

一种动态数据的新建模法及其预报应用*

贾晓静**

曹鸿兴

封国林

(中国气象科学研究院,北京 100081) (扬州大学物理学院,扬州 225009)

提 要

文章提出了一种新的动态数据建模法,利用观测的数据序列,先用双向差分原理反导出一个非线性常微分方程。以此作为微分动力核,然后运用自忆性原理建立预报模式,我们称之为数据机理自记忆模式(Data-based Mechanism Self-memory Model),简称为数忆模式,缩写为 DAMSM。多个实例计算表明,数忆模式的预报准确率是比较令人满意的,给出了长江三角洲夏季降水年际预报的实例。

关键词:自记忆 时间序列 年际预报 汛期降水

引 言

时间序列分析,也称动态数据分析有着广泛的应用,如用在气象、水文、商情、经济、地震等领域。现有的建模方案主要以自回归滑动平均(ARMA)为主体模型的 Box-Jenkins 方案、Pandit-Wu 方案,另有指数模型、门限回归模型等^[1]。1987 年曹鸿兴、魏凤英提出的均生函数模型也取得了相当好的预报效果。近几年来,曹鸿兴基于大气运动是一种不可逆过程的观点,提出了大气运动的自忆性原理,经过不断丰富与应用,证明是一种能提高预报准确率的有效途径^[2]。

近年来发展了一种称为基于数据的机理模式(Data-based Mechanism Model),就是先寻找与因变量关系密切的自变量,分析其间可能存在的函数关系,然后通过反演,建立一个以观测数据为基础的机理模式^[3,4]。

由于大气系统是一种非线性系统,而观测资料可以看做是描写大气非线性动力系统的一系列特解,假若我们有一组时间序列,便可以由此反导出一个描述大气运动的模型,林振山等^[5]根据观测资料反导出由三变量常微分方程组成的气候模式。在本文中先利用自记忆的思想设定一个可能的非线性常微分方程,然后运用双向差分法反导得到方程的具体形式,称为反导模式。随后再利用自忆性原理推导一个差分-积分方程,运用最小二乘法求得该方程离散化形式中的记忆系数,由此建立模式并进行预报试验。以长江三角洲 1951~1998 年的夏季降水量为例进行建模,将试报结果与反导模式和均生函数模式作比较,其结果表明,经过自记忆处理之后的模式,其预报效果有较大的改进。数忆模式的预报准确性也高于均生函数模型^[6]。

在气象中经常遇到只有一个长时期的观测序列,如由冰芯资料反演的 100~1000 年

* 国家自然科学基金(49875025)资助。

2000-04-03 收到,2000-07-13 收到修改稿。

** 现在在国家气象中心工作。

的温度序列,我国根据文献记载反推得的 500 年旱涝序列。我们就可用数忆模式法建模并用以制作预报。由于此方法只涉及到一定长度的数据序列,所以它可以运用到许多非气象领域,如水文、地震、经济等,应用前景是比较可观的。

1 数学原理

1.1 模式

设描述大气非线性系统的状态变量为 x , 有其一组离散观测数据:

$$X = \{x(i)\} \quad i = 1, 2, \dots, n$$

设变量 x 随时间变化的方程为:

$$\frac{dx}{dt} = a_1 x_t + a_2 x_{t-1} + \dots + a_p x_{t-p+1} + b_1 x_t^2 + b_2 x_{t-1}^2 + \dots + b_p x_{t-p+1}^2 \quad (1)$$

其中 $a_1, a_2, \dots, a_p, b_1, b_2, \dots, b_p$, 为待定系数, 式中 x 为连续变量。 $x_t, x_{t-1}, \dots, x_{t-p+1}$ 分别为变量在 $t, t-1, \dots, t-p+1$ 时刻的值。在此方程里我们假设回溯阶为 p , 即变量 x 的变化与 $t, t-1, \dots, t-p+1$ 个时刻的变量值有关。

