第44卷 第2期　　气 象 学 报　　Vol. 44, No. 2
1986年5月　　ACTA METEOROLCGICA SINICA　　May, 1986

相似离度及其使用技术*

李 开 乐

（广东省气象台）

提　　要

本文介绍一种新的相似比较的数学衡量标准——相似离度以及对它的使用技术。文中剖析了几种常用相似标准的效果之后，还给出了相似离度在预报中的使用技术和初步应用试验的效果。

一、引　　言

假如以平面直角坐标系的 x 轴和 y 轴分别表示因子序号和因子数值，那末，一个气象样本可以在该坐标系中被点绘成一条样本曲线（文中的插图所绘的实际上是折线）。显然，要全面比较两个样本的相似程度，就不但要看两根样本曲线彼此的靠近程度（即总数值的差异程度，简称值相似），还要看它们形状的相似程度（简称形相似）。

目前，常用相似比较的标准主要有相似系数（(1)式）、海明距离（(2)式）、欧氏距离（(3)式）等。经过剖析可知（请见第二节），除了海明距离尚能准确地比较出两样本的值相似程度（不能比较出形相似情况）之外，相似系数和欧氏距离都不理想。

本文引入一种新的衡量相似程度的统计量——相似离度，意思即相似性的差异程度，既考虑到样本之间的形相似情况，又体现了它们的值相似差异，是一个比较全面的相似标准，实际试用效果良好。本文的特点是，把所提出的原理概念紧密结合于实际使用中进行表述，并配有适当图表，使抽象的数学形式容易变得具体化。

二、几种相似标准的效果剖析

1. 相似系数

$$R_{ij}=\cos\theta_{ij}=\frac{\sum_{K=1}^{M}x_{iK}\cdot x_{jK}}{\sqrt{\sum_{K=1}^{M}x_{iK}^2}\cdot\sqrt{\sum_{K=1}^{M}x_{jK}^2}} \tag{1}$$

式中，x 表示因子数值，M 表示每个气象样本取 M 个因子，K 表示因子变量（$K=1,2,$

* 本文于1984年7月9日收到。1984年12月8日收到修改稿。

$\cdots M)$，i，j 表示两个不同的样本。后面各式的符号含义相同。

假设有 5 个简单样本的三个因子数值如表 1，分别点绘出它们的样本曲线如图 1。据(1)式：

$$R_{12}=\frac{1\times2+3\times4+2\times3}{\sqrt{1^2+3^2+2^2}\cdot\sqrt{2^2+4^2+3^2}}=0.99258$$

$$R_{13}=0.951258$$

$$R_{14}=-0.7142857$$

$$R_{15}=0.98198$$

表 1

样本序号	x_1	x_2	x_3
1	1	3	2
2	2	4	3
3	10	12	11
4	−3	−1	−2
5	2	5	5

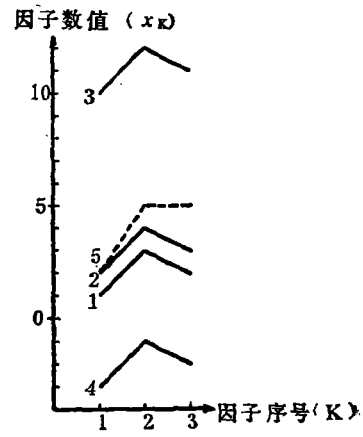


图 1

从图 1 显见，样本曲线 1,2,3,4 的形状是完全相似的，它们与 5 却差异明显，但是，计算结果反而 $R_{12}\nsim R_{14}$，$R_{12}\approx R_{15}$，可见(1)式难以正确判断样本之间的形相似情况。另一方面，样本曲线 1,2 之间的距离(总数值差异)与 1,3 之间的距离显然大不相同，但 $R_{12}\approx R_{13}$，可见(1)式又缺乏正确判断样本值相似程度的能力。

如果把表 1 中因子的原数值换成距平值，检验结果也同样不好。唯独把因子数值换成各样本内的自距平数值，就有良好的形相似比较功能，但仍然不具备值相似的比较能力。有关变值检验计算的详细情况不在此详述。

