

系统聚类方法对大气污染地区的划分

高胜云,王拥兵*,张丽霞

(安庆师范大学数学与计算科学学院,安徽 安庆 246133)

摘要:当前国内经济快速发展的同时带来了环境质量下降等问题,地区间发展落脚点不同,受到的大气污染程度不同。从主要的大气污染监控物PM_{2.5}、PM₁₀、一氧化碳、二氧化碳、臭氧和二氧化硫6个方面构建大气污染地区分类的聚类评价指标体系,建立地区分类的聚类分析模型。并以北京、上海、海口、西安、佛山、酒泉、拉萨和重庆8个城市为例,应用评价指标的数据进行聚类分析,根据聚类结果提出具有针对性的环境保护建议。

关键词:大气污染;地区分类;系统聚类分析;环境保护

中图分类号:X823 **文献标志码:**A **文章编号:**1673-1891(2019)02-0070-04

Division of Air Pollution Areas by Systematic Cluster Method

GAO Shengyun, WANG Yongbing, ZHANG Lixia

(School of Mathematics and Computational Sciences, Anqing Normal University, Anqing, Anhui 246133)

Abstract: The present rapid development of the domestic economy brings about problems such as the decline of environmental quality. The regional difference in the development focus leads to different degree of regional atmospheric pollution. Based on the six major atmospheric pollutants of PM_{2.5}, PM₁₀, carbon monoxide, carbon dioxide, ozone and sulfur dioxide, we construct an evaluation index system for air pollution area classification and build a cluster analysis model. We employ the data of the evaluation index to make a cluster analysis of eight cities: Beijing, Shanghai, Haikou, Xi'an, Foshan, Jiuquan, Lhasa and Chongqing, and then propose targeted environmental protection solutions based on the cluster analysis results.

Keywords: air pollution; regional classification; systematic cluster analysis; environmental protection.

0 引言

随着第三次工业革命的快速发展,环境保护与大气污染成为了热门话题^[1]。在我国不同地区的发展速度和发展重点不相同,大气污染物含量造成的环境影响存在差异。从国内的大气污染现状来看,国内各城市的大气污染呈现较为明显的复合型污染特征^[2]。目前全国各地环境监测站的常规监测就有二氧化硫、氮氧化物、一氧化碳、可吸入颗粒物和细颗粒物等物质^[3]。严峻的环境污染现状已经逐渐成为制约地区经济可持续发展的主要障碍,过度排放的工业“三废”、生活垃圾、汽车尾气和化肥农药等污染物使得环境日趋恶化。

聚类分析方法是统计学理论中重要的方法之一,是按照某一标准来对样本点进行分类,从而能够对样本数据进行分析判断,提供针对性的意见和决策。文献[4-6]中举例分析了多种聚类方法,并对

它们的优缺点进行比较,当样本的性质不同时,不同的聚类方法对聚类结果产生的影响不同。为分析大气污染物的地区划分问题,需要提取各地区常规监测物质数据。本文收集了8个城市的六种大气污染物的含量,从污染物含量的数据上来看,直观的数据无法分析大气污染物的地区划分问题。因此,本文利用系统聚类法对大气污染物的含量进行地区分类,科学有效的解决污染地区的划分问题。

1 系统聚类分析

变量的分类对应着聚类模型的应用,聚类分析模型利用不同的分类方法,试图寻找变量间存在的内在关系,找到某些统计量来度量两两变量或者多个变量之间的关联性,并将统计量作为变量划分的依据^[7]。

