

# PM<sub>2.5</sub>浓度空间估算的神经网络与克里格方法对比

许 珊<sup>1</sup> 邹 滨<sup>1</sup> 王 敏<sup>2</sup> 刘 宁<sup>1</sup>

1 中南大学地球科学与信息物理学院,湖南 长沙,410083

2 湖南省第三测绘院,湖南 长沙,410004

**摘 要:**针对人工神经网络与克里格插值在PM<sub>2.5</sub>浓度空间估算中精度随样本点数量与耦合因素不同差异较大的问题,基于相关分析与径向基函数(radial basis function, RBF)筛选PM<sub>2.5</sub>空间变异关键影响因素,对比不同比例训练样本下普通克里格插值(ordinary Kriging, OK),仅考虑地理坐标RBF神经网络,耦合关键因素的协同克里格插值(CoKriging, CK)及RBF神经网络(CoRBF)的效果差异,并基于最优方法开展PM<sub>2.5</sub>浓度空间制图。结果表明:4种方法均能有效实现PM<sub>2.5</sub>浓度空间估算,且精度随训练样本比例增大而波动上升。考虑关键因素人口密度的CoRBF最能表现数据变化趋势,而CK在误差指标上更优越。基于CK与CoRBF的PM<sub>2.5</sub>浓度空间估算结果较好展示了污染的分异特征,前者较后者更平滑。

**关键词:**径向基函数;人工神经网络;克里格插值;大气污染;空间估算

**中图分类号:**P208

**文献标志码:**A

PM<sub>2.5</sub>是城市大气污染的主要污染物之一,长期暴露在PM<sub>2.5</sub>高污染水平的环境中会增加人们罹患心血管、呼吸系统等疾病甚至死亡的风险<sup>[1-2]</sup>,掌握城市PM<sub>2.5</sub>浓度的空间分布特征,从而有针对性地制定污染防控与风险规避的措施成为当前相关领域研究人员共同关注的热点。

目前,国际上普遍通过大气环境定点监测方式获取各城市PM<sub>2.5</sub>污染状况,但受污染源分布、自然条件以及人为活动等空间异质性的影响,各城市内部稀疏监测站点位置观测数据的空间代表性往往十分有限,仅能反映该监测站点周围一定范围内的PM<sub>2.5</sub>污染特征<sup>[3]</sup>。为此,研究人员引入了地统计、数值模拟、遥感反演、高级统计建模等一系列方法获取城市PM<sub>2.5</sub>污染连续空间分布,以了解其分异规律<sup>[4-8]</sup>。其中,地统计学估值方法克里格(Kriging)插值以区域化变量理论为基础,以变异函数为核心,在充分利用样本点空间位置信息的基础上提供了一种简单有效的线性无偏最优估值方法,成为最早引入该领域也是应用最为广泛的空间估算方法之一。

随着人工智能领域的飞速发展,人工神经网络、贝叶斯最大熵估计、随机森林等机器学习模

型因具有较强灵活性、成长性和普适性以及擅长通过学习解决复杂问题的能力,成为大气污染物浓度空间模拟的新兴前沿<sup>[9-11]</sup>。其中,人工神经网络通过从信息处理角度对人脑神经元网络进行抽象,建立由大量处理单元互联组成的非线性、自适应、并行分布式的信息处理系统,克服了传统机器学习模型在处理非结构化信息方面的缺陷,在大气污染浓度预测与空间估算方面已有较多应用<sup>[9,12-14]</sup>。

然而,作为大气污染浓度空间估算机器学习经典方法的人工神经网络与作为地统计方法重要代表的克里格插值,在不同数据情景下(如不同样本点数量与耦合因素),两种方法的优势是否会随之发生改变,这些都是高效、准确了解PM<sub>2.5</sub>污染空间变异特征亟待回答的关键问题。

本研究结合美国本土PM<sub>2.5</sub>年均浓度与气象、人口、土地、交通专题数据,在基于相关分析与径向基函数(radial basis function, RBF)筛选PM<sub>2.5</sub>污染空间变异关键影响因素的基础上,通过构建和比较不同数据情景下耦合与未耦合决定性影响因素的RBF神经网络与克里格插值方法,探讨PM<sub>2.5</sub>空间估算的不同方法在不同数据条件

