正规化方差分析及其在长期天气预报中的应用

陈創买

（气象学系）

用方差分析寻找时间序列周期的方法，在气象水文的中、长期预报中已被广泛使用，并取得一定的效果。在概率统计意义下，气象序列变化的周期极为复杂[1,2]，而普通方差分析在实际应用中受到较大限制。本文旨在探讨一种既能较好地消除年气候变化趋势的影响，又易于掌握非一年整倍数周期变化的正规化方差分析法。

一、基本原理

设以月(或季、旬、候等，下同)为时间单位的气象要素连续观测序列，按年份排列：

$$\begin{pmatrix} X_{11}, & X_{12}, & \cdots\cdots & X_{1M_2} \\ X_{21}, & X_{22}, & \cdots\cdots & X_{2M_2} \\ \cdots & \cdots & \cdots & \cdots \\ X_{M_1 1}, & X_{M_1 2}, & \cdots\cdots & X_{M_1 M_2} \end{pmatrix} \quad (1)$$

式中横行代表年份(共M_1年)，纵列代表月份(共M_2个月)。或写成：

$$(X_{ij}),\ i=1,2,\cdots\cdots M_1;\ j=1,2,\cdots\cdots M_2 \quad (2)$$

为了更有效地消去年气候变化趋势的影响，我们将(1)式按纵列作正规化(标准化)处理[3]，即计算

$$\widetilde{X}_{ij}=(X_{ij}-\overline{X}_j)/S_j \quad (3)$$

$$(i=1,2,\cdots\cdots M_1;\ j=1,2,\cdots\cdots M_2)$$

其中

$$\overline{X}_j=\frac{1}{M_1}\sum_{i=1}^{M_1}X_{ij};\ (j=1,2,\cdots\cdots,M_2) \quad (4)$$

代表月平均值，而

$$S_j=\sqrt{\frac{1}{M_1}\sum_{i=1}^{M_1}(X_{ij}-\overline{X}_j)^2} \qquad (5)$$

$$(j=1, 2, \cdots\cdots M_2)$$

代表月标准差值。

将(3)式的全部值按逐月时间顺序排列，就成为正规化时间序列：

$$(\widetilde{X}_t); \quad t=1, 2, \cdots\cdots, (M_1\times M_2) \qquad (6)$$

为了充分利用近期最新信息，应将(1)式中自X_{M1M2}项以后新观测到的$L+L_1$个资料(称为序列的尾数)，除留下L_1个作检验之用外，逐个按(3)式作正规化处理，加入到(3)式序列的末尾，于是有

$$(\widetilde{X}_t), \quad t=1, 2, \cdots\cdots, (M_1\times M_2+L) \qquad (6')$$

设$\widetilde{X}_t$序列的总(离差)平方和为$\widetilde{e}_O$，则可证明：

$$\widetilde{e}_O=\sum_{t=1}^{M_1\times M_2+L}\widetilde{X}_t^2\simeq M_1\times M_2+L \qquad (7)$$

令$M_1\times M_2+L=N$，则$\widetilde{e}_O\simeq N$。

我们将$\widetilde{X}_t$序列看成是由若干个具有概率统计意义的显著周期函数和平稳随机函数叠加而成的，于是便可应用方差分析方法，从长度为$l=2,3,\cdots\cdots,[\frac{N}{2}](=\frac{N}{2}$，当$N$为偶数；或$=(N-1)/2$，当$N$为奇数)的各种可能周期中，通过以周期的$F$—分布下侧概率值为参数的周期显著性检验去识别显著周期的办法进行周期普查，并将这种显著周期从$\widetilde{X}_t$序列中分离出来。这项工作是逐次进行的。

首先，计算F_l—统计量，即取$l=2, 3, \cdots\cdots, [\frac{N}{2}]$，将$\widetilde{X}_t$序列的$N$个数据，按长度$l$截成$T$段，每段$l$个组(但最后一段可能不足$l$个组)，排列成$T$行(段)$l$列(组)如表1所示。

然后计算：

$$F_l=\frac{(B_l-C)}{(A-B_l)}\frac{(N-l)}{(l-1)} \qquad (8)$$

其中
$$\begin{cases} A=\sum\limits_{v=1}^{l}\sum\limits_{u=1}^{T_v}\widetilde{X}_{uv}^2 \\ B_l=\sum\limits_{v=1}^{l}\left(\frac{1}{T_v}\left(\sum\limits_{u=1}^{T_v}\widetilde{X}_{uv}\right)^2\right) \\ C=\frac{1}{\sum\limits_{v=1}^{l}T_v}\left(\sum\limits_{v=1}^{l}\sum\limits_{u=1}^{T_v}\widetilde{X}_{uv}\right)^2 \end{cases} \qquad (9)$$

