*

最优子集回归在福建热带气旋年频数预测中的应用

高建芸 许金镜

(福建省气候中心, 福州)

摘 要

用相关分析方法普查了影响福建省热带气旋(TC)年频数与前期北半球500hPa和100hPa月平均位势高度场、北太平洋海温场以及500hPa月环流特征量的关系, 分析影响福建TC年频数的前期大气环流及海温场特征, 寻找与福建TC年频数关系密切的预测因子, 采用"两段筛选"的思路, 选用逐步回归筛选出M个(10个左右)预报因子, 再用最优子集回归建立预报模型, 其效果较为理想。

关键词: 最优子集回归, 热带气旋年频数, 预测。

一、资料和预报因子的选取

1. 资料

选取1951~1996年北半球500hPa月平均高度场(5°×10°网格点)、北半球100hPa月平均高度场(10°×10°网格点)、太平洋月平均海温场(286个5°×5°网格点)和500hPa月环流特征量作为预报因子的分析素材, 取1951~1996年影响福建热带气旋(简称TC)年频数为预报对象。

2. 预报因子的选取

对上一年1月至当年4月每个因子场进行相关普查, 寻找预报福建TC年频数的预报因子。所选因子满足以下条件:

(1) 相关系数大于0.29, 达信度标准0.05。

(2) 选取三个以上格点的相关系数大于0.29连成一片的区域为显著相关区。若相关区格点数较多, 以相关系数较大且连成一片的区域为相关区, 相关区内所有格点值平均构

* 本文于1998年8月收到。

造出一个预报因子序列。

二、方法简述

鉴于逐步回归用于建立预测模型时存在一些缺点和问题: (1) 逐步回归可视为按选取方差贡献大的因子为准则进行的一种子集回归, 因此, 所建立的模型不一定是全局最优。(2) F 临界值不好确定。(3) 回归方程的检验流于形式。而最优子集回归采取合理途径、穷尽所有预报因子的搭配, 选择回归效果最好的子集回归, 确保筛选出的预报因子组合是最优。这也是最优子集回归正逐步替代逐步回归的原因。

本文采用双评分准则作为模式识别准则[1], 定义:

$$CSC = S1 + S2 \tag{1}$$

其中:

$$S1 = (N - K)(1 - Q_K / Q_Y)$$

$$S2 = 2I = 2[\sum_{i=1}^{G}\sum_{i=1}^{G} n_{ij} 1nn_j + N 1nN - (\sum_{i=1}^{G} n_{i\cdot} 1nn_{i\cdot} + \sum_{i=1}^{G} n_{\cdot j} 1nn_{\cdot j})]$$

式中 $S1$ 为数量评分, 即为精评分, $S2$ 为趋势评分, 即为粗评分。N 为样本长度, K 为统计模式中变量个数, Q_K 为模型的残差平方和, Q_Y 为模型总离差平方和。由此可见, 双评分准则旨在使模型拟合的精度越好, 趋势亦准。

用双评分准则作为模式判别准则的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 CSC 值, 然后确定最大值, 如果 $CSC(X_{i1}X_{i2}\cdots X_{ik}) = \max$, 则其所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。

本文采用“两段筛选”的思路, 选用逐步回归控制 F 检验值, 筛选出 M 个 (10 个左右) 预报因子, 再计算 M 个变量的全部可能回归寻找最优子集, 建立预报模型, 实验证明, 最优子集回归的预报效果比逐步回归好。

对 TC 预测评分采用平均绝对误差 (E) 和预测效率 (RF)[2], 其中:

$$E = 1/N_0 \sum | X_i^f - X_i^0 | \tag{2}$$

$$RF = (N + f_1 \times N_1 + f_2 \times N_2)/(N_0 + f_1 \times N_1 + f_2 \times N_2) \times 100\% \tag{3}$$

(2) 式中 E 为平均绝对误差，X_i^f 表示第次的预测值，X_i^0 表示第次预测相应的实际值，N_0 为预测总次数，显然 E 越小越好。(3) 式中 RF 为预测效率，N_0 为预测正确（指预测与实况的绝对误差小于 0.5 个）的次数。N_1 和 f_1 分别为一级异常（指预测与实况的绝对误差大于等于 0.5 个且小于 1.5 个）的样本数和权重系数（一级异常气候概率的倒数），N_2 和 f_2 分别为二级异常（指预测与实况的绝对误差大于等于 1.5 个且小于 2.5 个）的样本数和权重系数（二级异常气候概率的倒数）。预测效率（RF）在此 0~ 100% 之间，且越大越好。

