研究论文

逐步回归周期分析的改进方案及其在气候预测中的应用

魏凤英　张先恭　曹鸿兴

（气象科学研究院）

提　要

考虑均值生成函数随机性的强弱，本文引进了筛选周期因子的新标准，提出了逐步回归周期分析的改进方案。运用东太平洋海温、长江流域和华南地区汛期降水量、太阳黑子和年轮指数序列进行实例计算，结果表明，本方案不但有较高的长期预测效果，且有一定的分析时间序列隐含周期的能力。

一、问题的提出

1983年魏凤英等曾提出利用逐步回归技术，筛选时间序列生成周期因子的计算方案，即逐步回归周期分析[1]。用这一方案建立的模型可以作较长时间的预报，且有较高的拟合精度和较好的预报效果。几年来，一些台站将它作为长期预报的常规方法之一[2-3]。但是，在应用过程中我们发现用这一方法所提取的主周期常常是长周期。这不但会造成所选取的周期有时在物理意义上得不到合理的解释，且长周期的周期序列随机性较大，由此所建预报方法的稳定性也会较差。

设一时间序列

$$x(t)=\{x(1),\ x(2),\ \cdots x(N)\}$$

式中N为样本大小。把序列x(t)按一定的时间间隔计算平均值，即

$$\bar{x}_l(i)=\frac{1}{n_l}\sum_{j=0}^{n_l-1}x(i+jl)$$

$$i=1,\cdots,l \quad 2\leqslant l\leqslant M \qquad (1)$$

式中n_l为满足$n_l\leqslant[\frac{N}{l}]$的最大整数，$M=[\frac{N}{2}]$为不超过$\frac{N}{2}$的最大整数。我们称$\bar{x}_l(i)$为均值生成函数(简称均生函数)[4]。

对均生函数$\bar{x}_l(i)$作周期性延拓，即令

$$f_l(t)=\bar{x}_l(i)\quad t=i(\mathrm{mod}(l))$$

$$t=1,2,\cdots,N \qquad (2)$$

这里mod表示同余。由此构造出M－1个外延的均生函数，记为f_i，i＝1,2,…,M－1。

我们以x(t)生成的外延均生函数f_i作为预报因子，建立关于x(t)的模型，即设

$$x(t)=a_0+\sum_{i=1}^{M-1}a_if_i(t)+e(t) \qquad (3)$$

这里a_0,a_i为待定系数，f_i为延拓均生函数，e为白噪声。

采用逐步回归技术估计a_0，a_i，从而挑选序列的主要周期。由筛选出的周期所对应的均生函数建立预报方程。假设筛选出K个周期，则

$$\hat{x}(t)=a_0+\sum_{i=1}^{K}a_if_i(t)$$

$$(i=1,2,\cdots,K\quad t=1,2,\cdots,N) \qquad (4)$$

若作q步预报，将入选的均生函数外延q步，即

$$\hat{x}(N+q)=a_0+\sum_{i=1}^{K}a_i f_i(N+q)$$

$$(q=1,2,\cdots) \qquad (5)$$

以上即为逐步回归周期分析的基本步骤。

逐步回归周期分析（称原方案）是用均生函数 $f_i(t)$ 与原序列 $x(t)$ 间的相关系数计算方差贡献，并依次选取方差贡献的最大值来确定入选周期的。从（1）式可知，长周期的均生函数是由2—3个数据相加求平均得到的。通常情况下，它与原序列的相关系数可能要比短周期即由多个数据平均求得的均生函数要高，方差贡献亦大。因此，被选取的机会相对比短期要多些。

基于上述分析，我们这里提出对这一方法的改进方案即引入新的筛选标准，并与原方案进行了比较分析。

二、改进方案

设长度为l的均生函数的方差贡献为 U_l，在 U_l 上添加关于周期长度的“惩罚”系数，即令

$$V_l=\alpha_l U_l \quad \alpha_l=\frac{N}{l}$$

$$(l=2,3,\cdots,M) \qquad (6)$$

N为样本大小。长度为l的均生函数是 $\left[\frac{N}{l}\right]$ 个数据的平均。也就是l越大，$\frac{N}{l}$ 越小，当 $l=\frac{N}{2}$ 时，只有两个数据求平均，相应的均生函数随机性亦大。而我们期望用随机性较小稳定性较大的均生函数建立方程。当l较小时，α_l 较大即对方差贡献施加较大权重。随着l不断增大，α_l 逐渐变小，也就是给予长周期适当的惩罚，以期筛选出隐含于序列中的周期，避免总是长周期入选。

三、改进方案与原方案的比较分析

以1951—1985年赤道东太平洋（0—10°S，180°—90°W）秋季（9—11月）海表温度为例，分析在方差贡献上施加对周期长度惩罚后的效果。这里N=35，$M=\left[\frac{35}{2}\right]=17$。表1给出了原方案、改进方案及功率谱分析提取的主要周期。

