**引用格式**:陈剑铭,王晴,徐鑫,等. 利用气象和空气污染因素预测呼吸系统疾病死亡的机器学习应用——以北京市海淀区为例[J]. 北京化工大学学报(自然科学版),2025,52(6):1-9.

CHEN JianMing, WANG Qing, XU Xin, et al. Application of machine learning to predict respiratory disease deaths using meteorological and air pollution factors: a case study of Haidian District, Beijing[J]. Journal of Beijing University of Chemical Technology (Natural Science),2025,52(6):1-9.

# 利用气象和空气污染因素预测呼吸系统疾病死亡的机器学习应用
## ——以北京市海淀区为例

陈剑铭[1#]　王　晴[2,3#]　徐　鑫[2,3]　马子昂[2,3]　伯　鑫[2,3]　李　杨[1*]

(1. 北京大学第三医院 病案科, 北京　100191; 2. 北京化工大学 化学工程学院, 北京　100029;
3. 北京化工大学 北化中国工业碳中和研究院, 北京　100029)

**摘　要**: 随着城市化和工业化的发展,日益严峻的空气污染形势和频发的极端天气事件对公共健康构成了重要威胁。本研究旨在评估气象因子及空气污染对呼吸系统疾病死亡的影响。以2014年1月1日至2024年7月31日中国北京市海淀区的气象数据、空气污染物数据和呼吸系统疾病死亡数据为研究数据集,利用随机森林(RF)模型分析气象因素和空气污染物对呼吸系统疾病死亡的影响,并结合SHapley Additive exPlanations (SHAP)开展呼吸系统疾病死亡的影响因素分析。斯皮尔曼相关分析和RF模型结果显示,$SO_2$浓度、$NO_2$浓度、$PM_{2.5}$和$PM_{10}$与呼吸系统疾病死亡呈正相关,最低气温与呼吸系统疾病死亡呈负相关,且该模型在冬季展现出较其他季节更优的预测性能。此外,模型的SHAP全局特征结果表明最低气温是影响呼吸系统疾病死亡的最主要因素。研究结果表明RF模型具有预测呼吸系统疾病死亡的潜力,能够有效结合气象和空气污染数据进行呼吸系统疾病的预测,结合SHAP能够进一步提升机器学习模型的可解释性。本研究可为政策制定者科学制定针对性的空气质量控制措施、极端气温健康预警及季节性呼吸系统疾病防控策略提供有力的支撑。

**关键词**: 空气污染; 气象因素; 随机森林; SHapley Additive exPlanations (SHAP); 呼吸系统疾病死亡; 相关性分析
**中图分类号**: R12　　**DOI**: 10.13543/j.bhxbzr.2025.06.001

## 引　言

近年来,空气污染和极端天气事件频发,对人类健康构成了严重威胁。已有研究表明,空气污染物(如$PM_{2.5}$、臭氧等)和极端气温(如高温和低温)与呼吸系统疾病的发病率和死亡率密切相关[1-2]。在全球范围内,24%的疾病负担和23%的死亡可归因于环境因素;从区域差异来看,发达国家只有17%的死亡可归因于环境因素,而发展中国家则高达25%[3]。美国独立非营利研究机构"健康影响研究所(HEI)"的报告[4]显示,2021年全球因空气污染导致的死亡人数高达810万。空气污染物,尤其是细颗粒物($PM_{2.5}$)和臭氧,能够直接进入人体的呼吸系统,引发炎症、氧化应激和肺功能下降。长期暴露于高浓度的$PM_{2.5}$环境中,会导致慢性阻塞性肺疾病(COPD)、哮喘、支气管炎等呼吸系统疾病的恶化,并显著增加死亡风险[5-7]。有研究发现,$PM_{2.5}$每增加10 μg/m$^3$,呼吸系统疾病死亡率将增加0.48%[8]。

此外,极端天气事件也被证实与呼吸系统疾病死亡风险的增加显著相关。例如,Du等[9]的研究表明,热浪与臭氧复合暴露事件会进一步加剧呼吸系

收稿日期:2025-04-10
基金项目:国家自然科学基金(72174125);中央高校基本科研业务费项目(PT2510);北京大学第三医院交叉联合专项基金(BYSYJC2024034)
#并列第一作者:陈剑铭,男,1967年生,主管技师;
王晴,女,2000年生,硕士生
*通信联系人
E-mail:13466326635@139.com

中国知网　https://www.cnki.net

统疾病死亡风险,尤其是在高温和高污染同时发生的地区。而低温环境也会诱发气道收缩和免疫功能的下降,进一步增加呼吸系统疾病患者的死亡风险。Hu 等[1]的研究表明,短期暴露于低温环境与支气管扩张症患者的死亡率显著相关,尤其是在寒冷季节和北方地区。此外,He 等[10]的研究显示,极端降雨事件与全因死亡、心血管疾病死亡和呼吸系统疾病死亡风险升高有关。本研究团队在临沂的研究也表明,空气污染物、气象因素对呼吸系统疾病具有一定的影响[11-15]。因此,评估气象和空气污染因素与呼吸系统疾病死亡之间的关系,对于制定有效的公共卫生干预措施至关重要。

近年来,机器学习和人工智能在空气质量预测和大气调查领域受到了相当大的关注,并广泛应用于临床决策中来进行空气污染相关疾病的诊断和管理[16]。随机森林(random forest, RF)作为一种强大的集成学习方法,被广泛应用于空气质量预测和健康风险评估领域,它能够处理高维数据、非线性关系以及数据中的噪声和异常值,非常适合用于分析空气污染、气象因素与呼吸系统疾病死亡之间的复杂关系。Kim 等[17]开发了一个基于随机森林的模型,用于估计城市内详细空间单元中与热相关的死亡率。Dowlatabadi 等[18]利用随机森林模型研究了 2017—2021 年空气污染和气象因素对马什哈德呼吸系统疾病死亡率的影响,结果表明空气污染和气象因素与呼吸系统疾病死亡密切相关。

