Statistics and Application 统计学与应用, 2022, 11(3), 653-659
Published Online June 2022 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2022.113070

Hans汉斯

基于数据分析的2019~2020北京市空气质量影响因素分析

吴向莉，李一格，靳　研，吴继垣

北京工业大学，北京

收稿日期：2022年5月29日；录用日期：2022年6月20日；发布日期：2022年6月29日

摘　要

本文通过对收集到的记录有AQI指数与二氧化硫、二氧化氮、PM10、PM2.5、一氧化碳和臭氧浓度的数据进行了描述性分析，并建立多元线性回归模型从而来得到六种物质与空气质量指数之间的关系，为空气质量改善提供学术依据。研究结果“两尘四气”两两变量之间大多具有明显的相关性，其中臭氧对AQI指数升高即空气污染程度增大具有最显著的影响，通过此研究结果本文认为在空气治理时应着重关注臭氧浓度的变化及其升高原因，从而得到更全面的科学治理策略。

关键词

空气质量，多元线性回归，描述性统计

Analysis of Influencing Factors of Air Quality in Beijing from 2019 to 2020 Based on Data Analysis

Xiangli Wu, Yige Li, Yan Jin, Jiyuan Wu

Beijing University of Technology, Beijing

Received: May 29th, 2022; accepted: Jun. 20th, 2022; published: Jun. 29th, 2022

Abstract

In this paper, we make a descriptive analysis of the collected AQI index, sulfur dioxide, nitrogen dioxide, PM10, PM2.5, carbon monoxide and ozone concentration data, and establish a multiple linear regression model to obtain the relationship between six substances and air quality index,

文章引用: 吴向莉, 李一格, 靳研, 吴继垣. 基于数据分析的 2019~2020 北京市空气质量影响因素分析[J]. 统计学与应用, 2022, 11(3): 653-659. DOI: 10.12677/sa.2022.113070

and provide an academic basis for air quality improvement. The results "two dust four gas" has obvious correlation between two variables, including the AQI index of the air pollution degree has the most significant effect, through this study results in this paper that in air management, attention should be paid to the change of ozone concentration and its rise, so as to get a more comprehensive scientific management strategy.

Keywords

Air Quality, Multiple Linear Regression, Descriptive Statistics

1. 引言

随着社会经济的发展，中国仍面临着许多难以解决的问题，环境问题为主要问题之一，环境污染问题对我国构建社会主义和谐社会的事业构成了严重的威胁和挑战，有了影响经济和制约社会的负面影响。空气质量指数(AQI)能够对空气质量进行定量描述，它描述了空气清洁程度或者污染的程度。环保局在计算空气质量时通过六个主要污染标准：二氧化硫、二氧化氮、PM10、PM2.5、一氧化碳和臭氧。AQI 发布时一般都取用 24 小时平均值，它将这六项污染物用统一的标准呈现。

近年来随着人们对空气质量关注度的提高，国内外都有学者就空气质量问题进行了研究并发表相关研究文献，例如 Neha Khanna (2000)采用多种大气污染物的综合评判的方法给出了一种新的空气污染指数(API)体系，并将此空气污染指数(API)与美国环保局(EPA)的污染标准指数(PSI)进行了对比[1]。Indrami Gupta 等(2006)选取印度的 4 个主要城市，分析了 10 年来总悬浮颗粒物和可吸入颗粒物的月平均值变化规律，指出这 4 个城市的总悬浮颗粒物(TSP)没有明显的减少趋势，但是 PM 呈递减或稳定趋势[2]。国内对空气污染的指数研究主要以时间序列为主：赵景波(2004)以北京、兰州、乌鲁木齐等 10 个城市城区空气质量作为研究对象，研究分析 2004 年这 10 个城市的总悬浮颗粒物、二氧化硫、氮氧化物的污染差异和污染状况[3]。鲁然英等(2006)通过分析 2001~2005 年的城市空气质量数据，指出了我国主要城市空气质量的时空分布状况[4]。

