哈尔滨大气中多环芳烃污染的FCM聚类算法

崔 嵩[1],李一凡[1,2],马万里[1],田崇国[3],贾宏亮[4],张 志[1],齐 虹[1],刘丽艳[1]

(1.哈尔滨工业大学 城市水资源与水环境国家重点实验室 国际持久性有毒物质联合研究中心,150090 哈尔滨,cuisong-bq@163.com;2. 加拿大环境部科学技术局,M3H 5T4 加拿大 多伦多;3.中国科学院烟台海岸带研究所,264003 山东 烟台;4. 大连海事大学 国际持久性有毒物质联合研究中心,116023 辽宁 大连)

摘 要: 为提高环境中持久性有毒物质残留水平的分析能力,以2007年1~4月哈尔滨市4个不同功能区域(市区—郊区—农村—偏远地区)的大气中多环芳烃(PAHs)的监测质量浓度为数据样本,运用模糊C均值聚类算法作为样本聚类的方法,研究该地区2007年春季大气中PAHs的分布特征,并得到不同样本的聚类中心以及样本的隶属度矩阵,为样本的特征分析提供量化依据.分类结果发现:城市工业区的污染状况最为严重,农村地区和介于市区和工业区的居民区的区域次之,而远离污染源的市内居民区、城市上风向和偏远地区污染较轻.分析结果表明,哈尔滨市区大气中PAHs主要来自于燃煤和工业排放,农村大气中的PAHs主要来自于农作物秸秆的燃烧.市区和农村地区大气中PAHs对于人类的健康存在较大潜在威胁.

关键词: 哈尔滨;大气;多换芳烃;模糊C均值聚类

中图分类号: X823 **文献标志码:** A **文章编号:** 0367-6234(2011)08-0065-05

Source identification and spatial characterization of polycyclic aromatic hydrocarbons in Harbin air by using fuzzy C-means clustering algorithm

CUI Song[1], LI Yi-fan[1,2], MA Wan-li[1], TIAN Chong-guo[3], JIA Hong-liang[4], ZHANG Zhi[1], QI Hong[1], LIU Li-yan[1]

(1. International Joint Research Center for Persistent Toxic Substances, State Key Laboratory of Urban Water Resource and Environment, Harbin Institute of Technology, 150090 Harbin, China, cuisong-bq@163.com; 2. Science and Technology Branch, Environment Canada, M3H 5T4, Toronto, Ontario Canada; 3. Yantai Institute of Coastal Ione Research, Chinese Academy of Sciences, 264003 Yantai, Shandong, China; 4. International Joint Research Center for Persistent Toxic Substances, Dalian Maritime University, 116023 Dalian, Liaoning, China)

Abstract: To improve the analysis ability of residue levels of persistent toxic substances (PTS) in environment, this paper investigated the distribution characteristics of air concentrations of polycyclic aromatic hydrocarbons (PAHs) during the Spring of 2007 (January to April) in 4 different functional areas (urban, suburban, rural, and remote areas) in and near the Harbin metropolitan by using the fuzzy C-means clustering algorithm, and got the cluster centers of different samples and a membership matrix which offered quantity foundation for analysis of samples description. Results showed that the contamination level in the urban industrial area was the highest, followed by those in the other urban places and rural area, and the contamination levels in the remote area and in the places on the windward were the lowest. PAHs in Harbin urban atmospheric were mainly from coal-burning and industrial emissions, those in rural areas were mainly from the burning of crop stalks. PAHs in air in this region have imposed a potential threat on human health.

Key words: Harbin; air; polycyclic aromatic hydrocarbons; fuzzy C-means clustering algorithm

收稿日期: 2009-12-28.
基金项目: 城市水资源与水环境国家重点实验室自主课题项目(2008DX01).
作者简介: 崔 嵩(1981—),男,博士研究生;
李一凡(1949—),男,教授,博士生导师.

