北京大学学报(自然科学版) 第 56 卷 第 5 期 2020 年 9 月
Acta Scientiarum Naturalium Universitatis Pekinensis, Vol. 56, No. 5 (Sept. 2020)
doi: 10.13209/j.0479-8023.2020.065

基于U-net神经网络模型的$PM_{2.5}$逐小时浓度值预测模型

李燚航[1] 翟卫欣[2,3,†] 颜寒祺[3] 朱道也[3] 童晓冲[4] 程承旗[5]

1. 北京大学城市与环境学院, 北京 100871; 2. 中国农业大学信息与电气工程学院, 北京 100083; 3. 北京大学前沿交叉学科研究院, 北京 100871; 4. 信息工程大学地理空间信息学院, 郑州 450052; 5. 北京大学工学院空天信息工程研究中心, 北京 100871;
† 通信作者, E-mail: pkuzhaiweixin@gmail.com

摘要 针对目前多数 $PM_{2.5}$ 预测模型泛化能力较差的问题, 提出基于 U-net 神经网络模型的 $PM_{2.5}$ 逐小时浓度值预测模型。该模型通过引入历史风场数据, 将离散的监测站点 $PM_{2.5}$ 浓度值插值为 $PM_{2.5}$ 网格图; 然后将 U-net 神经网络作为预测模型, 基于实验区域的 10 小时内的 $PM_{2.5}$ 网格图, 预测下一时刻的 $PM_{2.5}$ 网格图。该模型可以利用历史不同时刻提取的 $PM_{2.5}$ 浓度值网格图, 在预测区域内所有位置 $PM_{2.5}$ 浓度值的同时, 还可以提升预测的准确性以及对 $PM_{2.5}$ 浓度值突变情况的适应性。实验结果表明, 所提方法在 $PM_{2.5}$ 浓度值短时间突变情况下, 预测精度比传统方法有 10%左右的提升。

关键词 $PM_{2.5}$ 预测; 突变; 基于历史风速插值; 网格图; 神经网络

Prediction of $PM_{2.5}$ Hour Concentration Based on U-net Neural Network

LI Yihang[1], ZHAI Weixin[2,3,†], YAN Hanqi[3], ZHU Daoye[3], TONG Xiaochong[4], CHENG Chengqi[5]

1. College of Urban and Environmental Sciences, Peking University, Beijing 100871; 2. College of Information and Electrical Engineering, China Agricultural University, Beijing 100083; 3. Academy for Advanced Interdisciplinary Studies, Peking University, Beijing 100871; 4. Institute of Geospatial Information, Information Engineering University, Zhengzhou 450052; 5. Aerospace Information Engineering Research Center, Peking University, Beijing 100871;
† Corresponding author, E-mail: pkuzhaiweixin@gmail.com

Abstract Most of the previous $PM_{2.5}$ prediction models present unsatisfactory performance in several aspects, including predicting accuracy and generalization ability, especially in case of the sudden change in the value of $PM_{2.5}$ situation. Therefore, we propose a method based on the U-net neural network to predict the hourly $PM_{2.5}$ concentration value on the research area, attempting to improve the prediction performance. The proposed model includes two major steps. First, based on the inverse distance interpolation of historical wind field data, discrete station $PM_{2.5}$ values are interpolated into a $PM_{2.5}$ grid map; second, the U-net neural network is applied to train the prepared spatiotemporal grid data and make predictions. The model can use the $PM_{2.5}$ concentration values of the grid map extracted at different time stamps for the $PM_{2.5}$ prediction. The $PM_{2.5}$ concentration values at all locations in the research region can be achieved. Specifically, the prediction accuracy and the generalization ability of the model in case of sudden changes are revealed. Experimental results indicate that the proposed method has a 10% improvement in the prediction accuracy of $PM_{2.5}$ concentration values in the case of sudden change.

Key words $PM_{2.5}$ prediction; abrupt scenarios; interpolation of historical wind speed; grid graph; neural network

近年来, 大规模雾霾侵袭严重影响人们的生产生活和交通出行[1], 尤其对居民的身体健康造成较大伤害。$PM_{2.5}$(直径小于或等于 2.5 μm 的可吸入颗粒物), 是雾霾的主要成分[1]。对 $PM_{2.5}$ 的准确预测有利于提醒居民合理地安排出行及提前采取防护措施。因此如何快速、准确地预测雾霾的空间分布成

国家重点研发计划项目(2018YFB0505300, 2017YFB0503703)、广西科技重大专项项目(桂科AA18118025)、国防科技创新特区项目和中国博士后科学基金(2020M670024)资助
收稿日期: 2019–09–11; 修回日期: 2020–03–16

为研究的热点。

$PM_{2.5}$ 浓度值预测模型大致分为 3 类。

1) 大气扩散机理模型。Gibson 等[2]以 $PM_{2.5}$ 大气扩散机制为基础, 运用高斯烟羽扩散模型估算浓度值; 施加松等[3]研究一种新型的基于立体网格的放射性污染物扩散表达模型, 并以立体元胞机的方式, 对扩散过程进行有效的表达。

2) 遥感和 GIS 方法模型。利用加权回归和线性混合等方法来量化气溶胶光学厚度与 $PM_{2.5}$ 之间的相关关系, 用以反应大气污染状况[4–9]。由于使用的模型参数有温度、湿度、风向、季节等气象因子以及边界层高度和土地利用数据, 导致估算模型受地域和时间的限制, 需要在具体预测中选择适合研究区的模型参数来达到更高的估算精度[10–12]。

