

基于深度学习方法 PM_{2.5} 精细化时空估算模型

耿冰¹, 孙义博², 曾巧林³, 商豪律⁴, 刘霄宇⁵, 单菁菁^{1*} (1. 中国社会科学院生态文明研究所, 北京 100710; 2. 中国环境科学研究院生态研究所, 北京 100021; 3. 重庆邮电大学计算机科学与技术学院, 重庆 400065; 4. 中国科学院空天信息研究院数字地球重点实验室, 北京 100094; 5. 中国信息通信研究院, 北京 100191)

摘要: 提出一种基于深度学习方法的 PM_{2.5} 浓度时空估算模型(PM_{2.5}-DNN), 该模型基于葵花-8 卫星反演的 AOD 数据, 结合 PM_{2.5} 监测站和气象站观测数据对北京市地面 PM_{2.5} 浓度进行了逐时的高精度模拟, 同时将 PM_{2.5}-DNN 模型的模拟性能与当前的主流方法进行了对比研究。结果表明, 使用 PM_{2.5}-DNN 模型估算的北京地区 1km 分辨率每小时地面 PM_{2.5} 浓度与地表监测站观测数据对比的一致性较好, 模型估算精度可达到 $R^2=0.88$, 性能优于当前的主流方法。本文所提出的方法适用于区域尺度 PM_{2.5} 浓度时空分布细粒度建模与估算, 采用端到端的训练方式构建模型, 为精细的 PM_{2.5} 浓度估算提供了一个简便而有效的方法模型。

关键词: 大气细颗粒物浓度估算; 深度学习; 卫星遥感; 光学气溶胶厚度

中图分类号: X513 **文献标识码:** A **文章编号:** 1000-6923(2021)08-3502-09

Refined spatiotemporal estimation model of PM_{2.5} based on deep learning method. GENG Bing¹, SUN Yi-bo², ZENG Qiao-lin³, SHANG Hao-lv⁴, LIU Xiao-yu⁵, SHAN Jing-jing^{1*} (1. Research Institute for Eco-civilization, Chinese Academy of Social Sciences, Beijing 100710, China; 2. Institute of Ecological Environment, Chinese Research Academy of Environmental Sciences, Beijing 100012, China; 3. College of Computer and Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China; 4. Key Laboratory of Digital Earth Science, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China; 5. China Academy of Information and Communications Technology, Beijing 100191, China). *China Environmental Science*, 2021,41(8): 3502~3510

Abstract: The concentration distribution of fine particulate matter (PM_{2.5}) on the surface of the atmosphere has a strong temporal and spatial heterogeneity. Due to the limited spatial coverage of traditional PM_{2.5} monitoring sites, it is difficult to reflect the complexity of PM_{2.5} concentration in time and space. This paper proposed a temporal and spatial prediction model of ground PM_{2.5} concentration based on deep learning methods(PM_{2.5}-DNN). Based on the AOD data from Kuahua-8satellite and the observation data from PM_{2.5} monitoring and meteorological station, hourly high-precision simulations of the surface PM_{2.5} concentration in Beijing had been carried out. The results show that the 1km resolution hourly ground PM_{2.5} concentration in Beijing area estimated by the PM_{2.5}-DNN model had good consistency with the observation data from the surface monitoring station. The model estimation accuracy could reach $R^2=0.88$, which was better than the performance of current mainstream method. The method proposed in this paper was suitable for fine-grained modelling and estimation of the temporal and spatial distribution of PM_{2.5} concentration at a regional scale. The end-to-end training method is used to construct the model, which provides a simple and effective method model for fine PM_{2.5} concentration estimation.

Key words: PM_{2.5} concentration estimation; deep learning; satellite remote sensing; aerosol optical depth

为了应对日益严重的空气污染问题,我国大规模建立地面 PM_{2.5} 监测站点,对重污染天气进行监测及预警^[1-3]。众多学者也在此基础上开发了 PM_{2.5} 浓度估算模型^[4-8]。但是,地面 PM_{2.5} 监测站点仅提供空间上“点”尺度的观测,有限的空间范围不足以代表 PM_{2.5} 在空间上的异质性,而 PM_{2.5} 的空间信息对于研究空气污染与经济、地理及人口之间的关系至关重要。

近年来,随着卫星遥感技术的发展,使得区域尺度的污染信息获取成为可能,利用卫星反演的大气

气溶胶光学厚度(AOD)估算大空间尺度的地表 PM_{2.5} 浓度也已被广泛采用^[9-13]。

目前,已有多种卫星遥感数据反演的 AOD 产品被用于估算地表 PM_{2.5} 浓度的时空分布^[14-17]。

与此同时,多种类型的统计模型被提出并应用于 PM_{2.5} 浓度的估算中^[18-22],这些模型均以 AOD 作

收稿日期: 2020-12-28

基金项目: 中国博士后科学基金资助项目(2018M631684);国家自然科学基金资助项目(41590855)

* 责任作者, 研究员, shanj2008@126.com

为主要指示因子,结合气象观测及其他类型的统计参数估算地面 $\text{PM}_{2.5}$ 的时空分布.例如,早期的研究采用一元线性回归模型仅采用 AOD 作为指示因子来估算 $\text{PM}_{2.5}$ 浓度^[23];或更为复杂的采用多元或广义线性回归模型,考虑更多的地表及气象参数用以提高 $\text{PM}_{2.5}$ 的估算精度^[24-25].但是在真实的环境中, $\text{PM}_{2.5}$ 浓度的分布是一个与多种因素有关的非线性过程,在时间和空间上存在着强烈的可变性,因此学者们开发了更为复杂的模型对 $\text{PM}_{2.5}$ 浓度与 AOD 关系的时空变异性进行描述,例如地理(和时间)加权回归模型^[10]、混合效应模型^[26]以及广义加权混合模型等^[27].然而,本质上这些统计模型仍然是线性的,模型内部简化了 $\text{PM}_{2.5}$ 与 AOD 及其他指示因子之间的复杂关系,使 $\text{PM}_{2.5}$ 浓度估算结果仍然存在较大的不确定性.随着计算机技术的发展,机器学习(包括深度学习)方法以其强大的非线性建模能力越来越多地被用于 $\text{PM}_{2.5}$ 浓度的估算中^[28],例如支持向量回归模型^[29]、随机森林模型^[30]、人工神经网络模型^[24]、贝叶斯方法^[31]、广义回归神经网络模型^[32]以及深度信念网络^[33]等,这些模型在对 $\text{PM}_{2.5}$ 浓度的估算方面均表现出比传统统计模型更好的性能.在指示因子的选择方面,这些机器学习模型除了采用 AOD 和常规的气象观测参数之外,还使用了包括相邻时间和空间上观测的 $\text{PM}_{2.5}$ 信息、土地利用信息、植被指数信息、 NO_2 浓度信息、人口密度、海拔高程^[26, 33]以及路网密度信息等,这些信息或多或少与 $\text{PM}_{2.5}$ 浓度分布相关.考虑的影响因子越多,越能够提高 $\text{PM}_{2.5}$ 估算精度.但是,过多的人工设计的特征不仅耗时耗力,而且过于复杂的特征选择也不利于模型的工程化实施.此外,目前的模型大多仅对日平均 $\text{PM}_{2.5}$ 浓度进行估算,且空间分辨率相对粗糙(大于 3km).尽管此类模型可以有效降低目标函数的复杂性,但却忽略了 $\text{PM}_{2.5}$ 浓度每小时的时空变异性.针对以上问题,为了有效地开展 $\text{PM}_{2.5}$ 浓度精细化时空尺度(即每小时和 1km 的时空分辨率)估算,需要一种非线性表达能力更强并且容易实现工程化的模型.

