中国 31 个主要城市空气质量的聚类分析和主成分分析

虞 颖，孟彦菊

（云南财经大学 统计与数学学院，昆明 650221）

摘要：基于中国 31 个主要城市空气质量原始数据，进行处理之后得到 $PM_{2.5}$、PM_{10}、SO_2、CO、NO_2、O_3 的月平均浓度以及质量等级优和良的天数，依据这些数据做聚类分析和主成分分析。结果表明：福州、海口、贵阳、昆明、拉萨 5 个城市的空气质量较好，石家庄、太原、济南、郑州、西安 5 个城市的空气质量较差；主成分中 O_3 和 SO_2 的负影响比较大，总体主成分结果是 NO_2 对空气质量的影响最大。基于研究分析结果提出相应建议。

关键词：空气质量；聚类分析；主成分分析

中图分类号：F062.2 **文献标志码**：A **文章编号**：1671－1807(2022)05－0246－05

空气是人类赖以生存的最基本的环境要素之一。近几年来，经济和科技迅速发展，城镇化进程加快，人民生活水平得到了提高[1]。在社会面貌发生可喜变化的同时，空气污染程度越来越严重，空气质量问题在生态领域显得比较突出。空气污染会产生雾霾和酸雨，直接或间接威胁人类的健康，更会影响生态自然的和谐与可持续发展[2]。绿水青山就是金山银山，发展经济不应只局限于眼前的得利，也要考虑到子孙后代的生存与发展，因此要走绿色发展的道路，在发展的同时保护好生态环境。

空气污染与经济发展有着密切的联系，治理空气污染对于经济发展有好处，是生态可持续发展的必经之路[3-4]。空气质量的好坏反映了空气污染的程度，研究空气质量、改善空气质量不仅有利于更好地发展经济，而且对于人类的生命健康和生态自然保护有着重要的意义。同时为国家和政府制定相关改善空气质量的政策提供参考[5]。

1 研究方法

聚类分析就是将研究对象（样品或变量）按照各自特性进行合理分类的一种多元统计方法。目前聚类分析已广泛应用于经济、管理、医学、心理学、气象预报、地质勘探、生物分类等诸多领域。本文基于全国 31 个主要城市空气质量数据进行 K 均值聚类[6]。

主成分分析也称主分量分析，是由 Hotelling 于 1933 年提出的一种常用的多元统计方法。基本思想是用个数较少，但是保留了原始变量大部分信息的几个不相关的综合变量（即主成分）来代替原来较多的变量，从而可以简化数据，对原来复杂的数据关系进行简明有效的统计分析[7-9]。

2 城市空气质量聚类分析

2.1 数据说明

本文使用的数据是全国 31 个主要城市空气质量的数据，来源于 $PM_{2.5}$ 历史数据网站，包含月份、AQI、范围、质量等级、$PM_{2.5}$、PM_{10}、SO_2、CO、NO_2、O_3 和城市这些指标，数据时间跨度为 2013 年 12 月至 2021 年 4 月。

2.2 数据处理

基于收集到的数据，选取月份、质量等级、$PM_{2.5}$、PM_{10}、SO_2、CO、NO_2、O_3 和城市这些指标，分别计算各个城市 $PM_{2.5}$、PM_{10}、SO_2、CO、NO_2、O_3 的月平均浓度，根据质量等级计算质量等级优和良的天数，得到新的数据表，见表 1。

收稿日期：2022-01-28

基金项目：云南省国家社科基金(20BTJ001)。

作者简介：虞颖(1996—)，女，云南大理人，云南财经大学统计与数学学院，硕士研究生，研究方向为宏观经济统计分析；孟彦菊(1979—)，女，河南原阳人，云南财经大学统计与数学学院，教授，博士研究生导师，研究方向为统计理论与方法、宏观经济统计、投入产出方法等。

