第1卷第1期　　气象科学研究院院刊　　Vol. 1, No. 1
JOURNAL OF ACADEMY OF METEOROLOGICAL
1986年 6 月　　SCIENCE, S. M. A., CHINA　　Jun., 1986

带有周期分量的多元逐步回归*

魏凤英　赵　溱　张先恭

（气象科学研究院天气气候研究所）

一、引　　言

从气候资料中提取隐含周期是时间序列分析的目的之一。目前常用的方法有:谐波分析、周期图、谱分析和最大熵谱分析。谐波分析和谱分析在处理短序列时，分辨率很低。最大熵谱的优点是分辨率高，它对于揭露气候资料的周期是一个好方法，但很难用于预报。

根据方差分析原理有人提出了另一个提取周期的方法[1]。假设气候资料 x（t）是由若干周期函数和一个白噪声迭加而成的序列:

$$x(t)=\sum_{i=1}^{M} f_i(t)+E(t) \tag{1}$$

这里f_i（t）是周期为l_i的周期函数，$i=1,2,\cdots$，M，M为周期函数的个数，$E(t)$是白噪声。

设已知时间序列观测值为$x(1)$，$x(2)$，$\cdots$，$x(n)$，$n\geqslant 2M$，据此，要求出l_i及估计值$\hat{f}_i(t)$。这一方法在长期预报中得到普遍使用，称为“方差分析周期外推法”[2]。遗憾的是，当$x(t)$由两个以上周期函数迭加而成时，再用这个方法无疑是错误的。

基于这种情况，我们提出用多元逐步回归的方法来挑选气候资料的隐含周期。为了将此方法更好地应用到长期预报中去，我们还将模型作了改进。实例表明，这个方法不但为寻找时间序列的周期提供了方法，而且为长期预报提供了新的工具。

二、思　　路

设　　$x=x_1,\ x_2,\ \cdots,\ x_n$

为时间序列。

按试验周期的方法，将资料先以周期长度为 2 分组:

x_1　x_2

x_3　x_4

* 本文于1985年5月20日收到，1985年8月15日收到修改稿。

……

x_{n-1}　x_n

求出每列的平均值 $\overline{x}_1^{(2)}$，$\overline{x}_2^{(2)}$。

按周期长度为 3 分组：

x_1　x_2　x_3

x_4　x_5　x_6

……

x_{n-1}　x_n

求出每列的平均值 $\overline{x}_1^{(3)}$，$\overline{x}_2^{(3)}$，$\overline{x}_3^{(3)}$。

反复进行下去，直到分组的个数不大于资料的一半$[\frac{n}{2}]$。($[\frac{n}{2}]$表示$\frac{n}{2}$的整数部分，当 n 为偶数时，$[\frac{n}{2}]=\frac{n}{2}$，当 n 为奇数时，$[\frac{n}{2}]=\frac{n-1}{2}$)。

x_1　x_2　…… $x_{[n/2]}$

$x_{[n/2]+1}$　$x_{[n/2]+2}$ …… x_n

求出每列的平均值 $\overline{x}_1^{([n/2])}$，$\overline{x}_2^{([n/2])}$……，$\overline{x}_n^{([n/2])}$。

如果写成一般的表达式：

设

$$x=(x_1,\ x_2,\ \cdots,\ x_n)'$$

为时间序列，“′”表示转置。

构成周期序列：

$$X_1,\ X_2,\ \cdots\cdots,\ X_M \qquad 1\leqslant M\leqslant[\frac{n}{2}]-1$$

$$X_1=(x_1^1,\ x_2^1,\ \cdots,\ x_n^1)' \tag{2}$$

这里

$$x_1^1=x_3^1=x_5^1=\cdots=x_{2[n+1/2]-1}^1=\Big[\frac{1}{\frac{n+1}{2}}\Big]\sum_{i=1}^{[\frac{n+1}{2}]}x_{2i-1}$$

$$x_2^1=x_4^1=x_6^1=\cdots=x_{2[n/2]}^1$$

$$X_2=(X_1^2,X_2^2,\cdots X_n^2)' \tag{3}$$

这里

$$x_1^2=x_4^2=x_7^2=\cdots x_{3[n+2/3]-2}^2=\frac{1}{\Big[\frac{n+2}{3}\Big]}\sum_{i=1}^{[\frac{n+2}{3}]}x_{3i-2}$$

