

北京冬季气温的对应分析研究

黄嘉佑

(北京大学地球物理系)

提 要

对应分析是近年来在因子分析基础上发展起来的一种多元分析技术,它统一考虑空间和时间因素,使之在提取天气过程资料矩阵中的有用信息方面有更好的特点。

本文对对应分析法在分类中作了改进,提出向量夹角分角线的客观分类判据,并把这一方法应用到北京冬季气温分类上,结果表明它比其它分类法更直观、提取信息量更多,能较好地反映天气过程本质。

一、引 言

在气象要素场(或气象要素天气演变过程)的分类或分型的分析中,常常是取要素场上若干个空间点的某一时间段的历史资料样本,每个样本点称为时间点,由这些要素实测资料构成一资料矩阵。如果要分析研究时间点的分类,其做法通常是利用时间点上各空间点的相似性进行聚类分析^[7,11,12]或者利用因子分析中的 Q 型因子分析技术^[13]提取各时间点的公共因子,以各时间点在公共因子轴上的投影点在图上,用点子聚集的相似性进行分类。如果对象是针对空间点的地区分类,则以空间点为中心考察它们时间上的相似性用同样的方法进行分析,亦称为 R 型因子分析技术^[13]。这两个技术分别从不同需要出发来进行分类,它们并没有考虑时间与空间的联系方面,而实际的气象要素演变过程中时间和空间是有密切联系的。如何去综合分析它们的关联性来划分它们的不同类别正是对应分析方法所着重考虑的方面,因而使用这一方法来分析气象要素场或天气变化过程更具有其明显的优点。

对应分析是从1970年(由巴黎统计实验室教授 Benzeri 所提出)发展起来的一种因子分析技术^[1-2]。它综合 R 型及 Q 型因子分析的特点,从资料矩阵空间点和时间点之间联系性和对立性出发,它能在同一因子轴上同时表现时间和空间方面联系的密切程度,利用头两个公共因子作坐标轴,根据空间点和时间点在公共因子轴上的投影值点在平面图上的点子聚集状况作出分类,并作相应的物理意义解释^[3]。

在通常应用于地质矿物分类中的对应分析,仅根据因子平面图上点子聚集状况作分类^[2-3],这样任意性较大,而且在气象要素场或天气过程的分类中不象地质那样可以舍弃某些样品,而需要对全部样品进行划分。本文采用把空间点在因子平面图上看成为一个

向量,以向量夹角的分角线作为类别的分界线的客观判据,以改善这一方法所存在的任意性缺点。

本文使用这一方法对北京冬季气温变化过程(12月、1月、2月日平均气温)1951—1976年期间的资料,作时间点的分类。结果表明所分的各类具有明显的气温变化过程的气象特征,便于使用来作分类预报的依据,而且同一般的分类方法比较,还具有提取信息量高、直观等优点。

二、方 法

对气象要素场或天气过程的任一资料矩阵 $X(m \times n)$, 其中 m 表示时间点个数, n 表示空间点个数。资料阵各元素在全体中所占的比例为:

$$p_{ij} = x_{ij}/T \quad (1)$$

其中 x_{ij} 为原资料阵 X 中第 i 行 j 列的元素, p_{ij} 为新比例资料阵 P 中的元素, T 为原资料阵各元素之和,即

$$T = \sum_{i=1}^m \sum_{j=1}^n x_{ij} \quad (2)$$

然后对 P 阵每一行及每一列求和。令

$$\left. \begin{aligned} p_{i\cdot} &= \sum_{j=1}^n p_{ij} \quad (i = 1, 2, \dots, m) \\ p_{\cdot j} &= \sum_{i=1}^m p_{ij} \quad (j = 1, 2, \dots, n) \end{aligned} \right\} \quad (3)$$

考虑每个点在时间和空间上的相对权重比例,求一 Z 阵。在 Z 阵中的元素为

$$z_{ij} = \frac{p_{ij} - p_{i\cdot} p_{\cdot j}}{\sqrt{p_{i\cdot} p_{\cdot j}}} \quad (4)$$

Z 阵是一 $(m \times n)$ 矩阵,它相当于因子分析中资料标准化阵。从 Z 阵出发,计算

$$B = Z'Z \quad (5)$$

$B(n \times n)$ 阵就与因子分析中的 R 型对应,其中 $(\cdot)$ 表示矩阵转置。据矩阵分解定理, B 阵可分解为

$$B = V\Lambda V' \quad (6)$$

Λ 为 B 阵特征值组成的对角阵, V 为特征值所对应的特征向量组成的正交阵。

当然亦可类似于因子分析中的 Q 型,计算

$$A = ZZ' \quad (7)$$

对 A 同样可分解为

$$A = U\Lambda U' \quad (8)$$

可以证明, A 阵的对角阵 Λ 的非零元素与 B 阵的非零特征值一样, U 是 A 阵的特征值所对应的特征向量所组成的正交阵。如果 $m < n$ 时,解 A 阵的特征问题可以减少工作量或节省机器计算时间。

