DOI:10.19322/j.cnki.issn.1006-4710.2021.01.003

基于网格划分的空间关联区域 VOCs 浓度预测研究

陆秋琴，兰　琼，黄光球

（西安建筑科技大学 管理学院，陕西 西安 710055）

摘要： 为了提高 VOCs 浓度预测的准确性，实现污染物的精细化定位管理及提高环境治理效率，首先运用点云网格算法对研究区域进行划分，由克里金插值法预估出未设置监测点的网格数据，收集网格监测数据和预估数据。然后建立基于随机森林算法的 VOCs 预测模型，采用 Bootstrap 法选取训练样本子集，通过构建样本子集的决策树，得到 VOCs 污染物浓度预测结果，并对模型性能进行评价。结果表明：基于空间关联性和污染物特征的随机森林模型预测更加精准，将其与 BP 神经网络预测模型进行比较，二者的样本训练及预测的平均误差分别为 3.15％和 13.36％，随机森林模型的平均误差更小，模型效率更高。因此以网格划分为前提，将克里金插值法与随机森林回归模型相结合的 VOCs 预测方法，能为区域污染物治理和预警提供依据。

关键词： VOCs；网格划分；克里金法；随机森林

中图分类号： X511　　**文献标志码：** A　　**文章编号：** 1006-4710(2021)01-0016-09

Research on prediction of VOCs concentration in spatial correlation region based on grid division

LU Qiuqin, LAN Qiong, HUANG Guangqiu

(School of Management, Xi'an University of Architecture and Technology, Xi'an 710055, China)

Abstract: In order to improve accuracy of VOCs concentration prediction and realize pollutants' fine positioning management and enhance environmental governance efficiency, the study area was firstly divided by the point cloud grid algorithm, using the Kriging interpolation method to evaluate grid data without monitoring points and collect grid monitoring data & predicated data. Then, the VOCs prediction model was established based on the random forest algorithm, and the Bootstrap method was used to select the training sample subset, with the results of VOCs concentration prediction obtained by establishing the decision-making tree of this subset and the model performance further evaluated. The results indicated that the random forest model prediction based on spatial association and pollutant characteristics was more accurate. Compared with that of BP neural network predictive model, the average errors of sample training and predication were 3.15% and 13.36% respectively, but the average error of random forest model was smaller with a higher model efficiency. Therefore, based on grid division, the VOCs prediction method combining the Kriging interpolation method with random forest regression model can provide basis for areas in regards with pollutant control and early warning.

Key words: VOCs; gridding; Kriging method; random forest

随着工业化的快速发展，区域性大气污染日益突出，VOCs 的大幅度排放引发了诸多环境问题。作为 $PM_{2.5}$ 和臭氧等大气污染物的重要前体物，VOCs 能发生光化学反应并生成有害的二次有机气

收稿日期： 2020-07-05；　**网络出版日期：** 2020-10-20

网络出版地址： https://kns.cnki.net/kcms/detail/61.1294.N.20201019.1604.004.html

基金项目： 国家自然科学基金资助项目(71874134)；陕西省自然科学基础研究计划重点资助项目(2019JZ-30)

第一作者： 陆秋琴，女，博士，教授，研究方向为污染排放控制与管理研究。E-mail:luqiuqin@xauat.edu.cn

通信作者： 兰琼，女，硕士生，研究方向为污染排放控制与管理研究。E-mail:1443905029@qq.com

溶胶等物质[1-2]。部分 VOCs 易燃易爆,部分 VOCs 有毒,可以致癌、引起病变,严重危害人体健康[3],所以“十三五”生态环境保护规划将 VOCs 纳入大气污染防治的重要模块[4]。因此,对 VOCs 浓度进行预测研究,有助于掌握其发展和变化规律,对制定有效的污染防治对策具有重要意义。不同的研究方法拓展和推动了预测理论的发展,为其他行业的预测研究提供了参考。同时,该预测研究可为环境保护规划提供重要的数据积累,对开展污染控制有着积极的参考意义,也促进了公众参与和居民环保意识的提高。

当前,对 VOCs 等大气污染物浓度的预测研究主要是在其排放清单的基础上展开的,通过建立基准年的污染物排放清单,来实现其他时段的预测[5]。国内外学者还利用大气排放因子 S 型曲线预测大气污染物的未来排放趋势[6-7];除此之外,还有基于情景分析法的污染物浓度预测,通过识别关键不确定因素,构建几种可能出现的情景并分析内容[8];优化模型也是污染物浓度预测的常见方法[9-10]。已经提出的大气污染浓度预测模型主要有回归分析、灰色模型[11]、神经网络模型[12]、混沌模型[13]、基于时间序列的模型等[14],以及他们的组合和改进模型。最优定权组合法大气污染物浓度预测是基于多个空气质量模式,以各单项空气质量模式的组合预测误差平方和最小为原则,构建出针对大气污染的预测模型[15]。模糊综合评价方法一般都是结合预测模型来使用。通过模糊聚类分析,将影响环境质量的各因素按主次区分,预测时考虑主要因素[16]。