表 1　数 据 表

组序号 v / 段序号 u	1	2	……………………	l
1	$\tilde{x}_{11}$	$\tilde{x}_{12}$	……………………	$\tilde{x}_{1l}$
2	$\tilde{x}_{21}$	$\tilde{x}_{22}$	……………………	$\tilde{x}_{2l}$
⋮	⋮	⋮		……
T－1	$\tilde{x}_{(T-1)1}$	$\tilde{x}_{(T-1)2}$	……………………	$\tilde{x}_{(T-1)l}$
T	$\tilde{x}_{T1}$	$\tilde{x}_{T2}$	………… $\tilde{x}_{lv_0}$ *)	

这里 $A=\tilde{\theta}_0$，即代表 $\tilde{X}_t$ 序列的总离差平方和，C 为 $\tilde{X}_t$ 序列总和的平方与总数之商，B_l 代表 l 个组中各组内总和平方的平均之总和。T_v 代表在 l 个组中第 v 组的总项数，且 $\sum_{v=1}^{l} T_v = N$ 。

这个 F_l—统计量，当已知 l 个总体服从正态分布和方差相等时，在原假设 H_0：$\mu_1=\mu_2\cdots\cdots=\mu_l$（即各总体的均值相等）成立的条件下，服从自由度为 $n_1=l-1$，$n_2=N-l$ 的 F—分布，因此可以应用 F—检验。

但是，因为在 F—检验中，对一定信度 α 的 F—分布临界值 $F_\alpha(n_1, n_2)$ 是 n_1 和 n_2 的函数，不同的 n_1 和 n_2 值，就对应有不同的 $F_\alpha(n_1, n_1)$ 值，造成在选择显著周期时对比上的困难，即使是给出 $F_\alpha(n_1, n_2)$ 分布表也难以解决。因为这种分布表的信度 α 的等级不可能非常详细。在这种情况下，为了方便，往往给它以一固定的“标准值” F^*。但当自由度 n_1 和 n_2 变化较大时，这样处理是很不恰当的。因为这时 F^* 对应着相差较大的 α 值，即具有多种相差较大的对比水准。实际上，若计算由第 l 周期的 F_l，n_1 和 n_2 所确定的 F—分布的下侧概率值 $1-\alpha$（或上侧概率值，即信度 α），则可得到唯一的对比标准，那个周期的 F—分布下侧概率值 $(1-\alpha)$ 最大(或信度 α 最小)，在概率统计意义上，这个周期就应最显著，就最有可能被选择为显著周期。

因此，从长度为 $l=2$，3，……，$[\frac{N}{2}]$ 这些周期中选出显著周期，可不采用直接从 F_l 中挑选最大值作 F—检验，而可以只从中选出一定数量（例如10个）最

* 其中 $1\leqslant v_0\leqslant l$。

大的 F_l 值及其相应的自由度n_1和n_2值，计算出它们的F—分布下侧概率值，择其最大者进行F—分布置信水平*的检验，作为选择显著周期的依据。

于是，对于给定的正整数$K(\neq 0)$，取

$$F_{mk}=max\ \{F_l\} \tag{10}$$

$$(l=2,3,\cdots\cdots[\tfrac{N}{2}],\ l\neq m(k-1),m(k-2),\cdots\cdots)$$

$$k=1,2,\cdots\cdots,K$$

对于$(k-1)>(k-2)>\cdots\cdots\geqslant 0$成立。

将由上式取得的K个最大的F_l值及它们相应的自由度n_1和n_2按下式用递推法计算各周期的F—分布下侧概率[4] [5]

$$P_{mk}=P(F_{mk},n_1,n_2)=\int_0^{F_{mk}}f(n_1,n_2,F)dF \tag{11}$$

$$(k=1,2,\cdots\cdots K)$$

式中p_{mk}即称为自由度为n_1和n_2时统计量F_{mk}所达到的F—分布的置信水平。

$$f(n_1,n_2 F)=\frac{1}{\beta(\frac{n_1}{2},\frac{n_2}{2})}\ n_1^{\frac{n_1}{2}}\ n_2^{\frac{n_2}{2}}F^{\frac{1}{2}(n_2+n_1F)}(n_2+n_1F)^{-\frac{1}{2}(n_2+n_1F)} \tag{12}$$