聚类分析方法可分为系统聚类法、有序样品聚

类法、动态聚类法、模糊聚类法等^[8]。本文通过构建合适的聚类分析模型,结合实例进行分析。

系统聚类的过程解释是,平面内有随机的 n 个点,从左往右,依次将 n 个点记为 $x_1, x_2, \dots, x_n$ 。根据模型的计算方法,将随机点视为样本,首先聚类分析系统将自动生成两个异于样本点的聚类点,不妨分别记为A聚类点和B聚类点;其次,分别计算 n 个样本到随机生成点之间的距离,聚类分析通常将距离作为归类的依据,就某一个样本点,将它划分到距离它较近的一类中,这样就将一个样本进行了归类,依次计算下去,直到所有样本点都划分到类中;最后,平面内只留下了A类和B类。在完成初次归类后,聚类系统仍会随机生成两个点,不妨记为C聚类点和D聚类点,重复上述步骤,将继续计算A聚类点与B聚类点中已有样本点到C聚类点和D聚类点的距离。相同的步骤计算下去,直至所有样本点都聚为一类。

图1是聚类分析方法的简要示意图,下面将从定量的角度阐明聚类分析的具体过程。

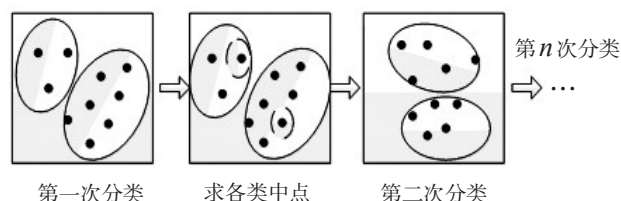


图1 聚类算法示意图

1.1 数据的处理

假设现有 n 个样品的 P 元观测数据,它们组成一个数据矩阵,记为 X ,那么:

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

上面的数据矩阵中,每一行表示一个样品,每一列则表示一个指标, x_{ij} 表示第 i 个样品的第 j 项指标的数值,其中 $i=1,2,\dots,p$ 。当 P 个指标对 n 个样品进行聚类时,经常会出现观测数据量纲不统一的问题,所以在进行分析之前首先需要对观测数据进行标准化,计算方法为:

$$x_{ij}^* = \frac{x_{ij} - E(x_{ij})}{\sqrt{D(x_{ij})}} \quad (1)$$

数据经过标准化处理后,得出新的期望和方差的大小为:

$$E(x_{ij}^*) = E\left(\frac{x_{ij} - E(x_{ij})}{\sqrt{D(x_{ij})}}\right) = \frac{E(x_{ij}) - E(x_{ij})}{\sqrt{D(x_{ij})}} = 0 \quad (2)$$

$$D(x_{ij}^*) = E(x_{ij}^*)^2 - 0 = \frac{E((x_{ij} - E(x_{ij}))^2)}{D(x_{ij})} = \frac{D(x_{ij})}{D(x_{ij})} = 1 \quad (3)$$

即数据处理后,期望均为0,方差都为1。

1.2 距离的计算

距离的计算是聚类分析方法的关键。距离是用来度量样品点之间的亲密程度的,计算距离的方法较多,比如说欧氏距离,绝对距离,马氏距离等,以欧式距离为例,不妨设两个样品构成的矩阵分别是 $x_i=(x_{i1}, x_{i2}, \dots, x_{in})$, $x_j=(x_{j1}, x_{j2}, \dots, x_{jn})$,则欧式距离的定义为:

$$d(x_i, x_j) = \left[\sum_{k=1}^n (x_{ik} - x_{jk})^2 \right]^{1/2} \quad (4)$$

用二阶范数表示为:

$$d(x_i, x_j) = \|x_i - x_j\| \quad (5)$$

1.3 系统聚类方法

系统聚类方法的原则取决于样品间距离的定义方式,距离的计算方式不同产生的系统聚类分析结果不同。

假设现有 n 个样品,用公式(4)计算每两个样品之间的距离,得到实对称矩阵:

$$D_0 = \begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1n} \\ d_{21} & d_{22} & \cdots & d_{2n} \\ \vdots & \vdots & & \vdots \\ d_{n1} & d_{n2} & \cdots & d_{nn} \end{bmatrix}$$

