tag{7}$$

式中β的下标b表示用于该特定解释变量的带宽, 其他变量的含义见GWR的说明。

通过前向逐步回归、排除多重共线性和交叉验证选择最有解释力的预测因子, 最终选用的自变量包括遥感反演NO_2浓度、第二产业产值、路网长度和、人口、平均坡度、数字高程、植被指数、10SI、2T、BLH、RH、SP、SSR和TP。经济、路网、人口和坡度采用采样点8 km范围内的局部值, 反演NO_2浓度、气象、高程和植被指数使用采样点的点值。以上所有模型均采用Python编程实现。

3　结果分析

为评估上述三个模型预测结果的质量, 对GWR使用留一法 (Leave One Out) 交叉验证 (CV), 对随机森林算法和MGWR使用80-Fold交叉验证。由于GWR高度依赖于空间权重矩阵, 常用的10-Fold交叉验证不太适用, 因为它将空间权重矩阵的维数减少了十分之一, 这会显著改变结果。因此, 留一法交叉验证是最佳选择。然而, MGWR使用留一法交叉验证可能需要几个月的时间。为了使得验证在可接受的时间跨度内完成, 故使用80-Fold交叉验证, 这一选择对空间权重矩阵维度的影响较小, 且完成时间也在可接受的范围 (约三天时间完成一次验证)。随机森林算法交叉验证浓度 (RF_NO_2)、GWR交叉验证浓度 (GWR_NO_2) 和MGWR交叉验证浓度 ($MGWR\_NO_2$) 的描述性统计见表1, 交叉验证浓度指交叉检验时对样本外点的浓度预测值。作为比较, 地面观察值 (Obs_NO_2) 以及NASA预测值 ($NASA\_NO_2$) 的统计描述也列于表中, 其中25%、50%和75%为分位数。由表可见, $NASA_NO_2$明显低估浓度, GWR_NO_2和RF_NO_2与地面观察值统计特征比较接近, GWR_NO_2稍微高估最大值, RF_NO_2稍微低估最大值, 说明随机森林算法不太擅长于解释极端值。MGMR的最大值与原值最接近, 但最小值出现负值。三种算法交叉验证浓度的直方图见图2。与图1相比, 三种算法均大幅度修正了NASA浓度的左偏问题。从整体上看, 随机森林算法的分布与地面观察值的分布最为接近。

表 1　地面观察值、NASA预测值及三种算法交叉验证浓度的统计描述

Table 1　Statistical description of observed，NASA and the cross validation concentrations

Variable	Count	Mean	Std	Min	25%	50%	75%	Max
Obs_NO2	1494	31.850	12.504	5.435	22.329	31.421	40.205	76.919
NASA_NO2	1494	22.490	14.862	0.600	9.600	19.550	34.400	64.800
GWR_NO2	1494	31.087	10.988	7.860	22.891	30.003	38.151	80.722
MGWR_NO2	1494	31.853	12.022	−0.544	22.822	30.789	40.035	73.671
RF_NO2	1494	31.962	10.479	8.515	24.035	30.955	38.975	64.301

注: Std表示标准差

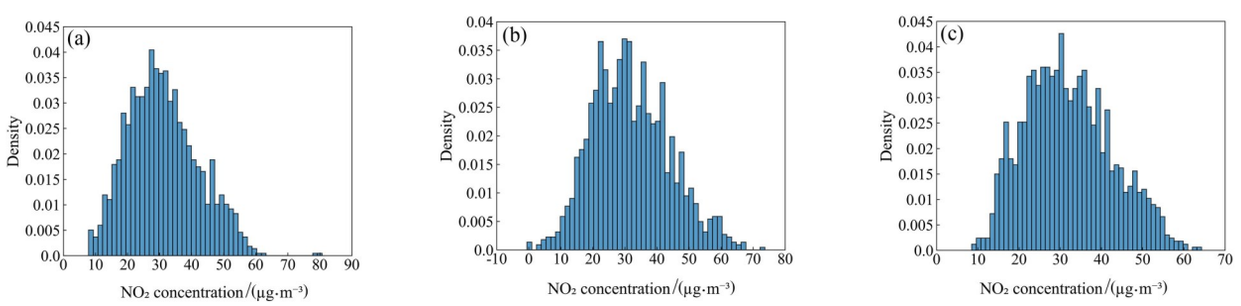


图 2　三种算法交叉验证浓度的直方图分布。(a) GWR_NO_2; (b) $MGWR\_NO_2$; (c) RF_NO_2

Fig. 2　The histograms of the cross validation concentrations from the three algorithms.

