# 常用相似性判据的检验和综合相似系数的使用

阎惠芳　李社宗　黄跃青　张　霞

(河南省郑州市气象局 ,郑州 450005)

摘要　文章对相似预报方法常用的几个相似性判据作了客观检验和讨论 ,同时提出了综合相似系数 ,2000 年 6 ~ 8 月黄淮主汛期郑州市气象台降水预报业务现场试验表明 ,综合相似系数在降水相似预报和 24h 暴雨预报中有较好的表现 。

关键词　相似预报方法　综合相似系数　暴雨预报

## 引言

相似方法是目前在短 、中 、长期天气预报业务中被广泛使用的方法 ,它是基于对当前大气的物理状况和发展过程与历史上相应时段内曾经出现过的大气的物理状况和发展过程的比较 ,从中找出涵盖静态和动态两方面均相似的个例 ,来推断未来天气的发展过程 ,按照历史上出现的一个或几个最相似的个例 ,预测今后即将出现的天气现象 。相似预报方法 ,尽管是一种老的天气学方法 ,但是它能够考虑天气发展过程的三维结构特征 ,且兼顾大气本身的线性和非线性变化规律 ,所以在数字化迅速发展的今天 ,仍有极其旺盛的生命力 。

相似方法中相似性判据对相似程度的衡量 ,乃至相似预报的质量 ,都起着至关重要的作用 。本文在对常用的几个相似性判据作客观检验和讨论的基础上 ,提出了综合相似系数 ,业务现场试验表明 ,该系数在降水相似预报中有较好的应用价值 。

## 1　几个常用的相似性判据

客观度量两个样品相似程度的判据主要有以下 4 类[1]:

(1) 相关距离系数( 以下简称相关系数)

$$R_{ij} = \frac{\sum_{k=1}^{m}(x_{ik} - \overline{x_i})(x_{jk} - \overline{x_j})}{\sqrt{\sum_{k=1}^{m}(x_{ik} - \overline{x_i})^2 \sum_{k=1}^{m}(x_{jk} - \overline{x_j})^2}} \qquad (1)$$

式中 $x$ 表示因子数值 , $m$ 表示某气象样品所取空间点数 , $k$ 表示空间点之序号( $k = 1 ,2 , . . . , m$) ,样品序号 $i , j$ 表示两个不同的样品(或两个不同时刻的样品) 。此处 $R_{ij}$ 是两个样品场之间的相关系数 ,用以衡量其间的相似程度 。

(2) 相似系数

$$S_{ij} = \frac{\sum_{k=1}^{m} x_{ik}x_{jk}}{\sqrt{\sum_{k=1}^{m} x_{ik}^2 \sum_{k=1}^{m} x_{jk}^2}} \qquad (2)$$

$S_{ij}$ 同样是衡量两个样品场 $x_{ik}$ 和 $x_{jk}$ 之间在时间或空间上的相似程度 ,和相关系数不同的是 ,相似系数用的不是因子的距平值 ,而是因子本身的数值 。

(3) 距离系数

距离系数有多种 ,但最常用的有两种 ,即绝对距离(或称海明距离)

$$A_{d_{ij}} = \sum_{k=1}^{m} | x_{ik} - x_{jk} | \qquad (3)$$

和欧氏距离

$$E_{d_{ij}} = \sqrt{\sum_{k=1}^{m}(x_{ik} - x_{jk})^2} \qquad (4)$$

这两种距离系数的意义很直观 ,一目了然 。

(4) 相似离度

文献[2] 提出了一种新的相似性判据 —— 相似离度

郑州市科技局“郑州市短期预报平台”项目资助
收稿日期 :2002 年 11 月 22 日 ;定稿日期 :2003 年 2 月 15 日
作者简介 :阎惠芳 ,女 ,1965 年生 ,高级工程师 ,从事天气预报研究