北京市是我国的首都,常具全国示范效应,是我国发展水平最高的城市,具有典型的污染特征,也曾长期面临严重雾霾问题和沙尘暴问题,且城市人口老龄化程度高,呼吸系统疾病负担显著。然而,至今尚未有研究使用机器学习方法探讨北京市气象和空气污染因素对呼吸系统疾病死亡的影响。因此,本研究旨在填补这一空白,探索气象因素和空气污染对北京市呼吸系统疾病死亡的潜在影响。本研究利用 RF 模型分析空气污染物、气象因素和呼吸系统疾病死亡之间的复杂联系。该模型采用北京大学第三医院病案科记录的北京市海淀区 2014 年 1 月 1 日至 2024 年 7 月 31 日的死亡患者数据,并使用模型评估指标(平均绝对误差、均方误差和决定系数)来评估回归模型的性能。研究结果有望为中国城市决策者制定降低呼吸系统疾病死亡率、提高城市居民健康和福祉的干预措施提供理论支持和应用参考。

## 1 数据与方法

### 1.1 数据来源

本研究的呼吸系统疾病每日死亡数据集来源于北京大学第三医院,该医院是北京市三级甲等综合医院。数据集时间范围为 2014 年 1 月 1 日至 2024 年 7 月 31 日。原始死亡数据由医院信息管理人员根据国家三级医院标准进行检查。根据医院数据集的特点,研究区域为北京市海淀区,并根据《疾病和有关健康问题的国际统计分类:第十次修订本(ICD-10)》[19],将直接死因为呼吸系统疾病(ICD-10 编码为 J00-J99)的死亡病例纳入本研究。数据集共包括北京市海淀区 2 034 例呼吸系统疾病死亡病例。

气象数据来自中国气象局北京市海淀站(东经 116.28°,北纬 39.98°),这些数据包括日平均气压、平均气温、最高气温、最低气温、平均相对湿度、最小相对湿度、平均风速、最大风速、最大风速的风向、极大风速、极大风速的风向、累计降水量、日照时数。空气污染物浓度来自中国环境监测总站海淀万柳站(东经 116.29°,北纬 39.96°),包含 2014 年 1 月 1 日至 2024 年 7 月 31 日收集的空气污染物($NO_2$、$SO_2$、CO、$O_3$、$PM_{10}$ 和 $PM_{2.5}$)的日均浓度数据。本研究对原始数据进行了质量控制,包括缺失数据检测、错误检测和可能的错误纠正,以支持数据的适用性和可用性。

### 1.2 斯皮尔曼相关系数

斯皮尔曼相关系数通过将原始数据转换为秩次,然后计算秩次之间的相关系数来衡量变量之间的相关程度。本研究中不同数据集并非正态分布,因此采用斯皮尔曼相关分析探讨空气质量和气象因素对呼吸系统疾病死亡病例的影响。

### 1.3 随机森林模型

随机森林是一种集成学习方法,它通过构建多个决策树来进行预测,在分类和回归任务中都有广泛的应用[20]。在随机森林中,每棵决策树都是独立并在随机选择的子样本上进行训练的,这样可以有效地减少过拟合的风险。随机森林通过将多个决策树的预测结果进行平均或加权平均,从而得到最终的结果[21]。本研究采用随机森林中的回归方法来分析气象因素和空气污染对呼吸系统疾病死亡的影响。已有研究表明,随机森林回归可以有效地

中国知网 https://www.cnki.net

捕获复杂的非线性关系,适用于处理多变量和高维数据[22-23]。同时,RF模型也已成功运用于疾病死亡预测[24-26]。

### 1.4　模型性能评价指标

为了评估RF模型的性能和效果,采用平均绝对误差$E_{MAE}$(MAE)、均方误差$E_{MSE}$(MSE)和决定系数($R^2$)来评估RF模型的预测精度。这些指标的计算方法如下[14]。

$$E_{MAE}=\frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i-y_i| \tag{1}$$

$$E_{MSE}=\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i-y_i)^2 \tag{2}$$

$$R^2=1-\frac{\sum_{i=1}^{n}(\hat{y}_i-y_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y}_i)^2} \tag{3}$$

式中,$i$表示第$i$天;$y_i$代表每天呼吸系统疾病死亡患者的数量;$\bar{y}_i$代表每天死亡患者数量的平均值;$\hat{y}_i$表示死亡患者预测值;$n$表示测试集中的样本数。

MSE是模型稳定性的一个指标;MAE表示预测值和真实值的平均绝对差异;$R^2$衡量模型预测值与真实数据的拟合程度。MSE和MAE值越小,表明模型的稳定性和准确性越高,$R^2$值越接近1表示模型预测得越准确。

## 2　结果与分析

### 2.1　数据特点

2014年1月1日至2024年7月31日,北京大学第三医院共报告北京市海淀区死亡病例8 797例,直接死因为呼吸系统疾病的共2 034例(表1)。如图1(d)所示,呼吸系统疾病死亡人数在2014—2020年间变化不大,2021—2023年间出现了一个剧增情况,其中2022年因呼吸系统疾病死亡人数最多,为298例(占总死亡人数的14.65%)。研究期间呼吸系统疾病男性死亡病例占比(1 219例,59.93%)高于女性死亡病例(815例,40.07%)。呼吸系统疾病死亡病例年龄以80岁及以上为主(65.44%),0~15岁占1.03%,16~59岁占5.51%,60~79岁占28.02%。

