第 2 卷 第 3 期　　大 气 科 学　　Vol. 2, No. 3
1978 年 9 月　　SCIENTIA ATMOSPHERICA SINICA　　Sept., 1978

逻辑组合因子的格筛选法

江西省气象台　江西师范学院数学系

在一般的事件概率回归估计予报方法[1]中，所有可能预报因子都是线性的．可是，实际经验表明，仅仅考虑预报量与预报因子之间的线性关系往往还不能确切反映客观规律．因此，寻求非线性因子对于提高预报效果是有意义的．利用布尔结构可以给出对非线性的逻辑代数组合因子进行筛选的方法，我们称为逻辑组合因子的格筛选法．受这一思路启发，我们开展了非线性事件概率回归估计的试验，效果比通常的事件概率回归好．

一、逻辑代数组合因子

转换成 (0, 1) 资料后，每个预报因子都可看成一个布尔变量．由布尔代数的三种基本运算："非(—)"、"加"(∨)"、"乘(∧)" 结合二变量 x_i, x_j 而成的每个布尔函数都描述了因子 x_i 和 x_j 所生成的一个组合因子，我们称这些组合因子为逻辑(代数)组合因子．每两个因子就可派生 16 个逻辑组合因子．特别，其中一个布尔函数

$$(\bar{X}_i \wedge X_j) \vee (X_i \wedge \bar{X}_j)$$

又叫做 x_i 和 x_j 的对称差"，其文氏图如右．

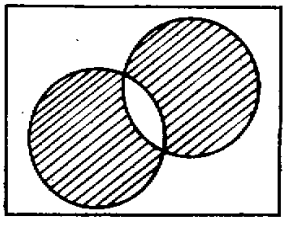

图 1

它也可以看成是布尔变量间类似于"加"、"乘"那样的又一种二元运算 (记为△) 的结果，其运算法则是：

$$0\triangle 0 = 0, 0\triangle 1 = 1, 1\triangle 0 = 1, 1\triangle 1 = 0.$$

为了方便，将上述四种运算的结果依次称为 x_i 的"逻辑非"、x_i 与 x_j 的"逻辑和"、"逻辑积"、"对称差"．其它组合因子也都描述了 x_i 与 x_j 的各种不同的配置情况．预报经验说明，在气象要素的不同配置下，可以发生完全不同的天气过程．逻辑组合因子不仅仅考虑了可能预报因子与预报量之间的线性关系，而且进一步考虑了非线性组合因子与预报量之间的关系以及组合因子相互间的联系．这就是逻辑组合因子格筛选法与通常线性回归的主要区别．

在有 n 个可能预报因子的情形，所有逻辑组合因子的个数就是变量布尔函数的数目 2^{2^n}．显然，企图逐个地检查这些组合因子与预报量的关系，即使 n 是不大的数，也是很难办到的．因此，必须确定一种切实可行的计算过程，使得对那些与预报量关系密切的组合因子的搜索范围尽可能地缩小．

"包含" 关系是布尔函数之间的半序关系．由于这一半序关系的存在，具有上述布尔

* 1977 年 4 月 5 日收到．

运算的 2^{2^n} 个组合因子的集合构成一个叫做**布尔格**的代数系[3]. 我们可以利用布尔函数间的这一半序关系来安排逻辑组合因子的筛选. 由于它是按组合因子在布尔格中的位置逐步进行的[2],因而可叫做"格筛选".

二、逻辑组合因子的格筛选过程

希望能用较小的计算量从所要 2^{2^n} 个组合因子中选出对预报量有显著作用的组合因子,并且体现出组合因子之间相互影响、相互制约的作用,比方说,剔除相形见绌的组合因子. 为此,以逐步回归分析方法,作为格筛选的工具.

我们安排格筛选的具体步骤如下:

1. 对 n 个可能预报因子用逐步回归分析方法挑选因子,建立第一个事件概率回归估计的预报方程. 这就是通常的线性回归.

2. 对于那些没有选上的可能预报因子,不是简单地丢掉,而是作出它们两两的对称差,将它们和进入第一个预报方程的因子一道进行第二次逐步回归,得到第二个预报方程.

3. 对于进入第二个预报方程的诸因子以及它们的逻辑非,两两生成逻辑和与逻辑积,即第二个预报方程中每两个因子 Z_i 与 Z_j 派生出八个因子,它们是: $Z_i \vee Z_j$、$\bar{Z}_i \vee Z_j$、$Z_i \vee \bar{Z}_j$、$\bar{Z}_i \vee \bar{Z}_j$、$Z_i \wedge Z_j$、$\bar{Z}_i \wedge Z_j$、$Z_i \wedge \bar{Z}_j$、$\bar{Z}_i \wedge \bar{Z}_j$. 将它们和进入第二个预报方程的因子一道,进行第三次逐步回归分析,得到第三个预报方程.

4. 对进入第三个预报方程的诸因子(这里可以含有原始因子、对称差、逻辑和、逻辑积)重复上一步的做法,生成高一级的组合因子,再用第四次逐步回归进行挑选,建立第四个预报方程.

