第3卷 第1期 气象科学研究院院刊 Vol. 3, №1
1988年4月 JOURNAL OF ACADEMY OF METEOROLOGICAL SCIENCE Apr. 1988

基于预测均方误差为最小的梅雨期长度统计预报模型

陈孝源 俞善贤
（浙江省气象局气象科学研究所）

提 要

本文使用北半球500百帕月平均网格点高度场资料。应用预测均方误差 $MSEP_x$ 准则所建立的预测模型，对我省北部地区1953—1985年历年梅雨期长度进行预测。结果表明：用 $MSEP_x$ 准则建立的预测模型对异常年份的梅雨期长度进行预测，其预测精度有明显的提高。

一、引 言

应用一般常用的多元回归方法建立的预报模型对未来预报量进行预测，往往存在一定的局限性，其主要表现在对异常年份进行预测时，预测误差较大。这除了与输入变量的选取有关外，其重要的原因之一是在预测过程中，用固定的回归模型去预测变化着的系统状态，其预测误差必然会随着预测时间的增长而加大。另一方面，用多元统计方法所建立的预报模型中的因子，只是反映了历史样本的平均状况。

为了克服上述缺点，本文应用了由 D. M. Allen 于 1971 年提出的预测均方误差 $MSEP_x$ 准则。$MSEP_x$ 准则是着眼于预测平方的均值，对每个给定的预测点 x，在该处的预测偏差平方的均值就只与所选自变量有关。$MSEP_x$ 准则在于选择这样的自变量子集，使得在点 x 的预测偏差平方的均值达到最小。这样，针对不同的预测点，所建立的预测方程也是不同的。

二、预测均方误差 $MSEP_x$ 准则和计算方法[1]

1. 预测均方误差 $MSEP_x$ 准则

设有 M 个自变量 Z 和因变量 Y 的 N 个样本，当选定一组自变量时，相当于用某一矩阵去乘原设计矩阵，对未选入自变量亦相应于用一矩阵去乘原设计矩阵。例如，当选定自变量 x_1, x_2, x_3，则对选入变量和未选入变量相当于用以下矩阵 Q 和 R 右乘原设计矩阵，

本文1986年1月8日收到，1987年6月8日收到再改稿。

$$Q=\begin{pmatrix}1&0&0\\0&1&0\\0&0&1\\0&0&0\\\vdots&\vdots&\vdots\\0&0&0\end{pmatrix}\qquad R=\begin{pmatrix}0&0&\cdots&0\\0&0&\cdots&0\\0&0&\cdots&0\\1&0&\cdots&0\\0&1&\cdots&0\\\vdots&\vdots&&\vdots\\0&0&\cdots&1\end{pmatrix}$$

且满足条件 $R'Q=0$，则有

$$Q(Q'X'XQ)^{-1}Q'=(X'X)^{-1}-(X'X)^{-1}R(R'(X'X)^{-1}R)^{-1}R'(X'X)^{-1} \tag{1}$$

(证明参看文献[1]P112—113)

记

$$Z'=x'(X'X)^{-1}R(R'(X'X)^{-1}R)^{-1}R' \tag{2}$$

由(1)式易知，在点 x 处的预测值为

$$\hat{y}_Q=x'Q\hat{\beta}_Q=x'Q(Q'X'XQ)^{-1}Q'X'Y$$
$$=(x-Z)'\hat{\beta}$$

这里 $\hat{\beta}$ 为全模型中的 β 的最小二乘估计。为了计算在点 x 处预测偏差 $D=\hat{y}_Q-x'\beta$ 平方的均值。这时

$$Var(\hat{y}_Q)=(x-Z)'(X'X)^{-1}(x-Z)\sigma^2$$
$$=(x'S^{-1}x+Z'S^{-1}Z-2x'S^{-1}Z)\sigma^2 \tag{3}$$

其中 $S=X'X$ 。由于

$$Z'S^{-1}Z=x'S^{-1}R(R'S^{-1}R)^{-1}R'S^{-1}R(R'S^{-1}R)^{-1}R'S^{-1}x$$
$$=x'S^{-1}R(R'S^{-1}R)^{-1}R'S^{-1}x$$
$$=Z'S^{-1}x$$

因此

$$Var(\hat{y}_Q)=(x'S^{-1}x-Z'S^{-1}Z)\sigma^2$$

又因

$$E(\hat{y}_Q-x'\beta)=E(\hat{y}_Q)-x'\beta$$
$$=(x-Z)'\beta-x'\beta$$
$$=-Z'\beta$$

故最后得到在点 x 处预测偏差平方的均值为

$$\mathrm{MSEP}_X(Q)=E(D^2)=E((\hat{y}_Q-x'\beta)^2)$$
$$=Var(\hat{y}_Q)+(E(\hat{y}_Q-x'\beta))^2$$
$$=\sigma^2(x'S^{-1}x-Z'S^{-1}Z)+(Z'\beta)^2$$