三、影响福建TC 年频数的因子特征分析

热带气旋发生于热带海洋上，其生命史的大部分又生存于海洋，因此，海温与TC 的生成、强度及移动路径有相当密切的关系。同时，热带气旋作为一种天气系统还受到海气相互作用的影响，受到大气环流的直接制约和引导。因此，海温与大气环流是影响TC 年频数异常的最重要的前期因子。本文利用相关分析方法普查了影响福建 TC 年频数与前期北半球500hPa 和 100hPa 月平均位势高度场、北太平洋海温场以及 500hPa 月环流特征量的关系，为了综合评估每个月各种因子场与预测量的相关程度，定义了以下高相关因子场的标准[1]：

(1) 因子场的最高相关系数 Rmax 绝对值大；

(2) 因子场的高相关因子（$r_i \geq r_c, r_c = 0.29$，显著性信度标准 $\alpha = 0.05$）总数 n 较大。

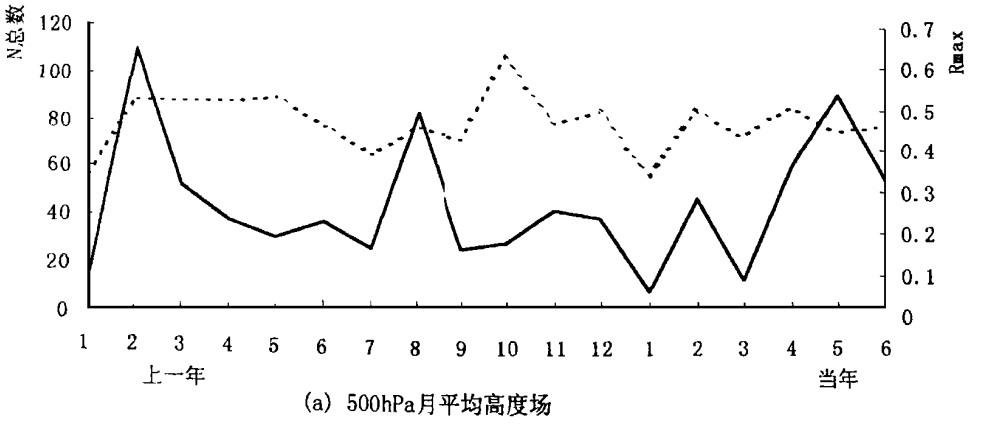
以下分别讨论影响福建TC 年频数的前期因子 500hPa、100hPa 月平均高度场和北太平洋海温场的特征。

1. 前期 500hPa 高度场

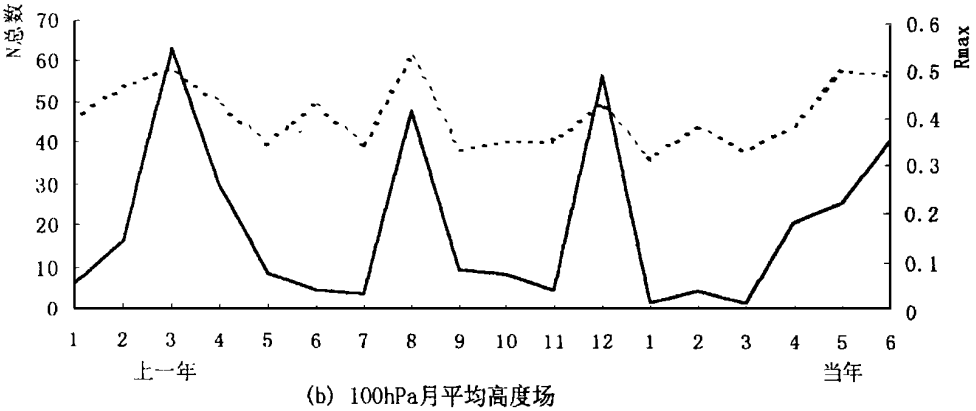
图1 (a) 为前一年 1 月至当年 6 月 500hPa 月平均高度场与影响福建TC 年频率的最高相关系数 Rmax 值和高相关因子总数 n 的时间分布图。由图可见，上一年 2 月为最高相关月份，高相关因子总数 n 为各月之冠，最高相关系数 Rmax 值仅次于 10 月，当年5 月和上一年 8 月为较高相关月份，Rmax 与 n 两条曲线都出现较大值。