5. 如此反复地生成和挑选,得出一个又一个预报方程. 直到后一预报方程的复相关系数比前一预报方程的更低,或后一预报方程的标准误差比前一预报方程的更大,那就说明再做下去所得的方程效果不会更好,于是计算停止. 选定所得的一串报预方程中效果最好(一般就是复相关系数最高)的一个作为最后确定的预报方程.

6. 按历史概括率最高的原则确定判据(当估计概率超过 0 与 1 的界限之处,则取小于零的值为零,大于 1 的值为 1).

由于每一批(组合)因子生成高一级的逻辑和、逻辑积时,必须两两组合起来(第二步生成对称差也是这样),因而每一步参加逐步回归挑选的组合因子的数量一般是较多的. 这样,不仅可能超过电子计算机的容量,而且实践证明,当参加筛选的因子数超过样本数时,逐步回归将无法进行. 我们采取以下办法克服这一困难: 将每次生成的所有逻辑和、逻辑积(在第二步是对称差)按它们对预报量的相关系数绝对值的大小排列,取排在前面的

$$\frac{m}{K} - C$$

个和上一步进入预报方程的 C 个因子一道进行逐步回归,其余的舍去. 这里 m 表示样本数. 于是,总可保证每次参加逐步回归的因子数为样本数的 K 分之一. K 视样本数的大

小而定，例如 $m = 100$，K 取 3 或 4；$m < 50$，K 取 2.

上述计算序骤可用框图描述如下：

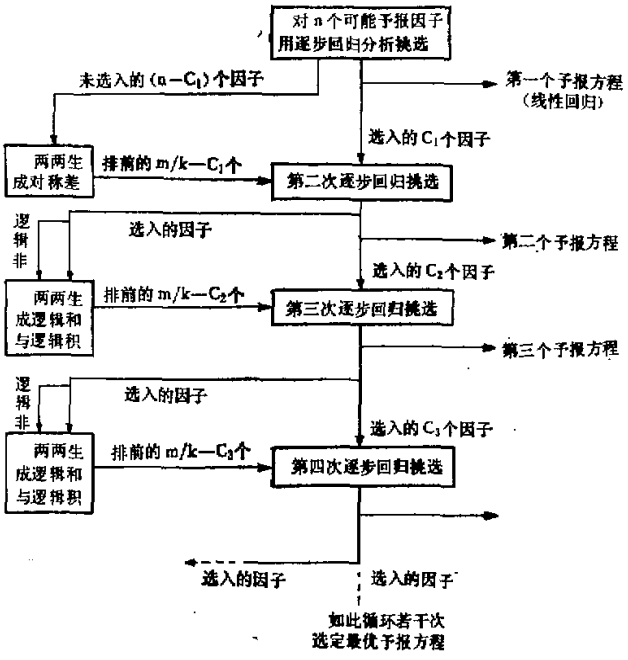

图　2

三、试 验 实 例

江西 5—7 月连续四天以上的暴雨过程，雨量大而集中，容易造成大范围洪涝，是我省主要灾害性天气之一．在四到六天前适时提供连续暴雨过程的趋势预报，对加强防汛工作是有益的．为此我们采用前述方法进行试作．

1. 500 毫巴环流起始特征：

通过普查，发现在连续暴雨形成前期，500 毫巴上东亚东部中高纬度上有一个较为稳定的北脊南槽或阻塞形势建立，以此环流特征作为连续性暴雨中期预报起始条件．

(1) 500 毫巴 47 区 401 站高度减 31 区 088 站高度值、50 区 953 站高度减 24 区 843 站高度值、54 区 102 站高度减 30 区 054 站高度值．这三者中必须连续三天差值有一个 $\leqslant 6$ 位势什米．

(2) 500 毫巴 50 区 953 站高度减 24 区 843 站高度值、54 区 102 站高度值减 30 区 054 站高度值．这两个高度差值必须连续三天有一个 $\geqslant 15$ 位势什米．

上述两条中任何一条满足，就符合东亚北脊南槽型．

2. 可能预报因子的挑选和预报方程的建立：

根据过去预报这类连续性暴雨的实践经验，注意这类环流形势中冷暖空气活动，锋区

和付热带高压强弱及其环流形势变化等特点，选取了若干因子．在大量原始资料制作点聚图的基础上，以尽可能减少漏报、空报，提高因子与预报量的相关程度为准，进行 0.1 分档．逐一计算诸因子与预报量的单相关系数，并考虑到因子的物理意义，确定以下 19 个因子作为参加逻辑组合因子的格筛选的原始的可能预报因子：