称为F—分布密度函数，$\beta(\frac{n_1}{2},\frac{n_2}{2})$称为$\beta$函数。

从p_{mk} 中选取最大的F—分布下侧概率值：

$$Q_{k1}=max\ \{p_{m1},p_{m2}\cdots\cdots p_{mk}\} \tag{13}$$

对于给定的F—分布置信水平Q，若

$$Q_{k1}\geqslant Q \tag{14}$$

则可接受k_1为第一个显著周期。求出第一个显著周期的k_1个振幅值$\Delta X_{w1}(k_1)$，$w_1=1,2,\cdots\cdots,k_1$。并将它们从$\widetilde{X}_t$序列中提取出来，得，

$$\widetilde{X}_{1t}=\widetilde{X}_t-\Delta X_{w1}(k_1) \tag{15}$$

$$(t=1,2\cdots\cdots N;\ w_1=t-[\frac{t-1}{k_1}]\times k_1)$$

---

* 通常将F—分布下侧概率称为F—置信水平，而将其上侧概率称为信度

式中$\left[\frac{t-1}{k_1}\right]$表示取不大于$(t-1)/k_1$ 的最大整数。$\widetilde{X}_{1t}$ 序列称为残余序列。其平方和

$$\widetilde{e}_1=\sum_{t=1}^{N}\widetilde{X}_{1t}^2 \tag{16}$$

与原正规化序列平方和$\widetilde{e}_0$的比值百分数

$$\widetilde{E}_1=\widetilde{e}_1\times 100/\widetilde{e}_0\quad(\%) \tag{17}$$

表示残余误差的相对大小。

对残余序列$\widetilde{X}_{1t}$重复做(7)式以后的工作，并依次算出$\Delta X_{w2}(k_2)$，$\Delta X_{w3}(k_3)$，……和残余误差相对百分数$\widetilde{E}_2$，$\widetilde{E}_3$，……，而第q次残余误差相对百分数可写为

$$\widetilde{E}_q=\widetilde{e}_q\times 100/\widetilde{e}_0(\%) \tag{18}$$

它反映了前面q个周期的总方差贡献的大小，$\widetilde{E}_q$越小贡献越大。当对给定的置信水平Q，出现$Q_{kq}<Q$，或者所需挑选的显著周期个数已经满足（例如$\widetilde{E}_q<10$），则周期的挑选工作结束。假定共选择Z个显著周期。

用上述方法提取的Z个周期，由于随机函数的干扰，它们的振幅值常常是不稳定的。为了得到稳定的周期振幅值，须进行“周期提纯”。

将选出的各种长度的周期的经过“周期提纯”后的振幅进行叠加，便可对原正规化序列进行拟合：

$$\hat{\widetilde{X}}_t=\Delta X_{w1}(k_1)+\Delta X_{w2}(k_2)+\cdots\cdots+\Delta X_{wZ}(k_Z) \tag{19}$$

$$\left(t=1,2,\cdots\cdots N;\ w_q=t-\left[\frac{t-1}{k_q}\right]\times k_q;\ q=1,2,\cdots\cdots Z\right)$$

式中$\left[\frac{t-1}{k_q}\right]$表示取不大于$(t-1)/k_l$ 的最大整数。

然后根据(3)式进行量级变换，把$\hat{\widetilde{X}}_t$序列还原成与原始数据序列相同量级的量：

$$\hat{X}_{ij}=\hat{\widetilde{X}}_t\cdot S_j+\overline{X}_i \tag{20}$$

$$(t=M_2(i-1)+j\ ;\ i=1,2,\cdots\cdots M_1\ ;\ j=1,2,\cdots\cdots M_2$$

及 $j=1,2,\cdots\cdots L(<M_2)$，当 $i=M_1+1$时）

$\hat{X}_{ij}$序列与原序列X_{ij}的差（残差）为

$$e_{ij}=\hat{X}_{ij}-X_{ij} \tag{21}$$

$$(i=1,2,\cdots\cdots M_1\ ;\ j=1,2,\cdots\cdots M_2\ 及$$

$$j=1,2,\cdots\cdots L\,(<M_2)\text{；当}i=M_1+1\text{时）}$$

总残差估计值（估计标准差）为

$$S_{总}=\sqrt{\frac{1}{(N-1)}\left(\sum_{j=1}^{M_2}\sum_{i=1}^{M_1}e_{ij}^2+\sum_{j=1}^{L}e^2_{(M_1+1)j}\right)} \tag{22}$$