图2 (a) 为前一年 2 月 500hPa 月平均高度场与影响福建TC 年频数的相关场，图中有两个较大的正相关区和一个较大的负相关区，其中最高、最大的相关区位于 20~ 40°N，10 ~ 110°E 的阿拉伯半岛和印度半岛上，高相关中心位于伊朗高原，中心值为 0.52，另一正相关区位于格陵兰岛西部的巴芬湾；最大的负相关区位于鄂霍次克海和堪察加半岛附近地区，最大相关系数为– 0.42。可见当前一年 2 月位于格陵兰岛西部的巴芬湾的极涡偏弱(强)，阿拉伯半岛和印度半岛 500hPa 月平均高度场出现正（负）偏差，同时鄂霍次克海和堪查加半岛附近地区上出现负（正）偏差，则次年影响福建TC 年频数偏多（少）。

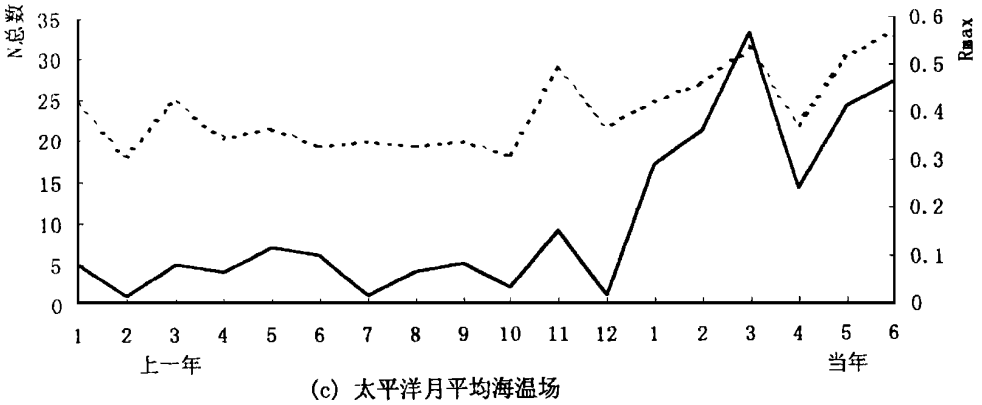
前一年8 月在高纬度地区东西伯利亚海和格陵兰岛北部各有一个较大的正相关区，而



(a) 500hPa月平均高度场



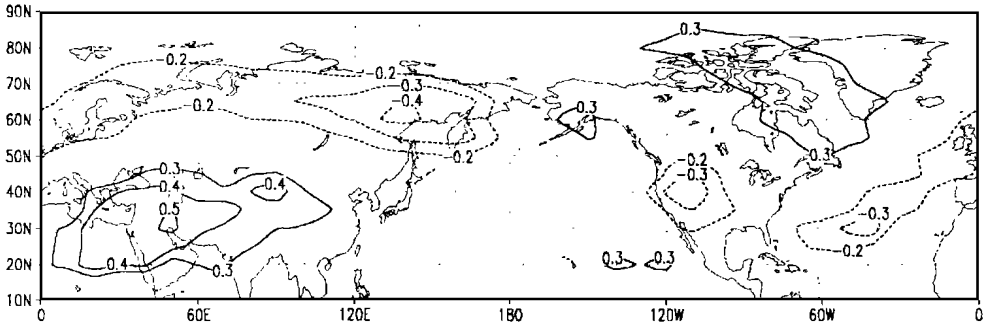
(b) 100hPa月平均高度场



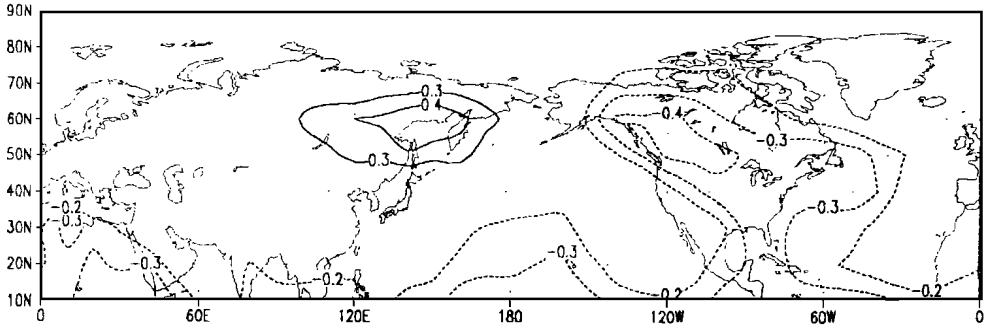
(c) 太平洋月平均海温场

图 1 前期大气环流和海温因子与影响福建省热带气旋年频数相关的时间分布

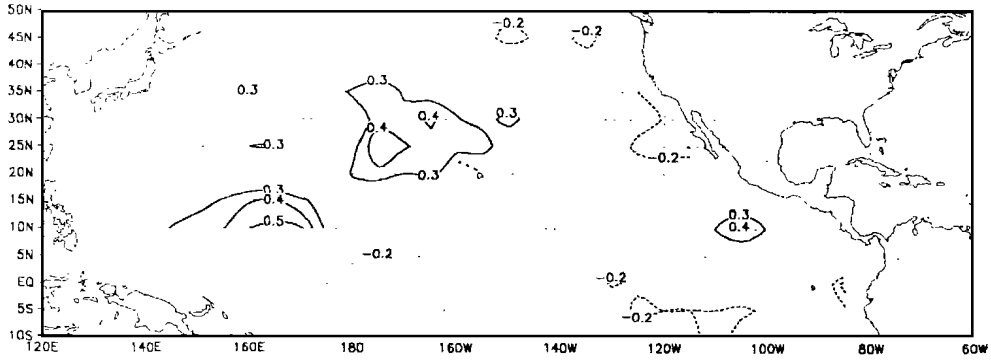
(实线为N，虚线为R_{max})



(a) 前一年 2 月 500hPa 月平均高度场



(b) 前一年 3 月 100hPa 月平均高度场



(c) 当年 3 月海温场

图 2　前期大气环流和海温因子与影响福建省热带气旋年频数相关场

