最优子集回归模型在低纬高原雨季开始期预报中的应用试验①

张万诚　郑建萌　　解明恩
（云南省红河州气象局，蒙自　661100）　（云南省气象局）

提　要

将所有可能子集回归模型用于云南雨季开始的预报试验，选取有实际预报意义的500hPa、海温头年1～12月及同年1～3月网格点资料作为共同影响因子，分别与云南各站降雨量建立最优子集回归模型，对1998年、1999年的雨季开始期雨量进行独立预报表明：模型预报准确率优于逐步回归模型。从实况检验来看，效果较好，前期1～3月预报因子和模型适合低纬高原降水预测。

关键词：　最优子集　雨季开始　试验

引　言

目前用于短期气候预测的统计模型大多是基于残差平方和(RSS)准则下的逐步回归进行因子筛选，这种处理方法在业务预报中被广泛运用，但在实际应用中和理论上都发现有不足之处[1]。例如：在选入或剔除变量时的F检验，在理论上并不能以任何概率保证所选变量的显著性，因而不能认为涉及的F检验是正确的；由逐步回归方法所决定的变量子集，可能比包含相同变量子集的残差平方和要大得多；逐步回归所建立的模型不一定是全局最优。因而，势必造成预报模型模拟效果好，而实际预报效果差的现象。

本文将最优子集回归应用于低纬高原地区雨季开始期预报，所选因子为前期高空和海温网格点资料。并作了1998年、1999年两年的雨季开始期雨量预报检验。其结果表明：模型预报准确率优于逐步回归模型。从实况检验来看，效果较好，前期1～3月预报因子和模型适合低纬高原雨季开始期降水预测。

1　方法和预报原理

挑选回归方程的一个最彻底的办法就是将全部自变量按所有不同的排列组合与因变量建立全部可能的回归方程，从所有可能的回归方程中确定一个效果最好的子集回归，效果最好的子集回归称为最优子集回归。它与逐步回归的区别在于所建立的模型是全局最优的，逐步回归中置信度很高的回归方程不一定是最优的方程，而最优方程经常是显著性方程。选择自变量就是要确定哪一个子集回归效果最好，对于确定最优子集回归的方法，可先根据要求设计一个统计量S，一般S与子集回归方程的效果关系为S越小对应的回归方程效果越好，每一个子集回归都能算出它的S值。因此，最优子集回归的计算，就是按照一定的顺序求出一切可能子集回归的S值，然后，从其中确定最小值，如果：

$$S(X_{i1}\ X_{i2}\ \cdots\ X_{ik}) = \min S$$

则$S(X_{i1}\ X_{i2}\ \cdots\ X_{ik})$对应的子集回归方程

$$Y = \beta_0 + \beta_{i1}X_{i1} + \beta_{i2}X_{i2} + \cdots + \beta_{ik}X_{ik}$$

就是最优子集回归方程。对于回归模型识别准则，可采用平均残差平方和准则（S_k）、预报平方和准则（press）、AIC、BIC和C_p准

① 本文得到国家“九五”重中之重科技攻关项目云南专题“云南短期气候预测系统的研究（96—908—05—08）”资助。

则[2]。本文采用双评分准则[3]，其定义为：

$$CSC = S_1 + S_2 \tag{1}$$

$$S_1 = nR^2 = n\left(1 - \frac{Q_k}{Q_y}\right) \tag{2}$$

$$Q_k = \frac{1}{n}\sum_{i=1}^{n}(y(t) - \hat{y}(t))^2$$

$$Q_y = \frac{1}{n}\sum_{i=1}^{n}(y(t) - \bar{y}(t))^2$$

$$S_2 = 2I = 2[\sum_{i=1}^{I}\sum_{j=1}^{I}n_{ij}\ln n_{ij} + n\ln n - (\sum_{i=1}^{I}n_i\ln n_i + \sum_{j=1}^{I}n_j\ln n_j)] \tag{3}$$

式中 S_1 为数量评分，即为精评分，S_2 为趋势评分，称为粗评分，n 为样本长度，Q_k 为残差平方和，Q_y 为模型的总离差平方和。I 为预报趋势类别数，n_{ij} 为 i 类事件与 j 类估计事件的列联表中的个数，其中

$$n_j = \sum_{i=1}^{I}n_{ij} \qquad n_i = \sum_{j=1}^{I}n_{ij}$$

双评分准则旨在使模型拟合的精度更好，趋势亦准。显然，当CSC达最大时相应的回归模型为最优，用CSC达到最大为准则选取最优子集回归。

2 预报试验结果

2.1 隔年前期因子预报雨季雨量

2.1.1 对5月雨量的预测

云南雨季的早迟直接关系到工农业生产特别是大春的栽种，一般说来5月雨量多的年份，雨季也来得早；5月雨量少的年份，雨季也来得晚。因此每年5月雨量多少的预报一直是各级领导十分关心的问题。取云南15个地州站1952～1997年5月雨量为预报对象，预报因子为1951～1997年1～12月500hPa高度，北太平洋海温网格点资料为共同影响因子，按前述方法进行建模。