$$c_{ij} = \frac{1}{2}(s_{ij} + d_{ij}) \tag{5}$$

其中形系数

$$s_{ij} = \frac{1}{m}\sum_{k=1}^{m}|x_{ijk} - e_{ij}| \tag{6}$$

值系数 $d_{ij}$ 实际就是绝对距离的平均值

$$d_{ij} = \frac{A_{d_{ij}}}{m} = \frac{1}{m}\sum_{k=1}^{m}|x_{ijk}| \tag{7}$$

其中, $x_{ijk} = x_{ik} - x_{jk}$

因子之间差值的总平均

$$e_{ij} = \frac{1}{m}\sum_{k=1}^{m}x_{ijk} \tag{8}$$

相似离度可以理解为距离系数的某种延伸,它考虑了形相似和值相似,看来是较全面的。不过,我们不得不遗憾地指出,实际应用相似离度未能得到理想的效果。

以上相似性判据从不同角度描述样品的相似程度,因此它们的描述能力和确切性受到人们的关注。文献[2]特别设计了两个例子共 8 个样品,分别计算了相似系数及绝对距离与欧氏距离后得出结论:相似系数难以正确判断样本之间的形相似的程度,欧氏距离体现不出曲线之间的真实距离,只有绝对距离能准确反映样品之间的值相似差异。本文也从文献[2]中设计的 8 个典型样品入手,逐个检验上面提到的 4 类 5 种相似性判据,通过剖析,再引入一个描述相似性的新判据——综合相似系数。

## 2 相似性判据的初步检验

表 1 是引用文献[2]中构造的比较有特点的 8 个简单样品(相应的曲线略)。样品 2 和样品 6 与样品 1 形完全相似,值也比较接近;样品 5 与样品 1 形不太相似,但值介于样品 2 和 6 之间;样品 3 和 4,前者值偏高,后者值偏低;样品 7 和 8 与样品 1 完全反相。利用式(1)~(8),计算结果列于表 2 中。

表 1 用于相似性判据初步检验的 8 个典型样品

| 样品序号 | $X_1$ | $X_2$ | $X_3$ |
|---|---|---|---|
| 1 | 1 | 3 | 2 |
| 2 | 2 | 4 | 3 |
| 3 | 10 | 12 | 11 |
| 4 | -3 | -1 | -2 |
| 5 | 2 | 5 | 5 |
| 6 | 3 | 5 | 4 |
| 7 | 1 | -3 | 2 |
| 8 | 3 | 1 | 4 |

表 2 5 个相似性判据初步检验结果比较

| 相比较的两样品序号 | 相关系数($R$) | 相似系数($S$) | 欧氏距离($E_d$) | 绝对距离($A_d$) | 相似离度 | | |
|---|---|---|---|---|---|---|---|
| | | | | | $e$ | $s$ | $c$ |
| 1 和 2 | 1.0 | 0.9926 | 1.7321 | 3.0 | -1.00 | 0.00 | 0.50 |
| 1 和 3 | 1.0 | 0.9513 | 15.5885 | 27.0 | -9.00 | 0.00 | 4.50 |
| 1 和 4 | 1.0 | -0.7143 | 6.9282 | 12.0 | 4.00 | 0.00 | 2.00 |
| 1 和 5 | 0.8660 | 0.9820 | 3.7417 | 6.0 | -2.00 | 0.67 | 1.34 |
| 1 和 6 | 1.0 | 0.9827 | 3.4641 | 6.0 | -2.00 | 0.00 | 1.00 |
| 1 和 7 | -0.7559 | -0.2857 | 6.0 | 6.0 | 2.00 | 2.89 | 2.45 |
| 1 和 8 | -0.6547 | 0.7338 | 3.4641 | 6.0 | -0.67 | 1.78 | 1.89 |

由表 1 和表 2 可得出如下几点看法:

(1)相似离度兼顾形相似情况和值相似的差异,思路是清晰的,由表 2 中该判据从小到大的排序来看,结果基本合理。但因其受绝对距离的影响颇大,以致对样品 4 这样全是负值的因子特点未能充分反映出来。再者,完全反相的样品 7 和 8 与样品 1 的相似度排在样品 1 和 3 之前,不能认为是合理的。

