第40卷 第4期 气 象 学 报 Vol. 40, No. 4
1982 年 11 月 ACTA METEOROLOGICA SINICA Nov., 1982

相关系数稳定性分析方法及其应用*

朱 盛 明

(江苏省气象科学研究所)

相关和回归的方法,目前仍是制作统计天气预报的主要方法,也是把定性估计推进到定量预报的主要手段。通过相关分析,确定变量之间关系的密切程度,选取天气学意义明显的相关区,或寻求具有一定物理内容的预报因子,在此基础上使用回归等方法进行综合推断。人们在普查相关和对回归方程中因子进行筛选时,常常给定了较高的显著水平,如 $\alpha=0.1,0.05$ 等。但把最后建立的回归方程,付诸实际应用后,方程的可靠性和稳定性变化很大,准确率通常只有 50%左右。1973 年我们曾在分析拟合和预报显著差异的成因之后,提出了相关系数稳定性分析方法,并对梅雨预报方程进行了改进。后来,此方法又推广应用于春季灾害性、关键性天气预报方程。在业务预报中,这一方法已应用七年,预报效果一般优于常规方法。实践证明,相关系数稳定性分析方法对提高统计预报稳定性,具有一定效果。

1. 相关系数的非平稳性

气象要素之间的相关系数,可以看成是一个随机过程 r_t。检验其平稳性的方法很多,我们不妨采用直观的方法,对 52—72 年共 21 年样本,以 10 年为长度计算滑动相关系数,对梅雨有关的 298 个因子进行普查,发现 r_t 随时间的波动是极普遍的。以梅雨开始日期与 11 月雨日数、12 月≥8 级风日数、2 月份平均气压、2 月份积雪日数的相关系数随时间变化曲线为例。如图一所示,相关系数随时间的变化是相当明显的。图 1 a 中 56—65 年的相关系数为−0.25,显著水平不足 0.20。而 62—71 年升为+0.65,显著水平超过0.05。1 b 和 1 c 有类似的情况。图 1 d 则相对比较稳定,56—65 年相关系数从最高点的+0.65,显著水平超过 0.05,下降到 59—68 年的+0.38,而显著水平仍超过 0.10。

相关系数非平稳变化说明以样本资料求得较好的相关系数,既不能保证过去的年代里始终有这样的相关,更不能预测未来是否仍保持不变。我们在选择因子时,实际上是选取了高相关的峰点(或谷点),但对于不稳定的因子,相关系数一达到高点就会很快下降,原来的回归系数已不能表示因子与因变量之间的关系,预报与拟合的效果就会有很大差

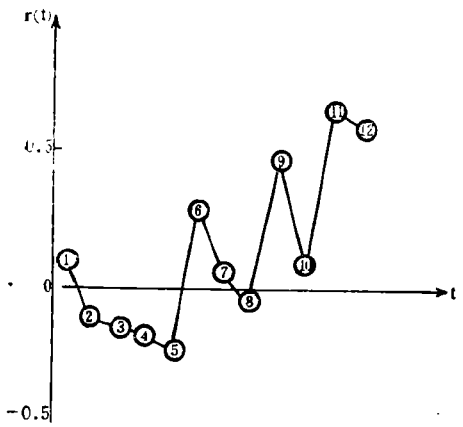
* 本文于 1981 年 3 月 3 日收到,1981 年 9 月 4 日收到修改稿。

异。如 1 *b* 中若以 52 到 65 年(1—5)这一段资料拟合建立方程，则复相关系数一定很低，不能反应 68 年以后(8—12)较高的相关关系。反之，对 1 *c* 而言若用 52 到 70 年 (1—10) 资料建立的方程，拟合度虽高，但 71 到 72 年以后(11—12)的使用效果一定很差。

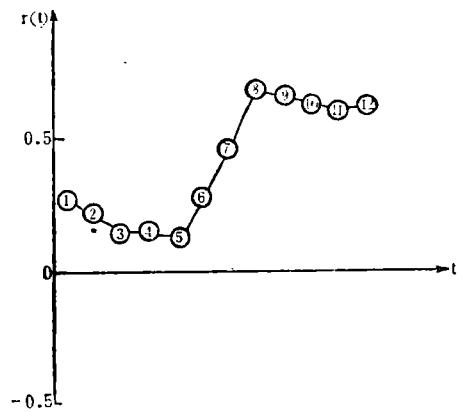
这就是说，我们在对相关系数作统计检验时，应用的是样本相关系数 r 。它与不依赖于时间的总体相关系数 ρ 之间的关系，可表示为

