

重庆市空气质量影响因素的实证分析

张力芝¹, 凌秀²

¹重庆对外经贸学院数学与计算机学院, 重庆

²荣昌高新技术产业开发区发展促进中心, 重庆

收稿日期: 2024年7月18日; 录用日期: 2024年8月28日; 发布日期: 2024年9月5日

摘要

随着工业化和城市化的快速发展, 空气质量问题日益凸显, 成为影响人们生活质量和可持续发展的因素, 因此, 研究空气质量的影响因素, 对于改善空气质量、保护生态环境具有重要意义。本文旨在探讨影响重庆市空气质量的各个因素, 选取重庆空气质量在线检测分析平台的相关数据, 运用统计学方法进行预处理和深入分析, 将PM_{2.5}、PM₁₀、SO₂、NO₂、CO及O₃作为自变量, 空气质量指数作为因变量, 构建多元线性回归模型, 研究其因素对空气质量的影响, 并识别重庆市近一年来空气质量的主要污染物, 从而提出改善空气质量的针对性建议, 为重庆市进行空气质量的改善提供科学依据和策略。

关键词

空气质量, 影响因素, 多元线性回归分析

Empirical Analysis of Influential Factors of Air Quality in Chongqing Municipality

Lizhi Zhang¹, Xiu Ling²

¹School of Mathematics and Computer Science, Chongqing College of International Business and Economics, Chongqing

²Rongchang High-Tech Industrial Development Zone Development Promotion Center, Chongqing

Received: Jul. 18th, 2024; accepted: Aug. 28th, 2024; published: Sep. 5th, 2024

Abstract

With the rapid development of industrialization and urbanization, air quality issues are becoming

文章引用: 张力芝, 凌秀. 重庆市空气质量影响因素的实证分析[J]. 自然科学, 2024, 12(5): 998-1005.

DOI: 10.12677/ojns.2024.125109

more and more prominent, and have become an important factor affecting people's quality of life and the sustainable development of society. Therefore, it is of great significance to study the influencing factors of air quality in order to improve air quality and protect the ecological environment. The purpose of this paper is to discuss the various factors affecting air quality in Chongqing Municipality, and relevant data from the Chongqing Air Quality Online Detection and Analysis Platform have been selected. And this data was pre-processed and analyzed in depth using statistical methods. In this paper, the $PM_{2.5}$, PM_{10} , SO_2 , NO_2 , CO and O_3 were used as independent variables, and the air quality index was used as the dependent variable to construct a multiple linear regression model to investigate the effects of their factors on air quality, and identify the main pollutants in the air quality of Chongqing Municipality in the past year, so as to put forward targeted recommendations to improve air quality, and provide a scientific basis and strategy for Chongqing Municipality to carry out air quality improvement.

Keywords

Air Quality, Influencing Factors, Multiple Linear Regression Analysis

Copyright © 2024 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

空气质量是衡量城市生活环境质量的重要指标, 对居民健康和生态平衡具有深远的影响。随着经济的高速发展, 工业化的推进和城市化进程的加快, 空气污染问题早已成为全球性热点, 各国都在采取相应措施进行空气质量的改善。重庆, 作为中国西南地区的经济中心, 其独特的地理位置和气候条件为大气污染物的扩散带来了较大的挑战。近年来, 虽然重庆市的空气质量有所改善, 但依旧存在许多问题, 一些污染指标仍然处于比较高的水平。

随着人们对空气质量的重视程度不断提高, 国内众多研究者们时刻关注着空气质量问题, 通过探讨各种方法对该问题进行了相关研究与分析, 力争找到最优方案来改善大气环境。文献[1]选取 2014~2022 年 AQI 作为研究对象, 通过探究了六大污染物、五个气象因子和 14 个经济变量对空气质量的影响, 构建 DT、RF、GBDT 和 XGBoost 模型对 AQI 进行预测, 结果表明: 这四种模型都有较好的预测效果, 且 XGBoost 和 RF 的预测效果最佳。文献[2]针对西北地区大气污染特征与影响因素, 利用克里金插值和核密度等方法分析了 AQI 和六种污染物的时空变化特征, 同时通过构建 PSR 模型和利用皮尔逊相关分析法分别探究了社会经济因素和气象因素对大气污染的影响。文献[3]通过探究空气质量影响因素进行描述性统计分析, 利用 TOPSIS 法对 2019~2021 年重庆市的空气质量进行综合评价, 并建立 BP 神经网络和随机森林模型对空气质量进行预测, 根据结论提出相应的改善建议。文献[4]在重庆市大气污染物分布特征及影响因素分析中, 基于四年的 AQI 和六大污染源, 对六种污染物的月度、季度和年度的变化特征进行描述性统计分析, 建立 AQI 和污染物的回归方程, 分析主要污染物在各个分位点上的变化, 指出 $PM_{2.5}$ 和臭氧是影响重庆市空气质量的主要原因。除了上述方法以外, 还有一些学者提出一些模型与方法来对全国各地空气质量问题进行分析与预测, 如 Bootstrap 方法[5]、深度学习[6] [7]、神经网络[8]-[10]等。