表 1　全国 31 个主要城市空气质量数据

城市	$PM_{2.5}$ 浓度/($\mu g/m^3$)	PM_{10} 浓度/($\mu g/m^3$)	SO_2 浓度/($\mu g/m^3$)	CO 浓度/($\mu g/m^3$)	NO_2 浓度/($\mu g/m^3$)	O_3 浓度/(mg/m^3)	优和良天数
北京	60.26	86.10	9.29	0.97	42.80	95.11	48
天津	62.51	99.18	21.44	1.23	45.91	92.04	47
石家庄	82.44	143.02	34.46	1.27	49.01	92.90	30
太原	60.61	118.25	46.15	1.30	44.56	86.87	48
呼和浩特	40.42	92.42	26.42	1.27	39.73	86.29	76
沈阳	54.45	92.55	43.30	0.97	41.39	89.66	66
长春	48.52	80.45	23.37	0.87	38.08	83.33	72
哈尔滨	55.25	82.22	29.61	0.92	41.33	73.25	65
上海	42.01	58.21	12.04	0.74	42.26	99.90	83
南京	47.83	84.01	14.74	0.90	44.36	101.30	69
杭州	45.70	74.66	11.52	0.84	43.28	93.79	81
合肥	55.52	82.37	11.93	0.89	39.36	83.35	66
福州	26.49	49.83	6.16	0.71	27.10	87.31	89
南昌	39.65	71.91	15.43	0.92	32.37	84.73	85
济南	68.03	131.29	32.13	1.08	46.42	106.46	34
郑州	69.52	121.47	23.09	1.24	49.43	96.45	36
武汉	56.08	85.94	14.60	1.03	46.37	90.91	60
长沙	54.22	68.51	13.55	0.92	35.49	84.24	69
广州	35.20	56.15	10.88	0.89	45.21	91.93	86
南宁	36.64	62.01	11.54	0.93	31.79	75.17	87
海口	19.63	36.11	5.27	0.62	13.17	74.02	89
重庆	46.64	71.79	12.90	0.96	41.26	68.73	77
成都	55.33	89.26	11.48	0.96	47.45	87.87	64
贵阳	33.39	55.35	14.07	0.70	24.58	73.20	87
昆明	28.10	52.85	14.36	0.87	30.01	81.73	89
拉萨	18.63	48.49	7.73	0.65	20.30	98.61	88
西安	63.37	119.93	18.06	1.36	48.80	82.20	47
兰州	45.28	105.35	19.85	1.21	50.29	89.33	73
西宁	42.45	87.67	24.89	1.35	37.53	82.54	75
银川	42.11	94.87	39.45	1.05	37.01	91.75	71
乌鲁木齐	60.58	108.12	13.93	1.28	47.04	69.00	57

后面进行的聚类分析和主成分分析都是基于数据处理之后得到的。

2.3　系统聚类和 *K* 均值聚类

用 R 软件对全国 31 个主要城市空气质量进行聚类分析[10]。先采用类平均法做系统聚类,分别绘制合并距离为 38 和 35 的两条水平线。

如果取合并距离为 38,则 31 个城市可分为 4 类。

第 1 类:西安、乌鲁木齐。

第 2 类:太原、石家庄、济南、郑州。

第 3 类:南京、合肥、武汉、成都、北京、天津、兰州、沈阳、银川、哈尔滨、长春、呼和浩特、西宁。

第 4 类:长沙、重庆、上海、广州、杭州、南昌、海口、拉萨、南宁、贵阳、福州、昆明。

如果取合并距离为 35,则 31 个城市可分为 5 类。

第 1 类:西安、乌鲁木齐。

第 2 类:太原、石家庄、济南、郑州。

第 3 类:南京、合肥、武汉、成都、北京、天津、兰州、沈阳、银川、哈尔滨、长春、呼和浩特、西宁。

第 4 类:长沙、重庆、上海、广州、杭州、南昌。

第 5 类:海口、拉萨、南宁、贵阳、福州、昆明。

系统聚类结果如图 1 所示。

系统聚类两种情形下除长沙、重庆、上海、广州、杭州、南昌、海口、拉萨、南宁、贵阳、福州、昆明这几个城市外,其余城市的分类相同。合并距离 38 情况下,将这些城市分为一类,合并距离为 35 情况下,则将这些城市分为两类。