3) 机器学习和深度学习模型。Yin 等[13]运用 ARIMA 回归模型(一种统计学习方法), 将 PM_{10} 和风速、风向作为自变量, 预测北京的 $PM_{2.5}$ 日浓度值。Tang 等[14]基于北京的 $PM_{2.5}$ 浓度值时间序列数据, 提出图模型的方法, 以每个 $PM_{2.5}$ 监测站点为节点, 站点之间相互影响的权重为边, 基于构造的图模型, 使用时序模型 LSTM 模型来预测。Yi 等[15]将地理相关性和深度学习相结合, 将预测的站点周围划分成扇形网格, 根据站点的风向计算不同扇形网格区域 $PM_{2.5}$ 的浓度值, 作为站点周围的 $PM_{2.5}$ 信息, 以此预测该站点的 $PM_{2.5}$ 浓度值。深度学习的模型主要基于卷积神经网络模型和递归神经网络模型。Vahdatpour 等[16]利用拍摄的天空照片作为输入, 预测当地当时的空气污染等级(优、良、轻度污染、中等污染和重度污染)。他们的模型主要利用 3 层带有 5 × 5 卷积核的卷积层来捕捉图像特征, 实验证明了卷积神经网络模型的效果优于用小波变换提取特征后随机森林分类的效果。Tong 等[17]考虑到空气污染物浓度变化的时序性, 引入递归神经网络来预测未来时刻的污染物浓度。除预测下一时刻的污染物浓度外, 他们发挥 LSTM 对于较长时间序列的预测优势, 将以往单一时刻的预测拓展到 5~10 小时的较长时序预测。

大气扩散机理模型的预测方法可以保证预测的准确性和可解释性, 但是依赖于非常多的输入、复杂的传播理论和公式, 计算复杂度高、效率低, 且扩散模型适合点污染源的扩散, 对于区域内的相互扩散并不适用。机器学习的图模型方法考虑到站点之间关联性对预测结果的影响, 但仅以站点之间的距离作为站点间的关联, 没有考虑风对站点之间污染物扩散的影响。扇形网格的方法以站点为圆心, 把站点周围的 $PM_{2.5}$ 浓度值作为输入, 预测该站点的 $PM_{2.5}$ 浓度值。此方法考虑到周围环境的污染物信息, 但由于其网格的尺度是单一的, 所以对大尺度的环境信息利用不够充分。利用遥感数据反演的预测模型可以弥补地基观测空间覆盖小、空间分布信息少的不足, 但近地遥感卫星很难实现连续长时序的观测, 且受天气和云层遮挡的影响较大, 难以获取连续不间断的遥感数据。

在预测某一区域当前时刻的 $PM_{2.5}$ 浓度值时, 应从时间和空间两个维度来考虑, 即过去一段时间的 $PM_{2.5}$ 的浓度值以及邻近空间的 $PM_{2.5}$ 的浓度值都会对该区域未来时刻的 $PM_{2.5}$ 浓度值预测起决定性作用。本文以北京市作为研究区域, 由于各个站点记录数据较为完整, 研究区域内多为商业或住宅, 各个子区域扩散情况的差异性较小, 适合作为研究案例。另外, 北京位于污染物排放较多的京津冀区域, 且三面环山, 扩散条件不好, 容易产生 $PM_{2.5}$ 浓度值发生突变的现象。传统的预测模型是基于空间离散点的时序模型(例如 ARIMA 统计模型或深度学习的 LSTM 模型), 这些模型在预测过程中无法有效地融合周围区域的信息对预测点 $PM_{2.5}$ 浓度值的影响, 而是过度依赖预测点的历史 $PM_{2.5}$ 浓度值, 导致当 $PM_{2.5}$ 浓度值发生突变时, 预测趋势往往会趋于平滑, 误差较大。使用遥感影像的方式可以弥补这一不足, 但是遥感影像的尺度过大且单一, 还要通过建立反射率与污染物浓度的模型来预测, 增加了预测的误差和不稳定因素。目前对缓变平滑的 $PM_{2.5}$ 浓度值预测精度已经很高, 但尚不能很好地解决实际问题, 因为人们并不关注长时序平滑的 $PM_{2.5}$ 浓度变化, 而 $PM_{2.5}$ 浓度值发生突然升高或降低的情况恰恰是人们关注的。对于周围站点信息的更合理高效的利用是提高 $PM_{2.5}$ 突变预测准确性的重点, 也一直是预测 $PM_{2.5}$ 的一个难点。

$PM_{2.5}$ 监测站点分布较为稀疏, 在北京 16411 km^2 范围内有 35 个, 平均 468 km^2 有一个。本文提出通过不同形式的插值方式, 将北京市城区内部的 27 个稀疏站点的 $PM_{2.5}$ 数据, 插值为一个无缝无叠的均匀网格覆盖的空间模型。该模型中的每一个网格都代表一个区域, 记录该区域的 $PM_{2.5}$ 浓度值, 每一个网格中的值表示该区域的 $PM_{2.5}$ 浓度值, 生成网格图, 摒弃每个站点单独预测的形式, 改为网格图

的形式来预测。有学者通过网格化的方式进行过实验，基于空间网格分析多尺度人文地理特征以及污染物空间分布的社会经济影响因素[18–19]，使用网格作为控制尺度的工具，从不同的尺度来认知人类社会活动的空间分布模式以及污染物的分布对社会经济的影响。因此，对于 $PM_{2.5}$ 的预测，改为预测一张网格图，即同时预测整个区域的 $PM_{2.5}$ 浓度值，这是因为相邻或相关站点的浓度值会互相影响。生成最接近真实情况的网格图是准确利用周围站点的关键，也是使用深度网络训练的基本保证。