收稿日期:2019-12-27

项目资助:国家重点研发计划(2016YFC0206205);国家自然科学基金(41871317);中南大学创新驱动计划(2018CX016)。

第一作者:许珊,博士生,主要从事城市大气污染时空精细模拟研究。shan\_xu@csu.edu.cn

通讯作者:邹滨,博士,教授。210010@csu.edu.cn

下的适用性,从而为有效制定大气污染防治措施、规避污染暴露风险提供科学支持。

## 1 研究方法 with 数据

考虑 PM<sub>2.5</sub> 浓度空间分布的平稳特征,以美国本土为研究区域,包括 48 个州和哥伦比亚特区,位于 67°W~125°W、25°N~49°N 之间。研究包括数据获取与处理、关键因素选取、模型构建与精度评定 3 个主要步骤(如图 1 所示)。为降低后续实验数据处理与模型构建难度,考虑到德克萨斯作为本土面积最大州,具有丰富土地利用类型,地势西高东低,与研究区整体趋势基本一致,具有一定代表性。首先,以德克萨斯州为例对气象、人口、土地利用、交通等影响因子进行筛选,确定关键影响因素。然后,基于美国本土 PM<sub>2.5</sub> 数据,构建不同数据情景下普通克里格插值(ordinary Kriging, OK),仅以监测站点的地理坐标为变量的 RBF 神经网络,以关键影响因素为辅助变量的协同克里格插值(CoKriging, CK),以及耦合地理坐标与关键影响因素变量的 RBF 神经网络(称为 CoRBF,以示区别),并开展精度分析对比。在此基础上,综合考虑监测站点间距离与辅助数据空间分辨率,选取最优方法开展研究区全域范围 PM<sub>2.5</sub> 浓度的空间分布制图。

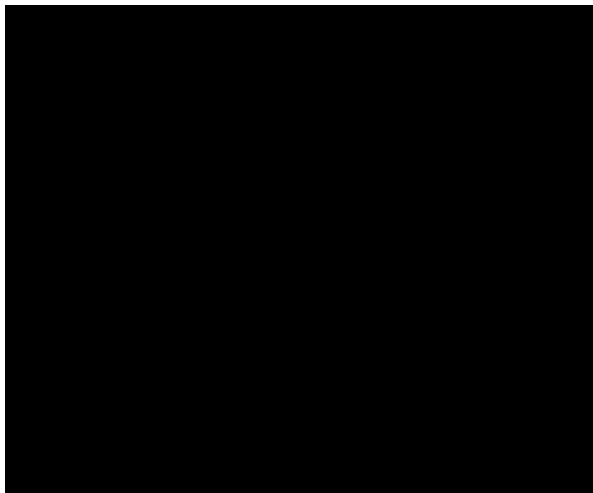


图1 研究设计

Fig.1 Study Design

### 1.1 数据获取与处理

PM<sub>2.5</sub> 浓度数据来源于美国环境保护局(United States Environmental Protection Agency, EPA),通过对 2010 年日均数据进行处理求取年均值。美国本土各县人口总数采自美国人口普查局(US Census Bureau),结合各县面积计算得

到人口密度。气象数据包括年均温度、湿度、降水和风速,采自美国国家气候数据中心(National Climatic Data Center, NCDC)和国家海洋与大气管理局(National Oceanic and Atmospheric Administration, NOAA)网站。土地利用类型数据源于美国国家地质调查局土地覆盖数据库(United States Geological Survey, USGS),并按照需求重分类为水域、居住区、开放空间、植被覆盖区(包括草地与林地)、湿地与耕地,空间分辨率为 30 m×30 m。交通数据涵盖主要街道,采自 ESRI 公司国家街道与地理编码数据库。

研究区 2010 年有 1 321 个 PM<sub>2.5</sub> 监测站,为方便后续精度评价工作的开展,以训练样本比例 5% 为起点,5% 为间隔,逐渐增加样本点数量(至 95%),研究过程中采用 ArcGIS 子集要素工具对上述数据随机分为训练样本和检验样本两组。

针对辅助要素,首先利用 ArcGIS 叠加分析,以 PM<sub>2.5</sub> 监测站为中心,提取、计算各中心点 100 m、200 m、300 m、500 m、800 m、1 000 m、1 500 m、2 000 m、2 500 m、3 000 m、3 500 m、4 000 m、4 500 m、5 000 m 范围内道路长度、距最近道路距离、各土地利用类型面积占比。然后,对比分析 2010 年美国本土 1 000 个气象监测点气温、降水量等年均值反距离加权(inverse distance weighted, IDW)、OK 与样条函数 3 种经典方法的插值精度,选取十折交叉验证精度最优的 IDW( $R^2$ : 0.89~0.90)得到气象要素空间分布,进而提取监测站点对应数值。最后,提取监测点对应县级人口密度,按位置匹配对应 PM<sub>2.5</sub> 年均浓度构建数据样本。

### 1.2 关键因素选取

基于德克萨斯州 54 个监测点的数据样本,以皮尔逊相关系数为指标,初步快速筛选出与 PM<sub>2.5</sub> 浓度相关度较大的影响因子。参考相关研究,认为与 PM<sub>2.5</sub> 浓度在 0.01 水平上显著相关且相关系数大于 0.6 的因素为强相关因素<sup>[9]</sup>。在此基础上,结合监测站点的地理坐标,构建筛选出的强相关因素单个因素与组合因素的 PM<sub>2.5</sub> 浓度 RBF 网络,采用十折交叉验证定量评价各个模型的精度,从而实现 PM<sub>2.5</sub> 浓度空间变化的关键因素进一步筛选。