从矩阵 D_0 的非主对角线上找到最小距离,记为 D_{pq} ,那么类 G_p, G_q 就可以合并为一个新类 $G_r=(G_p, G_q)$,在矩阵 D_0 中去掉类 G_p, G_q 所在的行和列,再加上新类 G_r 与其余类之间的距离,得到 $n-1$ 阶的新矩阵 D_1 ,对新矩阵 D_1 重复上述计算步骤,得到新矩阵 D_2 ,依次计算下去,直到所有样品聚为一个大类结束。

在合并类的过程中需要记下被合并样品的编号和两类合并时的水平,最后绘制聚类谱系图。

2 基于系统聚类方法的地区污染物分类

2.1 系统聚类要素的构建与处理

在解决大气污染的地区分类问题时,聚类要素的选择十分重要,它影响着分类结果的准确性和科学性^[8]。首先对聚类的要素进行处理,将地区作为聚类的对象,大气污染物含量作为聚类要素。

我国目前监控的空气污染物包括:可吸入颗粒物(直径小于 $10\mu\text{m}$ 的颗粒物,PM10)、氮氧化物、臭氧、二氧化氮、二氧化硫等^[9-12],本文选取大气污染物的聚类分析指标分别是:PM2.5, PM10, 一氧化碳, 二氧化碳, 臭氧和二氧化硫,选取了8个城市作为聚类对象,分别是:北京,上海,佛山,海口,西安,酒泉,拉萨和重庆,表1给出具体的数据情况(数据来

源于 2019 年 3 月 18 日 24:00 中国天气网实况数据,单位: $\mu\text{g}/\text{m}^3$, CO 为 mg/m^3 。

表 1 大气污染物评价指标

地区	PM2.5	PM10	一氧化碳	二氧化碳	臭氧	二氧化硫
北京	143	175	1.23	109	34	8
上海	43	49	0.68	70	40	5
佛山	65	137	1.20	83	3	13
海口	17	33	0.50	10	45	4
西安	25	87	0.61	82	18	6
酒泉	39	124	0.65	33	61	16
拉萨	22	61	0.54	22	92	5
重庆	45	72	0.86	69	28	8

在对聚类的要素数据进行处理时,处理的方法有总和标准化、标准差标准化、极大值标准化和极差的标准化等。以下采用标准差标准化对表 1 中的数据进行处理。

标准差标准化的处理需要计算每一个聚类要素的数学期望和方差,计算公式为:

$$x_{ij}^* = \frac{x_{ij} - E(x_{ij})}{\sqrt{D(x_{ij})}}, (i=1,2,\cdots,8; j=1,2,\cdots,6).$$
 (6)

利用标准差标准化处理方法,计算得到标准化后的数据见表 2。

表 2 大气污染物评价指标标准化

地区	PM2.5	PM10	一氧化碳	二氧化碳	臭氧	二氧化硫
北京	2.290 9	1.696 5	1.554 3	1.433 1	-0.224 7	-0.029 4
上海	-0.169 1	-0.886 7	-0.361 4	0.298 3	-0.004 6	-0.734 0
佛山	0.372 1	0.917 5	1.449 8	0.676 5	-1.361 7	1.145 1
海口	-0.808 7	-1.214 7	-0.988 3	-1.447 6	0.178 8	-0.968 9
西安	-0.611 9	-0.107 6	-0.605 2	0.647 4	-0.811 5	-0.499 1
酒泉	-0.267 5	0.650 9	-0.465 9	-0.778 4	0.765 7	1.849 7
拉萨	-0.685 7	-0.640 7	-0.849 0	-1.098 4	1.902 8	-0.734 0
重庆	-0.119 9	-0.415 2	0.265 6	0.269 2	-0.444 7	-0.029 4

2.2 聚类距离的计算

系统聚类模型有多种距离的计算方法,包括:最短距离法、最长距离法、类平均法、中间距离法、重心法和 ward 法^[13]。本文使用最短距离聚类法进行求解,计算方法为:

$$D_{rk} = \min(D_{pk}, D_{qk}).$$
 (7)

