

一种新的气候分型区划方法

丁裕国¹ 张耀存² 刘吉峰¹

¹ 南京信息工程大学, 江苏省气象灾害重点实验室, 南京 210044

² 南京大学大气科学系, 南京 210093

摘 要 从理论上证明统计聚类检验 (CAST) 与旋转经验正交函数或旋转主分量分析 (REOF/RPCA) 用于气候聚类分型区划的关联性。研究表明, CAST 在一定的意义上可认为是 REOF/RPCA 用于气象要素场 (气候) 分型区划的理论基础。由此, 作者提出 CAST 与 REOF/RPCA 相结合的一种新的分型区划方法, 并用仿真随机模拟资料和实例计算验证了理论与实际结果的一致性, 从而证实了这种分型区划方法的有效性及其优点。

关键词 主分量分析 统计聚类检验 气候区划

文章编号 1006-9895 (2007) 01-0129-08

中图分类号 P462

文献标识码 A

A New Cluster Method for Climatic Classification and Compartment Using the Conjunction Between CAST and REOF

DING Yu-Guo¹, ZHANG Yao-Cun², and LIU Ji-Feng

¹ Nanjing University of Information Science & Technology, Jiangsu Key Laboratory of Meteorological Disaster, Nanjing 210044

² Department of Atmosphere Sciences, Nanjing University, Nanjing 210093

Abstract It is theoretically demonstrated that the Cluster Analysis with Statistical Test (CAST) by Yao C. S. (1997) has a close relationship with the rotated empirical orthogonal function or principal component analysis (REOF/RPCA) to climatic classification and cluster compartment. Then, a new cluster method for climatic classification and compartment is proposed, which is theoretically based on CAST combined with REOF/RPCA in order to decide to accept or reject the assumption of a climatic category according to the χ^2 test from CAST. Euclidean distance coefficient d_{ij} (or c_{ij}) defined by the standard raw data in classical cluster analysis is related to the correlation coefficient of the raw data, both of them possess certain function relationship. Therefore, the calculation steps of the new cluster method proposed are, first, putting each principal component from RPCA to a meteorological variable field instead of the elected each central station corresponding to climatic classifications, second, determining the number of station (or variation) in every climatic category by using the χ^2 test from CAST. Then, the PCA or EOF is once again made at the stations with large loading, corresponding to every climatic category, and on the basis of the above result, the last climatic cluster and classification may be obtained.

The result of the research shows that the CAST may considered a theoretical basis of the REOF/RPCA applied to climatic classification and cluster compartment, because REOF is a data-analytic technique which is based on the Principal Components Analysis (PCA) developed by Hotelling (1933), and it is not only applied to climatic classification and compartment but also used in other problems. In this paper, the theoretical relationship between CAST and REOF is verified by means of the statistical simulation experiment and real observational data calculations. For

收稿日期 2005-07-28, 2005-11-01 收修定稿

资助项目 国家自然科学基金资助项目 40675043, 江苏气象灾害重点实验室项目 KLME050209

作者简介 丁裕国, 男, 1941 年出生, 教授, 长期从事统计气候和气候变化诊断预测以及陆面过程参数化的研究。

E-mail: dygnim@jionline.com

example, based on the daily data of maximum and minimum temperatures (1957—2001) from 203 meteorological stations in China by means of CAST combined with REOF/RPCA, the patterns of compartment for the extreme temperature sequences from 203 stations in China are obtained. This example also verifies that the new cluster method in this paper for climatic classification is more reasonable and reliable to determine each climatic category by using REOF (or RPCA) twice.

Key words principal component analysis, cluster analysis with statistical test, climatic classification and cluster compartment

1 引言

多年来, 学者们运用旋转经验正交函数或旋转主分量分析 (REOF/RPCA) 进行各种气象要素场的分型区划, 其效果很好^[1~5], 但往往并不能解释其数学物理意义。20 世纪 90 年代, 么枕生^[6]根据气候统计理论提出了一种新的具有统计检验的聚类分析 (cluster analysis with statistied test, 简称 CAST) 用于气候分型区划。这种更为客观的数值分类法, 改进了一般数值分类统计方法中不考虑聚类样品距离系数的统计显著性的重大缺陷^[6]。为了更好地发展这一新的学术观点, 本文将论证 CAST 与 REOF/RPCA 用于气候聚类分型区划的理论联系。根据“统计聚类检验”思想, 我们完全可导出 REOF/RPCA 气候分类区划方法, 这样就从理论上找到了用 PCA 一类方法作为气候分类区划的理论依据。同时, 也印证出 CAST 与 PCA 用于气象场分类区划具有等价性和互补性的事实。为了减少计算中可能产生的误差, 本文将 CAST 与 PCA 相结合, 提出了一种新的分型区划方法。