1 数据生成

网格图是通过网格插值的形式，将非结构的 $PM_{2.5}$ 站点数据结构化的重要方式。网格图层分为结构化的网格和非结构化的网格。结构化的网格中，每一个网格具有相同数量和大小的相邻网格，非结构化的网格为不同尺度、不同形状的格网。结构化的网格划分较为简单，对于二维的平面，通过正方形型网格，无缝无叠铺排，将某一区域结构化为网格图层。本文选用结构化网格图生成 $PM_{2.5}$ 网格图。

1.1 高斯扩散模型

本文基于传统的大气扩散模型机理，首先将离散的 $PM_{2.5}$ 站点数据结构化，以便后续的预测。高斯扩散模型[20]是最常见的大气扩散模型，是依据平流扩散的微分方程来模拟污染物扩散过程。在风向、风速和平流的湍流扩散系数恒定的前提下，高斯扩散模型得到的解服从标准的正态分布，一般用来模拟理想条件下的大气中污染物的扩散和分布。标准的高斯扩散模型的表达式为

$$C(x,y,z,t)=\frac{Q\cdot e^{-(x-\mu t)^2/2\sigma_x^2}}{(2\pi)^{3/2}\sigma_x\sigma_y\sigma_z}\cdot e^{\frac{-(y-vt)^2}{2\sigma_y^2}}\cdot \left[e^{\frac{-(z-h-wt)^2}{2\sigma_z^2}}+e^{\frac{-(z+2H+h-wt)^2}{2\sigma_z^2}}\right], \quad (1)$$

其中，t 为时间(s)，Q 为污染释放率(mg/s)，u，v 和 w 为风速分解后沿着 3 个坐标方向的矢量值，σ_x，σ_y 和 σ_z 为沿着 3 个坐标方向的扩散系数，h 为污染源高度(m)，H 为混合层高(m)。

高斯扩散模型主要用来模拟单点扩散源在风向、风速以及扩散系数都恒定条件下的扩散过程。当不考虑污染物高程方向的变化以及风速在除风向方向外的两个方向为 0(即 v=w=0 且 z=0)时，方程化简为

$$C(x,y,z,t)=\frac{Q\cdot e^{-(x-\mu t)^2/2\sigma_x^2}}{(2\pi)^{3/2}\sigma_x\sigma_y\sigma_z}\cdot e^{\frac{-y^2}{2\sigma_y^2}}\left[e^{\frac{-h^2}{2\sigma_z^2}}+e^{\frac{-4H^2}{2\sigma_z^2}}\right]。 \quad (2)$$

根据式(2)，假设在不同站点之间的污染物扩散互相独立，且扩散过程中各个站点的风向和风速恒定，推算出不同的站点扩散到待插值位置的浓度值，求和得到该点浓度的估计值。由于风向和风速不停地变化，不同的污染源在扩散时具有较强的相关性。并且，这种计算方式十分复杂，适用于单点污染源的局部扩散过程，如果要用于 $PM_{2.5}$ 的插值，还需要改进。

1.2 常用插值方法

常用的空间插值方式包括最近邻插值、反距离加权插值、克里金插值以及基于以上 3 种方法发展的其他自定义插值方法。

最近邻插值(nearest monitor)[21]较为简单，将待插值点的值直接设为距离该点最近采样点的值。目前，主流的天气应用就是通过最近邻插值，得到当前位置的 $PM_{2.5}$ 浓度值。

反距离加权插值(IDW)法[22]的插值函数为

$$F(x,y)=\frac{\sum_{k=1}^{n}\frac{f_k}{(d_k(x,y))^2}}{\sum_{k=1}^{n}\frac{1}{(d_k(x,y))^2}}=\sum_{n}^{k=1}f_k\cdot w_k(x,y), \quad (3)$$

其中，$d_k(x, y)$ 表示第 k 个采样点与待插值点之间的距离这里使用的是欧式距离，(x, y) 表示待插值点处的坐标，f_k 表示采样点处的数值。权重函数的计算方式为

$$w_k(x,y)=\frac{\prod_{i=k}^{n}(d_i(x,y))^2}{\sum_{k=1}^{n}\prod_{i=k}^{n}(d_i(x,y))^2}, \quad (4)$$

$$d_i(x,y)=\sqrt{(x-x_i)^2+(y-y_i)^2}。 \quad (5)$$

克里金插值法[23]是较常用的空间插值的方法，广泛应用于大气科学和环境科学，目的是求得待插值点处最优的线性无偏估计，也称为空间插值最优无偏估计器。与简单的反距离加权方法相比，不仅能利用待插值点与采样点之间的空间距离，还会隐式地利用二者的空间位置以及待插值点与周围采样点之间的相对位置关系。

传统的插值方式中，度量插值得到的数据与真实数据的距离时，均使用欧式空间的距离，这个距

离对一般的地理要素是合理的，但对污染物的扩散不适用。污染物的扩散与一个区域以及周围区域的气象条件和地形地貌等扩散条件紧密相关，如 A 和 B 两个区域相邻很近，但是中间存在地形起伏或存在变化的风向和风速[24]，则会影响 A 与 B 之间的传播距离，且 A 到 B 以及 B 到 A 的传播距离不相等，并实时变化。

1.3 基于风向风速的插值方法

本文研究的是二维空间的污染物的扩散，所以选择二维结构化网格，每个网格有 8 个相邻的网格。如图 1 所示，网格 A 向其邻域网格扩散的方位角为 0°, 45°, 90°, 135°, 180°, 225°, 270°和315°。