在大气污染物方面,$PM_{2.5}$、$NO_2$、CO和$SO_2$的浓度在研究期间呈下降趋势,$PM_{10}$和$O_3$的浓度年际变化起伏较大(图1(a))。在气象参数方面,日平均气温、最高气温、最低气温、平均气压、平均风速、最大风速、极大风速和日照时数在平均水平附近波动,年际变化不大,平均相对湿度、最小相对湿度、最大风速的风向、极大风速的风向、累计降水量在研究期间波动较大(图1(b)、(c))。

表1　呼吸系统死亡病例特征

Table 1　Characteristics of respiratory deaths

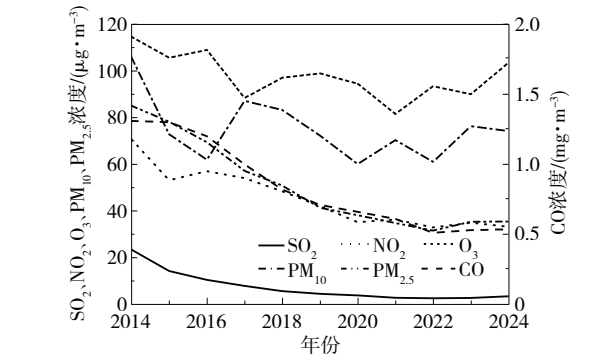
| 年份 | 呼吸系统疾病(J00-J99)死亡人数 | 人数 | | | | | |
|---|---|---|---|---|---|---|---|
| | | 男性 | 女性 | 0~15岁 | 16~59岁 | 60~79岁 | ≥80岁 |
| 2014 | 216 | 116 | 100 | 2 | 21 | 78 | 115 |
| 2015 | 147 | 82 | 65 | 2 | 13 | 57 | 75 |
| 2016 | 156 | 97 | 59 | 1 | 16 | 47 | 92 |
| 2017 | 148 | 91 | 57 | 1 | 10 | 55 | 82 |
| 2018 | 156 | 96 | 60 | 6 | 10 | 49 | 91 |
| 2019 | 174 | 102 | 72 | 2 | 10 | 45 | 117 |
| 2020 | 186 | 111 | 75 | 1 | 2 | 45 | 138 |
| 2021 | 148 | 90 | 58 | 5 | 9 | 38 | 96 |
| 2022 | 298 | 187 | 111 | 0 | 10 | 54 | 234 |
| 2023 | 296 | 185 | 111 | 0 | 8 | 77 | 211 |
| 2024 | 109 | 62 | 47 | 1 | 3 | 25 | 80 |
| 总数 | 2 034 | 1 219 | 815 | 21 | 112 | 570 | 1 331 |

### 2.2　相关分析结果

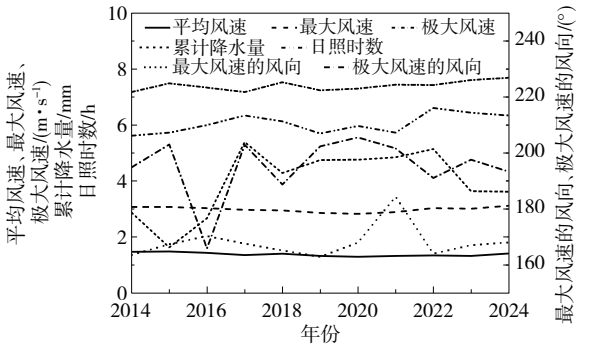
图2为呼吸系统疾病死亡与气象、大气污染因素的斯皮尔曼相关分析结果。在空气质量因素与呼吸系统疾病死亡的相关关系中,$SO_2$、$NO_2$、$PM_{10}$和$PM_{2.5}$浓度与呼吸系统疾病死亡率呈正相关,相关系数分别为0.018、0.013、0.017和0.008。已有研究表明这些空气污染物与呼吸系统疾病的死亡率相关。例如,Liu等[27]发现,在全球652个城市中,环境$PM_{2.5}$和$PM_{10}$的浓度与呼吸系统疾病的每日死亡率密切相关。Huang等[28]的研究发现,长期单独和联合暴露于高浓度$PM_{10}$、$PM_{2.5}$、$SO_2$和$NO_2$与非恶性呼吸系统疾病死亡风险增加显著相关。$NO_2$是典型的交通源大气污染物,主要由化石燃料燃烧和汽车尾气排放产生,流行病学的相关研究说明$NO_2$短期暴露与呼吸系统急性健康效应存在因果关联[27]。Meng等[29]的研究结果提供了短期暴露于$NO_2$与总死亡率、心血管疾病和呼吸道疾病风险增加之间的独立和线性关联的关键证据。