按数据机理模式的原意, 应在式(1)中置入数据序列中隐含的物理项。如序列有周期性, 可加入正弦、余弦项, 序列增长显著应加入指数项, 若有其他因素影响显著, 应纳入该因素作为方程的自变量。作为一种新建模法的初步尝试, 本工作只考虑了线性项与平方项。

1.2 用双向差分法求系数

假设数据序列为等时间间隔采样, 设 $\Delta t = (t+1) - t = 1$, 改写方程(1)为差分方程

$$\Delta x = a_1 x_t + a_2 x_{t-1} + \dots + a_p x_{t-p+1} + b_1 x_t^2 + b_2 x_{t-1}^2 + \dots + b_p x_{t-p+1}^2 \quad (2)$$

由数值差分知, 差分 Δx 可以有向前差分和向后差分^[6], 分别表达为

$$\Delta_b x_k = x_k - x_{k-1} = a_1 x_{k-1} + \dots + a_p x_{k-p} + b_1 x_{k-1}^2 + \dots + b_p x_{k-p}^2 + \varepsilon_{bk} \quad (3)$$

$$\Delta_f x_k = x_{k+1} - x_k = a_1 x_k + \dots + a_p x_{k-p+1} + b_1 x_k^2 + \dots + b_p x_{k-p+1}^2 + \varepsilon_{fk} \quad (4)$$

由式(3), (4)可得:

$$\varepsilon_{bk} = (x_k - x_{k-1}) - (a_1 x_{k-1} + \dots + a_p x_{k-p} + b_1 x_{k-1}^2 + \dots + b_p x_{k-p}^2) \quad (5)$$

$$\varepsilon_{fk} = (x_{k+1} - x_k) - (a_1 x_k + \dots + a_p x_{k-p+1} + b_1 x_k^2 + \dots + b_p x_{k-p+1}^2) \quad (6)$$

$\varepsilon_{bk}, \varepsilon_{fk}$ 分别为向后差分误差和向前差分误差。

双向差分原则就是使向前差分预报误差与向后差分预报误差之和达到最小, 运用最小二乘法求系数。即使:

$$\sum (\varepsilon_{bk}^2 + \varepsilon_{fk}^2) \rightarrow \min \quad (7)$$

将式(5), (6)代入式(7), 得到式(7)的表达式, 且令:

$$\mathcal{E}^2 = \sum_{k=1}^n (\varepsilon_{bk}^2 + \varepsilon_{fk}^2) \quad (8)$$

然后根据最小二乘原理, 对式(8)分别对 $a_1, a_2, \dots, a_p, b_1, b_2, \dots, b_p$ 求导, 得到 $2p$ 个方程组, 对上述 $2p$ 个方程组, 可以令

$$\begin{aligned}
 Z_1 &= \sum_{k=1}^n [(x_k - x_{k-1}) x_{k-1} + (x_{k+1} - x_k) x_k] \\
 &\quad \vdots \\
 Z_p &= \sum_{k=1}^n [(x_k - x_{k-1}) x_{k-p} + (x_{k+1} - x_k) x_{k-p+1}] \\
 Z_{p+1} &= \sum_{k=1}^n [(x_k - x_{k-1}) x_{k-1}^2 + (x_{k+1} - x_k) x_k^2] \\
 &\quad \vdots \\
 Z_{2p} &= \sum_{k=1}^n [(x_k - x_{k-1}) x_{k-p}^2 + (x_{k+1} - x_k) x_{k-p+1}^2]
 \end{aligned} \tag{9}$$