(a) GWR_NO_2; (b) $MGWR\_NO_2$; (c) RF_NO_2

GWR、MGWR、随机森林算法的交叉验证结果如图3所示, 其中随机森林算法2指不含经济人口路网因子的随机森林算法, 加入随机森林算法2是为了检验局部化经济人口路网因子的贡献。由图可以看出, 图3 (c) 中随机森林算法的交叉验证R^2最大, 为0.7365, 但斜率偏小, 为0.7192, 这意味着它的预测精度最高, 但略微低估浓度。由于使用可变带宽, MGWR在斜率上表现最好, 为0.8047, 但交叉验证R^2低于随机森林算法, 也低于固定带宽的GWR, 且MGWR的浓度预测值出现负值。MGWR的预测质量在很大程度上取决于所使用的带宽, 这一带宽虽然可以通过迭代算法得到, 但实际上很难与模型完美匹配, 有时人为指定另一带宽反而效果更好。GWR的估计精确度在两者之间。图3 (d) 的交叉验证R^2低于图3 (c), 因此经济人口路网因子提高了模型的拟合度。

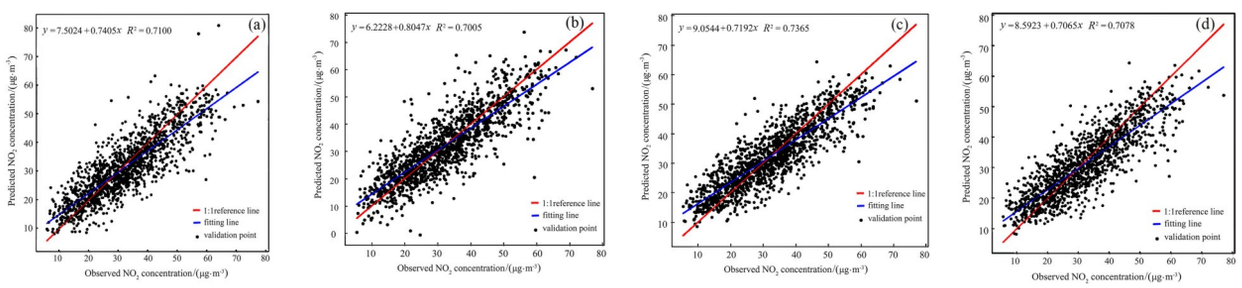


图3 GWR、MGWR和随机森林算法交叉检验结果。(a) GWR; (b) MGWR; (c) 随机森林算法; (d) 随机森林算法2

Fig. 3 Cross validation results of the GWR, MGWR and random forest. (a) GWR; (b) MGWR; (c) random forest; (d) random forest 2

表2中给出更多交叉验证结果, 包括交叉验证均方根误差 (RMSE)、交叉验证平均绝对误差 (MAE)、交叉验证平均绝对百分比误差 (MAPE) 和交叉验证相关系数 (CV r)。第一行是NASA提供的NO_2浓度原始估计值的质量评估, 可见三个模型都不同程度地提高了NASA浓度估计值的质量。在这三个模型中, 随机森林算法的相关系数和拟合优度最高, 且RMSE和MAE最低, 表明随机森林算法是最好的模型。比较表中含与不含经济人口路网因子的交叉检验结果, 可以发现, 经济人口路网因子约把交叉验证R^2了提高0.0287 (边际贡献, 未含与其他预测因子重叠的部分), 含经济人口路网因子的随机森林模型约把原NASA浓度的R^2提高了0.2552, 因此在修正NASA NO_2浓度方面, 局部化经济人口路网因子的贡献至少为11.24% (0.0287/0.2552), 加上重叠的贡献就有可能大于此比例。