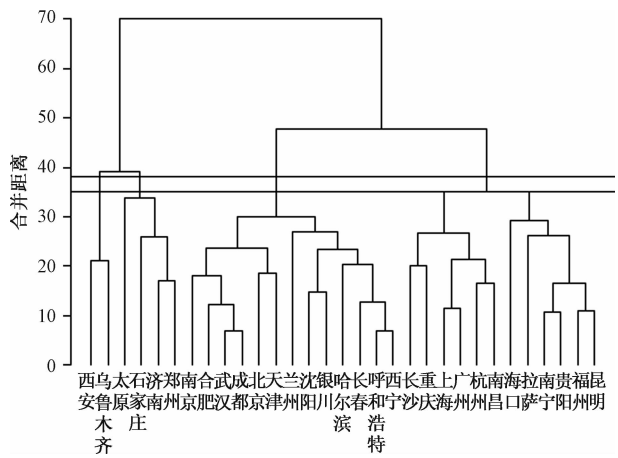

图 1　系统聚类结果

对这 31 个城市空气质量进行 *K* 均值聚类,*K* 取 4 时聚为 4 类[11]。

第 1 类:北京、天津、呼和浩特、沈阳、长春、哈尔滨、南京、合肥、武汉、成都、兰州、西宁、银川、乌鲁木齐。

第 2 类:上海、杭州、南昌、长沙、广州、南宁、重庆。

第 3 类:福州、海口、贵阳、昆明、拉萨。

第 4 类:石家庄、太原、济南、郑州、西安。

聚类结果见表 2。

表 2　聚类分析 *K*=4 聚类结果

北京	天津	呼和浩特	沈阳	长春	哈尔滨	南京	合肥
1	1	1	1	1	1	1	1
武汉	成都	兰州	西宁	银川	乌鲁木齐	上海	杭州
1	1	1	1	1	1	2	2
南昌	长沙	广州	南宁	重庆	福州	海口	贵阳
2	2	2	2	2	3	3	3
昆明	拉萨	石家庄	太原	济南	郑州	西安	
3	3	4	4	4	4	4	

注:类间平方和在总平方和中的占比(between_SS/total_SS)=77.8 %。

K 取 5 时聚 5 类。

第 1 类：呼和浩特、沈阳、长春、哈尔滨、兰州、西宁、银川。

第 2 类：福州、海口、贵阳、昆明、拉萨。

第 3 类：北京、天津、南京、合肥、武汉、成都、乌鲁木齐。

第 4 类：石家庄、太原、济南、郑州、西安。

第 5 类：上海、杭州、南昌、长沙、广州、南宁、重庆。

聚类结果见表 3。

表 3　聚类分析 K=5 聚类结果

呼和浩特	沈阳	长春	哈尔滨	兰州	西宁	银川	福州
1	1	1	1	1	1	1	2
海口	贵阳	昆明	拉萨	北京	天津	南京	合肥
2	2	2	2	3	3	3	3
武汉	成都	乌鲁木齐	石家庄	太原	济南	郑州	西安
3	3	3	4	4	4	4	4
上海	杭州	南昌	长沙	广州	南宁	重庆	
5	5	5	5	5	5	5	

注：类间平方和在总平方和中的占比(between_SS/total_SS)=81.9 %。

K 均值聚类两种情形下除北京、天津、呼和浩特、沈阳、长春、哈尔滨、南京、合肥、武汉、成都、兰州、西宁、银川、乌鲁木齐这几个城市外，其余城市的分类相同。K 取 4 情况下，将这些城市分为一类；K 取 5 情况下，则将这些城市分为两类。