深度学习方^[34]作为当前最先进的机器学习技术之一,以其优异的非线性表达能力在许多领域都取得了超过传统机器学习方法的显著成果.目前,已有研究人员采用深度学习方法来估算 $\text{PM}_{2.5}$ 浓度的时空分布^[33,35-36],但是相关的模型规模仍然相对

较小,并且很大程度上依旧依赖于人工特征选择,并没有充分利用深度学习方法通过更深更宽的网络结构来表达高度复杂目标函数的优点.因此,本文以北京市 2017 年的 $\text{PM}_{2.5}$ 观测数据为基础,提出了一种典型的深度学习模型($\text{PM}_{2.5}$ -DNN),仅采用卫星遥感反演的 AOD 数据以及常规的气象观测要素(例如气温、地表温度、风速、风向、相对湿度、压强以及能见度)来估算 $\text{PM}_{2.5}$ 浓度的时空分布.

1 数据与模型

1.1 研究区与数据来源

1.1.1 研究区概况 北京市位于华北平原,中心经纬度为 116.41°E , 39.92°N ,总面积约 16410.54km^2 ,2019 年常住人口约 2153.6 万人.其地处暖温带半湿润半干旱季风区,气候四季分明,夏季炎热多雨,冬季寒冷干旱,春秋短促.本文采用了北京市 2017 年全年观测的每小时 $\text{PM}_{2.5}$ 浓度数据和气象观测数据.研究区域地理空间范围以及 $\text{PM}_{2.5}$ 站点和气象观测站点的分布见图 1 所示.

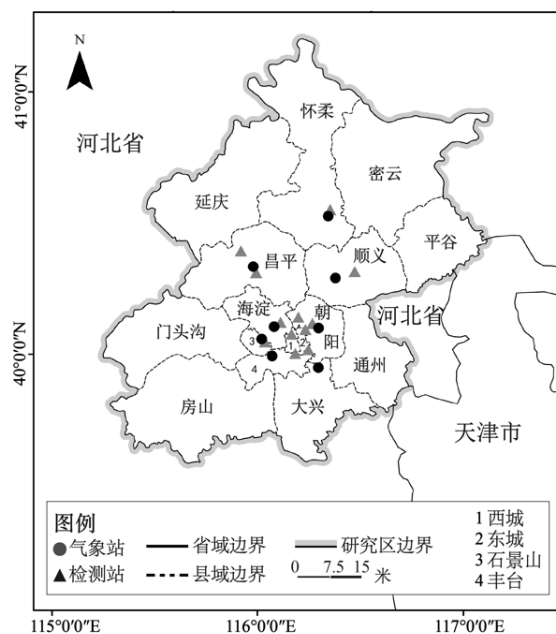


图1 研究区及 $\text{PM}_{2.5}$ 监测站和气象观测站点分布

Fig.1 Distribution of study area and $\text{PM}_{2.5}$ monitoring stations and meteorological observation stations

1.1.2 卫星 AOD 数据 本文卫星遥感数据采用葵花-8 气象卫星数据.葵花-8 属于第 3 代地球静止气象卫星,其观测范围为东西 $80^\circ\text{E}\sim 160^\circ\text{W}$,南北 $60^\circ\text{N}\sim$

60°S,距离地面高度 35800km,星下点位于 140.7°E^[17].卫星搭载的主要传感器为 AHI(高像素葵花成像仪),可见光最高分辨率为 0.5km,红外-近红外最高空间分辨率为 1km,最高时间分辨率为 10min,是目前全球最先进的气象观测传感器之一.

葵花-8 号卫星 AOD 数据产品的反演采用了 Yang 等提出的新暗目标算法(New-DT),该算法主要利用葵花-8 卫星的可见光和近红外波段数据反演空间分辨率为 1km 的每小时 AOD 数据产品.

1.1.3 环境及气象观测数据 2017 年 PM_{2.5} 监测站点数据来自国家环境监测中心网站,共采用了北京市 12 个 PM_{2.5} 监测站点.2017 年北京市气象观测数据来自于中国气象数据中心网站,本文所用到的主要气象观测数据包括气温(A_temp)、地表温度(S_temp)、风速(wing_S)、风速方向(wind_D)、相对湿度(RH)、地表压强(SP)和能见度(VIS).为了获得与 PM_{2.5} 浓度相关性最佳的气象观测数据,本文选用了在 PM_{2.5} 监测站半径为 5km 区域内的气象观测站,最后共选择了 8 个气象观测站点.PM_{2.5} 监测站点与气象观测站点的分布如图 1 所示.

1.1.4 数据预处理 由于模型的构建需要将 PM_{2.5} 浓度数据、AOD 数据以及气象观测数据一一对应,因此需要对所有数据进行预处理,使之形成时间和空间上一致的数据集,用于对所构建的 PM_{2.5} 浓度估算模型进行训练与验证.其中 AOD 数据提取 PM_{2.5} 监测站点位置所在栅格的数值.对于气象观测数据,由于 PM_{2.5} 监测站点的分布与气象观测站点的分布不同,本文使用了 PM_{2.5} 周围 5km 范围内的气象观测站点观测均值作为与之匹配的气象观测数据.此外,由于较小的太阳高度角会导致大气路径变长,使得大气漫反射的比例变大,从而导致卫星遥感反演的 AOD 数据产品的精度变差.因此,为了有效避免太阳高度角对 AOD 数据反演精度的影响,本文仅使用北京时间 9:00~16:00 之间 8h 的葵花 8 号卫星影像反演 AOD 数据,并同时获取与 AOD 数据时间上一致的 PM_{2.5} 浓度与气象观测数据.最后,去掉数据中的缺失值以及由于降雨影响导致的无效数据,并将数据进行归一化处理(零均值,单位方差).经过预处理后,共得到可用数据 17059 条.

