第42卷 第2期 气 象 学 报 Vol. 42, No. 2
1984 年 5 月 ACTA METEOROLOGICA SINICA May, 1984

矢量场的聚类分析及模糊分类*

陆 龙 骅

（国家气象局气象科学研究院）

提 要

本文给出了矢量场的聚类分析及模糊分类方法，分析了一个实例，并对这两种方法作了初步讨论。

一、引 言

对气象要素场的天气学分类，过去大多是给定分类标准后用主观方法进行的。近年来，随着聚类(Cluster)分析和模糊(Fuzzy)分类方法的发展，使气象要素场的分类有了新的数值分析方法。等压面高度、温度、气压、降水等气象要素场，是标量场，可以直接套用一般的聚类分析[1]或模糊分类[2]方法。有些场，如等压面上的风，是一个二维矢量场。若需要同时考虑多个气象要素，则为多维矢量场。本文即研究矢量场的聚类分析及模糊分类。

二、方 法

为方便起见，我们用等压面风场（二维矢量场）的聚类分析和模糊分类来说明这一问题。

设风场资料为$\{\vec{V}_{l,m}\}$，$1\leqslant l\leqslant N$，$1\leqslant m\leqslant P$，l 和 m 分别为风场样品及资料取值点的序号，N为风场样品个数，P为风取值的格点数。

1．样品间的平均距离及模糊关系

在空间点m上，第 l 个样品的风可以分解成纬向风 $u_{l,m}$ 和经向风 $v_{l,m}$ 两部分，有：

$$\vec{V}_{l,m}=u_{l,m}\vec{i}+v_{l,m}\vec{j} \tag{1}$$

式中 $\vec{i}$ 及 $\vec{j}$ 分别为纬向和经向的单位矢量。第 l 个风场 $\vec{V}_l$、纬向风场 u_l 和经向风场 v_l，分别由 $m:1\to P$的 $\vec{V}_{l,m}$、$u_{l,m}$ 和 $v_{l,m}$ 组成。

第 l 个 u 场或 v 场与第 k 个 u 场或 v 场间的相关系数可表示为：

* 本文于 1981 年 7 月 24 日收到，1983 年 5 月 4 日收到修改稿。

$$\gamma^{u}_{l,k}=\frac{\sum_{m=1}^{P}\left(u_{l,m}-\frac{1}{P}\sum_{m=1}^{P}u_{l,m}\right)\left(u_{k,m}-\frac{1}{P}\sum_{m=1}^{P}u_{k,m}\right)}{\sqrt{\sum_{m=1}^{P}\left(u_{l,m}-\frac{1}{P}\sum_{m=1}^{P}u_{l,m}\right)^{2}\cdot\sum_{m=1}^{P}\left(u_{k,m}-\frac{1}{P}\sum_{m=1}^{P}u_{k,m}\right)^{2}}} \tag{2}$$

$$\gamma^{v}_{l,k}=\frac{\sum_{m=1}^{P}\left(v_{l,m}-\frac{1}{P}\sum_{m=1}^{P}v_{l,m}\right)\left(v_{k,m}-\frac{1}{P}\sum_{m=1}^{P}v_{k,m}\right)}{\sqrt{\sum_{m=1}^{P}\left(v_{l,m}-\frac{1}{P}\sum_{m=1}^{P}v_{l,m}\right)^{2}\cdot\sum_{m=1}^{P}\left(v_{k,m}-\frac{1}{P}\sum_{m=1}^{P}v_{k,m}\right)^{2}}} \tag{3}$$

一般情况下，纬向或经向风场本身之间的相关系数很大，因此在求相关时，u、v 用的是相对于各样本平均值的距平场，即：

$$\Delta u_{l,m}=u_{l,m}-\frac{1}{N}\sum_{l=1}^{N}u_{l,m} \tag{4}$$

$$\Delta v_{l,m}=v_{l,m}-\frac{1}{N}\sum_{l=1}^{N}v_{l,m} \tag{5}$$

定义第 l 及 k 个 u 场之间和 v 场之间的距离为在 P 维空间中两向量间的夹角。有：

$$\theta^{u}_{l,k}=\arccos(\gamma^{u}_{l,k}) \tag{6}$$

$$\theta^{v}_{l,k}=\arccos(\gamma^{v}_{l,k}) \tag{7}$$

若 δu 及 δv 分别为 u 场或 v 场的权重，则第 l 个风场 $\vec{V}_l$ 及第 k 个风场 $\vec{V}_k$ 之间的平均距离可定义为：

$$\theta_{l,k}=\delta u\cdot\theta^{u}_{l,k}+\delta v\cdot\theta^{v}_{l,k} \tag{8}$$