(只给| 相关系数| ≥0. 30 的曲线)

西西伯利亚平原有一负相关区，研究还发现上一年 10 月巴基斯坦附近有一高相关区，最大相关系数为 0.63，这表明上一年 10 月印缅槽及其东侧的东亚低槽的强弱与次年影响福建TC 年频数关系非常密切。当上一年 10 月印缅槽及其东侧的东亚低槽偏西并加深（减弱），次年影响福建TC 年频数偏少（多）。另外，当年 5 月份高纬地区出现负相关区，日本附近有个最大相关系数为 0.45 的正相关区，25°N 以南是一负相关区。

此外，自上一年 3 月开始至当年 6 月在北大西洋的西南部（15~ 30°N，10~ 50°W）范围内持续出现一正相关区，其中最大相关系数以当年 4 月最大为 0.51，上一年 12 月次之为0.5。

综上所述，前期 500hPa 月平均高度场的极涡活动频繁的高纬地区、鄂霍次克海附近地区、日本附近和西亚（阿拉伯半岛和印度半岛）以及北大西洋的西南部是与影响福建TC 年频数密切相关的关键区。

2. 前期 100hPa 高度场

由图 1（b）可见，前一年 3 月为最高相关月份，前一年 12 月次之，前一年 8 月再次。图2 为前一年 3 月 100hPa 月平均高度场与影响福建热带气旋年频数的相关场，图中最大正相关区位于鄂霍次克海附近地区，最大负相关区位于北美洲，最大相关系数分别为0.48、- 0.44。前一年 8 月在西半球格陵兰岛西部的巴芬湾和前一年 12 月在东半球高纬各存在一个正相关区，最大相关系数分别为 0.53 和 0.42。当年 4~ 6 月，主要相关区域从朝鲜半岛向东北移动，经库叶岛至鄂霍次克海附近。