表1

方法＼序号（周期长度）	1	2	3	4
原方案	15	14	13	12
改进方案	7	6	4	5
功率谱	7	5	4	

从表中可以看出，原方案选取的主要周期均为较长的周期，而改进方案选取的主要周期与功率谱分析的大致相同，尤其第一周期都是长度为7年的周期。大量的研究工作证明，赤道东太平洋地区的海温确实存在着3—7年的周期变化。可见，改进方案提取的周期比原方案的结果更可信。

四、在气候预测中的应用实例

1．长江流域和华南地区汛期降水预报

取1951—1987年长江流域17个站（南京、合肥、上海、杭州、安庆、屯溪、九江、汉口钟祥、岳阳、宜昌、常德、宁波、衢县、贵溪、南昌、长沙）和华南地区15个站（厦门、梅县、汕头、曲江、河源、广州、阳江、湛江、海口、桂林、柳州、梧州、南宁、北海、百色）平均6—8月降水量。其中N=37，M=18。

长江流域的预报方程为：

$$\begin{aligned}\hat{x}(t)=&-1187.4600+0.9426f_3(t)\\&+0.6561f_5(t)+0.5062f_7(t)\\&+0.6075f_{11}(t)+0.7229_{13}(t)\end{aligned} \qquad (7)$$

即为由长度3年、5年、7年、11年和13年的均生函数构成的回归模型。方程的拟合均方根误差RMSE=53.0102。

华南地区的预报方程为：

$$\hat{x}(t) = -1698.6410 + 0.9951f_2(t) + 0.6376f_7(t) + 0.6351f_8(t) + 0.5101f_9(t) + 0.6980f_{13}(t) \quad (8)$$

方程的拟合均方根误差RMSE=42.0963。

图1分别给出长江流域（a）和华南地区（b）的方程拟合曲线和实况曲线。我们知道，降水量序列是起伏较大难以拟合的序列，但从我们所建方程的均方根误差和图1看出，用这种方案建模的拟合效果是令人满意的。尤其像1954年、1969年、1980年和1983年的长江大水、华南地区1959年、1966年、1968年和1973年的多雨，都可以很好地反映出来。对极值能有如此之好的拟合是其它统计方法很难办到的。

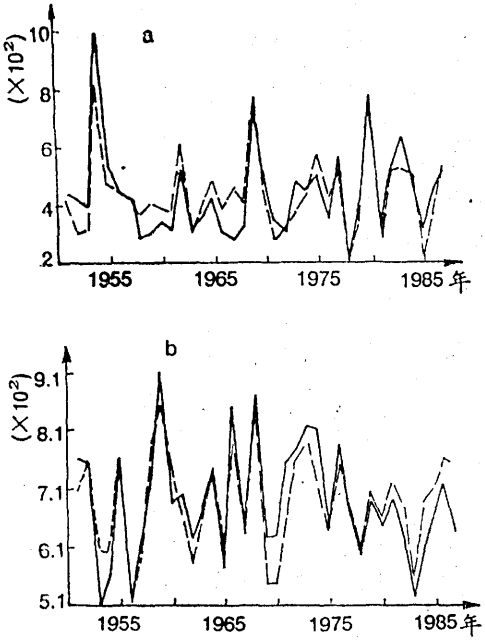

图1 汛期降水量变化曲线

实线为实况值，虚线为计算值；

(a) 为长江流域，(b) 为华南地区

1988年的预报结果与实际值列在表2。从表中可以看出，这两个地区的预报值与实况数值分别相差38mm和27mm。用距平的正、负符号来表示趋势，预报的距平符号与实况的符号是一致的。

表2 1988年预报值和实况值（mm）

项目 / 地区	预报值	实况值	多年平均值
长江流域	442	480	489
华南	642	615	671

2．太阳黑子周期分析及其预报

众所周知，太阳黑子最显著的周期是11年。为了进一步考查改进后方案提取隐含周期及作长期预测的能力，作为例子我们截取1936—1979年太阳黑子数进行分析，建立了预测模型，并对1980—1986年进行了试报。图2中实线是1936—1986年年平均太阳黑子数变化曲线。我们清楚地看出，太阳黑子明显地遵循着11年左右的周期循环。用改进方案筛选的主要周期、复相关系数和均方根误差列于表3。令人鼓舞的是，提取的第一显著周期正是众所公认的11年周期。值得注意的是，入选的第五个周期是22年。22年周期是太阳黑子磁周期的一种反映。而这一周期用通常的统计手段往往难以揭露出来。从这一例的分析结果足见改进方案具有一定的分析隐含周期的能力。

表3

序号	1	2	3	4	5
周期长度	11	9	8	5	22
复相关系数	0.88	0.93	0.95	0.95	0.97
均方根误差	18.01	14.06	13.25	12.42	9.72

