第 2 卷 第 1 期　　大 气 科 学　　Vol. 2, No. 1
1978 年 3 月　　SCIENTIA ATMOSPHERICA SINICA　　March, 1978

统计天气预报中相关系数的不稳定性问题

林 学 椿
（中国科学院大气物理研究所）

提　　要

统计天气预报中，不论多复杂的统计数学预报模型，都离不开预报量和预报因子的相关程度及其对未来时间的稳定性。本文着重讨论了相关系数的不稳定性。我们定义了滑动相关系数，计算表明相关系数是随时间变化的；利用相关系数的这种时间变化，可以改善未来时间的预报，而且以 10—20 年的滑动年数效果较好。

一、引　　言

统计天气预报在最近二十年内，有了较大的发展，取得了一定的效果[1]。在业务预报中，特别是气象要素的预报，更显出其优越性。在我国由于业务预报的需要，近十年来统计天气预报已普遍地开展起来，成为广大台站业务预报的一种重要而有效的手段.

统计预报经过几年试验，目前趋向于预报的物理化，力求预报因子和预报量之间有明显的物理内容，使二者之间有稳定的关连，以提高预报的准确率. 但往往有这样的情况：从大量的历史资料中挑选预报因子，可以很正确地拟合预报量，但预报却经常失败，预报准确率大大地低于历史资料的扦验准确率，说明预报因子和预报量之间对未来时间有不稳定的关系. 不管多元迴归，概率迴归还是判别分析以及更复杂的统计预报模型，要作好预报都离不开预报因子和预报量之间的相关程度及其对未来时刻的稳定性. 因此相关系数的稳定性对统计预报有决定意义. 本文主要讨论相关系数的不稳定性以及在预报上的应用.

二、一 个 例 子

为了简化问题，先从直线迴归方程(1)谈起

$$y = a + bx \tag{1}$$

以地面 7 月太平洋高压南界为 y 轴，地面 1 月西风指数为 x 轴，自 1873—1905 年(资料分别取自文献[2][3])，其分布如图 1. 可以看出 x 与 y 之间的关系并不好. 如果把资料分成 1873—1855 年和 1886—1905 年二段（分别为图中圆点和黑点），可以看出这二段都有近似直线关系，分段的迴归方程如下：

$$y_1 = 12.0 - 0.23x_1 \qquad (1873—1885\ 年) \tag{2}$$

$$y_2 = 22.9 + 0.32x_2 \qquad (1886—1905\ 年) \tag{3}$$

1977 年 7 月 23 日收到修改稿.

它们的相关系数分别为 $r_1 = -0.36$，接近于 10% 信度；$r_2 = +0.60$ 已超过 1% 信度，可见直线回归方程(2)、(3)，以一定的信度分别表示了 y 与 x 之间的直线关系．求得回归方程不是唯一的目的．我们关心的是根据自变量 x 来预报 y 的值以及预报精度如何？由方程(2)(3)可以看出同样的 x、y，由于时段的不同，它们的差别是很大的，连回归系数的符号都已经相反了，这种差别是由于抽样的偶然原因还是存在着本质的不同呢？

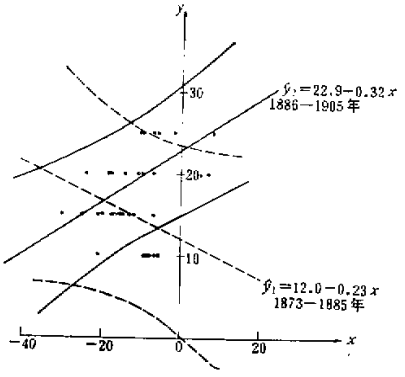

图 1 1 月地面西风指数与 7 月地面太平洋高压南界的直线回归(虚线为 1873—1886 年，实线为 1886—1905 年)

考虑因变量 y 的变化：

$$S_{yy} = \Sigma(y - \bar{y})^2 = \Sigma[(y - \hat{y}) + (\hat{y} - \bar{y})]^2$$
$$= \Sigma(y - \hat{y})^2 + \Sigma(\hat{y} - \bar{y})^2 + 2\Sigma(y - \hat{y})(\hat{y} - \bar{y})$$