综上所述, 空气质量研究是环境保护和人类健康的重要组成部分, 对于实现可持续发展至关重要, 通过科学研究, 人们可以更好地理解和解决空气污染问题, 为未来创造一个更清洁、更健康的环境。本文通过整理空气质量影响因素的相关文献, 选取 2023 年 7 月至 2024 年 6 月这一年的空气质量数据, 通

通过对空气质量指数和六种污染物数据的预处理和分析, 构建多元线性回归模型, 定量评估近一年内重庆市空气质量的主要污染物, 并针对结论提出相应的建议。

2. 多元线性回归分析

多元回归分析是一种统计技术, 用于研究两个或两个以上的解释变量与一个响应变量之间的关系, 通过将一个变量确定为因变量, 其余一个或多个变量作为自变量, 在相关变量之间建立线性或非线性的数学模型关系, 并利用样本数据进行分析, 简要而言, 就是拟合多个影响因素对一个变量的影响。

2.1. 多元线性回归模型

设因变量(响应变量)为 Y , 影响因变量的 k 个自变量(解释变量)为 X_1, X_2, \dots, X_k , 假设每一个自变量 X 对因变量 Y 的影响是线性的, 那就意味着只要其他自变量保持不变, Y 的均值就会随着自变量 X_i 的变化而均匀变化, 即:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \mu \quad (1)$$

其中, β_0 为无任何影响因素下的空气质量指数, $\beta_1, \beta_2, \dots, \beta_k$ 是待确认的估计值(待估参数), 把 $\beta_0, \beta_1, \beta_2, \dots, \beta_k$ 称为回归参数, μ 为随机误差, 代表模型未解释的部分, 常假定

$$\begin{cases} E(\mu) = 0 \\ \text{Var}(\mu) = \sigma^2 \end{cases} \quad (2)$$

称

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k \quad (3)$$

为理论回归模型。

对于一个实际问题, 若获得 n 组观测数据 $(X_{i1}, X_{i2}, \dots, X_{ik}; Y_i)$ ($i = 1, 2, \dots, n$), 则线性回归模型数学形式可表示为:

$$\begin{cases} Y_1 = \beta_0 + \beta_1 X_{11} + \beta_2 X_{21} + \dots + \beta_k X_{k1} + \mu_1 \\ Y_2 = \beta_0 + \beta_1 X_{12} + \beta_2 X_{22} + \dots + \beta_k X_{k2} + \mu_2 \\ \vdots \\ Y_n = \beta_0 + \beta_1 X_{1n} + \beta_2 X_{2n} + \dots + \beta_k X_{kn} + \mu_n \end{cases} \quad (4)$$

写成矩阵形式, 为

$$Y = X\beta + \mu. \quad (5)$$

2.2. 拟合优度检验

多元线性回归的拟合优度检验主要用于评估模型对观测数据的拟合程度, 即模型解释了多少响应变量的变异。度量拟合优度的统计量通常选用可决系数 R^2 和调整可决系数 \bar{R}^2 , 即

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \quad (6)$$

$$\bar{R}^2 = 1 - \frac{(1 - R^2)(n - 1)}{n - k - 1} \quad (7)$$

其中, $SSE = \sum_{i=1}^n (\hat{y}_i - y_i)^2$ 为残差平方和, $SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$ 为回归平方和, $SST = \sum_{i=1}^n (y_i - \bar{y})^2$ 为总离差平方和。

方和, 且 $SST = SSE + SSR$, n 为样本数, k 为解释变量个数。

可决系数等于回归平方和在总平方和中所占的比率, 即回归方程所能解释的因变量变异性的百分比。 R^2 越大, 说明多元线性回归模型的拟合效果越好, 且 R^2 的最优值为 1, 通常的取值范围为 $[0, 1]$, R^2 越靠近 1, 表明模型中的自变量对因变量的解释能力越强, 模型拟合得也就越好, R^2 越靠近 0, 表明模型拟合得越差, 一般可决系数 < 0.4 , 可看作拟合效果不好。