表2最后一列为拟合R^2 (Reg R^2), 即回归模型的可决系数, 一般拟合R^2大于交叉验证R^2 (CV R^2)。有的研究只给出拟合R^2, 因此表中也给出拟合R^2, 以供比较。除了拟合R^2这一列, 其他列均为交叉验证结果。可以发现MGWR模型的拟合R^2最大, 高达0.94, 斜率也更接近于1, 说明采用多尺度带宽后, 大幅度提高了模型的拟合度。但是, 它的交叉验证R^2只有0.7005, 低于其他模型, 说明可能存在过度拟合问题。因此, 拟合R^2高并不一定说明模型样本外的预测能力强。Meng等[18]基于上海数据的研究也给出了两种R^2的差异, 比如拟合R^2为0.82的模型, 交叉验证R^2只有0.75。随机森林算法不同于计量回归模型, 没有计算拟合R^2, 但可计算交叉验证R^2。

基于监测站点数据得到模型的参数后, 将模型应用于县级城市NO_2浓度的预测。值得指出的是, 县级浓度预测的结果并不一定在表1给出的地面监测站的取值范围之内。下面以西藏自治区为例, 比较三种算法的预测浓度以及NASA原预测浓度, 结果如图4所示, 图中Pred NO_2表示NO_2的预测浓度。由于缺少地面环

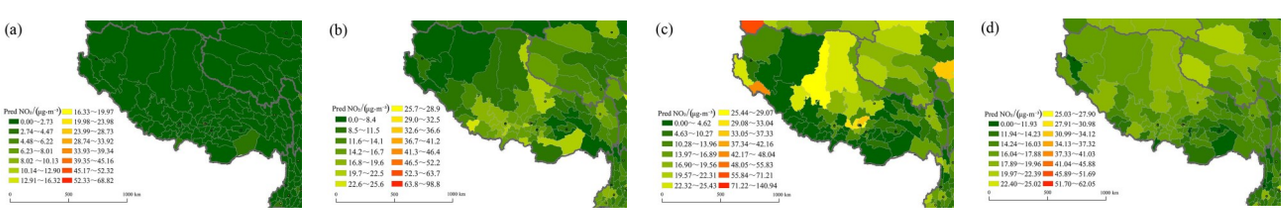
表 2 模型交叉验证结果的比较

Table 2 The comparison of cross validation results of the models

Model	CV r	CV R^2	Slope	RMSE	MAE	MAPE	Reg R^2
NASA	0.6937	0.4813	0.8246	14.3809	11.8655	0.4066	NA
GWR	0.8426	0.7100	0.7405	6.7895	5.2021	0.1904	0.7880
MGWR	0.8370	0.7005	0.8047	7.0156	5.3053	0.2002	0.9400
Random forest	0.8582	0.7365	0.7192	6.4223	4.9850	0.1953	NA
Random forest 2	0.8413	0.7078	0.7065	6.7986	5.2283	0.1994	NA

注: NA 表示该指标不适用于对应的模型

境监测站, 西藏地区的浓度相对于其他地区更难于预测 (有的研究甚至把西藏地区留空), 因此该地区是检验模型质量的关键区域。图 4 (a) 表明, 在缺少近邻因子纠正的情况下, NASA 原预测浓度显著偏低, 缺少变化, 且这种情况与事实情况不符, 比如在有监测站的那曲县, 地面监测值为 24.5 μg/m^3, 而 NASA 预测的浓度值只有 1.4 μg/m^3。图 4 (c) 中 MGWR 的预测浓度出现一些不该出现的极端值 (图中偏红的区域), 这是由于模型过度拟合, 其参数不太适应样本外的点, 从而导致极端值的出现。图 4 (b) 和图 4 (d) 的浓度分布较理想。相比之下, 随机森林算法得到的相邻地区浓度过渡更为自然, 且在靠近喜马拉雅山的地区也没有出现 GWR 浓度图中的偏黄地区 (GWR 图中同等颜色代表更高浓度), 故随机森林算法更为准确。