在这些文章的基础上本研究拟通过分析 2019~2020 年北京市空气质量数据并对其进行回归建模，探究其变化趋势及空间特征并可以通过其中某几项或一项污染物的浓度变化预测 AQI 的变化，为北京市空气质量改善提出建议。

2. 数据处理及描述性分析

本文所用数据来源于中国 AQI 网站，包含 AQI 及 PM2.5、PM10、二氧化硫、二氧化氮、一氧化碳、O_3_8h 浓度数据，数据选取范围为：2019 年、2020 年两年全年数据[5]。

描述性统计能展示数据最基本的统计特征，下文通过展示各物质的统计学特征展示其 AQI 指数和其余六种物质的变化趋势及分布特征，从而对北京市的空气质量进行初步了解。

依据所收集的数据进行描述性分析和初步处理的步骤为：1) 数据处理，检查缺失值并去除缺失值项；2) 对 AQI 及“两尘四气”进行描述性统计分析；3) 画出空气质量饼图观察两年各空气质量等级的占比；4) 作 AQI、及“两尘四气”随时间变化的时间序列图，观察变化规律；5) 作 AQI 及“两尘四气”两两

之间做相关性分析[6]。

Table 1. Statistical analysis of “Two Dust and Four Gas AQI” description in Beijing from 2019 to 2020
表 1. 北京 2019~2020 年“两尘四气 AQI”描述统计分析

统计量	AQI	PM2.5 (μg/m^3)	PM10 (μg/m^3)	SO_2 (μg/m^3)	NO_2 (μg/m^3)	CO (mg/m^3)	O_3 (μg/m^3)
最小值	18	3	2	2	5	0.1	5.0
最大值	267	217	292	21	101	2.7	283.0
均值	82.81	39.86	62.12	4.03	32.92	0.667	96.218
标准偏差	43.271	32.317	38.490	2.539	16.331	0.3591	57.8948
方差	1872.392	1044.404	1481.489	6.445	266.686	0.129	3351.806
偏度	1.137	1.914	1.564	2.155	1.007	1.608	0.792
峰度	1.310	5.127	3.595	6.141	0.851	4.143	-0.070

表 1 中的标准差和方差的值可以看出各类物质浓度数据的离散程度，NO_2、CO 浓度和 SO_2 浓度的标准偏差和方差很小，说明两者浓度数据的离散程度很小即各物质在随着时间的推移变化，这三种物质浓度的变化不大，且其浓度均值均达标，说明北京市在这三种物质治理上取得了显著成效；AQI、PM2.5、PM10 和 O_3 的标准偏差和方差很大，说明对应数据的离散程度很大即随着时间变化较大，在空气治理时应注重 PM2.5、PM10 和 O_3 增大时原因以及时间特征。AQI 的偏度均大于 0，并且偏度值差值在 1.363 内，说明它们的数据分布呈现是右偏，及直方图中有一条长尾拖在右侧，偏斜程度相当。AQI 及“两尘四气”中除 O_3 以外峰度值均大于 0，为尖顶峰，说明总体数据分布与正态分布相比较为陡峭。而 O_3 的峰度值小于 0，说明 O_3 数据分布与正态分布相比较为平缓[7]。

Table 2. Correlation of substances
表 2. 各物质相关性

	PM2.5	PM10	NO_2	SO_2	O_3	CO
PM2.5	1	0.712**	0.617**	0.467**	0.002	0.861**
PM10		1	0.571**	0.477**	0.049	0.565**
NO_2			1	0.560**	−0.328**	0.684**
SO_2				1	−0.164**	0.583**
O_3					1	−0.093*
CO						1