设第 k 个公共因子轴为 f_k , 第 j 个空间点在第 k 个因子轴上的坐标用下式求出

$$q_{jk} = \frac{v_{jk}}{\sqrt{p_{\cdot j}}} \quad j = 1, 2, \dots, n \quad (9)$$

其中 v_{jk} 为 B 阵第 k 个特征向量的第 j 个分量, 第 i 个时间点在第 k 个因子轴上的坐标为:

$$t_{ik} = \frac{1}{\sqrt{\lambda_k}} \sum_{j=1}^n y_{ij} q_{jk} \quad i = 1, 2, \dots, m \quad (10)$$

其中

$$y_{ij} = \frac{p_{ij}}{p_{\cdot j}} - p_{\cdot j} \quad \begin{cases} i = 1, 2, \dots, m \\ j = 1, 2, \dots, n \end{cases} \quad (11)$$

λ_k 为第 k 个特征值, 特征值的个数与 m 、 n 中最小者一致, 每个公共因子轴的方差贡献大小可用下式求出 (设 $m < n$)

$$G(k) = \lambda_k / \sum_{i=1}^m a_{ii} \quad k = 1, 2, \dots, m \quad (12)$$

其中 a_{ii} 为矩阵 A 的主对角线上元素。

如果头两个公共因子轴的方差贡献都比较大时 (例如它们之和达到占总方差的 80% 以上), 分别用头两个因子轴 (对应两个最大特征值) 构成二维平面图, 在图上点出空间及时间观测点并进行分类分析。

空间点在因子轴的平面上可看成为一个向量, 它们之间夹角余弦近似反映空间点之间的相关系数。时间点在平面图上亦可看成为向量, 它们之间夹角余弦以及和空间点向量之间的夹角余弦反映时间之间以及和空间点之间关系密切状况。因此, 以空间点向量为基础, 作向量间夹角的分角线, 两分角线之间区域内的时间点归为一类, 该类反映所夹的空间点的特征 (见图 1)。当几个空间点向量在图上比较接近时, 说明这几个空间点性质近似可以综合为一个代表点, 这时利用向量加法在图上作这几个向量的平均向量作为代表向量再参加分类 (见图 3)。

在使用气象要素距平值作对应分析时, 要注意由于距平值的特点会使 (2) 式 T 值变成接近或等于 0, 从而使 P 阵无法构成, 这时需对距平值加上适当的非零常数以避免发生这种情况。

三、分析结果

取北京冬季 (12—2 月) 月平均气温 1951—1976 年资料构成资料阵 ($m = 26, n = 3$), 以各月距平值出发作对应分析, 结果如图 1。

根据空间点向量 (图 1 中以月份数字表示) 相互分立的特点分为三类, 以两个月向量的分角线作为分类界线。

I 类: 以 12 月正距平为特点, 反映的天气过程是冬季由暖到冷的下降趋势。

II 类: 以 1 月相对其余月份较高值为特点, 冬季长期气温变化为升后降。

III 类: 以 2 月相对高值为特点, 各月演变为上升趋势。

这几类各年个例月气温变化曲线如图 2 所示。这种既反映各月特点又反映整个天气过程的分类法在天气过程分析中是有很大好处的。因为在一般的冬季气温预报中,往往以 1 月作为代表整个冬季来作预报,以 1 月的距平正负来估计冬季的暖冷,显然这不能完全反映冬季的天气过程演变特点。例如,1955 年 1 月气温距平值为 0.8°C ,1975 年为 0.9°C ,似乎都是暖年。但从图 2 可见,1955 年是先暖后冷,而 1975 年则是先冷后暖,它们反映不同的天气过程。往往在一般的寻找因子作单月预报,预报关系不稳定看来与此有密切关系。而对应分析的分类结果却能很好地把这两年区别开来。

本文再取北京冬季(11—3 月)1951—1976 年气温资料进行对应分析,结果见图 3。从图中可见,由于 11 月、1 月和 3 月的向量比较集中,用向量加法求其平均向量(图中用(11, 1, 3)表示),以它作为一类,结果亦分为三类。与 12—2 月的例子比较,在 I 类中仅