矢量场的平均距离除了可用(8)式来定义外，也可用其它方式定义。例如，我们也可以把第 l 及 k 个矢量场的平均距离定义为：

$$\theta'_{l,k}=\arccos(\gamma^{x}_{l,k}) \tag{9}$$

其中 $\gamma^{x}_{l,k}$ 为 $\vec{V}_l$ 与 $\vec{V}_k$ 的相关系数，

$$\gamma^{x}_{l,k}=\frac{\sum_{m=1}^{2P}(x_{l,m}-\bar{x}_{l,m})(x_{k,m}-\bar{x}_{k,m})}{\sqrt{\sum_{m=1}^{2P}(x_{l,m}-\bar{x}_{l,m})^{2}\cdot\sum_{m=1}^{2P}(x_{k,m}-\bar{x}_{k,m})^{2}}} \tag{10}$$

式中

$$x_{l,m}=u_{l,m},\quad x_{k,m}=u_{k,m},\quad 1\leqslant m\leqslant P \tag{11}$$

$$x_{l,P+m}=v_{l,m},\ x_{k,P+m}=v_{k,m},\ 1\leqslant m\leqslant P \tag{12}$$

$$\bar{x}_{l,m}=\frac{1}{2P}\sum_{m=1}^{2P}x_{l,m},\quad \bar{x}_{k,m}=\frac{1}{2P}\sum_{m=1}^{2P}x_{k,m} \tag{13}$$

当 $\theta_{l,k}$（或 $\theta'_{l,k}$）等于 0 时，$\vec{V}_l$ 及 $\vec{V}_k$ 场之间的平均距离为 0；而当 $\theta_{l,k}$（或 $\theta'_{l,k}$）等于 90° 时，$\vec{V}_l$ 与 $\vec{V}_k$ 正交，也可认为 $\vec{V}_l$ 及 $\vec{V}_k$ 之间不相关。

样品间的模糊关系 $\{\tilde{R}_{l,k}\}$，$1\leqslant l\leqslant N$，$1\leqslant k\leqslant N$，可由平均距离 $\theta_{l,k}$ 变换而得：

$$\tilde{R}_{l,k}=1-\frac{\theta_{l,k}}{180^\circ} \tag{14}$$

此时的 $\tilde{R}_{l,k}$ 为第 l 个样品对第 k 个样品的相似系数，当 $\theta_{l,k}=0$ 时，$\tilde{R}_{l,k}=1$，而当 $\theta_{l,k}=180^\circ$ 时，$\tilde{R}_{l,k}=0$。

样品间的平均距离及模糊关系分别具有反自反性、自反性及对称性，即对 $1\leqslant l\leqslant N$，$1\leqslant k\leqslant N$，有：

$$\text{反自反性：}\quad \theta_{l,l}=0 \tag{15}$$

$$\text{自反性：}\quad \tilde{R}_{l,l}=1 \tag{16}$$

$$\text{对称性：}\quad \theta_{l,k}=\theta_{k,l} \tag{17}$$

$$\tilde{R}_{l,k}=\tilde{R}_{k,l} \tag{18}$$

由此求得的二维矢量场各样品间的平均距离$\{\theta_{l,k}\}$及模糊关系$\{\tilde{R}_{l,k}\}$，可作为进行聚类分析和模糊分类的基础，对多维矢量场也可作类似的处理。

2．聚类过程

1）初始临界距离：

置信度为 α 的初始临界距离 θ_a 定义为：

$$\theta_a=\arccos\gamma_a=\arccos\left(\frac{t_a}{\sqrt{P-2+t_a^2}}\right) \tag{19}$$

其中 γ_a 为置信度为 α 时的临界相关系数；t_a 则是置信度为 α、自由度为 $P-2$ 的 t 分布值。α 为 0.1、0.01、0.001 时的 γ_a 及 θ_a 在图 1 中给出。