又令

$$\mathbf{Z} = \begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_{2p} \end{pmatrix} \quad \mathbf{S} = \begin{pmatrix} S_{1,1} & S_{1,2} & \cdots & S_{1,2p} \\ S_{2,1} & S_{2,2} & \cdots & S_{2,2p} \\ \vdots & \vdots & \ddots & \vdots \\ S_{2p,1} & S_{2p,2} & \cdots & S_{2p,2p} \end{pmatrix} \quad \boldsymbol{\alpha} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ b_p \end{pmatrix}$$

$$\begin{aligned}
 S_{1,1} &= \sum_{k=1}^n (x_{k-1} x_{k-1} + x_k x_k) \cdots S_{1,2p} = \sum_{k=1}^n (x_{k-1} x_{k-p}^2 + x_k x_{k-p+1}^2) \\
 &\quad \vdots \\
 S_{p,1} &= \sum_{k=1}^n (x_{k-1} x_{k-p} + x_k x_{k-p+1}) \cdots S_{p,2p} = \sum_{k=1}^n (x_{k-1} x_{k-p}^2 + x_{k-p+1} x_{k-p+1}^2)
 \end{aligned} \tag{10}$$

$$\begin{aligned}
 S_{p+1,1} &= \sum_{k=1}^n (x_{k-1} x_{k-1}^2 + x_k^2 x_k) \cdots S_{p+1,2p} = \sum_{k=1}^n (x_{k-1}^2 x_{k-p}^2 + x_k^2 x_{k-p+1}^2) \\
 &\quad \vdots
 \end{aligned}$$

$$S_{2p,1} = \sum_{k=1}^n (x_{k-p}^2 x_{k-1} + x_{k-p+1}^2 x_k) \cdots S_{2p,2p} = \sum_{k=1}^n (x_{k-p}^2 x_{k-p}^2 + x_{k-p+1}^2 x_{k-p+1}^2)$$

则 $2p$ 个方程可以写成矩阵的形式:

$$\mathbf{Z} = \mathbf{S} \boldsymbol{\alpha} \tag{11}$$

可以由多种方法来求式(11)中的未知系数,如最小二乘法,矩阵法等等。系数求出来以后,取某种标准进行筛选,如取相对方差做判据,方程(1)中,一次项系数为 a_k ,二次项系数为 b_k ,令: $C_k = |a_k \vdots b_k|$,表示任意系数,则取判据:

$$\alpha_k = \frac{c_k^2}{\sum_k c_k^2}$$

当 α_k 小于一个规定的值时,可以剔除此项,即认为该项在方程(1)中起的作用较小,由此来确定关于变量 x 的常微分方程的项及系数,在本文中取 $\alpha_k = 0.25$ 。将确定的常微分方程的右端记为 F ,由于方程右端均为已知,所以用它就可以做预报,只要把方程左边的微分变成差分即得。

2 自记忆建模

将上文确定的 $dx/dt = F$ 作为一个动力核,运用自忆性原理,求得一个新的预报模式。为此引进记忆函数 $\beta(\tau)$,对方程求 t_0 至 t 的加权积分:

$$\int_{t_0}^t \beta(\tau) \frac{\partial x}{\partial \tau} d\tau = \int_{t_0}^t \beta(\tau) F(x, \tau) d\tau \tag{12}$$

$$\text{其中 } F(x, \tau) = \sum_{i=1}^{p'} a_i x_i + \sum_{k=1}^{p''} b_k x_k^2 \quad p' \leq p, p'' \leq p \quad (13)$$

式中 x_i 和 x_k^2 为符合标准入选的线性项和平方项,经重新顺次编号得到的自变量。 p' 与 p'' 分别为其入选项数,运用微积分中的中值定理,回溯阶取为 p ,导得一个自忆性方程:

$$\beta x_t - \beta_p x_{t-p} - \sum_{i=-p}^0 x_i^m (\beta_{+1} - \beta) - \int_{t-p}^t \beta(\tau) F(x, \tau) d\tau = 0 \quad (14)$$

