

问题讨论

模糊聚类与背景值检验

叶以宁 解恩惠 王恩增

(哈尔滨科技大学)

陈万峰

(黑龙江省环境保护所)

于洪涛 宋国利

(黑龙江省环境监测中心站)

松嫩平原环境背景值的课题组在六万多平方公里的松嫩平原上,设立了 300 多个取样点,每个取样点各采集了上、中、下三层的样本,每个样本作了十几项化学分析,共有二万个左右分析数据。这些数值的可靠性是最后获得地区或土壤类型背景值以及土壤背景值图的基本依据。数值的可靠性首先依赖于取样的科学性与代表性,以及分析过程的准确性。然而对样品分析数据进行数学方法的检验,剔除可能已被污染的样品,是十分必要的。

一、分析数据的三维性及目前的各种检验法

松嫩平原环境背景值的二万个数据,都是由三个参量决定的。这些参量是:样点剖面号、土层号码及分析对象。例如对于剖面号 210,土层号 1,Cu 的含量浓度,就确定了一个数值:“17.7”。因此,如果将剖面号、层数及分析对象看作抽象的三维坐标,每一个分析数据则是定义域在这个三维空间上的函数。这些分析数据之间的内在联系,决定了这个函数的性质。

目前国内外的各种土壤背景值检验法很多^[1],从三维空间的角度来看,共同特点都是将三维空间的某两维固定起来,考虑一维变化过程中数据函数的变化规律(或者用几何语言说“考虑三维抽象空间中一条与坐标轴平行的“直线”),以这些变化规律为基础,来建立数据的检验方法。

仅考虑剖面号的变化,不涉及土层和分析对象之间关系的检验法是最常用检验法。平均值加减标准差法,4d 法等是这类方法的典型方法^[1,2]。另外还有中值法^[10],非参数估计法^[13]以及近来人们推崇的 Konlmogorov 检验法^[11,12]。这些方法在分布规律及样本数目,类型不同的问题中互见长短。

不同元素在土层中上下移动的规律是土壤学者们已经探索过的领域。这就是,用三维抽象空间语

言来说,“第二维”方法。日本学者浅见辉男将这种方法使用到沉积物质的污染研究方面,取得值得借鉴的结果^[19]。中国科学院土壤背景值协作组在广东省作了这一方面的尝试,亦取得可喜的结果^[5,6]。富集系数法^[20]是这类型方法中较独特的一种。但因化验条件要求太高,目前尚难推广。

“第三维”方法——用不同元素之间的联系来考察背景值的分布规律,是新近发展起来的一大类新方法。国内亦有这方面的工作^[8]。

这些方法各有千秋,对土壤背景值的研究都做出了重要的贡献。这些方法都是各自从某一特定的维出发的。而分析数据是三维抽象空间的函数,因此我们能否首先在抽象三维空间中,利用现代数学观点,决定一种度量,然后对这些点进行“立体”式地考察这种三维抽象空间中定义的函数变化规律,由此找到一种更为有力的检验方法,作者在这方面作了一些尝试,现将工作作一简单介绍。

二、用模糊聚类进行数值检验的基本思路

平均值加减标准差方法,在检验方法中具有代表性,是最通用的方法,它的优点除了易于掌握、便于推广外,还在于这种方法反映了异常值的主要统计规律——显著地偏离平均值,但这种方法在理论上的缺陷是明显的。当我们用平均值加减二倍(确切地说应为 1.96 倍)的标准差时,我们所得到的置信区间是 95%。因此我们将落在这个区域以外的值作为异常值时,就会发生概率统计学上所谓的“第一类错误”。由概率论的常识可知,此时对于 200 个样本来说,即使它们都是正常值,也有十个左右分布在认定的异常值区间,故而被剔除^[16]。但单纯为了减少第二类错误而扩大置信区间,又增加了犯第二类错误的概率。一些应该剔除的异常值将不被剔除。其它同类型的方法亦存在类似的弊病。

因此为了提高检验准确性,有必要引进新的三维空间理论为基础的检验法。

1. 度量——样点的距离和相似度的确定

要把环境问题抽象为数学定量问题,以便使用各种数学工具进行研究。引进适当和正确的度量是关键的一步^[2]。最常用的度量有两个:样本距离及样本的相似度。相似度是模糊聚类进行计算的基本依据。

距离和相似度的定义方法很多。方开泰先生在文献[3]、[4]中列举了十种相似度和五种距离。我们采用的方法如下:

由于我们讨论的问题是有害元素的污染问题,因此考虑两个样本差异时,我们选用 Hg、As、Zn、Cd、Ni、Pb、Cu 和 Cr 作为样本的八个坐标。由于酸碱度对有害元素的存留有一定影响,而且测试方法容易,因此取它作为第九个坐标。各种元素在自然界中含量不同,因此用它们在样本中的浓度大小绝对值来衡量样本差别,将使自然界中含量较多的元素和含量较少元素处于不平等的地位。而且 pH 值和它们之间也没有共同量纲。因此首先进行无量纲处理,使之标准化。

我们分别使用 L_1 距离(或称欧氏距离)相关的相似度和非参数方法直接定义相似度两种方法、非参数方法,即参考文献[3]中的方法 V。文献[3]指出,这种方法较适合作为样品之间的相似系数。这种方法在原理上接近于将空间看作 L_1 空间,而 L_1 空间则是保留空间各坐标独立影响最大,而仍满足三角不等式的空间。在实际应用中也证实了用确定相似度的灵敏度高,临界点性质表现明显,因此最后结果被正式采用。

2. 模糊聚类,异常状态的特征检验

模糊聚类分析,是综合了模糊数学原理和聚类分析的方法而得到的一种新型的聚类方法。目前在许多领域中,都得到探索性的应用[14,17]。模糊聚类可以在不同的相似程度上给出等价类。

模糊聚类的原理,有点象中国农村的套亲戚,若甲和乙存在亲戚关系,乙和丙亦存在亲戚关系,我们则认为甲和丙之间存在亲戚关系。模糊聚类是通过 Fuzzy 等价关系及 Fuzzy 矩阵运算来实现的。这种方法可以根据问题的性质,不同的要求及计算过程来确定聚类标准的选择,以达到最佳分类,具有很大灵活性。同时,聚类标准一旦选定,分类则与聚类的过程无关,因此具有客观性。这种聚类方法的缺点是分析对象大类之间如果具有过渡类型,则很难分开。正如农村久居一地的各个家族之间,如果按亲

戚关系分类,则无法将其分开。而我们样本之间类别恰是这种情况,所以这种方法不适宜作为样本分类。尽管如此,模糊聚类在此时表现出一种有趣的性质,这种性质我们称之为“临界点性”,恰好适用于异常值的剔除。

“临界点”是模糊聚类方法应用于类别之间没有明显区别的对象时所表现的这样一种特性,将分类的相似标准从 1 逐渐减少到 0 时,最初绝大多数的元都聚合成一大类,分离出个别元来,这些个别元各自成为一类。随着分类标准的逐渐减少,被分离出来的独立元逐渐增加到了某个“临界点”附近时,聚成一大类的那些元突然散成了大量自己成类的独立元和几个很小类。因此我们取比临界点稍微大一些数作为分类的相似标准时,多数元聚成一类,少数元被分离出来。这些被分离元与聚成一类多数元在相似性上有差异,因此可以将其作为进一步检验的样本。

正如在一个村庄里,久居于这个村庄的居民,通过亲戚的办法,任意两个之间总有亲戚关系。而“异常”值则是刚到这个村庄不久的外乡人,用这种方法恰巧能将其分离出来。土壤虽然千差万别,但它们都在多少个世纪相似的自然条件下形成的,它们存在内在的相似性是必然的。而异常值则是近些年来,工业化社会刚在这些地层覆盖物上打下的印记,异常样本自然是定居不久的“外乡人”。

我们采用的做法是:用相似度的平均值加 1.5 倍标准差作头一次聚类标准,然后依照已经确定好的相似度进行模糊聚类。聚出来的类中最大的一类如果超过了样本总数的 60%,则聚类标准被认定为临界点,而超过 60% 的这大大类被认为是正常类,而不在这大大类中的其余样本,则与多数样本不相似,特性异常而被选出,进入下一步检验。

如果聚出的最大一类小于总样本数的 60%,则认为聚类标准过高,将聚类标准减去 0.2 倍标准差作为新的聚类标准,然后重新进行模糊聚类和检验。这样一次又一次降低标准,直至最大一类大于样本总数 60%,即临界点被找出为止。

如果聚出的最大一类突然跳过样本总数 60%—95% 区间,即重复过程中找不到一大类相似样本,占总样本数的 60%—95% 之间,则认为在这种相似度情况下,模糊聚类的临界点性表现不明显,必须重新定义合适的相似度。