1.2 模型构建方法

1.2.1 深度网络模型(PM_{2.5}-DNN) 深度学习属于

机器学习中的一类模型,它通过构建非常深的神经网络来学习输入数据的多级表示特征,从而表达复杂及抽象的概念或模式^[34].与传统神经网络模型相比,深度学习模型通常由超过三层的隐含层构成.并且可以在没有进行预训练的情况下直接对深度网络进行端到端的监督训练.

本文采用 AOD 数据和常规的气象观测参数作为输入变量,通过深度学习模型直接构建输入变量与 PM_{2.5} 浓度之间的高时空分辨率(每小时,1km)关系模型.模型的输入参数为:葵花 8 卫星反演的 AOD、空气温度(A_temp)、地表温度(S_temp)、风速(wind_S)、风向(wind_D)、相对湿度(RH)、压强(SP)、能见度(VIS)以及年积日(DOY),模型的输出为 PM_{2.5} 浓度的估算值.对模型隐含层神经单元数据的确认采用启发式搜索方法,将隐含层的神经单元个数设为 10,并以 10 为步长进行迭代的训练和验证,并统计验证误差,直至验证误差稳定且不再降低,然后以相同的方式来确定多个隐含层的神经元个数,直至整个模型的验证误差不再降低为止.模型的训练采用 ReLU 作为激活函数,输出层采用线性函数作为激活函数,最终确定的深度学习网络模型结构如图 2 所示,结构为 9-300-300-100-20-1.

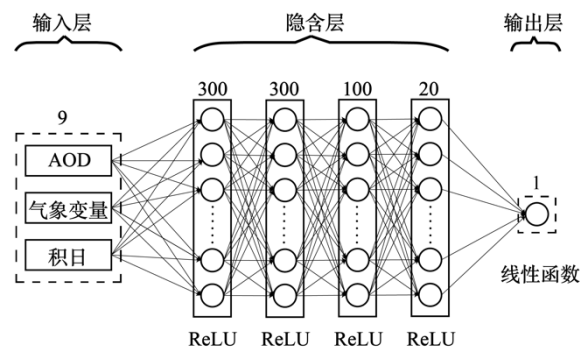


图 2 本文所采用的深度网络模型结构

Fig.2 The structure of the deep network model

模型的训练和应用流程如图 3 所示,在训练阶段通过采用误差反向传播方法对输入的训练数据集进行学习,获得能够表征输入数据时空特征的模型内部参数,并对模型的训练结果进行验证,在获得可靠的验证精度后即可应用训练好的模型,通过输入 AOD 影像及栅格化后的气象观测数据,对 PM_{2.5} 浓度的时空分布进行估算.

1.2.2 线性混合效应模型(LME) 线性混合效应

模型(LME)是目前估算 PM_{2.5} 时空分布的主要方法之一,是既包含了固定效应参数又包含随机效应的统计回归模型.其中固定效应表示模拟因子对 PM_{2.5} 的多年平均影响状态,而随机效应则用于解释 PM_{2.5} 与 AOD 以及其他气象因子之间的日变化关系,以随机截距或者随机系数的形式表示.线性混合效应模型可以表达为:

$$\begin{aligned} \text{PM}_{2.5, it} = & (b_0 + b_{0,t}) + (b_1 + b_{1,t})\text{AOD}_{it} + \\ & (b_2 + b_{2,t})\text{HP}_{it} + (b_3 + b_{3,t})\text{A_temp}_{it} + (b_4 + \\ & b_{4,t})\text{RH}_{it} + (b_5 + b_{5,t})\text{wind_D}_{it} + \\ & (b_6 + b_{6,t})\text{wind_S}_{it} + (b_7 + b_{7,t})\text{S_temp}_{it} + b_8 \cdot \\ & \text{DOY} + \varepsilon_{it} \quad (b_{0,t}, b_{1,t}, b_{2,t}) \sim N[(0, 0, 0), \Psi] \end{aligned} \quad (1)$$

式中:PM_{2.5, it} 表示第 i 个监测站点在时间 t 的 PM_{2.5} 浓度值; b_0 和 $b_{0,t}$ 分别代表模型的固定截距和随机截距; $b_1 \sim b_9$ 和 $b_{1,t} \sim b_{9,t}$ 分别代表各自变量参数的固定效应斜率和各变量的随机效应斜率; ε_{it} 第 i 个监测站点在时间 t 的随机误差项, $b_{0,t}$ 、 $b_{1,t}$ 和 $b_{2,t}$ 为其参数; Ψ 为随时间变化的随机效应方差-协方差矩阵.

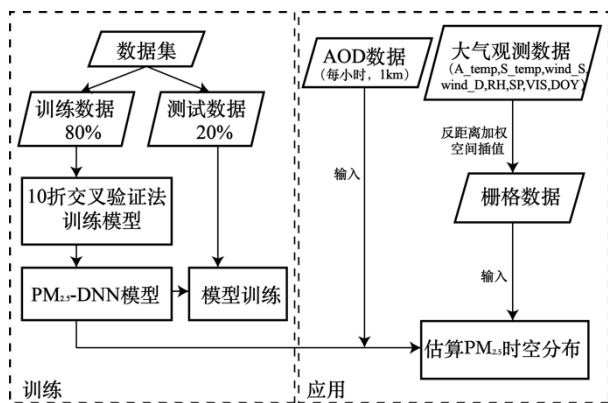


图3 模型训练和应用流程图

Fig.3 Model training and application flow chart

1.2.3 地理加权回归模型(GWR) 地理加权回归模型(GWR)区别于传统回归方法之处在于不同辅助变量的回归系数不再是利用全局信息所获得的常量进行估算,而是用邻近观测值进行局部加权回归从而得到相关系数,并考虑了数据的空间位置.其公式可表达为:

$$\begin{aligned} \text{PM}_{2.5, i} = & \beta_{0, i} + \beta_{1, i} \cdot \text{AOD} + \beta_{2, i} \cdot \text{HP} + \beta_{3, i} \cdot \\ & \text{A}_{\text{temp}} + \beta_{4, i} \cdot \text{RH} + \beta_{5, i} \cdot \text{wind}_D + \beta_{6, i} \cdot \text{wind}_S + \\ & \beta_{7, i} \cdot \text{S}_{\text{temp}} + \beta_{8, i} \cdot \text{DOY} + \varepsilon_i \end{aligned} \quad (2)$$

式中:PM_{2.5, it} 表示第 i 个监测站在时间 t 的 PM_{2.5} 浓度

值, $\beta_{0, i}$ 为模型的固定截距; $\beta_{1, i} \sim \beta_{8, i}$ 分别代表各自变量参数的回归系数, ε_i 为第 i 个监测站点的回归残差.