显然，把 $\hat{X}_{ij}$ 序列延长下去，当$t>N$时便是试报和预报了。

二、实例：用正规化方差分析作广州地区逐月降水量预报

1．资料：广州1951—1977年逐月降水量等级（表2）

表 2

年份＼降水等级＼月份	1	2	3	4	5	6	7	8	9	10	11	12
1951	7	5	17	18	14	28	13	18	15	10	13	3
……	…	…	…	…	…	…	…	…	…	…	…	…
1977	4	2	5	6	18	22	14	12	20	14	2	5

表1中第i级降水量R（毫米）的界限定义为：

$$(i-1)_{\times}n_0+\frac{(i-1)i}{2}<R\leqslant i_{\times}n_0+\frac{i(i+1)}{2}$$

式中取基数$n_0=4$毫米。因此，第1级是0—5毫米，第二级是5.1—11毫米，第三级是11.1—17毫米，…… 这种降水分级标准是为了对全年降水量进行统一分级而设计的。目的是增加稀少的冬季降水的权重。

2．资料处理：根据27年月降水量等级资料计算得月平均值和月标准差值（表3）

表 3

项目＼月份	1	2	3	4	5	6	7	8	9	10	11	12
平均值	5.1	6.7	8.9	13.9	19.6	20.4	16.9	17.1	14.9	7.9	5.0	4.1
标准差	3.7	3.7	4.0	5.2	6.3	4.6	4.7	5.8	6.8	4.6	4.0	3.0

我们取27年共324个月的降水量等级值参加分析。用各月的历年距平值除以该月标准差值，便得各月的正规化序列值 $\widetilde{X}_{ij}$（$i=1,2,\cdots\cdots27$；$j=1,2,\cdots\cdots12$）。然后将$\widetilde{X}_{ij}$序列按逐年逐月时间顺序排列起来就成为我们所需要的正规化序列：

$$\widetilde{X}_t=\widetilde{X}_{ij}\text{；}\ (t=12(i-1)+j;\quad i=1,2,\cdots\cdots27;\ j=1,2,\cdots12)$$

3．用方差分析对$\widetilde{X}_t$序列作周期分析

给定F—分布置信水平(下侧概率)$Q=0.95$；并限定在这一检验水平下，在序列的残余误差相对百分数$\widetilde{E}_q$的值未下降到2.5%以前最多只挑选8个周期，让计算机自动挑选。根据上述限制，三个条件中只要有一个条件不满足，即停止周期的挑选工作。结果找到8个显著周期，其有关参数如表4（各周期的振幅值从略）。

表　4

周期顺序	1	2	3	4	5	6	7	8
周期長度(月)	121	118	93	53	143	87	43	25
F—置信水平	0.9968	0.9884	0.9957	0.9982	0.9981	0.9967	0.9992	0.99996
残差平方和$\widetilde{e}_i$	169.3	93.1	57.5	42.9	19.2	12.2	9.4	7.7
$\widetilde{e}_i/\widetilde{e}_0$(%)*	52.3	30.1	17.8	13.5	5.9	3.8	2.9	2.4

* $\widetilde{e}_0=N=324$.

从所找到的8个周期中，值得注意的是121(月)和118(月)这两个准十年周期所占的分量最大，约占总方差的70%，它们比较接近太阳黑子活动的平均11年周期。其次是93(月)的近8年周期、53(月)的准四年半周期和143(月)的准十二年周期。这三个周期又约占总方差的24%，87(月)周期接近地极摆动的准7年周期，43(月)周期与地球自转角速度振动的3.5年周期相近，最后是25(月)的准两年周期，但这后面三个周期的分量都很小。

4．拟合预报和误差估计

从上述选8个周期中看出，前7个周期已使序列的残余误差相对百分数下降到3.0%以下，已基本上能概括原正规化序列，因此我们仅取前7个周期进行拟合，

并由计算机按照（20）式自动还原到原始数据量级。由此得出的部分拟合结果和实况对照如图。外延预报结果及与实况对照误差如表5。

由此得到总残差的估计值为$S_{总}=0.85\leqslant 1$。

表 5　预报值（1978年1月至7月）

月 份	1	2	3	4	5	6	7
預报值	3.5	4.5	13.5	18.9	15.4	10.6	10.2
实 况	4	5	12	18	22	19	12
誤 差	−0.5	−0.5	1.5	0.9	−6.6	−8.4	−1.8