### 1.3 模型构建与精度评定

#### 1.3.1 RBF 构建

RBF 神经网络是一种常见的前馈神经网络,具有逼近精度高、网络规模小、学习速度快和全局收敛等优点<sup>[14]</sup>。它由输入层、隐含层和输出层

3个神经元层组成。每一层都包含大量互不相连的简单神经元,输入层神经元一般采用高斯函数将输入数据传递到隐含层,再由隐含层将数据激活放大后传递到输出层并输出。其中,高斯函数表达式为:

$$R_i(x) = \exp \left[ -\frac{\|x - c_i\|^2}{2\delta_i^2} \right] \quad (1)$$

式中,  $i=1, 2, \dots, m$ ;  $c_i$  为第  $i$  个基函数的中心;  $\delta_i$  为第  $i$  个感知的变量;  $\|x - c_i\|$  为  $(x - c_i)$  的范数, 表示  $x$  与  $c_i$  间的距离。

首先以地理位置(或地理位置与关键要素)作为输入层,  $PM_{2.5}$  浓度作为输出层; 然后设计隐含层, 研究证明 RBF 神经网络能以任意精度逼近任意连续函数, 当隐含层神经元数目足够多, 具有一个隐含层的 RBF 神经网络就具有较高的精度<sup>[14-15]</sup>。因此本研究采用 3 层模型, 对不同训练比例下的网络进行性能测试, 确定不同训练样本下隐含层神经元个数。最后, 对网络进行训练, 并对比效果。

### 1.3.2 克里格插值

普通克里格方法是满足假设的区域化变量的线性估计, 它假设数据的变化呈正态分布, 每一估值都是由其邻近观测值加权平均计算而得<sup>[16]</sup>。观测值离估值点越近, 其结果影响越大, 权重选择必须保证最终估值方差最小。OK 方法可以简单表达为:

$$\hat{Z}(s_0) = \sum_{i=1}^n w_i Z(s_i) \quad (2)$$

式中,  $Z(s_i)$  为已测得的第  $i$  个位置的属性值;  $w_i$  为在第  $i$  个位置上测得值的权重, 其取决于已知点的拟合模型、到插值点的距离和插值点周围的已知样点的空间关系, 由变异函数给出, 各权重之和为 1;  $s_0$  为待插值的位置;  $n$  为已知样本点的数量。

协同克里格方法把区域化变量理论的最佳估值方法从单一属性发展到两个以上的协同区域化属性, 利用与另一个变量之间的空间相关性实现估值<sup>[17]</sup>。基本表达式如下:

$$\hat{Z}(s_0) = \sum_{i=1}^n w_i Z_1(s_i) + \sum_{j=1}^p w_j Z_2(s_j) \quad (3)$$

式中,  $Z_1(s_i)$ 、 $Z_2(s_j)$  分别为已测得位置主变量和辅变量的属性值;  $w_i$ 、 $w_j$  为对应权重, 由协变异函数给出;  $s_0$  为待插值的位置;  $n$ 、 $p$  为主变量和辅变量已知样点的数量。

首先判断训练样本数据是否服从或近似服从

正态分布, 对于不满足要求的数据进行 Box-Cox 或对数变换; 然后以方差最小为原则, 选择方差变异的指数函数类型; 经过多次设置比较, 选择合适的搜索半径和邻近数据点数; 最后基于训练样本采取普通/协同克里格插值方法对输入的训练样本数据进行插值, 并最终提取检验样本各点  $PM_{2.5}$  估算值。

### 1.3.3 精度评定

采用检验样本的模型估算浓度与观测值之间的均方误差(mean square error, MSE)、平均绝对误差(mean absolute error, MAE)、相对误差(relative error, RE)、决定系数( $R^2$ )以及一致度指数(index of agreement, IOA) 5 个指标来度量各个模型的空间估算精度。各指标计算公式如下:

$$MSE = \frac{1}{n} \sum_i (O_i - S_i)^2 \quad (4)$$

$$MAE = \frac{1}{n} \sum_i |O_i - S_i| \quad (5)$$

$$RE = \frac{1}{n} \sum_i \left( \frac{|S_i - O_i|}{O_i} \right) \times 100\% \quad (6)$$

$$R^2 = \frac{\left[ \sum_i (S_i - \bar{S})(O_i - \bar{O}) \right]^2}{\sum_i (O_i - \bar{O})^2 \sum_i (S_i - \bar{S})^2} \quad (7)$$

$$IOA = 1 - \frac{\frac{1}{n} \sum_i (S_i - O_i)^2}{(|S_i - \bar{O}| + |O_i - \bar{O}|)} \quad (8)$$

式中,  $i$  指样本点编号;  $O_i$  与  $S_i$  分别是检验样本点的  $PM_{2.5}$  年均浓度观测值与估算值;  $\bar{O}$  与  $\bar{S}$  指各点  $PM_{2.5}$  年均浓度观测值与估算值的平均值;  $n$  指检验样本点的数量。上述指标中, MSE、MAE 和 RE 值越小,  $R^2$  和 IOA 越接近于 1, 表示估算精度越高。