以全模型中 σ^2 和 β 的最小二乘估计 $\hat{\sigma}^2$ 和 $\hat{\beta}$ 代替上式中的 σ^2 和 β，并去掉与 Q 无关的 $\hat{\sigma}^2x'S^{-1}x$ 得到 MSEP_X 的估计：

$$\widetilde{\mathrm{MSEP}}_X(Q)=(Z'\hat{\beta})^2-Z'S^{-1}Z\hat{\sigma}^2$$

2. 计算步骤

已知 M 个自变量 X 的 N 个样本

$$X=\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1M} \\ x_{21} & x_{22} & \cdots & x_{2M} \\ \vdots & \vdots & & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{NM} \end{pmatrix}$$

(1)将 X 矩阵中心化后，求其协方差矩阵得到

$$A=X'X=\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1M} \\ a_{21} & a_{22} & \cdots & a_{2M} \\ \vdots & \vdots & & \vdots \\ a_{M1} & a_{M2} & \cdots & a_{MM} \end{pmatrix}$$

其中

$$a_{ij}=\sum_{k=1}^{n}x_{ki}x_{kj}-n\overline{x_i}\cdot\overline{x_j} \qquad i,j=1,2,\cdots M$$

$$\overline{x_i}=\frac{1}{n}\sum_{k=1}^{n}x_{ki} \qquad i=1,2,\cdots M$$

然后求协方差矩阵 $X'X$ 的逆矩阵得到 $(X'X)^{-1}$。

(2)计算 Z，将(1)式两边左乘 x'，右乘 $X'X$，再利用(2)式得到

$$Z'=x'-x'Q(Q'X'XQ)^{-1}Q'(X'X) \tag{4}$$

我们可利用扫描运算来计算 Z。

扫描运算：设有 M 阶方阵 $A=(a_{ij})_{M\times M}$，以 a_{ii} 为枢轴的运算，记为 S_iA，定义为一个新方阵 $C=(c_{ij})_{M\times M}$，其中

$$\begin{aligned} c_{jk}&=a_{jk}-a_{ik}a_{ji}/a_{ii} && \text{当}j\neq i,\ k\neq i \\ c_{ji}&=-a_{ji}/a_{ii} && \text{当}j\neq i \\ c_{ij}&=a_{ij}/a_{ii} && \text{当}j\neq i \\ c_{ii}&=1/a_{ii} \end{aligned}$$

这里假定 $a_{ii}\neq 0$。

下面我们来计算 Z，设选入变量的下标为 $i_1,i_2,\cdots,ip$，未选入变量的下标为 $j_1,j_2,\cdots jr$。记

$$C=S_{i1}S_{i2}\cdots S_{ip}(X'X)$$

则有

$$Z_l=x_l-\sum_{k=1}^{p}x_{ik}C_{ikl} \qquad l=j_1,\cdots,jr \tag{5}$$

$$Z_{i1}=Z_{i2}=\cdots Z_{ip}=0$$

可以证明(4)式和(5)式求得的 Z 是相同的。计算因子所有组合的 $MSEP_x$ 值，以 $MSEP_x$ 为最小作为准则，选取预报方程 $\hat{y}_Q=(x-Z)'\hat{\beta}$

三、实　例

1. 自变量的选取

国内有关梅雨研究的论著表明[2]：梅雨期维持时间的长短与500百帕高度场的环流形势有关。为此，我们对北半球11月—3月500百帕月平均网格点高度场进行相关普查，对相关系数在0.4以上的相关区取其均值作为选入因子。这样从12—3月选入4个因子，它们分别为以下区域：

x_1：　10—15°N，50—65°E，12月。　x_2：　45—65°N，160°E—175°W，1月。

x_3：　15—35°N，20°W—25°E，2月。　x_6：　10—25°N，55—70°E，3月。

以同样的方法对11—3月西北太平洋水域37格点海水表层月平均温度进行相关普查，引进2个因子。它们分别为以下区域：

x_4：　35°N，150—155°E，11月。　x_5：　45°N，150—155°E，12月。

2. 计算结果与结果分析

作为例子，我们应用我省北部地区1953—1985年实况梅雨期长度资料和前述的6个自变量。为了与逐步回归方法进行比较，我们在预测某一年的值时，首先将该点的历史样本去掉，同时应用$MSEP_x$准则和逐步回归方法所建立的预报方程预测1953—1985年历年梅雨期长度。在逐步回归中我们对F检验选入因子和剔除因子的F值取为8.5，其预测值分别记为$\hat{y}(k)$，$y^*(k)$，预测误差记为$\hat{e}(k)$，$e^*(k)$，计算结果见表1。