2.4　两种聚类分析的比较

系统聚类和 K 均值分为 4 类的结果略有不同，系统聚类和 K 均值分为 5 类的结果也有差异。

分为 4 类时，无论在系统聚类或者 K 均值聚类情况下，太原、石家庄、济南、郑州这 4 个城市在同一类；南京、合肥、武汉、成都、北京、天津、兰州、沈阳、银川、哈尔滨、长春、呼和浩特、西宁这 13 个城市在同一类；长沙、上海、广州、杭州、南昌、南宁这 6 个城市在同一类；重庆、海口、拉萨、贵阳、福州、昆明这 6 个城市在同一类。

分为 5 类时，无论在系统聚类或者 K 均值聚类情况下，太原、石家庄、济南、郑州这 4 个城市在同一类；南京、合肥、武汉、成都、北京、天津这 6 个城市在同一类；兰州、沈阳、银川、哈尔滨、长春、呼和浩特、西宁这 7 个城市在同一类；长沙、重庆、上海、广州、杭州、南昌这 6 个城市在同一类；海口、拉萨、贵阳、福州、昆明 5 个城市在同一类。

类间平方和在总平方和中的占比越大越好，在做 K 均值聚类分析时该指标可用于确定较优的聚类数 K，可由小到大改变 K 的值，找出使该占比达到最大的 K，K 也不能太大，否则分类太琐碎。在 K 取 4 时，类间平方和在总平方和中的占比为 77.8%；在 K 取 5 时，类间平方和在总平方和中的占比为 81.9%。因此选择将全国 31 个主要城市按空气质量进行分类最好分为 5 类。

3　城市空气质量主成分分析

3.1　多元回归和逐步回归

依据处理之后的全国 31 个主要城市空气质量的数据表先做线性回归分析。回归结果见表 4。表中 x_1、x_2、x_3、x_4、x_5、x_6 分别对应 $PM_{2.5}$、PM_{10}、SO_2、CO、NO_2、O_3 浓度。

表 4　线性回归分析结果

变量	Estimate	Std. Error	t value	Pr(>|t|)	显著性
(Intercept)	142.756	10.736	13.30	1.5×10^{-12}	***
x_1	−1.020	0.154	−6.63	7.3×10^{-7}	***
x_2	−0.243	0.120	−2.02	0.054 88	·
x_3	0.159	0.107	1.48	0.152 54	
x_4	−9.112	9.101	−1.00	0.326 68	
x_5	0.736	0.185	3.98	0.000 55	***
x_6	−0.321	0.110	−2.93	0.007 39	
R^2	0.947				
P 值	3.97×10^{-14}				

从上述输出结果可以看出，回归方程是非常显著的，R^2 为 0.947，模型拟合效果很好，但 x_2、x_3 和 x_4 的回归系数没有通过显著性检验(在 0.05 的显著性水平下)。回归方程为

$$y = 142.756 - 1.020x_1 - 0.243x_2 + 0.159x_3 - 9.112x_4 + 0.736x_5 - 0.321x_6 \quad (1)$$

也可进行逐步回归，逐步回归结果见表 5。

表 5　逐步回归结果

变量	Estimate	Std. Error	t value	Pr(>|t|)	显著性
(Intercept)	135.331 1	7.763 4	17.43	1.7×10^{-15}	***
x_1	−0.947 3	0.135 5	−6.99	2.5×10^{-7}	***
x_2	−0.328 2	0.085 1	−3.85	0.000 72	***
x_3	0.161 4	0.107 2	1.51	0.144 84	
x_5	0.658 4	0.167 6	3.93	0.000 60	***
x_6	−0.262 8	0.093 2	−2.82	0.009 23	**
R^2	0.945				
P 值	6.55×10^{-15}				