很容易证明上式右边最后一项为零

$$S_{yy} = \Sigma(y - \bar{y})^2 = \Sigma(y - \hat{y})^2 + \Sigma(\hat{y} - \bar{y})^2 = \delta + Q$$

即 y 的变化可以分解为两部分，右边第二项 $Q = \Sigma(\hat{y} - \bar{y})^2$ 表示回归直线与平均数 $\bar{y}$ 之差的平方和．由回归方程(1)可知，它反映了在 y 的平方和中由于 x 和 y 的线性关系而引起的 y 变化部分，称之回归平方和．而右边第一项 $\delta = \Sigma(y - \hat{y})^2$ 表示观测点与回归直线之间的距离平方和，称之残差平方和．根据最小二乘方的原则，这个量是所有直线中最小的一个，它是除了 x 对 y 的线性影响之外的其他一切因素对 y 的平方和的贡献．它愈小，回归效果也就愈好，因此 y 的预测区间就可以写成[4]

$$a + bx \pm t_{0.05}\sqrt{\frac{\delta}{n-2}\left(\frac{n+1}{n} + \frac{(x-\bar{x})^2}{s_{xx}}\right)}$$

其中

$$S_{xx} = \Sigma(x - \bar{x})^2$$

这就是说对于固定的 x 值，y 落在这个区间的概率为 95%．可以看出预测区间与残差平方和的大小有关，残差平方和愈大，预测区间也愈大，反之残差平方和愈小，预测区间也愈小，y 的精度也愈高．同时它还和 x 值距离它的平均数的大小有关，距离越大，预测区间也大，因此它不是直线，而是曲线．回归方程(2)、(3)的预测区间分别如图 1 中的二条虚曲线和实曲线．对 1886—1905 年有 4 点在回归方程(2)的预测区间之外，而对 1873—1885 年则有 6 点在回归方程(3)的预测区间之外．这些落在预测区间之外的点子是由于抽样的结果还是反映二个回归方程之间有显著差别呢？对回归系数 b_1、b_2 进行统计扦验，其统计量 t 为：

$$t = \frac{|b_1 - b_2|}{\sqrt{\frac{\delta_1 + \delta_2}{n_1 + n_2 - 4}\left[\frac{1}{s_{x_1x_1}} + \frac{1}{s_{x_2x_2}}\right]}} = 2.78$$

查自由度 $f = n_1 + n_2 - 4$ 的 t 分布表 $t_{0.05} = 2.05$，$t_{0.01} = 2.76$，所以 $t > t_{0.01}$，是有显著差别的．同样可以对这二个相关系数作统计扦验[5]，可以证明它们之间的差异也是显著的．

通过这个例子可以看到，相同的 x、y 用不同时段得到的相关系数（或迴归系数）是不同的，尽管这时段分别的相关系数是可信的，但它们之间的差异有时会很大，甚至是显著的．相关系数的这种变化，引起了迴归方程的不稳定性．

三、滑动相关系数

为了进一步说明相关系数的变化，可以定义滑动相关系数来讨论这个问题．

$$R_{n,t}=\frac{\sum_{i=t-n+1}^{t}(x_i-\bar{x}_t)(y_i-\bar{y}_t)}{\sqrt{\sum_{i=t-n+1}^{t}(x_i-\bar{x}_t)^2\cdot\sum_{i=t-n+1}^{t}(y_i-\bar{y}_t)^2}} \tag{4}$$

其中

$$\bar{x}_t=\frac{1}{n}\sum_{i=t-n+1}^{t}x_i \qquad \bar{y}_t=\frac{1}{n}\sum_{i=t-n+1}^{t}y_i$$

$t=n,n+1,n+2\cdots\cdots$．　n 为滑动年数

为了便于说明问题，又使讨论不失一般性，用距平符号来计算滑动相关系数，则(4)式可以写成

$$R_{n,t}=\frac{\sum_{i=t-n+1}^{t}x_iy_i}{n}$$

$$=\frac{1}{n}(x_{t-n+1}y_{t-n+1}+x_{t-n+2}y_{t-n+2}+\cdots\cdots+x_{t-1}y_{t-1}+x_ty_t) \tag{5}$$