注: 此图基于国家自然资源部标准地图服务系统的标准地图 [审图号: GS(2016)1570 号] 绘制, 底图无修改

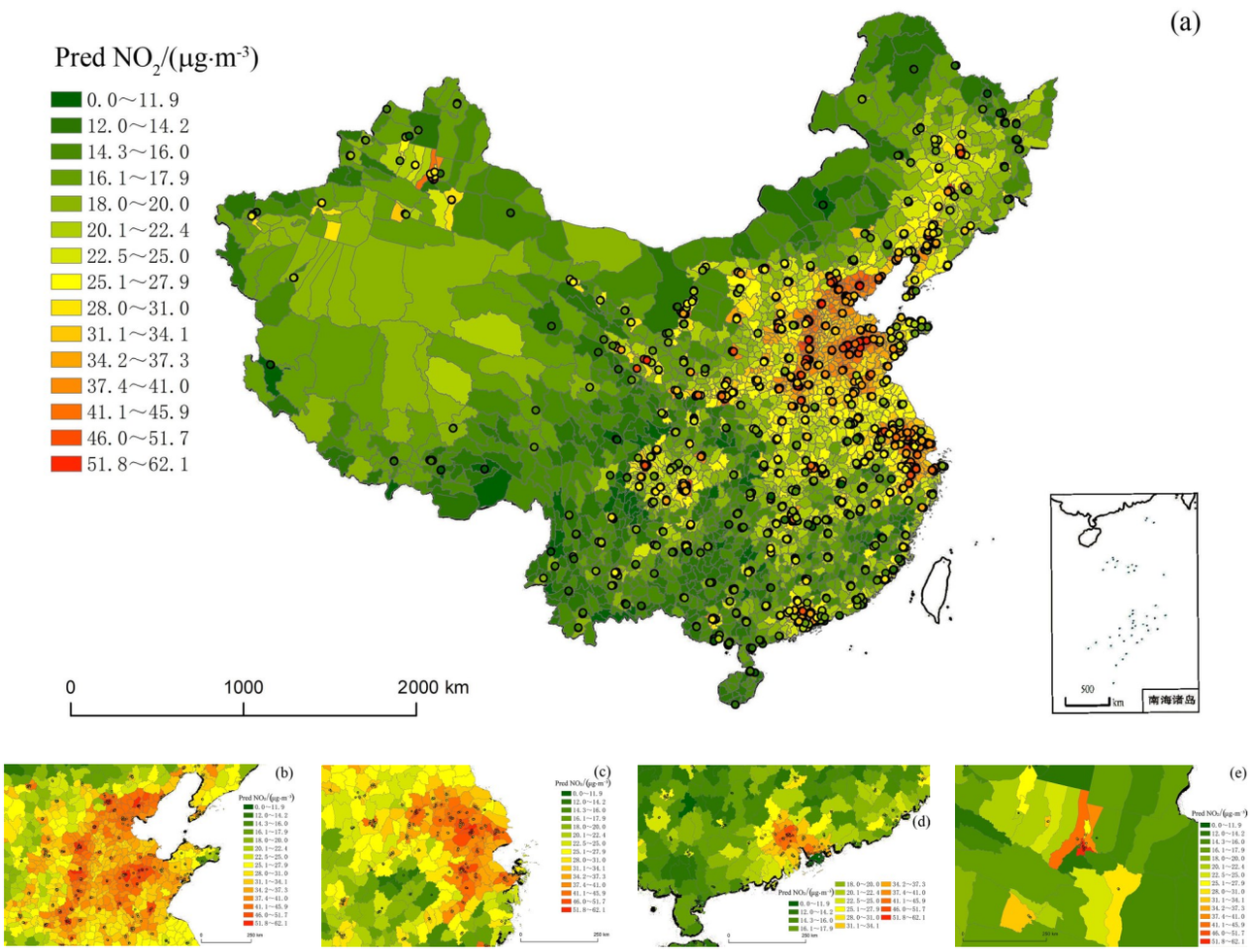
图 4　西藏自治区四种预测方法预测浓度的比较。(a) NASA_NO_2; (b) GWR_NO_2; (c) MGWR_NO_2; (d) RF_NO_2