以网格 A 向邻域网格 B 扩散为例，单个网格向周围网格扩散的公式为

$$\text{DIFF-DIS} = (Q(\text{A}) + Q(\text{B})) \cdot L_{\text{AB}}, \tag{6}$$

$$Q(\text{A}) = W(\text{A}_1) + W(\text{A}_2), \tag{7}$$

$$Q(\text{B}) = W(\text{B}_1) + W(\text{B}_2)。 \tag{8}$$

式(6)中的 DIFF-DIS 为在扩散过程中网格 A 与网格 B 的距离，即从 A 扩散到 B 的难易程度；Q(A)和 Q(B)分别表示网格 A 与网格 B 的历史风向和风速对 A 到 B 扩散距离的影响程度，L_{AB} 为网格 A 与网格 B 之间的距离，若二者为边相邻，则 L_{AB}=1；若二者为角相邻，则 $L_{\text{AB}} = \sqrt{2}$。Q 函数由两个部分组成，分别是网格 A，B 间一个小时前和两个小时前的风向和风速的影响程度。相邻网格位于下风向时，历史风速越大，则两个网格的距离越小；位于上风向时，历史风速越大，二者的距离越大。风向和风速对扩散的影响如式(9)所示。

$$W(\text{A}_i) = 1 \cdot \text{e}^{-\left|F_{\text{A}_i} \cdot \cos\left(R\left(D_{\text{A}_i}, D_{\text{AB}}\right)\right)\right|^2 \cdot -\text{sgn}\left(\cos\left(R\left(D_{\text{A}_i}, D_{\text{AB}}\right)\right)\right)}, \tag{9}$$

$$R(D_{\text{A}_i}, D_{\text{AB}}) = |(180 - |D_{\text{A}_i} - D_{\text{AB}}|)|。 \tag{10}$$

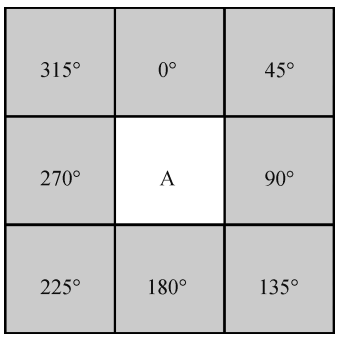

图 1　正方形网格示意图
Fig. 1　Square grid schematic

式(9)中，i 表示 i 小时前；1 表示从 i 到 $i-1$ 时刻影响延续一个小时；F_{A_i} 为网格 A 在 i 小时前的风速；D_{A_i} 为网格 A 在 i 小时前的风向的方位角，D_{AB} 为网格 A 指向网格 B 的方位角；$R(D_{\text{A}_i}, D_{\text{AB}})$表示某一历史时刻，网格 A 的风向与网格 A 到某一邻域网格传播方向的夹角；$|F_{\text{A}_i} \cdot \cos(R(D_{\text{A}_i}, D_{\text{AB}}))|$ 为某一时刻网格 A 的风速沿 A→B 扩散方向分量的模长，代表这一时刻的风速对扩散过程的影响程度。式(9)的具体形式参考了高斯烟羽扩散模型中风速的影响模式，根据扩散模型来计算风速对相邻网格内污染物扩散的影响。当风向与相邻网格的扩散方向夹角的余弦值为正(即网格 B 在网格 A 的下风方向)时，网格 A 的风速越大，从 A 向 B 的传播越容易，其扩散距离就越小；当风向与相邻网格的扩散方向夹角的余弦值为负数(即网格 B 在网格 A 的上风方向)时，网格 A 的风速越大，从 A 向 B 的传播越快，其扩散距离越小。

至此，构建出网格图中网格之间 $PM_{2.5}$ 扩散的图模型，网格与邻接网格之间的扩散距离构成邻接矩阵。基于历史风场的反距离加权插值时，使用的距离为图模型中的最短距离。计算图模型中的最短距离时，本文使用启发式 A*搜索算法。A*搜索算法是一种在图形平面上，从多个节点的路径中，求出通过成本最低的算法，计算待插值网格到采样点网格的最短 $PM_{2.5}$ 扩散距离。如图 2 所示，左上角的网格为某个 $PM_{2.5}$ 监测站点，右下角的网格为没有 $PM_{2.5}$ 监测站点而待插值的网格，两个网格之间的最短距离用粗线表示，这个距离既可以被视为基于历史风速下的两个网格的最短距离，同时也可以被视为 $PM_{2.5}$ 在两个网格之间传播经过的路径。图 3 给出基于网格图预测 $PM_{2.5}$ 浓度值的流程。

2 模型训练

以图的方式同时预测整个区域的前提是，把稀疏的不等间距排列的站点数据插值成为稠密的等间距的网格数据，插值过的结构化数据才能作为神经网络的输出。深度学习中计算机视觉的方法在图片预测中应用广泛，预测整个区域等价于整幅图片的预测，可以借鉴计算机视觉的方法来做网格图的预测。数据的插值方式也是对周围空间数据有效利用以及预测准确性的保障。