图 2 给出最短距离法的示意图。

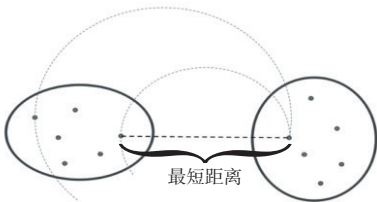


图 2 最短距离法示意图

表 2 数据已经通过聚类要素的处理,下面利用 MATLAB 软件辅助进行距离的计算。

首先,将每一个城市作为一个样品,先计算这 8 个城市之间的欧式距离,用 D_0 表示距离矩阵,则 D_0 为:

$$D_0 = \begin{bmatrix} 0 & & & & & & & \\ 4.269\ 3 & 0 & & & & & & \\ 2.746\ 7 & 5.821\ 6 & 0 & & & & & \\ 4.186\ 5 & 4.595\ 9 & 5.615\ 1 & 0 & & & & \\ 3.651\ 7 & 3.513\ 4 & 2.011\ 6 & 1.300\ 2 & 0 & & & \\ 3.288\ 3 & 2.480\ 7 & 1.144\ 1 & 4.820\ 7 & 3.041\ 2 & 0 & & \\ 3.359\ 9 & 5.115\ 6 & 2.409\ 7 & 2.646\ 1 & 3.575\ 3 & 1.874\ 4 & 0 & \\ 2.627\ 7 & 3.278\ 8 & 3.289\ 3 & 1.262\ 2 & 3.171\ 9 & 2.790\ 5 & 3.080\ 7 & 0 \end{bmatrix}$$

在计算完每行的欧式距离之后,使用最短距离法对类间的距离进行聚类,得出的结果为:

$$D = \begin{bmatrix} 2.000\ 0 & 8.000\ 0 & 1.144\ 1 \\ 5.000\ 0 & 9.000\ 0 & 1.262\ 2 \\ 4.000\ 0 & 7.000\ 0 & 1.874\ 4 \\ 10.000\ 0 & 11.000\ 0 & 2.011\ 6 \\ 3.000\ 0 & 12.000\ 0 & 2.409\ 7 \\ 1.000\ 0 & 13.000\ 0 & 2.746\ 7 \\ 6.000\ 0 & 14.000\ 0 & 2.790\ 5 \end{bmatrix}$$

矩阵 D 的解释为:在 1.144 1 的水平下,类 2 和类 8 合并成为类 9;在 1.262 2 的水平下,类 5 和类 9 合并成为类 10;在 1.874 4 的水平下,类 4 和类 7 合并成为类 11;在 2.011 6 的水平下,类 10 和类 11 合并成为类 12;在 2.409 7 的水平下,类 3 和类 12 合并成为类 13;在 2.746 7 的水平下,类 1 和类 13 合并成为类 14;在 2.790 5 的水平下,类 6 和类 14 合并成为类 15。

2.3 结果谱系图的得出与分析

根据 MATLAB 软件计算得到谱系图,可以直观的得到 8 个地区的分类结果,见图 3。

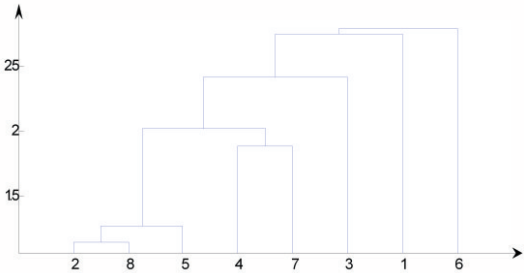


图 3 系统聚类谱系图

根据系统聚类的谱系图可以得到空气污染物含量对地区的划分结果,分成四类的结果见表 3。

表 3 地区分类结果

类别	欧氏距离
1	佛山
2	上海、海口、西安、拉萨、重庆
3	北京
4	酒泉

2.4 谱系图的分析

本文利用系统聚类模型,在对大气污染物取样数据分析的基础上,对8个地区的污染特征进行了分类,现将8个地区分成四类,佛山单独为一类,这一类的地区中臭氧含量浓度较低,一氧化碳含量、二氧化硫含量和PM10含量较高;上海、海口、西安、拉萨、重庆是一类,这一类中与第一类相反,一氧化碳含量和二氧化硫含量相对较低,而臭氧含量相对较高,其余几项污染物指标居中;北京单独为一类,这一类中除二氧化硫含量居中外,剩余污染物含量都排名靠前,环境状态较为严峻;最后,酒泉单独为一类,这一类中PM2.5含量较低,但PM10含量、二