其中 $x_i=\begin{cases}+1\\-1\end{cases}$　$y_i=\begin{cases}+1 & \text{当 } x_i,y_i \text{ 为正距平时}\\-1 & \text{当 } x_i,y_i \text{ 为负距平时}\end{cases}$　$t=n,n+1\cdots\cdots$．

图 2-4 是按(5)式的计算结果．图 2 是上海站 5 月降水量距平符号与次年上海站5—8 月总降水量距平符号的相关．由图可见，相关系数是随时间变化的．由 n = 5 到 n=80 都有这种变化，只不过随着滑动年数 n 的增加，相关系数变化的振幅不断减少，即使用 80 年来滑动，相关系数还是可以由 0.10 变到−0.03．对各个不同的滑动年数，相关系数的变化也是不同的，n = 5 的相关系数变化就比较大，从 +1.00 变到 −1.00，波动数目也比较多．从 n = 10 到 n = 25，相关系数的变化比较稳定，基本上由二个波组成．n ≥ 30，相关系数的变化就比较小，只有一个波．用同样的方法，计算了 30 组这类相关系数，发现相关系数都是随时间变化的．为了减少篇幅，仅举图 3-4 为例．图 3 是 10 年滑动相关系数．图 4 是 30 年滑动相关系数．可见对各种不同气象要素，相关系数都是随时间变化的．

大家都知道，大气过程是极复杂的，受着各种不同尺度的物理因子的影响．长期天气过程存在着各种不同长度的周期，如世纪周期，35—40 年周期，22 年周期，11 年周期，以及低纬度的 2 年周期等等．所谓周期，只有在统计意义下才成立．换言之，所谓周期并不是指在一定时间间隔内必然重复出现的现象，而是指在统计条件下的重复出现．例如有