深度学习中的 CNN (convolutional neural networks)网络是计算机视觉[25]、图像处理领域最常用

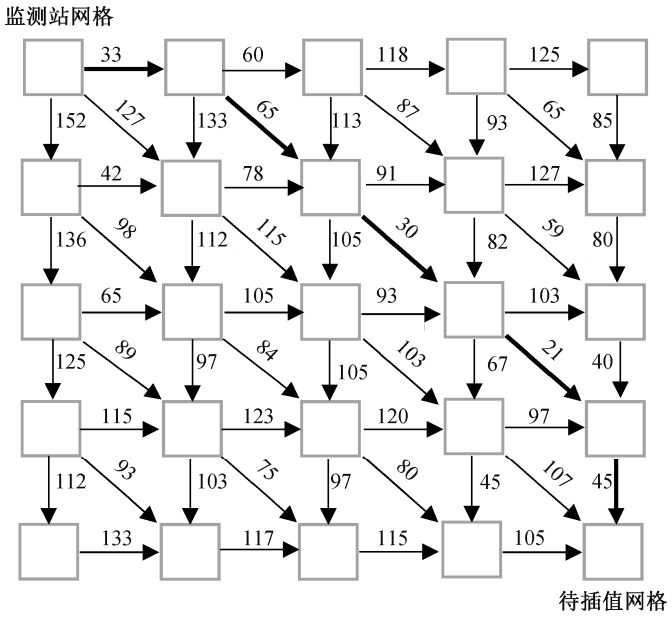

图 2 某一时刻待插值网格与监测站网格间的最短距离

Fig. 2 Shortest distance between interpolated grid and monitoring station grid

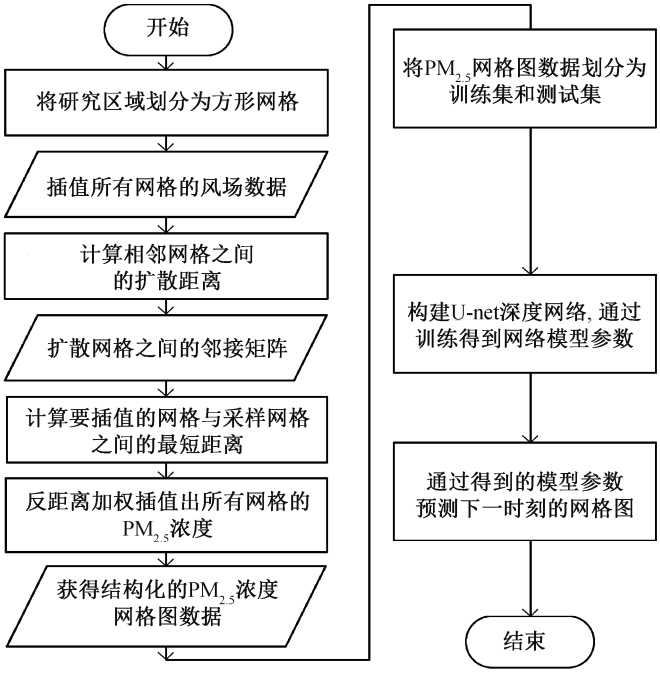

图 3 预测 $PM_{2.5}$ 浓度流程

Fig. 3 Prediction flow chart of $PM_{2.5}$

的一种神经网络模型, 在提取结构化的空间信息时有非常突出的表现, 在图像纹理检测和语义分割等任务中扮演重要的角色。由于地理空间区域的预测与图像区域空间的预测本质上是同构的, 只是存在尺度的差异, 因此可以借鉴计算机视觉方法进行网格图预测。与以往单个站点的预测方式不同, 基于结构化的网格图的预测不是各个站点独立地预测下一时刻的 $PM_{2.5}$ 浓度, 而是以网格图的形式同时预测下一时刻整个研究区域所有网格的浓度值。

如图 4 所标, 本文选取 U-net 网络[26]模型作为

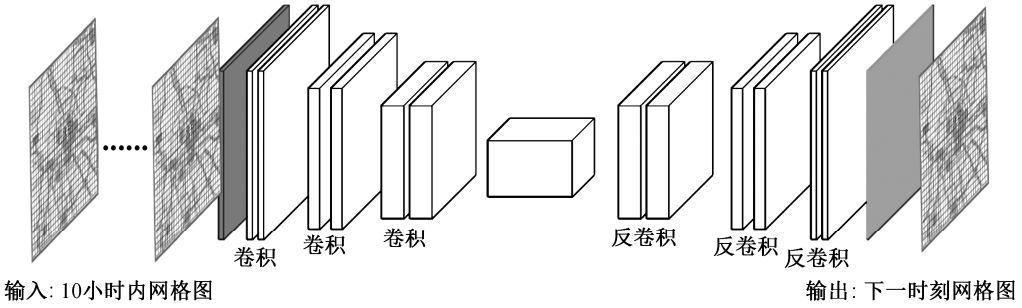

图 4　用于 $PM_{2.5}$ 网格图预测的 U-net 神经网络

Fig. 4　U-net Neural Network for Prediction of $PM_{2.5}$ Grid Graph

基础模型, 利用全卷积进行特征提取, 即利用卷积层和池化层进行特征提取, 再利用反卷积层还原图像尺寸。模型由压缩通道(contracting path)和扩展通道(expansive path)组成。压缩通道用于不同尺度的特征提取, 扩展通道用于基于提取的特征扩展成输入数据的维度进行预测。U-net 共进行多次上采样, 并在同一层级的特征图中使用跨层连接, 而不是直接在高级语义特征上进行监督和损失的反向传播, 从而保证原始输入的信息可以更少损失地传到预测的网络模型[27], 恢复得到的特征图融合更多的低层次高分辨率的特征, 也使得不同尺度的特征得到融合, 提高预测的准确性[28]。

3 数据与结果分析

3.1 数据介绍

本文采用的实验数据为来自北京市环境监测保护中心的北京市空气污染物 $PM_{2.5}$ 浓度值的历史记录数据, 共收集到北京市内共 35 个站点从 2015 年 1 月 1 日至 2017 年 12 月 31 日的 $PM_{2.5}$ 浓度值历史观测数据, 观测数据的时间间隔为 1 小时。