少 1957 年,II 类仅少 1970 年,这两年个例归到第 III 类。说明分类的结果在两个实例中基本是一致的,表明对应分析方法能提取主要信息,能反映天气过程的实质表现在不同的类别中。

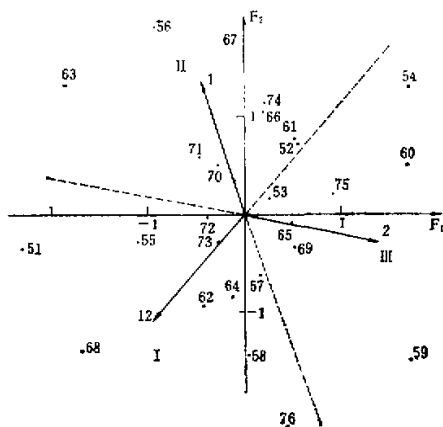


图 1 北京 12—2 月气温对应分析分类图

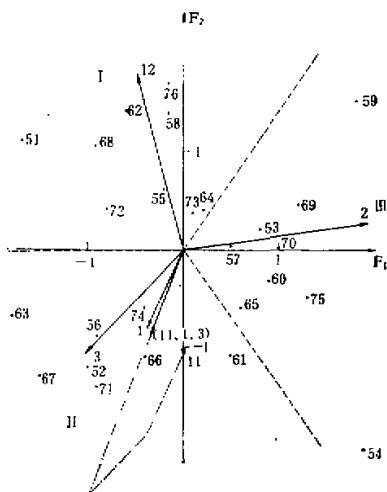


图 3 北京 11—3 月气温对应分析分类图

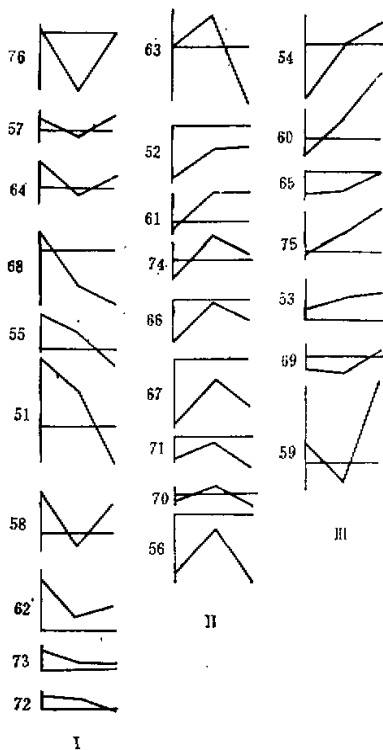


图 2 北京 12—2 月气温分类个例图

四、几种方法的比较

分类优劣的判别标准, Jaquet 等人^[5]已有详细论述。本文只取其中一个判据, 即用 Wilks 的 Λ 比值, 它是类内协方差阵的行列式与总协方差阵行列式之比, Λ 值愈小说明分类情况愈好。

本文对北京冬季 (12—2 月) 气温一例用距平值的对应分析 (A)、原值的对应分析 (B)、 $\cos \theta$ 平均权重串组法 (C)^[7]、距离系数 d 平均权重串组法 (D)^[7] 等四种客观分类法进行比较 (其中 $\cos \theta$ 及 d 为一般的相似系数和距离系数)。分别计算分类后的 Λ 值 (见表 1)。从表 1 可见, 距平值的对应分析在分类中是比较好的, 它的 Λ 值通过 F 分布信度 1% 的显著性检验 (见 [8] 中 227 页的检验)。D 法尽管 Λ 值最小, 但反映天气过程的物理意义不如 A 法好。例如 1953、61、75 年在曲线变化外形上比较相似, 在 D 法中是归到一类, 强调的仅是演变过程相似。而对应分析则照顾到各月正负距平的特点而把 1961 年分到另一类。又如 57 年与 59 年尽管变化外形相似, 但在对应分析中强调 1957 年 12 月是正距平, 变化近似于 I 类, 而 1959 年 2 月是正距平明显, 变化类型是上升趋势故归为 III 类。这些不一致正说明对应分析不但考虑时间点上的相似还考虑空间点的特点, 这点和一般仅用时间点上的相似为依据的相似分类法是不同的。

表 1 四种方法的 Λ 值

方 法	A	B	C	D
A	0.17	0.21	0.36	0.10

对应分析与因子分析的 R 型 (E) 及 Q 型 (F) 进行比较, 其明显的优点是客观。用文献 [4] 的方法对同一例作分析, 发现时间点在图上分布分散, 空间点向量又集中在第一因子轴的一边无法较客观地分类。用 F 法进行分析勉强可分为三类, 但仍带主观成份, 而且在头两个因子提取方差量占总方差比例上明显不如对应分析的 A、B 法, 比较结果见表 2。