在气象因子与呼吸系统疾病死亡的相关关系中,最低气温、平均气温、最高气温、平均气压和平

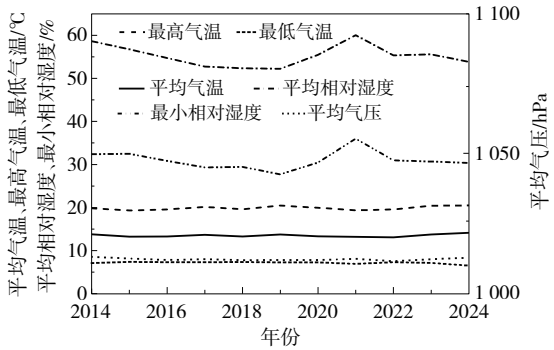
中国知网　https://www.cnki.net



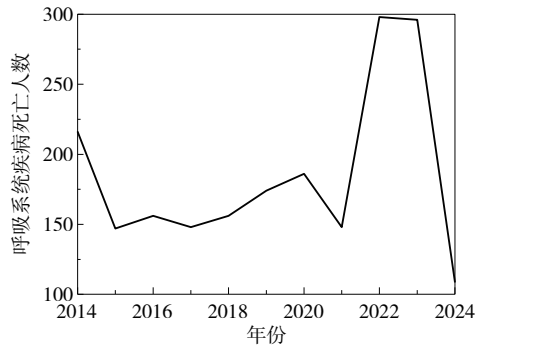

(a) 空气污染物浓度年均变化趋势




(b) 风速、风向等气象参数年均变化趋势




(c) 气温、湿度等气象参数年均变化趋势




(d) 呼吸系统疾病死亡人数年际变化趋势

图1　2014—2024年空气污染物浓度、气象参数和呼吸系统疾病死亡人数年均变化趋势

Fig. 1　The average annual variation trends of air pollutant concentrations, meteorological parameters and the number of deaths from respiratory diseases from 2014 to 2024

均相对湿度与呼吸系统疾病死亡相关性较大,都呈负相关(相关系数分别为-0.088、-0.086、-0.083、-0.069和-0.058)。平均风速和最大风速与呼吸系统疾病死亡呈正相关(相关系数分别为0.022和0.02)。说明呼吸系统疾病死亡与气温、气压、湿度和风速有一定的关系。气温较低时,鼻腔局部血管收缩,鼻腔内局部分泌的免疫球蛋白明显减少,为病毒入侵提供了有利条件[14]。另外,低温冷却黏膜层,寄宿在上呼吸道的病毒更稳定,有利于其进一步扩增进而引起病毒再次感染。He等[30]的研究表明,非适宜的环境温度(尤其是低温)可增加不同感染原因肺炎死亡风险,并带来相当大的疾病负担,而且极低温引起流感相关肺炎死亡的相对风险更高。此外,气温、湿度等气象因素能直接影响病毒飞沫在空气或物体表面的沉降和蒸发过程,从而对病毒在环境中的存活和传播起到关键作用。有研究表明,病毒的传播效率与相对湿度之间存在U形曲线关系[31]。在低湿度环境中,咳嗽或打喷嚏时排出的较小致病性飞沫在空气中停留的时间更长,从而有助于其传播[32]。气压影响呼吸系统疾病死亡的可能原因在于当气压下降时,氧分压会相应减低,导致人体缺氧,从而影响人体的呼吸系统。关于风速与呼吸系统疾病之间联系的研究有限,推测风速主要通过影响空气污染物的浓度来影响呼吸系统疾病死亡。

### 2.3　RF模型结果及模型评价

本文将呼吸系统疾病死亡人数作为回归目标。数据集中80%的数据用于训练,20%用于测试。使用MAE、MSE和$R^2$这3个指标来表征模型性能,结果显示MAE为0.67,MSE为0.99,$R^2$为0.16。图3显示了RF模型估计的测试样本中的死亡人数。模型的预测值越接近真实值,其准确性就越高。这些结果表明,基于RF模型的方法在预测呼吸系统疾病死亡风险方面展现出潜在的应用价值。

在此基础上,进一步研究不同季节下(春季3—5月,夏季6—8月,秋季9—11月,冬季12—次年2月)气象因素和空气污染对呼吸系统疾病死亡的影响。同样使用RF模型,将呼吸系统疾病死亡人数作为回归目标,80%的数据用于训练,20%用于测试,使用MAE、MSE和$R^2$3个指标来表征模型性能。结果发现冬季模型的训练效果相对较好,其MAE为0.97,MSE为3.0,$R^2$为0.24。