$$x_2^2 = x_5^2 = x_8^2 = \cdots x_{3[n+1/3]-1}^2 = \frac{1}{\left[\frac{n+1}{3}\right]} \sum_{i=1}^{\left[\frac{n+1}{3}\right]} x_{3i-1}$$

$$x_3^2 = x_6^2 = x_9^2 = \cdots x_{3[n/3]}^2 = \frac{1}{\left[\frac{n}{3}\right]} \sum_{i=1}^{\left[\frac{n}{3}\right]} x_{3i}$$

……

$$X_k = (x_1^k, x_2^k, \cdots, x_n^k)' \tag{4}$$

这里

$$x_1^k = x_{k+2}^k = x_{2k+3}^k = \cdots = x_{(k+1)[n+k/k+1]-k}^k = \frac{1}{\left[\frac{n+k}{k+1}\right]} \sum_{i=1}^{\left[\frac{n+k}{k+1}\right]} x_{(k+1)i-k}$$

$$x_2^k = x_{k+3}^k = x_{2k+4}^k = \cdots = x_{(k+1)\left[\frac{n+k-1}{k+1}\right]-(k-1)}^k = \frac{1}{\left[\frac{n+k-1}{k+1}\right]} \sum_{i=1}^{\left[\frac{n+k-1}{k+1}\right]} x_{(k+1)i-(k+1)}$$

……

直到X_M。这里X'为X的转置。其中[ ]表示不超过方括号内数的最大整数。

为了寻找$x(t)$的隐含周期l，按照方差分析作统计检验：

$$F^l = \frac{S^l/(l-1)}{S/(n-l)} \tag{5}$$

这里

$$S^l = \sum_{i=1}^{l} N_i (x_i^l - \bar{x})^2$$

$$S = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$N_i = \left[\frac{n+l-i+1}{l+1}\right]$$

当$F^{l}>F_{\alpha}$（α是给定的显著性水平），则认为$x(t)$存在值为$x_1^{l}, x_2^{l}, \cdots, x_n^{l}$的隐含周期分量。通常$F^{l_1}=\max\limits_{1\leqslant l\leqslant m}\{F^{l}\}$是第一个被考虑的周期。一旦$l_1$被确定，则剩余序列为：

$$x_1-x_1^{l_1}, x_2-x_2^{l_1}, \cdots, x_n-x_n^{l_1} \tag{6}$$

然后对剩余序列重复上述步骤，在这个过程中发现第二、第三、…，第k个周期l_2、l_3、…、l_k。

假设寻找到周期k，

$$\tilde{x}(t)=\sum_{i=1}^{k}x^{l_i}(t) \tag{7}$$

是$x(t)$的一个近似。

$$\tilde{x}_{n+1}=\sum_{i=1}^{k}x_{n+1}^{l_i} \tag{8}$$

是$x(t)$的一个外推。

在最简单的情况下，$k=1$即

$$x(t)=f(t)+E(t) \tag{9}$$

这里$f(t)$是一个长度为l的周期函数，$E(t)$是白噪声。

相应的序列是：

$$\begin{aligned}&f_1+E_1, f_2+E_2, \cdots, f_l+E_l\\&f_1+E_{l+1}, f_2+E_{l+2}, \cdots, f_l+E_{2l}\\&f_1+E_{2l+1}, f_2+E_{2l+2}, \cdots, f_l+E_{3l}\end{aligned} \tag{10}$$

显然，第一列$f_1+E_1, f_1+E_{l+1}, f_1+E_{2l+1}\cdots$是服从正态分布$N(f_1, \sigma)$，第二列$f_2+E_2$，$f_2+E_{l+2}$，$f_2+E_{2l+2}\cdots$是服从正态分布$N(f_2, \sigma)$[3]。于是，使用方差分析作出气象序列$x_1, x_2, \cdots, x_n$存在一个隐含周期$l$的统计推断是非常巧妙的。当(1)中$M\geqslant 2$时，它就是错误的。为方便起见，

