

统计预报中的组合因子方法

张家诚 赵 溱 许协江

(中央气象局研究所)

排列组合是一种常见的统计方法,气象中已有一些应用,作了一些试验,但这些试验还是极为初步的^[1]。我们这里总结的是这种方法进一步试验的结果。

为了说明用排列组合方法制作预报的原理与性能,我们用北京降水量作为一个例子。为了照顾到资料的连续性和可能作为因子的资料的长度,取1875—1973年共99年作为分析长度。

因子与预报量如下:

X_1 上年太阳黑子相对数

X_2 上年长江武汉站的年平均流量

X_3 当年3月12日—14日上海平均日最低气温

X_4 上海冬季温度(上年12月、当年1、2月月平均气温的均值)

X_5 上年上海年降水量

Y 当年北京年降水量

选取3月12日—14日上海日最低气温的原因是在上海百年气温年变程中,这三天温度特殊偏低,形成气候上的一个奇异点。

排列组合需要对各因子分级,我们这里分级按均匀分布,1、2、3级各占1/3。各因子的均值、均方差及分级标准为:

	平 均 值	均 方 差	1、2 级之间	2、3 级之间
x_1	51.4	42.6	63.7	25.8
x_2	23333	3210.8	24200	21950
x_3	3.5	2.8	4.6	1.8
x_4	4.5	1.0	5.0	4.2
x_5	1139.8	193.0	1207.4	1027.8
y	636	223.5	692.7	520.0

根据分级标准,可把资料改成分级资料列成表一:

1977年5月19日收到。

表一 各因子逐年分级资料

年份	x_1	x_2	x_3	x_4	x_5	y	年份	x_1	x_2	x_3	x_4	x_5	y	年份	x_1	x_2	x_3	x_4	x_5	y	年份	x_1	x_2	x_3	x_4	x_5	y
1875	2	2	1	2	3	3	1901	3	3	3	2	3	3	1927	1	1	1	3	1	2	1953	2	1	1	2	1	2
76	3	2	1	3	1	2	02	3	3	1	1	2	2	28	1	2	3	1	2	2	54	3	3	3	1	3	1
77	3	1	3	3	3	3	03	3	3	1	2	3	3	29	1	3	3	2	2	1	55	3	1	2	3	1	1
78	3	3	2	3	3	1	04	3	2	2	1	2	2	1930	1	3	1	2	3	3	56	2	2	3	1	2	1
79	3	1	1	2	2	1	05	2	2	2	3	3	3	31	2	2	1	2	2	2	57	1	3	3	3	1	3
1880	3	3	1	3	1	3	06	2	1	2	2	1	2	32	3	1	2	2	1	2	58	1	3	1	2	1	2
81	2	3	3	3	2	2	07	2	1	1	2	1	3	33	3	2	3	3	3	1	59	1	3	2	2	3	1
82	2	2	3	1	1	2	08	2	1	3	2	1	2	34	3	2	2	2	3	2	1960	1	3	2	1	1	2
83	2	1	2	3	1	1	09	2	2	2	1	2	3	35	3	2	1	1	3	3	61	1	3	1	2	2	2
84	1	1	3	3	2	2	1910	2	1	2	3	1	2	36	2	1	1	3	3	3	62	2	2	1	2	2	3
85	2	3	3	3	2	2	11	3	1	1	2	2	1	37	1	2	2	1	3	1	63	2	2	2	3	3	1
86	2	1	2	3	2	1	12	3	1	3	2	1	1	38	1	1	3	2	3	1	64	2	3	2	2	1	1
87	3	1	2	3	2	1	13	3	2	3	2	1	2	39	1	1	2	1	1	2	65	3	1	1	1	3	3
88	3	1	1	2	2	3	14	3	2	2	1	2	1	1940	1	3	1	2	2	2	66	3	2	1	1	2	2
89	3	2	3	3	3	1	15	3	3	3	1	2	1	41	1	3	2	1	2	3	67	2	3	1	3	2	2
1890	3	1	3	1	1	1	16	2	2	2	1	1	3	42	2	3	1	1	1	3	68	1	2	2	3	3	3
91	3	1	3	2	3	2	17	2	3	1	3	1	1	43	2	3	2	2	3	3	69	1	1	3	1	3	1
92	2	2	3	2	1	1	18	1	2	2	3	2	3	44	3	2	2	1	2	3	1970	1	3	3	2	2	2
93	1	2	1	3	3	1	19	1	1	2	2	1	2	45	3	2	1	3	3	3	71	1	2	3	2	2	3
94	1	2	2	1	2	1	1920	2	2	3	3	2	3	46	2	2	2	1	1	1	72	1	3	3	2	3	3
95	1	2	1	3	3	3	21	2	1	3	1	3	3	47	1	3	1	2	1	2	73	1	3	1	1	3	1
96	1	3	3	2	3	2	22	2	1	1	1	1	1	48	1	2	1	2	1	2	74	2	1	2	3	2	2
97	2	2	2	3	2	2	23	3	2	2	3	2	3	49	1	1	1	1	2	1	75	2	2	1	1	2	3
98	2	1	3	1	2	2	24	3	3	3	2	3	1	1950	1	1	3	1	1	1	76	3	2	2	3	1	2
99	2	3	3	2	3	3	25	3	1	1	3	3	1	51	1	1	2	2	1	3							
1900	3	3	2	2	1	3	26	2	3	1	2	3	3	52	1	3	3	1	1	2							