1.2.4 支持向量回归模型(SVR) 支持向量回归(SVR)模型是支持向量机模型在回归问题上的应用模型,该模型会在尽可能拟合现有数据的情况下考虑模型的泛化误差,从而尽量避免模型的过拟合,使得学习到的模型能够在未知数据上具有良好的预测性能.SVR 模型在 PM_{2.5} 模拟方面已有了较多的应用,以下对其原理进行简单叙述.

给定数据集 $\{(x_i, y_i), i=1, \dots, m\}$, x 为输入因子, y 为输出因子, SVR 模型可以表示为:

$$y = f(x) = w^T \phi(x) + b \quad (3)$$

式中: w 是权重向量; $\phi(x)$ 是将输入数据从输入空间映射到特征空间的核函数; b 是常数项.对模型的训练是使风险函数最小,风险函数可表示为:

$$L(f) = \frac{1}{2} \|w\|^2 + C \sum_{i=1}^m L[y_i, f(x_i)] \quad (4)$$

式中:等号右侧第一项是对模型复杂程度的惩罚项;第二项是模型输出值与真实值之间误差的惩罚项;常数 C 是用来调整惩罚比例的系数.本文使用径向基函数(RBF)为核函数,表示为:

$$\phi(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right) \quad (5)$$

在对 SVR 模型的训练中,采用了格网搜索方法来确定模型超参数(σ, C),即设定 σ 和 C 的范围在 4 到 -4 之间,以 0.8 为间隔进行遍历计算,找到模型验证误差最小的超参数组合即为最优参数.

1.2.5 随机森林回归模型(RFR) 随机森林(RF)算法是通过集成学习的思想将多颗决策树集成的一种分类与回归算法.随机森林引入集成学习思想和随机子空间思想,通过实现样本选取随机性和特征选取随机性,对样本单独构建决策子树,结合集成学习思路将各决策子树的结果按照一定规则汇总作为最后输出.对于回归模型而言,汇总规则为取平均值.

对一组由决策子树 $\{h(x, \theta_t), t=1, 2, \dots, T\}$ 构成的决策组合模型.其中 θ_t 为服从独立同分布的随机变量, x 为自变量, T 为决策子树的个数.回归模型的估算结果为:

$$\bar{h}(x) = \frac{1}{T} \sum_{t=1}^T \{h(x, \theta_t)\} \quad (6)$$

式中: $\bar{h}(x)$ 表示回归估算结果; $h(x, \theta_t)$ 为基于 x 和 θ_t 的输出.训练随机森林的过程就是训练各个决策树的

过程,最后将各个决策树取平均,就可得到一个随机森林结果.

2 结果与讨论

2.1 深度网络模型 PM_{2.5}-DNN 的训练和验证

将数据集随机分隔为训练数据集(80%)和测试数据集(20%)分别应用于模型的训练阶段和测试阶段.对模型的训练采用带动量项的随机梯度下降算法,并采用从均匀分布中采样的方式对 PM_{2.5}-DNN 模型的初始权重进行初始化.此外,超参数的设置也会显著的影响模型训练结果,超参数主要包括学习速率,动量以及为防止过度拟合而引入的 dropout 和正则化因子^[37].本文对超参数的设置采用了格网搜索的方法,即设置各超参数取值范围,然后以一定的步长进行迭代训练,找出获得最佳验证性能的数值作为后续模型训练时采用的超参数.本文最终确定的模型训练超参数设置情况为:批量大小为 100,学习速率为 0.04,动量为 0.8,dropout 为 0.1,L2 正则化因子为 0.0002,同时采用早停法来避免模型的过度拟合.对模型的训练采用 10 折交叉验证方法,即将训练数据集随机平均分为 10 等份,轮流将其中 9 份作为训练数据,1 份作为验证数据进行模型训练,最后将 10 次训练的验证结果的均值作为模型的性能^[37].

采用相关性系数(R^2)、均方根误差(RMSE)、平均预测误差(MPE)以及相对预测误差(RPF)作为模型性能的评价指标.

模型训练结构如图 4 所示,在模型的训练阶段,模型的总体性能分别为: $R^2=0.95$, $RMSE=10.6\mu g/m^3$, $MPE=1.64\mu g/m^3$, $RPE=25.47\%$;在模型的测试阶段,模型的总体性能为: $R^2=0.88$, $RMSE=18.78\mu g/m^3$, $MPE=0.73\mu g/m^3$, $RPE=42.98\%$.通过对比训练阶段和测试阶段的 R^2 可以看出,模型在训练阶段存在一定的过拟合现象,导致了所构建的 PM_{2.5}-DNN 模型在测试阶段对高浓度范围 PM_{2.5} 的低估和低浓度范围 PM_{2.5} 的高估.但同时测试结果的相关性系数和误差水平仍然表现的非常稳健,表明本文所构建的 PM_{2.5}-DNN 模型对 PM_{2.5} 浓度的估算仍然非常有效.

2.2 深度网络模型与其他模型的对比

从表 1 可见,机器学习方法(包括 SVR,RFR 和 PM_{2.5}-DNN 模型)表现出了比传统方法(LME 和 GWR)更好的估算性能.PM_{2.5}-DNN 与 SVR 和 RFR

模型相比表现出了更加优异的估算性能,这主要得益于 PM_{2.5}-DNN 模型的深层结构,能够从大量数据中直接学习高度复杂的函数关系.同时,本结果也表明,深度网络模型在 PM_{2.5} 浓度的估算方面具有极高的应用潜力,可以直接通过端到端的方式进行训练,即直接使用容易获得的观测因子进行建模便能够获得最佳的估算性能.