	因 子 意 义	分 档 标 准
x_1	酒泉 500 毫巴 20 时高度值	[576, 584]
x_2	20 时 500 毫巴上 40°—45°N; 90°—105°E 区域中最高温度值与 40°—45°N; 105°—120°E 区域中最低温度值的差值	≥4
x_3	南昌 500 毫巴 20 时气温	[−4, −2]
x_4	500 毫巴 20 时福州高度值减贵阳高度值	[−1, 3]
x_5	500 毫巴 20 时太平洋付高脊线与 120°E 交点所在纬度	<20°
x_6	汕头 500 毫巴 20 时高度值	[584, 589]
x_7	广州 500 毫巴 20 时高度值	[584, 589]
x_8	700 毫巴 20 时 51777 站减 51076 站气温差值	[0, 13]
x_9	700 毫巴 20 时 52、53 区中各站最大高度值	≥12
x_{10}	南昌 700 毫巴 20 时气温	≤11
x_{11}	700 毫巴 20 时福州高度值减兰州与济南高度差	≤10
x_{12}	700 毫巴 20 时芷江气温	≤10
x_{13}	南宁 700 毫巴 20 时露点 24 小时的差值	<−1
x_{14}	庐山 08 时地面气温	[10, 18]
x_{15}	衡山与庐山风向风速编码和	≥15
x_{16}	重庆 08 时海平面气压	≥1006.4
x_{17}	兰州 08 时海平面气压	[1005.5, 1011.4]
x_{18}	南昌 08 时海平面气压	[1006.1, 1009.7]
x_{19}	福州 08 时海平面气压	[1006.0, 1010.9]

诸因子观测值在分档标准栏所列范围之内时，转换为 1；否则，转换为 0．

用电子计算机按前述步骤进行逻辑组合因子格筛选，并确定第三个方程

$$y = -0.098 + 0.233x_7 + 0.290x_{13} + 0.254x_{16} - 0.196(x_{10}\Delta x_{18}) + 0.229(x_1 \wedge x_8) + 0.331(x_8 \wedge x_{11}) y \geqslant 0.7,$$

为预报方程．按历史概括率最高的原则确定临界值为 0.7，若则未来 4—6 天开始有一次 4 天或 4 天以上连续性暴雨出现．

3. 历史检验、回报和使用情况：对 59—72 年（68 年除外）5 月 20 日至 7 月 15 日进行历史检验，符合起始场条件共 335 天，逐日计算 y 值，如连续数日符合临界值，则按一次过程起报．结果共报 39 次过程，报对 26 次连续暴雨，13 次空报；历史上共有连续暴雨 29 次，仅有 3 次未报出．以 73 年汛期作回报，共预报 3 次过程，报对 2 次，空报 1 次．

近三年使用于实际预报中，取得了较好的效果：

年 度	预 报 次 数	出现连续暴雨次数	空 报	漏 报
1974	9	7	2	0
1975	3	2	1	0
1976	4	4	0	0

实践表明，用本文所述方法作连续暴雨过程的预报，效果比较稳定，具有一定参考价值.

四、讨　论

1. 格筛选与通常作一次逐步回归主要区别是：(1) 逐步回归分析总是在给定的一组可能预报因子中考虑引进或剔除，优选一个线性的预报方程. 而格筛选却不限于原始因子，它能不断地提供新的组合因子，逐次地挑选，逐次地建立含有逻辑组合因子的非线性预报方程；(2) 通常逐步回归对没有选上的可能预报因子就不再考虑，而格筛选法将进一步考虑它们两两的对称差. 实践证明，这些对称差中往往可能提供显著信息.

2. 在格筛选的实际运用中，不一定要选取复相关系数最高的那个方程. 为了使用方便，可以选取复相关系数接近最高，而所含的组合因子又比较简单的方程(往往是第三或第四个方程). 经验表明，有时复相关系数提高很小，预报效果几乎一样，而组合因子却可以变得复杂得多. 没有必要过分追求高级组合因子对预报量的影响，这不仅是因为描述高级组合因子的布尔函数一般都很复杂，而且这种组合因子的物理意义也不易弄清楚.

3. 由于将观测资料转换成 (0, 1) 特征资料，使得无法作真实纪录或无法用数字表达的天气资料都可应用，遗漏的、不完整的资料常常可被有规则地处理，预报方程的使用也方便. 关于资料转换的标准，应以尽可能减少漏报、空报，提高因子与预报量的相关程度为原则. 对于连续暴雨、寒潮等灾害性天气的预报，特别应当以尽可能少漏报或不漏报为转换标准.

4. 在逐次生成逻辑和、逻辑积的过程中，可能出现某两个组合因子的逻辑和(积)是恒等于 1(0) 的布尔函数. 显然，这样的组合因子对预报是毫无意义的，可以认为它与预报量的相关系数为零，而不参加筛选，以免因其均方差为零在计算相关系数时发生溢出.

5. 格筛选法和一般用 (0, 1) 特征资料的其它方法一样，当因子的真实资料接近转换标准时，预报往往难以决断.

6. 格筛选只是对已提出的可能预报因子的一种数学处理，根本的问题仍在于加强对预报对象的物理分析和天气气候分析，以期选出物理意义比较明确的可能预报因子. 此外，有一些意义较明确的物理组合因子，也可当成原始因子参加挑选和组合.

参　考　资　料

[1] 王宗皓、李麦村等编著，天气预报中的概率统计方法，第九章，1974年，科学出版社。
[2] R. G. Miller, *SLAM*, A Screeing Lattice Algorithm for Non-linear Regression Estimation of Event Probabilities.
[3] Bikhoff and Mac Lane, A survey of Modern Algebra.