式(14)是一个 p 阶差分-积分方程,即 x 在时刻 t 的变化与它过去 p 个时刻的值有关。在(14)中,积分用求和代替,微分变为差分,中值简单地用两个时刻的值表示,即:

$$x_i^m = \frac{1}{2} (x_{i+1} + x_i) \equiv y_i \quad (15)$$

则式(12)可以写为:

$$x_t = \sum_{i=-p-1}^{-1} a_i y_i + \sum_{i=-p-1}^{-1} \theta_i F(x, i) \quad (16)$$

称

$$a_i = (\beta_{+1} - \beta) \quad \theta_i = \beta / \beta$$

为记忆系数,用式(13)与式(15)计算出 F 与 y_i 的时间序列,即可用类似于上节中的最小二乘公式求出记忆系数,就可用式(16)进行预报。

3 记忆系数的估计

假设有一组观测资料,取其中 L 个时次的资料 $x_{t1}, x_{t2}, \dots, x_{tL}$,用最小二乘法求记忆系数。其中 L 的选取与回溯阶 p 有关,至少需要 $2p$ 长度的观测资料才能求出记忆系数。所以,序列的长度不能太短。

记:

$$X_t = \begin{bmatrix} x_{t1} \\ x_{t2} \\ \vdots \\ x_{tL} \end{bmatrix} \quad \alpha = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} \quad Y = \begin{bmatrix} y_{1,1} & y_{1,2} & \cdots & y_{1,p} \\ y_{2,1} & y_{2,2} & \cdots & y_{2,p} \\ \vdots & \vdots & \cdots & \vdots \\ y_{L,1} & y_{L,1} & \cdots & y_{L,p} \end{bmatrix} \quad (17)$$

其中向量 X_t 是实际的观测资料, Y 是根据式(15)计算得到的, α 向量是待求的未知系数。

向量 $\Theta_{p \times 1}$ 的表达式与 $\alpha_{p \times 1}$ 的表达式形式类似,矩阵 $F_{L \times p}$ 的矩阵表达式与 $Y_{L \times p}$ 类似。则式(16)可以写成矩阵形式:

$$X_t = Y\alpha + F\Theta \quad (18)$$

令

$$M = \begin{bmatrix} Y \\ F \end{bmatrix} \quad W = \begin{bmatrix} \alpha \\ \Theta \end{bmatrix} \quad (19)$$

则式(18)可以写成:

$$X_t = MW \quad (20)$$

用最小二乘法求出系数矩阵 W 后,就可以进行预报。

4 长江三角洲降水预报

为了说明数忆模式的优效性以及证明运用自忆性原理之后,预报模式有了较大的改进,我们计算了几个实例。

长江三角洲降水用上海、杭州、南通、苏州、南京等 10 站组成,数据序列为 1951 ~ 1998 年 6 ~ 8 月的降水总量:

{513, 447, 461, 764, 485, 601, 631, 356, 307, 508, 385, 519, 419, 320, 487, 323, 221, 290, 497, 474, 365, 423, 372, 556, 516, 408, 460, 169, 391, 717, 407, 517, 456, 488, 424, 511, 669, 403, 548, 413, 847, 382, 660, 361, 534, 622, 516, 469}

样本量 n 为 48,回溯阶 p 取为 9。我们用 1951 ~ 1993 年的数据资料来建模,用最小二乘法计算出系数后,经过筛选,得到式(1)的表达式为:

$$\frac{dx}{dt} = -0.926 x_{t-1} + 0.2 x_{t-3} + 0.22 x_{t-4} + 0.58 x_{t-5} - 0.219 x_{t-9} + 0.244 x_{t-1}^2$$

这样得到的模式我们称之为反导模式,将得到的方程作为动力核,运用自忆性原理建立预报模式,我们称为数据机理-自记忆模式(Data-based Mechanism Self-memory Model),简称为数忆模式,缩写为 DAMSM 模式。