综上所述，前期 100hPa 月平均高度场的极涡活动频繁的高纬地区、鄂霍次克海附近地区和北美大陆是与影响福建 TC 年频数密切相关的关键区。

3. 前期 SST 场

图 1（c）表明，当年 3 月和 6 月为最高相关月份。从当年 1 月开始西北太平洋低纬地区（10~ 15°N，140~ 175°E）持续出现相关系数超过 0.05 信度标准的正相关区，至当年 6 月达最强，最大相关系数为 0.57，可见西太平洋热带气旋源地的海温与影响福建省热带气旋年频数呈相当稳定的正相关。3 月在西风飘流区、5 月在黑潮区皆出现正相关区，6 月赤道东太平洋区出现负相关区（最大为- 0.41），位于 NIN03 区（5°S~ 5°N，90~ 150°W）。

综上所述，前期北太平洋海温与影响福建TC 年频数密切相关的关键区位于西北太平洋低纬地区、赤道东太平洋和西风飘流区。

四、影响福建热带气旋年频数的预测模型

1. 预测模型的建立

本文在建立最优子集回归模型时，用 1951~ 1993 年资料建模，1994~ 1996 年 3 年用

于预测效果检验。预测因子分别选取上一年1月至当年4月的500hPa高度场、100hPa高度场、海温场和500hPa月环流特征量，先采用不同资料分三类构造分类预测方程，最后作出集成预报方程。拟合和预测效果的检验标准采用双评分准则（CSC）和预报效率(RF)。三类预测方程分别是：

Ⅰ类：500hPa月平均高度场

Ⅱ类：100hPa月平均高度场

Ⅲ类：海温场和500hPa月环流特征量

分类因子分别选出Ⅰ类53个，Ⅱ类34个，Ⅲ类18个，采用“两段筛选”的思路，先用逐步回归控制F检验值，各筛选出M个（10个左右）预报因子，再计算M个变量的全部可能回归寻找最优子集，建立分类预报模型，三类预测方程分别为：

$$Y1 = -10.753 + 0.326X1 - 0.185X2 - 0.272X3 + 0.069X4 - 0.167X5 + 0.064X6 + 0.605X7 - 0.11X8 - 0.118X10 - 0.105X11 + 0.056X12 \quad (4)$$

$$Y2 = 17.519 - 0.086X1 + 0.136X3 + 0.075X8 - 0.117X11 - 0.053X12 - 0.144X13 \quad (5)$$

$$Y3 = -108.43 + 0.219X1 + 0.176X3 + 0.136X5 - 0.093X6 - 0.069X7 + 0.047X8 + 0.051X9 + 1.797X10 \quad (6)$$

然后由三类预测方程的预测结果，应用全回归分析得到预测集成方程。集成方程为：

$$Y = -0.549 + 0.718Y1 + 0.244Y2 + 0.144Y3 \quad (7)$$

三类预测方程因子的情况见表1。

表1　分类预测方程的因子情况

方程类型	因　子	时　　间	点　数	中心位置	最大相关系数
Ⅰ类	X1	上一年2月	20	30°N，40°E	0.52
	X2	上一年2月	5	40°N，110°W	0.40
	X3	上一年3月	7	15°N，30°E	0.43
	X4	上一年3月	10	65°N，170°W	0.52
	X5	上一年6月	6	25°N，40°W	0.45
	X6	上一年8月	6	70°N，10°E	0.36
	X7	上一年8月	7	30°N，60°W	0.39
	X8	上一年8月	15	80°N，60°W	0.38
	X10	上一年11月	8	45°N，20°W	-0.47
	X11	上一年12月	8	35°N，80°W	-0.37
	X12	当年2月	4	40°N，10°E	0.37

(续表 1)

