

基于模糊 C 均值聚类和 Fisher 判别分析的 城市空气质量评价^{*}

尤游

(安徽机电职业技术学院公共基础教学部, 安徽 芜湖 241000)

摘 要:为实时评价城市空气质量等级,以长三角城市群 41 个城市为研究对象,基于样本城市空气质量影响指标,首先利用模糊 C 均值聚类算法对城市空气质量等级进行划分,再通过 Fisher 判别分析法基于训练样本进行回判,由获得的混淆矩阵来衡量聚类分析的可靠性,同时可以利用构建好的判别函数来快速判别非样本城市的城市空气质量等级.仿真结果表明将模糊 C 均值聚类和 Fisher 判别分析用于城市空气质量等级评价可信度较高,其误判率仅为 2.4%,应用该综合模型可以为城市空气质量的防控治理提供参考.

关键词:模糊 C 均值聚类; Fisher 判别分析; 误判率; 空气质量等级; 长三角城市群

中图分类号: TP391; X823 **文献标识码:** A **文章编号:** 1673-2103(2022)05-0040-06

DOI: 10.16393/j.cnki.37-1436/z.2022.05.017

引言

随着工业化、城镇化进程的逐渐加快和经济的高速发展,生态环境的安全问题越来越受到全社会的关注和重视. 2018 年习近平总书记在全国生态环境保护大会上强调要“加快构建生态文明体系”“全面推动绿色发展”“坚决打赢蓝天保卫战是重中之重”. 在整个生态文明体系中,空气质量是最重要的生态指标^[1],空气质量的好坏严重影响城市的知名度和投资竞争力. 近年来,城市雾霾天气频繁出现,已经严重威胁人们的日常生活和身心健康. 为此,构建科学便捷的空气质量评价模型可以帮助实现对城市空气质量的可视化质量预测,为大气污染防治治理提供理论参考依据,同时也能推动城市绿色经济发展和可持续发展,促进生态文明建设^[2-3].

针对城市空气质量评价,目前国内有不少学者已展开研究,并取得了一些成果. 如张茹等^[3]以徐州市为例,分别运用层次分析法和主成分分析法来对比评价城市空气质量;陈颖等^[4]以山西省 11 个地级市为例,基于聚类分析和主成分分析构建城市空气质量评价模型;郑霞等^[5]以长沙市为例,提出一种基于组合赋权集对分析的空气质量评价方法,仿真结果表明该方法评价结果波动性小、稳定性强;候甜甜等^[6]利用主成分分析选取影响空气质量的主要因素,然后进行费希尔(Fisher)判别分析,最终构建主成分的费希尔判别分析模型. 以上文献主要针对某个城市或某个省来讨论,未能覆盖大型的城市群. 随着城市化进程的推进和发展逐渐进入成熟阶段,城市群已成为当今世界城市化与区域发展的新趋势^[7],研究城市群的空气质量对区域经济和环境的协调发展有重要的促进作用.

长三角城市群作为国内最具代表性的城市群之一,研究其区域大气环境,深化城市间联防联控,有利于推动长三角区域的高质量一体化发展. 文章在前人研究的基础上,以长三角城市群为研究对象,选取 6 种污染物(SO₂、NO₂、PM₁₀、CO、O₃、PM_{2.5})浓度和空气质量达到和好于二级的天数比例作为影响城市空气质量

* 收稿日期:2022-06-22

基金项目:2021 年安徽省高校自然科学研究重点项目(KJ2021A1523);2021 年安徽省职业与成人教育学会教育教学研究规划重点课题(Azcyj2021017);2020 年安徽省省级质量工程教学研究项目(2020jyxm0310);2020 年安徽省高校自然科学研究重点项目(KJ2020A1107,KJ2020A1058)

作者简介:尤游(1988—),女,安徽安庆人,讲师,硕士,研究方向:数学建模和多元统计.

的指标. 考虑到待分类对象的不确定性, 首先选用模糊 C 均值聚类(FCM)算法对城市空气质量进行聚类得到分类类别, 但由于分类类别属于离散型变量, 想要构建其与多个连续性自变量之间的关系则需要借助于判别分析^[8], 所以文章进一步引入 Fisher 判别法来构建线性判别函数, 并依据距离判别法进行回判和效果检验, 同时获得其他待判城市的判别结果.