## 1.4 $PM_{2.5}$ 浓度空间制图

对比 4 种方法的精度, 选取 95% 比例下空间估算效果最优的方法开展  $PM_{2.5}$  浓度空间制图。 $PM_{2.5}$  监测站点间距离范围较大 ( $0.0004^\circ \sim 2.58^\circ$ ), 经统计分析知, 接近 35% 站点间距离小于  $0.1^\circ$ , 而辅助数据空间分辨率差异较大, 人口数据以县为单位, 综合考虑, 选定  $0.1^\circ \times 0.1^\circ$  为  $PM_{2.5}$  浓度制图空间分辨率。对于人工神经网络方法, 首先采用 ArcGIS 要素类工具生成  $0.1^\circ \times 0.1^\circ$  渔网点, 提取点位对应地理坐标与辅助变量值后, 利用训练好的 RBF 神经网络估算对应  $PM_{2.5}$  浓度; 对于克里格方法, 则利用 ArcGIS 的地统计分析工具开展制图。



## 2 结果分析

### 2.1 关键影响因素

经相关性分析筛选出人口密度、降水和站点周围 500 m 缓冲区范围内道路总长度 3 个强相关因素。基于这 3 个因素与地理坐标( $X, Y$ )及组合

构建的 PM<sub>2.5</sub> 年均浓度空间估算 RBF 神经网络模型交叉检验结果如表 1 所示。不同单因素与多因素组合构建的 RBF 神经网络估算效果都存在一定的差异,人口密度对 PM<sub>2.5</sub> 年均浓度空间变异的贡献作用不可忽视。

表 1 基于 RBF 网络的关键影响因素筛选

Tab.1 Selection of Key Factors Based on RBF Network

关键影响因素	MSE/( $\mu\text{g}\cdot\text{m}^{-3}$ )	MAE/( $\mu\text{g}\cdot\text{m}^{-3}$ )	RE/%	$R^2$
$X, Y$	0.91	0.54	6.11	0.67
$X, Y$ 、人口密度	0.42	0.52	4.14	0.90
$X, Y$ 、降水	0.95	0.78	7.55	0.76
$X, Y$ 、道路总长度	1.26	0.89	7.98	0.55
$X, Y$ 、人口密度、降水	1.08	0.86	8.05	0.63
$X, Y$ 、人口密度、道路总长度	1.11	0.84	8.50	0.66
$X, Y$ 、降水、道路总长度	0.85	0.73	6.12	0.86
$X, Y$ 、人口密度、降水、道路总长度	0.87	0.85	8.25	0.82

从表 1 可知,基于人口密度单因素的 RBF 神经网络估算效果最优,而基于道路总长度单因素的最差。前者 MSE、MAE 与 RE 值均为所有网络中最小值,分别为  $0.42 \mu\text{g}/\text{m}^3$ 、 $0.52 \mu\text{g}/\text{m}^3$ 、 $4.14\%$ ,  $R^2$  则最大,达到 0.90;后者  $R^2$  最小(0.55), MSE ( $1.26 \mu\text{g}/\text{m}^3$ ) 与 MAE ( $0.89 \mu\text{g}/\text{m}^3$ ) 最大。基于双因素的 RBF 神经网络中,降水和道路总长度耦合下的模型比其他组合效果更好,总体精度位列所有网络中的第二,其 MSE、MAE、RE 和  $R^2$  值分别为  $0.85 \mu\text{g}/\text{m}^3$ 、 $0.73 \mu\text{g}/\text{m}^3$ 、 $6.12\%$  和 0.86,而基于人口密度和道路总长度的效果最差。基于人口密度、降水和道路总长度三者耦合下的 RBF 神经网络空间估算精度整体第三,这与已报道的 PM<sub>2.5</sub> 浓度在多因素耦合情况下效果更好的结论不一致<sup>[9]</sup>。究其原因:一方面,已有研究是对不同类型变量进行对比,并没有进一步对比分析单个因素或多个因素耦合对模型的影响;另一方面,在建模过程中只考虑因素对大气污染浓度的单独作用(即因子作用相互独立),而忽略了输入变量间可能存在的交互作用。事实上,当两个及以上因素共同作用时,还可能会增加或减弱对因变量的解释力<sup>[18]</sup>。例如邹滨等<sup>[19]</sup>的研究表明,气象因素的引入会增强土地利用因素对 PM<sub>2.5</sub> 浓度空间分布的解释作用,而 Guo 等<sup>[20]</sup>发现相对湿度会弱化气溶胶光学厚度与 PM<sub>2.5</sub> 浓度的相关关系。

