均生函数—最优子集回归在高温极值预测中的应用

张德宽　杨贤为　邹旭恺

（国家气候中心，北京　100081）

提　要

根据南京等城市1961年以来历年气温极值资料序列，采用均生函数—最优子集回归法设计的短期气候预测模型不但能较好地拟合历史实况，而且对未来1～5年的演变趋势也具有一定的预报能力。

关键词： 均生函数　最优子集回归　高温极值

引　言

随着全球气候变暖和我国城市化进程的加快，近年来我国许多城市夏季气温居高不下，年极端最高气温不断向上攀升，城区和郊区的温差也逐渐扩大。例如，上海城区龙华站和郊区崇明站7、8月份的平均气温在上世纪60年代仅差0.1℃，到90年代扩大到1.0℃[1]；北京夏季城区和郊区的温差普遍在1.2℃以上[2]。近年来我国高温时空分布的显著特点是高温天气开始早，范围广，历时久。以2001年为例，京、津、冀等地在5月中下旬就出现了3～7天日最高气温为35～41℃的酷热天气，6、7月份高温范围扩及东北至长江流域的广大地区，嫩江、齐齐哈尔出现了40℃以上的极端高温[3]。

鉴于高温极值一定程度上反映了某地夏季的炎热程度[4]，本研究采用均生函数—最优子集回归法对我国若干城市的高温极值序列进行拟合与预测，结果比较令人满意。

1　资料选取

本研究以长江沿岸南京、武汉、宜昌、重庆等城市1961～1996年历年出现的高温极值为计算样本，来构造预测模型，然后根据这些预测模型分别获取上述城市1997～2001年高温极值的预测值。由于以下的计算思路及方法步骤将以南京为实例详细展开，故表1只列出南京站历年高温极值，其余各站不一一罗列。

表1　南京1961～1996年历年高温极值

年份	高温极值(℃)
1961～1970	38.3 35.9 36.2 37.5 37.0 40.5 39.0 36.6 36.6 36.6
1971～1980	38.1 36.7 35.7 36.0 36.0 37.6 36.7 39.7 36.5 36.0
1981～1990	37.9 34.4 38.2 36.4 36.8 36.7 35.2 38.5 36.4 36.8
1991～1996	37.0 37.9 35.7 38.0 37.2 36.5

2　均生函数的构建和统计

2.1　定义

设样本量为 n 的一个时间序列

$$x(t) = x(1), x(2), \cdots, x(n) \quad (1)$$

$x(t)$的平均值为

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x(i) \quad (2)$$

对于式(1)定义均生函数[5]

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

$$(i = 1, \cdots, l)(l = ,1,2, \cdots, m) \quad (3)$$

上式中 $n_l = \mathrm{INT}(n/l)$，$m = \mathrm{INT}(n/2)$ 或 $\mathrm{INT}(n/3)$，INT表示取整数，据式(3)可得 m 个均生函数，作周期性延拓可得：

$$f_l(t) = \bar{x}_l(i) \quad (4)$$

式中，$t = i[\mathrm{mod}(l)]$　$t(1,2,\cdots,n)$

这里mod表示同余，据式(4)可构造出周期为 l 的 m 列周期函数。

2.2 实例计算

现以南京1961～1996年历年高温极值时间序列为样本(见表1)来进行计算。这里 $n=36$，m 取 $\mathrm{INT}(n/3)=12$，利用式(3)可获得一个上三角矩阵 H：

