我国月地面温度场的自回归滑动平均模型

曹鸿兴 黄文杰
（气象科学研究院气候研究所）（气象科学研究院情报所）

王棉棉
（北京工业学院）

提 要

对1952—1980年我国连续的月地面气温用时间序列ARMA(p,q)模型进行随机建模。月温度场由60个站组成，用经验正交函数加以展开，取不同的样本长度即348，336和300月，以便考察经验正交展开的稳定性。前四个主成分，即z_1,z_2,z_3,z_4取为多维时间序列的变数，因为它们的总方差贡献达99.26%。在这四个主成分序列中的决定性周期用周期图和最大熵方法加以揭露。对一维变量z_i，$(i=1,2,3,4)$的ARMA(p,q)的模型识别用Pandit－Wu方法进行，这样就可求得实验模型。用z_i模型的外推值来预报月温度场。距平预报的命中率评分为78.3%，高于目前的业务长期天气预报。

一、引 言

随着Box和Jenkins[1]的书的出版，时间序列分析日臻完善，其中尤以自回归滑动平均模型（ARMA）日益被广泛地应用于各个领域，如自动控制、预测预报等。Box－Jenkins方法是以相关分析为基础的，模型识别、定阶和模型参数估计等都着眼于统计观点，但这种建模方法中的某些环节需凭藉人的判断，对计算机自动建模是不方便的。以后Pandit和吴贤铭[2]从系统分析的角度来研究时间序列，把序列看成是具有独立随机输入的动态系统的响应，提出了节省计算量的建模方法，使随机模型的建立能在计算机上自动实现。

本文提出了一个运用随机建模预报大范围月平均气温的方案。首先把60个站的1952－1980年的逐月地面气温场用经验正交函数（EOF）展开，视每一个EOF时间分量为一维数据序列，用Pandit－Wu方法分别建模，月气温场的预报则归结为运用这些模型作外推。从所得试报结果来看，预报评分高于现有的业务月预报水平。

二、建模原理

设要素场序列

本文1987年3月18日收到，6月19日收到修改稿。

$$X=\begin{bmatrix} x_1(1) & x_1(2)\cdots x_1(n) \\ x_2(1) & x_2(2)\cdots x_2(n) \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ x_m(1) & x_m(2)\cdots x_m(n) \end{bmatrix} \equiv (X(1), X(2), \cdots, X(n)) \tag{1}$$

式中 n 为样本长度，m 为测站数或网格点数。对要素场 X 用经验正交函数（缩写为 EOF，或称自然正交函数）展开，

$$X=EZ \tag{2}$$

称 E 为空间分量，Z 为时间分量或主成分。易求得 Z 的数量表达式

$$z_i(t)=\sum_{k=1}^{m} e_{ik}x_k(t) \quad i=1, 2, \cdots m, t=1, 2, \cdots n \tag{3}$$

式中 e_{ik} 表示第 k 个站点上的第 i 个 EOF 值。

把方差贡献小的主成分略去，即只取 $s<m$ 个主成分来对要素场进行拟合，就达到了压缩时间序列维数的目的。

对每一个主成分序列 $z_i(t), i=1,2,\cdots,s$，分别建立自回归滑动平均（ARMA）模型

$$z_i(t)=\varphi_1 z_i(t-1)+\varphi_2 z_i(t-2)+\cdots+\varphi_p z_i(t-p)+a_t-\theta_1 a_{t-1}-\theta_2 a_{t-2}-\cdots-\theta_q a_{t-q} \tag{4}$$

式中，$\varphi_1,\varphi_2,\cdots\varphi_p$ 为 p 阶自回归参数，$\theta_1,\theta_2,\cdots\theta_q$ 为 q 阶滑动平均参数，$a_t,a_{t-1},\cdots,a_{t-q}$ 为白噪声序列。

由 ARMA 模型可以推导得一种等价形式，即逆转形式

$$a_t=z_i(t)-\sum_{j=1}^{\infty} I_j z_i(t-j) \quad i=1, 2, \cdots, s \tag{5}$$

式中，$I_j,(j=1,2,\cdots)$ 称为逆函数，它具有负指数下降的性质，$I_0=-1$。逆转形式是由随机序列的加权和来表示白噪声序列。利用逆函数 I_j 与模型参数 $\varphi_i,\theta_j(i=1,2,\cdots p;j=1,2,\cdots q)$ 间的线性关系可以发展出确定参数初值的估计法——逆函数法。