2）初始场

由样品之间的最小距离确定初始场 N_0。即为存在着某一 N_1，有：

$$\theta_{N0,N1}=\min_{l\neq k}(\theta_{l,k}),1\leqslant l\leqslant N,1\leqslant k\leqslant N, \tag{20}$$

把第 N_0 个场作为本类入选的第一个场，作为聚类的起点。

3）归并

各样品与本类(这里为第 N_0 个流场)的距离为：

$$\tilde{Q}_{1,k}=\theta_{N0,k},1\leqslant k\leqslant N \tag{21}$$

临界距离为

$$\tilde{Q}_1^R=\theta_a \tag{22}$$

式中 k 为样品场序号，序号 1 的位置为进行聚类运算的总次数。

对于 $k:1\to N$，若排除已选 N_0 后 $\tilde{Q}_{1,k}$ 的最小值 $\tilde{Q}_{1,N1}=\min\limits_{1\leqslant k\leqslant N}(\tilde{Q}'_{1,k})\leqslant\tilde{Q}_1^R$，则可认为 N_1 与 N_0 属于同一类。

此时，各样品与 N_0 和 N_1 组成的新类之间的平均距离和临界距离分别为：

$$\tilde{Q}_{2,k}=\frac{1}{3}(\theta_{N0,N1}+\theta_{N0,k}+\theta_{N1,k}) \tag{23}$$

$$\tilde{Q}_2^R=\frac{1}{3}(\theta_{N0,N1}+2\,\theta_a) \tag{24}$$

同样，若 $\tilde{Q}_{2,N2}=\min\limits_{1\leqslant k\leqslant N}(\tilde{Q}'_{2,k})\leqslant\tilde{Q}_2^R$，$\tilde{Q}'_{2,k}$ 表示在求最小时把已选的各流场(N_0,N_1)排除在外，则认为 N_2 与 N_0、N_1 属于同一类，继续归并。

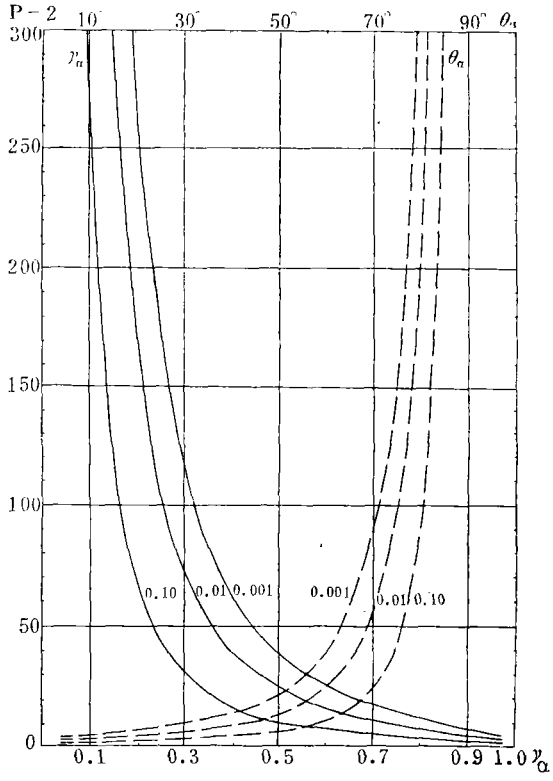


图 1 临界相关系数 γ_a 及初始临界距离 θ_a

在归并了 k_0 个样品后，若逐次归并入本类的各场为 $\{C_i\}$，$C_1=N_0$，$C_2=N_1$，$C_3=N_2\cdots$，$C_{k0}=N_{k0-1}$，k_1 为聚类运算的次数，则各样品与新类之间的平均距离为：

$$\tilde{Q}_{k1,k}=\left(\tilde{Q}_{k1-1,c_{k_0}}\cdot\frac{k_0(k_0-1)}{2}+\sum_{i=1}^{k_0}\theta_{C_i,k}\right)\frac{2}{k_0(k_0+1)} \tag{25}$$

相应的临界距离为：

$$\tilde{Q}^R_{k1}=\left(\tilde{Q}_{k1-1,c_{k_0}}\cdot\frac{k_0(k_0-1)}{2}+k_0\theta_a\right)\frac{2}{k_0(k_0+1)} \tag{26}$$

