



## (七) 分类筛选预报的极差法

张 强

七、八、九三讲主要介绍分类筛选因素的预报方法。用一个实例来说明这个方法的基本思想和相应的计算步骤、方法。

例：考察六个因子对8、9月份降水总量的关系。资料见表7.1：

第一步：逐个考察每个因子 $x_i$ 与预报量 $y$ 的关

表7.1 新安江水库的六个前期因素和8、9月份降水量的数据

| 数<br>年<br>份 | 因子及<br>预报量<br>值 | $x_1$<br>上一年<br>小雪一大雪<br>降水 | $x_2$<br>入 梅<br>日 期 | $x_3$<br>12月—当年<br>1月雷暴<br>天 数 | $x_4$<br>上一年阴历<br>十月降水 | $x_5$<br>上一年谷雨<br>节气降水 | $x_6$<br>上一年白露<br>节气降水 | $y$<br>8、9月总的<br>降 水 量 |
|-------------|-----------------|-----------------------------|---------------------|--------------------------------|------------------------|------------------------|------------------------|------------------------|
| 1954        | 45.6 (15)       | 双日 (2)                      | 2 (3)               | 118.1 (17)                     | 28.0 (1)               | 79.1 (15)              |                        | 94.2                   |
| 1955        | 51.8 (16)       | 双日 (2)                      | 0 (1)               | 59.2 (8)                       | 112.9 (10)             | 3.3 (2)                |                        | 169.9                  |
| 1956        | 2.0 (4)         | 双日 (2)                      | 0 (1)               | 64.5 (10)                      | 79.5 (7)               | 63.4 (13)              |                        | 381.5                  |
| 1957        | 0.3 (1)         | 单日 (1)                      | 0 (1)               | 8.6 (2)                        | 73.5 (5)               | 88.6 (16)              |                        | 519.2                  |
| 1958        | 24.2 (11)       | 单日 (1)                      | 0 (1)               | 58.8 (7)                       | 248.7 (17)             | 105.3 (17)             |                        | 170.1                  |
| 1959        | 1.0 (2)         | 单日 (1)                      | 1 (2)               | 19.3 (3)                       | 249.4 (18)             | 35.7 (8)               |                        | 182.9                  |
| 1960        | 33.7 (14)       | 单日 (1)                      | 0 (1)               | 97.3 (15)                      | 63.1 (3)               | 52.9 (11)              |                        | 270.2                  |
| 1961        | 65.8 (17)       | 单日 (1)                      | 0 (1)               | 65.6 (11)                      | 71.5 (4)               | 46.6 (9)               |                        | 234.8                  |
| 1962        | 6.9 (6)         | 单日 (1)                      | 0 (1)               | 59.4 (9)                       | 75.3 (6)               | 50.9 (10)              |                        | 280.0                  |
| 1963        | 23.2 (10)       | 单日 (1)                      | 0 (1)               | 48.1 (4)                       | 127.6 (12)             | 17.7 (5)               |                        | 207.2                  |
| 1964        | 27.2 (12)       | 双日 (2)                      | 1 (2)               | 52.9 (5)                       | 133.2 (13)             | 33.8 (7)               |                        | 100.2                  |
| 1965        | 3.4 (5)         | 单日 (1)                      | 0 (1)               | 3.4 (1)                        | 118.7 (11)             | 66.5 (14)              |                        | 396.7                  |
| 1966        | 14.2 (8)        | 双日 (2)                      | 2 (3)               | 107.3 (16)                     | 153.5 (15)             | 4.3 (3)                |                        | 45.5                   |
| 1967        | 10.6 (7)        | 单日 (1)                      | 1 (2)               | 90.6 (13)                      | 147.9 (14)             | 0.0 (1)                |                        | 32.7                   |
| 1968        | 19.5 (9)        | 单日 (1)                      | 0 (1)               | 89.5 (12)                      | 211.7 (16)             | 11.4 (4)               |                        | 140.0                  |
| 1969        | 28.1 (13)       | 双日 (2)                      | 5 (4)               | 133.2 (18)                     | 49.7 (2)               | 58.5 (12)              |                        | 235.7                  |
| 1970        | 1.0 (3)         | 单日 (1)                      | 0 (1)               | 55.9 (6)                       | 103.6 (8)              | 21.0 (6)               |                        | 200.1                  |
| 1971        | 80.3 (18)       | 双日 (2)                      | 0 (1)               | 93.0 (14)                      | 110.2 (9)              | 109.1 (18)             |                        | 219.3                  |