5. 结果分析

从所选用7个显著周期(未经周期提纯)对324个月的降水量等级的拟合情况来看，结果是令人满意的，就拟合误差来说，误差超过2级的有10个月次，其中误差超过2.4级的有4个月次，最大误差为—3.1级出现在1974年9月，其它年月的误差值绝大多数出现在(−1.0～1.0)之间，图1曲线清楚地表明这一点。对月雨量超过550毫米(30级以上)的特大降水量的四个月次的拟合也较理想。这四个月次的拟合值分别为1975年5月36.4、1967年8月29.8、1959年6月31.4和1957年5月30.1；它们的实况分别为36、31、33和31，最大误差仅−1.9。从表4看出，对1978年上半年月降水量等级的预报，1—4月份是准确的，7月份是基本准确的，5—6月份是报错的，

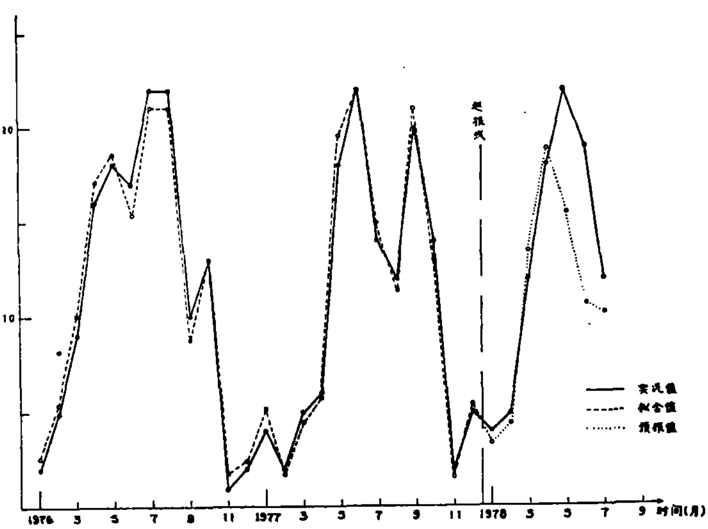

图　部分拟合和预报曲线

三、几点结论

1. 正规化方差分析是一种能够消除序列的年气候变化趋势，并抓住那些非一年整倍数周期变化的方法。它可适用于降水，温度、湿度、气压等多种气象要素的月（季、旬、候）的长期预报。

2. 提出采用F一分布下侧概率值作为具有唯一标准的周期显著性检验参数，它比人们惯用的取某一固定值的F_l参数严格和明确。但F—分布下侧概率不易计算，故这一参数只在电子计算机计算时适宜。

3. 周期个数的确定可取以当序列的残差平方和相对百分数$\widetilde{E}_q$下降到某个百分数(例如10%)以下为原则。当$\widetilde{E}_q$小于这个百分数时即停止周期的挑选工作。实践证明，通常需6—8个周期。但不是所有序列都能找到这么多的周期。

参考文献

〔1〕么枕生，气候統計，科学出版社，1964.

〔2〕李麦村，現代統計預报的进展，現代气象学若干問題的进展，科学出版社，1975.

〔3〕王宗皓，李麦村等，天气預报中的概率統計方法，科学出版社，1974.

〔4〕Peizer, D. B. and Pratt, J. W., A normal approximation for binomial, F, beta, and other common, related tail Probabilities, *J, Amer, Statist, Assoc,* 63 (1969), 1416——1482.

〔5〕中国科学院計算中心概率統計组，概率統計計算，科学出版社，1978.

Normalized Variance Analysis and its Application in Weather Prediction

Ghen Chuangmai

Abstract

This paper deals With a method of variance analysis for meteorological time series. In order to remove climatic yearly frend, the series under examination are normalized beforehand for each season, month, decade or pented. Significance tests are introduced for the periods thus determined, using the left-side probability of F-distribution as the parameter. Electronic computer is employed to extract automatically various components of such periods as are statistically meaningful, especially those periods of lengths not of integral multiples of one year. On the basis of such analysis the future frend of the series is predicted. Actual calculation shows that the result is fairly satisfactory.