以上研究还存在一些不足:①由于资金、地理条件等限制,对 VOCs 并不能做到全方位监测,所获取的数据和信息不太完整;②研究主要集中在数量预测方面,较少通过划分区域精细到每一个网格进行研究;③预测过程中较少考虑气象指标等因素对预测结果的影响。为了解决上述问题,本文提出基于网格划分的空间关联区域 VOCs 浓度预测方法,以实现区域内 VOCs 精细化预测研究。

1 网格划分与编号

1.1 区域坐标集合

根据选定区域建立相应的坐标系,建立原则为其中的每一点都能用坐标表示,可以取所选范围比例尺为坐标刻度,获取不同地方的坐标,形成区域坐标集合 R_c:

$$R_c = \{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\} \quad (1)$$

式中:(x_i, y_i)表示选定区域中的第 i 个坐标,用二维平面坐标表示,其中 $i=1,2,\cdots,n$;n 表示区域坐标点总个数。

1.2 点云网格划分算法与编号

点云网格划分算法是利用点与点之间的距离关系来实现网格划分,基于一点搜索临近点形成线段,根据线段中点临近检索第三点,连接三点形成一个三角网格。对其新边进行中点临近检索,依次形成网格体系,具体步骤如下。

1)获取区域坐标点集合 R_c,初始化一个种子网格。基于点 $p_1 = (x_{m_1}, y_{m_1})$ 进行临近检索到第二个坐标点 $p_2 = (x_{m_2}, y_{m_2})$,连接两点形成线段 $L(p_1, p_2)$,再基于线段 L 的中点临近检索第三点 $p_3 = (x_{m_3}, y_{m_3})$,连接点 p_3 形成第一个三角网格,如图 1 所示。将网格形成过程中产生的每条边存入集合 E_l,开始时 $E_l = \varnothing$。

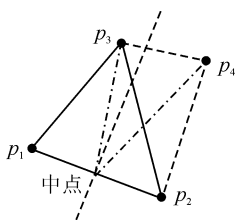


图 1 种子网格

Fig. 1 Seed grid

$$E_l = E_l \cup (p_1, p_2) \cup (p_4, p_3) \cup \cdots \cup (p_i, p_j) \quad i, j = 1, 2, \cdots, n \quad (2)$$

2)在种子网格的基础上进行网格扩充,利用中点检索,形成原始网格。从边集合 E_l 中获取未进行中点检索的边 L_h($h=1,2,\cdots, l$;l 为边的数量),其端点坐标为 $p_i = (x_{m_i}, y_{m_i})$、$p_j = (x_{m_j}, y_{m_j})$,计算其中点坐标 $C_{i,j}$;从集合 R_c 检索距离点 $C_{i,j}$ 最近且未形成边的点,中点边与新点构造出两条新边,形成一个新的三角网格,并将新产生的边存入集合 E_l 中。重复该步骤,直到边集合 E_l 中不再提供外边中点检索为止。

$$C_{i,j} = \left(\frac{x_{m_i} + x_{m_j}}{2}, \frac{y_{m_i} + y_{m_j}}{2}\right), i \neq j \quad (3)$$

$$|AB| = \sqrt{(x_{m_i} - x_{m_j})^2 + (y_{m_i} - y_{m_j})^2}, i \neq j \quad (4)$$

3)原始网格扩展,形成新网格。第二步结束形成一个原始网格,检索集合 R_c 是否存在未形成边的点,如果存在,则寻找新的种子网格重复第一、第二步,直到集合 R_c 不再有未形成边的点为止,如图 2 所示,此种情况下所选区域中存在大量的坐标点。在形成网格过程中,如果出现中断现象,只需重复上述第一、第二步形成新的网格即可。

4)编制网格顺序码,标识网格信息。在初始化种子网格时,将初始化的第一个三角网格编号为

001,表示该区域的第一个网格。在网格扩充时,根据网格划分步骤以及检索点算法,对形成的新网格依次编号,最后输出编号后的区域网格以及网格编号信息 $[(p_i,p_j,p_k),Num](k=1,2,\cdots,n)$,如图 2 所示。其中$(p_i,p_j,p_k)$表示形成该网格的三个坐标点,即 $p_i=(x_{m_i},y_{m_i})$、$p_j=(x_{m_j},y_{m_j})$、$p_k=(x_{m_k},y_{m_k})$,Num 表示网格编号,其编号值范围为 0～999的整数。

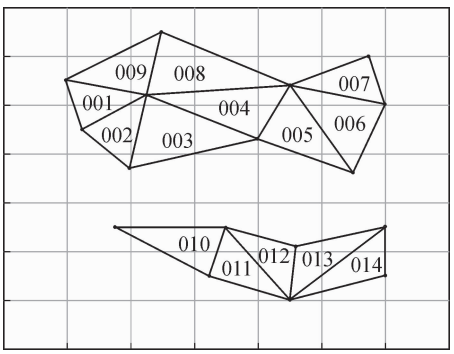


图 2 原始网格及网格编号图

Fig. 2 Original grid and grid number diagram

1.3 点云网格划分的优点

1) 不规则划分。根据所取点不规则形成大小不一的三角网格。

2) 划分区域选点灵活。根据划分需求可以随意选取点,选点过程能有效避免山川、河流等地理条件的限制。

3) 自动编码。在划分过程中自动编码表示网格,达到网格唯一性和明确性的要求。

4) 点利用率高。在网格划分中采取三点为一的原因是可以将研究区域中所有的点全部划分完,不会遗留未划分的点。

2 空间关联区域数据预估与收集

2.1 空间关联区域数据预估

1) 网格数据预估原理

在实现 VOCs 精细化监管的过程中,将区域划分成网格,在网格内设置监测点,监测设备在固定时段对网格内 VOCs 污染物进行监测,能够准确地标识该网格内 VOCs 污染物的监测浓度值。但由于网格数众多,并不是每一个网格都会设置监测点,为了收集和计算未设置监测点的网格数据,以及预估其污染物发展态势,采取克里金插值法,通过已知网格数据及其与未知网格之间的空间关联性来预估未知网格数据。