注 ( ) 中的数字表明由小到大排列时相应的次序。

系，按 $x_1$ 取值的大小，由小到大的顺序，将 $y$ 的值按这个顺序排列，对 $y$ 的值考虑它的最优2分割。对 $x_1$

和 $y$ ，列出表7.2。

用极差分割法找到最优的2分割是(写年份)：(1957,

表7.2

|              |   |
|--------------|---|
| 依 $x_1$ 的大小排 | 0.3 1.0 1.0 2.0 3.4 6.9 10.6 14.2 19.5 23.2 24.2 27.2 28.1 33.7 45.6 51.8 65.8 80.3<br>(1) (2) (3) (4) (5) (6) (7) (8) (9) (10) (11) (12) (13) (14) (15) (16) (17) (18) |
| 年 份          | 1957 1959 1970 1956 1965 1962 1967 1966 1968 1963 1958 1964 1969 1960 1954 1955 1961 1971   |
| $y$ 值        | 519.2 200.1 396.7 32.7 140.0 170.1 235.7 94.2 234.8<br>182.9 381.5 280.0 45.5 207.2 100.2 270.2 169.9 219.3   |

1959, 1970, ……1962) (1967, 1966, 1968, ……, 1971)，相应的最大极差是 $519.2 - 182.9 = 336.3$ 。对其余的因子 $x_2, \dots, x_6$ 和 $y$ 的关系，也完全一样。要注意的是对 $x_1$ 和 $y$ ， $y$ 只有一个2分割，即依入梅日期

是单日还是双日分组，得相应的最大极差是 $519.2 - 32.7 = 486.5$ 。又如对 $x_3$ 和 $y$ ，注意因子的值可以不写，就得表7.3。

实际上，此时要比较的2分割只有三个，最优2分割

表 7.3

| 依 $x_3$ 的大小排<br>(年份) | (1) (1) (1) (1) (1) (1) (1) (1) (1) (1) (1) (1) (1) (2) (2) (2) (3) (3) (4)                                 |
|----------------------|---|
|                      | 1955 1956 1957 1958 1960 1961 1962 1963 1965 1968 1970 1971 1959 1964 1967 1954 1966 1969                   |
| $y$ 值                | 169.9 519.2 270.2 280.0 396.7 200.1 182.9 32.7 45.5<br>381.5 170.1 234.8 207.2 140.0 219.3 100.2 94.2 235.7 |

是 (1955, 1956, 1957, ..., 1970, 1971) (1959, 1964, 1967, 1954, 1966, 1969), 最大极差是 519.2  $- 140.0 = 379.2$ 。 $x_4$ 、 $x_5$ 、 $x_6$  与  $y$  的关系同上计算,

列出表 7.4、7.5、7.6。

表 7.4

| 依 $x_4$ 的大小排<br>年份 | 1965 1957 1959 1963 1964 1970 1958 1955 1962 1956 1961 1968 1967 1971 1960 1966 1954 1969                   |
|--------------------|---|
| $y$ 值              | 396.7 182.9 100.2 170.1 280.0 234.8 32.7 270.2 94.2<br>519.2 207.2 200.1 169.9 381.5 140.0 219.3 45.5 235.7 |

最优 2 分割是 (1965, 1957) (1959, 1963, 1964, 1967)  $32.7 = 348.8$ 。

1970, 1958, 1955, ..., 1969), 最大极差是 381.5 -

表 7.5

| 按 $x_5$ 的大小排<br>年份 | 1954 1969 1960 1961 1957 1962 1956 1970 1971 1955 1965 1963 1964 1967 1966 1968 1958 1959                   |
|--------------------|---|
| $y$ 值              | 94.2 270.2 519.2 381.5 219.3 396.7 100.2 45.5 170.1<br>235.7 234.8 280.0 200.1 169.9 207.2 32.7 140.0 182.9 |

最优 2 分割是 (1954, 1969, ..., 1963) (1964, 1967, 1966, ..., 1959), 相应的最大极差是 519.2 -

表 7.6

| 按 $x_6$ 的大小排<br>年份 | 1967 1955 1966 1968 1963 1970 1964 1959 1961 1962 1960 1969 1956 1965 1954 1957 1958 1971                   |
|--------------------|---|
| $y$ 值              | 32.7 45.5 207.2 100.2 234.8 270.2 381.5 94.2 170.1<br>169.9 140.0 200.1 182.9 280.0 235.7 396.7 519.2 219.3 |

最优 2 分割是 (1967, 1955, ..., 1965, 1954) (1957, 1958, 1971), 最大极差是 396.7 - 32.7 = 364。

比较这六个最优 2 分割的最大极差, 发现依  $x_1$  的大小排所得的最优 2 分割最大极差最小, 是 336.3。于是采用这个最优 2 分割, 把全部 18 年的资料分成两组, 分组剔去因子  $x_1$ , 剩下其余的五个因子。