2 理论证明

所谓 CAST 聚类, 就是具有显著性检验标准的聚类分析。一般可有两种计算方案: 均匀聚类和中心聚类^[6, 7]。对于地理空间上的气候区划来说, 用中心聚类方案更适合。

假设样本容量为 n , 具有 p 个网格点 (或测站) 所构成的某一气象变量场 $\mathbf{X}$ (约定序列已标准化), 可以证明, 其两两格点或测站之间的欧氏距离^[6, 7]为:

$$d_{ij} = \left\{ \sum_{k=1}^n (x_{ki} - x_{kj})^2 \right\}^{1/2} = \{2n(1 - r_{ij})\}^{1/2}, \quad (1)$$

式中, x_i 和 x_j 分别为气象场中不同序号 $i=1, 2, \dots, p-1$ 和 $j=1, 2, \dots, p$ 的各变量。而 r_{ij} 为两者的线性相关系数, n 为其样本容量。文献^[1]已

证明, 属于各个同类样品之间的欧氏距离 d_{ij} 为随机变量, 其抽样分布具有统计特征量,

$$E(d_{ij}) = \sqrt{2n(1 - \rho_{ij})}, \quad (2)$$

$$\text{var}(d_{ij}) = 1 - \rho_{ij}, \quad (3)$$

这里, ρ_{ij} 为理论相关系数 (总体相关系数), 在同类样品中, 由于欧氏距离 d_{ij} 符合正态分布, 其极大似然估计值可写为

$$\hat{d} = \sum_{i < j} w_{ij} d_{ij} / \sum_{i < j} w_{ij}, \quad i = 1, 2, \dots, q-1; \\ j = 1, 2, \dots, q, \quad (q \leq p) \quad (4)$$

其中,

$$w_{ij} = \frac{1}{\text{var}(d_{ij})} = \frac{1}{1 - r_{ij}}. \quad (5)$$

由此构造服从自由度为 $\eta-1$ 的 χ^2 分布统计量^[1],

$$\chi^2 = 2n \sum_{i < j} w_{ij} (c_{ij} - \hat{c})^2, \quad (6)$$

其中,

$$c_{ij} = \frac{d_{ij}}{\sqrt{2n}} = \sqrt{1 - r_{ij}}, \quad (7)$$

$$\eta = \frac{1}{2}q(q-1), \quad (8)$$

于是, 有

$$\begin{cases} w_{ij} c_{ij} = \frac{1}{\sqrt{1 - r_{ij}}}, \\ \hat{c} = \sum_{i < j} w_{ij} c_{ij} / \sum_{i < j} w_{ij} \end{cases}. \quad (9)$$

在 (4)~(8) 式中, q 为同类样品中变量的个数, 对于气象变量场 $\mathbf{X}$ 而言, q 即为变量场中同一区 (型) 的网格点数 (或站点数)。 η 是与自由度有关的 (6) 式中求和项数。聚类统计检验 (CAST) 方法就是依据 (6) 式逐一计算出拟并入该区域的相应站点的 χ^2 统计量, 并检验零假设 $H_0: d_{ij} = \hat{d}$ 和备择假设 $H_1: d_{ij} \neq \hat{d}$ 。在给定信度 α 下, 可寻求临界区间

$$\chi^2 \geq \chi_{1-\alpha}^2 (\eta-1), \quad (10)$$

如若

$$\chi^2 < \chi_{1-\alpha}^2 (\eta-1), \quad (11)$$

则表明该测站可归于同类区域，否则，可在一定信度下判定其不属于该类区域。根据 (1) 式，我们可直接计算出每两两测站间的相关系数，从而构成相关系数矩阵。在此基础上，选择中心站点，依据 (1)~(11) 式，就可编制计算程序进行分类。显然，为了并入某一区类，我们希望加入的站点满足 $\chi^2 < \chi^2_{1-\alpha}(\eta-1)$ ，即 χ^2 值愈小愈好。但 CAST 存在着一个不确定性的问题：即如何选取聚类中心站点并无客观标准。在同一气候区内，如果选择不同的中心站点，其分区结果可能并不完全一致。那么，怎样才能使选出的中心站点更能代表该气候区的典型特征？