从图2可以看出，模型的拟合值与观测值是相当一致的。拟合相对误差为$\varepsilon = \frac{9.72}{73.98} = 1.22\%$。1980—1986年预报与实况曲线比较，趋势基本是正确的。若以多年年平均值

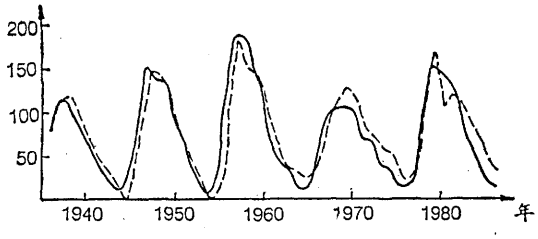

图 2　1936—1986年年平均太阳黑子数

实线为观测值，虚线为拟合值，其中1980—1986年为试报值。

73.98作为衡量趋势预报的标准，7年中除1983年报错外，其余6年均正确。事实上，1983年预报与观测数值很接近，仅差17.5。

3．用年轮指数作气候预测的试验

为了预测未来我国气候的变化趋势，我们取1043—1977年祁连山圆柏的最后年表作为代用资料[5-6]，用上述方法进行了试算。该年表共有935年，即N = 935，取M = $\left[\frac{935}{2}\right]$ = 467。

表4为计算中的某些统计结果。从表中可看出，用10个周期拟合，大约可解释原序列80%左右的方差。从复相关系数的增长率看，238，188，74，26及141年等周期的贡献较大，这与文献〔5〕用方差分析所得结果是一致的。

表4

周　期	329	26	23	188	2
复相关系数	0.731	0.748	0.757	0.806	0.807
均方根误差	0.265	0.261	0.257	0.231	0.230
周　期	10	74	238	47	141
复相关系数	0.809	0.826	0.875	0.879	0.895
均方根误差	0.229	0.222	0.189	0.186	0.175

图3为近千年来每10年祁连山圆柏年轮指数的实际值和计算值。由图可见，两条曲线基本上是一致的。下图中1980—2100年为预测值。由于祁连山圆柏采自我国西部地区，而西部高原地区的气候变化往往比东部地区提前10—30年[7]。因此，我们在用这个

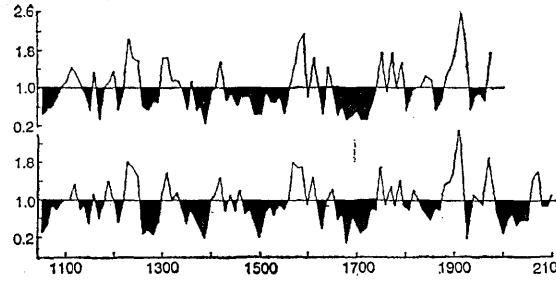

图 3　近千年来每10年祁连山圆柏年轮指数的实际值(上)和计算值(下)

代用资料预测东部大范围气候变化时，将时间坐标后延20年来考虑。根据外推曲线分析，从1970年代开始的增温将持续到21世纪初，从2010年代到2070年代将再次处于一个偏冷的时期，2080年代后气温将再次回升。当然，这仅仅是根据气候自身的变化规律用上述方法外推出来的结果，没有考虑二氧化碳等人为因素的影响。

五、结语

采用新筛选标准的逐步回归周期分析在上述实例计算中显示了较高的长期预测效果，在分析时间序列隐含周期方面也显示了一定的能力。因此，这一方法可以作为数据分析及气候预测的工具。当然，它的分析和预测效果还有待更多应用实例的进一步检验。

参考文献

〔1〕魏凤英、赵凑、张先恭，逐步回归周期分析，气象，1983，2期。

〔2〕李邦宪，因子筛选与周期分析相结合的逐步回归双重分析预报模型，气象，1988，6期。

〔3〕王春乙、潘亚茹，我国北方主要产麦区冬小麦产量海温业务预报模式，数学的实践与认识，1989，1期。

〔4〕曹鸿兴、魏凤英，基于均值生成函数的时间序列分析，数值计算与计算机应用，待发表。

〔5〕张先恭等，祁连山圆柏年轮与我国气候变化趋势，全国气候变化学术讨论会文集，科学出版社，1981。

〔6〕刘光远等，祁连山圆柏的最后年表，气象，1984，11期。

〔7〕汤懋苍等，青藏高原及其四周的近代气候变化，高原气象，7卷1期，1988。

An improved scheme for the period analysis using stepwise regression and its application to climatic prediction

Wei Fengying Zhang Xiangong Cao Hongxing

(Academy of Metcorological Science)

Abstract

Considering the randomness of homogeneous out-growth function, the new criterion for screening factors is used in the period analysis and an improved scheme of stepwise regression is brought out. From the case computations, such as the sea surface temperature in the equatorial Eastern Pacific area, the precipitation over the regions along Changjiang River and in southern China during the rainy season, the sunspot and the tree ring index, it is found that the scheme is not only suitable for long-term prediction but also capable of detecting the potential periods in a time series.