为了探索影响 AQI 指数的六种因素两两之间线性关系强弱，从而探索六种因素彼此对彼此变化影响的强弱，本文将对其进行相关性分析。表 2 为相关性分析结果，将北京 2019~2020 空气质量数据导入 SPSS 软件，进行变量之间的相关分析，通过此步可以看出两变量之间的相关性，经过“分析－相关－双变量”过程[6]，结果 PM2.5 与 CO 的相关系数为 0.861，说明它们具有极强的正相关；PM2.5 和 PM10、NO_2 和 SO_2 的相关系数分别为 0.712、0.617 和 0.467，说明它们也具有较强的正相关性；PM10 与 NO_2、SO_2 和 CO 的相关系数分别为 0.571、0.477 和 0.565，说明它们具有较强的正相关性；NO_2 与 SO_2、CO 的相关系数分别为 0.560、0.684，说明它们具有较强的正相关性；SO_2 和 CO 的相关系数为 0.583，两者之间也具有较强的相关

性。其它变量间的相关系数小于0.4，说明它们之间相关性很弱[6]。两物质之间存在正相关证明其中一种物质浓度的增大也会在一定程度上使另一种物质的浓度增大，当正相关系数越大，这种影响越明显，反之亦然。

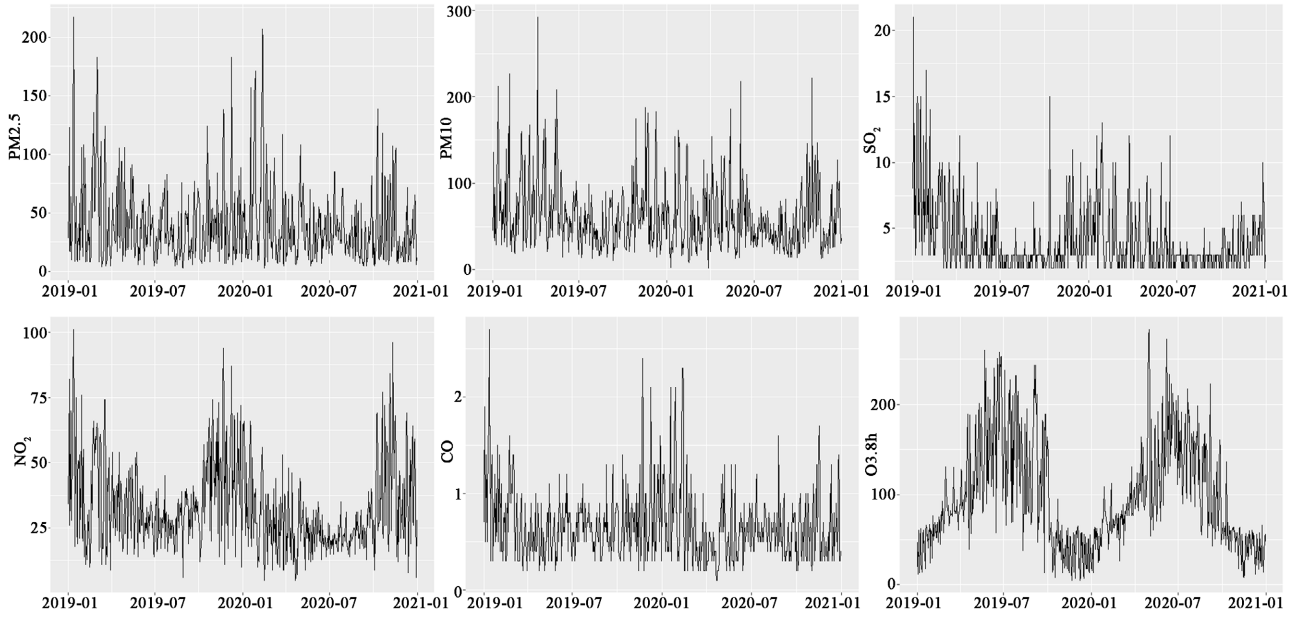

Figure 1. Scatter plot of “Two Dust and Four Gas” changes over time in Beijing
图 1. 北京市“两尘四气”随时间变化的散点图