表1　　基于$MSEP_x$准则和逐步回归方法的计算结果

k	年	y(k)	$\hat{y}(k)$	$y^*(k)$	$\hat{e}(k)$	$e^*(k)$	K	年	y(k)	$\hat{y}(k)$	$y^*(k)$	$\hat{e}(k)$	$e^*(k)$
1	1953	33	29	29	−4	−4	18	1970	32	31	16	−1	−16
2	1954	78	76	52	−2	−26	19	1971	31	36	35	5	4
3	1955	34	29	29	−5	−5	20	1972	13	15	15	2	2
4	1956	45	45	30	0	−15	21	1973	14	14	13	0	−1
5	1957	20	19	14	−1	−6	22	1974	38	29	29	−9	−9
6	1958	7	10	16	3	9	23	1975	30	29	29	−1	−1
7	1959	10	15	22	5	12	24	1976	30	35	44	5	14
8	1960	18	21	21	3	3	25	1977	19	11	11	−8	−8
9	1961	8	18	23	10	15	26	1978	11	12	11	1	0
10	1962	21	18	18	−3	−3	27	1979	20	25	25	5	5
11	1963	20	23	23	3	3	28	1980	41	34	34	−7	−7
12	1964	5	6	6	1	1	29	1981	9	2	−5	−7	−14
13	1965	14	6	5	−8	−9	30	1982	25	30	30	5	5
14	1966	31	27	28	−4	−3	31	1983	30	20	20	−10	−10
15	1967	20	25	25	5	5	32	1984	19	28	28	9	9
16	1968	19	27	26	8	7	33	1985	12	17	24	5	12
17	1969	19	15	15	−4	−4							

从应用以上两种不同准则所建立的预报模型的计算结果（表1）可以看出，应用

$MSEP_x$准则进行预测与实况梅雨期长度趋势基本一致。特别是对异常年份的梅雨期长度预测反映较灵敏，预测效果较好见图1。

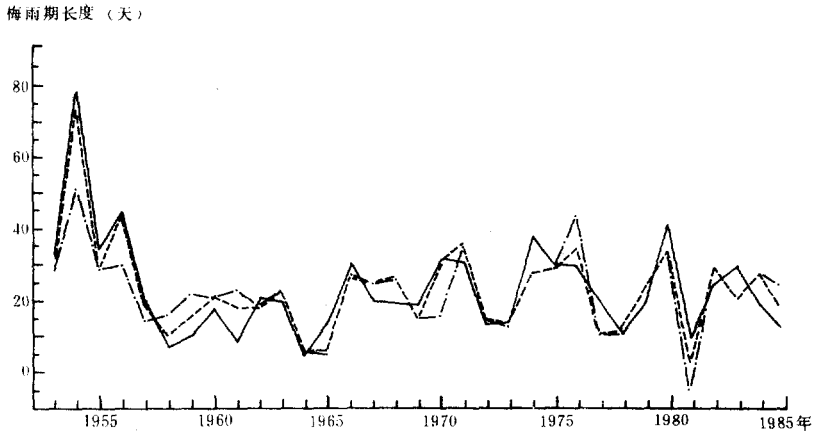

图1 基于$MSEP_x$准则和逐步回归方法的试验结果与实况曲线
——实况值---应用$MSEP_x$准则预报值－·－逐步回归方法预报值

应用$MSEP_x$准则一个明显的事实是在选模型中的自变量个数较少，其预测精度反而提高。例如1957年，应用$MSEP_x$准则所建立的预测模型只引进1个因子，其预测值为19天，实况值为20天，误差1天。而逐步回归方法所建立的预测模型引进x_1,x_2,x_3,x_4,x_5 5个因子，预测值为14天，误差6天。又如1959年，前者引进x_5 1个因子，预测值为15天，实况为10天，误差5天。后者引进和1957年同样的5个因子，预测值为22天，误差12天。特别是1954年，梅雨期持续时间达78天，基于$MSEP_x$准则的预测模型中引进x_1,x_2,x_3,x_4 4个因子，而逐步回归方法所建立的预测模型中引进x_1,x_2,x_5,x_6 4个因子，但预测精度有明显的差异，前者预测误差只有2天，后者为26天。

四、结 论

1. 应用$MSEP_x$准则所建立的预测模型来预测未来梅雨期长度，对于提高预测的趋势准确性和预测精度，有明显的效果，特别是对异常年份的预测效果尤佳。

2. 基于$MSEP_x$准则所建立的预测模型能较优地反映预测点的系统状态。在该模型中丢掉了那些对因变量的影响确实存在的自变量，由于这些自变量的引入，使得β_k难于估计，因此，在模型中丢弃它们后，使余下的自变量的回归系数估计的精度反而提高。这说明，把那些虽然与因变量有关，但关系很小或难于掌握的自变量从模型中去掉是有利的。

参 考 文 献

[1] 陈希孺、王松桂，近代实用回归，广西人民出版社，1984年。

[2] 陶诗言、徐淑英，夏季江淮流域持久性旱涝现象的环流特征，气象学报，**32**，1－10，1962。