1 资料和方法

1.1 研究区域概况

长三角城市群位于我国华东地区, 由浙江、江苏、安徽和上海三省一市主要的地级市组成, 具体包括浙江省的杭州、宁波、温州、嘉兴、湖州、绍兴、金华、衢州、舟山、台州、丽水, 江苏省的南京、无锡、徐州、常州、苏州、南通、连云港、淮安、盐城、扬州、镇江、泰州、宿迁, 安徽省的合肥、淮北、亳州、宿州、蚌埠、阜阳、淮南、滁州、六安、马鞍山、芜湖、宣城、铜陵、池州、安庆、黄山和上海等 41 个城市.

长三角城市群总人口约 2.2 亿, 该地区经济发展活跃, 制造业发达, 城镇化率高, 创造的 GDP 总量约占全国总值的 25%, 是我国经济发展最活跃的区域之一, 也是我国“一带一路”与长江经济带的重要交汇地带.

2020 年 8 月习近平总书记在扎实推进长三角一体化发展座谈会上强调, 要紧扣“一体化”和“高质量”两个关键词抓好重点工作, 推动长三角一体化发展不断取得成效. 其中长三角一体化发展具体包括经济一体化、科技一体化、设施一体化和生态一体化^[9]. 目前长三角一体化已经上升为国家战略, 伴随着区域经济的快速发展, 引发的大气污染问题也日益突出. 根据《2020 年中国生态环境状况公报》显示, 长三角城市群中有 34 个城市优良天数比例在 80%~100% 之间, 7 个城市优良天数比例在 50%~80% 之间, 平均超标天数比例约为 14.8%. 在全国 168 个主要城市空气质量排名中, 长三角城市群的省会城市杭州、南京、合肥以及上海排名均在 80 名以后, 相比较其他空气优良城市其大气环境还需要进一步的改善^[10].

1.2 变量选取和数据来源

本研究以长三角城市群 41 个城市为研究对象, 选取 X_1, X_2, \dots, X_7 7 个指标, 分别为二氧化硫(SO₂)年平均浓度(ug/m³)、二氧化氮(NO₂)年平均浓度(ug/m³)、可吸入颗粒物(PM₁₀)年平均浓度(ug/m³)、一氧化碳(CO)日均值第 95 百分位浓度(mg/m³)、臭氧(O₃)日最大 8 小时第 90 百分位浓度(ug/m³)、细颗粒物(PM_{2.5})年平均浓度(ug/m³)和空气质量达到和好于二级的天数比例(%).

文中数据来源于 2021 年浙江省、江苏省、安徽省统计年鉴以及中国统计年鉴和相关气象网站(<http://www.tianqihoubao.com/>), 通过查询整理获得长三角城市群 2020 年全年空气质量指标数据. 依据统计结果并结合《2020 年中国生态环境状况公报》可以获得长三角城市群空气质量的总体情况. 41 个城市的空气平均优良天数比例为 85.2%, 而 2020 年全国 168 个地级及以上城市平均优良天数比例为 80.7%, 且 2020 年公布的 168 个城市环境空气质量排名前 20 个城市中包含长三角地区的舟山市、黄山市、丽水市和台州市.

2 相关理论

2.1 模糊 C 均值聚类算法

模糊 C 均值聚类是由 Bezdek 于 1981 年提出的聚类算法, 该算法基于隶属度大小来量化样本属于某个聚类的程度, 进一步优化目标函数获得最小值. 设样本数据集为 $X = \{x_1, x_2, \dots, x_n\}$, 其中每个样本 x_i 对应有 t 个指标属性, 将样本集分为 $s(1 < s < n)$ 类, 则 s 个类的模糊聚类中心为 $V = \{v_1, v_2, \dots, v_s\}$. 令 ω_{ij} 表示样本 x_i 属于类别 v_j 的隶属度, 由此可以结合隶属度、聚类中心和样本集定义聚类目标函数为

$$\Phi_\lambda(\Omega, V) = \sum_{i=1}^n \sum_{j=1}^s \omega_{ij}^\lambda \|x_i - v_j\|^2 \quad (1)$$

其中 $\Omega = \{\omega_{ij}\}_{s \times n}$ 为隶属度矩阵; λ 表示模糊加权因子, 一般认为 $1.15 \leq \lambda \leq 2.15$ 算法效果最好, 常见的 λ 取值为 $2^{[11-12]}$; $\|x_i - v_j\|$ 表示样本 x_i 到聚类中心 v_j 的欧式距离.