文献[8]已经证明，EOF 能客观地识别出同类振荡型的空间分布，即旋转前或旋转后的载荷场上，每一个空间点所对应的变量只与一个主分量存在高相关，那么该向量场的高值区就代表了该区域的典型特征。因此，文献[7]认为具有高荷载的站点就是该气候区的中心，选取其作为聚类分析的中心站点最具代表性。

根据 REOF 理论，可以证明各主分量的最高荷载中心实际上可作为聚类分区的中心站点。因为上述具有 p 个网格点（或站点）的气象变量场序列 $\mathbf{X} = (x_1, x_2, \dots, x_p)'$ ，实质上就是具有某种学科意义的多元正态变量。假定对资料已作标准化预处理，即有

$$E(\mathbf{X}) = (\mu_1, \mu_2, \dots, \mu_p)' = 0, \quad (12)$$

或者

$$\mathbf{X} \sim N(0, \mathbf{\Sigma}), \quad (13)$$

其中 $\mathbf{\Sigma}$ 为多元变量的协方差阵。

显然，其各变量间相关系数矩阵式应为

$$\mathbf{R} = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1p} \\ r_{21} & r_{22} & \cdots & r_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ r_{p1} & r_{p2} & \cdots & r_{pp} \end{bmatrix}. \quad (14)$$

由于 $\mathbf{R} = \mathbf{R}'$ 为实对称矩阵，根据对称阵的对角化和正交变换理论^[9]，有

$$\mathbf{L}'\mathbf{R}\mathbf{L} = \mathbf{A}, \quad (15)$$

定义主分量

$$\tilde{\mathbf{Y}} = \mathbf{L}'\tilde{\mathbf{X}}, \quad (16)$$

则有

$$\tilde{\mathbf{Y}}'\tilde{\mathbf{Y}} = \mathbf{L}'\tilde{\mathbf{X}}\tilde{\mathbf{X}}'\mathbf{L} = \mathbf{L}'\mathbf{R}\mathbf{L} = \mathbf{A}, \quad (17)$$

其中， $\tilde{\mathbf{X}}$ 、 $\tilde{\mathbf{Y}}$ 均为标准化变量。根据 (1) 和 (7) 式

可得

$$r_{ij} = 1 - \frac{d_{ij}^2}{2n} = 1 - c_{ij}^2. \quad (18)$$

考虑 (18) 式的关系，(17) 式又可等价地写为

$$\tilde{\mathbf{Y}}'\tilde{\mathbf{Y}} = \mathbf{L}'\tilde{\mathbf{X}}\tilde{\mathbf{X}}'\mathbf{L} = \mathbf{L}'\mathbf{R}\mathbf{L} = \mathbf{L}'(\mathbf{J} - \mathbf{C})\mathbf{L} = \mathbf{A}, \quad (19)$$

上式中， $\mathbf{J}$ 代表了元素全为 1 的方阵， $\mathbf{C}$ 代表了元素为 $c_{ij}^2 (i, j=1, \dots, p)$ 的矩阵，显然，相关系数矩阵 $\mathbf{R}$ 是距离系数 c_{ij} 平方的函数矩阵。 $\tilde{\mathbf{Y}}$ 为标准化主分量，(19) 式中相应的特征值就是标准化主分量方差，

$$\text{var}(\tilde{\mathbf{Y}}) = \mathbf{L}'\tilde{\mathbf{X}}\tilde{\mathbf{X}}'\mathbf{L} = \tilde{\mathbf{A}} = \text{diag}(\tilde{\lambda}_1, \tilde{\lambda}_2, \dots, \tilde{\lambda}_p). \quad (20)$$