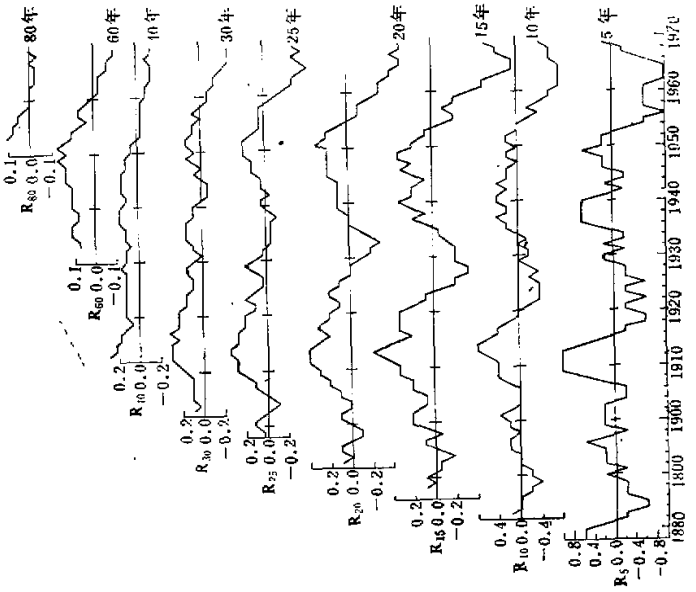

图 2 上海 5 月降水量对次年上海 5—8 月降水总量距平的滑动相关系数。

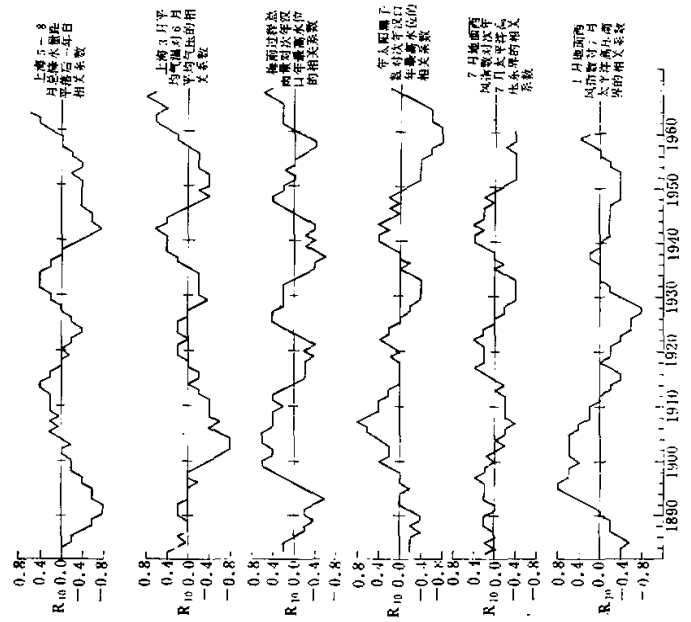

图 3 10 年滑动相关系数

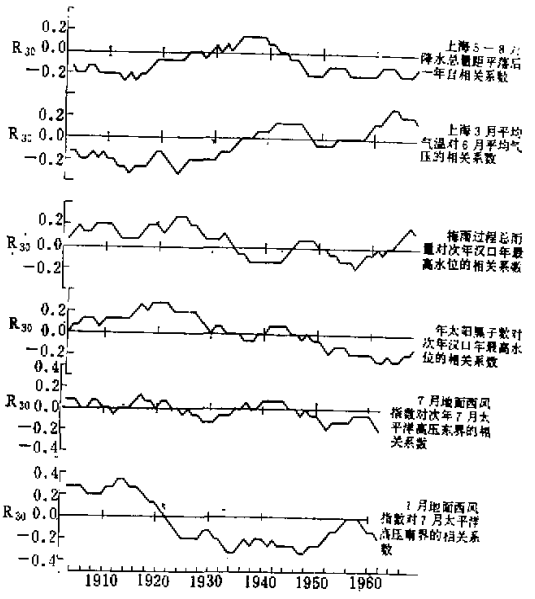

图 4　30 年滑动相关系数

名的太阳黑子 11 年周期，并不是说每隔 11 年太阳黑子数都出现高峰(或低谷)，而是指平均讲有11年左右峰值(或谷点)重复出现，具体说每次峰值(或谷点)的出现是不同的. 可以隔 8 年，也可以隔 12 年，而且值也是不同的. 另一方面，一个周期在不同的时段也是有变化的. 例如有人[8]在研究南方涛动，对达尔文站的季节地面气压作谱分析时，发现在 1882—1926 年 2.5 年左右周期非常显著，到 1921—1962 年 2.5 年左右周期已不明显了，而 5 年左右周期变得更重要了. 因此，不同长度而且随时间变化的各种周期迭加在一起，组成了复杂的气象要素时间序列，对这样复杂的时间序列求相关，相关系数随时间的变化应该说是普遍的现象. 例如二个简单的正弦波(图 5)只要它们的位相角或周期不同，按(5)式计算相关系数，必然是随时间变化的，更何况这样复杂的气象要素时间序列. 东非维多利亚湖的水位自 1899 年以后基本上随太阳黑子数而升降，二者的相关系高达0.87，这个关系一直当做日地关系中的典型事例被许多作者所引用.但自1927 年以后，湖面水位峰值在每11年黑子周期中出现两次，即 1932，1937，1942，1947，1952，1957 和 1964 年，这些峰值都是在黑子极小值和极大值年份或附近，也就是说，湖水水位的周期在 1927 年以后缩短了，致使与太阳黑子的关系不一致，甚至变得相反了[9]. 有人研究南方涛动时，发现同时期的 6—8 月气压之间，达尔文站和檀香山站的相关系数在

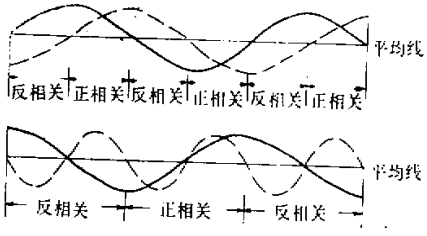

图 5　相关系数变化可能原因示意图.