模糊 C 均值聚类算法的核心是通过不断迭代获得目标函数的最小值, 从而得到最优的隶属度矩阵 Ω^* 和最佳聚类中心 V^* . 具体计算流程如下^[12-14]:

Step1: 导入样本数据并标准化, 确定聚类个数 s 和模糊加权因子 λ , 设定最大迭代次数 θ_{\max} 和目标函数

的终止阈值 ϵ , 初始化隶属度矩阵 $\Omega^{(0)}$;

$$\text{Step2: 计算聚类中心 } v_j, v_j = \frac{\sum_{i=1}^n (\omega_{ij})^\lambda x_i}{\sum_{i=1}^n (\omega_{ij})^\lambda}, j = 1, 2, \dots, s, \text{ 这里每次新的聚类中心都是通过上一次迭}$$

代的隶属度计算得到并不断循环反复;

Step3: 进一步更新隶属度矩阵 Ω , 其中 $\omega_{ij} = \left[\sum_{p=1}^s \left(\frac{\|x_i - v_j\|}{\|x_i - v_p\|} \right)^{\frac{2}{\lambda-1}} \right]^{-1}$, 从而通过式(1)获得新的目标函数 $\Phi_\lambda(\Omega, V)$;

Step4: 根据给定的终止阈值 ϵ , 判断是否 $\|\Delta W\| \leq \epsilon$ 或者迭代次数超过 θ_{\max} , 如果满足条件则迭代终止, 认为此时算法收敛, 目标函数 $\Phi_\lambda(\Omega, V)$ 达到最优, 可根据最优的 Ω^*, V^* 确定样品的类别; 如果不满足条件则返回到 Step2 继续迭代, 直至满足条件;

Step5: 根据隶属度最大原则输出聚类结果即获得样品类别. 若 $\omega_{ik} = \max_{1 \leq j \leq s} \{\omega_{ij}\}, 1 \leq i \leq n$, 则将第 i 个样品归属于第 k 个聚类.

2.2 Fisher 判别分析

Fisher 判别的核心思想是投影^[14], 试图寻找一个最优投影向量或者最优判别函数, 使得样本数据投影到该方向上, 基于组内离散度尽可能小而组间离散度尽可能大的原则确定判别函数, 再根据判别函数确定样品类别. 假设有 l 个总体 G_1, G_2, \dots, G_l , 观测样本为 $x_{i1}, x_{i2}, \dots, x_{iq_i}$ ($i = 1, 2, \dots, l$), 则样本数据 x_{ij} 的组间离差平方和和组内离差平方和分别为^[14-15]

$$SSA_x = \sum_{i=1}^l q_i (\bar{x}_i - \bar{x})^2, SSE_x = \sum_{i=1}^l \sum_{j=1}^{q_i} (x_{ij} - \bar{x}_i)^2 \quad (2)$$

设投影向量为 p , 则观测样本投影数据为 $y_{ij} = p'x_{ij}$, 该线性关系即为对应的判别函数. 则投影数据 y_{ij} 的组间离差平方和和组内离差平方和分别为

$$SSA_y = \sum_{i=1}^l q_i (\bar{y}_i - \bar{y})^2 = p'SSA_x p, SSE_y = \sum_{i=1}^l \sum_{j=1}^{q_i} (y_{ij} - \bar{y}_i)^2 = p'SSE_x p \quad (3)$$