由此可见，对气象变量场相关系数矩阵的主分量正交变换即 PCA 的运算过程就等价于对场内各变量所构成的距离系数矩阵的某种函数的正交变换过程。在标准化变量条件下，任一主分量的载荷值 l_{hj} 实质上就是主分量与原变量的相关指标，可以证明 $l_{hj} \sqrt{\lambda_h}$ 代表了主分量振型与原序列振型的相关系数，实际上 PCA 是从原变量场序列中逐一提取代表原气象变量场的各种主要振荡信号型^[3, 7, 8]。气候型的分类区划，实质上就是按载荷值的高低，将与某一主分量最为相似的高相关（或高荷载区）振荡型聚为同一类型；反之则属于其他类型。载荷场经过正交旋转变换，可使载荷场方差加大，即载荷值分异性更明显，这就使原来与某一主分量最为相似的高荷载区更加突出，而其余的低值区相形见绌^[5]。采用 CAST 作气候分类，以中心聚类为最优^[6]。同一气候区内各站点间彼此相似，其中任何站点的气候特征都可认为是与某一个所谓中心站点的气候特征最为相似。在理论上，总可假定存在某一中心站点的变量序列，其在时间域上的振荡型就是相应的主分量振荡型。如前所述，既然荷载值 l_{hj} 实际上代表了主分量与原变量振荡型的相似程度（以相关系数为度量）。而根据 (5)~(8) 式可见，站间距离可以直接转化为站间相关系数。因此，利用 χ^2 统计检验对距离系数的检验等价于对相关系数的检验。其聚类判据也是等价的。

此外，任一个气象变量场的 PCA 收敛性都与场的相关结构有关^[9]，一般地说，场内各站均有高相关，即两两相关密切且均匀，则收敛必然较快（其前 1 或 2 个主分量即可占有相当大的方差百分

比),即整个场的随机性减小,反之,如场内各站相关较差,必然使整个场的随机性增大,必然收敛慢。在特殊情况下,假定场内有 p 个变量,其相关矩阵中各变量间的相关系数都很大,接近 1,则可证明其特征方程近似为

$$\lambda^{p-1}(\lambda - p) = 0, \quad (21)$$

由此得到

$$\begin{cases} \lambda_1 = p, \\ \lambda_2 = \lambda_3 = \cdots = \lambda_p = 0, \end{cases} \quad (22)$$

即相应的最大特征值已接近或等于其总方差,而其余特征值则接近 0;相反,假定场内 p 个变量,其各变量间的相关系数都很小,接近为 0,则其特征方程近似为

$$(1 - \lambda)^p = 0, \quad (23)$$

由此得到

$$\lambda_1 = \lambda_2 = \cdots = \lambda_p = 1, \quad (24)$$

即表明该场序列无中心站点可言,区域各站不能聚为一类。由此可见,将已经用 CAST 或 PCA 分类得到的各区域再分别用 PCA 展开,其第 1 主分量特征值(即方差值)必然有上述(22)与(24)式的结果。因此,利用 χ^2 统计检验对距离系数的检验恰好等价于对相关系数的检验,其聚类判据必然也是等价的。因此,由 CAST 或 PCA 聚类所得高相关或高载荷的变量子集,必然满足(22)与(24)式的结果。根据上述理论,针对各主分量所对应的高载荷变量子集(或子场)重新再做二次 PCA 展开,就成为对 PCA 与 CAST 聚类结果一致性的验证方法。

当然,根据(14)式,对任一主分量而言,也可将其线性组合式写为下列形式:

$$y_{jt} = \sum_{h=1}^q l_{hj} x_{ht} + \sum_{s=q+1}^p l_{sj} x_{st}, \quad (25)$$

其中, $h=1, \cdots, q$ 为高载荷子集序号, $s=q+1, \cdots, p$ 为低载荷子集序号。这样,主分量与各变量的相关密切程度可按其高低载荷区分为两部分。从理论上,利用线性回归分析,就可建立相应的线性回归方程,以便客观地验证和确定 PCA 与 CAST 聚类结果的一致性,这就是另一种验证方法。因篇幅所限,对此,笔者将另文探讨。

综上所述,基于 PCA 与 CAST 相结合用于气候分型区划的一种新的聚类方法,其步骤可归纳为:(1)对气象变量场首先作 PCA 或 RPCA/

REOF,并以各主分量作为气候区划聚类的中心变量;(2)其次,根据与其相应的荷载场的荷载值大小,采用 CAST 的 χ^2 统计检验,按给定信度确定接受或拒绝的变量逐步引入属于各自的区域;(3)为了对各拟分区域已引入变量子集是否符合 CAST 检验标准作验证或鉴别,再次重新对各拟分区的变量作新一轮的 PCA,由此最终确定各个分区或聚类的结果。