2) 克里金插值法预估过程

克里金插值被称为空间最优无偏估计器,它是以变异函数理论和结构分析为基础[17],所选变异函数由数学期望、随机场内特定点的数学期望、方差运算组成。克里金插值法会根据所选的变异函数模型进行模拟,最终对待估点进行预估。

设区域网格坐标点 p_i 处设有监测点,监测值为 $V(p_i)$,$i=1,2,\cdots,n$,则未设置监测点 p_0 的估计值可以通过周围 n 个监测点的监测值 $V(p_i)$ 求得,即

$$V^*(p_0)=\sum_{i=1}^{n}\lambda_i V(p_i) \tag{5}$$

式中:λ_i 为监测点 p_i 的权重,λ_i 的取值不仅要考虑监测点与预测点之间的距离,而且需结合二者的空间分布关系来确定,样点分布如图 3 所示。

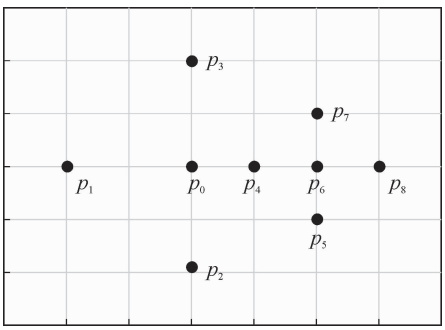


图 3 样点分布图

Fig. 3 Sample distribution

设 p_0 为待估计点,已知其邻域内有 $p_1,p_2,\cdots,p_8$ 共 8 个采样点,其位置如图 3 所示,各点的权重分别是 $\lambda_1,\lambda_2,\cdots,\lambda_8$,由于图中 p_1、p_2、p_3、p_6 到 p_0 的距离相同,并且有 p_2 与 p_3、p_1 与 p_6 关于 p_0 对称,则有 $\lambda_2=\lambda_3$,但由于样点 p_5、p_7、p_8 与 p_6 丛聚在一起,这种丛聚作用降低了样点 p_6 对待估计点 p_0 的影响,p_1 是一个单独的样点不存在丛聚影响,而且点 p_6 与 p_0 之间存在点 p_4,由于点 p_4 距离点 p_0 更近,对 p_6 存在屏蔽效应,所以 $\lambda_1>\lambda_6$。

要得到无偏最优估计值,必须满足下面两个条件:

a) 无偏估计,即 $E=[V(p_0)-V^*(p_0)]=0$

b) 估计方差最小,即

$Var[V(p_0)-V^*(p_0)]=\min$

则要求权重 λ_i 满足下列方程:

$$\gamma(p_i,p_0)=\sum_{i=1}^{n}\gamma(p_i,p_j)\lambda_i+\mu \tag{6}$$

式中:$V^*(p_0)$ 表示待估点 p_0 的克里金插值结果,是已知点 $V(p_i)$ 的加权和;E 表示真实值 $V(p_0)$ 与插值 $V^*(p_0)$ 的差的最小无偏估计;$\gamma(p_i,p_j)$ 为监测点 p_i 与 p_j 之间的半变异值;$\gamma(p_i,p_0)$ 是监测点 p_i 与内插值 p_0 之间的变异值;μ 是与方差最小化有关的拉格朗日乘数;$\sum_{i=1}^{n}\lambda_i=1$。由此方程计算出权重 λ_i,代入式(5)中即可求出待估计点 p_0 处的内插值 $V(p_0)$。

2.2 数据收集及预处理

1）数据收集

现有的监测设备不仅可以监测到 VOCs 的浓度（即单位体积排放量），而且可以分析出该区域内 VOCs 不同组成成分的含量，并将监测数据上传至服务器进行存储，对于设有监测点的网格，通过监测设备获取到 VOCs 监测值，并按照统一格式处理。已知监测点的监测数据，通过克里金插值法计算未设有监测点网格的 VOCs 组成成分预估值。将网格监测数据与网格预估数据合并，得到区域网格的 VOCs 污染物浓度值，如表 1 所示。

表 1 区域网格 VOCs 污染物浓度值

Tab. 1 VOCs emission value of regional grid

污染物	污染物浓度值/($\mu g \cdot m^{-3}$)			
	001 网格	003 网格	…	028 网格
苯	$V_{001(1)}$	$V_{003(1)}$	…	$V_{028(1)}$
甲苯	$V_{001(2)}$	$V_{003(2)}$	…	$V_{028(2)}$
⋮	⋮	⋮	⋱	⋮
苯乙烯	$V_{001(12)}$	$V_{003(12)}$	…	$V_{028(12)}$