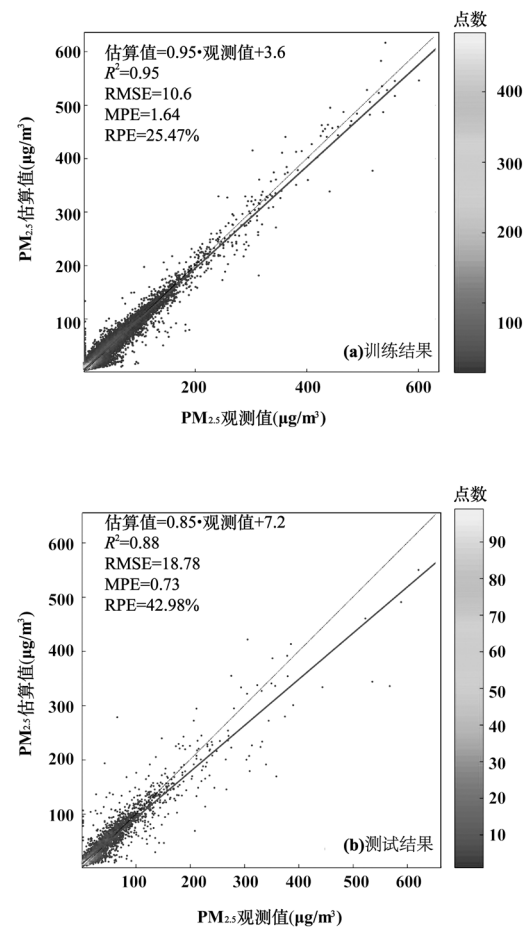


图 4 PM_{2.5}-DNN 模型的总体性能评价

Fig.4 Performance evaluation of PM_{2.5}-DNN model

表 1 不同 PM_{2.5} 估算模型之间的结果对比

Table 1 Results comparison of different PM_{2.5} estimation models

模型	模型训练				模型测试			
	R^2	RMSE	MPE	RPE (%)	R^2	RMSE	MPE	RPE (%)
LME	0.71	25.1	15.5	52.7	0.63	29.0	18.1	60.9
GWR	0.88	14.6	10.0	31.4	0.76	23.3	16.7	56.3
SVR	0.90	15.1	7.1	36.2	0.83	22.3	10.9	43.4
RFR	0.94	10.9	5.5	26.0	0.87	18.9	9.8	43.3
PM _{2.5} -DNN	0.95	10.6	1.64	25.47	0.88	18.78	0.73	42.98

2.3 北京市 PM_{2.5} 浓度的精细化时空分布估算

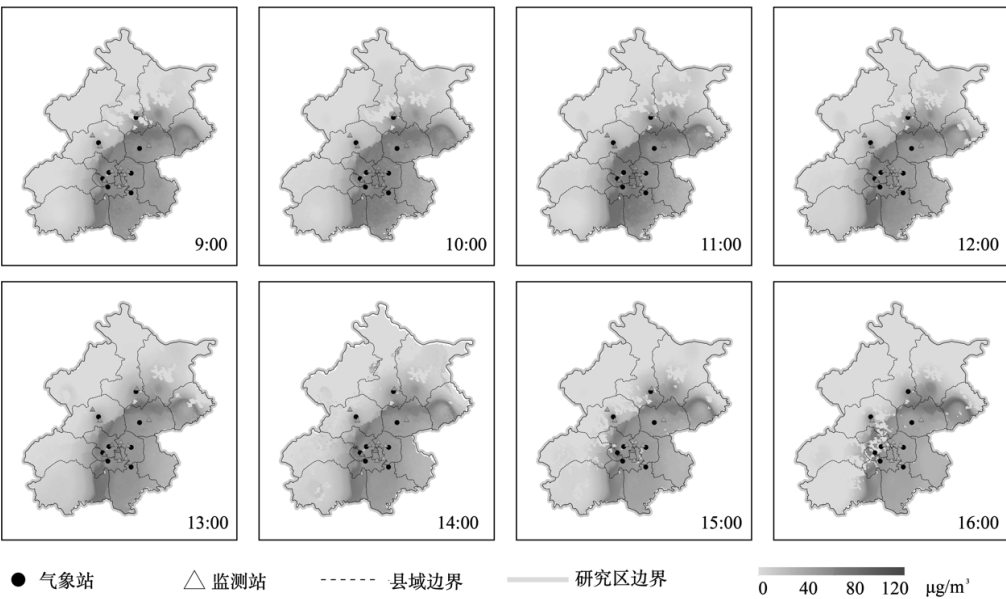


图 5 每小时 PM_{2.5} 浓度模拟结果与监测站点的对比

Fig.5 Hourly PM_{2.5} concentration estimated by PM_{2.5}-DNN model compared with the monitoring site

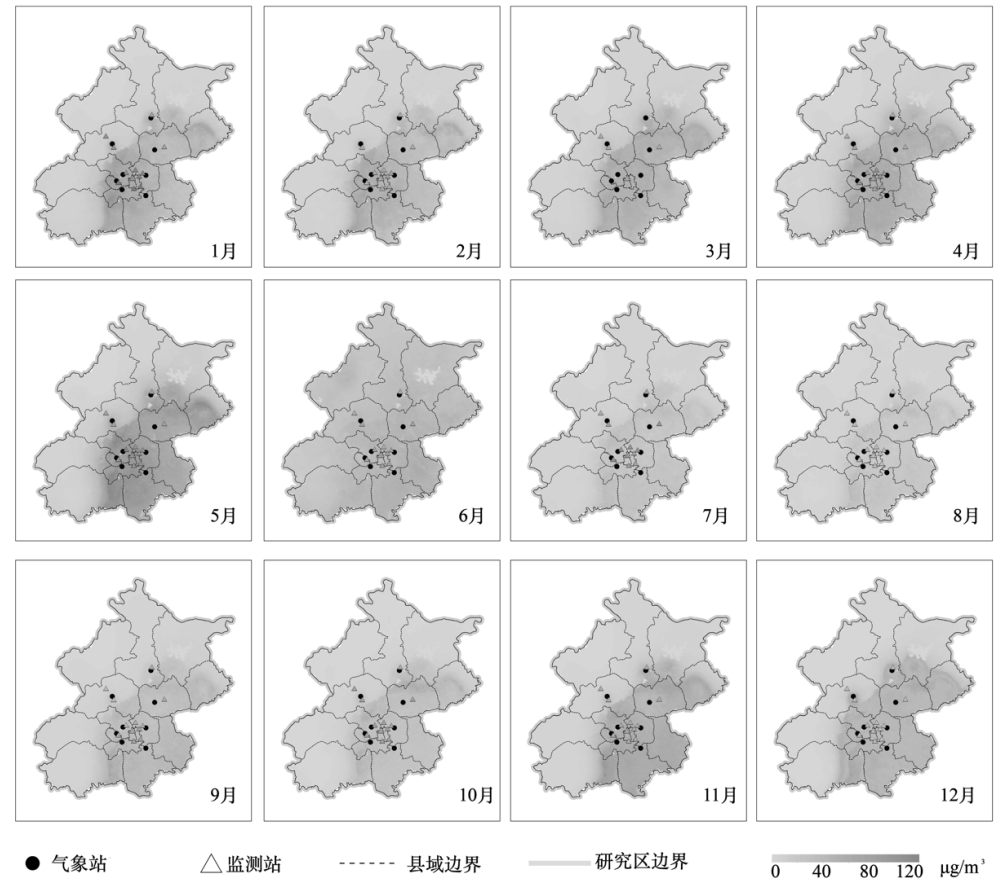


图 6 PM_{2.5}-DNN 模型生成的 2017 年各月份 PM_{2.5} 浓度分布

Fig.6 PM_{2.5} concentration distribution per month in 2017 generated by PM_{2.5}-DNN model