根据表一,很容易将资料按各种因子排列起来,可以用电子计算机操作,也可以用边洞卡或手工进行。我们首先是用电子计算机按五因子排列,但为了比较直观地说明问题,我们任意选了三个因子,作成表二。以下我们就从解释这张表格入手,再进而说明其他问题。

从五个因子中任选三个因子,共有 $c_3^5 = 10$ 种选法,可制成10张这样的表格。用计算机操作甚方便。本文所列 X_1, X_2, X_3 的表格系手工作成,约费时一小时。

制表的方法:第一步根据表一的资料确定每年在表二中的位置,记下年号,年后括弧内的数字即预报值。为了扩大表的用途,我们把五年的预报值按顺序列上。因此,利用这一张表,就有制作五年预报的可能。

以后,统计各个预报年中预报值出现的概率,将可以作为预报指标使用的情况用记号标出来(如表二用√表示)。例如,在 $X_1 = 1, X_2 = 1, X_3 = 1$ 的情况下(简称111格),预报当年的降水级别历史概括率可达4/4。但预报第二、三年只能排除3级而无法分辨1、2级。第五年排除1级,无法分辨2、3级。第四年则没有参考价值。

不仅表中27种排列可以分别给出预报指标,而且还可以任意结合起来,增加统计次数。例如131、231、331三格中的第三年合在一起,9次中完全排除了出现3级的可能。

而在第五年 9 次中排除 3 级的可能性达 8/9。但如果把这三格分开,则有的不能达到统计要求。事实上,这种合并也等于剔除了作用不大的因子。在第一例中剔除了 X_1 。又例如 111 和 113 两种情况的第二年 8 次中完全排除 3 级,可是在 112 中却有 3 级。这说明在 $X_1 = 1$ 和 $X_2 = 1$ 的背景下,3 级的出现呈非线性关系,它的出现概率只是在 $X_3 = 2$ 的情况下最大,向两端减小。(当然,这里只是为了解释方便,从统计学上作出结论次数尚太少)。诸如此类,27 格共有 2^7 种结合方式,可以进行细致而广泛的分析,得出各种组合状态下的关系。从这个角度来看,这种表格显示了分析中的灵活性。