由上述可见，(1)式这个相似标准是不理想的。

表 2

样本序号	x_1	x_2	x_3
1	1	3	2
2	3	5	4
3	1	−3	2
4	3	1	4

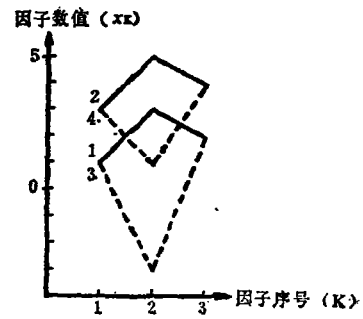


图 2

2. 海明距离

$$d_{ij}=\sum_{K=1}^{M}|x_{iK}-x_{jK}| \tag{2}$$

假设有4个样本的三个因子分配情况如表2和图2，据(2)式：

$$d_{12}=|1-3|+|3-5|+|2-4|=6$$
$$d_{13}=|1-1|+|3-(-3)|+|2-2|=6$$
$$d_{14}=|1-3|+|3-1|+|2-4|=6$$

显然，从样本曲线的形状看，1，2之间完全相似，1，3之间或1，4之间都相差甚远，计算结果却无法体现这种情况。然而，$d_{12}=d_{13}=d_{14}=6$，这对于比较它们之间在总数值的差异上是准确的。由此可见，(2)式虽无法衡量样本之间的形相似情况，却能准确地比较出它们的值相似差异。

3. 欧氏距离

$$d_{ij}=\sqrt{\sum_{K=1}^{M}(x_{iK}-x_{jK})^2} \tag{3}$$

仍以表2的资料为例，据(3)式：

$$d_{12}=\sqrt{(1-3)^2+(3-5)^2+(2-4)^2}=\sqrt{12}=3.4641$$
$$d_{13}=\sqrt{(1-1)^2+[3-(-3)]^2+(2-2)^2}=\sqrt{6^2}=6$$
$$d_{14}=\sqrt{(1-3)^2+(3-1)^2+(2-4)^2}=\sqrt{12}=3.4641$$

由2中的讨论，知道(3)式的计算结果体现不出样本曲线之间的真实距离。看曲线的形状，1，2之间完全相似，1，4之间相差甚远，但$d_{12}=d_{14}$，这就与实际不符。由此可见，(3)式也是个很不理想的相似标准。