从输出结果可见，回归方程显著，系数 x_3 不显著，R^2 为 0.945，模型拟合效果很好，逐步回归所得方程为

$$y = 135.331 - 0.947x_1 - 0.328x_2 + 0.161x_3 + 0.658x_5 - 0.263x_6 \quad (2)$$

3.2　主成分分析和主成分回归

做主成分回归分析,先求样本相关系数矩阵,结果见表 6。

表 6　样本相关系数矩阵

变量	x_1	x_2	x_3	x_4	x_5	x_6	y
x_1	1.000	0.884	0.492	0.678	0.814	0.220	−0.934
x_2	0.884	1.000	0.661	0.842	0.775	0.277	−0.903
x_3	0.492	0.661	1.000	0.555	0.378	0.164	−0.510
x_4	0.678	0.842	0.555	1.000	0.698	0.027	−0.682
x_5	0.814	0.775	0.378	0.698	1.000	0.321	−0.696
x_6	0.220	0.277	0.164	0.027	0.321	1.000	−0.327
y	−0.934	−0.903	−0.510	−0.682	−0.696	−0.327	1.000

表 7　主成分回归结果

变量	Comp. 1	Comp. 2	Comp. 3	Comp. 4	Comp. 5	Comp. 6
Standard deviation	1.955	0.999	0.819	0.533 9	0.426 8	0.206 29
Proportion of Variance	0.637	0.166	0.112	0.047 5	0.030 4	0.007 09
Cumulative Proportion	0.637	0.803	0.915	0.962 5	0.992 9	1.000 00

前 3 个主成分累积贡献率已超过 91%,故选择前 3 个主成分就够了[12]。也可采用矩阵形式来做主成分回归。前 3 个主成分分别为

$$z_1^* = 0.463x_1^* + 0.496x_2^* + 0.352x_3^* + 0.440x_4^* + 0.444x_5^* + 0.160x_6^* \quad (3)$$

$$z_2^* = 0.132x_3^* + 0.307x_4^* - 0.123x_5^* - 0.934x_6^* \quad (4)$$

$$z_3^* = 0.260x_1^* + 0.846x_3^* + 0.435x_5^* - 0.157x_6^* \quad (5)$$

下面计算样本主成分,并将第 1、第 2 和第 3 主成分得分放入数据库的后 3 列,记作 z_1、z_2 和 z_3,再作响应变量 y 关于 3 个主成分 z_1、z_2 和 z_3 的回归分析,结果见表 8。

表 8　主成分回归分析结果

变量	Estimate	Std. Error	t value	Pr(>|t|)	显著性
(Intercept)	68.194	1.450	47.03	$<2\times10^{-16}$	***
z_1	−7.775	0.749	−10.38	6.3×10^{-11}	***
z_2	−2.332	1.758	−1.33	0.196	
z_3	−5.675	2.553	−2.22	0.035	*
R^2	0.809				
P 值	7.52×10^{-10}				

可见,作 y 关于 3 个主成分 z_1、z_2 和 z_3 的回归分析效果理想,回归方程和其中两个回归系数是显著的,R^2 为 0.809,主成分回归方程为

$$y = 68.194 - 7.775z_1 - 2.332z_2 - 5.675z_3 \quad (6)$$

可以利用主成分与原来自变量间的关系 $z^* = \boldsymbol{P}^{\mathrm{T}}x^*$ 将主成分还原为原来的自变量,将主成分 z_1、z_2 和 z_3 还原为原始变量后所得回归方程为

$$y = 132.999\ 7 - 0.292\ 9x_1 - 0.149\ 7x_2 + 0.075\ 9x_3 - 16.484\ 5x_5 - 0.511\ 2x_6 \quad (7)$$

这个回归方程是从主成分回归方程变形而来的,最初回归方程中 x_2、x_3 和 x_4 的回归系数不显著,主成分所得回归方程更为合理。从方程可以看出,x_5 所对应的指标 NO_2 对空气质量的影响最大。

4　降低空气污染的对策建议

4.1　从污染源上进行控制

工业发展的同时造成了空气污染,工业污染物的排放是空气污染的重要原因之一。从污染源上进行控制需要监管工业污染源的排放,控制城市道路施工和扬尘。对于工业高耗能、高污染项目进行评估,严格控制高耗能、高污染项目的实施。对城市道路施工进行管理,降低施工扬尘污染;对道路运输车辆进行管理,严格查处车辆扬尘的现象,加强道路保洁和洒水降尘工作力度。