为保证得到的网格图数据质量, 本文选择分布较为集中的北京市六环以内的 27 个 $PM_{2.5}$ 监测站点(116.104°—116.655°E, 39.618°—40.127°N)。以 27 个监测站点围成的最小外包矩形为研究区域。以 0.02°×0.02°经纬度网格紧密铺排的结构化网格划分方式, 将矩形区域划分成结构化的 28×28 的规则网格。对于每一个时刻的监测数据, 都可以通过本文提出的插值方式进行划分, 从而得到所有网格的浓度值。

经过分析北京市 3 年的历史风场数据, 得到研究区域历史风速的中位数为 0.6179 m/s, 本文主要考虑两个小时的历史风场对污染物扩散的影响, 按照该风速, 一小时扩散的距离为 2224.44 m。在北京市区所处的区域中, 0.01°经线长约为 1100 m, 每小时扩散的距离大约为 0.02°的经线长, 故本文选择 $0.02° \times 0.02°$ 的经纬度网格划分方式。

网格内的浓度值与历史风速强相关, 如果划分的尺度小于 $0.02° \times 0.02°$, 则每小时的扩散距离大于相邻网格的距离, 会导致构建扩散网格图时出现错误; 如果采用更大的尺度来划分, 则有可能导致一小时后并没有扩散到邻域的网格中, 网格图的状态没有变化, 致使计算资源的浪费, 也可能造成距离较近的两个站点被划分到同一个网格内, 造成同一网格上有两份数据的问题。

3.2 预测方法的对比

对 27 个站点从 2015 年 1 月 1 日到 2017 年 12 月 31 日之间每小时的监测数据, 采取基于历史风速的插值, 每一个时刻得到一个 28×28 的矩阵, 矩阵中的每一个值表示对应网格处的 $PM_{2.5}$ 浓度值。对矩阵中按照时间顺序排列的数据进行划分, 前 80%的数据划分为训练集, 用于训练, 后 15%划分为验证集, 用于对模型的交叉验证, 最后 5%为测试集, 用于测试。使用自回归移动平均模型[29]和长短时循环神经网络[30]两类时序模型进行预测。

自回归移动平均模型常用于进行时间序列的预测。其原理是在将非平稳时间序列转化为平稳时间序列的过程中, 将因变量仅对其滞后值及随机误差项的现值和滞后值进行回归。长短时循环神经网络则利用过去一段时间内某事件的特征来预测未来一段时间内该事件的特征。与回归分析模型的预测不同, 时间序列模型更依赖事件发生的先后顺序。

基于 U-net 神经网络(使用基于历史风场数据的插值)、ARIMA 方法和 LSTM 方法得出的 9 个站点在 2017 年 12 月 10 日 $PM_{2.5}$ 浓度值发生突变的情况下, 预测结果与真值的对比, 情况如图 5 所示。