氧化硫含量和臭氧含量较高。

3 结论

根据以上的计算分析过程,能够发现地区间的污染物含量存在较大差异,根据系统聚类分析结果,能够针对性的对这些地区提供环境保护的方针政策。若继续利用聚类分析模型对大气污染物进行分类研究,进一步可以考虑不同的聚类方法对分析结果的影响。本文通过系统聚类方法对8个地区的归类可以看出,科学的统计方法在社会发展中的应用广泛,合理的数学模型辅以分析方法能够更加客观和科学的对问题进行深度的研究。

参考文献:

- [1] 樊琳琳.大气治理中的个体责任探究[J].西昌学院学报(自然科学版),2018,32(2):36-38.
- [2] 吕连宏,罗宏,张型芳.近期中国大气污染状况、防治政策及对能源消费的影响[J].能源与环境,2015,37(8):9-15.
- [3] 陈健鹏,李佐军.中国大气污染治理形式与存在问题及若干政策建议[J].发展研究,2013,10:4-14.
- [4] 童金茂,李利利,文燕梅.2008-2016年我国篮球研究热点问题的聚类分析——以12种体育核心期刊为例[J].西昌学院学报(自然科学版),2016,30(3):29-33.
- [5] 李慕容.基于聚类分析的安徽企业R&D资源分布研究[J].合肥工业大学学报(自然科学版),2004,27(11):1464-1468.
- [6] 钱堃,包为民,李偲松,等.K均值聚类分析方法在洪水预报中的应用[J].水电能源科学,2012,30(5):41-44.
- [7] 王全亮,张月芬.系统聚类法在网络学习行为分析中法应用研究[J].中国教育信息化,2016,8:90-92.
- [8] 高激.广东省自然地理环境聚类分析模型及其应用——以21个地级市为例[J].现代经济信息,2013,9:444.
- [9] 潘本锋,宫正宇.环境空气质量指数在应用中存在的问题及建议[J].中国环境监测,2015,1(11):64.
- [10] 尚子激,程一帆,康延臻,等.全国10个代表性城市空气污染指数与气象条件的关系[J].兰州大学学报(自然科学版),2018,54(1):98-103.
- [11] 李小飞,张明军,王圣杰,等.中国空气污染指数变化特征及影响因素分析[J].环境科学,2012,33(6):1936-1943.
- [12] 高会旺,陈金玲,陈静.中国城市空气污染指数的区域分布特征[J].中国海洋大学学报,2014,44(10):25-34.
- [13] 于闽,周翔.系统聚类分析法在统筹区域土地利用分区研究中的应用[J].国土资源刊,2008,1:44-46.

(责任编辑:曲继鹏)

(上接第3页)

- [10] 唐文莉,高健彭,彭镇华.毛竹(*Phyllostachys edulis*)光系统I基因LhcaPe02全长的克隆与序列分析[J].安徽农业大学学报,2008,35(2):153-155.
- [11] 黄培堂.分子克隆实验指南[M].科学出版社.2002:27-28.
- [12] 高志民,李雪平,彭镇华,等.竹子捕光叶绿素a/b结合蛋白基因全长的克隆和序列分析[J].林业科学.2007,43(3):34-38.
- [13] 马荣群,陈红运,黄文胜,等.实时定量PCR方法检测转基因产品[J].植物检疫,2002,16(1):61-64.

(责任编辑:曲继鹏)