多环芳烃(PAHs)作为环境中一种广泛存在的持久性有毒物质(PTSs),具有致癌性、生物富集性和长距离迁移特性,并且有些PAHs还具有致癌、致畸及致突变等"三致"效应.在众多污染

源中,垃圾焚烧、煤的燃烧、炼油厂、焦化厂以及汽车等机动车辆尾气的排放是 PAHs 的主要来源[1]. PAHs 存在于大气、土壤、植物和水体等环境介质中,可以通过皮肤接触、呼吸作用及食物链进入人体,对人体造成潜在的危害,因此,开展 PAHs 的研究对环境影响评价具有重要意义[2-4].

聚类分析就是将所研究的数据对象,根据其相似性分组成多个簇或类,使其在同一簇中组分之间的相似度较高,而在不同簇中的组分差别则较大[5]. 模糊聚类算法与硬聚类(每一元素只能归属于某一类)算法相比较而言,在作样本分析时能够很好地体现样本特征,同时模糊聚类引入了隶属度的概念,能够更好地给出样本分属于各类的隶属度,从而为制定相关的决策提供有价值的参考依据. 在模糊 C 均值(Fuzzy C - Means, FCM)聚类算法中,每一个数据点都按照一定的隶属度隶属于某一聚类中心. 这一聚类技术,首先由 Dunn 于 1974 年提出,并由 Jim Bezdek 于 1981 年改进[6]. 该方法提供了一种如何将多维空间分布的数据点分组成特定数目群的途径,并且该方法首先随机选取若干聚类中心,所有数据点都被赋予对聚类中心一定的模糊隶属度,然后通过迭代方法不断修正聚类中心,迭代过程中以极小化所有数据点到各个聚类中心的距离与隶属度值的加权和为优化目标[7]. 在实际应用中,由于模糊 C 均值聚类算法是基于距离的算法,这就使得聚类结果可能会受到数据集中孤立点的影响. 目前,FCM 已应用于多种领域,如医学诊断、目标识别、沉积物污染特征、岩石分类、顾客关系管理[8]等,但该方法用于环境影响评价的报道很少. 本文采用模糊 C 均值聚类算法,以 2007 年春季哈尔滨大气气相中 16 种 PAHs 为属性指标,在技术上对不同功能区进行了聚类,并对聚类结果进行了分析,以期能够更好地研究该地区大气中 PAHs 的分布特征,为进一步研究 PAHs 污染所造成的区域性环境污染评价进行初探,为环境保护部门制定相应的决策提供科学依据.

1 研究方法

1.1 模糊 C 均值聚类算法的基本原理

本文将模糊 C 均值聚类算法应用于区域性环境污染评价,其基本原理为:首先假设聚类目标 $X = \{x_1, x_2, \cdots, x_n\} \in R^s$ 是存在于一个 s 维实数空间 R^s 中的有限样本数据集,n 表示样本数据集中的元素个数,本研究中样本数据集为 8 个,每个样本数据集中的元素个数为 16 个,即不同的采样点所监测到的大气中多环芳烃的种类. 设 c_i 为 FCM 聚类算法把样本数据聚为 c 个模糊类时,每一个类 i 相应的聚类中心,$d_{ij} = \| c_i - x_i \|$ 为欧几里得距离,表示为 c_i 中的第 i 个聚类中心与第 j 个数据点间的距离,$m \in [1, \infty)$ 为加权指数,$\boldsymbol{U} = \{u_{ij}\}$ 为隶属度,是一个 $c \times n$ 维的矩阵,且它们须:

$$\sum_{i=1}^{c} u_{ij} = 1, \forall j = 1, \cdots n, \tag{1}$$

$$J(\boldsymbol{U}, c_1, \cdots, c_c) = \sum_{i=1}^{c} J_i = \sum_{i=1}^{c} \sum_{j=1}^{n} J_{ij}^m d_{ij}^2. \tag{2}$$

FCM 聚类算法的基本过程就是求在约束条件式(1)成立的情况下,式(2)取最小值时的数值解. 式(1)表示样本中每个元素属于各类的隶属度之和为 1;式(2)则表示判定距离之和的目标函数. 该算法的特点是,类数 c 需要事先给出才能进行下一步运算[9-10].