2014—2024年冬季呼吸系统疾病死亡人数各

中国知网　https://www.cnki.net

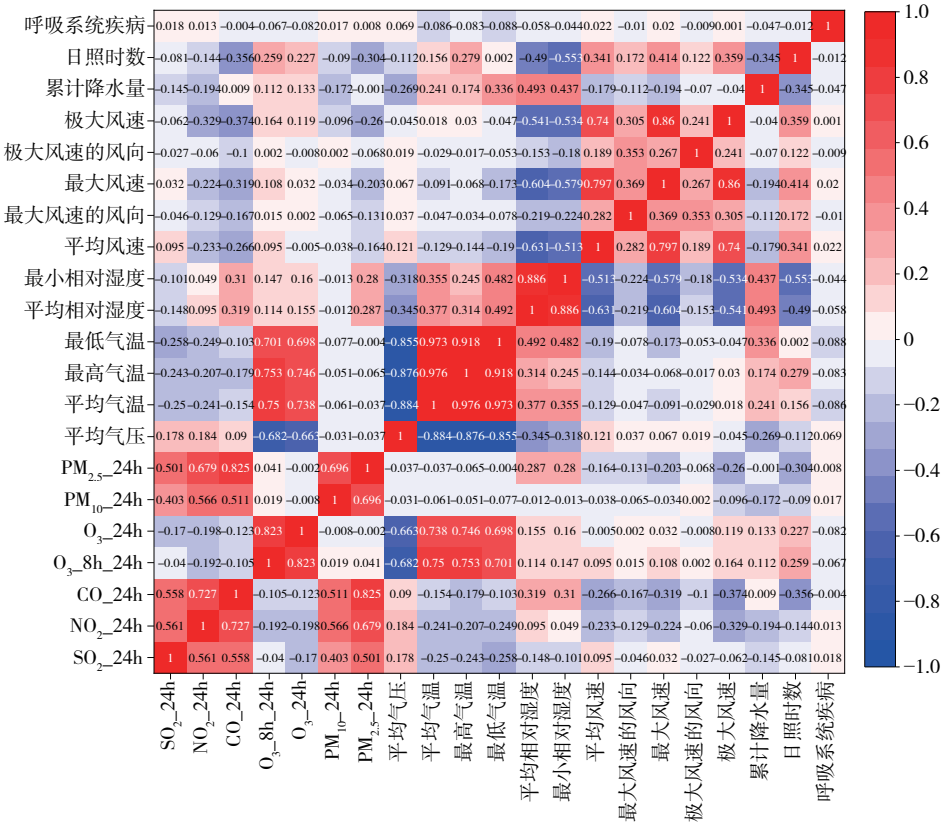



图2　斯皮尔曼相关性结果

Fig. 2　Spearman correlation results

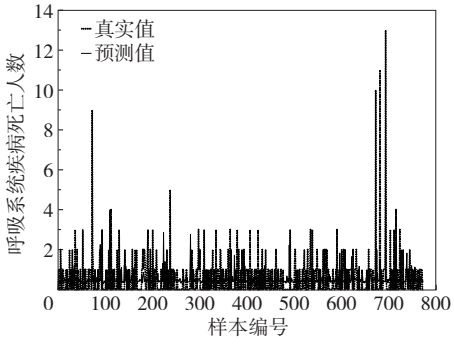



图3　真实值和预测值的散点图

Fig. 3　Scatter plot of true values and predicted values

月统计情况如图4所示,其中2022年12月和2023年1月呼吸系统疾病死亡人数远高于其他月份,而大气污染物和气象因素在这两年的变化却并不明显(图1),说明该现象与环境中的大气污染物和气象因素关系不大。据官方报道显示,全国发热门诊的诊疗量在2022年12月达到高峰,之后呈持续下降趋势,同时,在院的新冠感染者和重症患者数量也呈现出连续下降的趋势[32]。本研究发现的呼吸系统疾病死亡人数剧增的2022年12月和2023年1月正好处于新冠疫情感染高峰期,结合国家流感中心的周报数据显示[33],我国流感样病例在2022年末出现一次爆发高峰,国内流感样病例明显高于往年同期。已有研究表明,同时感染新冠病毒和流感病毒的成年人的重症和死亡风险会大大增加[34]。因此该时段呼吸系统疾病死亡人数的异常剧增的现象可能是由新型冠状病毒(COVID-19)感染大流行主导,并叠加了季节性流感在北方地区秋冬季活跃流行的共同影响所致。这两种呼吸道传染病的叠加流行,显著增加了呼吸系统疾病相关的死亡风险。

## 2.4　模型可解释性分析

为进一步增强机器学习算法模型的可解释性,本研究采用SHapley Additive exPlanations (SHAP)[35-36]开展呼吸系统疾病死亡的影响因素分析,旨在揭示气象因素和空气污染因子对模型预测的影响。模型的SHAP全局特征分析如图5所示,图中横坐标表示SHAP值,纵坐标表示不同特征,红色表示样本特征值大,蓝色表示样本特征值小。结果表明,最低气温是影响呼吸系统疾病死亡的最主要因素,其特

中国知网　https://www.cnki.net

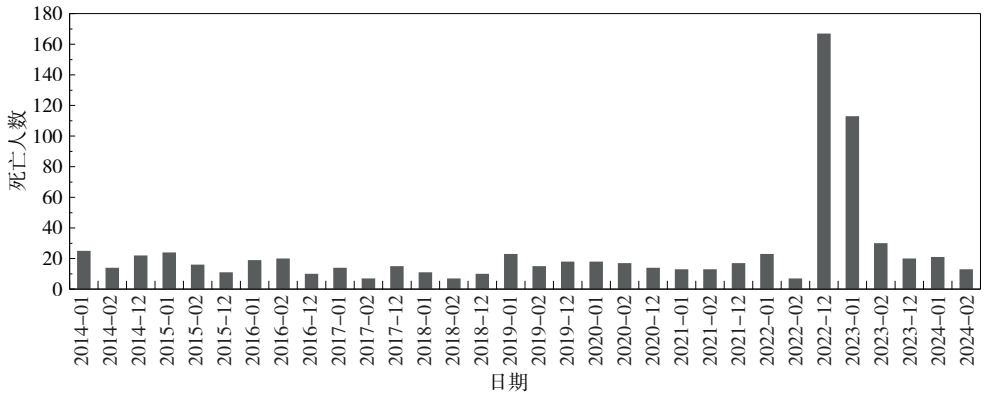



图4　2014—2024年冬季呼吸系统疾病死亡人数各月统计

Fig. 4　Deaths from winter respiratory diseases in 2014—2024 by month

征值越小,模型预测呼吸系统疾病死亡风险越大,与前文相关性分析中得到的最低气温与呼吸系统疾病死亡呈负相关结果相对应。同时,气象因素中,平均气温、平均气压和最高气温是影响呼吸系统疾病死亡的关键因素。这也表明气温相关特征(最低气温、平均气温、最高气温)整体的影响力最强。此外,空气污染因子中CO、$SO_2$污染物的影响力较高,且呈正相关影响。