Fig. 4　The comparison of the predicted concentrations from the four approaches for Tibet. (a) NASA_NO_2; (b) GWR_NO_2; (c) MGWR_NO_2; (d) RF_NO_2

综上所述, 采用综合表现最佳的随机森林算法预测县级城市的 NO_2 浓度, 结果如图 5 所示。图中的圆圈表示环境监测站的位置, 圆圈中的颜色代表监测浓度值的大小。县边界内的颜色表示随机森林预测 NO_2 浓度的大小。由图可见, 模型预测值和地面站浓度监测值高度一致。模型还给出了没有监测站的县的 NO_2 浓度预测值。值得注意的是, 由于在本研究使用城区所在位置, 因此图中县边界内的颜色仅代表该县城区的 NO_2 浓度, 而不代表农村地区。如果使用农村的位置来收集解释变量, 则可预测农村的 NO_2 浓度。但是, 由于农村地区可供选择的点太多, 因此不再一一列出各村或乡镇的预测结果。

预测结果表明, NO_2 在山区、西部地区和北部边境地区的浓度较低, 南部多雨地区的浓度也较低, 而华北平原地区 [图 5 (b)] 的浓度较高, 赵冉等[22]发现该地区的污染与人为生活源排放的相关系数非常高。京津冀和长三角 [图 5 (c)] 为 NO_2 浓度高且扩散范围比较广的地区。珠三角的浓度也比较高, 但扩散范围有限, 可能跟当地无大面积的平原有关。广东省 [图 5 (d)] NO_2 污染浓度较高的区域大概在佛山和广州交界处, 以及珠江入海口附近, 这一点与王肖汉等[23]的研究结论基本一致。王肖汉等[23]还指出珠三角的污染与经济状况、产业结构、常住人口数量及机动车保有量具有一定的相关关系, 这与本研究用到的解释变量是一致的。总的来

说, 经济活动强度高和人口密集的地区 NO_2 浓度高, 比如在新疆, 地广人稀, NO_2 浓度较高的区域就集中在乌鲁木齐市附近 [图5 (e)]。



注: 此图基于国家自然资源部标准地图服务系统的标准地图 [审图号: GS(2016)1570 号] 绘制, 底图无修改

图 5 随机森林算法预测出的县级 NO_2 浓度。(a) 全国; (b) 华北平原; (c) 长三角; (d) 广东省; (e) 新疆乌鲁木齐

Fig. 5 County-level NO_2 concentrations predicted by the random forest algorithm. (a) Nationwide; (b) North China Plain; (c) Yangtze River Delta; (d) Guangdong Province; (e) Urumqi, Xinjiang

相对于已有的研究, 本研究的预测精确度有所提高。比如 Harper 等[8]的交叉验证 R^2 为 0.44 和 0.65, 低于本研究的 0.7365。Meng 等[18]的交叉验证 R^2 为 0.75, 与本研究的相差不大, 但他们的研究针对上海单个城市, 一般针对单个大城市的研究的交叉验证 R^2 高于全国范围的研究。类似, 国外小区域的研究, 比如 Lee 和 Koutrakis[24]估算美国新英格兰地区 NO_2 浓度, 交叉验证 R^2 等于 0.79, 也远高于其他大范围的研究。Chen 等[12]针对整个欧洲的交叉验证 R^2 为 0.57 至 0.62, 低于本研究的交叉验证 R^2。Gu 等[25]的模拟结果和地面观察值的相关系数为 0.8 和 0.78, 低于本研究的相关系数 0.8582。