FCM 聚类算法的具体运算步骤如下:1)给出拟定的分类数 c 和相应的加权指数 m;2) 初始化隶属矩阵 $\boldsymbol{U}$,用值在[0,1] 区间的随机数进行初始化,同时使其满足约束条件(1);3) 计算相应的 c 个聚类中心 $c_i, i = 1, \cdots c$;4) 计算目标函数(2),如果某个给定的阈值大于 J 值,或者 J 值的改变量相对于上一次仍小于某个阈值时,则停止运算;否则重新计算隶属矩阵 $\boldsymbol{U}$, 并返回上一步骤.

1.2 聚类有效性函数

聚类有效性的判别是对聚类结果客观合理性的验证,应用 MATLAB7.1 提供的模糊逻辑工具箱(Fuzzy Logic Toolbox)中的 fcm 函数可以很好地对样本数据进行聚类,此时,只需要输入样本所分的类数 c ,即初始变量,就可以得出相应的结果. 但是,有关初始变量 c 值的给定,在某种程度上存在一定的主观性,因为 c 值的不同会导致不同的聚类结果,即使在同一 c 值的情况下,也可能产生不同的结果. 这是由于算法结果本身过分地依赖于初始给定值,而初始值的给定在具体的计算过程中又是随机的,有时也会不可避免地造成局部最优而并非能够达到全局最优[11-13],而这就可能导致计算结果对真实情况产生偏差,此时,就需要考虑聚类的有效性问题. 而聚类的有效性问题一般可以通过建立有效性函数的方法来加以解决. 这种函数通常用于衡量聚类的分离度和紧密度,并以此来判定聚类的有效性. 1991 年由 XIE,X L 和 BENI,G 共同提出的模糊聚类有效性函数,可以很好地解决这个问题,具体判别如下:

$$S = \frac{\sum_{i=1}^{c}\sum_{k=1}^{n}(\mu_{ik})^m \| V_i - V_k \|^2}{n[\min_{i,j} \| V_i - V_k \|^m]}. \quad (3)$$

其中:c 为所划分的类数,n 为所采集数据点的个数,V_i 为第 i 类的聚类中心,V_k 为第 k 类的聚类中心,$\mu_{ik}(i=1,2,\cdots,c;k=1,2,\cdots,n)$ 为第 k 个点属于第 i 类的模糊隶属度,$\min_{i,j} \| V_i - V_k \|^2$ 表示第 i 个聚类中心和第 j 个聚类中心之间的距离最小值,用来衡量类间的分离度;$\left(\sum_{i=1}^{c}\sum_{k=1}^{n}(\mu_{ik})^m \| V_i - V_k \|^2\right)/n$ 表示每一点到其相应聚类中心的平均偏差和,用来衡量每一类的紧密度. 通常一个好的聚类结果应该是,具有相同特征的数据点尽可能地划分为同一类,具有不同特征的数据点尽可能地划分为不同的类别,即 S 值越大,表明所有的聚类都是分离的,S 越小,表明所有聚类紧密且相互独立,聚类效果越好.

根据大量研究和使用经验,以及相关理论依据给出[14],通常 c 的范围在$[2,\sqrt{n}]$,通过计算 c 在其取值范围 $2 \leqslant c \leqslant \sqrt{n}$ 内,每个整数 c 所对应的 S 值,进一步比较取不同 c 值所对应的 S 值,当 S 值达到最小时所对应的 c 值即为所求的聚类数,此时,所取的类数 c 获得聚类的效果最好,且符合客观实际,从而减少计算可能导致的结果偏差.

2　实例分析

2.1　数据获得

选择哈尔滨地区为研究区域,大气样品的采集、16 种 PAHs 的分析等详见文献[1]. 2007 年春季(从 1 月末至 4 月末),在哈尔滨地区,根据不同功能区特点,系统布设市区 3 个(UR)、郊区 1 个(SU)、农村 3 个(RU)和偏远地区 1 个(BA)等 8 个采样点,利用聚氨酯泡沫材料(PUF)被动采样器进行大气样品的采集,采样时间从放置被动采样器至样品取回,为一次监测并按监测天数取平均值,样品的分析以及 PAHs 质量浓度数据的获得均在哈尔滨工业大学国际持久性有毒物质联合研究中心(IJRC - PTS)实验室进行. 有关研究表明,应用 PUF 被动采样器进行污染物大气质量浓度的监测,能够获得大气中污染物的准确质量浓度,但是也受很多客观因素的影响,其中采集气体的体积为主要影响因素,采样体积一般可以通过采样速率和采样时间计算得出. 大量的研究表明,PUF 的采样平均速率通常在 3 ~ 4 m^3/d[15-16],因此,假定该地区采样速率为 3.5 m^3/d,从而得到每立方米 PAHs 的具体质量浓度数值. 采样点分布和详细情况见文献[1],哈尔滨市大气中 16 种 PAHs 的平均质量浓度见表 1.