三、相 似 离 度

为了较客观地比较样本之间的相似性，我们提出一种新的衡量相似程度的统计

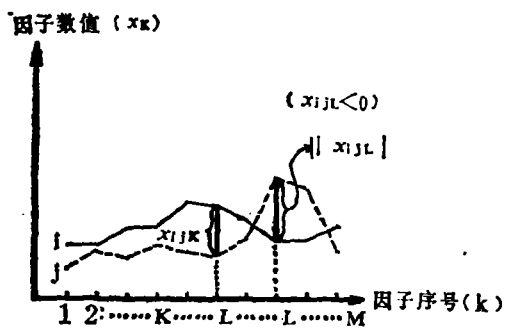


图 3

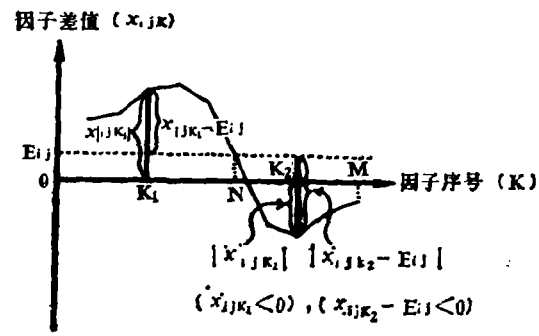


图 4

量——相似离度，符号C_{ij}表示i，j两个样本的相似离度。图3中绘出i，j两条样本曲线。C_{ij}可表示为：

$$C_{ij}=\frac{1}{2}(S_{ij}+D_{ij}) \qquad (4)$$

其中，

$$S_{ij}=\frac{1}{M}\sum_{K=1}^{M}|x_{ijK}-E_{ij}| \qquad (5)$$

$$D_{ij}=\frac{1}{M}\sum_{K=1}^{M}|x_{ijK}| \qquad (6)$$

$$x_{ijK}=x_{iK}-x_{jK} \qquad (7)$$

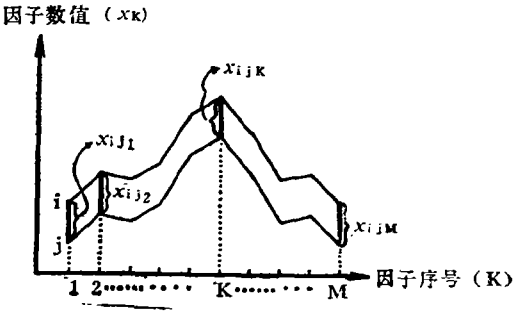


图 5

$$E_{ij}=\frac{1}{M}\sum_{K=1}^{M}x_{ijK} \qquad (8)$$

在此，E_{ij}表示了 i 样本对 j 样本中所有因子之间的总平均差值（见图4）。

D_{ij}实际上就是海明距离对因子容量M求平均值，结合图 3 和图 4 容易看出其具体意义。从二、2 节的讨论知道它能准确地反映出两样本之间在总平均数值上的差异程度，在此称为值系数。

S_{ij}能反映出两个样本中的各个因子之间的差值 x_{ijK}对E_{ij}的离散程度（参见图 4）。显然，只当各个 x_{ijK}都等于E_{ij}，即

$x_{ij1}=x_{ij2}=\cdots=x_{ijM}=E_{ij}$，才可能 $S_{ij}=0$，这时，两根样本曲线的形状将完全相似（如图 5）。相反，若各个 x_{ijK} 对于 E_{ij} 的离散程度越大，则两根样本曲线的形状就越不相似。由此可知，S_{ij} 能较好地反映两样本间的形相似程度，称之为形系数。

综上所述，相似离度由形系数 S_{ij} 和值系数 D_{ij} 两项共同决定。由(5)和(6)式可知，$S_{ij}\geqslant 0, D_{ij}\geqslant 0, \therefore C_{ij}\geqslant 0$。具体地说，当两根样本曲线相重合时，由 (7) 式可知 $x_{ijk}=0$ (K=1,2,$\cdots$,M)，所以可由(8)式得 $E_{ij}=0$，从而由(5)、(6)和(4)式可得 $S_{ij}=0$，$D_{ij}=0$，$C_{ij}=0$。由此可知，当两根样本曲线相重合时，S_{ij}，D_{ij}和C_{ij} 都达到最小值 0，这是最理想的相似情况。对一般的情况，从图 4 可见，在 $K=K_1$ 处，$|x_{ijK_1}-E_{ij}|$比$|x_{ijK_1}|$ 小一个 $|E_{ij}|$值，但在$K=K_2$处，$|x_{ijK_2}-E_{ij}|$却比$|x_{ijK_2}|$大一个$|E_{ij}|$值，所以S_{ij}与D_{ij}的数量级相同，因此，本文的C_{ij}值只由两者取简单平均求得。具体证明请见四、1节。

四、预报操作技术

引入相似离度，目的在于应用。本节着重介绍在天气预报中对相似离度的使用技术。

1．数据标准化

众所周知，由于不同因子的变化域一般都不相同，必须先把因子数据标准化，才能对它们进行比较。

这里的数据标准化方法采用下式进行：

$$x'_K=\frac{x_K-x_{K\min}}{x_{K\max}-x_{K\min}} \quad (K=1,2,\cdots,M) \qquad (9)$$