4 结 论

随机森林算法、GWR 和 MGWR 之间的比较表明, 就提高 NASA NO_2 反演浓度的精度而言, 三者中最佳

的模型为随机森林算法, 其中经济人口路网因子至少贡献11.24%。GWR模型存在着估算浓度超出正常范围的问题, 且其随地理位置改变的系数容易出现符号错误, 比如会出现污染物排放量的系数为负的问题。总体而言, GWR的估计精度不如随机森林算法。随机森林算法估计出来的浓度在地面观测浓度的范围之内, 随机森林算法能较好地处理异质性问题, 其算法比较稳健, 能避免过度拟合。但是, 随机森林算法也存在不足, 其拟合线的斜率比较低, 预测浓度偏低, 难以拟合极端值, 但这也是基于卫星遥感数据预测浓度的常见问题之一。MGWR由于采用多尺度的带宽, 可以部分纠正浓度低估问题。然而, 其存在过度拟合问题, 对样本外的点预测结果不理想。此外, MGWR被证明是一个非常耗时的算法。经过综合比较, 本预测模型选择随机森林算法。由于模型考虑了局部的经济、人口、地理和气象因素, 体现了人口密度和经济活动对NO_2浓度的影响, 故预测精度较高。

参考文献:

[1] Li M, Zhang Q, Kurokawa J, *et al*. MIX: A mosaic Asian anthropogenic emission inventory for the MICS-Asia and the HTAP projects [J]. *Atmospheric Chemistry & Physics*, 2015, 15: 34813-34869.

[2] Bechle M J, Millet D B, Marshall J D. National spatiotemporal exposure surface for NO_2: Monthly scaling of a satellite-derived land-use regression, 2000—2010 [J]. *Environmental Science & Technology*, 2015, 49(20): 12297-12305.

[3] Chan K L, Khorsandi E, Liu S, *et al*. Estimation of surface NO_2 concentrations over Germany from TROPOMI satellite observations using a machine learning method [J]. *Remote Sensing*, 2021, 13(5): 969.

[4] Cooper M J, Martin R V, McLinden C A, *et al*. Inferring ground-level nitrogen dioxide concentrations at fine spatial resolution applied to the TROPOMI satellite instrument [J]. *Environmental Research Letters*, 2020, 15(10): 104013.

[5] Knibbs L D, Hewson MG, Bechle M J, *et al*. A national satellite-based land-use regression model for air pollution exposure assessment in Australia [J]. *Environmental Research*, 2014, 135: 204-211.

[6] Novotny E V, Bechle M J, Millet D B, *et al*. National satellite-based land-use regression: NO_2 in the United States [J]. *Environmental Science & Technology*, 2011, 45(10): 4407-4414.

[7] Larkin A, Geddes J A, Martin R V, *et al*. Global land use regression model for nitrogen dioxide air pollution [J]. *Environmental Science & Technology*, 2017, 51(12): 6957-6964.

[8] Harper, A, Baker, P N, Xia, Y, *et al*. Development of spatiotemporal land use regression models for $PM_{2.5}$ and NO_2 in Chongqing, China, and exposure assessment for the CLIMB study [J]. *Atmospheric Pollution Research*, 2021, 12(7): 101096.

[9] Shi Y Y, Li R D, Qiu J, *et al*. Spatial distribution simulation and underlying surface factors analysis of NO_2 concentration based on land use regression [J]. *Journal of Geo-Information Science*, 2017, 19(1): 10-19.
施媛媛, 李仁东, 邱 娟, 等. 基于LUR的二氧化氮浓度空间分布模拟及其下垫面影响因素分析 [J]. 地球信息科学学报, 2017, 19(1): 10-19.

[10] Anand J S, Monks P S. Estimating daily surface NO_2 concentrations from satellite data—A case study over Hong Kong using land use regression models [J]. *Atmospheric Chemistry and Physics*, 2017, 17(13): 8211-8230.

[11] He Q, Kai Q, Cohen J B, *et al*. Spatially and temporally coherent reconstruction of tropospheric NO_2 over China combining OMI and GOME-2B measurements [J]. *Environmental Research Letters*, 2020, 15(12): 125011.