表 1　哈尔滨市大气中 16 种 PAHs 平均质量浓度[1]　　ng · m^{-3}

样本名称	Nap	Acy	Ace	Flo	Phe	Ant	Flu	Pyr	BaA	Chr	BbF	BkF	BaP	LcdP	DahA	BghiP
UR1	0.960 0	0.237 1	1.962 9	5.037 1	32.074 3	1.005 7	13.500 0	10.411 4	1.060 0	1.982 9	1.380 0	0.397 1	0.251 4	0.662 9	0.128 6	0.597 1
UR2	1.851 4	0.545 7	2.517 1	4.362 9	43.762 9	1.720 0	20.648 6	14.105 7	1.848 6	3.082 9	2.254 3	0.668 6	0.542 9	1.708 6	0.345 7	1.540 0
UR3	1.254 3	0.485 7	3.248 6	5.848 6	60.202 9	2.145 7	23.788 6	23.005 7	2.088 6	3.562 9	2.462 9	0.711 4	0.540 0	1.454 3	0.288 6	1.317 1
SU1	0.442 9	0.082 9	1.088 6	4.320 0	23.534 3	0.580 0	8.351 4	5.294 3	0.465 7	0.957 1	0.625 7	0.165 7	0.068 6	0.242 9	0.048 6	0.200 0
RU1	0.951 4	0.217 1	2.951 4	7.017 1	37.637 1	1.240 0	14.034 3	10.245 7	1.097 1	2.040 0	1.608 6	0.457 1	0.348 6	0.760 0	0.148 6	0.714 3
RU2	1.120 0	0.191 4	2.837 1	7.197 1	33.502 9	1.234 3	10.882 9	9.288 6	1.008 6	1.614 3	1.100 0	0.302 9	0.222 9	0.560 0	0.134 3	0.520 0
RU3	0.688 6	0.125 7	1.962 9	4.062 9	43.728 6	1.074 3	18.720 0	9.628 6	0.885 7	1.762 9	1.100 0	0.308 6	0.180 0	0.577 1	0.111 4	0.488 6
BA1	0.725 7	0.080 0	1.617 1	5.137 1	27.162 9	0.654 3	9.785 7	4.962 9	0.380 0	0.880 0	0.545 7	0.148 6	0.040 0	0.222 9	0.045 7	0.177 1

2.2　数据处理与分析

1)使用 MATLAB7.1 中的 FCM 函数对以上数据进行聚类运算,并进行聚类有效性的判定,得到如下结果:

对聚类有效性函数(3),确定类数 c. 一般地,取经验值 $m = 2$,分母权值均为 1,由于本文采样点数有限,当类数 $2 \leqslant c \leqslant 2\sqrt{2}$ 时,c 只能取 2,由此得到聚类中心矩阵为

$$\begin{vmatrix} 4.648 & 1.446 & 9.287 & 17.11 & 173.4 & 5.964 & 73.99 & 56.50 & 5.884 & 10.18 & 7.106 & 2.067 & 1.568 & 4.606 & 0.9198 & 4.134 \\ 2.942 & 0.5622 & 7.174 & 19.75 & 107.7 & 3.278 & 39.77 & 28.09 & 2.785 & 5.184 & 3.599 & 1.006 & 0.6257 & 1.670 & 0.3506 & 1.516 \end{vmatrix}$$

然而当 n 为 3 时,得到的聚类中心矩阵为

4.391	1.697	11.36	20.46	210.3	7.495	83.11	80.29	7.295	12.45	8.605	2.486	1.887	5.083	1.009	4.604
4.094	1.042	8.318	16.96	147.2	4.688	63.71	39.74	4.472	8.021	5.675	1.642	1.212	3.569	0.7118	3.195
2.728	0.4802	6.365	18.74	100.3	2.918	36.56	24.89	2.379	4.520	3.035	0.8406	0.4664	1.383	0.2906	1.216

聚类中心点在各个维的取值均表征了该类的特征,由此可以看出,当 $n = 3$ 时,即分为3类时,特征比较明显,此时 S 值也为最小.