表 1 对 VOCs 主要成分依次划分了编号：苯为 1 号、甲苯为 2 号、……、苯乙烯为 12 号，并结合单元网格顺序码，描述不同网格中不同成分的监测浓度值，如 $V_{001(1)}$ 表示 001 号网格中苯的浓度值、$V_{028(12)}$ 表示 028 号网格中苯乙烯的浓度值，依次收集得到区域网格 VOCs 污染物的浓度值。

2）数据预处理

数据预处理是对收集到的网格数据进行整理的过程，通过研究区域每个网格的 VOCs 污染物浓度数据，形成区域 VOCs 污染物数据集合：

$$D=\begin{Bmatrix} v_{11} & v_{12} & \cdots & v_{1j} \\ v_{21} & v_{22} & \cdots & v_{2j} \\ \vdots & \vdots & \ddots & \vdots \\ v_{i1} & v_{i2} & \cdots & v_{ij} \end{Bmatrix} \tag{7}$$

式中：D 表示整个研究区域网格 VOCs 组成成分浓度集合；v_{ij} 表示第 i 网格内第 j 类污染物的浓度值。

3 基于随机森林算法的VOCs 预测模型

3.1 数据集及模型结构

1）VOCs 预测模型特征

VOCs 预测模型特征可分为两大类型，VOCs 污染物和气象指标，具体特征如表 2 所示。

表 2 中 VOCs 污染物特征是指 VOCs 污染物的烷类、烃类、酯类、醇类、苯系物等具体监测成分；气象指标是指监测当天的气象特征。表 2 中所有特征形成特征向量集合 F。

表 2 VOCs 特征表

Tab. 2 VOCs characteristic table

特征类型	具体特征	符号
VOCs 污染物	苯	C_6H_6
	甲苯	C_7H_8
	⋮	⋮
	苯乙烯	C_8H_8
气象指标	温度	temp_v
	压强	pres_v
	风速	wind_v

2）VOCs 预测模型原始训练样本数据集合

基于研究区域 VOCs 污染物数据特征以及时间维度，形成区域 VOCs 数据集 V_D：

$$V_D=\begin{Bmatrix} \boldsymbol{\alpha}_{t1},\boldsymbol{\alpha}_{t2},\cdots,\boldsymbol{\alpha}_{ti} \\ \boldsymbol{\beta}_{t1},\boldsymbol{\beta}_{t2},\cdots,\boldsymbol{\beta}_{ti} \\ \boldsymbol{\gamma}_1,\boldsymbol{\gamma}_2,\cdots,\boldsymbol{\gamma}_m \end{Bmatrix} \tag{8}$$

式中：$\boldsymbol{\alpha}_{t1},\boldsymbol{\alpha}_{t2},\cdots,\boldsymbol{\alpha}_{ti}$ 和 $\boldsymbol{\beta}_{t1},\boldsymbol{\beta}_{t2},\cdots,\boldsymbol{\beta}_{ti}$ 是时序特征向量，分别表示某一时间段内区域 VOCs 污染物浓度集合和区域 VOCs 总浓度序列数据；$\boldsymbol{\gamma}_1,\boldsymbol{\gamma}_2,\cdots,\boldsymbol{\gamma}_m$ 是非时序特征向量，包含气象指标参数值、VOCs 污染物特征。

3）VOCs 预测模型构建

在上述数据处理的基础上，运用随机森林算法对研究区域 VOCs 浓度进行预测建模，建模过程如图 4 所示。

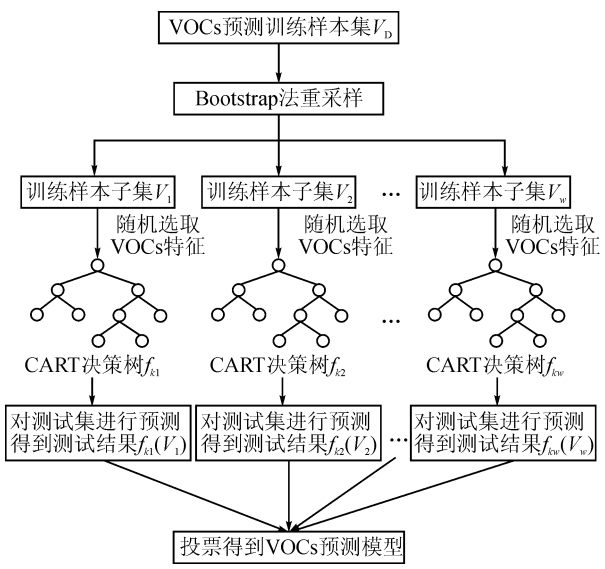


图 4 基于随机森林的区域 VOCs 预测建模过程

Fig. 4 Prediction modeling process of regional VOCs based on random forest

首先，利用 Bootstrap 方法从原始训练样本集

V_D中随机抽取多个训练样本子集,对每个子集分别进行决策树建模,然后利用测试集对各决策树进行测试,综合多棵决策树测试结果,通过投票得出最终的预测模型。

3.2 训练样本子集的随机选取

原始训练样本子集由两部分构成:一类为V_D中区域 VOCs 总量数据集合β_{ti},将其作为预测模型的输出;另一类为对应的区域网格 VOCs 污染物平均浓度集合α_{ti}和非时序特征数据集合γ_m,将其作为预测模型输入。