1882—1921 年为−0.66，到 1921—1950 为−0.12；科伦坡和圣地亚哥的相关系在 1875—1921 为−0.65 到 1921—1950 为−0.21[8]。印度西南季风雨量和其他气象要素的相关系数在不同时段也是不同的[6]。实际上我们知道气象要素的平均值在不同时期是不同的，是随时间变化的；同样它们的均方差也是随时间变化的，因此严格讲大气是一个非平稳过程。

统计预报有一个最基本的假定，即用样本资料求得预报因子和预报量的关系，假定在未来时刻是不变的。由上述讨论可知，相关系数是随时间变化的，用这种不稳定的相关系数去做预报实际上是很危险的，往往会造成失败。例如在图 2 的 15 年滑动相关系数中，挑出 1913 年的高点，其相关系数为 0.60，即相当于过去 15 年中有 12/15 是同符号的，用此作预报，一直到 1928 年，在这未来的 15 年中，报对的只有 5 次。用低点 1928 年的负相关作预报，一直到 1949 年，在这 21 年中报对的只有 8 次。同样用这些相关系数去求迴归方程作预报往往也要失败，例如上海 5—8 月总降水量距平符号的迴归方程如下

$$\hat{y} = -0.02 - 0.07x_1 + 0.08x_2 - 0.18x_3 + 0.04x_4 + 0.03x_5 + 0.20x_6 \qquad (6)$$

其中 $\hat{y}$ 为上海次年的 5—8 月总降水量距平符号；

x_1 为上海上年的 5—8 月总降水量距平符号；

x_2 为上海上年 5 月总降水量距平符号；

x_3 为上海上年 11 月总降水量距平符号；

x_4 为上海上年 6 月地面气压距平符号；

x_5 为上海上年 10 月地面气压距平符号；

x_6 为上海同年 3 月地面温度距平符号；

用 1873—1950 年资料作检验，其准确率为 0.64，还是可以用的，但对 1951—1969 年的资料作验证，其准确率下降为 0.26，还不如盲目预报高（见表 2）。原因很清楚，预报因子和预报量之间的关系，在未来 19 年已经变化了。是否可以利用相关系数的时间变化，来改进我们的预报？下面做一个最简单的试验。

四、滑动相关系数在预报中的应用

由图 2—4 可看到相关系数是时间的函数，但相关系数随时间的变化要远比气象要素随时间变化简单，特别是 10 年以上的滑动相关系数更是如此。很容易估计出未来相关系数的变化趋势，有了这个趋势就可以估计出未来时间预报量和预报因子是正相关还是反相关。

因为，由(5)式可知

$$R_{n,t+1} - R_{n,t} = \frac{1}{n}\sum_{i=t-n+2}^{t+1} x_i y_i - \frac{1}{n}\sum_{i=t-n+1}^{t} x_i y_i$$

$$= \frac{1}{n}(x_{t+1}y_{t+1} - x_{t-n+1}y_{t-n+1})$$

若 $R_{n,t+1} - R_{n,t} > 0$ 即相关系数趋势上升

则 $x_{t+1}y_{t+1} > 0$ 必须正相关

若 $R_{n,t+1} - R_{n,t} < 0$ 即相关系数趋势下降

则 $x_{t+1}\,y_{t+1} < 0$ 必须反相关

若 $R_{n,t+1} - R_{n,t} = 0$ 即相关系数不升不降

则 $x_{t+1}\,y_{t+1} = x_{t-n+1}\,y_{t-n+1}$ 即等号两端同号

具体的预报思路如下：利用单因子和预报量之间的相关系数变化可以估计出未来相关系数的趋势是上升还是下降，因而也就知道预报因子和预报量之间在未来时刻是正相关还是反相关，再以预报因子报出预报量 $\hat{y}$；n 个预报因子都可以作如此的预报，得到 n 个预报的预报量 $\hat{y}$，再对 n 个 $\hat{y}$ 求逦归方程使改善未来的预报．而相关系数的趋势估计则用最简单的方法．

凡符合下列二个条件者为峰点(或谷点)：

(1) 滑动相关系数要连续上升(或下降) 5 年或 1/2 滑动年数以上

(2) 接着要连续下降(或上升)二年．其转折点的相关系数是正的(或负的)，则此点定义为峰点(或谷点)