2)隶属度矩阵 $\boldsymbol{U}$. 隶属度矩阵 $\boldsymbol{U}$ 为一个 3×8 的矩阵,这表示8个不同功能区域分别属于3种类型的隶属度. 由于所划分的功能区域即采样区域数目有限,在此任取两个样本点作以分析.

样本1: $\boldsymbol{U}_1 = [0.004\,718, 0.022\,51, 0.972\,8]^{\mathrm{T}}$,

样本2: $\boldsymbol{U}_2 = [0.039\,00, 0.686\,9, 0.274\,1]^{\mathrm{T}}$.

从以上两个样本可以看出,矩阵 $\boldsymbol{U}_1$ 及 $\boldsymbol{U}_2$ 每一列的和均为1,符合每一样本的各类隶属度之和为1的前提. 因此,取样本中每一列的最大值,则最大值所在的行数就表示该样本属于其相应的类型. 例如样本1,其最大值在第3行,所以属于第3种类型;而样本2,取其最大值则属于第2种类型,但从结果中可以分析出,样本2还兼具有类型3的特征,说明这一样本有可能处于农村地区并且该地区介于城市与偏远地区之间.

通过以上运算还可具体得到单个样本所属类型,如表2所示.

表2 样本所属类型

类别	样本名称
第1类	UR3
第2类	UR2、RU1、RU3
第3类	UR1、SU1、RU2、BA1

3) 聚类结果分析. 从聚类结果可以看出,城区环境中PAHs的主要来源是人类活动,主要包括煤或石油化工等燃料的不完全燃烧. 在本研究中,大气中PAHs的质量浓度最大值出现在UR3,该采样点处在哈尔滨市内的主要工业区内,石油化工、金属冶炼等工业源排放及汽车等机动车的尾气排放,取暖燃煤,可能是导致该采样点高含量PAHs的原因. 其中取暖用煤的不完全燃烧可能是导致该地区春季大气中PAHs高质量浓度的主要原因,哈尔滨地区供暖一般在每年的4月中旬结束. 本研究中采集时间为1月末~4月末,正好处在冬春季采暖期,市区供暖燃烧用煤会产生大量PAHs,从而导致大气中PAHs质量浓度相应增加.

第二类结果,其污染源可能主要来自于农村地区(RU1,RU3)的供暖,东北地区农村冬春两季大多采用燃烧秸秆和木材取暖以及烹饪,有研究表明,冬季采取煤炭和秸秆等取暖普遍存在于中国北部平原地区[17]. 而市区内介于工业区和居民区之间的区域(UR2)属于这一类的原因可能是区域内的燃煤供暖及汽车尾气的排放.

第三类为城市上风向(SU1,RU2)、偏远地区(BA1)及远离污染源的市内居民区(UR1),由于北方冬春两季受北风的影响以及风力作用会对这些区域大气中的PAHs产生稀释作用,另外市内居民区内没有大量的PAHs排放源如工业排放、汽车尾气及燃煤供暖等,所以会导致这些区域大气气相中PAHs的质量浓度较低.

3 结 论

1) 相对于硬聚类方法,模糊C均值聚类能够很好地对所研究的不同功能区域进行聚类,当然,将模糊C均值聚类算法用于不同功能区域样本特征进行聚类还存在尚待解决的问题. 基于距离的模糊 c 均值聚类算法,由于样本点较少或存在孤立点可能会影响到聚类的效果,另外,该算法所存在的局限性在于,算法本身需要事先给出所需聚类的类数即参数 c,这就会导致算法结果对这个参数十分敏感,c 取值的不同,聚类的结果也会截然不同,因此,对于使用者来说,这个算法还需要根据实际情况加以判断. 而聚类有效性函数的搜索范围在其他文献的研究中能够起到很好的效果,在本文中却并不适用,可能是由于本文样本点较少所导致. 然而本文的实际聚类效果却比较理想.