(2)相似系数并非"形"相似系数,由式(2)知,相似系数用的是因子原值,所以相似系数大小,除了反映与场的线性相关外,还与场值本身的正、负和高、低的差异有关。例如,样品 1 与 4 相似系数很低,正是因为样品 4 的因子全是负值,和样品 1 的正值数列形成了明显差异。所以仅用"形"是否相似为标准来评价相似系数是不妥当的。上述例子相似系数的判别结果,还是基本符合天气经验的,不过样品 3 与样品 1 在数值上的差异较大,这方面反映有些不足。

(3)相关系数和相似系数不同之处在于使用了距平值,因此能反映出两个样品对各自因子均值的离散程度,可以认为它是一个十分理想的形相似系数。从位相上看,其正值是同相的,负值是反相的。样品 1 和 2、1 和 3、1 和 4、1 和 6 的相关系数达到 1.0,说明是完全相似的,而样品 1 和样品 7 与 8 的位相则是相反的。天气学预报方法中,形相似是十分重要的,预报员关注的槽脊位置及其深度或强度的相似,这对短期 24h 以内的预报尤为关键。

(4)距离系数对样品在因子数值上的差异较敏感。虽然绝对距离比欧氏距离能更好地反映总平均数值上的差异程度,但欧氏距离在相似判据中也有其实用性的一面。例如样品 1 和样品 5、6、7、8 之间的绝对距离都是 6.0,并不说明样品 1 和这 4 个样品之间的相似性是一致的,相反,样品 1 和 7、1 和 8,尤其是 1 和 7 之间,

存在着十分明显的差异,欧氏距离尽管不能区分样品 1 和 8 与样品 1 和 6 之间的差异,但可以显示出 1 和 7、1 和 5 与 1 和 6 之间的差异。

以上讨论阐明诸相似性判据各有特点,也各有缺点,为了在业务中方便使用,需要适当综合应用,这就是我们建立综合相似性判据的原因。

## 3　综合相似系数

相似离度虽然考虑了形相似和值相似,但是从上节讨论看出,其对样品处理存在明显缺陷。鉴于相似系数能把样品 1 和 6 与 1 和 8 以及 1 和 7 之差别展示出来,也能把样品 1 和 5 跟样品 1 和 6 的微小差别区分开来,尤其是能突出揭示如 1 和 4 这类因子符号截然相反的样品之差别,这个特点比较适合于天气学方法应用。显然人们会拒绝从负涡度区中寻找正涡度的"相似"样品,也不会为上升运动场找一个"形"相似的下沉运动场,因此,相似系数可以作为综合相似性判据的一部分。为提高相似系数对距离的识别能力,有必要综合距离系数,这里选用了欧氏距离。

距离系数有一个共同的缺限,即当样品各个因子不是同一类变量时,距离数值之间失去比较性,也就是说,各个因子对距离影响与它自身量纲有关。为此,通常因子需要采用标准化处理。考虑到在构造综合相似系数时,需要变换距离系数使之能与相似系数进行综合叠加,这里我们使用了相对欧氏距离、样本最大欧氏距离以及距离相似系数的概念。所谓相对欧氏距离是样品欧氏距离与样本最大欧氏距离之比:

$$R_{d_{ij}} = \frac{E_{d_{ij}}}{E_{d_{\max}}} \tag{9}$$

式中 $R_d$ 为相对欧氏距离,$E_{d_{\max}}$ 为样本最大欧氏距离

$$E_{d_{\max}} = \sqrt{\sum_{k=1}^{m}(x_{k\max} - x_{k\min})} \quad (k = 1, 2, \ldots, m) \tag{10}$$

式(10)中 $X_{k\max}$和 $X_{k\min}$分别为空间点 $k$ 上因子历史最大值和最小值。$R_{d_{ij}}$ 的极端最小值为 $R_{d_{ij}} = 0$,而当 $R_{d_{ij}} = 1$ 时,欧氏距离就达到了历史上最不相似的情况。这就是说,$R_{d_{ij}} \leqslant 1$,从而可定义距离相似系数为

$$E_{ij} = 1 - R_{d_{ij}} \tag{11}$$

且有 $0 \leqslant E_{ij} \leqslant 1$。于是定义综合相似系数的基础部分为相似系数和距离相似系数两者的平均值,它实际上是一个新的函数

$$C_{ij} = \frac{1}{2}(S_{ij} + E_{ij}) \tag{12}$$

在业务应用中,为突出形相似的作用,综合相似系数还可和相关系数的相似结果相叠加,即从各自挑选出的最相似的样本对应的未来天气现象中,以权重法综合两者的预报结果,作为综合相似系数的最终结果。