这里产生两个问题,一个是这种组合状态数量很大,不可能全部考虑。另一个问题是选取的指标是否反映了一定的气候规律,还是纯系随机碰上的。为此,我们设计了一种计算机上使用的方法,逐次考虑全部因子,剔除一个因子,剔除两个因子,……,直到统计显著性达到要求为止。而显著性以超过气候概率的 χ^2 值来衡量。对所有 99 年资料进行一次检查,费时不到半个小时,显著程度也较高(下面还要谈到)。

其次,这种表格又是一种极好的历史档案索引。如果所选用的指标代表了前期过程的主要特征的话,那么,具有同一特征的全部年份都已在表中列在一起,这就有可能把这些年份作为第一近似年,再去查阅其他历史资料,以挑选最相似的年份。举一个例子,我们要作 1961 年的预报,这一年出现在 112 格中(见表二),但这一格仅有 2 年的频数,其 1 级和 2 级两种情况均未占优势,因此不能据此作为预报指标。为了解决这个问题,根据表一中 X_1 、 X_3 、 X_5 的级别,可以找出 1940 和 1949 这两年与 1961 年相似,这时再扩大到五个因子,1940 年的因子级别排列为 13122,1949 年为 11112,而 1961 年为 13122,以 1940 年为最相似年。因为 1940 年为 2 级,所以预报 1961 年为 2 级。当然,这种分析是太简单了,还可以用更多的因子及其多种特征进行全面考虑。因此,即使在没有指标意义的情况下,这个表仍能提供一定信息量。在指标良好时,也可以根据这个线索查阅其他资料,以论证指标的可靠性。

此外,这种表尚十分便于业务应用。当做预报时,可把预报年填入相应格内,预报值用红字注在相应年的旁边,实况出现后再正式记入。这样,预报、检查和资料补充等工作单一化了,收到多快好省的效果。如果要将这种表用于大中小台预报的结合上,也可以将大台的指标甚至大台预报本身列为一个因子,其他因子的影响即是对大台预报的修正。

但是,这种表格有一个难以克服的缺点,就是包含的因子不多,而且也只能取其个别特征。例如,在表二中只取级别,未考虑变化,也未考虑如太阳黑子的位相等特征。为此,可以从此表推广,采用字典的形式。我们在计算机上打印了一份小字典,如表三所列,查起来很方便。

事实上,任何拼音文字的字典都是字母的排列组合的结果。采用字典的形式,因子数量就可大大扩大了。但对长期预报来说,限于资料年代,对短期预报也许更为适用。

不论字典或是较大的表格,在分析时仍有许多不便。边洞卡是便于分析的半手工方法,便于应用^[2]。边洞卡是一种普通的硬纸卡片,只是沿边均匀排列一圈小圆洞。平均每厘米可列两个洞。10 × 6 厘米的卡片可以有洞 60 个之多。每个因子都在卡片上占有一定位置的洞。如果采用二级制,一个洞即可表示一个因子。如果三级或四级则需两个洞表示。五到八级用三个洞。将洞剪开与不剪开表示二进位的 1 和 0,如图示:

表三 因子字典表

· · · · ·	· · · · ·	· · · · ·	· · · · ·
22233 级别 个例数 频率	1 1 0.5	2 0 0	3 1 0.5
年份 实况	1963 775.6		1905 482.0
32212 级别 个例数 频率	1 1 0.33	2 1 0.33	3 1 0.33
年份 实况	1914 721.0	1904 580.9	1944 476.0
· · · · ·	· · · · ·	· · · · ·	· · · · ·

$$\frac{1}{0} \quad \frac{2}{0} \quad \frac{1}{00} \quad \frac{2}{00} \quad \frac{3}{00} \quad \frac{4}{00}$$

二级 四级

只要用小棍穿入孔内,由于孔有剪开与不剪开的差别,自然就把级别分离开来。如果是多种因子各种级别的结合,可以顺次挑选。经过多次试验,即能选出较好的因子组合,记录下来,以备应用,不必每次重复。