$$H=\begin{Bmatrix}370 & 369 & 372 & 368 & 366 & 372 & 380 & 368 & 379 & 368 & 364 & 369\\ & 371 & 369 & 372 & 378 & 365 & 365 & 374 & 368 & 378 & 367 & 362\\ & & 370 & 370 & 365 & 363 & 371 & 366 & 370 & 370 & 366 & 358\\ & & & 371 & 372 & 371 & 377 & 370 & 375 & 383 & 386 & 379\\ & & & & 368 & 373 & 364 & 368 & 366 & 365 & 384 & 367\\ & & & & & 378 & 370 & 369 & 369 & 365 & 366 & 390\\ & & & & & & 366 & 376 & 362 & 378 & 365 & 375\\ & & & & & & & 371 & 373 & 362 & 375 & 368\\ & & & & & & & & 372 & 365 & 361 & 367\\ & & & & & & & & & 370 & 378 & 363\\ & & & & & & & & & & 363 & 378\\ & & & & & & & & & & & 365\end{Bmatrix}$$

式(4)中包含的12列均生函数是从表1所示的时间序列按1至12的时间间隔计算均值所派生出来的，其周期长度从左至右为1至12年。

3 最优子集回归建模

为了建立预报效果更好的模型，除了将原序列派生的均生函数作为预报因子备选外，还需对原序列作差分变换并计算相应的均生函数。

3.1 一阶差分序列

对于原序列式(1)，令

$$\Delta x(t)=x(t+1)-x(t)\quad(t=1,2,\cdots,n-1)\tag{5}$$

从上式可得一阶差分序列

$$x^{(1)}(t)=\Delta x(1),\Delta x(2),\cdots,\Delta x(n-1)\tag{6}$$

3.2 二阶差分序列

对于式(5)，令

$$\Delta^2 x(t)=\Delta x(t+1)-\Delta x(t)\quad(t=1,2,\cdots,n-1)\tag{7}$$

据此可得二阶差分序列

$$x^{(2)}(t)=\Delta^2 x(1),\Delta^2(2),\cdots,\Delta^2 x(n-2)\tag{8}$$

将原序列 $x(t)$ 的均生函数记为 $\bar{x}_l^{(0)}$，将一阶差分序列 $x^{(1)}(t)$ 和二阶差分序列 $x^{(2)}(t)$ 的均生函数分别记为 $\bar{x}_l^{(1)}(t)$ 和 $\bar{x}_l^{(2)}(t)$，利用式(4)可得它们的延拓序列 $f_l^{(0)}(t)$、$f_l^{(1)}(t)$ 和 $f_l^{(2)}(t)$。

3.3 累加延拓序列

在原序列起始值和一阶差分序列均生函数延拓序列的基础上，进一步建立累加延拓序列：

$$f_l^{(3)}(t)=x(1)+\sum_{i=1}^{t-1}f_l^{(1)}(i+1)\quad(t=2,3,\cdots,n;l=1,2,\cdots,m)\tag{9}$$

这样，从原序列可派生出 $4m$ 个均生函数延拓序列 $f_l^{(0)}(t)$、$f_l^{(1)}(t)$、$f_l^{(2)}(t)$ 和 $f_l^{(3)}(t)$ 作为自变量供选择。

3.4 粗选预报因子

对于本例而言，共有 $4\times12=48$ 个自变量，数量显然太多，应筛选与预报量关系较好的自变量作为预报因子。每个自变量与原序列的相关系数为0.00至0.69不等，令 $\alpha=0.05$，查相关系数临界值表[6]可知当相关系数≥0.32时为显著相关，从48个自变量中共选出13个自变量符合此条件，依次命名为 x_1～x_{13} 作为预报因子，预报量(高温极值)序列以 y 表示。

3.5 最优子集回归

在所有可能子集回归方程中，效果最好的一个子集回归称为最优子集回归。这里采用兼顾趋势和数量的双评分准则(CSC)[7]来识别模型，即

$$CSC = S_1 + S_2 \tag{10}$$

式中 S_1 为数量评分，S_2 为趋势评分。

$$S_1 = nR^2 \tag{11}$$

其中 R 为复相关系数。

$$S_2 = 2\Big[\sum_{i=1}^{I}\sum_{j=1}^{I} n_{ij}\ln n_{ij} + n\ln n - \Big(\sum_{i=1}^{I} n_i \ln n_i + \sum_{j=1}^{I} n_j \ln n_j\Big)\Big] \tag{12}$$