2.3.1 每小时 PM_{2.5} 浓度估算分析 为获得能够覆盖整个北京市区域的气象观测插值数据,本文采用了覆盖整个京津冀区域内的气象观测站点并采用基于反距离加权方法(IDW)进行插值获得北京市气象栅格数

据.然后将训练好的PM_{2.5}-DNN模型应用于卫星AOD栅格数据及插值生成的气象栅格数据生成2017年北京市每小时PM_{2.5}浓度栅格图.图5中,最大PM_{2.5}浓度分别为:89.92、82.37、91.42、89.89、89.73、83.27、80.6和91.62μg/m³.结合与地面PM_{2.5}监测站点的数据相对比,PM_{2.5}-DNN模型估算的PM_{2.5}浓度时空分布与地面实测结果均吻合较好,表明本文所构建的PM_{2.5}-DNN模型具有较好的时空估算性能.

2.3.2 各月份PM_{2.5}浓度分布估算分析 从各月份PM_{2.5}浓度分布来看,北京市2017年1~12月份最大PM_{2.5}浓度分别为:51.56、39.05、34.05、34、57.48、29.23、18.49、15.09、32.67、27.41、46.77和39.47μg/m³.在PM_{2.5}浓度时空分布中可知,冬季节污

染最为严重,夏季空气质量最优.研究结果也清晰地展示了2017年5月发生的北京市PM_{2.5}重度污染过程(图6).

2.3.3 各行政区PM_{2.5}浓度分布分析 北京市共16个行政区,但仅有12个观测站点,无法实现观测站点的行政区全覆盖,利用PM_{2.5}-DNN模型可生成更加精细的PM_{2.5}浓度时空分布数据.从计算结果可知(表2),2017年北京市PM_{2.5}浓度最大值出现在房山和丰台,其次为昌平、海淀、石景山,延庆PM_{2.5}浓度最低.城六区(东城、西城、朝阳、海淀、丰台、石景山)年平均PM_{2.5}浓度均高于其他地区,生态涵养区(门头沟、平谷、怀柔、密云、延庆、昌平和房山的山区)PM_{2.5}浓度较低.

表2 PM_{2.5}-DNN模型生成的2017年各行政区PM_{2.5}浓度分析
Table 2 Analysis of PM_{2.5} concentration in each administrative district in 2017 generated by PM_{2.5}-DNN model

序号	行政区名	最小值	最大值	平均值	方差	序号	行政区名	最小值	最大值	平均值	方差
1	房山	0.0000	31.1162	8.4032	8.9197	9	顺义	12.5507	25.4875	19.6331	2.0994
2	丰台	16.9736	30.6748	24.3146	2.6368	10	密云	0.0000	25.0423	5.7729	5.4797
3	昌平	1.9277	28.5647	8.8862	7.3120	11	西城	20.5477	23.4613	22.2668	0.6080
4	海淀	0.0000	28.5575	19.9157	6.0518	12	朝阳	16.5281	23.2015	20.9610	1.0962
5	石景山	18.2775	27.9831	22.2454	2.0354	13	通州	11.4474	23.1173	18.9646	1.4739
6	怀柔	0.0000	26.4295	4.5850	5.4393	14	东城	20.5213	22.6950	22.0519	0.3635
7	大兴	10.3162	25.7732	21.6578	2.3546	15	门头沟	1.0519	21.8149	4.1309	4.5128
8	平谷	3.5526	25.5354	15.3766	6.6629	16	延庆	0.0000	3.8201	1.5807	0.5024

同时,本文将所用12个观测站点的年度平均值与其所在栅格单元的PM_{2.5}-DNN模型模拟结果的年度平均进行了对比,其中 $R^2=0.89$,RMSE=20.04μg/m³,与模型的测试结果一致,进一步论证了所构建的PM_{2.5}-DNN模型的可靠性(图7).

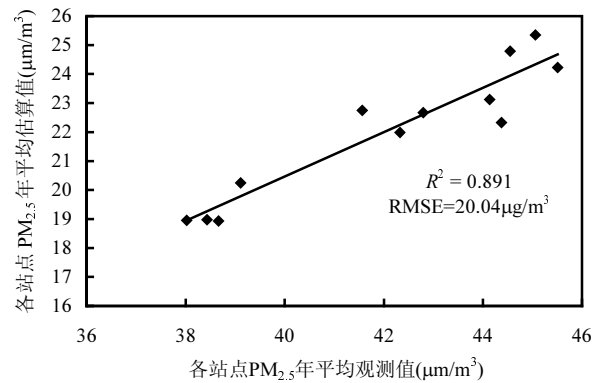


图7 各站点年平均观测值与估算值对比

Fig.7 Comparison of annual average observation values and estimated values of each station

2.4 讨论

与LME、GWR、SVR以及RFR相比,本文提出的基于深度网络模型的PM_{2.5}-DNN模型仅采用容易获得的观测因子就可以获得最佳的估算性能.结合卫星遥感反演的AOD数据,采用PM_{2.5}-DNN模型实现了PM_{2.5}浓度1km逐时的时空精细化模拟.

在模型应用方面,由于模型的建立主要是基于2017年北京市内的12个PM_{2.5}监测站点、8个气象观测站点以及葵花-8号卫星的AOD数据产品,考虑到其他年份或区域会出现不同的PM_{2.5}时空模式,模型应用的最佳方式是进行内插,因此本文仅对2017年北京市PM_{2.5}浓度的时空分布进行了模型估算.对于深度网络模型而且,随着模型的深度(隐含层数)和广度(每层神经元个数)的增加以及模型构建时输入数据量的增大,所训练的深度网络模型就越能更好地对复杂PM_{2.5}时空变化模式进行模拟.因此,采用更多年份及更多观测站点的数据进行深度网络模型的构建,实现全国区域多年份的PM_{2.5}浓度时空

精细化模拟是本文下一步的研究重点.此外,由于云层覆盖造成的 AOD 数据在空间上的数据缺失,造成了 $\text{PM}_{2.5}$ 浓度时空分布数据在部分空间上的不连续,这也是本文存在的不足之一.