按上述定义，出现峰点后则报相关系数下降，即未来 x，y 是反相关，直到谷点出现；谷点之后则报相关系数上升，即未来 x、y 是正相关，直到峰点．仍然用(6)式的预报量和因子 $x_1, x_2 \cdots\cdots x_6$，按上述定义估计未来相关系数的趋势．用 1873—1950 年资料作检验，其准确率如表 1．例如 R_2 栏中 20 年滑动相关系数预报的准确率为 0.64，即表示由 x_2 和 y 组成的 20 年滑动相关系数按上述方法预报 58 年中有 37 年是报对的．由表 1 可以看出大部分的准确率都在 0.60 以上．另外还可以看到随着滑动年数的增加，准确率有些下降，这是因为滑动相关系数和滑动平均有些相似，随着滑动年数的增大，而平滑了很多细节，使准确率下降．

表 1

	1873—1950					
	R_1	R_2	R_3	R_4	R_5	R_6
5 年滑动相关	0.75	0.63	0.70	0.74	0.74	0.74
10年滑动相关	0.76	0.66	0.74	0.71	0.76	0.66
15年滑动相关	0.62	0.65	0.65	0.70	0.65	0.68
20年滑动相关	0.62	0.64	0.62	0.66	0.64	0.60
25年滑动相关	0.60	0.62	0.70	0.70	0.64	0.66
30年滑动相关	0.63	0.58	0.71	0.63	0.69	0.56

有了未来时间的相关系数趋势，就可以报出预报量 $\hat{y}$，对 6 个 $x_1, x_2 \cdots\cdots x_6$ 的预报因子，就可得到 6 个预报量的估计值 $\hat{y}_1\,\hat{y}_2 \cdots\cdots \hat{y}_6$，对这 6 个估计量再求逦归，其方程如下

$$\hat{y}_5 = 0.03 + 0.17\hat{y}_1 + 0.27\hat{y}_2 + 0.08\hat{y}_3 + 0.31\hat{y}_4 + 0.20\hat{y}_5 + 0.35\hat{y}_6$$

$$\hat{y}_{10} = 0.01 + 0.34\hat{y}_1 + 0.10\hat{y}_2 + 0.26\hat{y}_3 + 0.19\hat{y}_4 + 0.32\hat{y}_5 + 0.12\hat{y}_6$$

$$\hat{y}_{15} = -0.17 + 0.21\hat{y}_1 + 0.29\hat{y}_2 + 0.28\hat{y}_3 + 0.26\hat{y}4 + 0.13\hat{y}_5 + 0.38\hat{y}_6$$

$$\hat{y}_{20} = -0.02 + 0.32\hat{y}_1 + 0.24\hat{y}_2 + 0.13\hat{y}_3 + 0.23\hat{y}_4 + 0.21\hat{y}_5 + 0.14\hat{y}_6$$

$$\hat{y}_{25} = 0.00 + 0.12\hat{y}_1 + 0.16\hat{y}_2 + 0.20\hat{y}_3 + 0.25\hat{y}_4 + 0.10\hat{y}_5 + 0.06\hat{y}_6$$

$$\hat{y}_{30} = 0.02 + 0.10\hat{y}_1 + 0.02\hat{y}_2 + 0.37\hat{y}_3 + 0.16\hat{y}_4 + 0.30\hat{y}_5 + 0.01\hat{y}_6$$