3 随机模拟数据的聚类试验

假设有气象场 $\mathbf{X}=(x_1, x_2, \cdots, x_{12})$ 的仿真随机模拟数据序列,其序列长度为 $n=50$ 。表 1 列出了按给定准周期加白噪声而产生的仿真随机数据序列所求得的其相关系数矩阵。首先,对其做 PCA 展开,并将各主分量作为聚类中心,根据 CAST 方法,分别按(7)至(9)式计算统计量 c_{ij} 和 $w_{ij}c_{ij}$ 以及 $\hat{c}$ 和 η 。由表 1 可见,变量 6、7、8、9 大致可聚为一类,根据(6)式及(10)、(11)式我们计算了 PC1 对变量 6、7、8、9 的 χ^2 统计量。结果表明,由于 $\chi_a^2(\eta-1=4)=7.815$,而

$$\chi^2(P_{c1} + 6 + 7 + 8 + 9) = 2n \sum_{i < j} w_{ij} (c_{ij} - \hat{c})^2 = 0.5918 < \chi_a^2.$$

可见,第一主分量 P_{c1} 可作为变量 6、7、8、9 的聚类中心。但如果再加入其他变量,则不能进入。例如,加入变量 12,由表 1,根据(1)~(6)式计算得到

$$\chi^2(P_{c1} + 6 + 7 + 8 + 9 + 12) = 2n \sum_{i < j} w_{ij} (c_{ij} - \hat{c})^2 = 66.01,$$

因为

$$\hat{c} = \frac{18.3207}{77.3423} = 0.2369,$$

$$w_{1,12} c_{1,12} = \frac{1}{\sqrt{1-r_{1,12}}} = \frac{1}{\sqrt{1.48}} = 0.8220,$$

$$w_{1,12} = \frac{1}{1-r_{1,12}} = 0.6756$$

$$\chi_a^2(\eta-1=5) = 11.070,$$

$$\chi^2 = 66.01 > \chi_a^2.$$

所以,上述结果证明,不能引进变量 12。显然,其他变量更不能引进了。

另一方面,由各主分量及其荷载向量(表 2)可见,前 4 个荷载场的高值中心(表中有下划线

的) 十分明显,基本上其载荷都在 0.50 左右,并与其他载荷分量有明显分异。计算表明,前 5 个主分量的方差百分比总和已达 96.0%以上,其中第 1~4 主分量分别为 32.73%、22.26%、18.99%、16.64%,总计为 90%左右。这就表明,按前 4 个主分量的荷载值聚类是完全合理的。为此,我们分别将其列于表 2 和表 3 中,显然对主分量荷载场旋转后(表 3),每一主分量高荷载值具有更大的分异。

这样,再对其所构成的 4 个子场作新一轮的 PCA 展开,结果表明,它们各自的第 1 主分量方差都达到相当大的百分比,其中变量 6~9 的 PC1 方差为 97.04%,变量 3~5 的 PC1 方差为 89.04%,变量 10~12 的 PC1 方差为 77.23%,变量 1~2 的 PC1 方差为 100%。这正是 (28) 式和 (31) 式所表明的一致性结果。如将变量 3~4 作为子场展开,其 PC1 的方差恰好为 100.0%。这就意味着,用 CAST

表 1 初始相关矩阵
Table 1 The correlation coefficient matrix of the raw data

序号 No.	1	2	3	4	5	6	7	8	9	10	11	12
1	1.00	<u>0.99</u>	−0.02	0.0	−0.01	0.01	0.00	0.00	0.01	0.00	0.00	0.01
2		1.00	−0.03	−0.02	−0.02	−0.02	−0.02	−0.01	−0.02	−0.02	−0.02	−0.01
3			1.00	<u>0.99</u>	<u>0.75</u>	−0.02	−0.04	−0.03	−0.04	−0.03	−0.04	−0.02
4				1.00	<u>0.75</u>	−0.01	−0.03	−0.02	−0.03	−0.02	−0.02	0.00
5					1.00	0.00	−0.02	−0.02	−0.04	−0.03	−0.03	−0.02
6						1.00	<u>0.98</u>	<u>0.94</u>	<u>0.89</u>	0.12	0.06	0.02
7							1.00	<u>0.99</u>	<u>0.95</u>	0.12	0.04	0.06
8								1.00	<u>0.99</u>	0.10	−0.01	0.07
9									1.00	0.07	−0.05	0.09
10										1.00	<u>0.85</u>	<u>0.74</u>
11											1.00	0.34
12												1.00