其中，x_K为样本的原因子数据，x'_K为标准化后的因子数值，$x_{K\max}$，$x_{K\min}$分别为所有历史样

本原因子数据的最大和最小值。

经过标准化后,可使所有样本的因子数据都变为 0—1 之间的数值。假定经过标准化后的 i , j 两个样本的因子数值仍然表示为 x_{iK}和$x_{jK}(K=1,2,\cdots,M)$,则有 $0\leqslant x_{ik}\leqslant 1$, $0\leqslant x_{jK}\leqslant 1$。据(7)式,即有

$$-1\leqslant x_{ijK}\leqslant 1 \tag{10}$$

∴ 由(6)式显见

$$0\leqslant D_{ij}\leqslant 1 \tag{11}$$

可以证明也有 $0\leqslant S_{ij}\leqslant 1$,从而有 $0\leqslant C_{ij}\leqslant 1$。从(8)式和(10)式可得

$$-1\leqslant E_{ij}\leqslant 1 \tag{12}$$

设在M个$(x_{ijK}-E_{ij})$值中(其中$K=1,2,\cdots M$),有N 个为正值或 0,则便有 $M-N$ 个为负值。把因子序号重新排列,使$(K_{ijK}-E_{ij})$为正值和 0 的全部排在前面,负值的排在后面。据(5)式便有

$$S_{ij}=\frac{1}{M}\sum_{K=1}^{M}|x_{ijK}-E_{ij}|$$

$$=\frac{1}{M}\sum_{p=1}^{N}|x_{ijp}-E_{ij}|+\frac{1}{M}\sum_{q=N+1}^{M}|x_{ijq}-E_{ij}| \tag{13}$$

其中,

$$X_{ijp}-E_{ij}\geqslant 0 \quad (p=1,2,\cdots N) \tag{14}$$

$$X_{ijq}-E_{ij}<0 \quad (q=N+1,N+2,\cdots M) \tag{15}$$

由(8)式:

$$E_{ij}=\frac{1}{M}\sum_{K=1}^{m}X_{ijK},$$

∴

$$\frac{1}{M}\sum_{p=1}^{N}(X_{ijp}-E_{ij})+\frac{1}{M}\sum_{q=N+1}^{M}(X_{ijq}-E_{ij})$$

$$=\frac{1}{M}\sum_{K=1}^{M}(X_{ijK}-E_{ij})$$

$$=\frac{1}{M}\sum_{K=1}^{M}X_{ijK}-E_{ij}$$

$$=0$$

$$\frac{1}{M}\sum_{p=1}^{N}(X_{ijp}-E_{ij})=-\frac{1}{M}\sum_{q=N+1}^{M}(X_{ijq}-E_{ij})$$

再从(14)和(15)式便可知:

$$\frac{1}{M}\sum_{p=1}^{N}|X_{ijp}-E_{ij}|=\frac{1}{M}\sum_{q=N+1}^{M}|X_{ijq}-E_{ij}|$$

(13)式又可写为:

$$S_{ij}=\frac{2}{M}\sum_{p=1}^{N}|X_{ijp}-E_{ij}| \tag{16}$$

或
$$S_{ij}=\frac{2}{M}\sum_{q=N+1}^{M}|X_{ijq}-E_{ij}| \tag{17}$$

其直观的意义可从图 4 容易看到。

由(12)式：$-1\leqslant E_{ij}\leqslant 1$，分开两种情况进行讨论。

1) 在 $0\leqslant E_{ij}\leqslant 1$ 的情况下。因为有(14)式：$X_{ijp}-E_{ij}\geqslant 0$，所以 $X_{ijp}\geqslant E_{ij}\geqslant 0$。又有(10)式：$-1\leqslant X_{ijK}\leqslant 1$，所以

$$0\leqslant E_{ij}\leqslant X_{ijp}\leqslant 1 \tag{18}$$

再从(10)式和(15)式：$X_{ijq}-E_{ij}<0$，又可知：

$$-1\leqslant X_{ijq}<E_{ij}\geqslant 0 \tag{19}$$

从(18)式可得

$$|X_{ijp}-E_{ij}|=X_{ijp}-E_{ij}\leqslant 1-E_{ij}\leqslant 1 \tag{20}$$

当 $2N\leqslant M$ 时，由(16)式可得

$$S_{ij}=\frac{2}{M}\sum_{p=1}^{N}|X_{ijp}-E_{ij}|\leqslant\frac{2}{M}\times N\times 1\leqslant 1。$$