其中 $\hat{y}_5$、$\hat{y}_{10}$……$\hat{y}_{30}$ 表示用 5、10……30 年滑动关系数作出的预报

表 2

		1873—1950年						1951—1969			
		实况			概括率	S_1	S_2	实况			概括率
		+	−	合计				+	−	合计	
一般回归	+	25	16	41				2	10	12	
予报	−	12	24	36				4	3	7	
	合计	37	40	77	0.64	0.27		6	13	19	0.26
5年滑动	+	30	4	34				2	4	6	
相关系数	−	6	33	39				4	9	13	
的予报	合计	36	37	73	0.86	0.73	0.62	6	13	19	0.58
10年滑动	+	31	5	36				1	1	2	
相关系数	−	3	29	32				5	12	17	
的予报	合计	34	34	68	0.88	0.76	0.67	6	13	19	0.68
15年滑动	+	26	5	31				3	2	5	
相关系数	−	5	27	32				3	11	14	
的予报	合计	31	32	63	0.84	0.68	0.56	6	13	19	0.74
20年滑动	+	22	6	28				2	4	6	
相关系数	−	7	23	30				4	9	13	
的予报	合计	29	29	58	0.78	0.49	0.38	6	13	19	0.58
25年滑动	+	20	6	26				2	3	5	
相关系数	−	7	20	27				4	10	14	
的予报	合计	27	26	53	0.75	0.50	0.35	6	13	19	0.63
30年滑动	+	21	8	29				2	3	5	
相关系数	−	4	15	19				4	10	14	
的予报	合计	25	23	48	0.75	0.50	0.31	6	13	19	0.63

用 1873—1950 年的资料作检验，用 1951—1969 年的资料作试报，如表 2. 表中的 S_1、S_2 为技术得分，定义如下[7]：

$$S = \frac{F - E}{N - E}$$

N为总的预报次数，F为预报成功的次数，E为盲目预报或其他方法预报的成功次数，而盲目预报的成功次数 $E_i = n_i c_i / N$，n_i 为联列表中第 i 级天气出现次数，c_i 为联列表中预报第 i 级天气的次数. S_1 是滑动相关系数预报与盲目预报的比较；S_2 是滑动相关系数预报与(6)式预报的比较.

从表中可以看出，不管那一种滑动相关系数预报比盲目预报要高 50%，比(6)式预报要高 30%以上，而 (6) 式预报比盲目预报高 27%. 对 1873—1950 年时段，准确率都在 75%以上，而对 1951—1969 年作试报，准确率有所下降，但还保持在 70% 左右. 准确率下降的原因是因为定义峰点(谷点)时，已经使用了峰点(谷点)出现后的两年，预报时只好牺牲这两年. 另外可以看出 10—15 年滑动相关系数预报比较稳定，效果也好，这也是可以理解的. 因为滑动年数增大，平滑了一些细节，使准确率下降，而滑动年数太短，则使相

关系数的波动增多，减少准确率．例如，5 年滑动相关系数的准确率下降是最大的．

五、讨　论

由上可知，大气中各种气象要素都是时间的函数，两个要素的相关系数也是时间的函数，随时间而变化．我们定义出滑动相关系数，计算表明相关系数是随时间变化的；利用相关系数的这种变化，可以改善预报，而且以 10—20 年的滑动年数效果较好．

由于相关系数是变化的，因此回归系数也不是常数，也是随时间变化的，可以设想，预报因子在不同的时段，所起的作用也是不同的．它的重要性也会变化，某一时段预报因子甲很重要，到另一时段，预报因子乙变得重要了，而甲变得次要了．如果用逐步回归，预报因子在不同时段都可能是不同的．这些问题将以后讨论．

本文承杨鉴初同志指导和鼓励，李麦林同志审阅全文并提出宝贵意见，特此致谢！

参　考　资　料

[1] 李麦村，统计预报的进展，近代气象学若干问题的进展，科学出版社，1975 年．
[2] 王绍武，近 90 年大气环流的振动(下)，气象学报，1965，第 35 卷，第 2 期．
[3] 朱炳海，东亚及西太平洋地区大气流场的演变，气象学报，1965，第 35 卷，第 1 期．
[4] 中国科学院数学研究所数理统计组编，回归分析，科学出版社，1975 年．
[5] 周华辛编，工业技术应用数理统计学，高等教育出版社，1954 年．
[6] Y. P. Rao, Southwest Monsoon, India Meteorological Department, 1976.
[7] 王宗皓、李麦村编著，天气预报中的概率统计方法，科学出版社，1974 年。
[8] A. J. Troup, The Southern Oscillation *Quart, J. R. Met. Soc.*, 1965, Vol. 91, No. 390.
[9] 张家诚等编著，气候变迁及其原因，科学出版社，1976 年．