$$r=\rho+\varepsilon,$$

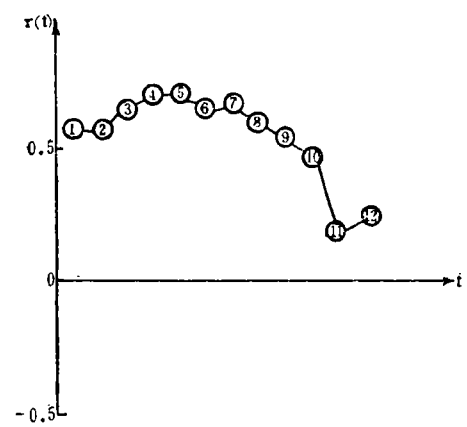
其中 ε 是随机扰动。



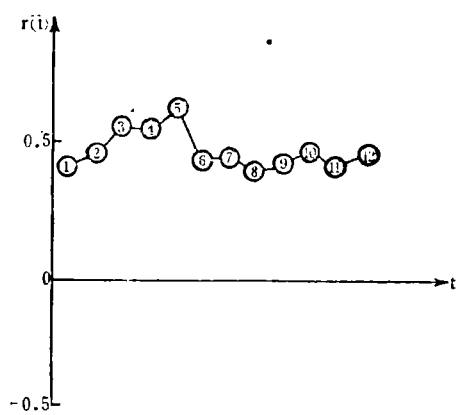
1 *a* 梅雨始日与上一年 11 月雨日



1 *b* 梅雨始日与上一年 12 月份≥8 级风日



1 *c* 梅雨始日与 2 月份平均气压



1 *d* 梅雨始日与 2 月份积雪日数

图 1 滑动相关系数 r_i 随时间变化曲线

(图中①②……⑪⑫分别表示 52—61、53—62……62—71、63—72 年的 r_i 值。)

当我们用统计检验判据把绝对值大的 r 分离出来时，不仅把 $|\rho|$ 较大、$|\varepsilon|$ 较小的关系录用，而且也可能把 $|\rho|$ 相对较小，而 $|\varepsilon|$ 较大的因子也一并引进来。

2. 相关系数稳定性分析方法

既然相关系数的非平稳性是客观存在，统计检验也无能为力，面对只有廿余年的短资料序列，又不好进行同族性检验[1]，更不容许我们按相关系数波动来分段取舍样本。预报服务的需要又决不能等待着非平稳随机过程在数学处理上得到完善解决后，再制作预报。应采用那种实用的方法来选取相关系数比较平稳的因子，使得我们所录用的因子与因变量的关系接近大多数统计方法的线性平稳性假设呢？

我们建议选用下列各统计量来考察相关系数的稳定性：

$r(q)$：以 q 年为长度的滑动相关系数，$m(m\geqslant q)$个样本共有 $k=m-q+1$ 个滑动相关系数。

对于 21 个样本，经试验各种 q 值，以 $q=10$ 年为宜，共有 12 个滑动相关系数作 $r_1(q)$、$r_2(q)\cdots r_{11}(q)$、$r_{12}(q)$。

$\overline{r(q)}$：k 个 $r(q)$的平均值即 $r(q)$的摆动水平。

$\sigma_{r(q)}$：$r(q)$偏离$\overline{r(q)}$的均方差。

$r(s)$：表示从 m 个样本中，分别舍去最后一个、二个、三个样本求得的相关系数，$S=m,\ m-1, m-2$。

$FZ(S)=\dfrac{1}{2}\ln\dfrac{1+r(s)}{1-r(s)}\cdot\sqrt{S-3}$：对 $r(s)$作真相关系数 $\rho=0$ 的检验。

为方便计，仍以图 1 中四个相关系数的变化曲线为例，并给出上述统计量如表 1。图 1 a 的$\overline{r(t)}$偏低，$\sigma_{r(q)}$ 也偏大，$r(s)$通过检验的显著水平也较低。虽然最后两个滑动相关系数 $r_{11}(q)$、$r_{12}(q)$均大于 0.60，也很容易变小。对于这一类因子，不论用何种方法都会被删去。再比较图 1 b 和 1 c、1 d，$r(s)$都在 0.40 左右或以上，1 b 和 1 c 的 $r(s)$有一半通过 0.05 的显著水平的检验，但图 1 b 的方差较大，1 c 则近期相关系数很低，比较起来虽然1 d 的 $r(s)$不是最高，但相关系数的变化比较平稳，而且近期还略有上升，所以是较

表 1　对应于图一中四条 r_i 曲线的 $\overline{r(q)}$、$\sigma_{r(q)}$、$r_{11}(q)$、$r_{12}(q)$ 以及 $r(s)$、$F_z(s)$ 的值

	图 1 a			图 1 b			图 1 c			图 1 d		
	$\overline{r(q)}$	$\sigma_{r(q)}$		$\overline{r(q)}$	$\sigma_{r(q)}$		$\overline{r(q)}$	$\sigma_{r(q)}$		$\overline{r(q)}$	$\sigma_{r(q)}$	
$k=10$	0.0214	0.2237		0.4072	0.2415		0.5883	0.0782		0.4634	0.1086	
$k=11$	0.0759	0.3003		0.4081	0.2300		0.5477	0.1586		0.4766	0.0965	
$k=12$	0.1515	0.3366		0.4453	0.2260		0.5118	0.1896		0.4785	0.0962	
$r_{11}(q)$	0.6525			0.6321			0.1539			0.4512		
$r_{12}(q)$	0.6292			0.6407			0.2147			0.4732		
	$r(s)$	FZ	α	$r(s)$	FZ	α	$r(s)$	FZ	α	$r(s)$	FZ	α
$s=19$	−0.1282	−0.5160	0.60	0.4449	1.9132	0.06	0.4022	1.7052	0.09	0.4006	1.6975	0.09
$s=20$	−0.3459	−1.4873	0.14	0.4754	2.1316	0.03	0.4357	1.9250	0.06	0.3772	1.6358	0.10
$s=21$	−0.2903	−1.2683	0.20	0.4758	2.1955	0.03	0.4368	1.9865	0.05	0.4107	1.8519	0.06