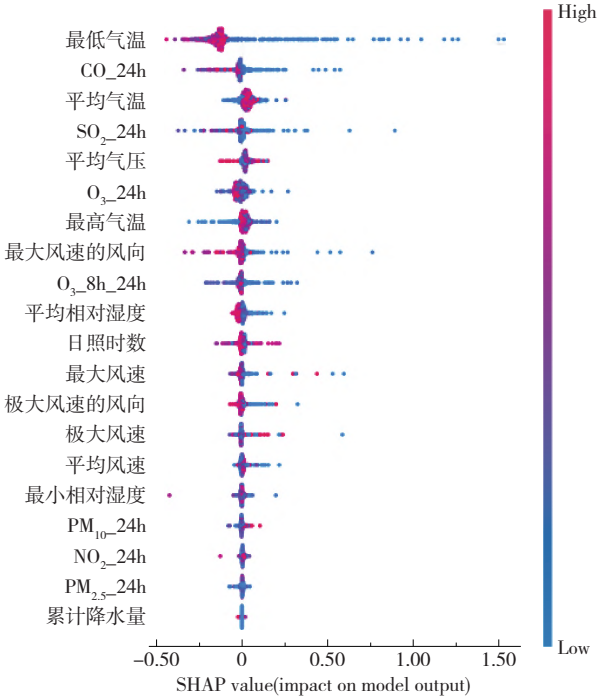



图5　模型SHAP全局特征

Fig. 5　Global SHAP feature analysis

### 2.5　不确定性分析

本研究分析的气象因素和空气污染对呼吸系统疾病死亡的影响存在一定的不确定性,具体如下。

(1)本研究涉及到13个气象和6个空气污染物变量(日平均气压、平均气温、最高气温、最低气温、平均相对湿度、最小相对湿度、平均风速、最大风速、最大风速的风向、极大风速、极大风速的风向、累计降水量、日照时数、CO浓度、$O_3$浓度、$SO_2$浓度、$NO_2$浓度、$PM_{2.5}$和$PM_{10}$),未考虑到其他因素影响,如患者身体素质、生活习惯、免疫力存在个体差异等。

(2)本研究使用的空气污染物浓度数据来源于中国环境监测总站海淀万柳站。站点遵循国家标准进行运维和质量控制,其数据具有官方权威性,但监测值仍可能存在一定程度的误差。本研究使用单个监测站点的测量值来代表海淀区的平均浓度,可能会降低模型的整体精度和泛化能力。

(3)本研究发现呼吸系统疾病死亡与气象和空气污染因素之间存在关联,但无法准确计算这些空气污染物带来的风险,气象和空气污染物影响呼吸系统疾病死亡的复杂机制需要进一步研究。

## 3　结论

本研究结合斯皮尔曼相关分析和RF建模,调查了气象因素、空气污染和呼吸系统疾病(J00–J99)死亡之间的关联。主要结论如下。

(1)在相关性分析中,$SO_2$浓度、$NO_2$浓度、$PM_{10}$和$PM_{2.5}$与呼吸系统疾病死亡呈正相关。对于气象因素,呼吸系统疾病死亡与气温、气压、湿度和风速有一定的关系,其中最低气温与呼吸系统疾病死亡相关性较大,呈负相关。未来需重视空气污染和极端天气(例如低温等)对呼吸系统疾病死亡的影响。此外,气象因素和空气污染物影响呼吸系统疾病死亡的复杂机制需要进一步的研究。

中国知网　https://www.cnki.net

(2)利用RF模型分析气象因素和空气污染对呼吸系统疾病死亡的影响结果表明,基于RF的方法有可能预测呼吸系统疾病的死亡(MAE和MSE分别为0.67和0.99)。结合SHAP方法对预测模型进行解释性分析,发现影响呼吸系统疾病死亡的关键因素是最低气温。在不同季节中,冬季的预测效果相对较好。2022年12月和2023年1月的呼吸系统疾病死亡人数远高于其他月份,但该时间段环境因素、气象因素未发生明显改变,推测该现象与新冠疫情相关。

本文研究结果可为政府卫生机构制定预防和控制措施提供支持。此外,本文研究证明了应用机器学习预测疾病死亡的潜力,并为未来研究其他空气污染物浓度与流行病学数据之间可能存在的关系提供了参考。本文仅关注了区域人群整体水平,未来的研究可进一步收集包含详细人口学特征和健康信息的个体化数据,针对儿童、老年人以及患有慢性呼吸系统疾病、心血管疾病等基础疾病的敏感人群和高危人群分别构建模型,深入探究气象和大气污染因素对这些群体的差异化影响,为精准化健康防护提供更强有力的科学依据。

**参考文献:**


[1] HU S L, XUE X W, XU J Y, et al. Association of short-term exposure to ambient air pollution and temperature with bronchiectasis mortality: a nationwide time-stratified case-crossover study [J]. eBioMedicine, 2024, 110: 105465.

[2] 张建宇. 中国环境健康面临的问题及国外经验借鉴[J]. 行政管理改革, 2017(7): 58-63.
ZHANG J Y. Problems faced by China's environmental health and lessons from foreign experiences [J]. Administration Reform, 2017(7): 58-63. (in Chinese)

[3] World Health Organization. Preventing disease through healthy environments: a global assessment of the burden of disease from environmental risks [R]. Geneva: World Health Organization, 2018.

[4] Health Effects Institute. State of global air 2024. special report [R]. Boston: Health Effects Institute, 2024.

[5] ZHAO Y X, AN X Q, SUN Z B, et al. Identification of health effects of complex air pollution in China [J]. International Journal of Environmental Research and Public Health, 2022, 19: 12652.