若排除已选各类样品后 $\tilde{Q}'_{k1,k}$ 的最小值 $\tilde{Q}_{k1,N_{k_0}}$ 仍不大于此时的临界距离，即：

$$\tilde{Q}_{k1,N_{k_0}}=\min_{1\leqslant k\leqslant N}(\tilde{Q}'_{k1,k})\leqslant\tilde{Q}^R_{k1} \tag{27}$$

则将 $N_{k0}=C_{k0+1}$ 也归并入同一类，重复上述过程再选，否则准正交换类。

4) 准正交换类

在进行了 k_1 次运算后，若 $\tilde{Q}'_{k1,k}$ 的最小值大于此时的临界距离，则取 $\tilde{Q}'_{k1,k}$ 中最接近 90°的第M个样品为下一类的初始场，也就是当

$$|90^\circ-\tilde{Q}_{k1,m}|=\min_{1\leqslant k\leqslant N}(|90^\circ-\tilde{Q}'_{k1,k}|) \tag{28}$$

时，则取换类后的 $N_0=C_1=M$，重复上述过程，直到所有N组样品都有所归属。

3. 模糊分类

由(14)式求得的模糊关系$\{R_{l,k}\}$，$1\leqslant l\leqslant N, 1\leqslant k\leqslant N$，是一个满足自反性和对称性的模糊相容关系，需要将其改造为同时具有传递性的模糊等价关系。

传递性定义为：

$$\tilde{R}\circ\tilde{R}\leqslant\tilde{R} \tag{29}$$

$\tilde{R}\circ\tilde{R}$为$\tilde{R}$与$\tilde{R}$的内积，其运算规则与普通的矩阵乘法相似，仅是将“+”与“×”分别改成“∨”(取最大值)与“∧”(取最小值)。若令$\tilde{R}\circ\tilde{R}=\tilde{R}^2$，则有：

$$\tilde{R}^2_{l,k}=\bigvee_{m=1}^{N}(\tilde{R}_{l,m}\wedge\tilde{R}_{m,k}), l,k=1,2,\cdots N, \tag{30}$$

在模糊数学中业已证明，如果$\tilde{R}$为模糊相容关系，则对于适当大的m，$\tilde{R}^m$是一个模糊等价关系。对此一般用平方幂的方法处理，即：

$$\tilde{R}^2=\tilde{R}\circ\tilde{R}, \tilde{R}^4=\tilde{R}^2\circ\tilde{R}^2, \tilde{R}^8=\tilde{R}^4\circ\tilde{R}^4\cdots \tag{31}$$

若在(31)中的某一步有：

$$\tilde{R}^m=\tilde{R}^{2m}=\tilde{R}^* \tag{32}$$

则$\tilde{R}^*$便是$\tilde{R}$的一个模糊等价关系，对其中的元素$\tilde{R}^*_{l,k}$给定某一λ水平后，就可以得到模糊分类结果[2,3]。

4．天气分析

聚类及模糊分类仅是数学处理，在这些工作完成后，应当对求得的各类平均矢量场进行天气学分析，寻找各类矢量场所代表的天气意义。

三、实　　例

我们对1979年5—7月逐候100毫巴35°N—35°S，0—180°E的流场进行了聚类分析和模糊分类。分析时$N=18$，$P=67$，均匀分布在上述区域中，着眼于纬向、经向风场的距平场。各合成风场间的平均距离由(8)式确定，$\delta u=\delta v=0.5$，对纬向、经向风同等看待。聚类分析时置信度α取0.1，模糊分类时λ为0.635。聚类和模糊分类都是在DJS-6上完成的。

1．聚类分析

计算出1979年5—7月逐候100毫巴流场之间的平均距离$\{\theta_{l,k}\}_{1\leqslant l\leqslant 18, 1\leqslant k\leqslant 18}$后，可以发现$\theta_{5,6}=\min\limits_{l\neq k}(\theta_{l,k})$，故可取第5候流场作为聚类的起点，$N_0=5$。

图2给出了$\alpha=0.1$，即初始临界距离为78.3°时的流场聚类结果，图中横坐标为从5月起算的候数，纵坐标为各流场间的平均距离，各类上方的罗马字为聚类序号。

由图2可见，1979年5—7月的100毫巴流场可分成7个类型。有意思的是，属于同一类的是相邻的数候，并没有发生时间上相隔甚远的两候归并为一类的情况，这与各环流型具有相对持续性的天气事实是一致的。属于同一类的各候，东风急流主中心平均位置及主要环流系统的配置情况是相似的。