式中 I 为预报趋势类别数，n_{ij} 为 i 类事件 j 类估计事件的个数。

表2给出南京不同自变量个数的高温极值最优子集组合及其 R 和 CSC 值。从中可以看出，由7个自变量组成的子集回归 CSC 值最大，其后随着自变量个数增加，不但 CSC 值不断降低，复相关系数也没有提高，由此确定最优子集回归方程为：

$$\hat{y} = -837.286 + 0.815x_1 + 0.801x_3 + 0.596x_4 + 0.517x_6 + 0.532x_7 + 0.166x_9 - 0.135x_{11} \tag{13}$$

表2 南京高温极值预测不同自变量个数最优子集

K	最优子集	R	CSC
1	x_7	0.69	32.72
2	x_6x_{12}	0.84	42.73
3	$x_1x_6x_7$	0.88	52.68
4	$x_1x_3x_6x_7$	0.91	62.20
5	$x_1x_3x_4x_6x_7$	0.94	68.83
6	$x_1x_3x_4x_6x_7x_{11}$	0.94	71.75
7	$x_1x_3x_4x_6x_7x_9x_{11}$	0.94	74.49
8	$x_1x_3x_4x_6x_7x_9x_{10}x_{11}$	0.94	73.65
9	$x_1x_3x_4x_6x_7x_8x_9x_{10}x_{11}$	0.94	72.83
10	$x_1x_3x_4x_6x_7x_8x_9x_{10}x_{11}x_{13}$	0.94	71.98
11	$x_1x_3x_4x_6x_7x_8x_9x_{10}x_{11}x_{12}x_{13}$	0.94	71.12
12	$x_1x_2x_3x_4x_6x_7x_8x_9x_{10}x_{11}x_{12}x_{13}$	0.94	70.23
13	$x_1x_2x_3x_4x_5x_6x_7x_8x_9x_{10}x_{11}x_{12}x_{13}$	0.94	69.33

将1961～1996年观测值构成的7个均生函数延拓序列代入式(13)可得到南京高温极值的拟合值，拟合均方根误差 $RMSE$ = 0.41℃，拟合值与实况相当接近，特别是1966年的峰值和1982年的谷值(见图1)完全吻合。1997～2001年的预报结果表明，1997、2000、2001年的预报误差为0.3～0.6℃；只有1999年误差较大达2.1℃。

图1 南京历年高温极值实况(实线)和拟合、预报(虚线)曲线

4 分析和讨论

采用同样的方法，分别获得武汉、宜昌、重庆等站高温极值的拟合值、预测值与实况的比较(图2)。在计算过程中，武汉共粗选12个自变量备选，最后选中6个因子建立最优子集回归方程；宜昌、重庆各粗选11、13个自变量备选，最后各确定6个和4个因子组成预报方程。

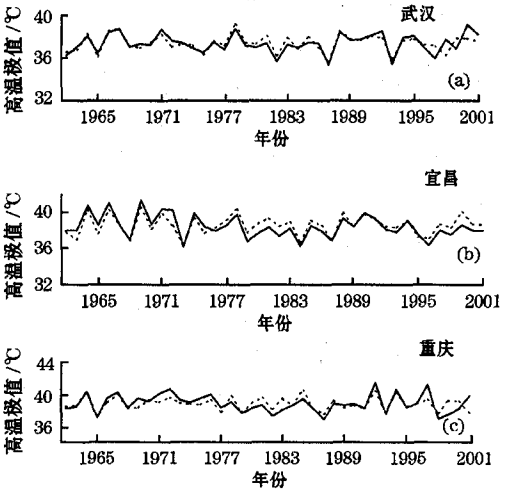

图2 武汉(a)、宜昌(b)、重庆(c)历年高温极值实况(实线)和拟合、预报(虚线)曲线

4.1 拟合结果分析和检验

由图2可见，武汉、宜昌、重庆等站1961～1996年的高温极值实况和模型的拟合值极为相似，一些周期性的波动趋势几乎完全一致。如武汉的高温从1987年的35.4℃跳跃到1988年的38.5℃，这两年的拟合值分