3 结论

3.1 提出一种基于深度学习方法的 $\text{PM}_{2.5}$ 浓度时空估算模型 $\text{PM}_{2.5}$ -DNN,该模型仅需要常规的气象观测数据(包括气温、地表温度、风速、风向、相对湿度、压强以及能见度)结合卫星遥感反演的 AOD 数据,就可以对地表 $\text{PM}_{2.5}$ 浓度进行高时空分辨率的估算.

3.2 深度学习模型在构建复杂关系模型中具有较强的性能,与线性混合效应模型、地理加权回归模型、支持向量回归模型、随机森林回归模型相比, $\text{PM}_{2.5}$ -DNN 模型表现出更高精度的估算性能,其测试结果 R^2 可以达到 0.88.

参考文献:

- [1] 何映月,张松林.2000~2016 年中国 $\text{PM}_{2.5}$ 时空格局 [J]. 中国环境科学, 2020,40(8):3284-3293.
He Y Y, Zhang S L. Spatiotemporal pattern of $\text{PM}_{2.5}$ from 2000 to 2016 in China [J]. China Environmental Science, 2020,40(8):3284-3293.
- [2] 李明升,任晓霞,于 洋,等.中国大陆城市 $\text{PM}_{2.5}$ 污染时空分布规律 [J]. 中国环境科学, 2016,36(3):641-650.
Li M S, Ren X X, Yu Y, et al. Spatiotemporal pattern of ground-level fine particulate matter ($\text{PM}_{2.5}$) pollution in mainland China [J]. China Environmental Science, 2016,36(3):641-650.
- [3] Zhan Y, Luo Y, Deng X, et al. Spatiotemporal prediction of continuous daily $\text{PM}_{2.5}$ concentrations across China using a spatially explicit machine learning algorithm [J]. Atmospheric environment, 2017,155: 129-139.
- [4] 郑国威,王腾军,杨友森,等.基于遗传小波神经网络的 $\text{PM}_{2.5}$ 浓度预测模型 [J]. 测绘与空间地理信息, 2018,41(9):248-250.
Zheng G W, Wang T J, Yang Y S, et al. $\text{PM}_{2.5}$ concentration prediction model based on genetic wavelet neural network [J]. Geomatics & Spatial Information Technology, 2018,41(9):248-250.
- [5] 梁 泽,王玥瑶,岳远素,等.耦合遗传算法与 RBF 神经网络的 $\text{PM}_{2.5}$ 浓度预测模型 [J]. 中国环境科学, 2020,40(2):523-529.
Liang Z, Wang Y Y, Yue Y W, et al. A coupling model of genetic algorithm and RBF neural network for the prediction of $\text{PM}_{2.5}$ concentration [J]. China Environmental Science, 2020,40(2):523-529.
- [6] 康俊锋,黄烈星,张春艳,等.多机器学习模型下逐小时 $\text{PM}_{2.5}$ 预测及对比分析 [J]. 中国环境科学, 2020,40(5):1895-1905.
Kang J F, Huang L X, Zhang C Y, et al. Hourly $\text{PM}_{2.5}$ prediction and its comparative analysis under multi-machine learning model [J]. China Environmental Science, 2020,40(5):1895-1905.
- [7] 李 锋,朱 彬,安俊岭,等.2013 年 12 月初长江三角洲及周边地区重霾污染的数值模拟 [J]. 中国环境科学, 2015,35(7):1965-1974.
Li F, Zhu B, An J L, et al. Modeling study of a severe haze episode occurred over the Yangtze River Delta and its surrounding regions during early December, 2013 [J]. China Environmental Science, 2015, 35(7):1965-1974.
- [8] 李焱航,翟卫欣,颜寒祺,等.基于 U-net 神经网络模型的 $\text{PM}_{2.5}$ 逐小时浓度值预测模型 [J]. 北京大学学报(自然科学版), 2020,56(5): 796-804.
Li Y H, Zhai W X, Yan H Q, et al. Prediction of $\text{PM}_{2.5}$ hour concentration based on U-net neural network [J]. Acta Scientiarum Naturalium Universitatis Pekinensis, 2020,56(5):796-804.
- [9] 付宏臣,孙艳玲,王 斌,等.基于 AOD 数据和 GWR 模型估算京津冀地区 $\text{PM}_{2.5}$ 浓度 [J]. 中国环境科学, 2019,39(11):4530-4537.
Fu H C, Sun Y L, Wang B, et al. Estimation of $\text{PM}_{2.5}$ concentration in Beijing-Tianjin-Hebei region based on AOD data and GWR model [J]. China Environmental Science, 2019,39(11):4530-4537.
- [10] Wu J, Yao F, Li W, et al. VIIRS-based remote sensing estimation of ground-level $\text{PM}_{2.5}$ concentrations in Beijing-Tianjin-Hebei: A spatiotemporal statistical model [J]. Remote Sensing of Environment, 2016,184:316-328.
- [11] 周 亮,周成虎,杨 帆,等.2000~2011 年中国 $\text{PM}_{2.5}$ 时空演化特征及驱动因素解析 [J]. 地理学报, 2017,72(11):2079-2092.
Zhou L, Zhou C H, Yang F, et al. Spatio-temporal evolution and the influencing factors of $\text{PM}_{2.5}$ in China between 2000 and 2011 [J]. Acta Geographica Sinica, 2017,72(11):2079-2092.
- [12] 陶金花,张美根,陈良富,等.一种基于卫星遥感 AOT 估算近地面颗粒物的方法 [J]. 中国科学:地球科学, 2013,43(1):143-154.
Tao J H, Zhang M G, Chen L F, et al. Method to estimate concentration of surface-level particulate matter from satellite-based aerosol optical thickness [J]. Science China: Earth Sciences, 2013,43(1):143-154.
- [13] 李成才,毛节泰,刘启汉,等.MODIS 卫星遥感气溶胶产品在北京市大气污染研究中的应用 [J]. 中国科学:D 辑, 2005,35(A01):177-186.
Li C C, Mao J T, Liu Q H, et al. Application of MODIS remote sensing aerosol products in Beijing air pollution research [J]. Science in China Ser. D Earth Sciences, 2005,35(A01):177-186.
- [14] 吴健生,王 茜.基于 AOD 数据反演地面 $\text{PM}_{2.5}$ 浓度研究进展 [J]. 环境科学与技术, 2017,40(8):68-76.
Wu J S, Wang X. Research progress of retrieval ground-level $\text{PM}_{2.5}$ concentration based on AOD data [J]. Environmental Science & Technology, 2017,40(8):68-76.
- [15] Xie Y, Wang Y, Zhang K, et al. Daily estimation of ground-level $\text{PM}_{2.5}$ concentrations over Beijing using 3km resolution MODIS AOD [J]. Environmental Science & Technology, 2015,49(20):12280-12288.
- [16] Paciorek C J, Liu Y, Moreno-Macias H, et al. Spatiotemporal associations between GOES aerosol optical depth retrievals and ground-level $\text{PM}_{2.5}$ [J]. Environmental Science & Technology, 2008, 42(15):5800-5806.
- [17] Yang F, Wang Y, Tao J, et al. Preliminary investigation of a new AHI aerosol optical depth (AOD) retrieval algorithm and evaluation with multiple source AOD measurements in China [J]. Remote Sensing, 2018,10(5):748.
- [18] 王 琴,张大伟,刘保献,等.基于 PMF 模型的北京市 $\text{PM}_{2.5}$ 来源的时