又从(19)式可得

$$|X_{ijq}-E_{ij}|\leqslant|X_{ijq}|+E_{ij}\leqslant 1+E_{ij} \tag{21}$$

当 $2N>M$ 时，由(13)、(20)、(21)式又可得

$$S_{ij}\leqslant\frac{N}{M}(1-E_{ij})+\frac{M-N}{M}(1+E_{ij})$$

$$=1+\frac{M-2N}{M}E_{ij}<1。$$

可知在 $0\leqslant E_{ij}\leqslant 1$ 的情况下，$S_{ij}\leqslant 1$。

2) 在 $-1\leqslant E_{ij}<0$ 的情况下。类似的，可有 $0>E_{ij}\leqslant X_{ijp}\leqslant 1$ 和 $-1\leqslant X_{ijq}<E_{ij}<<0$，

$\therefore$
$$|X_{ijp}-E_{ij}|\leqslant 1+|E_{ij}| \tag{22}$$

$$|X_{ijq}-E_{ij}|\leqslant 1-|E_{ij}|<1 \tag{23}$$

当 $2N\leqslant M$ 时，由(13)、(22)、(23)式可有

$$S_{ij}\leqslant\frac{N}{M}(1+|E_{ij}|)+\frac{M-N}{M}(1-|E_{ij}|)$$

$$=1+\frac{2N-M}{M}|E_{ij}|\leqslant 1。$$

当 $2N>M$ 时，由(17)、(23)式也有

$$S_{ij}=\frac{2}{M}\sum_{q=N+1}^{M}|X_{ijq}-E_{ij}|$$

$$<\frac{2}{M}\times(M-N)\times 1=1+\frac{M-2N}{M}<1。$$

可知在 $-1\leqslant E_{ij}<0$ 的情况下，也有 $S_{ij}\leqslant 1$。据(5)式，当 $X_{ij1}=X_{ij2}=\cdots=X_{ijM}=E_{ij}$ 时，可得 $S_{ij}=0$，即 S_{ij} 的最小值。

$$\therefore \quad 0 \leqslant S_{ij} \leqslant 1 \tag{24}$$

(11)和(24)式就是(4)式中 C_{ij} 只由 S_{ij} 和 D_{ij} 取简单平均的具体依据，而且可知也有

$$0 \leqslant C_{ij} \leqslant 1 \tag{25}$$

2．计算 C_{ij}

假定有 n 个历史样本，对于预报 0 天的样本（即现时样本）令 $i=0$，使用标准化资料顺序计算 $C_{0j}(j=1,2,\cdots,n)$，步骤可参见表 3。

表 3

因子差值 $x_{0jk}=x'_{0k}-x'_{jk}$ / 样本序号 j	x_{0j1}	x_{0j2}	$\cdots x_{0jk}\cdots$	x_{0jM}	E_{0j}	S_{0j}	D_{0j}	C_{0j}	b_j
1									
2									
⋮									
n									
									(Q)

3．预报要素的迭加

完成 C_{0j} 的计算之后，通过比较便可找到若干个最佳的相似个例。可以逐个地直接分析这些个例去作预报，也可以把这些个例的预报要素进行综合平均后再供预报参考。下面简单介绍一种权重迭加平均法供参考。

定义 b_j 为权重系数：

$$b_j=1-C_{0j}/Z \tag{26}$$

虽然从理论上可知相似离度的变化范围为 $0 \leqslant C_{ij} \leqslant 1$，但实际上，常常当 $C_{ij}=Z<1$ 时，两样本之间就已经变成了极端的不相似情况。为此，可根据经验按照一般槽脊的分布情况和某些因子数值的波动变化情况，主观地先确定一个数值 Z：

$$0<Z<1 \tag{27}$$

当 $C_{ij}=Z$ 时，就认为这两个样本已经最不相似。这个 Z 就是在某个具体的预报工具中，相似离度的允许限度的最大极限，称为相似极限。显然，当 $C_{0j}=0$ 时，0，j 两样本曲线相重合，即两样本最相似，这时，按(26)式可得 $b_j=1$，是权重最大值；当 $C_{0j}=Z$ 时，按相似极限的定义已被认为两样本的相似性最差，按 (26) 式可得 $b_j=0$，即不值得考虑与之迭加。