从图 1 可以看出散点图均没有表现出明显的上升和下降的趋势，因此能够得到北京市“两尘四气”随时间变化均没有明显的线性变化关系。然而，由上可知，图像存在不固定频率的上升和下降，并且有受到季节性因素的影响，即表明“两尘四气”随时间变化具有明显的周期性和季节性[6]。

O_3 随时间的变化表现出强烈的年度季节性，以及周期为 1 年的周期性，且数值冬季较小，夏季较大；PM2.5、PM10、CO、SO_2、NO_2 随时间的变化也都具有明显的季节性，冬季的数值较大，夏季的数值较小，同时，CO、SO_2。NO_2 也具有明显的以一年为周期的周期性。此外，2020 年 PM2.5、PM10 的数值与 2019 年的数值相比有所降低[6]。

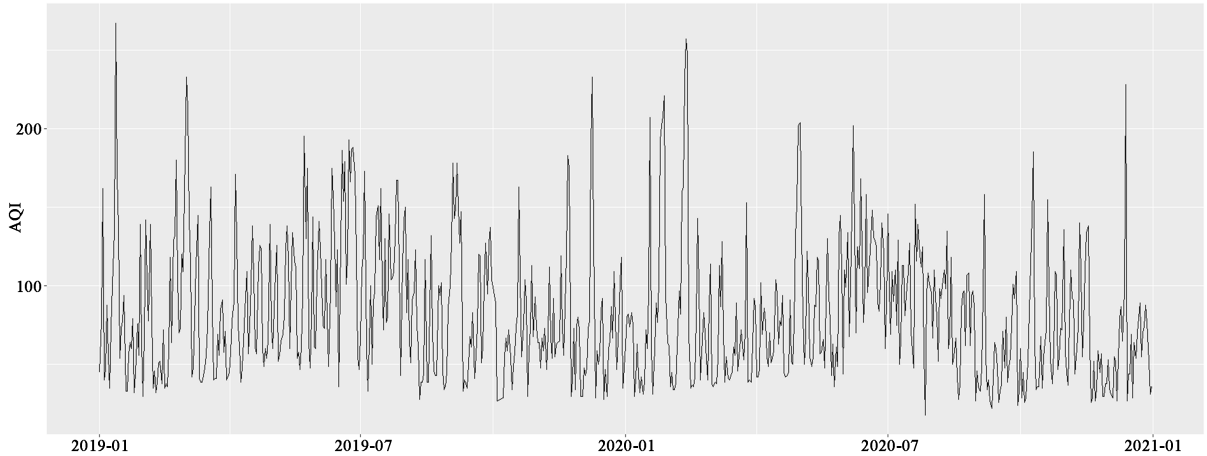

Figure 2. Scatter plot of AQI changes over time in Beijing
图 2. 北京市的 AQI 随时间变化的散点图

图 2 可以看出，AQI 随时间的变化无上升和下降的趋势，但是存在季节性和周期性[8]，且在春冬季波动较大，在空气治理时可以多关注春冬季各物质排放情况。

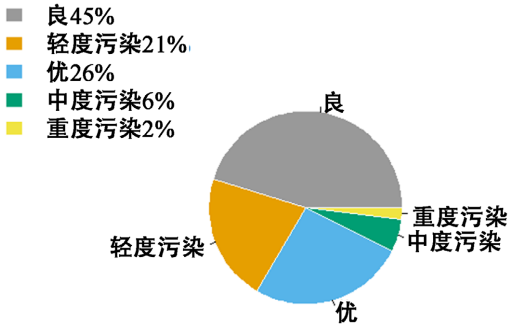

Figure 3. Pie chart of air quality distribution in Beijing
图 3. 北京市的空气质量分布饼图

本文运用饼图(图 3)对数据中的质量等级进行描述从而得到北京市两年总体空气质量等级分布的特征，结果显示空气质量为良的占比最大。其次质量等级为“优”和“轻度污染”占比均在百分之二十五左右。综上所述，北京市空气质量总体处于优良状态，但是空气污染指数为“强度污染”和“中度污染”的天数也较多。在经济持续发展的情况下，北京作为全国“政治中心”和超一线城市，更应该积极响应国家政策，强化绿色优先理念，探索全面科学的策略改善自身空气质量。