### 2.2 PM<sub>2.5</sub> 浓度估算值与观测值

基于上述结果,分别构建仅基于地理坐标的

RBF 神经网络、耦合地理坐标与人口要素的 RBF 神经网络 CoRBF、OK、耦合人口要素的协同克里格插值 CK 方法估算研究区 PM<sub>2.5</sub> 年均浓度。图 2 是训练样本比例分别为 15%、35%、55%、75%、95% 时,检验样本观测值与 4 种方法 PM<sub>2.5</sub> 估算浓度折线分布(每个数据点代表一个监测点的浓度值)。

由图 2 可知,不同比例 PM<sub>2.5</sub> 变化整体趋势基本一致,但在高低值区有差距。不同方法下模型效果之间存在差异,以训练样本比例为 55% 为例,OK 方法估算值起伏不大,无法表现出观测值的高值低值分布。这是由于克里格方法具有较强的平滑作用,影响了污染物空间变异的数据统计特征。CK 方法通过辅助变量融合了更多的空间信息,较 OK 方法在一定程度上略有起伏,空间插值效果得到提升。基于 RBF 神经网络方法的估算能力较 OK 方法好,在估算高低值方面具有一定优势,这种现象即使在训练样本点较少时也比较显著。说明 RBF 神经网络能够克服克里格插值的平滑效应,这主要是因为神经网络方法具有以任意精度逼近任意连续函数的能力。基于 CoRBF 神经网络较前 3 种方法,其估算效果显著增强,可以基本表现出观测值的变化趋势。这与 Caselli 等<sup>[21]</sup>、Asadollahfardi 等<sup>[22]</sup>利用神经网络预测大气颗粒物浓度结果一致,即耦合特征影响因素的神经网络预测效果比仅依靠污染观测数据效果好。同时,该结果也进一步佐证了 PM<sub>2.5</sub> 浓度的空间分布受人口密度的影响较大这一结论。



图2 检验样本的PM<sub>2.5</sub>浓度观测值与各方法估算值Fig.2 Observations and Estimations of PM<sub>2.5</sub> in Validation Sets Based on Four Methods

### 2.3 精度指标统计

图3为4种方法精度评价指标MSE、MAE、RE、 $R^2$ 与IOA随训练样本比例增加的变化情况(每个数据点代表给定训练样本比例下精度指标的值)。总体而言,PM<sub>2.5</sub>浓度空间估算的整体效果CK最优,而OK最差,研究采用方法的精度随着训练样本比例的增大而波动提升。

在相同训练比例下,除IOA外,训练方法整体效果基本均为CK>CoRBF>RBF>OK。例如,30%、60%与90%样本比例下,4种方法MSE依次为(5.00  $\mu\text{g}/\text{m}^3$ 、5.36  $\mu\text{g}/\text{m}^3$ 、5.60  $\mu\text{g}/\text{m}^3$ 、5.90  $\mu\text{g}/\text{m}^3$ )、(3.93  $\mu\text{g}/\text{m}^3$ 、4.34  $\mu\text{g}/\text{m}^3$ 、4.50  $\mu\text{g}/\text{m}^3$ 、4.88  $\mu\text{g}/\text{m}^3$ )与(2.45  $\mu\text{g}/\text{m}^3$ 、2.64  $\mu\text{g}/\text{m}^3$ 、2.67  $\mu\text{g}/\text{m}^3$ 、2.70  $\mu\text{g}/\text{m}^3$ );

决定系数 $R^2$ 分别为(0.50、0.51、0.53、0.57)、(0.52、0.54、0.56、0.62)与(0.64、0.67、0.70、0.70)。但是在个别比例存在波动,如75%比例下,基于OK方法及RBF方法的PM<sub>2.5</sub>估算浓度值与观测值的MAE分别为1.41  $\mu\text{g}/\text{m}^3$ 与1.45  $\mu\text{g}/\text{m}^3$ ;又如在50%比例下,基于CK与CoRBF方法浓度估值与观测值的 $R^2$ 分别为0.52与0.54。对于IOA,则当样本数量较少时( $\leq 50\%$ ),整体上CoRBF最高,RBF次之,而OK最低,当样本数量增加,CK>CoRBF>RBF且趋于接近。结果进一步证明融入人口因素的CK方法与CoRBF神经网络比OK与RBF方法PM<sub>2.5</sub>估算效果好,也说明了人口因素在估算PM<sub>2.5</sub>浓度时作为输入样本的必要

性。结合图 2 与图 3, 4 种方法中, CK 方法在研究区 PM<sub>2.5</sub> 浓度估算中偏离实际观测值最小, 模型模拟精度最高, 而 CoRBF 估算数据与 PM<sub>2.5</sub> 观测值一致性最高, 在研究区内最能表现出数据的变化趋势。这种现象产生的原因可能是估算点与

其他监测点在一定空间距离上存在关联性, CK 方法的每一估值都由其邻近观测值加权平均计算而得到, 在一定程度上提升了方法统计精度, 但同时, 由于这一平滑作用, 在保持观测数据的统计特征方面则稍逊一筹。