别为35.7℃和38.5℃，与实况相差无几；宜昌上世纪60年代有3年突破40℃高温，相应年份的拟合值也都在40℃以上，1976～1986年的拟合值虽然普遍比实况高出0.5℃左右，但两者的趋势是完全一致的；重庆上世纪60年代前期和90年代前期所出现的较大波动在拟合曲线上均有充分反映，两者的数值也比较贴近。

武汉、宜昌、重庆3个最优子集回归方程的均方根误差 *RMSE* 分别为0.35、0.66、0.63℃，复相关系数 *R* 分别为0.92、0.86和0.80。

另外，经计算可得南京、武汉、宜昌、重庆等回归方程的 *F* 分别为30.36、26.62、13.73和13.78，全部通过 $\alpha=0.01$ 的显著性检验。

4.2 预报结果验证

表3列出南京等4站1997～2001年高温极值预报值与实测值的误差，从表中20站年的预报误差来看，除个别站年（南京1999年）超过2.0℃外，其余站年均在2.0℃之内，其中11站年在1.0℃之内，表明所构造的预测模型未来1～5年内具有一定的预测能力。各站的平均误差为0.8～1.4℃之间，说明预测模型具有一定的稳定性，值得在业务应用中尝试。

表3 南京等4站预报值的绝对误差（℃）

站名	1997	1998	1999	2000	2001	平均
南京	0.3	1.3	2.1	0.2	0.6	0.9
武汉	1.2	1.6	1.0	1.4	0.8	1.2
宜昌	0.7	0.8	1.4	0.6	0.6	0.8
重庆	1.8	0.7	1.8	0.9	2.0	1.4
平均	1.0	1.1	1.6	0.8	1.0	1.1

4.3 讨论

（1）本方法通过原序列与该序列隐含的各种周期变化和演变趋势的关系来构造预测模型，要求原序列具有足够的长度；

（2）当预测时段出现反周期突变或出现前所未有的极值时，预报误差较大；

（3）适当缩短预报时效可望提高预报精度。

参考文献

1 丁金才，周红妹，叶其欣. 从上海市热岛演变看城市绿化的重要意义. 气象，2002，28(2)：22～24.

2 北京市气象局气候资料室. 北京城市气候. 北京：气象出版社，1992：3～18.

3 张尚印，宋艳玲. 夏季高温及其影响. 2001年全国气候影响评价. 北京：气象出版社，2002：41～44.

4 盛承禹. 中国气候总论. 北京：科学出版社，1986：222～224.

5 魏凤英. 现代气候统计诊断预测技术. 北京：气象出版社，1999：214～218.

6 屠其璞，王俊德，丁裕国等. 气象应用概率统计学. 北京：气象出版社，1984：251～255，531～533.

7 曹鸿兴，魏凤英. 估计模型维度的双评分准则及其应用. 数理统计与应用概率，1996，3：33～40.

Application of Mean Generating Function-Optimal Subset Regression to the Prediction of High Temperature Extremes

Zhang Dekuan Yang Xianwei Zou Xukai

(National Climate Center, Beijing 100081)

Abstract

According to the yearly high temperature extremes from 1961 to 1996 in Nanjing and other three cities, the short-range climate prediction models are developed by use of mean generating function and optimal subset regression methods. The results show that they can not only fit the historical sequences perfectly, but also possess predictive capability for coming 1～5 years' changes to a certain extent.

Key Words: Mean Generating Function (MGF) Optimal Subset Regression (OSR) high temperature extreme