除了可决系数和调整可决系数以外, 还可使用其它统计量来对模型的拟合优度进行检验, 如 F 检验和 t 检验, 其中 F 检验主要用于检验整个回归模型的显著性, 即所有变量作为一个整体是否对因变量有显著影响, t 检验主要用于检验单个自变量是否对因变量有显著影响。

3. 重庆空气质量影响因素的实证分析

3.1. 指标选取与数据来源

大气污染产生的原因既有自然原因, 也有人类原因, 例如火山喷发、森林火灾、工业废气、焚烧废气和机动车废气等这些因素所排放的污染物。大气污染的主要来源是生活和生产用煤, 其主要污染物是颗粒物和 SO_2 , 随着机动车辆地迅速增加, 大气污染特征正在由烟煤型转变为汽车尾气型, NO_2 和 CO 呈现加重趋势, 且随着近年来温度的逐渐回暖, 臭氧污染潜在的危机也处于随时爆发的状态, 因此, 本文根据已有文献与相应分析, 选取 $PM_{2.5}$ 、 PM_{10} 、 SO_2 、 NO_2 、 CO 及 O_3 作为大气污染的主要污染物[11]。衡量大气污染状况最直接的检测数据就是空气质量指数(AQI), 它是一种用于量化大气污染状态的非线性无量纲指标, 即 AQI 越大, 等级就越高, 表明大气环境的污染程度越高, 对人类身体的伤害也就越大。空气质量指数大小与空气质量的等级如表 1 所示。

Table 1. Comparison of air quality index and air quality class

表 1. 空气质量指数与空气质量等级对照表

空气质量指数	0~50	51~100	101~150	151~200	201~300	>300
空气质量等级	优	良	轻度污染	中度污染	重度污染	严重污染
AQI 级别	一级	二级	三级	四级	五级	六级

因此, 根据《环境空气质量标准》, 本文通过重庆空气质量在线检测分析平台收集整理了 2023 年 7 月 1 日至 2024 年 6 月 30 日这最近一年的重庆市每日空气质量指数(AQI)和大气污染物浓度数据, 其中大气污染物主要包含细颗粒物($PM_{2.5}$)、可吸入颗粒物(PM_{10})、二氧化硫(SO_2)、一氧化碳(CO)、二氧化氮(NO_2)及臭氧(O_3)这六大污染物, 共有 365 条数据, 相应的指标信息展示在表 2 中。

Table 2. AQI and top 6 pollutants meaning and units

表 2. AQI 和 6 大污染物含义及单位

变量	指标	指标含义与单位
X_1	$PM_{2.5}$	日平均浓度($\mu g/m^3$)
X_2	PM_{10}	日平均浓度($\mu g/m^3$)
X_3	SO_2	日平均浓度($\mu g/m^3$)
X_4	CO	日平均浓度(mg/m^3)
X_5	NO_2	日平均浓度($\mu g/m^3$)
X_6	O_3	日平均浓度($\mu g/m^3$)
Y	AQI	日平均值

3.2. 数据预处理

由于本文所用到的数据均来自重庆空气质量在线检测分析平台, 里面存在缺失值以及量纲不统一的情况, 需先对其进行预处理之后才能进行后续的分析。因此, 本文针对缺失数据采用均值替换的方式, 通过往年该月的污染物月平均浓度来对缺失数据进行填充, 针对自变量量纲不统一的情况, 本文采用数据标准化的方式进行处理。数据标准化的处理方式有多种, 如 Z-Score 标准化、最大最小值标准化以及归一化标准化等, 本文采用的是 Z-Score 标准化, 其表达式如下:

$$X'_i = \frac{X_i - \bar{X}_i}{\delta_i} \quad (8)$$

其中 \bar{X}_i 代表第 i 个自变量的均值, δ_i 代表对应的方差。

3.3. 模型的建立

基于以上研究, 本文以 2023 年 7 月至 2024 年 6 月 12 个月 365 天重庆空气质量日数据作为样本, 运用 SPSS 软件对样本数据进行分析, 初步建立模型, 其中 PM_{2.5}、PM₁₀、SO₂、NO₂、CO 及 O₃ 作为自变量, AQI 作为因变量, 建立多元线性回归方程:

$$Y_t = \beta_0 + \beta_1 X'_{1t} + \beta_2 X'_{2t} + \beta_3 X'_{3t} + \beta_4 X'_{4t} + \beta_5 X'_{5t} + \beta_6 X'_{6t} + \mu_t \quad (9)$$