注：表中有下划线的数值为显著相关
Note: The values with underlines are the significant correlation

表 2 PCA 的前 4 个主分量荷载值分布
Table 2 Distribution of the loading values of the first four principal components

主分量 Principal component	荷载值 Loading value											
	1	2	3	4	5	6	7	8	9	10	11	12
1	0.01	0.00	−0.07	−0.07	−0.07	<u>0.46</u>	<u>0.47</u>	<u>0.46</u>	<u>0.46</u>	−0.22	−0.14	−0.24
2	−0.08	−0.09	<u>0.55</u>	<u>0.54</u>	<u>0.54</u>	0.02	0.00	0.00	0.00	−0.18	−0.17	−0.20
3	−0.06	−0.10	0.15	0.17	0.15	0.15	0.18	0.19	0.20	<u>0.52</u>	<u>0.54</u>	<u>0.46</u>
4	<u>0.70</u>	<u>0.69</u>	0.08	0.10	0.08	0.03	0.02	0.02	0.01	0.03	0.03	0.04

注：其中有下列线的数值表示显著荷载值
Note: The values with underlines are the significant loading values

表 3 旋转后 PCA 的前 4 个主分量荷载值分布
Table 3 Distribution of the loading values of first four principal components for the rotated PCA

主分量 Principal component	荷载值 Loading value											
	1	2	3	4	5	6	7	8	9	10	11	12
1	0.01	0.00	−0.15	−0.13	−0.15	0.95	0.95	0.95	0.94	−0.44	−0.29	−0.49
2	−0.14	−0.16	0.94	0.93	0.92	0.03	−0.01	0.01	−0.01	−0.31	−0.28	−0.34
3	−0.10	−0.16	0.24	0.28	0.25	0.24	0.29	0.30	0.33	0.84	0.87	0.74
4	0.98	0.97	0.11	0.14	0.11	0.04	0.03	0.03	0.02	0.05	0.04	0.06

或 PCA 分类得到的各个区域(高载荷站点)再分别用 PCA 展开,必然得到(28)与(31)式的结果。而上述计算结果中,绝大部分其特征值已接近于 1.0,但也有较小的 PC1 方差,如变量 3~5,一旦去掉第 5 变量,则 3、4 变量的 PC1 方差即达 1.0,可见第 5 变量是处于区域边界过渡带。同理,其他类似情况则相仿。

4 实例验证

对我国境内夏季极端高温和冬季极端低温的年际振荡类型采用 CAST 与 RPCA 相结合的方法(站点分布图略)分型区划,现以前者为例,其步骤大致可归结为:(1)取 $n=45$ (年), $p=203$ (站),首先建立各站夏季极端高温年际记录的相关矩阵。(2)然后对其相关矩阵做 REOF/RPCA,将各主分量及所对应的高荷载区视为选取的聚类中心,根据(1)~(8)式计算相应的 χ^2 统计量,并给出一定信度 α 条件下分区的检验判据。(3)为了更准确合理地分区,对已分的各区边界(尤其是那些边界过渡区或交叉混合区)加以鉴定,检测分区的合理性。从给出的前 12 个主因子旋转后的荷载场空间分布可见,第 1 主分量高荷载区位于我国东北地区,荷载值大部分在 0.6 以上,最大可达 0.954。其余地区的荷载值均比较小。第 2 主分量高荷载区主要以四川盆地为中心,荷载绝对值达 0.91,其余各区荷载值都在 $-0.2 \sim 0.2$ 之间。第 3 主分量高荷载区则主要集中在我国的滇南地区,其余各区荷载值更小,大都在 $-0.1 \sim 0.1$ 之间,显示出这一地区极端温度变化的独特性。第 4~11 主因子的高荷载区则分别位于我国西北,南疆的东南部及华南沿海、天山及其以北的新疆地区、青藏高原的东南缘和西北地区中部、黄淮流域等地区。(4)对各个分区再作旋转主分量分析(ROF),就可更准确合理地鉴定出分区的合理性。表 4 列出了夏(冬)季极端高(低)温各分区第一主分量的方差贡献(%).不难看出,它们基本上符合(20)与(22)式的类似结果,当然,由于实测资料场多频振荡的复杂性,并不一定都能达到那么高的方差比率。图 1 和图 2 绘出了应用该方法对我国夏季极端高温和冬季极端低温的分区结果,图中所标区域序号与表 4 一致。这就从实际应用中证实了 CAST 与 REOF 相结合的方法用于气候分类区划的可行性。