图3是3—7候的平均流场，为聚类时得到的第Ⅰ类流场，它代表了东风急流建立时的形势。其主要特征是南亚高压中心位于中南半岛北部，在中南半岛南部有一支东风急流，其最大东风中心在南亚高压中心南侧约15个纬距。该类流场可作为夏季风爆发、东

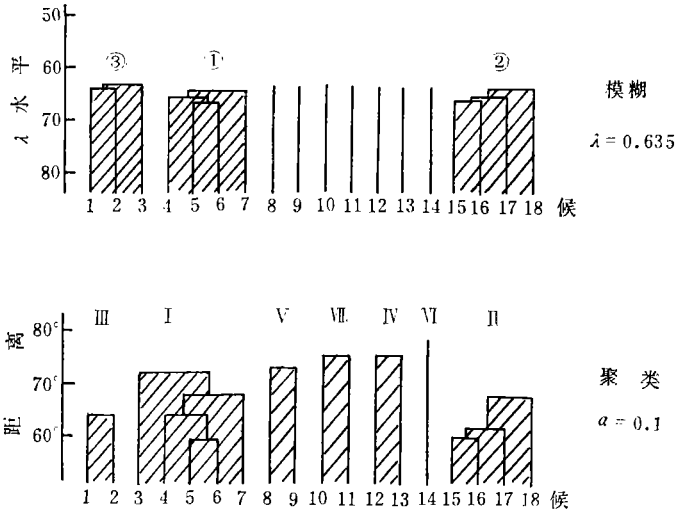


图 2 模糊及聚类划分结果

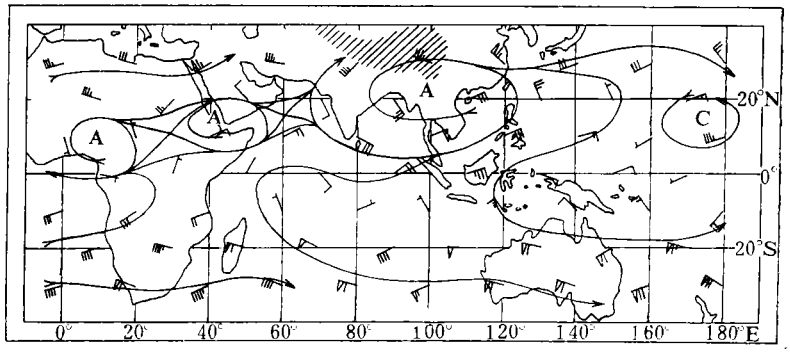


图 3 3—7 候 100 毫巴平均流场

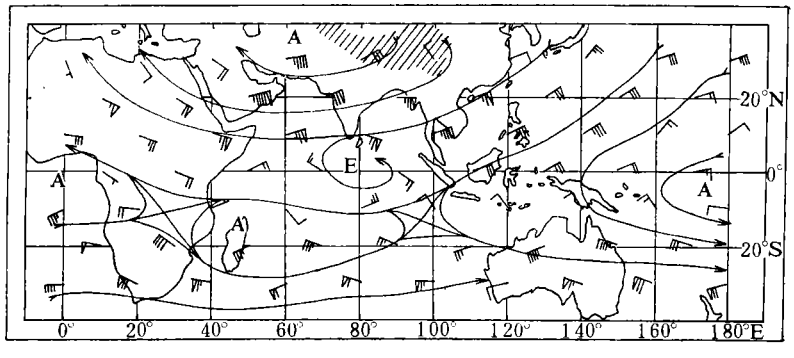


图 4 15—18候 100 毫巴平均流场

亚梅雨期开始前，初夏的典型流场。

图 4 是 15—18 候的平均流场，为聚类时得到的第 II 类流场。此时南亚高压在青藏高原西侧，东风急流轴在 15°N 附近，最大东风中心在阿拉伯南端。本类流场是盛夏流场的典形情况。

图 5 给出了 I、II 类流场和逐候流场的距离，在斜线区内该候流场与某类平均流场的距离小于临界距离。由图 5 可见，除同类内的各候外，I 或 II 类与其它各候的距离都大于相应的临界距离。I 和 II 类流场间的平均距离为 83°，是准正交的，说明在季节转换前后流场是有突变的。