式中, β_0 为无任何影响因素下的空气质量指数, $\beta_1, \beta_2, \dots, \beta_6$ 是待确认的估计值(待估参数), μ_t 为随机扰动项, $X'_{1t}, X'_{2t}, \dots, X'_{6t}$ 分别表示标准化后第 t 时刻 PM_{2.5}、PM₁₀、SO₂、CO、NO₂ 及 O₃ 的浓度数据。

3.4. 软件实现结果分析

据表 3 的数据结果显示, 可决系数 R^2 为 0.887, 调整后的可决系数为 0.885, 表明解释变量能够解释被解释变量 88.5% 的变异程度, 进而说明模型的拟合效果好。

Table 3. Descriptive statistics

表 3. 描述性统计

模型	R	R 方	调整后 R 方	标准估算的错误
1	0.942	0.887	0.885	9.706

Table 4. Analysis of variance

表 4. 方差分析表

模型	平方和	自由度	均方	F	显著性
回归	265342.222	6	44223.704	469.429	<0.001
残差	33820.489	359	94.207	—	—
总计	299162.710	365	—	—	—

表 4 展示了模型整体的显著性检验结果。据分析显示, 检验统计量 F 为 469.429, 显著性(P 值)小于 0.001, 则拒绝原假设, 说明该模型在显著性检验中获得通过, 即至少存在 PM_{2.5}、PM₁₀、SO₂、CO、NO₂ 及 O₃ 中的一个因素对空气质量有显著影响, 拟合效果有效。

表 5 展示了模型参数的显著性检验, 从中可看出, X'_2 和 X'_5 的 P 值均大于 0.05, 没有通过显著性检验, 说明该变量对对空气质量指数 AQI 的影响不显著, 可将其剔除, 且 X_1 与 X_2 的方差扩大因子大于 10,

意味着回归模型中存在多重共线性问题。因此, 综合利用所有变量进行多元回归的效果将会受到影响, 需对全模型进行优化或修正处理, 为得到更优的回归方程, 考虑采用逐步回归的方法。

Table 5. Least squares results for parameter estimations
表 5. 最小二乘法对参数估计结果

模型	未标准化系数		t	显著性	共线性诊断	
	B	标准错误			容差	VIF
(常量)	63.880	0.507	125.910	<0.001	—	—
X ₁ '	24.751	2.373	10.430	<0.001	0.046	21.817
X ₂ '	1.212	2.189	0.554	0.580	0.054	18.558
X ₃ '	-2.552	0.813	-3.140	0.002	0.391	2.559
X ₄ '	3.522	0.971	3.627	<0.001	0.274	3.653
X ₅ '	0.697	0.988	0.706	0.481	0.265	3.779
X ₆ '	17.049	0.626	27.243	<0.001	0.659	1.517

在剔除不显著的变量后, 可以得到经过逐步回归优化后的回归方程, 剔除变量后的描述性分析和参数估计结果如表 6、表 7 和表 8 所示。表 6 展示了优化后的可决系数, 通过拟合模型的可决系数 $R^2 = 0.887$, 调整后的可决系数仍为 0.885 可知, 该优化模型的拟合程度较高, 空气质量指数可以被这四种污染物数据很好地解释。通过表 7 可知, 检验统计量 F 为 705.891, 显著性(P 值)小于 0.001, 则拒绝原假设, 说明该模型在显著性检验中获得通过。表 8 展示了优化后的参数估计结果, 从中可看出, 常量与所有变量均通过了显著性检验, 且通过逐步回归之后不存在多重共线性问题, 因此, 该模型可用于最终的评判分析。

Table 6. Descriptive statistics after optimization
表 6. 优化后的描述性统计

模型	R	R 方	调整后 R 方	标准估算的错误
1	0.942	0.887	0.885	9.692

Table 7. Optimized analysis of variance
表 7. 优化后的方差分析表

模型	平方和	自由度	均方	F	显著性
回归	265249.829	4	66312.457	705.891	<0.001
残差	33912.882	361	93.942	—	—
总计	299162.710	365	—	—	—

Table 8. Optimized least squares results for parameter estimations
表 8. 优化后的最小二乘法对参数估计结果

模型	未标准化系数		t	显著性	共线性诊断	
	B	标准错误			容差	VIF
(常量)	63.880	0.507	126.088	<0.001	—	—

续表

X'_1	26.226	0.922	28.434	<0.001	0.303	3.305
X'_3	-2.159	0.702	-3.075	0.002	0.522	1.916
X'_4	3.508	0.904	3.879	<0.001	0.315	3.178
X'_6	16.972	0.601	28.234	<0.001	0.712	1.404