由于 ARMA 模型在理论上比较完善，对实际数据有较好的拟合效果，因此在月温度场的预报中采用 ARMA 模型。当一旦用观测值估计出模型参数 $\hat{\varphi},\hat{\theta}$ 并作出一步预报 $z_i(t+1)$，其中 $i=1,2,\cdots,s$，就可根据

$$x_i(t+1)=\sum_{j=1}^{s} e_{ij}z_j(t+1) \quad i=1, 2, \cdots, m \tag{6}$$

作出温度场的一步预报。

三、 我国月气温场的经验正交分解

月气温资料包括全国 60 个站时间为 1952 年 1 月至 1980 年 12 月，这就是说在作经验正交分解时站点数 $m=60$，样本长度 $n=348$，原始资料 X' 作中心化处理

$$x_i(t)=x_i'(t)-\overline{x_i'} \qquad i=1, 2, \cdots, m$$

$$\overline{x_i'}=\frac{1}{n}\sum_{t=1}^{n} x_i'(t) \qquad t=1, 2, \cdots, n \tag{7}$$

资料并没有象通常所做的那样，作消除年变化的处理。计算表明，这样做可以用少量几个经验正交函数即能高精度地拟合原始的场序列。

前四个经验正交函数分别给在图 1—4。由图可以看出，第一经验正交函数反映了我国气温随纬度的变化，等值线大体上与纬圈相平行，只是在江南地区有一向西南方向伸展的低值区。第二经验正交函数反映我国气温随经度的变化，等值线呈南北或西南－东北向，可能与我国的地形高度分布有关。第三经验正交函数的地域分布则把中国分成四块，分别为东北、西北、西南和江淮—华南，反映了我国的四种不同的大的气候类型。第四经验正交函数的分布型式与我国大陆度分布甚为一致[3]，可能反映了我国气温受海陆分布的影响。

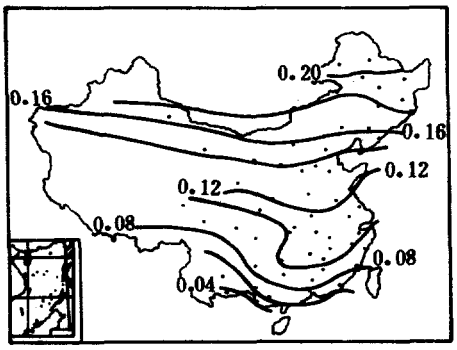

图 1　月气温场的第一经验正交函数
图中小点为所取资料的站点(下同)

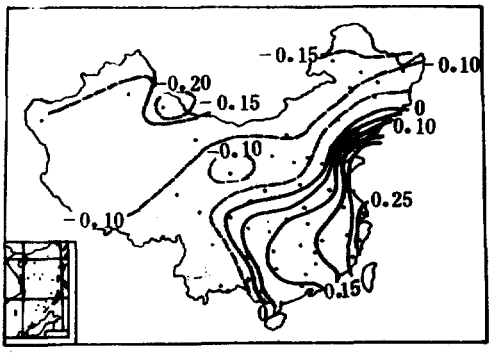

图 2　月气温场的第二经验正交函数

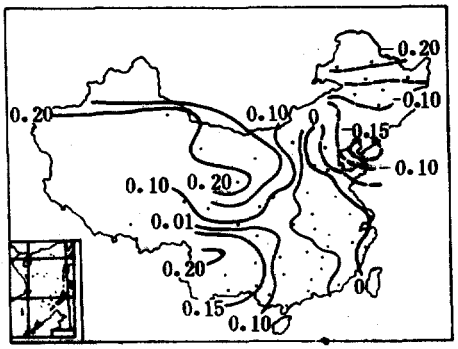

图 3　月气温场的第三经验正交函数

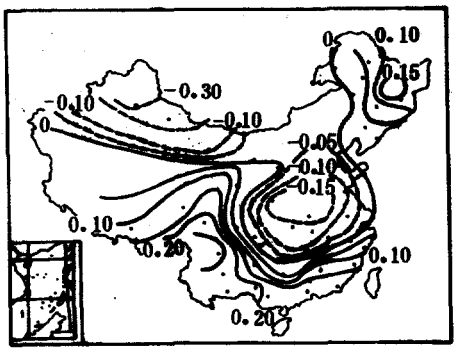

图 4　月气温场的第四经验正交函数

现采用向量的夹角余弦来考察经验正交函数的稳定性。

$$\cos\alpha_i = e_i^{*k} \cdot e_i^{*l} / \| e_i^{*k} \| \| e_i^{*l} \| \quad i = 1, 2, \cdots, p \tag{8}$$