3. 模型建立与检验

3.1. 因变量正态性检验，及数据处理

为了探索空气质量指数 AQI 和“两尘四气”浓度之间的关系，本文选用多因素回归分析以 AQI 指数做因变量其余六个指标做自变量进行回归拟合建模。在进行模型拟合之前将原数据分为测试集和训练集，选用 70%的数据作为训练集用来估计模型中的参数，使拟合得出模型能够反映现实，剩下的 30%的数据作为测试集可以对得到的模型进行评估。

通过对数据进行描述性分析认为本文所选择的 6 个自变量都会对 AQI 产生影响，于是考虑所有因素对 AQI 的影响。因为变换后的各变量的数量级差异较大，为了消除变量间的量纲关系，从而使数据具有可比性，所以回归前应用 R 软件中的 scale()语言对数据进行标准化。

3.2. 模型建立

以北京市 2019 年和 2020 年两年的 AQI 数据为因变量，以“两尘四气”含量为自变量，运用 R Studio 中的 lm()语言对数据进行回归分析但在进行回归建模时为了减小数据之间的绝对差异以及数据中部分绝对值的影响，本文对因变量取 log 对数，建立多元回归模型，结果如下表 3：

Table 3. Multiple regression model results table
表 3. 多元回归模型结果表

	回归系数	标准误差	t 统计量	P 值	显著性
常数	4.280600	0.009179	466.356	<2e−16	***
PM2.5 (μg/m^3)	0.096902	0.022060	4.393	1.37e−05	***
PM10 (μg/m^3)	0.151048	0.014945	10.107	< 2e−16	***

Continued

SO_2 (μg/m^3)	−0.040436	0.012522	−3.229	0.00132	**
NO_2 (μg/m^3)	0.117835	0.015383	7.621	1.27e−13	***
CO (mg/m^3)	0.111130	0.021464	5.178	3.26e−07	***
O_3_8h (μg/m^3)	0.287462	0.010593	27.138	< 2e−16	***
MultipleR^2	0.8301				
Adjusted R^2	0.828				
F	407.8				
P	< 2.2e−16				

输出的模型结果如下：

$$\ln(AQI) = 4.281 + 0.097\text{pm2.5} + 0.151\text{pm10} - 0.040SO_2 + 0.118NO_2 + 0.111CO + 0.287O_3\_8h$$

由模型可以看出各物质的回归系数中最大的为 O_3_8h，即在同样增加一单位的浓度的条件下当 O_3_8h 增大时 AQI 变化是最大的，即 O_3_8h 对 AQI 增大的影响是最显著的。

3.3. 模型检验

由上文的结果表明：F 统计量的值为 407.8，P 值 < 2.2e−16 < 0.05，此结果表明：5%的显著性水平下，可以认为所建立的回归方程显著有效，可决定系数 $R^2 = 0.8301$，调整后的 $R^2 = 0.828$，说明方程的拟合结果较好[9]。

此后对模型残差进行正态性检验，由于本文样本量较大所以选择通过 R Studio 中的 ks.test()语言即 K-S 方法对残差的正态性进行正态性检验，且由于这种方法默认的是检验是否服从标准正态，所以在检验前还需对模型残差数据标准化，检验输出结果为 D = 0.055509，P-value = 0.08738，由结果可得 P > 0.05 则残差的正态性检验也是通过的。

对拟合的结果进行 DW 检验(Durbin-Watson test)检验变量是否存在自相关，结果显示 DW = 1.7375，p-value = 0.0009763 < 0.05，说明误差一阶自相关，则在后面我们在 R 软件中用科克伦－奥克特法消除自相关性。后还需对模型进行共线性检验，本文通过 R 软件 car 包中的 vif()函数进行，输出结果如下表 4：