[6] 邓希希, 者炜. 气温和主要大气颗粒物对居民死亡影响的研究进展[J]. 疾病预防控制通报, 2024, 39(6): 85-89.
DENG X X, ZHE W. Progress in research on the effects of temperature and major atmospheric particulate matter on mortality in the population [J]. Bulletin of Disease Control & Prevention (China), 2024, 39(6): 85-89. (in Chinese)

[7] 陈亦晨, 李小攀, 孙良红, 等. 大气颗粒物($PM_{2.5}$和$PM_{10}$)与浦东新区居民慢性阻塞性肺疾病死亡的相关性研究 [J]. 中国全科医学, 2019, 22(12): 1419-1425.
CHEN Y C, LI X P, SUN L H, et al. Particulate matters ($PM_{2.5}$/$PM_{10}$) and COPD mortality in Pudong New Area of Shanghai [J]. Chinese General Practice, 2019, 22(12): 1419-1425. (in Chinese)

[8] 顾雯雯, 杨洛贤, 程庆林. 大气$PM_{2.5}$对呼吸系统疾病死亡率影响的Meta分析[J]. 预防医学, 2018, 30(11): 1100-1105, 1111.
GU W W, YANG L X, CHENG Q L. Meta-analysis of the effect of $PM_{2.5}$ on mortality of respiratory diseases[J]. China Preventive Medicine Journal, 2018, 30(11): 1100-1105, 1111. (in Chinese)

[9] DU H, YAN M L, LIU X, et al. Exposure to concurrent heatwaves and ozone pollution and associations with mortality risk: a nationwide study in China [J]. Environmental Health Perspectives, 2024, 132(4): 047012.

[10] HE C, BREITNER-BUSCH S, HUBER V, et al. Rainfall events and daily mortality across 645 global locations: two stage time series analysis [J]. The BMJ, 2024, 387: e080944.

[11] 冯美慧, 伯鑫, 杨静, 等. 2014—2019年临沂市大气颗粒物污染对慢性呼吸道疾病的影响研究[J]. 中国预防医学杂志, 2022, 23(3): 171-179.
FENG M H, BO X, YANG J, et al. Association of ambient particulate matter pollution with chronic respiratory disease in Linyi city in 2014—2019[J]. Chinese Preventive Medicine, 2022, 23(3): 171-179. (in Chinese)

[12] 李洪枚, 伍鹏程, 伯鑫, 等. 临沂市区主要大气污染物的污染特征及其对居民健康的影响[J]. 环境科学学报, 2020, 40(8): 2919-2934.
LI H M, WU P C, BO X, et al. Pollution characterization of major air pollutants and their impacts on resident health in Linyi City [J]. Acta Scientiae Circumstantiae, 2020, 40(8): 2919-2934. (in Chinese)

[13] 伍鹏程. 临沂市主要大气污染物人群健康效应研究[D]. 北京: 首都经济贸易大学, 2018.
WU P C. Study on health effects of main air pollutants in Linyi City [D]. Beijing: Capital University of Economics and Business, 2018. (in Chinese)

[14] YANG J, XU X, MA X T, et al. Application of machine


中国知网 https://www.cnki.net

learning to predict hospital visits for respiratory diseases using meteorological and air pollution factors in Linyi, China [J]. Environmental Science and Pollution Research, 2023, 30: 88431-88443.

[15] WU X, LI D, FENG M H, et al. Effects of air pollutant emission on the prevalence of respiratory and circulatory system diseases in Linyi, China [J]. Environmental Geochemistry and Health, 2021, 43: 4475 - 4491.

[16] MASMOUDI S, ELGHAZEL H, TAIEB D, et al. A machine-learning framework for predicting multiple air pollutants' concentrations via multi-target regression and feature selection [J]. Science of the Total Environment, 2020, 715: 136991.

[17] KIM Y, KIM Y. Explainable heat-related mortality with random forest and SHapley Additive exPlanations (SHAP) models [J]. Sustainable Cities and Society, 2022, 79: 103677.

[18] DOWLATABADI Y, ABADI S, SARKHOSH M, et al. Assessing the impact of meteorological factors and air pollution on respiratory disease mortality rates: a random forest model analysis (2017—2021) [J]. Scientific Reports, 2024, 14: 24535.

[19] 世界卫生组织. 疾病和有关健康问题的国际统计分类:第十次修订本 第一卷[M]. 2版. 董景五,主译. 北京: 人民卫生出版社, 2008.
World Health Organization. International statistical classification of diseases and related health problems: tenth revision, volume 1 [M]. 2nd ed. DONG J W, trans. Beijing: People's Medical Publishing House, 2008. (in Chinese)

[20] BREIMAN L. Random forests [J]. Machine Learning, 2001, 45: 5-32.

[21] TANG K, QIN M, HAN B B, et al. Identifying the influencing factors of soil nitrous acid emissions using random forest model [J]. Atmospheric Environment, 2024, 339: 120875.

[22] 张瑞杰, 周春艳, 陈辉, 等. 基于GF-5B卫星的随机森林模型反演京津冀地区$PM_{2.5}$[J]. 中国环境科学, 2024, 44(11): 5961-5970.
ZHANG R J, ZHOU C Y, CHEN H, et al. Inversion of $PM_{2.5}$ in the Beijing-Tianjin-Hebei region using a random forest model based on GF-5B satellite [J]. China Environmental Science, 2024, 44(11): 5961-5970. (in Chinese)

[23] JING W Q, LIU J Y, QIN Z M, et al. Evaluating and enhancing the applicability of thermal comfort indices in diverse outdoor environments using Bayesian and random forest regression [J]. Energy and Buildings, 2024, 324: 114858.

[24] 叶壮. 基于机器学习方法的脓毒症患者死亡预测分析[J]. 电脑编程技巧与维护, 2024(11): 138-140.
YE Z. Predictive analysis of sepsis death based on machine learning method [J]. Computer Programming Skills & Maintenance, 2024(11): 138-140. (in Chinese)

[25] 陈晓彤, 岑梓熹, 谭静仪, 等. 三种机器学习算法预测心力衰竭死亡的价值研究[J]. 医学信息, 2024, 37(11): 11-15.
CHEN X T, CEN Z X, TAN J Y, et al. Value of three machine learning algorithms in predicting death from heart failure [J]. Journal of Medical Information, 2024, 37(11): 11-15. (in Chinese)

[26] 郭九麟, 程涛, 师庆科. 基于机器学习的急诊消化道出血住院患者死亡风险评估[J]. 中国卫生信息管理杂志, 2023, 20(6):969-976.
GUO J L, CHENG T, SHI Q K. Risk assessment of death for hospitalized patients with emergency gastrointestinal bleeding based on machine learning [J]. Chinese Journal of Health Informatics and Management, 2023, 20(6): 969-976. (in Chinese)

[27] LIU C, CHEN R J, LEI J, et al. Ambient nitrogen dioxide and hospitalizations of full-spectrum respiratory diseases: a national case-crossover study [J]. Environment & Health, 2023, 1: 130-138.