4.2　控制空气污染物的排放

提倡绿色出行,控制私家车的数量,鼓励公民出门减少私家车的使用,尽量乘坐公共交通或骑自行车。以此降低汽车尾气排放,从而减少相应的氮氧化物、颗粒物和一氧化碳这些空气污染物的排放。加大公共交通的发展,增加地铁、公交线路以提高运载量,对机动车燃气改造,比如出租车和公交车实行油改气、油改电等措施,从根源上减少空气污染物排放。

4.3　加强环保宣传教育

提高民众环保意识,实现对广大民众的环保教育,需要国家与政府鼓励环境教育与环保建设,对

环保工作给予政策上的支持;通过媒体进行形式多样的环保建设宣传,培养公众的环境意识,提高全民参与环保的意识,使人们充分认识到生态环境污染对经济发展、社会稳定和人类生存的危害性。倡导公民从生活中一点一滴的小事做起,切实践行环保。

参考文献

[1] 陈颖,张仲伍.基于聚类分析和主成分分析的城市空气质量评价:以山西省 11 个地级市为例[J].山西师范大学学报(自然科学版),2020,34(4):72-78.

[2] 崔筱笛,郭民之,谭毅恒.全国环保重点城市空气质量状况的聚类分析[J].绿色科技,2020(4):1-4.

[3] 罗国梁.我国主要城市空气质量面板数据聚类分析[J].现代商贸工业,2014,26(7):8-9.

[4] 赵双蕊.聚类主成分分析在城市废气中污染物排放量分析中的应用[J].山西农经,2016(1):57-59.

[5] 郭云飞,林红飞,郑旭.中国城市空气质量指标的聚类分析[J].统计与管理,2016(8):80-81.

[6] 金仁浩,曾国静,王莎.基于聚类分析的北京市空气质量时空分布研究[J].环境保护与循环经济,2021,41(1):68-72.

[7] 刘萍.基于主成分分析的空气质量评价方法研究[J].环境保护与循环经济,2018,38(7):46-52.

[8] 毛宁,李益禛.基于主成分分析的全国主要城市空气质量评价[J].现代商贸工业,2014,26(10):49-50.

[9] 李成,李海波,高丹丹,等.主成分分析在城市大气环境质量评价中的应用[J].湖北大学学报(自然科学版),2016,38(6):567-571.

[10] 陈玉玲.基于实例的系统聚类分析法在环境空气质量评价中的应用[J].环境科学与管理,2010,35(8):159-162.

[11] 张宾,陈永佳.基于聚类和主成分分析的城市空气质量影响因素研究[J].经济师,2017(9):34,39.

[12] 刘海江,张海侠.基于主成分分析法的城市大气环境质量评价[J].中国资源综合利用,2019,37(12):141-143.

Clustering Analysis and Principal Component Analysis of Air Quality in 31 Major Cities in China

YU Ying, MENG Yanju

(School of Statistics and Mathematics, Yunnan University of Finance and Economics, Kunming 650221, China)

Abstract: Based on the original data of air quality in 31 major cities in China, the average monthly concentration of fine particulate matter, inhalable particulate matter, sulfur dioxide, carbon monoxide, nitrogen dioxide, ozone and the days with good quality grade were obtained. Based on these data, cluster analysis and principal component analysis were done. The conclusion is that the air quality of Fuzhou, Haikou, Guiyang, Kunming and Lhasa is high, and in Shijiazhuang, Taiyuan, Jinan, Zhengzhou and Xi'an is low. The negative effect of ozone and sulfur dioxide in the main component is large, and the overall main component result has the greatest effect of nitrogen dioxide on air quality. Corresponding suggestions are proposed based on the results of the research analysis.

Keywords: air quality; cluster analysis; principal component analysis