利用综合相似系数,针对样品 1 与其余 7 个典型样品间相似程度排序(表略)可知,样品 1 和 3 的序由相似系数中的第 4 位改变为综合相似系数的第 5 位,相似系数为负值的样品 1 和 7 以及 1 和 4 保持排序的最后两位,说明综合相似系数的结果是比较有特色的。

## 4　相似性判据在业务预报应用中的对比试验

为阐明上文所提出的各相似性判据的效果和综合相似系数的使用,就 2000 年 6 ~ 8 月黄淮主汛期郑州市气象台的业务预报现场对比试验结果作简要讨论。

相似预报效果除了受相似性判据影响之外,还与所选预报因子、因子精度、相似预报方法以及相似样本容量等有关。受资料收集的限制,制作 2000 年预报时,相似样本资料为 1994 ~ 1999 年,共 6 年。资料虽然太短,但对不同的相似性判据,在同样时段,用同样的因子和方法进行比较试验,还是可以得到有意义的结果的。

郑州市气象台业务中使用的相似预报方法是建立在大气演变动态相似基础上,称大气物理结构 4 级动态相似预报方法,着重揭示东亚大气动力过程发生、发展的特征,分析了与大气过程有关的 30 个物理量,4 个连续时次的动力过程。本文针对各相似性判据的检验,仅用了两级相似,即以前一日20:00为起始预报场,从 6 年汛期(6 ~ 8 月)552 天中客观选取 30 个最相似的样品作为第 1 级,在此基础上利用当日 08:00 的大气物理状态特征,从 1 级的 30 个相似样本中进一步筛选出 20 个最相似的样品作为第 2 级(所选的样品数目系四级动态方法需要,限于篇幅,这里不予讨论),在本次检验中,使用了第 2 级的前 3 ~ 6 个样品,并以此作为相似预报的依据。

预报检验所用的实况降水量,转化为郑州地区的雨型,雨型分 0 ~ 4,共 5 型,0 为无降水,1 为小雨

(0.1 ~ 9.9 mm) ,2 为中雨(10.0 ~ 24.9 mm) ,3 为大雨(25.0 ~ 49.9 mm) ,4 为暴雨( ≥50.0 mm) ,其划分原则为:郑州本站及市所辖 6 个县、市,共 7 站,有 3 站降水达到某量级以上,就定为该量级雨型。相似检验中用了 3 级雨型,即无雨为 0 级,小雨和中雨为 1 级,大雨和暴雨为 3 级。

预报检验所用的评分方法是参照国际上通用的 $T_s$ 评分[3]制定的,降水 $T_s$ 评分方法如下:

以 $A$ 表示预报无降水报对次数,以 $B$ 表示指定等级的降水漏报次数,以 $C$ 表示预报指定等级的降水空报次数,以 $D$ 表示预报指定等级的降水报对次数,

则

$$T_s = D/(D + C + B) \tag{13}$$

$$EH = (D + A)/(D + A + C + B) \tag{14}$$

$$FAR = C/(C + D) \tag{15}$$

$$DFR = B/(B + D) \tag{16}$$

式(13) ~ (15) 中 $T_s$ 即降水的风险评分(Threat score) ,评分时剔除了无降水事件,所以它适用于小概率事件的评分;$EH$ 为预报效率,它是通常意义上包含无降水在内的评分;是总体评分。$FAR$ 是降水空报率;$DFR$ 是降水漏报率。漏报率的计算式通常应是 $B/(B + A)$ ,但是对于诸如暴雨等小概率事件,$A$ 一般很大,$A$ 远大于 $B$,这时漏报率是个相当小的值,不能清楚地说明问题,因此改用 $B/(B + D)$ 替代。其中 $D$ 是报对的次数,$B$ 是漏报的次数,$B + D$ 是实际观测出现的预报事件总数,所以式(16) 含义是很清楚的。