为比较起见,我们对 1969 ~ 1993 年的降水量用反导模式与均生函数的方法拟合,得到反导模式的 RMSE 为 129.00 mm,均生函数的 RMSE 为 88.28 mm。随后用我们的 DAMSM 模式进行拟合,得到的均方根误差(RMSE)为 75.3 mm,并对 1994 ~ 1998 年降水做了试报,1994 ~ 1998 年预报值分别为 333.96, 465.25, 598.82, 470.15, 471.00, 实际观测值为:361, 534, 622, 516, 469。

图1是用 DAMSM 对 1969 ~ 1998 年的降水量的拟合和预报与实际降水量的比较,

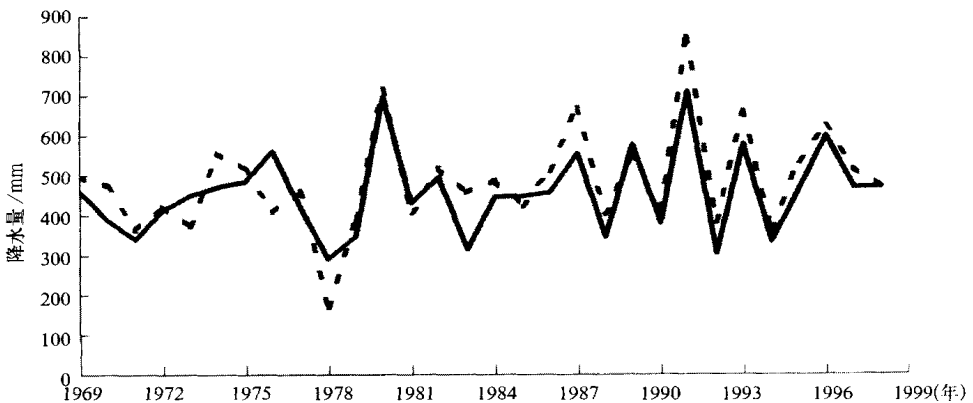


图1 DAMSM 模式对 1969 ~ 1998 年降水量的拟合预报与实际降水量的比较
(实线表示 DAMSM 模式降水量,虚线表示实际降水量)

其中序列 1 为由 DAMSM 模式拟合和预报的降水量,序列 2 为实测降水量,由图可以看出,DAMSM 模式的效果比较好。

另外,对民航运输量,太阳黑子等也进行了试报,效果也很好(图略),更多的实例有待于今后的计算。

参 考 文 献

- 1 丁裕国,江志红.气象数据时间序列信号处理.北京:气象出版社,1998.
- 2 曹鸿兴.大气运动的自忆性方程.中国科学 B 辑,1993,23(1):104~112.
- 3 Young P. Data-based mechanistic modeling, generalized sensitivity and dominant mode analysis. *Computer Physics Communication*, 1999, 117:113~129.
- 4 Steppeler J. The SO as an example of a simple, ordered subsystem of a complex chaotic system. *J. Climate*, 1997, 10(3):473~480.
- 5 林振山.天津局地气候的反演建模及研究.气象学报,1995,53(1):115~121.
- 6 魏凤英,曹鸿兴.长期预测的数学模型及应用.北京:气象出版社,1990.

NEW APPROACH TO DYNAMIC DATA MODELING AND ITS APPLICATION TO PRECIPITATION FORECASTING

Jia Xiaojing Cao Hongxing

(Chinese Academy of Meteorological Sciences, Beijing 100081)

Feng Guolin

(Department of Physical College, Yangzhou University, Yangzhou 225009)

Abstract

By use of an observed data series a new dynamic data modeling has been proposed. Taking a nonlinear ordinary differential equation which is retrieved from the data series based on the bilateral difference principle as a dynamic kernel, with the self-memorization principle a forecast model can be established, which is called the Data-based Mechanistic Self-memory Model (DAMSM). Some computing cases show that the forecasting accuracy of the DAMSM is quite satisfactory. An example of inter-annual precipitation prediction in summer in the Yangtze delta is given.

Key words: Self memory Time series Inter-annual precipitation prediction Flood season precipitation