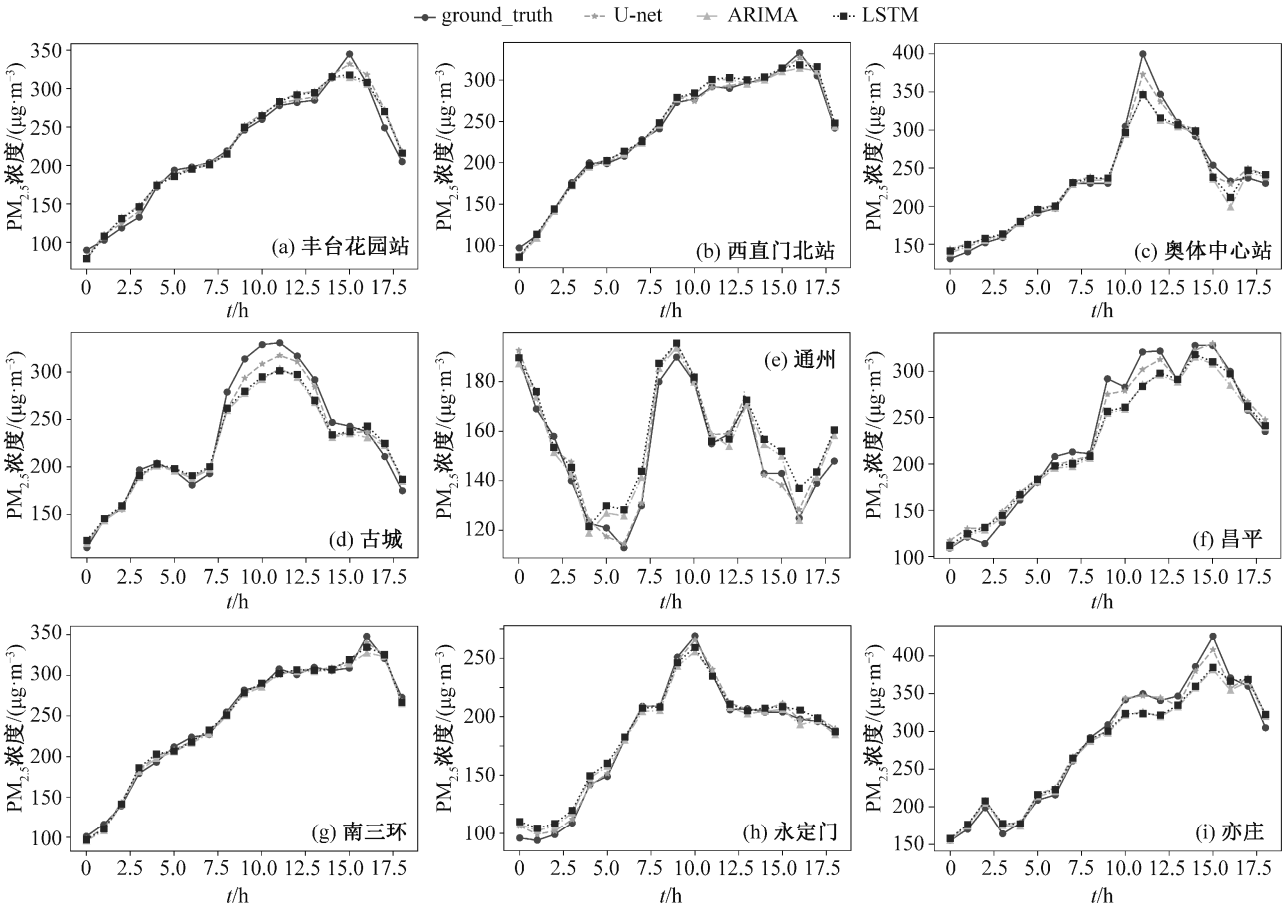

图 5 对 9 个站点未来 18 小时的预测

Fig. 5 Prediction of the next 18 hours for nine sites

3.3 结果分析

从图 5 可以看出，当 $PM_{2.5}$ 浓度值在短时间内突然升高时，基于单点预测的 ARIMA 和 LSTM 方法对 $PM_{2.5}$ 浓度值最高点的预测都出现很大的偏差，而基于历史风速插值的预测方法可以较好地预测 $PM_{2.5}$ 浓度值的走势，对 $PM_{2.5}$ 浓度峰值的预测更加准确。

测试集包括两个：测试集 1 具有较多的 $PM_{2.5}$ 浓度值突变，测试集 2 没有 $PM_{2.5}$ 浓度值突变。若当天 $PM_{2.5}$ 浓度值的标准差大于 50，则定义为有 $PM_{2.5}$ 浓度值突变的情况，反之则定义为没有 $PM_{2.5}$ 浓度值突变的情况。设置两个对照的数据集，可以对比基于网格图的方法对突变情况的预测优势。各个模型在两个测试集上的 RMSE(均方根误差)和 MAE(平均绝对误差)如表 1 所示。U-net (nearest)为使用最近邻插值得到 $PM_{2.5}$ 浓度值的网格图方法，U-net (based on wind field)为使用基于历史风速得到 $PM_{2.5}$ 浓度值的网格图方法。可以看出，在 $PM_{2.5}$ 浓度值的走势平滑、规律性较强的情况下，不同方法的预测结果相差不大，基于网格图的预测方法稍有改善。当测试集中存在多天的 $PM_{2.5}$ 浓度值突然变

表 1 不同方法在测试集上的预测结果

Table1 Prediction results of different methods in the test sets

测试集	方法	RMSE	MAE
测试集 1 (有突变)	ARIMA	18.8122	18.3714
	LSTM	14.3935	13.6171
	U-net (nearest)	12.2778	11.3251
	U-net (based on wind field)	10.5568	10.4123
测试集 2 (无突变)	ARIMA	12.1728	11.8917
	LSTM	11.2653	11.1746
	U-net (nearest)	11.0385	10.9535
	U-net (based on wind field)	9.6332	9.3172

化的情况时，传统方法的预测效果较差，基于历史风速插值的网格图预测方法准确率有约10%的提升。基于最近邻插值方式得到的预测结果误差明显大于基于历史风速的插值方法，说明对周围信息的利用是预测效果提升的主要因素。

4 结语

本文考虑历史风场数据，将离散的 $PM_{2.5}$ 浓度值插值为 $PM_{2.5}$ 网格图，使用 U-net 神经网络，输入实验区域 10 小时内的 $PM_{2.5}$ 网格图，预测下一时刻的 $PM_{2.5}$ 网格图，将图像预测的方法应用在 $PM_{2.5}$ 预测中。该方法避免了以点的形式预测导致无法有效利用该点周围空间信息的缺点，提出构建 $PM_{2.5}$ 的扩散网格图时采取直接预测 $PM_{2.5}$ 网格图，不仅可以利用时间轴上的 $PM_{2.5}$ 关联信息来预测，还可以利用时间轴上不同时刻提取的周围 $PM_{2.5}$ 的关联信息，保证预测结果的准确性和对突变情况的适应性。通过实验与传统的不利用扩散网格图的方法进行对比，分析突变情况下的预测准确率，证明基于 $PM_{2.5}$ 扩散的网格图预测优于独立利用站点历史数据的预测方法，能够保证预测结果的准确性以及对 $PM_{2.5}$ 浓度值突变情况的适应性，为 $PM_{2.5}$ 的预测提供了新的思路和方法。

参考文献

[1] 吕喆, 魏巍, 周颖, 等. 2015—2016 年北京市3次空气重污染红色预警 $PM_{2.5}$ 成因分析及效果评估. 环境科学, 2019, 40(1): 1–10

[2] Gibson M D, Kundu S, Satish M. Dispersion model evaluation of $PM_{2.5}$, NO_x, and SO_2, from point and major line sources in Nova Scotia, Canada using AERMOD Gaussian plume air dispersion model. Atmospheric Pollution Research, 2013, 4(2): 157–167

[3] 施加松, 余接情, 常芸芬, 等. 基于立体网格的放射性污染物扩散过程模拟与表达. 地理信息世界, 2019, 26(2): 52–59

[4] Kloog I, Nordio F, Coull B A, et al. Incorporating local land use regression and satellite aerosol optical depth in a hybrid model of spatiotemporal $PM_{2.5}$ exposures in the Mid-Atlantic States. Environmental Science & Technology, 2012, 46(21): 11913–11921