显然，对于500hPa月平均网格点资料因子数有576个，海温为286个，要将这些格点进行最优子集回归，目前台站上运用的计算机是不能实现的。据此，可采取以下方法进行挑选。

以昆明站为例：第一步，将昆明5月雨量和1～12月500hPa高度、海温资料进行相关分析，提取对昆明5月雨量影响较密切的因子105个，将这些因子再与昆明5月雨量进行逐步回归，筛选出 m 个变量（$m \leqslant 15$），本文共15个；第二步，计算 m 个变量与昆明5月雨量的全部可能回归找出最优子集，但应注意，如果第二步所得的最优子集恰好是第一步所选出来的 m 个变量的全模型，这时应放宽第一步的筛选条件，让第一步所筛选的变量数大于第二步所找出的最优子集的变量数。这是因为当第二步的最优子集是第一步的全模型时，表示第一步所选的局部最优子集要作为全局最优子集，而逐步回归方法正好不能保证永远达全局最优。

同理可得玉溪、大理、蒙自等15个站1998年5月雨量（见表1）。表中 R 是复相关系数、CSC 是双评分准则，因子数（n）表示模型的入选因子数。评分标准为预报值与实况值距平百分率符号相同或其中之一为零但两者误差绝对值≤15为预报正确，反之错误（以下同）。从表中可看出，最优子集回归模型预报准确率为66.7%，逐步回归模型的预报准确率为46.7%，这表明采用最优子集回归模型，虽入选因子数较少，但模型的预测精度并没有降低。

表1 1998年15站5月雨量距平百分率实况及预报值（%）

站名	实况	最优模型					逐步回归			
		预报	R	CSC	n	评定	预报	R	n	评定
蒙自	−40	−3	0.87	97.4	5	正确	−29	0.93	10	正确
昆明	−19	−42	0.98	101.5	11	正确	−47	0.98	13	正确
大理	14	−84	0.95	83.7	9	错误	−90	0.97	13	错误
沾益	−55	−9	0.88	77.8	5	正确	21	0.97	13	错误
昭通	−39	−50	0.96	77.3	11	正确	23	0.92	7	错误
玉溪	−53	−88	0.93	81.3	8	正确	−80	0.94	9	正确
保山	13	−85	0.92	66.5	8	错误	−90	0.94	11	错误
丽江	−10	122	0.77	42.6	11	错误	21	0.77	11	错误
东川	−68	−85	0.94	76.7	8	正确	−72	0.96	10	正确
文山	−52	−12	0.77	46.2	4	正确	−5	0.85	13	正确
楚雄	−61	−62	0.95	89.9	9	正确	−51	0.95	10	正确
潞西	36	−73	0.96	86.2	11	错误	−53	0.97	12	错误
景洪	−27	−8	0.92	73.5	8	正确	0	0.94	10	错误
临沧	16	−66	0.93	76.0	6	错误	−90	0.99	13	错误
思茅	−20	−91	0.95	91.0	8	正确	−89	0.96	10	正确

2.2 同年前期因子预报雨季雨量

2.2.1 对5月雨量的预测

取1951～1997年1～3月500hPa，海温网格点作为共同影响因子，预报对象仍为云

南5月雨量资料，样本长度为1951～1997。预报步骤同上，仍分两步进行。表2为1998年5月雨量。从表中可知，最优子集回归模型预报准确率为80.0%，逐步回归模型的预报准确率为60.0%。

表2 1998年15站5月雨量实况及预报值距平百分率（%）

站名	实况	最优模型					逐步回归			
		预报	R	CSC	n	评定	预报	R	n	评定
蒙自	−40	−44	0.91	78.7	9	正确	−70	0.93	11	正确
昆明	−19	−82	0.71	55.4	5	正确	−95	0.94	15	正确
大理	14	20	0.92	78.4	9	正确	−18	0.86	13	错误
沾益	−55	−53	0.88	83.4	13	正确	−60	0.87	12	正确
昭通	−39	−10	0.79	51.3	7	正确	−10	0.79	7	正确
玉溪	−53	2	0.81	60.7	11	错误	33	0.83	13	错误
保山	13	0	0.83	74.3	8	正确	−10	0.84	10	错误
丽江	−10	−30	0.94	105.3	12	正确	−81	0.94	12	正确
东川	−68	−44	0.91	73.5	8	正确	−30	0.93	9	正确
文山	−52	−9	0.91	65.4	9	正确	26	0.99	13	错误
楚雄	−61	−83	0.91	58.3	11	正确	−71	0.93	13	正确
潞西	36	−9	0.77	53.0	5	错误	−32	0.82	7	错误
景洪	−27	−39	0.95	63.1	8	正确	−39	0.95	8	正确
临沧	16	−32	0.93	83.9	9	错误	−32	0.93	9	错误
思茅	−20	−92	0.93	104.0	9	正确	−90	0.96	13	正确