图 3 基于各种方法的 PM<sub>2.5</sub> 浓度估算值与观测值间的精度对比

Fig.3 Accuracy Comparison of Between Observations and Estimations of PM<sub>2.5</sub> Based on Four Methods

随着样本比例的增加, 4 种方法的精度整体呈波动上升趋势, MSE、MAE 与 RE 分别从 5% 比例下的 6.00~6.60  $\mu\text{g}/\text{m}^3$ 、1.69~1.75  $\mu\text{g}/\text{m}^3$ 、25%~27% 下降至 95% 比例下的 2.44~3.59  $\mu\text{g}/\text{m}^3$ 、1.20~1.50  $\mu\text{g}/\text{m}^3$ 、14%~17%,  $R^2$  与 IOA 则分别从 0.42~0.46 与 0.77~0.80 上升至 0.64~0.86 与 0.87~0.92。当训练样本分别占样本总数的 5%、55%、95% 时, 基于 OK 方法的检验样本位置点的 PM<sub>2.5</sub> 浓度估算值与其相应观测值之间的 MSE 分别为 6.32  $\mu\text{g}/\text{m}^3$ 、4.70  $\mu\text{g}/\text{m}^3$ 、3.59  $\mu\text{g}/\text{m}^3$ , RBF 分别为 6.60  $\mu\text{g}/\text{m}^3$ 、4.30  $\mu\text{g}/\text{m}^3$ 、2.88  $\mu\text{g}/\text{m}^3$ 。CK 和 CoRBF 方法比 OK 和 RBF 方法 MSE 值分别降低了 0.30  $\mu\text{g}/\text{m}^3$ 、0.84  $\mu\text{g}/\text{m}^3$ 、1.15  $\mu\text{g}/\text{m}^3$  和 0.52  $\mu\text{g}/\text{m}^3$ 、0.20  $\mu\text{g}/\text{m}^3$ 、0.52  $\mu\text{g}/\text{m}^3$ 。MAE、RE 的下降趋势与之相似,  $R^2$ 、IOA 评价指标的上升趋势也进一步反映了这一结果。与文献[9] ( $R^2=0.39\sim0.81$ )、文献[23] ( $R^2=0.66$ ) 耦合地理特征要素开展城市 PM<sub>2.5</sub> 浓度神经网络空间估算研究结果相比, 当训

练样本比例达到 80% 以上, 本研究预测精度的优势性逐渐显现。然而, 在相邻比例区间内, 不同方法的精度存在一定波动, OK 方法波动最大, 如训练比例在 65%、50% 和 5% 时, 各方法 MSE 误差出现极大值, 训练比例在 5% 的估算效果比 10% 效果好。这是因为除数量外, 模型精度还受所选样本能否正确反映整体数据内在规律性及其分布特征的影响。

## 2.4 PM<sub>2.5</sub> 浓度空间分布

图 4(a)、图 4(b) 分别为 CK 与 CoRBF 神经网络估算得到的 PM<sub>2.5</sub> 浓度空间分布。对比监测点年均观测值 (图 4(c)) 可知, 两种方法估算得到的 PM<sub>2.5</sub> 浓度整体空间差异特征与观测值基本一致, 即东部高于西部, 西海岸及中西部存在范围较小的高值区, 进一步说明 CK 与 CoRBF 都能有效实现 PM<sub>2.5</sub> 浓度空间估算。CoRBF 神经网络的估算值体现了较为细致的空间差异, 这一现象在东部各州尤为显著。相较而言, CK 插值的 PM<sub>2.5</sub> 浓度空间分布具有更为显著的空间平滑特征。

图4 分布模式点集

Fig.4 Spatial Patterns Dens

从比关键影响因素人口密度的空间分布图4(d)来看,人口密度越大,PM<sub>2.5</sub>浓度的估算值与观测值都越高,东部地区人口密度的区县差异与CoRBF神经网络估算的PM<sub>2.5</sub>浓度差异几乎一致,说明相较于CK,人口要素对CoRBF神经网络的贡献作用更大。

### 3 结 语

本文对不同数据情景下人工神经网络与克里格插值对PM<sub>2.5</sub>浓度空间分布估算的效果差异进行分析,得到如下结论:

1) 人口因素在分析PM<sub>2.5</sub>空间变异方面比其他单因素及多因素耦合效果好;

2) RBF神经网络方法的PM<sub>2.5</sub>浓度空间估算效果整体优于OK方法,但在训练样本较少时差异较小,耦合人口关键影响因素的CoRBF神经网络与CK方法的PM<sub>2.5</sub>浓度估算结果比未耦合该因素的RBF与OK更优;

3) RBF、CoRBF神经网络、OK以及CK方法都能够有效实现PM<sub>2.5</sub>浓度的空间估算,且精度整体随训练样本比例增大而波动上升。在考虑人口密度这一辅助因素前提下,CoRBF神经网络方法在保持观测数据的统计特征方面是最好的模型,然而CK方法在误差指标上更显优越性;