为理想的因子。

对梅雨开始日期、梅雨期长度以及南京梅雨量等有关当年 2 月份以前的 298 个因子作了计算和统计分析，确定出通过稳定性分析的甲、乙二类因子的标准如表 2 所示。据表 2，把 298 个因子中，全样本相关系数通过显著水平 0.1 的因子作为初选因子，再进行相关系数稳定性分析，录用因子数见表 3。

表 2 通过稳定性分析的标准

	$\overline{r(q)}$	$\sigma_{r(q)}$	$r_{11}(q), r_{12}(q)$	$FZ(s)$
甲类因子	≥0.40	≤0.15	≥0.40	取显著水平 $\alpha=0.1$ 则分位值为 1.96
乙类因子(1)	<0.40 但 >0.35	>0.15 但 <0.20	<0.40 但 >0.35	
乙类因子(2)	≥0.40	≤0.15	<0.35 但 >0.30	

表 3 通过稳定性分析的因子表

	梅雨始日	梅雨长度	南京梅雨量
初选	40	40	44
乙类	14	28	18
甲类	12	19	9

分别对甲类、乙类因子求出逐步回归方程，自 1973—1979 共七年，在业务预报中，与常规方法(即用初选因子，直接进行逐步筛选回归，建立预告方程)同时使用。表 4 给出甲类与常规预报效果检验结果。用相关系数稳定性分析方法，所挑选的因子建立的回归方程预报效果比常规方法有所提高。初选因子数量大，但它的预报效果反而下降，表明建立预报效果较好的方程的关键在于初选因子的质量，特别是相关稳定性，而不在于投入因子的多少。

表 4 1973—1979 年梅雨预报检验表

因变量	方法 \ 年份	1973		1974		1975		1976		1977		1978		1979	
	项目	预报	实况	预报	实况	预报	实况	预报	实况	预报	实况	预报	实况	预报	实况
梅始日	常规	6.24	6.16	6.30	对有较大分歧意见	7.5	6.21	6.19	6.15	6.26	6.28	6.20	6.24	6.16	6.19
	甲类	6.21		6.20		6.26		6.15		6.25		6.23		6.18	
梅雨长度	常规	7 天	16天	2 天		14天	21天	13天	31天	17天	24天	5 天	2 天	16天	22天
	甲类	14天		14天		22天		29天		24天		5 天		18天	
南京梅雨量	常规	355.2	149毫米	278.8		210.3	474毫米	149.2	266毫米	167.5	146毫米	116.7	24毫米	383.4	229毫米
	甲类	172.3		352.0		444.8		293.0		329.1		259.4		857.1	

3. 小　结

73 年以来，在业务值班中应用七年，这期间包括弱梅雨的 73 年，也包括强梅雨的 75 年，以及空梅雨的 78 年，预报效果尚称理想。总预报准确率，用一般的常规方法为 40%，而用相关系数稳定性分析方法为 72%。为进一步检验这一方法的应用效果，75 年起曾用此法于春播期稳定通过 10°C、12°C、18°C初日，三麦渍害、终霜期、早稻秧苗期、生长期冷害以及低温连阴雨等共 8 个项目的预告，总的准确率为 72.5%，亦比一般方法好。

由于相关系数不稳定性而造成的历史资料拟合率高，而预报效果下降，是统计预报中急待解决的问题。林学椿[2]曾提出从滑动相关系数变化曲线来估计它未来的变化趋势是正相关还是负相关，然后用各种单因子分别求出预报量的估计值，再用集成的方法，求出预报值 $\hat{y}$。用独立样本检验证明此方法的预报效果优于常规方法，尤以 10—15 年的滑动相关系数预报效果最好。本文作者针对实际预报工作中大量存在的小样本问题，于 1973 年提出了这个定量的稳定性分析方法。实践证明，对改善预报稳定性能起一定的作用。但分析表明，一般高相关持续 3 到 4 年，以后原来相关较高的因子变得不重要了，而另一些次要因子，相关关系提高了，所以回归方程在用了 3 到 4 年后，回归系数亦将起变化，准确率就会出现下降的趋势。这是一个客观规律，任何处理方法都改变不了。而相关系数稳定性分析方法，正是讨论如何尽可能地符合这一客观规律，挑选出近期内相关显著而且相关关系稳定的因子。所以，一个方程在使用 3—4 年后要用相关系数稳定性分析方法，重新建立预报方程。

参 考 文 献

[1] Rao, C. R., Linear Statistical Inference and Its Application, Wiley, New York, 1973, 432—435.
[2] 林学椿，统计天气预报中相关系数的不稳定性，大气科学，2，No.1，55—63，1978。