第二步: 对每一组资料单独进行与第一步完全相似的计算。

第 1 组: 1957, 1959, 1970, 1956, 1965, 1962 共六年资料。分别算得如下结果:

| 按 $x_2$ 大小排<br>年份 | 1957 1959 1970 1965 1962 1956          |
|-------------------|--|
| $y$ 值             | 519.2 200.1 280.0<br>182.9 396.7 381.5 |

最优 2 分割是 (只有 1 个 2 分割) (1957, 1959, 1970, 1965, 1962) (1956), 最大极差是 519.2 - 182.9 = 336.3。

| 按 $x_3$ 大小排<br>年份 | 1956 1957 1965 1970 1962 1959          |
|-------------------|--|
| $y$ 值             | 381.5 396.7 280.0<br>519.2 200.1 182.9 |

此时只有一个 2 分割: (1956, 1957, 1965, 1970, 1962) (1959), 最大极差是 519.2 - 200.1 = 319.1。

| 按 $x_4$ 大小排<br>年份 | 1965 1957 1959 1970 1962 1956          |
|-------------------|--|
| $y$ 值             | 396.7 182.9 280.0<br>519.2 200.1 381.5 |

最优 2 分割是 (1965, 1957) (1959, 1970, 1962, 1956), 最大极差是 381.5 - 182.9 = 198.6。

| 按 $x_5$ 大小排<br>年份 | 1957 1962 1956 1970 1965 1959          |
|-------------------|--|
| $y$ 值             | 519.2 381.5 396.7<br>280.0 200.1 182.9 |

最优 2 分割是 (1957) (1962, 1956, 1970, 1965, 1959), 最大极差是 396.7 - 182.9 = 213.8。

| 按 $x_6$ 大小排<br>年份 | 1970  | 1959  | 1962  | 1956  | 1965  | 1957  |
|-------------------|-------|-------|-------|-------|-------|-------|
| $y$ 值             | 200.1 | 280.0 | 396.7 | 182.9 | 381.5 | 519.2 |
|                   |       |       |       |       |       |       |

最优 2 分割是 (1970, 1959, 1962) (1956, 1965, 1957), 最大极差是  $519.2 - 381.5 = 137.7$ 。

比较这五个最大极差, 最小值 137.7, 因此应该依  $x_6$  的大小顺序来分割  $y$  值, 得两组,

| 年 份     | 1970  | 1959  | 1962  | 1956  | 1965  | 1957  |
|---------|-------|-------|-------|-------|-------|-------|
| 降水值 $y$ | 200.1 | 182.9 | 280.0 | 381.5 | 396.7 | 519.2 |

两组内部的降水值已很接近, 就不必再往下分割了。再看第 2 组的数据, 把全部结果列成表 7.7, 以便对照。

表 7.7

| 依变量 $x_1$<br>的大小 | 最 优 2 分 割<br>(年份)                                    | 最 大 极 差                  |
|------------------|--|--------------------------|
| $x_2$            | (58, 60, 61, 63, 67, 68)<br>(54, 55, 64, 66, 69, 71) | $270.2 - 32.7 = 237.5$   |
| $x_3$            | (55, 58, 60, 61, 63, 68, 71)<br>(64, 67, 54, 66, 69) | $235.7 - 32.7 = 203.0$   |
| $x_4$            | (63, 64, 58, 55, 61, 68, 67)<br>(71, 60, 66, 54, 69) | $270.2 - 45.5 = 224.7$   |
| $x_5$            | (54, 69, 60, 61, 71)<br>(55, 63, 64, 67, 66, 68, 58) | $207.2 - 32.7 = 174.5^*$ |
| $x_6$            | (67, 55, 66, 68)<br>(63, 64, 61, 60, 69, 54, 58, 71) | $270.2 - 94.2 = 176.0$   |

可见最大极差最小的 2 分割依  $x_5$  分最好。于是依  $x_5$  的最优 2 分割把第 2 组数据又分成两组, 记为第 2.1 组与 2.2 组:

2.1 组 1954, 1960, 1961, 1969, 1971

2.2 组 1955, 1963, 1964, 1967, 1966, 1968, 1958

然后对 2.1 组和 2.2 组再分别单独进行相当于第一步的计算, 此时剔除因子  $x_6$ 。

在进行下一步的计算之前, 我们来指明一点。从刚才的计算过程可以看到, 在第一步中, 雷暴天数  $x_3$  是零还是几对降水的影响是不明显的。然而对第 2 组的资料, 虽然  $x_3$  还够不上最重要, 可是有无雷暴的影响是明显的, 雨量少的年份都集中在无雷暴的这一组。这里可以告诉我们怎样去比较好地总结群众的经验, 根据当地群众的农谚“有冬雷, 要干秋”, 是否说明冬雷的有无是一个重要的因素呢? 它的确“是”, 然而又不是无条件的“是”。