式中，n_k 和 n_l 为不同的样本长度，$e_i^{*k} \cdot e_i^{*l}$ 为向量的内积，$\| \cdot \|$ 为向量模。分别取样本长

度 $n_1=348$，$n_2=336$ 和 $n_3=300$ 来进行经验正交分解，所得值代入(8)式，求得的余弦值列于表 1。如果取 $a_0=\pi/12$，$\cos a_0=0.966$ 作为界限值，由表 1 可见，前四个经验正交向量是稳定的。这四个向量的方差累计贡献达到 99.26%，因此我们取与此相对应的四个时间分量 $z_1(t)$，$z_2(t)$，$z_3(t)$，$z_4(t)$作为多维时间序列的变量。

表 1 不同样本长度的经验正交向量间的夹角余弦

样本长度	n_1 与 n_2 间	n_1 与 n_3 间
1	0.9992	0.9964
2	0.9999	0.9994
3	0.9937	0.9944
4	0.9989	0.9913
5	−0.8865	0.9140
6	−0.9431	0.8507
7	−0.9082	−0.5451
8	−0.9956	0.7501

四、 ARMA 建模

引用后移算子 B，即

$$Bz_i(t)\triangleq z_i(t-1),\ B^2z_i(t)=BBz_i(t)=z_i(t-2),\ \cdots,\ B^nz_i(t)=z_i(t-n)$$

将(1)改写为

$$\Phi(B)z_i(t)=\Theta(B)a(t) \tag{9}$$

式中

$$\Phi(B)=1-\varphi_1B-\cdots-\varphi_pB^p$$

$$\Theta(B)=1-\theta_1B-\cdots-\theta_qB^q$$

用 Pandit—Wu 方法进行模型识别和参数估计[4]。这一方法的要点如下：

从系统的物理背景来分析，ARMA 模型可以由常微分方程制约的动力系统导出[2]。若时间序列数据为等时间间隔采样，则 ARMA (2, 1) 模型可得自一个自由度为 1 的弹簧—质块—阻尼系统，该系统由带白噪声的二阶常微分方程确定。一个自由度为 2 的系统由四阶常微分方程确定，离散化后便得 ARMA(4,3)模型，按归纳法，得知有 ARMA($2n$,$2n-1$)，$n=1,2,3,\cdots$，的规律。从计算量来看，模型阶数倍数为 2 是经济的，尤其对于高阶模型，可大大节省计算量，因为它不需要对 ARMA (p, q) 模型中 p, q 由小到大逐步增加的每种可能搭配即穷举法进行计算。

从 ARMA(2,1)模型开始，以 2 为增量逐步增加模型参数，对给定 ARMA($2n$,$2n-1$)，用 ARMA 序列参数的粗估计方法，如逆函数法确定相应的参数初值，然后用非线性阻尼最小二乘法使残差平方和函数极小求得参数的精估计值。用 F—检验考察 ARMA ($2n$, $2n-1$)与 ARMA($2n+2$,$2n+1$)模型残差平方和的差异。如果 F 值下降显著，则继续增加以 2 为倍数的阶数，重复上述各步。若 F 值下降不显著，停止增加阶数，ARMA ($2n^*$,$2n^*$ —