由表 8 可得出最终的影响重庆市空气质量的多元线性回归模型为

$$Y = 26.226X'_1 - 2.159X'_3 + 3.508X'_4 + 16.972X'_6 + 63.880 \quad (10)$$

根据该模型可知: $PM_{2.5}$ 、 SO_2 、 CO 及 O_3 是造成重庆市大气污染的主要原因, 其中 $PM_{2.5}$ 的影响最为显著, 其次是臭氧, 最后是 CO 和 SO_2 。

4. 结论与建议

4.1. 结论

本文以重庆 2023 年 7 月至 2024 年 6 月的空气质量日平均数据为研究样本, 将空气质量指数 AQI 作为因变量, 六项污染物($PM_{2.5}$, PM_{10} , SO_2 , NO_2 , CO 及 O_3)作为自变量, 初步建立多元线性回归方程, 发现 $PM_{2.5}$ 和 PM_{10} 存在共线性问题, NO_2 对因变量 AQI 的影响不显著, 因此使用逐步回归法, 剔除不显著的变量后, 将 $PM_{2.5}$ 、 SO_2 、 CO 及 O_3 作为自变量重新构建优化后的多元线性回归模型, 结果表明: 近一年内, $PM_{2.5}$ 和臭氧对重庆市空气质量的影响最为显著, 其次是 CO 和 SO_2 。

4.2. 建议

鉴于上述分析结果与重庆市发展现状, 为进一步改善重庆市空气质量问题提出一些相关建议:

第一, 从本文分析结果来看, $PM_{2.5}$ 对空气质量的影响最为显著, 因此, 应加强工业污染治理, 制定严格的排放标准和控制措施, 限制工业排放, 鼓励企业采用清洁能源和环保技术, 减少煤炭和石油产品的消耗, 并且可以在城市和工业区安装高效的空气净化装置, 以捕获和去除 $PM_{2.5}$ 。

第二, 臭氧对空气质量的影响也较为显著, 应采取有效措施去控制臭氧的产生与排放。由于地面臭氧是由氮氧化物(NO_x)和挥发性有机化合物($VOCs$)在阳光作用下发生光化学反应形成的, 因此可控制工业排放, 特别是石油、化工和印刷等行业, 鼓励使用低 $VOCs$ 含量的原料和产品, 推广电动汽车和混合动力汽车, 减少机动车的 NO_x 排放。

第三, CO 和 SO_2 也对空气质量有一定影响, 因此可推广使用清洁燃料, 如天然气和生物燃料, 减少汽油和柴油车辆的一氧化碳排放, 加强对燃煤电厂和工业锅炉的监管, 实施脱硫技术, 发展可再生能源, 如风能、太阳能和水能, 减少化石燃料的使用。

参考文献

- [1] 李佳成, 梁龙跃. 基于机器学习方法的空气质量预测与影响因素识别[J]. 计算机技术与发展, 2024, 34(1): 164-170.
- [2] 杨稳强. 西北地区大气污染特征及其影响因素分析[D]: [硕士学位论文]. 兰州: 西北师范大学, 2022.
- [3] 曾玲. 重庆市空气质量的影响因素分析及预测[D]: [硕士学位论文]. 重庆: 重庆理工大学, 2022.
- [4] 林静怡, 杨宜平. 重庆市大气污染物分布特征及影响因素分析[J]. 重庆工商大学学报(自然科学版), 2020, 37(5): 88-93.
- [5] 李晓童, 夏明月, 林善冬. 基于 Bootstrap 方法对北京空气质量的回归分析[J]. 河北北方学院学报(自然科学版), 2014, 30(4): 31-34.
- [6] 路凯丽, 杨露, 李涛. 基于集成深度学习模型的空气质量指数预测[J]. 南京信息工程大学学报, 2024, 16(1): 56-65.

- [7] 姚玉婷. 基于深度学习的空气质量指数预测[D]: [硕士学位论文]. 淮安: 淮阴工学院, 2022.
- [8] 诸鑫, 林孝先, 刘庆红, 等. 基于神经网络模型的绵阳市空气质量指数预测[J]. 绵阳师范学院学报, 2023, 42(8): 112-118.
- [9] 李乾, 乔栋, 李博文, 等. 基于 T-S 模糊神经网络的空气质量预测模型分析研究[J]. 内蒙古煤炭经济, 2022(17): 142-144.
- [10] 蒋洪迅, 田嘉, 孙彩虹. 面向 PM_{2.5}预测的递归随机森林与多层神经网络集成模型[J]. 系统工程, 2020, 38(5): 14-24.
- [11] 张方卉, 杨志远. 基于多元回归模型的空气质量影响因素实证分析[J]. 景德镇学院学报, 2018, 33(3): 7-9+45.