Table 4. Collinearity test results table
表 4. 共线性检验结果表

PM2.5	PM10	SO_2	NO_2	CO	O_3_8h
5.764652	2.645757	1.857355	2.803276	5.457233	1.329169

其结果都<10，说明共线性检验通过，自变量之间并不存在很高的共线性。在建立模型前对因变量作对数变换，通过 Q-Q 图结果认为对数变换后的因变量服从正态分布，从而残差满足正态性假设。

由上述检验结果我们只需对得到的模型用科克伦－奥克特法消除自相关性得到新的模型结果如下表 5。

根据多元线性回归模型的结果，可以从回归系数的大小和显著性看出各自变量对 AQI 的影响程度，可以得出臭氧含量对 AQI 的影响最大，其次为 NO_2 浓度，所以本文建议在治理改善空气质量时应该将重点放在这两种物质的治理上。

Table 5. Coefficient table after eliminating autocorrelation
表 5. 消除自相关性后的系数表

常数	PM2.5	PM10	SO_2	NO_2	CO	O_3.8h
4.280759	0.103368	0.151501	−0.037063	0.118773	0.098079	0.284994

4. 结论与建议

本文将自变量两两之间进行了相关分析发现：PM2.5 与 CO 之间、PM2.5 与 PM10、NO_2 和 SO_2 之间、PM10 与 NO_2、SO_2 和 CO 等之间都具有较强的正相关性，在两两变量相关性分析中还发现某些变量间的相关系数小于 0.4，即这些物质两两之间的相关性较弱。通过对各变量的时间序列图的分析，得到北京市“两尘四气”随时间变化均没有明显的线性变化关系。通过时间序列图可以分析出“两尘四气”随时间变化具有明显的周期性和季节性。

通过以 AQI 为因变量其余六个因素为自变量作回归分析，本文得出臭氧含量对 AQI 的影响最大。随着城市化和工业化发展，NOX 和 VOCs 等污染物排放到大气中导致空气中臭氧浓度升高[7]，且其是在“两尘四气”中对 AQI 影响最大且为正影响的因素，在空气质量改善时应找到臭氧浓度升高的原因，并采用对应管理的办法，在空气治理时应侧重采取科学可行措施降低空气中的臭氧浓度。同时学者也可积极关注研究臭氧含量与空气质量之间的关系，探索防止其浓度过高的方法。

致　谢

本文再次真诚感谢北京工业大学星火项目的支持。

参考文献

[1] Khanna, N. (2000) Measuring Environmental Quality: An Index of Pollution. *Ecological Economics*, **35**, 191-202.

[2] 胡芳芳. 北京市空气污染的空间统计分析[D]: [硕士学位论文]. 北京: 首都经济贸易大学, 2011.

[3] 赵景波, 张琪敏. 2004 年中国典型城市大气污染现状及污染差异[J]. 贵州师范大学学报(自然科学版), 2007, 25(2): 33-36.

[4] 田良, 鲁然英, 邢文听, 等. 2001~2004 年我国城市空气质量研究[J]. 干旱区资源与环境, 2005, 19(z1): 101-105.

[5] 董欢欢, 刘兴荣. 北京市空气质量分析[J]. 资源节约与环保, 2020(2): 43-45.

[6] 张宇玉. 基于逐步回归模型的空气质量数据的校准与统计分析[J]. 黑龙江生态工程职业学院学报, 2019, 34(5): 9-11+30.

[7] 贾俊平. 统计学[M]. 北京: 中国人民出版社, 2018.

[8] Hyndman, R.J. and Athanasopoulos, G. (2018) Forecasting: Principles and Practice. 2nd Edition, OTexts, Melbourne.

[9] 唐年胜, 李会琼. 应用回归分析[M]. 北京: 科学出版社, 2014.