3. 进一步检验

我们将形态不正常的样本挑出后,首先将其各个坐标(即各种污染元素浓度),分别进行以样本总

体为母本的平均值加减标准差检验,找出这些样本中显著偏离平均值的分析数据。

最后,我们用上下层比较法进行检验。重金属与土壤易发生吸附、螯合以及化学沉淀等反应,因此一旦外来的污染元素进入表土,则导致表土浓度低于底土(参见[1]P20),参照文献[19]中所使用的方法,将前两次检验中查出与多数样本在显著差异的样本中显著偏离平均值有害元素的浓度与底层同种元素的浓度比较,当比值大于1.2时,就最后确定该元素的浓度是异常值,样本受该元素污染。

三、计算机实现

我们将这种检验法编成计算机程序,并采用哈尔滨地区的100个采样点的分析数据进行检验,每个样点取上下两层,八种元素含量及pH值,共1,800个数据。并在微型MC—176苹果机上实现。

1. 程序框图

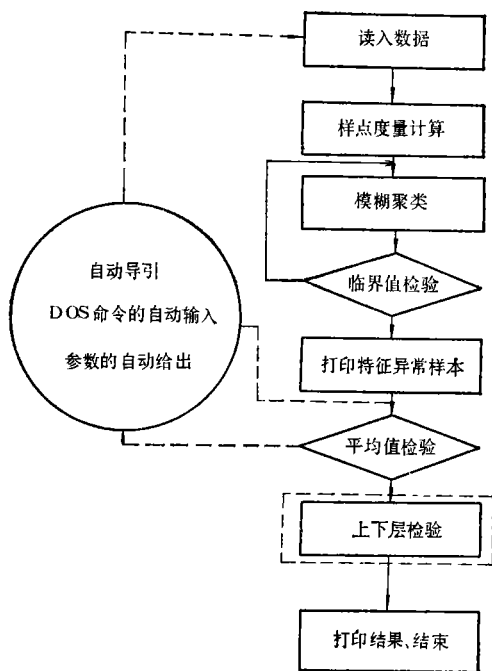


图1 程序框图

2. 程序(略)

3. 关于程序的几点说明

(1) 检验过程采用自动导引, DOS命令自动输入和参数自动给出。

(2) 由于使用机器内存量所限,每次用50个样

本为单位进行计算,然后自动逐批输入。

(3) 模糊聚类程序中 Fuzzy 矩阵计算量太大,用 MC—176 苹果机计算 50 个样本,约需 9 个小时。现采用作者的值日法*仅需 30 多分钟可完成,已证明两种方法的等价性。

(4) 上下层检验法用人工查对,尚未在计算机中实现。

4. 结果

程序运行一小时十分钟。

前 50 个样本经临界点检验,找出 17 个存疑样本。后 50 个样本找出 15 个存疑样本一共 32 个样本。

经平均值加减两倍标准差检验,共 18 个样本中,有不同数目的元素浓度显著偏离平均值。其中仅是酸碱度显著偏离平均值有 2 个样本,本身或母质砂土,而所含元素偏低有四个样本。

经上下层检验,最后确定,哈尔滨地区的 100 个样本中,有 6 个样本,受到多种有害元素的污染,3 种样本受 1 种元素污染。

四、讨论

1. 哈尔滨地区的 100 个土样,有 8 个在哈尔滨市区内,有 77 个在哈尔滨市郊区,其余 15 个在市属县。经检验被确定多种有害元素污染的土样中,有 4 个样本采于市区内,占市内样本总数 50%,有两个样本取样于郊区,占郊区总样本数 2.6%;市属县没有占。这个结果一方面可以看到城市多种污染的严重性,另一方面也说明了这种检验法在一定程度上正确地反映了环境背景值的异常状态。

2. 这种检验法虽然考虑到每一个数据在剖面号、土层号及分析对象三维坐标上的综合关系,但将其有机联系起来考虑亦不够。并且,土壤类型影响未加考虑,特别对透水性能良好的砂土,这种因素影响更为显著。

3. 上下层检验法,还须对不同的元素采用不同的方法,因为它们在土壤中垂直运动的规律有差别。继而即可完成上下层检验法的计算机实现。

参 考 文 献

- [1] 杨学义,杨国治,环境科学,4(2),17(1983).
- [2] 车宇瑚,数学的实践与认识,(1),66(1978).
- [3] 方开泰,同上,(1),45(1978).
- [4] 方开泰,同上,(2),54(1978).