进一步, 依  $x_5$  把资料分为两组, 再计算各组的分割。

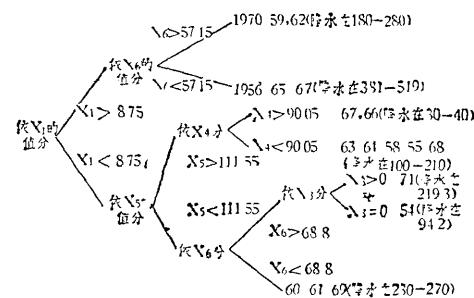
对第 2.1 组 1954, 1960, 1961, 1969, 1971 和对第 2.2 组 1955, 1963, 1964, 1967, 1966, 1968, 1958, 列表如下(表 7.8)。

从表中可以看出 2.1 组再分为 2 组, 而 (1954, 1971)

表 7.8

| 组号  | 依变量 $x_1$<br>的大小 | 最 优 2 分 割                     | 最 大 极 差 |
|-----|------------------|-------------------------------|---------|
| 2.1 | $x_2$            | (60, 61, ) (54, 69, 71)       | 141.5   |
|     | $x_3$            | (54, 69) (60, 61, 71)         | 141.5   |
|     | $x_4$            | (54, 69) (60, 61, 71)         | 141.5   |
|     | $x_6$            | (60, 61, 69) (54, 71)         | 125.1 √ |
| 2.2 | $x_2$            | (58, 63, 67, 68) (55, 64, 66) | 174.5   |
|     | $x_3$            | (55, 58, 63, 68, 64) (67, 66) | 107 √   |
|     | $x_4$            | (63, 64, 58, 55, 68) (67, 66) | 107 √√  |
|     | $x_6$            | (67) (55, 66, 68, 63, 64, 58) | 161.7   |

这一组虽然只有两年, 但出入很大, 还应分开, 依  $x_3$  或  $x_4$  都可以分开, 其效果是一样的。为了便于判断, 就用  $x_3$  (即冬天有无雷暴) 将它们分开。至于 2.2 组再分为两组后, 内部数据已经比较整齐了, 就不要再分了。因此得最后结果如附图。



附 图

图中的分界值是由最优 2 分割中第一段的最后一个点与后一段的最前一个点的数据的平均(例如 8.75 就是 1962、1967 两年  $x_1$  的值 6.9 和 10.6 的均值, 其余都类似)。这张图告诉了我们, 应将历史资料分成七种情况, 然后利用这张图作预报。而且已选出预报因子共 5 个:  $x_1$ ,  $x_3$ ,  $x_4$ ,  $x_5$ ,  $x_6$ .  $x_2$  没有选上, 被淘汰了。

下面我们来看如何预报。例如某两年五个因子的值是:

| $x_1$   | $x_3$ | $x_4$ | $x_5$ | $x_6$ |
|---------|-------|-------|-------|-------|
| 第一年 9.4 | 0     | 10.5  | 77.4  | 102.3 |
| 第二年 6.3 | 2     | 34.3  | 112.4 | 42.7  |

于是, 从图上找到  $x_1 > 8.75$ ,  $x_6 > 57.15$ , 所以第一年应归入 1970, 1959, 1962 这三年一类, 即与 1970, 1959, 1962 年相似, 降水量估计在 180—280 毫米之间。从图上找到第二年满足  $x_1 < 8.75$ ,  $x_6 > 111.55$ ,  $x_4 < 90.05$ , 因此它与 1963, 1964, 1958, 1955, 1968 这些年份相似, 降水量估计在 100—210 毫米之间。因此只要有了图, 作预报时是非常方便的。这个实例的全部过程就介绍完了, 下面我们对这个方法给出一些说明, 帮助读者理解这个方法。

1. 这个方法实际上是最优分割法的一种灵活运用, 如果不是用极差进行最优分割, 而是用方差进行

最优分割，它就是国外书上通常所说的A.I.D方法，下一讲我们将从一般公式和理论分析上来详细说明这一点。

2. 这个方法，每进行一次筛选，某一个因子入选后，下次分组考虑时，这个因子就不再使用了。当然，也可以分组后仍然还将这个因子再放入使用。实际上，如果  $x_1$  已入选了，将资料分成了两组，对每一组还要考虑依  $x_1$  的大小顺序来进行最优 2 分割，这不是相当于考虑未分组时依  $x_1$  的大小顺序来考虑最优 4 分割吗？可见，我们稍加改动，就可以把上面的方法推广到更一般的情形。我们不是考虑依  $x_1, \dots, x_k$  (因子) 的大小顺序来进行  $y$  的最优 2 分割，而是进行最优的 K 分割，K 的选取，视分割后的情况而定，这就得到了更进一步的分类筛选因子的预报方法。