从理论上看，因为 $0 \leqslant C_{ij} \leqslant 1$[(25)式]，又 $0<Z<1$[(27)式]，所以可能存在 $C_{0j}>Z$ 的情况，据(26)式，也就可能有 $b_j<0$ 的情况，这将使迭加计算复杂化。

实际上，寻找相似个例的过程，是从众多的历史样本中只选出少数的几个最佳的相似个例来，大多数的历史样本是被淘汰的。为使所选上的历史样本与 0 天样本之间尚存的差异性在各个因子上的分布比较均匀，即防止"畸形"样本的被选，在计算前拟先给定某些 L 值(可随因子的不同而不同)，规定：

$$0<L<Z \tag{28}$$

在 0，j 两样本的比较中，只要有某一个因子的差值大于 L，即 $|X_{0,K}|>L$（K 为 1,2,$\cdots$ M 中的任一整数），便可不必计算 C_{0j} 而立即把这个 j 样本剔除，这样还可大大减少计算量。由此，在所选上的有限几个相似样本中，必有 $|X_{0,K}|\leqslant L(K=1,2,\cdots M)$。

∴ 由四.1 节的证明过程便可知也必有

$$C_{0j}\leqslant L<Z \tag{29}$$

$$\therefore \quad 0<b_j\leqslant 1 \tag{30}$$

也就是说，在实际的计算中，$b_j\leqslant 0$ 的情况并不出现。

若考虑用 W 个相似样本迭加作 Y 要素未来 T 天的预报，迭加平均公式为：

$$Y_t=\frac{1}{\sum_{j=1}^{W}b_j}\cdot\sum_{j=1}^{W}b_jY_{jt} \tag{31}$$

$$(t=1,2,\cdots,T)$$

这里的 Y_{jt} 值取用原始数据，不再进行标准化。

4．相似度

因为有(30)式($0<b_j\leqslant 1$)，且当 $b_j=1$ 时最相似，所以，b_j 实际上也反映了两样本之间的相似程度。考虑

$$Q=\frac{1}{W}\sum_{j=1}^{W}b_j \tag{32}$$

Q 即表示了 W 个相似样本迭加平均后的综合样本与现时样本的相似程度，称相似度。当我们在使用这 W 个相似样本的 Y 要素迭加平均后的预报序列时，究竟它的参考价值有多大，就可以结合参考这个 Q 值。

五、初步应用试验的情况

初步的应用例子是，用 1961—1982 年的 500 hPa 逐日 08 时天气图，选取北支西风急流附近纬圈(55°N)和南支急流附近纬圈(30°N 和 40°N)的亚欧主槽脊的位置(经度)、高度、温度以及副热带高压的脊端经纬度等 14 项资料作为相似因子，经过数据标准化后按上述相似原理设计程序使用计算机进行计算。计算结果在 1983 年暴雨中期预报中有一定的参考价值，具体情况如下。

1．一次预报实例

以 1983 年 5 月 15 日(08 时)为预报 0 天，广东省雨型代码(0—5 共 6 个码分别代表无雨、小雨、中雨、大雨、暴雨、大暴雨)预报的计算结果如表 4。从表 4 中可见，不论 3 个主要的相似个例迭加，还是 4,5 或 6 个个例迭加的预报序列，都在 22—26 日前后出现雨型最大值。结合其他预报工具综合考虑，本气象台于 5 月 17 日对外发布了“22 日前后(开始)有一场大雨过程”的中期预报，实况是 23,24 日两天下了大雨(代码 3)，25 日下了暴雨(代码 4)。