设

$$x(t)=f_1(t)+f_2(t)+E(t) \tag{11}$$

这里f_1，f_2是周期为L，L^*的周期函数。

现在序列$x_1, x_2, \cdots, x_n$分成组：

$$\begin{aligned}&f_1(1)+f_2(1)+E(1), f_1(2)+f_2(2)+E(2), \cdots, f_1(l)+f_2(l)+E(l)\\&f_1(1)+f_2(l+1)+E(l+1), f_1(2)+f_2(l+2)+E(l+2), \cdots, f_1(l)+f_2(2l)+E(2l)\\&\cdots\cdots\end{aligned} \tag{12}$$

因为$L\neq L^*$，因此，$f_1(1)+f_2(1)\neq f_1(1)+f_2(l+1)$，那么，第一列$f_1(1)+f_2(1)+E(1)$，$f_1(1)+f_2(l+1)+E(l+1)\cdots$不来自同一母体。在这种情况下，再使用方差分析未免太牵强了。

基于上述情况，采用逐步筛选回归技术，将周期迭加模型(1)加以推广，建立多元回归方程。

$$X=a_0+\sum_{i=1}^{M}a_if_i+E \tag{13}$$

这里a_0为常系数，a_i为回归系数，f_i为周期序列(2)、(3)、(4)，E为白噪声。

自然采用逐步筛选回归技术估计a_i，a_0[4]。当估计值$\hat{a}_i$不为0时，认为气候序列X含有周期i的周期分量。我们称这个方法为“逐步回归周期分析”[5]。

为了将这个方法能够更好地应用到长期预报中去，我们将物理因子加入模型（13），即假设

$$X=a_0+\sum_{i=1}^{M}a_if_i+\sum_{j=1}^{k}b_jg_j+E \tag{14}$$

其中f_i为预报量X的周期分量，g_j为预报因子。当然，系数a_0，a_i，b_j仍用回归技术来估计。这样，不但顾及了预报量自身的周期变化对回归模型的影响，也考虑了其它物理因子的影响，使预报方程有了较明确的物理意义。

如果再将预报因子的周期分量也考虑进去，即将预报量的周期分量及各预报因子的周期分量同时作为附加因子加入回归模型。建立以X为预报量的模型。

$$X=a_0+\sum_{l=1}^{L}\sum_{i=0}^{M}a_i^lf_i^l+\sum_{j=1}^{k}b_jg_j+E \tag{15}$$

其中$f_0^1,f_0^2,\cdots,f_0^L\cdots f_M^1,f_M^2,\cdots,f_M^L$为预报量$X$和预报因子$g_1$，$g_2$，…，$g_k$的周期分量。$E$为白噪声。

采用通常的多元筛选技术，得到系数a_0，a_i^l，b_j的估计值。

通常的多元逐步回归模型，由于预报量和预报因子的相关关系呈现出一定的阶段性，某一段时期内的高相关在另一段时期内会消失，有时甚至会改变符号。因此，在各个不同的相关阶段的转折时期，预报往往失败。我们设想，气象要素的时间趋势主要是由周期性变化构成的，在模型（15）中，除了预报因子外，将预报量和各预报因子的周期分量作为附加因子加入回归模型。这样，由于预报量及预报因子的时间趋势而引起的相关不稳定性期望由选入的周期性因子加以调整。

三、计 算 步 骤

上述模型（13）、（14）和（15）可以根据不同需要使用。若分析气候资料的隐含周期，则用模型（13）简便一些；作长期预报，使用模型（14）或（15）更客观一些。下面简述一下模型（15）的计算步骤，模型（13）和（14）的计算步骤就包括其中了。

设预报量X有n个观测值，k个预报因子$g(t)$有n个观测值。

1．按照(2)、（3）和（4）式计算预报量和预报因子周期长度为2，3，…，M的周期序列。这里$M=\left[\frac{n}{2}\right]$。

2．把预报量和预报因子的周期函数，作为附加因子连同预报因子一起用通常的逐步回归方法一步步筛选，直到既没有再被选入也没有要剔除的因子为止。逐步回归的具体实施步骤参见文献［4］。

3．按（15）式建立预报方程。若筛选出预报量或某个预报因子的第 i 个周期因子，就认为序列存在 i 年显著周期。这些周期因子和所筛选出的预报因子一起进入预报方程。