1）即为所要的模型。同时进一步用 F- 检验和置信区间原则修改、简化模型。也可用 AIC 准则来确定模型阶数。

时间分量序列 $z_i(t)$ 中的周期用最大熵谱估计和周期图方法进行寻找，以确定时序模型中的周期趋势项。结果只有第一时间分量 $z_1(t)$ 中的 12 个月长度的周期是显著的，因此对 $z_1(t)$ 作差分计算，即

$$\Delta_{12}z_1(t) = z_1(t) - z_1(t-12)$$

这样，通过差分消除序列 $z_1(t)$ 中的季节性分量。根据我们过去的工作[5]，证明这样做在使序列平稳化上是有效的，建模是对序列 $\Delta_{12}z_1(t)$ 进行的[6]，分别计算了 ARMA (2, 1)，ARMA(4,3)，ARMA(6,5)，ARMA(8,7)，ARMA(6,4)模型，用 F－检验定阶为(6,4)，但用 AIC 准则定阶为(4,3)，最后选定模型 ARMA(4,3)，因为它的阶数低，便于实际使用，模型中的 7 个参数分别为

$\varphi_1 = 1.109 \pm 0.50$，$\varphi_2 = -1.237 \pm 0.52$，$\varphi_3 = -0.131 \pm 0.16$，$\varphi_4 = 0$，$\theta_1 = 0.970 \pm 0.52$，$\theta_2 = 0.968 \pm 0.49$，$\theta_3 = 0.013 \pm 0.48$。第二时间分量序列 $z_2(t)$ 的模型为 ARMA (3, 3)，其参数值为 $\varphi_1 = 1.868 \pm 0.00$，$\varphi_2 = -1.240 \pm 0.00$，$\varphi_3 = 0.142 \pm 0.00$，$\theta_1 = 1.577 \pm 0.01$，$\theta_2 = -0.747 \pm 0.01$，$\theta_3 = -0.082 \pm 0.01$。第三时间分量序列 $z_3(t)$ 的模型为 ARMA(4,3)，第四时间分量序列 $z_4(t)$ 的模型为 ARMA(6,5)，参数值不再列出。

五、预　报

用新息预报方法对时间分量进行递推[7]。记 $z_i(l)$ 为时刻 t 的 l 步预报，它的 Wold 展开为

$$\hat{z}_t(l) = \sum_{j=l}^{\infty} G_j b_{t+l-j}$$

式中 G_j 是 ARMA 模型中的格林函数，b_t 是白噪声，称

$$b_{t+l} \triangleq z_{t+l} - \hat{z}_t(l)$$

为时刻 $t+l$ 的新息。进一步可导得

$$\begin{aligned}
\hat{z}_{t+1}(l-1) &= \sum_{j=l-1}^{\infty} G_j b_{t+l-j} \\
&= G_{l-1}b_{t+1} + \sum_{j=l}^{\infty} G_j b_{t+l-j} \\
&= G_{l-1}b_{t+1} + \hat{z}_t(l) \\
\hat{z}_{t+2}(l-2) &= \sum_{j=l-2}^{\infty} G_j b_{t+l-j} \\
&= G_{l-2}b_{t+2} + \sum_{j=l-1}^{\infty} G_j b_{t+l-j} \\
&= G_{l-2}b_{t+2} + \hat{z}_{t+1}(l-1) \\
&= G_{l-2}b_{t+2} + G_{l-1}b_{t+1} + \hat{z}_t(l) \\
&\cdots\cdots\cdots\cdots \\
\hat{z}_{t+k}(l-k) &= \hat{z}_t(l) + \sum_{j=l-k}^{l-1} G_j b_{t+l-j} \qquad (l-k) > 0
\end{aligned} \tag{10}$$

式中 $G_j b_{t+l-j}$ 是时刻 $t+l-j$ 对 $\hat{z}_t(l)$ 的新息订正。b_{t+l-j} 是 $t+l-j$ 时刻的新息，G_j 就是 z_{t+l} 在 b_{t+l-j} 上的投影值。这样，利用(10)式，每过一个时刻得到的新息，即实测值对预报的贡献，立即吸收到预报中，随着预报时刻的临近，通过订正，使预报误差逐步减少，并不需要每过一个时刻把全部预报值重新算一遍。

作为示例，图 5*a* 给出 1981 年 5 月温度场的预报，图 5*b* 为实况。对照两张图，预报与实况相当一致，由华北伸向东北的暖脊能报出，只是 22℃的暖中心区未报出。