[12] Chen J, de Hoogh K, Gulliver J, *et al*. A comparison of linear regression, regularization, and machine learning algorithms to develop Europe-wide spatial models of fine particles and nitrogen dioxide [J]. *Environment International*, 2019, 130:

104934.

[13] Wang Z, Uno I, Yumimoto K, *et al*. Impacts of COVID-19 lockdown, Spring Festival and meteorology on the NO_2 variations in early 2020 over China based on *in situ* observations, satellite retrievals and model simulations [J]. *Atmospheric Environment*, 2021, 244: 117972.

[14] Lamsal L N, Martin R V, van Donkelaar A, *et al*. Ground-level nitrogen dioxide concentrations inferred from the satellite-borne Ozone Monitoring Instrument [J]. *Journal of Geophysical Research: Atmospheres*, 2008, 113(D16): D16308.

[15] Zheng K L, Huang Y, Yao X Y, *et al*. Correlation between $PM_{2.5}$, NO_2 and tourism activities, weather factors in Zhangjiajie City [J]. *Journal of Atmospheric and Environmental Optics*, 2020, 15(5): 347-356.
郑凯莉, 黄 毅, 姚小云, 等. 张家界市$PM_{2.5}$、NO_2与旅游活动及天气因素的相关性分析 [J]. 大气与环境光学学报, 2020, 15(5): 347-356.

[16] Breiman L. Random forests [J]. *Machine Learning*, 2001, 45(1): 5-32.

[17] Zhai L, Li S, Zou B, *et al*. An improved geographically weighted regression model for $PM_{2.5}$ concentration estimation in large areas [J]. *Atmospheric Environment*, 2018, 181: 145-154.

[18] Meng X, Chen L, Cai J, *et al*. A land use regression model for estimating the NO_2 concentration in Shanghai, China [J]. *Environmental Research*, 2015, 137: 308-315.

[19] Gilliland F, Avol E, Kinney P, *et al*. Air pollution exposure assessment for epidemiologic studies of pregnant women and children: Lessons learned from the Centers for Children's Environmental Health and Disease Prevention Research [J]. *Environmental Health Perspectives*, 2005, 113(10): 1447-1454.

[20] Health Effects Institute. Traffic-Related Air Pollution: A Critical Review of the Literature on Emissions, Exposure, and Health Effects [R]. Boston: Health Effects Institute, 2010, Special Report 17.

[21] Yang W B. *An Extension of Geographically Weighted Regression with Flexible Bandwidths* [D]. St Andrews: University of St. Andrews, 2014.

[22] Zhao R, Zhang C X, Wu Y, *et al*. Analysis of spatio-temporal variations of tropospheric nitrogen dioxide in the North China plain based on EMI [J]. *Journal of Atmospheric and Environmental Optics*, 2021, 16(3): 186-196.
赵 冉, 张成歆, 吴 跃, 等. 基于EMI观测华北平原对流层NO_2的时空变化研究 [J]. 大气与环境光学学报, 2021, 16(3): 186-196.

[23] Wang X H, Xu Y Z, Zhang C X, *et al*. Spatial-temporal variation of tropospheric NO_2 concentration in Pearl River Delta based on EMI observations [J]. *Journal of Atmospheric and Environmental Optics*, 2021, 16(3): 197-206.
王肖汉, 徐翼洲, 张成歆, 等. 基于EMI观测的珠三角地区对流层NO_2柱浓度时空变化特征分析 [J]. 大气与环境光学学报, 2021, 16(3): 197-206.

[24] Lee H J, Koutrakis P. Daily ambient NO_2 concentration predictions using satellite ozone monitoring instrument NO_2 data and land use regression [J]. *Environmental Science & Technology*, 2014, 48(4): 2305-2311.

[25] Gu J B, Chen L F, Yu C, *et al*. Ground-level NO_2 concentrations over China inferred from the satellite OMI and CMAQ model simulations [J]. *Remote Sensing*, 2017, 9(6): 519.