4) 基于CK插值与CoRBF神经网络的PM<sub>2.5</sub>浓度空间估算结果都能够较好地体现真实的污染分异特征,CK的平滑效应更为显著。

但出于对数据平稳特征的考虑,本研究仅以

美国本土PM<sub>2.5</sub>浓度数据为对象,在年均尺度开展了上述两种方法的比较。随着车载移动监测和众源监测等新兴、低成本污染监测方式以及多源时空大数据的逐渐兴起,未来还可以在更多区域,从季节、月均、日均等更为精细的尺度进一步探索各种大气污染浓度空间估算方法的适用条件。

### 参 考 文 献

- [1] Cohen A J, Brauer M, Burnett R, et al. Estimates and 25-Year Trends of the Global Burden of Disease Attributable to Ambient Air Pollution: An Analysis of Data from the Global Burden of Diseases Study 2015 [J]. *The Lancet*, 2017, 389(10 082): 1 907-1 918
- [2] Di Q, Wang Y, Zanobetti A, et al. Air Pollution and Mortality in the Medicare Population [J]. *New England Journal of Medicine*, 2017, 376(26): 2 513-2 522
- [3] Apte J S, Messier K P, Gani S, et al. High-Resolution Air Pollution Mapping with Google Street View Cars: Exploiting Big Data [J]. *Environmental Science & Technology*, 2017, 51(12): 6 999-7 008
- [4] Li Shenxin, Zou Bin, Liu Xingquan, et al. Pollution Status and Spatial-Temporal Variations of PM<sub>2.5</sub> in China During 2013—2015 [J]. *Research of Environmental Sciences*, 2017, 30(5): 678-687 (李沈鑫, 邹滨, 刘兴权, 等. 2013—2015年中国PM<sub>2.5</sub>污染状况时空变化[J]. 环境科学研究, 2017, 30(5): 678-687)
- [5] He J, Zhang Y, Wang K, et al. Multi-Year Application of Wrf-Cam5 over East Asia-Part I: Compre-



- hensive Evaluation and Formation Regimes of O<sub>3</sub> and PM<sub>2.5</sub> [J]. *Atmospheric Environment*, 2017, 165: 122-142
- [6] Fang X, Zou B, Liu X, et al. Satellite-Based Ground PM<sub>2.5</sub> Estimation Using Timely Structure Adaptive Modeling [J]. *Remote Sensing of Environment*, 2016, 186: 152-163
- [7] Zou B, Pu Q, Bilal M, et al. High-Resolution Satellite Mapping of Fine Particulates Based on Geographically Weighted Regression [J]. *IEEE Geoscience & Remote Sensing Letters*, 2017, 13(4): 495-499
- [8] Jiao Limin, Xu Gang, Zhao Suli, et al. LUR-Based Simulation of the Spatial Distribution of PM<sub>2.5</sub> of Wuhan [J]. *Geomatics and Information Science of Wuhan University*, 2015, 40(8): 1 088-1 094(焦利民, 许刚, 赵素丽, 等. 基于 LUR 的武汉市 PM<sub>2.5</sub> 浓度空间分布模拟[J]. 武汉大学学报·信息科学版, 2015, 40(8): 1 088-1 094)
- [9] Zou B, Wang M, Wan N, et al. Spatial Modeling of PM<sub>2.5</sub> Concentrations with a Multifactorial Radial Basis Function Neural Network [J]. *Environmental Science & Pollution Research*, 2015, 22(14): 1-10
- [10] Xiao L, Lang Y, Christakos G. High-Resolution Spatiotemporal Mapping of PM<sub>2.5</sub> Concentrations at Mainland China Using a Combined BME-GWR Technique [J]. *Atmospheric Environment*, 2018, 173: 295-305
- [11] Liu Y, Cao G, Zhao N, et al. Improve Ground-Level PM<sub>2.5</sub> Concentration Mapping Using A Random Forests-Based Geostatistical Approach [J]. *Environmental Pollution*, 2017, 235: 272-282
- [12] Mishra D, Goyal P, Upadhyay A. Artificial Intelligence Based Approach to Forecast PM<sub>2.5</sub> During Haze Episodes: A Case Study of Delhi, India [J]. *Atmospheric Environment*, 2015, 102: 239-248
- [13] Adams M D, Kanaroglou P S. Mapping Real-Time Air Pollution Health Risk for Environmental Management: Combining Mobile and Stationary Air Pollution Monitoring with Neural Network Models [J]. *Journal of Environmental Management*, 2016, 168: 133-141
- [14] Ordieres J B, VergarA E P, Capuz R S, et al. Neural Network Prediction Model for Fine Particulate Matter (PM<sub>2.5</sub>) on the US — Mexico Border in El Paso (Texas) and Ciudad Juárez (Chihuahua) [J]. *Environmental Modelling & Software*, 2005, 20(5): 547-559
- [15] Park J, Sandberg I W. Universal Approximation Using Radial-Basis-Function Networks [J]. *Neural Computation*, 1991, 3(2): 246-257
- [16] Stein M L. Interpolation of Spatial Data: Some Theory for Kriging [M]. New York: Springer-Verlag, 1999: 23-27
- [17] Myers D E. Matrix Formulation of Co-Kriging [J]. *Journal of the International Association for Mathematical Geology*, 1982, 14(3): 249-257
- [18] Wang J F, Li X H, Christakos G, et al. Geographical Detectors-Based Health Risk Assessment and Its Application in the Neural Tube Defects Study of the Heshun Region, China [J]. *International Journal of Geographical Information Science*, 2010, 24(1): 107-127
- [19] Zou Bin, Xu Shan, Zhang Jing. Spatial Variation Analysis of Urban Air Pollution Using GIS: A Land Use Perspective [J]. *Geomatics and Information Science of Wuhan University*, 2017, 42(2): 216-222 (邹滨, 许珊, 张静. 土地利用视角空气污染空间分异的地理分析[J]. 武汉大学学报·信息科学版, 2017, 42(2): 216-222)
- [20] Guo J, Xia F, Zhang Y, et al. Impact of Diurnal Variability and Meteorological Factors on the PM<sub>2.5</sub>-Aod Relationship: Implications for PM<sub>2.5</sub> Remote Sensing [J]. *Environmental Pollution*, 2017, 221: 94-104
- [21] Caselli M, Trizio L, Gennaro G D, et al. A Simple Feedforward Neural Network for the PM<sub>10</sub> Forecasting: Comparison with a Radial Basis Function Network and a Multivariate Linear Regression Model [J]. *Water Air & Soil Pollution*, 2009, 201(1-4): 365-377
- [22] Asadollahfardi G, Zangooei H, Aria S H. Predicting PM<sub>2.5</sub> Concentrations Using Artificial Neural Networks and Markov Chain: A Case Study Karaj City [J]. *Asian Journal of Atmospheric Environment*, 2016, 10(2): 67-79
- [23] Yao L, Lu N. Spatiotemporal Distribution and Short-Term Trends of Particulate Matter Concentration over China, 2006 — 2010 [J]. *Environmental Science & Pollution Research International*, 2014, 21(16): 9 665-9 675