3. 从上面的讨论可以看出，对于因子，我们进行分割时只利用到它的大小顺序，而在分类图中才用到了它的数值(取平均值)。实际上，在分类图中也可以去掉数值，要的是顺序，即依  $x_i$  的大小排列时，它的顺序在第几个之前还是之后，对于每一个新的资料，如能判定它的  $x_i$  因子的顺序号，就完全可以用分类图作预报，因此，这个方法可以处理有顺序而不能定量地描述的因子，例如对云量，我们可以用无云、少云、多云来反映顺序，而无需去定量地测定它，这对于选预报因子和进行预报都是很方便的。

4. 从上面的讨论还可以看出，在分类筛选预报中，预报的对象的值是我们考察的重点，而它的顺序是随各个因子来改变的。在前三讲中，我们详细介绍了最优分割法，它可以对多个变量进行最优分割。所以，很自然，这里介绍的分类筛选预报自然可以对多个预报量同时给出预报值。这一点和我们熟知的一些预报方法是不一样的，这是分类筛选法的一个很重要的优点。

5. 读者可能会问：既然分类筛选预报是最优分割法的一种灵活运用，那么它与第六讲中的预报方法是否真的有区别呢？会不会只是形式上的不同，实质上这两种灵活运用是一样的？对于这个问题，需要仔细分析比较后才能回答。上一讲的方法，是按预报量的大小顺序来对因子进行最优分割，最后，比较因子的分割和预报量事先给定的分割是否一致；这一讲的方法是按因子的大小顺序来对预报量进行最优分割，按分割后的类型进行预报。看起来，这两种方法是不一样的。从筛选因子的角度来看，这两种方法都能筛选因子，但是筛选的过程和方法也不一样。从理论上说，我们还不知道这两种方法是否一样，或主要差别是什么，然而，从直觉上，从具体的实例来看，这两种方法似乎是有共同的基本思想。打一个比方，和我们熟悉的回归分析方法相比较，前者相当于按预报量分组回归，后者相当于按因子来分组回归，回归的基本思

想是相同的，然而回归的手段是不一样的，它们各有自己的优缺点，可以互相参照比较，而不是相互排斥的，但也不是完全一样的。当然，这些看法仅仅是经验，还有待于作理论上的进一步分析和研究。

6. 最后再说明一点。既然最优分割法可以考虑分块最优分割和多中心最优分割，把它用在分类筛选方法上也是一样适用的，这就不再重复了。

总之，对一种方法，只要我们真正掌握了它的基本思想和处理方法，往往一旦灵活运用后，可以变化出很多很复杂很有意思的新方法。因此，希望读者认真把一种方法弄通，自己就可以变化、发展，得到新的更好的方法。

本讲例子的数据是选用长江流域规划办公室编的《一九七五年长江流域长期水文气象预报讨论会技术经验交流文集》中第11页—第15页，新安江水电厂气象组的文章“用一组农谚预报干旱期降水量趋势”一文中的材料。因文中没有给出72、73、74年各因子的具体数值，就无法作出真实的预报来和实测值比较。读者可以选用本站资料来作一次分析和预报，看看这种方法的效果究竟如何。

### 测报经验点滴

## 避免气温读数误差

在每次观测前巡视仪器时，先读一次干球温度表读数，做到事前胸中有数，在进行观测温度后还要复读一遍，比较一下是否有误。然后与最低温度表酒精柱读数比较是否相差过大。最后与温度计读数进行比较。经过如此两读、两比，可以杜绝干球温度表读数的  $1^{\circ}\text{C}$  和  $5^{\circ}\text{C}$  差。

(云南勐龙气象站 杜香)



### 花粉母细胞减数分裂

期 在水稻幼穗发育过程中，当花粉母细胞形成后，即进入减数分裂，经过这种分裂以后，染色体比原来的数目减少一半，形成四分体细胞。一个花粉母细胞自开始第一次分裂，再经过第2次分裂，直到形成四分体，大约1—2天时间。从外形上看，幼穗长度达到成穗长度的一半时为花粉母细胞减数分裂期。在分裂前适当追肥，可减少颖花的退化，而保证结实。