根据方差分析理论得当目标函数 $f(p) = (p'SSA_x p)/(p'SSE_x p)$ 取得最大值时, 此时得到的投影向量 p 最佳. 为保证解的唯一性^[16], 假定 SSA_x/SSE_x 为单位矩阵 E , 求偏导推出

$$(SSE_x)^{-1} \cdot SSA_x p = \lambda p \quad (4)$$

此时求出 $(SSE_x)^{-1} \cdot SSA_x$ 的最大特征值即为目标函数 $f(p)$ 的最大值, 对应的特征向量即为最佳投影向量, 从而求出线性判别函数为 $y_{ij} = p'x_{ij}$ ($i = 1, 2, \dots, l; j = 1, 2, \dots, q_i$). 依据判别函数依次获得观测样本的投影矩阵 Y 和对应的组均值投影矩阵 \bar{Y} , 基于距离判别法向量间距离最小原则可获得样本判别结果.

2.3 判别效果检验

同时为了验证上述判别函数是否合理, 需要进行效果检验. 这里判别分析具有回判功能, 基于训练样本通过回代求解出判别分析对应的混淆矩阵, 由该矩阵可以读出每个类别的样品正确判别个数和误判个数. 假设聚类 v_i 中样品个数为 N_i , 误判个数为 ρ_i , 则误判率^[15] $\eta_i = \frac{\rho_i}{N_i} \times 100\%$, 一般以误判率来衡量判别函数的优良性和聚类分析的可靠性.

3 MATLAB 仿真实验结果分析

3.1 FCM 聚类结果分析

基于长三角 41 个城市样本数据, 利用 MATLAB2016 进行模糊聚类. 这里确定 3 个聚类, 设定模糊加权因子为 2, 最大迭代次数为 100, 目标函数的终止阈值为 10^{-5} . 经过 23 次迭代后目标函数获得最小值, 根据隶属度最大原则得到 FCM 聚类结果如表 1 所示.

表 1 41 个城市模糊 C 均值聚类结果

类别	城市
第一类	宁波、温州、绍兴、金华、衢州、舟山、台州、丽水、盐城、宣城、安庆、黄山
第二类	上海、杭州、嘉兴、湖州、南京、无锡、常州、苏州、南通、连云港、扬州、镇江、泰州、合肥、滁州、六安、马鞍山、芜湖、铜陵、池州
第三类	徐州、淮安、宿迁、淮北、亳州、宿州、蚌埠、阜阳、淮南

根据《2020 中国生态环境状况公报》公布的全国 168 个城市空气质量排名, 其中前 20 名中有舟山、黄山、丽水、台州 4 个城市全部聚类为第一类, 说明第一类空气质量最好; 再根据各省发布的 2020 年生态环境状况公报可知, 第三类空气质量较差, 第二类居中, 聚类结果符合实际。

3.2 Fisher 判别分析

基于上述聚类结果, 接下来依据 41 个城市的空气质量指标数据进行回判并对待判城市进行聚类评价。得到的判别式函数分别为

$$y_1 = 0.0059x_1 - 0.0112x_2 + 0.0172x_3 + 0.999x_4 - 0.0051x_5 + 0.000863x_6 - 0.0392x_7$$

$$y_2 = 0.0101x_1 + 0.0195x_2 - 0.0102x_3 + 0.9997x_4 + 0.0051x_5 + 0.0042x_6 + 0.0058x_7$$

类均值投影矩阵代表 3 个类的类中心位置, 如表 2 所示。由程序运行结果可以读出两个判别式的贡献率分别为 83.14% 和 16.86%, 且由表 2 混淆矩阵可看出训练样本中第一类的 12 个城市和第二类的 20 个城市均得到正确判别, 第三类的 9 个城市中仅有 1 个城市错判, 即第三类的“淮安市”误判到第二类, 回判综合正确率为 97.6%, 误判率仅为 2.4%。由此可见模糊聚类结果可信度较高。

表 2 判别分析对应的混淆矩阵和投影矩阵

名称	混淆矩阵			类均值投影矩阵	
	From/To	第一类	第二类	第三类	判别式 1
第一类	12	0	0	-2.9799	2.3679
第二类	0	20	0	-2.3821	2.6889
第三类	0	1	8	-1.5691	2.3709