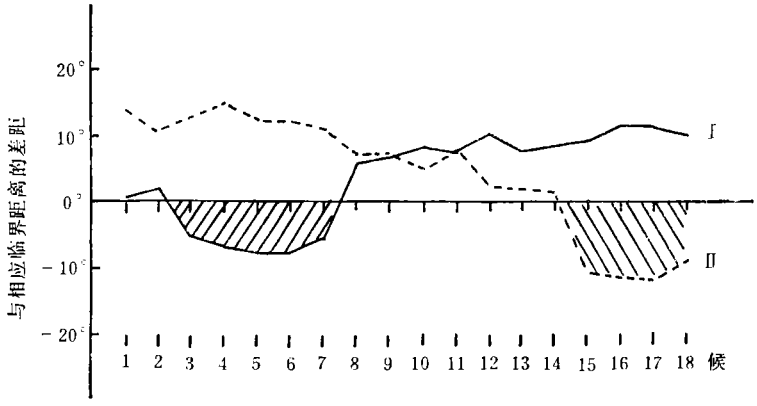


图 5　I、II 类流场与逐候流场的距离

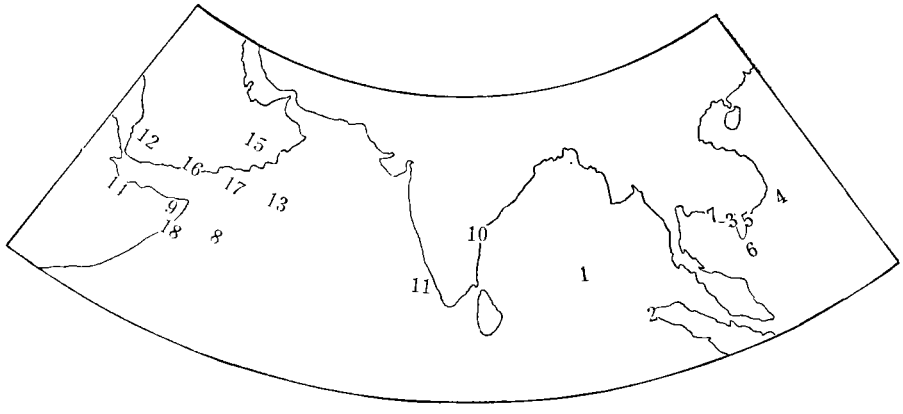


图 6　1979年5—7 月逐候 100 毫巴东风主中心位置
（图中数字为该候东风主中心所在位置）

其它各类为过渡类，除 1—2 候时东风急流尚未建立外，东风急流主中心在中南半岛—阿拉伯之间，而南亚高压则位于中南半岛—伊朗高原。

2．模糊分析

由(14)式求得的模糊关系$\{\tilde{R}_{i,k}\}$经平方幂变换后有：$\tilde{R}^{8}_{i,k}=\tilde{R}^{16}_{i,k}=\tilde{R}^{*}$。此时的$\tilde{R}^{*}$是$\{\tilde{R}_{i,k}\}$的一个模糊等价关系，具有传递性。

当 $\lambda=0.635$ 时，由 $\tilde{R}^*$ 的 λ 截集，可将全部样品分成 10 类(图 2)，其中②类与聚类时的 II 类完全相同，都能作为季节转换后盛夏流场的代表。①和③类也包括了从 5 月第 1 候到 6 月第 1 候的 7 个候，除了模糊分类时第 3 候划入第①类外，与聚类结果完全相同，也都是季节转换前的流场。8—14 候也都属过渡类。

由此可见，流场聚类分析和模糊分类的结果大体相同，都能反映主要环流系统及其活动特征，天气意义是清楚的。

至于 5 月第 3 候，刚好处于 I、III 类流场转折处，其与 I 类流场的平均距离为 68°，而与 III 类流场的平均距离为 70°，都小于相应的临界距离。因此，从数学上来说，将它分在 I 或 III 类都是可以的。但如果考虑到第 3 候离 I 类流场的距离要近一些，而且更重要的是 5 月第 3 候刚好是热带东风急流在中南半岛南端开始建立的一个候，3—7 候在该处都维持着风速达 20 米/秒的东风急流中心，1—2 候时东风急流尚未建立，最大东风风速只有 12—13 米/秒，东风中心位置也与 3—7 候有显著差异(图 6)。从天气意义上来讲，聚类分析时将第 3 候归入第 I 类流场，把 3—7 候放在一起，较之模糊分类时将此划入③类流场为好。