表 2 头两个因子提取方差量比较

方 法	A	B	E	F
G (%)	99.99	99.99	79.65	93.66

五、讨 论

一般对小范围局地要素场 (或时间不太长的天气过程) 作对应分析所得的头两个公共因子常可以满足方差贡献较大的要求, 从而可以使用平面图形聚类方法, 这样具有直观, 可以全面了解点子相似程度的优点。但是如果分析对象是较大范围的要素场或者时间较长的天气过程时, 有时并不能很好地用头两个因子得到大部分方差解释。对于这种情况可以取使解释方差累积贡献达到 80% 以上的 s 个 ($s > 2$) 公共因子, 由它们组成一个 s

维的公共因子空间,把每一个空间点或时间点看成为 s 维空间中的向量,用一般的聚类分析方法(如文中 C 、 D 方法)首先把空间点进行聚类,使之分成不同类型,求出各类空间点的中心,然后利用矩心串组法^[12]以空间点为中心,把时间点围绕它们距离最近为原则进行分类。这一过程完全与二维平面聚类类似,只不过推广到高维空间去而已。

利用对应分析对天气过程分类结果作为预报量进行预报,实际是预报一个天气过程,在因子选取和预报效果的稳定性方面似乎会比一般的单月预报效果好些。本文对北京冬季(12—1月)气温距平值分类结果选取 10 个夏季长期环流因子(资料选自[10]),用逐步判别法^[9]选得 2 个因子,建立各类预报方程为:

$$y_I = \ln 0.39 - 6.49 + 0.25x_1 + 0.51x_2$$

$$y_{II} = \ln 0.34 - 10.14 + 0.26x_1 + 0.76x_2$$

$$y_{III} = \ln 0.27 - 2.54 + 0.11x_1 + 0.38x_2$$

其中 x_1 为 8 月 W 型因子, x_2 为 7 月太平洋副高西伸脊点位置(去掉 100 的数值)因子。试报 1977 年为 III 类,实况确为 III 类。

参 考 文 献

- [1] Teid, H., Correspondence factor analysis an outline of its method, *Journal of the International Association for Mathematical Geology*, 7, 3—12, 1975.
- [2] 中国地质科学院地矿所四室方法组, 对应分析(油印本), 1977.
- [3] Teil, H. and J. L. Cheminee, Application of correspondence factor analysis to the study of major and trace elements in the Erta Ale Chain (Afar, Ethiopia), *Journal of the International Association for Mathematical Geology*, 7, 13—30, 1975.
- [4] Gabriel, K. R., Analysis of meteorological data by means of canonical decomposition and biplots, *J. Appl. Meteor.*, 11, 1071—1077, 1972.
- [5] Jaout, Froidevanx and Vernet, Comparison of automatic classification methods applied to lake geochemical samples, *Journal of the International Association for Mathematical Geology*, 7, 237—266, 1975.
- [6] McCammon, R. B. Concepts in Geostatistics. Springer-Verlag Berlin Heidelberg. New York. 1975.
- [7] John, C. Davis, Statistics and Data Analysis in Geology. John Wiley & Sons. New York. 1973.
- [8] Cooley, W. W. and Lohnes, P. R. Multivariate Data Analysis. John Wiley & Sons, inc, New York. 1971.
- [9] 中国科学院计算所杨自强, 逐步判别(油印本).
- [10] 中央气象台长期天气组, 长期天气预报技术经验总结, 附录, 1976.
- [11] 中国科学院地质研究所, 数学地质引论.
- [12] 曹鸿兴, 串组分析、多元分析资料汇编(II), 中国科学院数学研究所资料室印, 1975.

A STUDY OF THE TEMPERATURE IN WINTER OF BEIJING BY THE USE OF THE CORRESPONDENCE FACTOR ANALYSIS METHOD

Huang Jia-you

(Department of Geophysics, Beijing University)

Abstract

The correspondence factor analysis method is a multi-variable analysis technique, which has been developing in recent years based on the foundation of the factor analysis. In this technique both space factor and time factor are taken into account, so that it has advantage in the aspect of extracting the useful information from the data matrix.

In this article the author has made some improvements on the correspondence factor analysis method and proposed an objective classification criterion for the bisector of the angle made by the two vectors. Using this method to classify the temperature in winter of Beijing, the results show that this method is better than other classification methods. It is more audio-visual, the amount of information extracted is more than that by other methods, and it can describe the characters of the synoptical process quite well.