- 空分布特征 [J]. 中国环境科学, 2015,35(10):2917-2924.
- Wang Q, Zhang D W, Liu B X, et al. Spatial and temporal variations of ambient PM_{2.5} source contributions using positive matrix factorization [J]. China Environmental Science, 2015,35(10):2917-2924.
- [19] 周 丽,徐祥德,丁国安,等.北京地区气溶胶 PM_{2.5} 粒子浓度的相关因子及其估算模型 [J]. 气象学报, 2003,61(6):761-768.
- Zhou L, Xu D, Ding G A, et al. The correlation factors and pollution forecast model for PM_{2.5} concentration in Beijing area [J]. Acta Meteorologica Sinica, 2003,61(6):761-768.
- [20] 贺 祥,林振山,刘会玉,等.基于灰色关联模型对江苏省 PM_{2.5} 浓度影响因素的分析 [J]. 地理学报, 2016,71(7):1119-1129.
- He X, Lin S, Liu H Y, et al. Analysis of the driving factors of PM_{2.5} in Jiangsu province based on grey correlation model [J]. Acta Geographica Sinica, 2016,71(7):1119-1129.
- [21] 王 敏,邹 滨,郭 宇,等.基于 BP 神经网络的城市 PM_{2.5} 浓度空间预测 [J]. 环境污染与防治, 2013,35(9):63-66.
- Wang M, Zou B, Guo Y, BP artificial neural network-based analysis of spatial variability of urban PM_{2.5} concentration [J]. Environment Pollution & Control, 2013,35(9):63-66.
- [22] 宋国君,国满丹,杨 啸,等.沈阳市 PM_{2.5} 浓度 ARIMA-SVM 组合预测研究 [J]. 中国环境科学, 2018,38(11):4031-4039.
- Song G J, Guo X D, Yang X, et al. ARIMA-SVM combination prediction of PM_{2.5} concentration in Shenyang [J]. China Environmental Science, 2018,38(11):4031-4039.
- [23] Chu D A, Kaufman Y, Zibordi G, et al. Global monitoring of air pollution over land from the Earth Observing System-Terra Moderate Resolution Imaging Spectroradiometer (MODIS) [J]. Journal of Geophysical Research: Atmospheres, 2003,108(D21).
- [24] Gupta P, Christopher S A. Particulate matter air quality assessment using integrated surface, satellite, and meteorological products: 2. A neural network approach [J]. Journal of Geophysical Research: Atmospheres, 2009,114(D20).
- [25] 孙 成,王 卫,刘方田,等.基于线性混合效应模型的河北省 PM_{2.5} 浓度时空变化模型研究 [J]. 环境科学研究, 2019,32(9):1500-1509.
- Sun C, Wang W, Liu F T, et al. Spatial-temporal simulation of PM_{2.5} concentration in Hebei Province based on linear mixed effects model [J]. Research of Environmental Sciences, 2019,32(9):1500-1509.
- [26] Ma Z, Liu Y, Zhao Q, et al. Satellite-derived high resolution PM_{2.5} concentrations in Yangtze River Delta Region of China using improved linear mixed effects model [J]. Atmospheric Environment, 2016,133:156-164.
- [27] Kloog I, Koutrakis P, Coull B A, et al. Assessing temporally and spatially resolved PM_{2.5} exposures for epidemiological studies using satellite aerosol optical depth measurements [J]. Atmospheric Environment, 2011,45(35):6267-6275.
- [28] 李 娟,尉 鹏,戴学之,等.基于机器学习方法的西安市数值模拟优化研究 [J]. 环境科学研究, 2021,34(4):872-881.
- Li J, Wei P, Dai X Z, et al. Optimization of numerical simulation in Xi'an based on machine learning methods [J]. Research of Environmental Sciences, 2021,34(4):872-881.
- [29] Liu B-C, Binaykia A, Chang P-C, et al. Urban air quality forecasting based on multi-dimensional collaborative Support Vector Regression (SVR): A case study of Beijing-Tianjin-Shijiazhuang [J]. PloS one, 2017,12(7):1-17.
- [30] Hu X, Belle J H, Meng X, et al. Estimating PM_{2.5} concentrations in the conterminous United States using the random forest approach [J]. Environmental Science & Technology, 2017,51(12):6936-6944.
- [31] Yu W, Liu Y, Ma Z, et al. Improving satellite-based PM_{2.5} estimates in China using Gaussian processes modeling in a Bayesian hierarchical setting [J]. Scientific reports, 2017,7(1):1-9.
- [32] Li T, Shen H, Zeng C, et al. Point-surface fusion of station measurements and satellite observations for mapping PM_{2.5} distribution in China: Methods and assessment [J]. Atmospheric Environment, 2017,152:477-489.
- [33] Hu X, Waller L A, Al-Hamdan M Z, et al. Estimating ground-level PM_{2.5} concentrations in the southeastern US using geographically weighted regression [J]. Environmental research, 2013,121:1-10.
- [34] Bengio Y. Learning deep architectures for AI [M]. Now Publishers Inc, 2009.
- [35] 郑 毅,朱成璋.基于深度信念网络的 PM_{2.5} 预测 [J]. 山东大学学报(工学版), 2014,44(6):19-25.
- Zheng Y, Zhu C Z. A prediction method of atmospheric PM_{2.5} based on DBNs [J]. Journal of Shandong University (Engineering Science), 2014,44(6):19-25.
- [36] 马井会,曹 钰,余钟奇,等.深度学习方法在上海市 PM_{2.5} 浓度预报中的应用 [J]. 中国环境科学, 2020,40(2):530-538.
- Ma J H, Cao Y, Yu Z Q, et al. The application of deep learning method in Shanghai PM_{2.5} prediction [J]. China Environmental Science, 2020,40(2):530-538.
- [37] Ian Goodfellow, Yoshua Bengio, Aaron Courville. Deep learning [M]. The MIT Press, 2016.

作者简介: 耿 冰(1987-),女,吉林长春人,博士,主要从事国土规划和信息技术研究.发表论文 10 余篇.