表 4

雨型代码（实况和预报） 个例和选加 \ 预报时效 日期	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
第 1 个相似个例实况 1977.5.20	2	1	1	1	2	3	4	5	5	4	5	1	2	2	4
第 2 个相似个例实况 1977.5.19	3	2	1	1	1	2	3	4	5	5	4	5	1	2	2
第 3 个相似个例实况 1979.5.11	3	4	4	3	2	2	1	1	1	1	3	2	2	3	1
前 3 个个例选加预报	2.7	2.3	2.0	1.7	1.7	2.3	2.7	3.3	3.7	3.3	4.0	2.7	1.7	2.3	2.3
第 4 个相似个例实况 1977.5.28	5	4	5	1	2	2	4	2	2	3	2	1	2	1	1
前 4 个个例选加预报	3.2	2.7	2.7	1.5	1.7	2.3	3.0	3.0	3.3	3.3	3.5	2.3	1.7	2.0	2.0
第 5 个相似个例实况 1980.5.7	2	1	1	3	2	1	3	4	3	3	1	1	1	0	3
前 5 个个例选加预报	3.0	2.4	2.4	1.8	1.8	2.0	3.0	3.2	3.2	3.2	3.0	2.0	1.6	1.6	2.2
第 6 个相似个例实况 1974.5.8	2	2	1	1	0	2	2	1	0	1	1	0	1	1	1
6 个个例选加预报	2.8	2.3	2.2	1.7	1.5	2.0	2.8	2.8	2.7	2.9	2.7	1.7	1.5	1.5	2.0
未来实况	0	0	1	2	3	2	2	3	3	4	2	2	3	2	1

当然，在使用计算结果时，还可根据兴趣进一步逐个地分析研究各个相似个例的具体天气形势的异同点去考虑预报结论。

2．一般效果统计

在 1983 年 4—6 月间共上机计算 16 次，对各次计算的未来第 3—15 天的预报结果进行评分统计。评分的标准是：按 0—5 共 6 级雨型，预报和实况同级者得 100 分，误差 1 级者得 50 分，误差≥2 级者得 0 分。另外，实况在预报的±1 天内出现者才能按标准评分，在±1 天内不出现者只能得 0 分。评分统计的结果如表 5。

表 5

准确率% \ 预报时效天数 选加个例数	3	4	5	6	7	8	9	10	11	12	13	14	15	平均
3	75.0	78.1	71.9	78.1	78.1	81.3	75.0	81.3	75.0	78.1	75.0	78.1	71.9	76.7
4	71.9	84.4	78.1	75.0	78.1	84.4	81.3	81.3	78.1	78.1	81.3	81.3	71.9	78.9
5	68.8	78.1	81.3	75.0	68.8	81.3	90.6	84.4	78.1	81.3	75.0	84.4	75.0	78.6
6	78.1	84.4	84.4	75.0	71.9	84.4	90.6	87.5	81.3	75.0	78.1	87.5	75.0	81.0

参 考 文 献

[1] 张尧庭、方开泰，《多元统计分析引论》第八章，科学出版社，1983 年。

[2] 黄嘉佑，《气象统计预报试用教材》第五章，北京大学地球物理系气象专业。

[3] 张尧庭等编，《气象资料的统计分析方法》第十四、十五讲，农业出版社，1979 年。

[4] 王得民等编，《气象统计预报》第四章，南京气象学院，1980 年。

[5] 《模糊数学在气象中的应用会议文集》第四章，无锡，1981 年。

[6] 谭冠日，《气象站数理统计预报方法》第六章，科学出版社，1978 年。

[7] 曹鸿兴“天气预报中的客观分类问题”，中央气象局气科所编《数理统计天气预报文集》，农业出版社，1977 年。

A NEW SIMILARITY PARAMETER AND ITS APPLICATION

Li Kaile

(Guangdong Meteorological Observatory)

Abstract

A new similarity parameter and its application are introduced in this paper. The author has compared several similarity parameters and has introduced a new parameter. We found that it is useful in forecasting practice.