利用 Bootstrap 方法从V_D随机选取w个训练样本子集$V_1,V_2,\cdots,V_w$,用于构建w棵分类回归树(CART)。由于训练样本集的选取采用有放回的采样方法,在采样过程中会有 36.8%的原始样本不会出现在采集的样本集合中,这些数据称为袋数(out-of-bag,OOB),对 CART 决策树的误差进行估计。对误差估计取平均,便可得到随机森林的泛化误差估计值,由此可以对 VOCs 浓度预测模型的精度进行量化度量[18]。

3.3 CART 决策树的构建

对每个训练样本子集,采用 CART 算法生成一棵决策树,共生成w棵决策树。为保证决策树构建的随机性,采用随机子空间思想,从 VOCs 特征集合F中随机选取m个特征作为随机特征变量,参与决策树节点分裂过程,其中$m\leqslant \log_2(M+1)$,而M表示特征集合F的集合长度。此外,整个随机森林中决策树的棵数w需根据预测结果来调整。

3.4 VOCs 浓度预测结果投票及性能评价

1) VOCs 浓度预测结果

当w棵树构建完成后,利用测试集对数据进行仿真。将测试集数据V_k作为输入,得到各决策树模型预测的结果序列$\{f_{k1}(V_1),f_{k2}(V_2),\cdots,f_{kw}(V_w)\}$,基于随机森林算法的预测模型最终预测输出的 VOCs 浓度采用投票方式产生:

$$F_k(V_k)=\operatorname{argmax}\sum_{i=1}^{w}I(f_{ki}(V_k)=Y_k)\quad k=1,2,\cdots,n \tag{9}$$

式中:F_k为组合预测模型;f_{ki}为单棵决策树预测模型;I为示性函数;Y_k为各决策树预测的结果序列。将预测模型进行线性组合,即可得到区域 VOCs 浓度预测模型。

2) 性能评价指标

采用通用的模型误差、拟合程度、效率作为度量指标,进行多模型量化评估,如平均相对误差(*MRE*)和决定系数(R^2)。其中R^2表示模型输入变量对输出变量的解释程度,也称为拟合优度,取值在 0 到 1 之间。*MRE* 越小,R^2越接近于 1,说明模型准确度越高。

$$MRE=\frac{1}{n}\sum_{i=1}^{n}\left|\frac{\widehat{Q}_i-Q_i}{Q_i}\right| \tag{10}$$

$$R^2=1-\sum_{i=0}^{n}(Q_i-\widehat{Q}_i)^2/\sum_{i=0}^{n}(Q_i-\bar{Q}_c)^2 \tag{11}$$

式中:Q_i表示真实值;$\widehat{Q}_i$表示预测值;$\bar{Q}_c$表示真实值的均值;n为样本数。此外,计算模型训练时间*Time*,训练时间越小,说明模型效率越高。

4 案例分析

4.1 数据源

以西安市某区域涉及 VOCs 排放的企业为研究对象,企业清单来源于北极星网站,时间跨度为 2018 年 6 月至 2018 年 12 月。VOCs 具体浓度数据通过企业年报、地方统计年鉴以及天气后报网站获得。将研究区域划分成不同大小的网格,收集设有监测设备网格的污染物数据,通过克里金插值估计法计算出未设监测设备网格的污染物数据,形成 VOCs 数据集V_D。

4.2 研究区域网格划分及数据收集

1) 网格划分

通过点云网格算法对西安市某区进行网格划分并且对网格进行编号。首先获取该区的坐标点集合,初始化种子网格,然后在种子网格的基础上继续扩充,形成新的网格,以此类推,将整个区域的网格划分完毕,并编制网格顺序码,标识网格信息,结果如图 5 所示。

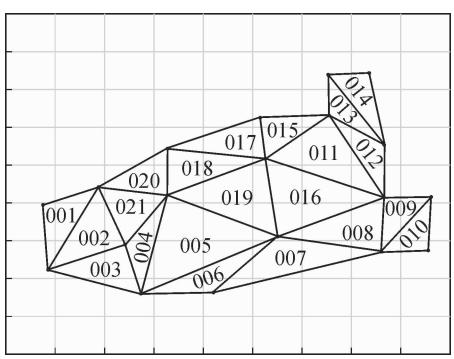


图 5 西安市某区网格划分及编号图

Fig. 5 Grid division and numbering map of a district in Xi'an

2) 数据集

研究区域中有部分网格设有监测点,由监测点获取到网格 VOCs 监测数据,包括 VOCs 污染物组分中的甲苯、乙烯、苯乙烯等 12 种物质,具体监测数值如表 3 所示。