## Performance Comparison of Artificial Neural Network and Kriging in Spatial Estimation of PM<sub>2.5</sub> Concentration

XU Shan<sup>1</sup> ZOU Bin<sup>1</sup> WANG Min<sup>2</sup> LIU Ning<sup>1</sup>

1 School of Geosciences and Info-Physics, Central South University, Changsha 410083, China

2 The Third Surveying and Mapping Institute of Hunan Province, Changsha 410004, China

**Abstract:** Performance of artificial neural network modeling and Kriging interpolation in PM<sub>2.5</sub> concentration estimation varies with sample sizes and predictor variables change. This paper analyzes the performance of ordinary Kriging (OK), radical basis function (RBF) networks based on geographic coordinates, CoKriging and RBF with the key factor(s) (CK and CoRBF) selected by correlation analysis and RBF network, using different training sets with various sizes. The spatial distribution of PM<sub>2.5</sub> concentration is then estimated by the best performed method. Results show that RBF, CoRBF, OK, and CK can all be used to estimate PM<sub>2.5</sub> concentration efficiently, and their accuracies improved unstably as the number of training sites increase. CoRBF with the key factor of population illustrates the largest variation of PM<sub>2.5</sub> concentration, while CK has the highest coefficient of determination ( $R^2$ ) and index of agreement (IOA) and the lowest mean square error (MSE), mean absolute error (MAE), and relative error (RE). Correspondingly, the spatial pattern of CK estimated PM<sub>2.5</sub> concentration is smoother than CoRBF estimated PM<sub>2.5</sub> concentration, while they both are very similar to site measurements and reveal detailed information.

**Key words:** radical basis function; artificial neural network; Kriging interpolation; PM<sub>2.5</sub>; spatial estimation

**First author:** XU Shan, PhD candidate, specializes in fine spatial-temporal modeling of urban air pollution. E-mail: shan\_xu@csu.edu.cn

**Corresponding author:** ZOU Bin, PhD, professor. E-mail: 210010@csu.edu.cn

**Foundation support:** The National Key Research and Development Program of China (2016YFC0206205); the National Natural Science Foundation of China (41871317); the Innovation Driven Program of Central South University (2018CX016).

**引文格式:** XU Shan, ZOU Bin, WANG Min, et al. Performance Comparison of Artificial Neural Network and Kriging in Spatial Estimation of PM<sub>2.5</sub> Concentration[J]. Geomatics and Information Science of Wuhan University, 2020, 45(10): 1642–1650. DOI: 10.13203/j.whugis20180482(许珊, 邹滨, 王敏, 等. PM<sub>2.5</sub>浓度空间估算的神经网络与克里格方法对比[J]. 武汉大学学报·信息科学版, 2020, 45(10): 1642–1650. DOI: 10.13203/j.whugis20180482)