[28] HUANG W Z, ZHOU Y, CHEN X, et al. Individual and joint associations of long-term exposure to air pollutants and cardiopulmonary mortality: a 22-year cohort study in Northern China [J]. The Lancet Regional Health-Western Pacific, 2023, 36: 100776.

[29] MENG X, LIU C, CHEN R J, et al. Short term associations of ambient nitrogen dioxide with daily total, cardiovascular, and respiratory mortality: multilocation analysis in 398 cities [J]. The BMJ, 2021, 372: n534.

[30] HE Q L, LIU Y N, YIN P, et al. Differentiating the impacts of ambient temperature on pneumonia mortality of various infectious causes: a nationwide, individual-level, case-crossover study [J]. eBioMedicine, 2023, 98: 104854.

[31] HE Y C, LIU W J, JIA N, et al. Viral respiratory infections in a rapidly changing climate: the need to prepare for the next pandemic [J]. eBioMedicine, 2023, 93: 104593.

[32] 光明网. 全国发热门诊和急诊高峰已过[EB/OL]. (2023-01-15)[2024-12-20]. https://epaper.gmw.cn/gmrb/html/2023-01/15/nw.D110000gmrb_20230115_2-02.htm.
Guangming Online. The peak of national fever clinics

中国知网 https://www.cnki.net

and emergency departments has passed[EB/OL]. (2023-01-15)[2024-12-20]. https://epaper.gmw.cn/gmrb/html/2023-01/15/nw.D110000gmrb_20230115_2-02.htm. (in Chinese)

[33] 中国国家流感中心. 流感周报[EB/OL]. (2022-12-28)[2024-12-20]. https://ivdc.chinacdc.cn/cnic/zyzx/lgzb/.
Chinese National Influenza Center. Influenza weekly report[EB/OL]. (2022-12-28)[2024-12-20]. https://ivdc.chinacdc.cn/cnic/zyzx/lgzb/. (in Chinese)

[34] SWETS M C, RUSSELL C D, HARRISON E M, et al. SARS-CoV-2 co-infection with influenza viruses, respiratory syncytial virus, or adenoviruses [J]. The Lancet, 2022, 399(10334): 1463-1464.

[35] 陈小昆,左航旭,廖彬, 等. 融合XGBoost与SHAP的冠心病预测及其特征分析模型[J]. 计算机应用研究, 2022, 39(6): 1796-1804.
CHEN X K, ZUO H X, LIAO B, et al. Coronary artery disease prediction and feature analysis model based on XGBoost and SHAP[J]. Application Research of Computers, 2022, 39(6): 1796-1804. (in Chinese)

[36] HAQUE M A, CHEN B, KASHEM A, et al. Hybrid intelligence models for compressive strength prediction of MPC composites and parametric analysis with SHAP algorithm[J]. Materials Today Communications, 2023, 35: 105547.

# Application of machine learning to predict respiratory disease deaths using meteorological and air pollution factors: a case study of Haidian District, Beijing

CHEN JianMing[1#]　WANG Qing[2,3#]　XU Xin[2,3]　MA ZiAng[2,3]　BO Xin[2,3]　LI Yang[1*]

(1. Medical Records Department, Peking University Third Hospital, Beijing 100191;
2. College of Chemical Engineering, Beijing University of Chemical Technology, Beijing 100029;
3. BUCT Institute for Carbon-Neutrality of Chinese Industries, Beijing University of Chemical Technology, Beijing 100029, China)

**Abstract**: With the development of urbanization and industrialization, increasingly severe air pollution and frequent extreme weather events pose a significant threat to public health. The objective of this study was to evaluate the effects of meteorological factors and air pollution on respiratory deaths. The study dataset includes meteorological, air pollutant, and respiratory disease death data from Haidian District, Beijing, China, from January 1, 2014 to July 31, 2024. A random forest (RF) model was used to analyze the effects of meteorological factors and air pollutant levels on respiratory disease mortality, and the factors influencing respiratory disease mortality were analyzed in combination with SHapley Additive exPlanations (SHAP). The results of Spearman correlation analysis and the RF model showed that $SO_2$, $NO_2$, $PM_{2.5}$ and $PM_{10}$ concentrations were positively correlated with respiratory mortality, whereas the minimum temperature was negatively correlated with respiratory mortality. In addition, the model showed better prediction performance in winter than in other seasons. Furthermore, the SHAP global feature results indicate that the minimum temperature is the most significant factor affecting mortality from respiratory diseases. The results show that the RF model has the potential to predict deaths from respiratory diseases, since it can effectively combine meteorological and air pollution data. Combined with SHAP, it can further enhance the interpretability of the machine learning model. This study should provide strong support for policymakers in scientifically formulating targeted air quality control measures, health warnings about extreme temperatures, and prevention and control strategies for seasonal respiratory diseases.

**Key words**: air pollution; meteorological factors; random forest; SHapley Additive exPlanations (SHAP); death from respiratory diseases; correlation analysis

(责任编辑:吴万玲)

中国知网　https://www.cnki.net