相似样品的选取有两种方式:依照相似程度从高到低的顺序选最佳相似样品,24h 降水预报和 24 ~ 48h 的暴雨预报选前 3 个最佳样品,48 ~ 72h 降水预报选取前 6 个最佳样品。

其中某日相似样品降水者达半数或以上时,就认为该日有降水,否则无降水。降水等级由实际降水样品的平均值求出。表 3 是 2000 年 6 ~ 8 月各相似性判据 24h 相似预报的 $T_s$ 评分结果。

表 3 2000 年 6 ~ 8 月各相似性判据的 24h 相似预报评分

| 相似性判据 | 有无降水预报 | | | | 分级降水预报 | | | | 暴雨预报 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $T_s$ | $EH$ | $FAR$ | $DFR$ | $T_s$ | $EH$ | $FAR$ | $DFR$ | $T_s$ | $EH$ | $FAR$ | $DFR$ |
| 综合相似系数 | 0.512 | 0.772 | 0.214 | 0.405 | 0.465 | 0.750 | 0.259 | 0.246 | 0.333 | 0.935 | 0.400 | 0.444 |
| 相关系数 | 0.488 | 0.761 | 0.222 | 0.432 | 0.442 | 0.739 | 0.240 | 0.486 | 0.250 | 0.935 | 0.333 | 0.714 |
| 相似系数 | 0.400 | 0.674 | 0.394 | 0.459 | 0.380 | 0.663 | 0.406 | 0.486 | 0.333 | 0.935 | 0.400 | 0.571 |
| 欧氏距离 | 0.250 | 0.609 | 0.478 | 0.676 | 0.188 | 0.576 | 0.591 | 0.743 | 0.083 | 0.880 | 0.833 | 0.857 |
| 绝对距离 | 0.295 | 0.663 | 0.350 | 0.649 | 0.205 | 0.620 | 0.526 | 0.735 | 0.091 | 0.891 | 0.800 | 0.857 |
| 相似离度 | 0.186 | 0.620 | 0.429 | 0.784 | 0.140 | 0.598 | 0.571 | 0829 | 0.000 | 0.902 | 1.000 | 1.000 |

由表 3 可以看出,综合相似系数评分在有无降水、分级降水和暴雨预报 3 个项目中的评分都是最高的。分级降水预报 $T_s$ 达到 0.46 以上,有无降水预报 $T_s$ 超过 0.5,大雨和暴雨预报 $T_s$ 达到 0.3 以上,在样本容量如此小的情况下,得到这个结果还算比较理想。值得注意的是,相关系数的评分也比较高,分级降水预报和有无降水预报高达 0.4 以上,降水效率均在 70 %以上,空漏报率也不高,从而说明对于 24h 相似预报而言,槽脊位相的相似是相当重要的;这就是为什么我们要在 24h 业务预报中突出形相似以提高预报准确率,集成相似系数和相关系数的最佳相似结果的原因。此外,24h 预报评分表现最差的是相似离度和欧氏距离。48 ~ 72h 预报,尤其是 72h 预报本应是 4 级动态预报第 4 级所要解决的问题。本试验利用前两级相似外延,制作 48h 以上时效的降水预报,准确率降低是必然的。为说明各种相似性系数随预报时效延长降低的程度,这里列出 48h 降水相似预报的结果于表 4 中。