上述预报结果检验表明，同年前期预报因子效果高于隔年预报因子，以下将用同年前期预报因子作为试验。

2.3.2 对5～6月雨量的预测

取1951～1997年1～3月500hPa高度、海温网格点作为共同影响因子，预报对象为云南5～6月雨量资料。

表3为云南1998年5～6月雨量，从表中可知，最优子集回归模型预报准确率为60.0%，逐步回归模型的预报准确率为46.7%。

表3 1998年15站5～6月雨量实况及预报值距平百分率（%）

站名	实况	最优模型					逐步回归			
		预报	R	CSC	n	评定	预报	R	n	评定
蒙自	−5	−33	0.79	70.1	5	正确	−42	0.89	10	正确
昆明	110	−24	0.80	42.0	5	错误	−30	0.92	9	错误
大理	52	−39	0.86	82.5	8	错误	−45	0.88	9	错误
沾益	−44	−73	0.93	91.4	12	正确	−73	0.93	12	正确
昭通	3	3	0.85	65.0	8	正确	−1	0.9	10	错误
玉溪	36	12	0.93	91.8	10	正确	1	0.92	7	正确
保山	−36	−4	0.76	70.0	5	正确	−4	0.76	5	正确
丽江	27	4	0.88	73.0	7	正确	−12	0.91	10	正确
东川	−30	−79	0.83	48.5	6	正确	−93	0.85	7	正确
文山	18	−4	0.90	60.0	8	错误	−2	0.90	10	错误
楚雄	28	−21	0.92	72.9	11	错误	−37	0.93	12	错误
潞西	4	−8	0.87	57.3	7	错误	−37	0.94	13	错误
景洪	1	21	0.90	59.0	7	正确	23	0.91	9	正确
临沧	−30	−25	0.94	72.0	9	正确	−25	0.95	11	正确
思茅	−15	16	0.88	60.3	7	错误	33	0.91	9	错误

上述预报结果于1998年4月19日全省汛期会商会上作了讨论，根据5月雨量、5～6月雨量的分布趋势，得到了一个全省雨季开始期正常偏晚，雨水偏少的结论，与实况相比，除滇西部分站等预报错外，其余降水趋势分布预报较好；1999年4月19日，运用最优子集模型作5月雨量、5～6月雨量的预测，并认为1999年全省雨季开始期偏早，实况是除滇西北部分站等预报错外，大部分站预报正确，5月雨量、5～6月雨量的预报准确率分别为66.7%、66.7%。

综上所述，将高空环流因子与海温因子作雨季降水预报，较好地考虑了海气相互作用，其因子选取有一定物理意义；对雨季开始预报，以同年前期1～3月预报因子最好。

3 讨论与小结

（1）最优子集回归的优点在于能选到全局最优的子集，而各种逐步算法，虽然速度快，仅能做到局部最优，而且经常也是全局最优，但总不能保证每次达到全局最优，从本文分析中也证实了这一点。而且逐步算法的最优子集方程与F临界值的大小有关[2]。

（2）在作云南雨季降水预报时，最优子集回归模型对低纬高原地区的降水预报有较好的预报能力，从实际运用来看，效果较好，有关更广泛的应用，有待今后作进一步的应用研究。

（3）由于雨量受地形等因素影响极大，特别是在低纬高原地区显得更突出。因此，在作雨量预报时，因子的选取最好能综合考虑高空和海温形势或其他外界因子，如较好的考虑海气相互作用等。当然预报模型的选取与因子的选择一样是至关重要的。

参考文献

1 俞善贤，汪锋. 试用最优子集与岭迹分析相结合的方法确定回归方程. 大气科学，1998，12：382～388.

2 施能. 气象科研与预报中的多元分析方法. 北京：气象出版社，1995，89～91.

3 魏凤英，曹鸿兴. 长期预测的数学模型及其应用. 北京：气象出版社，1990，9～90.

Application of Optimum Subset Regression to Forecast the Beginning of the Rainy Season in the Low Latitude Plateau

Zhang Wancheng Zheng Jianmeng
(Honghe Meteorological Office, Mengzi, Yunnan Province, 661100)
Xie Ming'en
(Yunnan Meteorological Bureau)

Abstract

All potential subset regression models were used to forecast experiment of the beginning of the rainy season in Yunnan province. The grid point data of 500 hPa and SST during the periods from Jan. to Dec. next before this year and from Jan. to Mar. this year were selected for the influent factors. The factors combined with rainfall at weather stations in Yunnan province were used to built the optimum subset regression model. The model was applied to forecast the rainfall of rainy season in 1998 and 1999. The model forecast accurate is better than the stepwise regression.

Key Words: optimum subset regression beginning of rainy season low latitude rainfall forecast