另外从三类的 2 个判别式得分绘制出的散点图来看, 3 个类别的分离效果较好, 具体如图 1 所示, 进一步验证样本城市空气质量等级分类结果是合理的。

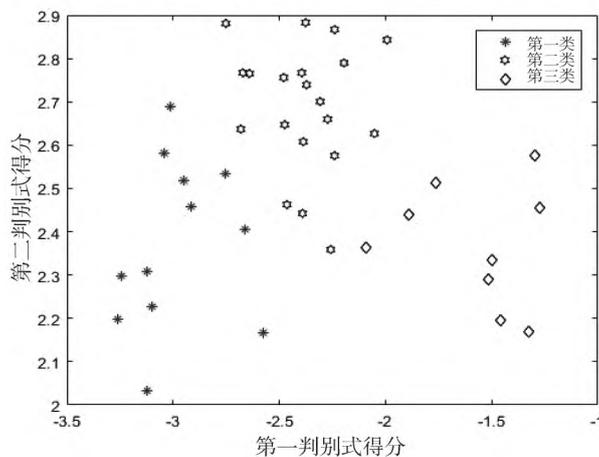


图 1 判别式得分对应的散点图

3.3 待判城市结果分析

这里选取太原、福州、南昌、济南、郑州、武汉、长沙和昆明 8 个城市为待判城市, 分别导入空气质量指标

数据,通过代入判别函数获得投影数据矩阵,再根据 3 个类的类均值投影矩阵,分别计算对应的欧氏距离,根据距离最小原则就能判别 8 个城市的类别。

具体计算结果如表 3 所示,这里 d_i 表示待判城市的投影向量与第 i 类类中心的欧氏距离. 从而获得 8 个待判城市的类别结果依次是第三类、第一类、第二类、第三类、第三类、第二类、第二类和第一类,该判别结果与生态环境部公布的全国城市空气质量排名相吻合. 其中福州市和昆明市属于空气质量排名前 20 个城市,判为第一类,空气质量最好;太原和济南属于空气质量排名后 20 个城市,属于第三类,空气质量最差. 根据上述判别式函数,也可以判别其他城市的类别,以实现城市空气质量的精准防控。

表 3 待判城市判别结果

待判城市	投影数据 y_1	投影数据 y_2	欧氏距离 d_1	欧氏距离 d_2	欧氏距离 d_3	判别结果
太原	-0.280 4	3.411 2	2.894 1	2.222 4	1.656 2	第三类
福州	-3.195 9	2.291 4	0.229 2	0.905 7	1.628 8	第一类
南昌	-2.594 9	2.485 2	0.402 5	0.294 5	1.032 1	第二类
济南	-0.525 3	3.058 4	2.549 9	1.893 2	1.249 9	第三类
郑州	-0.894 1	2.902 0	2.153 1	1.503 2	0.858 9	第三类
武汉	-2.212 9	2.802 5	0.881 6	0.203 8	0.775 1	第二类
长沙	-2.266 2	2.715 3	0.793 8	0.118 9	0.777 5	第二类
昆明	-3.158 5	2.392 6	0.180 3	0.831 0	1.589 5	第一类

4 结束语

由于样本城市空气质量类属的不确定性,模糊 C 均值聚类算法可以基于隶属度的大小快速的对样本城市进行空气质量等级归类,缺点在于不能评价聚类结果的优劣性,且不能对非样本城市进行聚类^[17]. 所以文中引入 Fisher 判别法,在聚类分析的基础上,依据判别式函数来评判聚类分析的可靠性,进一步判别待判城市的空气质量等级,两者结合对城市空气质量进行判别,可以提高空气质量分类评价的准确性. 文中通过收集 2020 年长三角地区 41 个城市的空气质量相关指标数据,对 41 个样本城市的空气质量进行等级评价,并对照《2020 中国生态环境状况公报》进行类比,分析验证其评价结果合理,具有一定的参考价值. 以此可以对其他待判城市进行快速判别,该模型有利于提高环保部门对大气污染的风险信息研判和预警能力。

参考文献:

- [1]王娜娜. 随机森林模型在北京市首要污染物研究中的应用[D]. 北京:北京工业大学,2019:1-4.
- [2]薛士琼. 基于 BP 神经网络的空气质量预测及可视化的实现[D]. 天津:天津大学,2015:1-3.
- [3]张茹,张学杨,陆洪光,等. 基于层次分析和主成分分析的城市空气质量评价——以徐州市为例[J]. 安全与环境工程,2017,24(3):103-107.
- [4]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.
- [5]郑霞,胡东滨,李权,等. 基于组合赋权集对分析的空气质量评价——以长沙市为例[J]. 安全与环境工程,2021,28(1):226-232.
- [6]侯甜甜,户亚慈. 基于主成分判别分析的全国主要城市空气质量评价[J]. 平顶山学院学报,2020,35(5):16-21.
- [7]张国兴,温俊娜,林伟纯,等. 城市群建设改善还是恶化了城市空气质量? ——基于双重差分模型的实证检验[J]. 系统工程理论与实践,2022,42(5):1245-1259.
- [8]李志,翁克瑞,杨娟,等. 基于 FCM-Fisher 判别分析的难采储量分类[J]. 科技管理研究,2013,33(1):241-248.
- [9]曾刚. 长三角城市协同发展能力评价及其区域一体化深化路径研究[J]. 华东师范大学学报(哲学社会科学版),2021,53(5):226-236+242.
- [10]肖严华,侯伶俐,毛源远. 经济增长、城镇化与空气污染——基于长三角城市群的实证研究[J]. 上海经济研究,2021(9):

57-69.

- [11] Pal N R, Bezdek J C. On cluster validity for the fuzzy C-mean model[J]. IEEE Transactions on Fuzzy Systems, 1995, 3(3):370-379.
- [12] 殷士勇. 基于模糊 c-均值与核 Fisher 判别分析的不平衡数据分类方法[J]. 武汉大学学报(工学版), 2014, 47(6):849-853.
- [13] 周贵宝. 基于空间插值和模糊 C 均值聚类的沥青路面气候分区研究[J]. 武汉理工大学学报(交通科学与工程版), 2021, 45(4):793-798.
- [14] 谢中华. MATLAB 统计分析与应用:40 个案例分析(第 2 版)[M]. 北京:北京航空航天大学出版社, 2015:325-378.
- [15] 赵伟, 刘启蒙, 柴辉婵, 等. 基于 Fisher 判别法和质心距理论的突水水源识别[J]. 科学技术与工程, 2020, 20(9):3552-3556.
- [16] 陈恋, 袁梅, 向维, 等. PCA-Fisher 判别模型在煤层底板突水预测中的应用[J]. 数学的实践与认识, 2021, 51(6):103-111.
- [17] 常丽娜, 王颖俐, 王瑶. 基于 K-均值聚类和贝叶斯判别的城市空气质量等级分类及预测[J]. 太原师范学院学报(自然科学版), 2021, 20(2):41-46.

Urban Air Quality Assessment Based on Fuzzy C-means Clustering and Fisher Discriminant Analysis

YOU You

(Public Basic Teaching Department, Anhui Technical College of Mechanical and Electrical Engineering, Wuhu Anhui 241000, China)

Abstract: In order to evaluate the urban air quality grade in real time, this paper studies 41 cities in the urban agglomeration of the Yangtze River Delta. Based on the air quality impact indicators of the sample cities, firstly, the air quality grade is divided by using Fuzzy C-means Clustering Algorithm, and then, Fisher Discriminant Analysis Method is used to re-judge. The obtained confusion matrix measures the reliability of the clustering analysis. At the same time, the urban air quality grade of non-sample cities is quickly discriminated by using the constructed discriminant function. The simulation results show that Fuzzy C-means Clustering and Fisher Discriminant Analysis for urban air quality grade evaluation are of high reliability, and the misjudgment rate is only 2.4%, which can provide reference for urban air quality prevention and control.

Key words: Fuzzy C-means Clustering; Fisher Discriminant Analysis; misjudgment rate; air quality grade; the urban agglomeration of the Yangtze River Delta

(责任编辑:王晓知)