四、实　　例

1．用模型（13）分析了1951—1980年北京1月平均温度资料，找出12，13，9年三个显著周期。用分辨率较高的最大熵谱方法分析了同样的资料，也得到这三个显著周期，模型(13)能够提取隐含周期得到证实。我们将计算结果与方差分析作一比较（见表1），虽然方差分析也可以得到大多数相同的结果，但显著性水平却很低。模型（13）不仅可以分析出时间序列的隐含周期，而且拟合、预报效果也相对好些（见图1）。

表1　方差分析周期外推法与模型(13)计算结果比较

周期顺序	分析方法	周期长度	F检验 $F^{(1)}$	$F_{0.05}$
1	方差分析周期外推法	12	1.86	2.67
	模型（13）	12	30.65	
2	方差分析周期外推法	13	2.09	2.60
	模型（13）	13	22.39	
3	方差分析周期外推法	8	1.87	3.43
	模型（13）	9	11.41	3.13

2．用模型（15）对上海5—8月总降水量作了拟合预报试验。通过对1、2月份500百帕各网格点高度作相关普查，得到参加筛选的预报因子有：30°N、10°E和45—50°N、100—60°W两地区500百帕1月高度距平之平均和50—55°N、110—120°E地区500百帕2月高度距平之平均这三个预报因子；另外附加因子有：预报量即上海5—8月总降水量1954—1981年

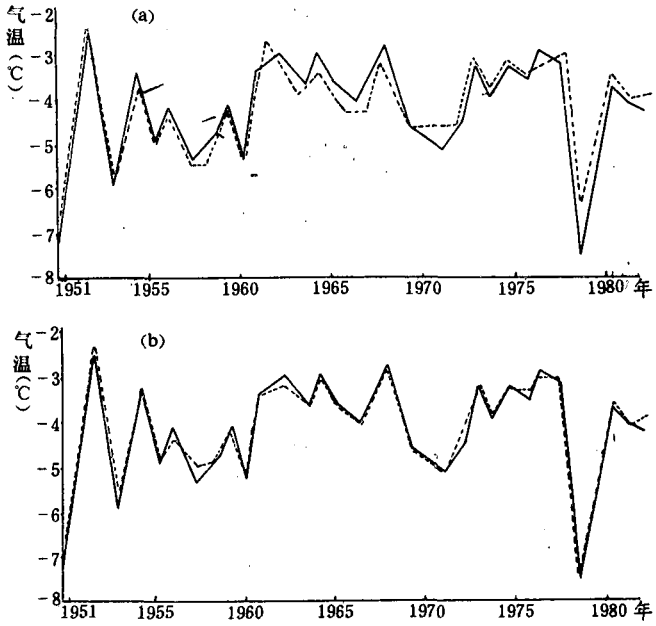

图1　(a)北京1月份气温实况曲线(实线)与用方差分析周期外推法的拟合曲线（虚线）；
(b) 北京1月份气温实况曲线（实线）与用模型（13）的拟合曲线（虚线）
1981、1982年的为预报曲线

28年的2,3,……,14年的13个试验周期,以及上述三个预报因子的13个试验周期函数，共计55个因子。

取$F=4.0$,回归方程共选取了四个因子，即上海5－8月降水的14年周期和13年周期，北非副热带地区高度以及高度变化的14年周期（见表2）。有意思的是除北非副热带地区高度的遥相关外，在所选取的因子中，周期分量占很大比重。无论预报量或预报因子均有14年周期，虽然目前还没有找到14年周期的物理根据，但在分析长江汉口站流量、宜昌水位和西太平洋台风年总数时[6]，均发现有14年周期。

表2 回归方程选取的因子

序号	选入的因子
1	预报量的14年周期
2	预报量的13年周期
3	第1个预报因子（30°N、10°E 500百帕高度距平之平均）
4	第1个预报因子的14年周期

表2的结果说明，影响上海5－8月总降水量的主要因素是本身的气候变化，而其它物理因子的影响只占一定份量。而本模型比较充分地揭露了这一特征。因此预报效果比通常使用的逐步回归分析要好一些。为便于比较，将同样的资料用通常的逐步回归模型进行了计算（见图2）。图中虚线为1982年的预报结果,用通常的逐步回归模型预报为541毫米,用模型(15)预报为611毫米,实况为601毫米。