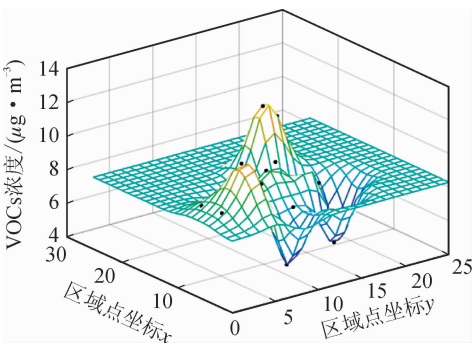
表 3 监测点 VOCs 污染物浓度值

Tab. 3 VOCs pollutant value at monitoring point

污染物	污染物浓度值/(μg·m^{-3})						
	001 网格	003 网格	007 网格	010 网格	015 网格	…	021 网格
苯	1.76	1.63	1.37	0.48	2.69	…	1.15
甲苯	2.99	3.17	2.68	1.06	4.52	…	2.44
乙烯	5.32	4.38	3.33	2.21	1.02	…	3.81
异戊烷	2.44	2.48	3.98	0.86	1.84	…	3.48
⋮	⋮	⋮	⋮	⋮	⋮	⋱	⋮
苯乙烯	0.26	0.78	1.2	1.53	2.03	…	2.02

4.3 空间关联区域数据预估

根据网格坐标点及 VOCs 污染物浓度值，构建一个 40×40 的网格，标注范围为 1～40，即使网格间距为 1。创建矩阵 $\boldsymbol{S}$ 和 $\boldsymbol{Y}$ 分别存储坐标值和观测值(即 VOCs 污染物浓度值)用于预测，根据其预估点和已知数值网格坐标点的空间位置，形成预测值表面，如图 6 所示。



注：黑色点表示原始散点数据

图 6 预测值表面和原始散点数据

Fig. 6 Predicted surface and original scatter data

根据图 6 中预测值表面，结合每个点的拟合误差值，求解出待估点的预估值，拟合误差值如图 7 所示。

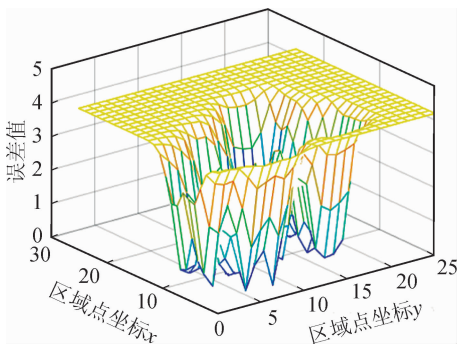


图 7 拟合误差值

Fig. 7 Fitting error value

在 λ_i 满足式(6)的条件下，将其相关数值代入式(5)计算出未设有监测点网格的 VOCs 污染物预估值，具体数值如表 4 所示。

表 4 预估点 VOCs 污染物浓度值

Tab. 4 VOCs pollutant value at estimated point

污染物	污染物浓度值/(μg·m^{-3})						
	005 网格	012 网格	013 网格	014 网格	018 网格	…	020 网格
苯	0.99	0.59	1.11	1.07	0.72	…	0.93
甲苯	1.2	0.44	1.16	1.16	0.88	…	1.12
乙烯	3.0	2.97	2.92	3.23	2.21	…	2.79
异戊烷	2.22	0.79	1.56	1.81	2.65	…	1.72
⋮	⋮	⋮	⋮	⋮	⋮	⋱	⋮
苯乙烯	1.2	0.8	1.3	1.4	0.9	…	1.1

4.4 随机森林模型预测

1) 模型构建及变量相关性分析

通过上述数据收集，获得 1 237 组 VOCs 浓度数据，按式(8)处理得到数据集 V_D 形成原始训练样本集，将其划分为训练集和验证集，构建随机森林回归模型预测 VOCs 污染物浓度。VOCs 特征集合 F 作为变量参与决策树的分裂，模型预测中每个特征所起的作用不同，其相关系数如表 5 所示。

表 5 VOCs 部分特征相关系数表

Tab. 5 Table of correlation coefficients of some features

污染物	C_6H_6	C_7H_8	C_3H_6	C_5H_{12}	VOCs
C_6H_6	1.00	0.60	0.86	0.76	0.83
C_7H_8	0.61	1.00	0.92	0.87	0.52
C_3H_6	0.86	0.73	1.00	0.82	0.73
C_5H_{12}	0.77	0.72	0.82	1.00	0.66

根据相关系数表,VOCs 与异丁烷以及环戊烷的线性相关性最大,相关系数达到了 0.8 以上,但是异戊烷与丙烯、甲苯之间的相关系数也达到了 0.8 以上,即各因素之间存在多重共线性,不满足相互独立条件,不能直接进行线性回归,所以采用随机森林预测。

2) 模型训练、验证和评估

将原始数据集合分为训练集和验证集,由式(10)、(11)分别进行模型的训练和验证,并对模型训练和验证结果进行评估,如表 6 所示。

表 6 模型评估参数表

Tab. 6 Model evaluation parameter table

评估方法	训练集	验证集
R^2	0.998 58	0.986 06
均方差	1.730 80	3.707 708
绝对差	0.341 55	0.879 92
解释度	0.997 48	0.996 12

表 6 中训练集和验证集的相关评估参数值相差很小,其决定系数 R^2 以及解释度均达到了 98% 以上,表明模型在自变量不发生变化的情况下,因变量的变异概率极小。模型训练过程中,各特征参数的重要性如图 8 所示。