2)市区工业区内的受污染程度最大,而上风向和偏远地区的污染相对较小. 由于本文的研究范围为冬春季,分类结果也可对我国北方同时期不同功能区域大气污染分布特征的研究起到借鉴意义. 另外,分类结果也表明,市区工业区和农村地区由于季节性燃煤和秸秆的燃烧导致PAHs在大气中的质量浓度增多,对人类的健康可能存在较大的潜在威胁,这也可为进一步对人类健康风险评价的研究起到一定的参考作用.

参考文献:

[1] 马万里,李一凡,孙德智,等. 哈尔滨市大气气相中多环芳烃的研究[J]. 环境科学,2009,30(11):49-54.

[2] WANG X P, XU B Q, KANG S C, *et al.* The historical residue trends of DDT, hexachlorocyclohexanes and polycyclic aromatic hydrocarbons in an ice core from Mt.

Everest, central Himalayas, China[J]. Atmos Environ, 2008, 42(27):6699 – 6709.

[3] HUNG H, BLANCHARD T P, HALSALL C J, *et al.* Temporal and spatial variabilities of atmospheric polychlorinated biphenyls (PCBs), organochlorine (OC) pesticides and polycyclic aromatic hydrocarbons (PAHs) in the Canadian Arctic: results from a decade of monitoring [J]. Sci Total Environ, 2005, 342(1/3):119 – 144.

[4] SIMCIK M F, EISENREICH S J, LIO Y P J. Source apportionment and source/sink relationships of Polycyclic Aromatic Hydrocarbons (PAHs) in the coastal atmosphere of Chicago and Lake Michigan[J]. Atmos Environ, 1999, 33 (30):5071 – 5079.

[5] HSU T H. An application of fuzzy clustering in group – positioning analysis [J]. Proc Natl Sci, Counc ROC (C), 2000,10(2):157 – 167.

[6] MICHALOPOULOS M, DOUNIAS G D, THOMAIDIS N T. Decision making using fuzzy C – means and inductive machine learning for managing bank branches performance [EB/OL]. http://citeseer. nj. nec. com/458829, html, 2002.

[7] 吴晓莉,林哲辉. MATLAB 辅助模糊系统设计[M]. 西安:西安电子科技大学出版社,2002:158 – 159.

[8] 梁静国,张亚光,戈华. CRM 中的模糊 C 均值(FCM)客户聚类算法研究[J]. 哈尔滨工程大学学报,2004,25(2):257 – 260.

[9] XIE X L,BENI G. A validity measure for fuzzy clustering [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1991,13(8): 841 – 847.

[10] HAN J, MICHELINE K. 数据挖掘:概念与技术[M]. 范明,孟小峰,译. 北京:机械工业出版社,2001:126 – 128.

[11] 张智星,张春在,[日]水谷英二. 神经 – 模糊和软计算[M]. 西安:西安交通大学出版社,2000:93 – 97.

[12] 刘小览,赵英凯,陆金桂. 数据挖掘中 Fuzzy C – means 的自适应聚类算法[J]. 南京化工大学学报:自然科学版,2001,23(5):23 – 25.

[13] 吴成茂,范九伦. 基于数据划分最大信息的聚类有效性函数[J]. 西安电子科技大学学报:自然科学版,2001,28(6):781 – 784.

[14] 于剑,程乾生. 模糊聚类方法中的最佳聚类数的搜索范围[J]. 中国科学:E 辑, 2002,32(2):274 – 280.

[15] ZHANG Z, LIU L, LI Y F, *et al.* Analysis of polychlorinated biphenyls in concurrently sampled Chinese air and surface soil [J]. Environ Sci Technol, 2008, 42 (17): 6514 – 6518.

[16] SHOEIB M, HARNER T. Characterization and comparison of three passive air samplers for persistent organic pollutants [J]. Environ Sci Technol, 2002, 36(19): 4142 – 4151.

[17] LIU S Z, TAO S, LIU W X, *et al.* Atmospheric polycyclic aromatic hydrocarbons in north China: a winter – time study [J]. Environ Sci Technol, 2007, 41(24): 8256 – 8261.

(编辑　刘　彤)