[5] Li X, Zhang C, Li W, Evaluating the use of DMSP/OLS nighttime light imagery in predicting $PM_{2.5}$ concentrations in the Northeastern United States. Remote Sensing, 2017, 9(6): 620–613

[6] He Q, Huang B. Satellite-based mapping of daily high-resolution ground $PM_{2.5}$ in China via space-time regression modeling. Remote Sensing of Environment, 2018, 206: 72–83

[7] 李佳霖, 樊子德, 邓敏. 顾及风向和风速的空气污染物浓度值插值方法. 地球信息科学学报, 2017, 19(3): 382–389

[8] Li Longxiang, Gong Jianhua, Zhou Jieping, et al. Spatial interpolation of fine particulate matter concentrations using the shortest wind-field path distance. PLOS ONE, 2014, 9(5): e96111

[9] Chen L, Shi M, Li S, et al. Quantifying public health benefits of environmental strategy of $PM_{2.5}$ air quality management in Beijing-Tianjin-Hebei region, China. Journal of Environmental Sciences, 2017, 57(7): 33–40

[10] 卢德彬, 毛婉柳, 杨东阳. 基于多源遥感数据的中国 $PM_{2.5}$ 变化趋势与影响因素分析. 长江流域资源与环境, 2019, 28(3): 651–660

[11] 薛文博, 武卫玲, 许艳玲, 等. 基于 WRF 模型与气溶胶光学厚度的 $PM_{2.5}$ 近地面浓度卫星反演. 环境科学研究, 2016, 29(12): 1751–1758

[12] 董佳丹, 陈晓玲, 孙昆, 等. 利用 MODIS 资料监测湖北省 $PM_{2.5}$ 的 3 种模型对比. 测绘科学, 2019, 44(10): 35–42

[13] Yin Q, Wang J, Hu M, et al. Estimation of daily $PM_{2.5}$ concentration and its relationship with meteorological conditions in Beijing. Journal of Environmental Sciences, 2016, 48(10): 161–168

[14] Tang P, Wang H, Kwong S. Deep sequential fusion LSTM network for image description. Neurocomputing, 2018, 312: 154–164

[15] Yi Xiuwen, Zhang Junbo, Wang Zhaoyuan, et al. Deep Distributed Fusion Network for Air Quality Prediction // 24th ACM SIGKDD International Conference. London, 2018: 965–973

[16] Vahdatpour M S, Sajedi H, Ramezani F. Air pollution forecasting from sky images with shallow and deep classifiers. Earth Science Informatics, 2018, 11: 413–422

[17] Tong W, Li L, Zhou X, et al. Deep learning $PM_{2.5}$ concentrations with bidirectional LSTM RNN. Air Quality, Atmosphere & Health, 2019, 12(4): 411–423

[18] 翟卫欣, 段杰雄, 童晓冲, 等. 基于空间网格的多尺度人文地理特征分析. 测绘学报, 2016, 45(S1):

85–89

[19] 段杰雄, 翟卫欣, 程承旗, 等. 中国 $PM_{2.5}$ 污染空间分布的社会经济影响因素分析. 环境科学, 2018, 39(5): 2498–2504

[20] Han L, Zhou W, Li W, et al. Impact of urbanization level on urban air quality: a case of fine particles ($PM_{2.5}$) in Chinese cities. Environmental Pollution, 2014, 194(1): 163–170

[21] Katja P, Christina D B, Stefan B, et al. Air quality and chronic stress: a representative study of air pollution ($PM_{2.5}$, PM_{10}) in Germany. Journal of Occupational and Environmental Medicine, 2019, 61(2): 144–147

[22] Heo J, Adams P J, Gao H O. Public health costs accounting of inorganic $PM_{2.5}$ pollution in metropolitan areas of the United States using a risk-based source-receptor model. Environment International, 2017, 106: 119–126

[23] Liu Zhao, Xie Meihui, Tian Kun, et al. GIS-based analysis of population exposure to $PM_{2.5}$ air pollution — a case study of Beijing. Journal of Environmental Sciences, 2017, 9(3): 35–47

[24] Bao Chengzhen, Chai Pengfei, Lin Hongbo, et al. Association of $PM_{2.5}$ pollution with the pattern of human activity: a case study of a developed city in eastern China. Journal of the Air & Waste Management Association, 2016, 66(12): 1202–1213

[25] Li Huan, Fan Hong, Mao Feiyue. A visualization approach to air pollution data exploration — a case study of air quality index ($PM_{2.5}$) in Beijing, China. Atmosphere, 2016, 7(3): 35–37

[26] Zhang P, Hong B, He L, et al. Temporal and spatial simulation of atmospheric pollutant $PM_{2.5}$ changes and risk assessment of population exposure to pollution using optimization algorithms of the back propagation-artificial neural network model and GIS. International Journal of Environmental Research and Public Health, 2015, 12(10): 12171–12195

[27] Denton E L, Chintala S, Fergus R, et al. Deep generative image models using a Laplacian Pyramid of adversarial networks // Neural Information Processing Systems. Montreal, 2015: 1486–1494

[28] Isola P, Zhu J Y, Zhou T, et al. Image-to-image translation with conditional adversarial networks // CVPR 2017. Hawaii, 2017: 2–5

[29] Zhang G P. Times series forecasting using a hybrid ARIMA and neural network model. Neurocomputing, 2003, 50: 159–175

[30] Shi X, Chen Z, Wang H, et al. Convolutional LSTM network: a machine learning approach for precipitation now casting // Neural Information Processing Systems. Montreal, 2015: 802–810