(下转第 34 页)

* 谢恩惠等,模糊聚类的一种机器算法(待发表)。

74%。值得指出的是在监测期间,氮氧化物的浓度都比较低。

8. 回归方程的检验: 在原始数据中任意抽出 5 天的数据不参加回归分析, 然后选择几个得到的回归方程, 计算这 5 天的 O_3 浓度值, 将计算值和实测值进行了比较, 结果见图 5 和表 2。利用当天的气象变量数据, 以及回归方程:

$$\begin{aligned} [O_3] = & 29.1 - 0.903[T] \\ & - 0.0797[RH] + 0.0385[T]^2 \\ & - 0.00134[RH]^2 + 64.3[K]^2 \end{aligned}$$

得到的 118 个计算值与实测值的相关系数为 0.63; 利用前一天的气象变量数据以及回归方程:

$$\begin{aligned} [O_3] = & 16.3 + 0.428[T] - 0.115[RH] \\ & + 30.6[K] \end{aligned}$$

得到的 118 个计算值与实测值的相关系数为 0.68。

表 2 日最大值回归方程的计算值与实测值比较

日期	4 月 19 日	5 月 2 日	5 月 15 日	5 月 20 日	6 月 4 日
实测值	62	39	41	64	50
计算值	48	38	46	60	48
相对误差	-22.6%	-2.6%	12.1%	-6.2%	-4.0%

由以上结果可以看出计算值和实测值之间吻合的还是比较好的。利用上面第二个回归方程, 平均日变化回归方程以及日最大值回归方程进行污染预报工作的尝试也是可行的。

三、小结

本实验结果表明, 在本地区春末、夏初时节大气中的臭氧浓度比较低。臭氧与气温、 K 值有较好的正相关关系, 与湿度一般呈负相关, 在臭氧与 K 值的相关关系中, 有两个小时的滞后效应。

同一地区在臭氧前体物的排放量基本保持稳定的情况下, 臭氧浓度的变化主要受气象条件的影响。在实验中发现大风天气的影响尤为突出, 昼夜大风可使全天的臭氧浓度一直保持在几个 ppb 上下。

由回归方程得到的计算值和实测值之间一般吻合得较好, 特别是利用前一天的气象数据计算第二天的臭氧浓度, 得到的相关系数远远高于 99% 的置信水平。

为了更好地了解大气中臭氧浓度和气象变量之间的统计相关性, 做好污染预报工作, 应该积累更多的长期监测数据和考虑更多的气象变量。本实验仅从现有条件对臭氧与气温, 湿度、 K 值的统计相关性做了初步探讨, 有待于今后进一步深入。

参 考 文 献

- [1] 荷兰飞利浦公司, PW9771 型臭氧监测仪说明书 (1979 年)
- [2] Jones, F. L. et al., *J. Air Pollution Control Association*, 31, 42-5 (1981)
- [3] 紫金山天文台, 《天文普及年历 (1981 年)》, (1980 年)
- [4] John E. Wals, *ibid.*, AD-71793.
- [5] 汪培庄, 模糊聚合论及其应用, 73-84 页 上海科学技术出版社, 1983 年.
- [6] M. A. Stephens, *GRA(USA)* AD-761308.
- [7] 林少宫, 基础概率与数理统计, 147-152 页 人民教育出版社, 1963 年.
- [8] 彭祖贻 孙毓玉, 模糊数学, (1), 125(1982).
- [9] 方开泰 马逢时, 应用数学学报, 5(4), 339(1982).
- [10] 浅见辉男等, 日本土壤肥料杂志, 50(2) 138-148 (1979).
- [11] 陈静生, 环境科学丛刊, 4(8), 1-12(1983).
- [12] (上接第94页)
- [13] 中国科学院土壤背景值协作组, 环境中若干元素的自然背景值及其研究方法, 23-34 页, 科学出版社, 1982 年.
- [14] 刘付强, 同上, 60-63 页.
- [15] 杨国治, 同上, 82-87 页.
- [16] 梁伟, 同上, 88-91 页.
- [17] John E. Wals, *GRA(USA)* AD-781499.
- [18] *ibid.*, AD-717934.
- [19] Jasper Paur Hendren, *GRA(USA)* AD-704780.
- [20] Frank E. Grubbs, *ibid.*, AD-781499.