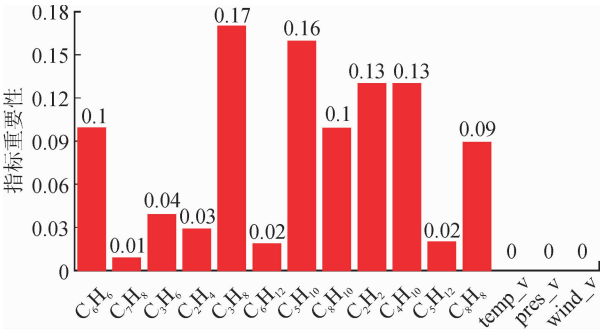


图 8 VOCs 特征影响系数表

Fig. 8 VOCs characteristic influence coefficient table

图 8 表明,VOCs 污染物的预测中,烷烃类污染物重要性比较强,相对而言温度及压强作用比较小。

3) VOCs 污染物浓度预测

从设有监测点网格中选取 19 组数据作为预测集输入模型,得到各决策树的预测结果序列,再根据式(9)投票筛选出最优预测结果,预测结果如表 7 所示。

表 7 VOCs 污染物浓度预测结果

Tab. 7 VOCs pollutant prediction results

网格	监测点	污染物	浓度实际值/($\mu g \cdot m^{-3}$)	浓度预测值/($\mu g \cdot m^{-3}$)	
				克里金插值	随机森林
001	有	VOCs	56.7	55.37	57.41
003	有	苯	1.0	0.86	0.92
007	有	异丁烷	6.0	4.59	5.21
005	无	VOCs	—	63.14	65.24
012	无	苯	—	0.73	0.62
020	无	异丁烷	—	6.12	7.03

4) 模型比较

本文是基于网格空间特性以及随机森林回归模型实现 VOCs 污染物浓度预测,现将预测结果与常用的 BP 神经网络预测结果进行比较,如表 8 所示。

表 8 给出了不同网格在两种预测模型下的 VOCs 污染物预测值,未设置监测点的网格 VOCs 污染物实际值用克里金插值预测结果代替;分别采用相对误差和平均相对误差对两种模型进行分析。由表 8 可知,随机森林模型和 BP 神经网络模型的 VOCs 总量预测值的平均误差分别是 3.15% 和 13.36%,由此可见,随机森林回归模型误差更小。

表 8 不同预测方法的结果对比

Tab. 8 Comparison of the results by different prediction methods

网格	监测点	污染物	浓度监测值/($\mu g \cdot m^{-3}$)	克里金插值浓度/($\mu g \cdot m^{-3}$)	随机森林		BP 神经网络	
					预测浓度/($\mu g \cdot m^{-3}$)	相对误差/%	预测浓度/($\mu g \cdot m^{-3}$)	相对误差/%
001	有	VOCs	56.7	—	57.41	1.25	59.63	5.17
003	有	苯	1.0	—	0.92	−8	1.23	23
007	有	异丁烷	6.0	—	5.21	−13.17	5.03	−16.17
005	无	VOCs	—	63.14	65.24	3.33	70.2	11.18
012	无	苯	—	0.73	0.62	−15.07	0.53	−27.40
020	无	异丁烷	—	6.12	7.03	14.87	5.32	−13.07
VOCs 平均相对误差					3.15%		13.36%	

5 结 论

本次预测是根据区域空间关联性以及 VOCs 污染物特征,对其浓度进行的精细化预测,意在解决监测设备不能普及部署以及区域之间污染物的流动影响问题。

1) 各区域之间的污染物存在相互影响。克里金插值法通过网格的空间地理位置来预估未设置监测点的网格数据;随机森林模型基于污染物特征之间的相关关系预测污染物的浓度,随机森林模型预测的结果更加精准。

2) 和 BP 神经网络模型相比,随机森林模型误差更小,其 VOCs 总浓度预测值的平均误差为 3.15%。模型构建过程考虑了气象指标对预测结果的影响,更能体现出 VOCs 特征之间的关联性及相互影响作用。

3) 运用基于随机森林算法的预测模型预测区域 VOCs 总浓度,同时也可以预测其组成成分的浓度(如苯、甲苯、苯乙烯等),将其与国家 VOCs 排放控制标准限值进行对比,当超出限值时,结合区域网格编号信息[(p_i, p_j, p_k), Num]获得其坐标信息(p_i, p_j, p_k),而坐标定位位置可为管理者超前管控提供依据。

参考文献:

[1] 张敬巧,吴亚君,李慧,等. 廊坊开发区秋季 VOCs 污染特征及来源解析[J]. 中国环境科学,2019,39(8):3186-3192.
ZHANG Jingqiao, WU Yajun, LI Hui, et al. Characteristics and source apportionment of ambient volatile organic compounds in autumn in Langfang development zones [J]. China Environmental Science, 2019, 39(8): 3186-3192.

[2] 鲁晓晗,王丽涛,马笑,等. 邯郸市 VOCs 变化特征及 O_3 和 SOA 生成潜势[J]. 环境科学与技术,2019,42(3):30-37.
LU Xiaohan, WANG Litao, MA Xiao, et al. Change characteristics of VOCs and their formation potential of O_3 and SOA in Handan city[J]. Environmental Science & Technology, 2019, 42(3): 30-37.