方程类型	因 子	时 间	点 数	中心位置	最大相关系数
II 类	X1	上一年 2 月	5	40°N，110°W	- 0.43
	X3	上一年 3 月	7	60°N，140°E	0.50
	X8	上一年 8 月	5	60°N，150°E	0.34
	X11	上一年 12 月	3	40°N，90°W	- 0.36
	X12	当年 2 月	3	40°N，50°W	- 0.38
	X13	当年 4 月	3	10°N，90°E	- 0.38
III 类	X7	上一年 1 月	3	10°N，180°E	- 0.42
	X8	上一年 11 月	3	50°N，170°W	0.33
	X1	上一年 11 月	4	10°N，150°W	0.49
	X9	当年 1 月	3	15°N，115°W	0.36
	X3	当年 1 月	3	15°N，115°W	0.36
	X5	当年 3 月	15	25°N，165°E	0.50
	X6	当年 4 月	5	10°N，165°E	0.36
	X10	上一年 2 月	亚欧西风环流指数		0.34

2 效果分析

图 3 给出影响福建 TC 年频数变化曲线和集成预测方程 (4) 拟合和预报的曲线，从中可以看出，除个别年份有些偏差外，基本趋势拟合很好，尤其异常年份与实测值基本一致。由表 2 可知，无论是模拟还是预测，集成预测模型皆优于分类预测模型，其复相关系数 (为 0.97) 和 CSC 评分 (为 86.25) 最大，均方根误差 (为 0.48) 为最小，从 1994~1996 三年试预报情况来看，平均绝对误差为 1.34，1995 年预测最好平均绝对误差为 0.58，预测效率为 86%，说明该方程具有一定的预测能力。分类预测模型中 I 类 (即用 500hPa 月平均高度场为因子) 效果要好于 II 类和 III 类。

表 2 样本方程的拟合评分和预测评分

方程类型	因子个数	复相关系数	均方根误差	CSC 评分	平均绝对误差	RF (%)
I 类	11	0.96	0.53	69.46	1.38	81
II 类	6	0.85	1.03	70.27	1.83	33
III 类	8	0.89	0.89	53.06	2.01	88
集成	3	0.97	0.48	86.25	1.34	86

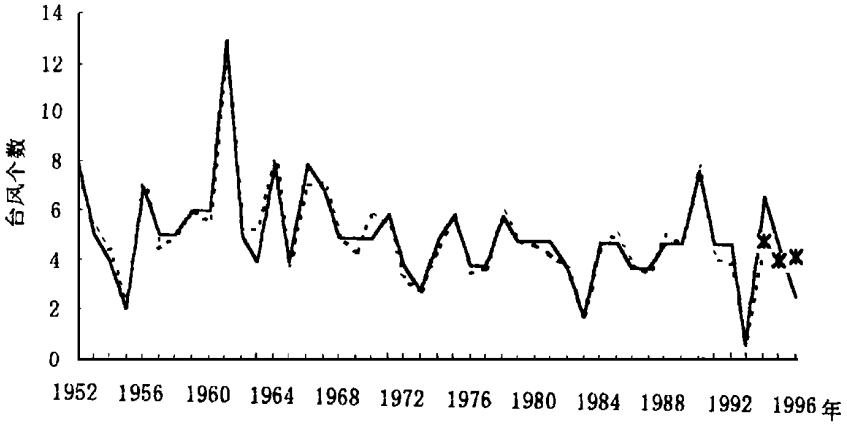


图 3　影响福建省热带气旋年频数变化

(实线为观测值，虚线为拟合值，* 号为预测值)

五、结　　论

1. 500hPa 月平均高度场在前一年 10 月巴基斯坦附近出现正偏差，当年 5 月高纬度地区出现负偏差，日本附近地区是正偏差；同时，当年 4~ 6 月 100hPa 月平均高度场鄂霍次克海附近地区出现正偏差，则预示当年可能有较多的TC 影响福建省，反之亦然。

2. 1~ 6 月西北太平洋低纬地区海温持续出现正距平，6 月赤道东太平洋海温出现负距平，则当年影响福建 TC 的年频数偏多。

3. 利用最优子集回归方法建立TC 预报模型，由于此方法确保筛选出的预报因子组合是最优，其效果较为理想，能很好地模拟影响福建 TC 年频数的变化趋势，具有一定的预测能力。

参　考　文　献

[1]　曹鸿兴等，1998：最优子集回归，讲义。

[2]　雷小途，1998：热带气旋频数预测的研究进展和业务预测水平，大气科学研究与应用（十四），198。