表 4 2000 年 6 ~ 8 月各相似性判据的 48h 相似预报评分

| 相似性判据 | 有无降水预报 | | | | 分级降水预报 | | | | 暴雨预报 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $T_s$ | $EH$ | $FAR$ | $DFR$ | $T_s$ | $EH$ | $FAR$ | $DFR$ | $T_s$ | $EH$ | $FAR$ | $DFR$ |
| 综合相似系数 | 0.293 | 0.554 | 0.564 | 0.528 | 0.241 | 0.522 | 0.611 | 0.611 | 0.125 | 0.924 | 0.500 | 0.857 |
| 相关系数 | 0.231 | 0.565 | 0.571 | 0.667 | 0.173 | 0.533 | 0.640 | 0.750 | 0.143 | 0.935 | 0.000 | 0.857 |
| 相似系数 | 0.214 | 0.522 | 0.625 | 0.667 | 0.161 | 0.489 | 0.690 | 0.750 | 0.000 | 0.931 | 1.000 | 1.000 |
| 欧氏距离 | 0.196 | 0.554 | 0.600 | 0.722 | 0.137 | 0.522 | 0.696 | 0.800 | 0.000 | 0.891 | 1.000 | 1.000 |
| 绝对距离 | 0.235 | 0.576 | 0.556 | 0.667 | 0.196 | 0.554 | 0.600 | 0.722 | 0.000 | 0.902 | 1.000 | 1.000 |
| 相似离度 | 0.091 | 0.565 | 0.667 | 0.889 | 0.091 | 0.565 | 0.667 | 0.889 | 0.000 | 0.924 | 0.000 | 1.000 |

表 4 说明,随着预报时效的延长,相似预报 $T_s$ 评分一般下降很快,但综合相似系数还大体上保持着微弱的优势。$T_s$ 评分下降最快的是相关系数,$T_s$ 评分由 24h 的 0.44 猛降至 48h 的 0.17。绝对距离的评分下降比较慢。48h $T_s$ 评分表现较差的还是相似离度和欧氏距离。48h 大雨和暴雨的预报只有相关系数和综合相似系数有反映,但 $T_s$ 评分仅略大于 0.1,说明 48h 以上的大到暴雨预报,相似预报能提供的有价值的信息很少,这是相似预报应用中一个值得研究的问题。

## 5　结论

本文通过典型样品检验和业务现场对比试验,剖析了常用相似性判据的特点和性能,显示出相似性判据对相似预报质量的关键作用,并初步得出以下几点认识:

(1) 对于特定对象的相似预报,并非任何一种相似性判据都能获得较好的结果。根据预报对象性质,适当选取相似性判据十分必要。

(2) 相似性判据中有些偏重于值相似,有些偏重于形相似。相关系数的相似预报方法,因对形相似方面描述细致,在降水相似预报中是很有用的,尤其对 24h 预报尤为重要。

(3) 建立相似性判据时尽量考虑到形相似的同时,也要关注值的相似,不过根据问题性质不同可有所偏重。把相似系数、距离系数和相关系数适当综合起来,建立的综合相似系数在降水相似预报以及 24h 暴雨预报中,有较好的表现。

(4) 大气具有多态性,因而 48h 以上时效的相似预报 $T_s$ 评分迅速下降。我们不能期望一旦找到一个好的相似性判据,就可以随意向后外延预报。

(5) 48h 以上大到暴雨相似预报是一个亟待深入探讨的问题。

本文是针对降水问题讨论的,所得结果可供类似情况参考。相似预报不能预报超出历史样本以外的情况,受限于资料,本文不能涵盖因子序列间更多的组合,影响预报质量是可以理解的。当有可能扩大样本容量时,预报评分应当还会有所提高。

## 参考文献


1　黄嘉佑.气象统计分析与预报方法.北京:气象出版社,1990.255 - 257

2　李开乐.相似离度及其使用技术.气象学报,1986,44(2):174 - 182

3　Mcloy M C. Severe storm forecast results from the PROFS 1983 forecast experiment. *Bull. Amer. Meteoro. Soc.*,1986,63:155 - 164


# Tests for Conventional Similarity Criterions and Application of Composite Similar Coefficient


Yan Huifang　Li Shezong　Huang Yueqing　Zhang Xia

(Zhengzhou Meteorological Bureau, Henan Province, Zhengzhou 450005, China)



**Abstract:** Several similarity criterions most frequently used are tested and discussed objectively. A composite similar coefficient is introduced. The field experiments indicate that the coefficient is highly applicable in the analogue forecasting of precipitation.



**Key words:** similarity criterion, analogue forecasting method, composite similar coefficient