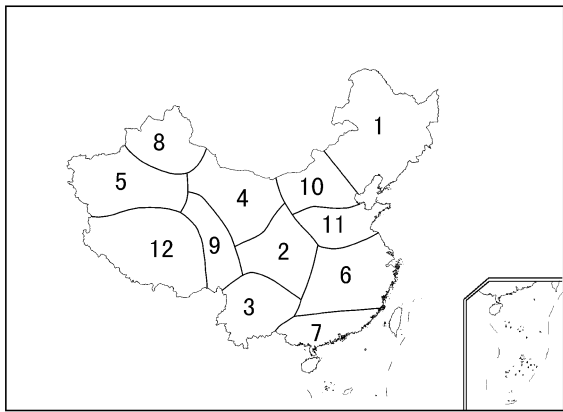


图 1 由新聚类方法得到的我国夏季极端高温区划示意图

Fig. 1 The compartment scheme for the maximum temperature in summer over China by the new cluster method

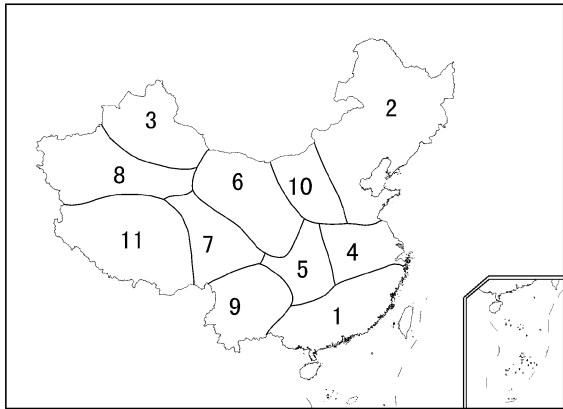


图 2 由新聚类方法得到的我国冬季极端低温区划示意图

Fig. 2 The compartment scheme for the minimum temperature in winter over China by the new cluster method

5 结论与评述

(1) 由于 REOF 方法本身并不仅仅只用于气候分类区划,正是因为这一点,本文才从 CAST 与 REOF 的联系中阐明 REOF 具有分型区划功能的基本原理与 CAST 方法的内涵完全一致。从理论上证明么枕生^[6,7]提出的具有统计检验的聚类分析(CAST)方法恰好可作为通常 PCA/EOF 对气象变量场分型区划的理论基础。文献[8]又证明 PCA 分型区划实际上是对变量场序列各种振荡型的空间分布的区划,利用旋转的 RPCA/ROF 对荷载场所作的旋转变换,则更能达到此目的。理论上总可假定,存在着某一中心站变量序列,其在时间域上的振荡型就是相应的主分量振荡型。既然荷载值 L_{hj}

表 4 夏(冬)季极端高(低)温区第一主分量方差贡献百分比

Table 4 The percentage of variances for the first principal component of the max temperature (min temperature) in summer (winter) in each region

高温区 $T_{\max}$ region	地理位置 Location	方差贡献 Variance contribution (%)	低温区 $T_{\min}$ region	地理位置 Location	方差贡献 Variance contribution (%)
1	东北区 Northeast China	58	1	华南及东南沿海 South China and the coastal region of Southeast China	68
2	四川、重庆 Sichuan, Chongqing	62	2	东北及华北东部 Northeast China and the eastern North China	72
3	西南地区 Southwest China	51	3	北疆地区 North Xinjiang	69
4	西北地区中部 Middle Northwest China	54	4	淮河和长江中下游 The Huaihe River and the middle and lower rea- ches of the Yangtze River	71
5	新疆南部 Southern Xinjiang	72	5	华中区 Middle China	65
6	华东区 East China	60	6	内蒙古西部及华中北部 Western Inner Mongolia and Northern Mid- dle China	70
7	华南沿海区 The coastal region of South China	56	7	青海及西藏东部 Qinghai and eastern Tibet	68
8	天山及其以北的新疆 Tian Shan and Xinjiang north of it	61	8	南疆地区 South Xinjiang	69
9	西藏东南部 Southeastern Tibet	56	9	西南地区 Southwest China	67
10	华北北部及内蒙古中部 Northern North China and Middle Inner Mongolia	69	10	华北西部和内蒙古中部 Western North China and Middle Inner Mon- golia	72
11	黄淮流域 Huanghe-Huaihe valley	73	11	青藏高原区 The Tibetan Plateau	72
12	青藏高原区 The Tibetan Plateau	72			