用两种评分指数检验预报结果。距平符号百分率

$$H_1 = \frac{R_1}{m}$$

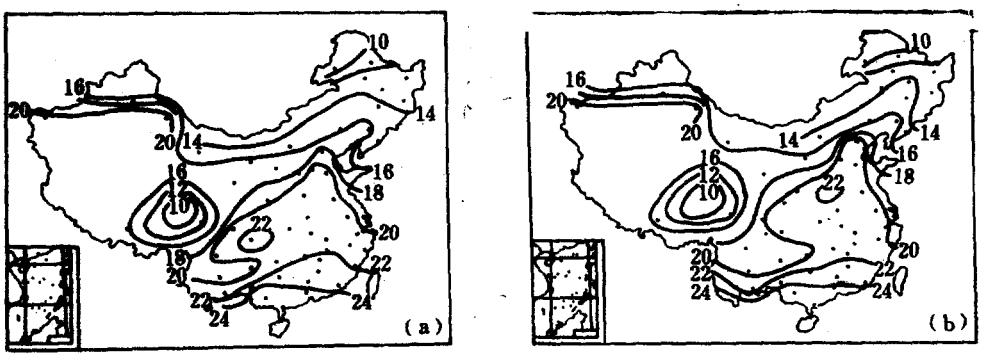

图 5 1981 年 5 月月气温场(a)预报(b)实况(单位：C)

式中 R_1 是预报的距平符号和观测的距平符号相同的个数，m＝60，是站数。绝对值差

$$H_2 = \frac{R_2}{m}$$

式中 R_2 是预报值与观测值的绝对值差小于 1℃的个数。对 1981 年 1 月至 5 月的气温场的评分给在表 2。距平符号百分率的平均值为 78.3%，高于目前的长期业务预报的水

表 2 月温度场预报评分

评分指数 \ 月份	1月	2月	3月	4月	5月	平均
H_1(%)	83.3	96.7	80.0	50.0	81.7	78.3
H_2(%)	80.0	85.0	63.3	55.0	83.3	73.3

平[8]。可见，本方法可以作为月气象场预报的客观方法用在业务工作中，尽管建模时需化费较多计算机时间，但在作实际预报时计算量是不大的。可取的是能报出一个内在统一的要素场，避免了各站分别建立统计模型，预报出现矛盾难以综合的问题。

参 考 文 献

[1] Box, G. E. P. , and G. M. Jenkins, Time series analysis, forecasting and control. Holden Day. San Francisco. 1970.

[2] Pandit. S. M. , and S. M. Wu. Time series and system analysis with application. John Wiley and Sons. New York. 1983.

[3] 张家诚、林之光. 中国气候. 上海科学技术出版社. 121. 1985 年.
[4] 项静恬等. 动态数据处理. 气象出版社. 223—230. 1986 年.
[5] 曹鸿兴. 局地天气预报的数据分析方法. 气象出版社. 86—88. 1983 年.
[6] 黄文杰等. 时间序列的 ARIMA 季节模型在长期预报中的应用. 科学通报. 25. 22. 1030—1032. 1980.
[7] 安鸿志等. 时间序列的分析与应用. 科学出版社. 193—200. 1983 年.
[8] 叶愈源. 十年来长期天气预报的检查. 气象. 10. 1980.

AN AUTOREGRESSIVE MOVING AVERAGE MODEL OF MONTHLY SURFACE TEMPERATURE FIELD OVER CHINA

Cao Hongxing
(Institute of climatology, AMS)
Huang Wenjie
(Institute of Meteorological Scientific and Technical Information, AMS)
Wang Mianmian
(Department of Automatic Control, Beijing Institute of Technology)

Abstract

For designing a time series of the successive monthly surface temperature in 1952—1980 over China a stochastic model is formulated by an ARMA (p,q) model. The monthly temperature fields composed of records at 60 stations are expanded by means of EOF (Empirical Orthogonal Function) with the various sample sizes 348, 336 and300 months so as to examine the stability of EOF expansion. Then the first four principal components, i. e. z_1, z_2, z_3, z_4 are taken as the variables of multivariate time series, since their total variance contribution is 99. 26%. The major periods in first four principal components series are also revealed by using the periodogram and the maximum entropy method. The model identification of ARMA (p, q) for univariate variables z_i ($i = 1, 2, 3, 4$) is made by the Pandit—Wu method, then, the empirical models are obtained.

The extrapolations of z_i' s models are used for predicting a monthly temperature field. The score of hit ratio of departure forecasts is 78. 3%, which is better than that of the recent operational long—range weather forecast.