四、讨 论

1．本文提出的矢量场的聚类分析与通常所用的聚类分析方法[1]的差异在于，在聚类时引进了矢量场各样品之间的平均距离、临界距离及准正交换类。聚类时得出的相邻两类是准正交的。例如，在上面所给的例子中，I 与 II 类流场的平均距离为 83°，而 II 与 III 类流场的距离为 94°，都是准正交的。这样能很快地完成聚类过程，当然尚不能保证(也并不需要)所分各类相互之间均为准正交。由于所划各类典型流场是由实际存在的流场平均而得的，故其天气意义清楚。

定义了矢量场之间的平均距离后，矢量场的聚类分析与标量场就没有什么差别了。各矢量场之间的平均距离可以有多种定义方法，能根据需要加以选择。本文(8)式的定义能考虑各分量在分类时的权重，而(9)式则是把各分量同等看待的。

2．在由各矢量场之间的平均距离建立了各矢量场之间的模糊关系后，矢量场的模糊分类与天气过程的模糊分类[2]相同。一般情况下，置信度 α 时的聚类分析与某一 λ 时的模糊分类结果大体相同。

3．在进行模糊划分时，为了使模糊关系具有传递性，要用平方幂的方法进行变换。这在数学上是严格的，但物理和几何意义却不甚清楚，计算也费时。一般情况下，在经平方幂变换后的模糊等价关系中，各样品之间的关系是扩大了的，如此处理后，就有可能把某些类别的范围扩大了。因此，应特别注意所得各类平均场的物理和天气意义是否清楚。

4．气象上的大量问题是模糊的，模糊概念的引入对研究和解决这个问题具有十分重要的意义。相对于取 0、1 两值的确切子集而言，我们同样可以把由各样品间平均距离组成的矩阵 $\{\theta_{i,k}\}$ 看作模糊矩阵。在此矩阵中属于“0”时距离为 180°，而属于“1”时距离为 0°，$\theta_{i,k}$ 同样可以取 0—180°中的任意值。因此聚类分析实质上同样是一种模糊分类，由于

引进了某一置信度时的临界距离，用最小距离归类及准正交换类，所以物理和天气意义很清楚，计算量也不大。就处理分类问题而言，似乎聚类分析要比一般的模糊分类方法好一些。

5．聚类分析方法发展的历史不长，理论上还没有完整的体系。模糊数学诞生于1965年[4]，发展的时间更短，同样也是很不成熟的。这两种方法用于矢量场还仅是个尝试，分析方法也比较粗糙。例如，在由(8)式建立各矢量场的平均距离时，δu 和 δv 的选择完全取决于经验，如何更好地选择某一种权重，能对所考虑的问题获最佳聚类或模糊分类结果；在给定 λ 水平后，模糊划分的结果是唯一的，而在同一置信度 α 时，如果不用样品间的最小距离来决定初始场 N_0，则初始场选择不同，聚类结果也会有一些差异；λ 的统计意义是什么，置信度 α 及 λ 水平的数值如何选取合适等，均有待于进一步研究。但总的讲来，上述两种方法均不失为简便易行、客观定量、行之有效的矢量分析方法，它们所存在的各种问题，尚有待于在实践中克服和改进。

参 考 文 献

[1] 方开泰，聚类分析(I)、(II)，数学的实践与认识，1978年第1期，60—68；1978年第2期，54—62。
[2] 曹鸿兴、陈国范，天气过程的模糊划分，科学通报，**25**，457—460，1980。
[3] 汪培庄，模糊数学简介(I)、(II)，数学的实践与认识，1980年第2期，45—59；1980年第3期52—63。
[4] Zadeh, L. A., Fuzzy Sets, *Information and Control*, **8**, 338—353, 1965.

CLUSTER ANALYSIS AND FUZZY CLASSIFICATION IN THE VECTOR FIELD

Lu Longhua

(Academy of Meteorological Science, State Meteorological Administration)

Abstract

In this paper, cluster analysis and fuzzy classification methods for vector field are presented, with an example and preliminary discussion.