[3] 邹文君,刘杰,鲍仙华,等. 国内外涂料制造工业挥发性有机物排放标准比较[J]. 环境科学研究,2019,32(3):380-389.
ZOU Wenjun, LIU Jie, BAO Xianhua, et al. Comparison of emission standards of volatile organic compounds for coating industry in domestic and foreign areas[J]. Research of Environmental Sciences, 2019, 32(3): 380-389.

[4] 徐遵主,陆朝阳,张纪文,等. 长三角典型城市工业 VOCs 处理技术应用状况分析[J]. 环境工程,2020,38(1):54-59.
XU Zunzhu, LU Chaoyang, ZHANG Jiwen, et al. Application status of industrial VOCs treatment technologies in typical cities of the Yangtez River Delta region [J]. Environmental Engineering, 2020, 38(1): 54-59.

[5] 黄成,安静宇,鲁君,等. 长三角区域非道路移动机械排放清单及预测[J]. 环境科学,2018,39(9):3965-3975.
HUANG Cheng, AN Jingyu, LU Jun, et al. Emission inventory and prediction of non-road machineries in the Yangtze River Delta region, China[J]. Environmental Science, 2018, 39(9): 3965-3975.

[6] STREET D G, DEVANE M K, LU Z F, et al. All-time releases of mercury to the atmosphere from human activitie [J]. Environmental Science & Technology,

2011, 45(24): 10485-10491.

[7] ZHU C Y, TIAN H Z, CHENG K, et al. Potentials of whole process control of heavy metals emissions from coal-fired power plants in China[J]. Journal of Cleaner Production, 2016(114): 343-351.

[8] 安文辉,曹国良,蒙小波,等. 情景分析法预测西安市大气污染物排放总量[J]. 环境工程,2016,34(8):99-103.
AN Wenhui, CAO Guoliang, MENG Xiaobo, et al. Scenario analysis to predict the total emission of air pollutants in Xi'an[J]. Environmental Engineering, 2016, 34(8): 99-103.

[9] GARCIA-GUAANO D, CABAL H, LECHON Y, et al. Long-term behavior of CO_2 emissions from cement production in Spain: scenario analysis using an energy optimisation model[J]. Journal of Cleaner Production, 2015(99): 101-111.

[10] MORROW W R, HASANBEIGI A, SATHAYE J, et al. Assessment of energy efficiency improvement and CO_2 emissions reduction potentials in India's cement and iron & steel industries[J]. Journal of Cleaner Production, 2014(65): 131-141.

[11] WU D H, GAO C. Short-term wind power generation forecasting based on the SVM-GM approach[J]. Electric Power Components and Systems, 2018, 46(11-12): 1250-1264.

[12] ZHU K, ZHANG N N, SHI Y, et al. Within-project and cross-project just-in-time defect prediction based on denoising autoencoder and convolutional neural network[J]. IET Sofeware, 2020, 14(3): 185-195.

[13] 祝媛,黄盛. 基于小波和 BP 神经网络的大气污染物混沌预测[J]. 西南科技大学学报,2013,28(3):24-27,39.
ZHU Yuan, HUANG Sheng. Chaotic prediction of atmospheric pollutants based on wavelet and BP neural network[J]. Journal of Southwest University of Science and Technology, 2013, 28(3): 24-27,39.

[14] 严宙宁,牟敬锋,赵星,等. 基于 ARIMA 模型的深圳市大气 $PM_{2.5}$ 浓度时间序列预测分析[J]. 现代预防医学,2018,45(2):220-223.
YAN Zhouning, MOU Jingfeng, ZHAO Xing, et al. Time series prediction of atmospheric $PM_{2.5}$ concentration in Shenzhen based on ARIMA model[J]. Modern Preventive Medicine, 2018, 45(2): 220-223.

[15] 谢磊,铁治欣,宋飞扬,等. 基于最优定权组合法的大气污染物 SO_2 预测[J]. 计算机系统应用,2019,28(3):80-87.
XIE Lei, TIE Zhixin, SONG Feiyang, et al. Prediction of atmospheric pollutant SO_2 based on optimal weighted combination method [J]. Computer Systems & Applications, 2019, 28(3): 80-87.

[16] 陆婷,朱家明,陈涛,等. 基于模糊综合评价对京津冀大气污染的分析[J]. 哈尔滨商业大学学报(自然科学版),2019,35(4):503-507.
LU Ting, ZHU Jiaming, CHEN Tao, et al. Analysis of air pollution control in Beijing-Tianjin-Hebei based on fuzzy comprehensive evaluation [J]. Journal of Harbin University of Commerce (Natural Sciences Edition), 2019, 35(4): 503-507.

[17] 李如仁,李广超,陈伟,等. 京津冀气溶胶数据普通克里金插值研究[J]. 沈阳建筑大学学报(自然科学版),2020,36(1):179-185.
LI Ruren, LI Guangchao, CHEN Wei, et al. A study on ordinary Kriging interpolation of aerosol data in Jing-Jin-Ji area [J]. Journal of Shenyang Jianzhu University(Natural Science), 2020, 36(1): 179-185.

[18] LIU Y, CAO G F, ZHAO N Z, et al. Improve ground-level $PM_{2.5}$ concentration mapping using a random forests-based geostatistical approach[J]. Environmental Pollution, 2018(235): 272-282.

(责任编辑 周 蓓)