实际上代表了主分量与原变量振荡型的相似程度(以相关系数为度量),而站间距离又可以直接转化为站间相关系数。因此,利用 χ^2 统计检验对距离系数的检验就等价于对相关系数的检验,其聚类判据也是等价的。这就是为什么 REOF 可用于气象变量场分型区划的理论基础。

(2) 仿真随机数据模拟表明,对气象变量场的分型区划可将上述两种方案结合,第一步,先对其作 PCA/EOF 或 RPCA/REOF;第二步,再用各主分量为聚类中心,通过荷载相关阵用 CAST 方法聚类;第三步,对各高荷载区(或 CAST 聚类)再作新一轮 PCA/EOF,以便最终验证分区的合理性。

(3) 实例计算表明,采用统计聚类检验

(CAST) 与旋转主分量(RPCA)变换方法相结合,就可更准确合理地鉴定出分区的合理性,其分区的可靠性加大。因此,在给定场的各个变量相关系数矩阵的基础上,用各主分量作为聚类中心,借助于 CAST 方法与 RPCA 分类结果完全一致,再以相应分区的子场 PCA 所得第一主分量方差贡献作为验证,就可准确完成聚类区划。

参考文献 (References)

[1] Horel J D. A rotated principal component analysis of the inter-annual variability of the Northern Hemisphere 500 mb height field. *Mon. Wea. Rev.*, 1981, **109**: 2080~2092

[2] Richman M B. Review article rotation of principal components. *J. Climate*, 1986, **6** (3): 293~333

[3] Preisendorfer R W. *Principal Component Analysis in Mete-*

- orology and Oceanography*. Elsevier, Amsterdam, 1988. 425pp
- [4] 黄嘉佑. 气象统计分析与预报方法. 北京: 气象出版社, 1990. 182~188
Huang Jiayou. *Meteorological Statistical Methods in Analysis and Prediction* (in Chinese). Beijing: China Meteorological Press, 1990. 182~188
- [5] 周家斌, 黄嘉佑. 近年来中国统计气象学的进展. 气象学报, 1997, **55** (3): 297~305
Zhou Jiabin, Huang Jiayou. Advances in statistical meteorology in China in recent years. *Acta Meteor. Sinica* (in Chinese), 1997, **55** (3): 297~305
- [6] 么枕生. 用于数值分类的聚类分析. 海洋湖沼通报, 1994, (2): 1~12
Yao C S. A method of cluster analysis for numerical classification. *Transactions of Oceanology and Limnology* (in Chinese), 1994, (2): 1~12
- [7] Yao C S. A new method of cluster analysis for numerical classification of climate. *Theor. Appl. Climatol.*, 1997, **57**: 111~118
- [8] 么枕生. 载荷相关模式用于气候分类与天气气候描述. 气候学研究——气候与环境. 北京: 气象出版社, 1998. 1~9
Yao Zhensheng. A loading correlation model for climatic classification with climatic descriptions in synoptic climatology. *Climatology Studies—Climate and Environment* (in Chinese). Beijing: China Meteorological Press, 1998. 1~9
- [9] 丁裕国, 梁建茵, 刘吉峰. EOF/PCA 诊断气象变量场问题的新探讨. 大气科学, 2005, **29** (2): 307~313
Ding Yuguo, Liang Jianyin, Liu Jifeng. New research on diagnoses of meteorological variable fields using EOF/PCA. *Chinese Journal of Atmospheric Sciences* (in Chinese), 2005, **29** (2): 307~313
- [10] 么枕生, 丁裕国. 气候统计. 北京: 气象出版社, 1990. 624~629
Yao Zhensheng, Ding Yuguo. *Climatic Statistics* (in Chinese). Beijing: China Meteorological Press, 1990. 624~629