通常的逐步回归方法只考虑物理因子对预报量的影响，这样对周期性很强的预报量就必然要影响预报效果。这一例正说明了这一点。

3．由于模型（15）随着资料长度的增长,附加因子数迅速增加。这样，需要的计算时间增多，而且有些计算机的容量不够。因此，模型（14）更便于业务预报。我们用模型（14）试作了1984年长江中下游地区、华北地区6－8月逐月降水预报。对照1984年中央气象台《气象月报》提供的实况，对预报进行了检查（见表3）。从表3中距平百分率的符号可以看出，降水比常年偏多或偏少的趋势，大多数能预报出来。

表3 6—8月逐月降水预报距平百分率

距平百分率 / 月份	华北地区		长江中下游地区	
6	预报	140	预报	26
	实况	8	实况	12
7	预报	－9	预报	－100
	实况	－16	实况	－13
8	预报	0	预报	59
	实况	20	实况	2

五、结束语

根据多元回归分析我们提出这一方法，这是用多元分析解决时间序列问题的初步尝试。它为提取时间序列隐含周期提供了一个比较简便的方法，为更好地将逐步回归技术应用到长期预报中去提供了新的途径。但是，它的理论根据还不充分，也具有一定的局限性。从时间序列中提取隐含周期，解决长期预报这一复杂问题，仍有待进一步研究和探索。

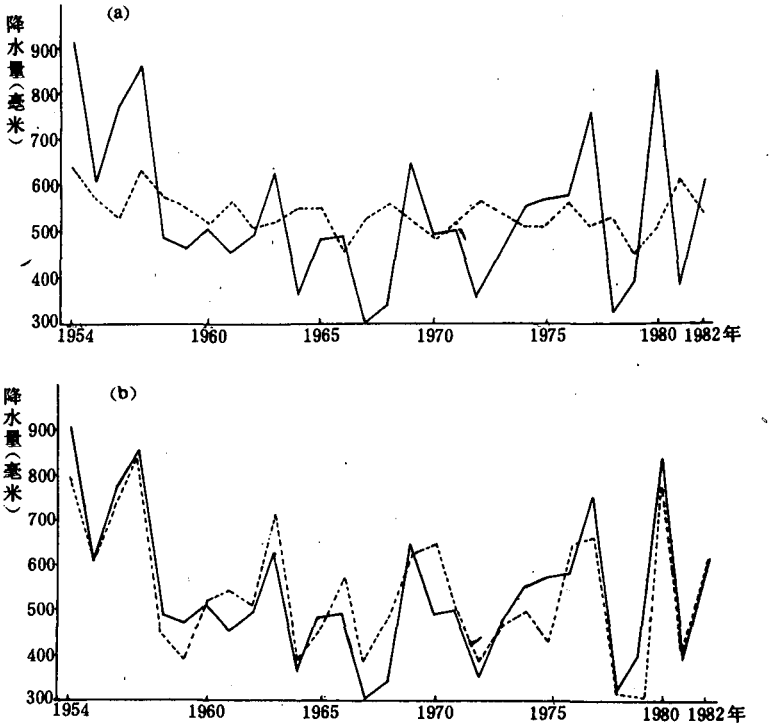

图 2　(a) 上海 5 — 8 月降水量实况曲线（实线）与通常的逐步回归分析的拟合曲线（虚线）；
(b) 上海 5 — 8 月降水量实况曲线（实线）与模型（15）的拟合曲线（虚线）
1982年的为预报曲线

参 考 文 献

[1] Brooks, C. E. P., Carruthers N., Handbook of Statistical Methods in Meteorology, London, Her Majesty's stationary office, 1953.
[2] 章少卿、丁士晟等，一种简化了的时间序列预报方法的讨论，气候学术会议文件，1964年。
[3] 中国科学院数学研究所统计组编，方差分析，科学出版社，1977年。
[4] 中国科学院数学研究所数理统计组编，回归分析方法，科学出版社，1975年。
[5] 魏凤英、赵溱、张先恭，逐步回归周期分析，气象，第 2 期，1983年。
[6] 张先恭等，长江汉口站流量水位分析及其与太阳活动关系的初步探讨，长江流域长期水文气象预报讨论会文集，100—105，1975年。