字典与边洞卡两种形式都可增加预报量的内容,以扩大使用价值。

为了说明该方法的使用效果,我们同时又作了逐步回归和逐步判别,拟合效果如下:

1. 逐步回归 取 $F_1 = F_2 = 0$, 各因子选入次序是 X_2, X_1, X_3, X_4, X_5 , 复相关系数 $R = 0.29$, 将拟合值化成级别, 拟合效果如下表。报错 60 次, Hedike 评分为 0.09, 拟合效果甚差。

试报: 1974 年 657.8 (2 级), 实况为 591.2 (2 级)。1975 年 580.9 (2 级), 实况为 411.9 (3 级)。

表四 逐步回归法拟合表

次 数 级 别	级 别	原 级 别			合 计
		1	2	3	
计 算 级 别	1	10	6	5	21
	2	23	26	25	74
	3	0	1	3	4
合 计		33	33	33	99

2. 逐步判别 取 $F_1 = F_2 = 2.0$, 选入 X_2, X_3 两个因子, 拟合效果比回归要好一些 (见表五)。报错 52 次, Hedike 评分 0.21。

试报: 1974 年 1 级, 实况为 2 级。1975 年 3 级, 实况为 3 级。

表五 逐步判别法拟合表

拟合次数 级 别	级 别	原 级 别			合 计
		1	2	3	
计算 级别	1	14	10	10	34
	2	9	16	6	31
	3	10	7	17	34
合 计		33	33	33	99

3. 用本文所述方法 考虑 5 个、4 个、……因子,直到通过 χ^2 检验 (与气候概率差异显著) 为止,拟合效果如表六。报错 39 次, Hedike 评分为 0.41, 效果比逐步判别又有所提高。这个道理是不难理解的。设因子为 $x_1, x_2, \dots, x_p$, 预报量为 y , 它们有如下关系:
 $y = f(x_1, x_2, \dots, x_p)$

表六 组合因子法拟合表

次 数 级 别	级 别	原 级 别			合 计
		1	2	3	
计算 级别	1	17	5	3	25
	2	7	24	11	42
	3	9	4	19	32
合 计		33	33	33	99

“回归”是将 F 值限制在线性函数类, y 值落在一个超平面附近。“判别”则是用几个平面把空间分割成几个区域,使之对应预报量出现的几种状态。这两种方法都不可能将因子与预报量之间的实在关系确切地表达出来。本文使用的方法,在只考虑级别的情况下是比较好的。但由于计算量的关系,只考虑了一部分因子的组合状态,并不能保证是最优的。

表七 完全拟合表

次 数 级 别	级 别	原 级 别			合 计
		1	2	3	
计算 级别	1	33	0	0	33
	2	0	33	0	33
	3	0	0	33	33
合 计		33	33	33	99

此外,我们用了—一个方法对此进行检验,如果预报与实况完全一致,则它的分布应该如表七所示。对此表频数分布的偏差越小,说明该方法使用效果越好,今暂称差数为“全差”(即全相关差),三个方法的平均“全差”值比较如下:

逐步回归: $120/9 = 13.3$

逐步判别: $104/9 = 11.6$

本文方法: $78/9 = 8.7$

以上检验说明,本文方法的平均“全差”值最小,因此比另两种方法要好.

根据上述方法,我们在 1976 年初预报北京 1976 年降水不会出现 3 级(无明显干旱),实况为 2 级(正常).

统计预报方法是在不断向前发展的,排列组合原理在一些兄弟学科已取得显著成绩,这一原理究竟在气象预报中使用价值如何,尚待实践检验,本文只从个例上进行了若干试验,希望批评指正.

参 考 资 料

- [1] 张家诚, F. 鲍尔学派的长期天气预报方法,《气象学若干问题的进展》,科学出版社,1963 年.
- [2] 张先恭、林复旦、张家诚,手选边洞卡在长期天气预报中的应用,《气象学报》,第 36 卷,第 2 期,1966 年,6 月.