用方差分析选择预报因子

黑龙江泰来县气象科

在天气预报中，预报因子的选择是很重要的。方差分析是通过对试验数据的分析来辨别矛盾主次的一种数学方法，通过它可以抓住影响事物变化的主要矛盾。在天气预报工作中，可以用这种方法来选择对预报对象影响明显的主要预报因子。

先将要选择的某个因子X_k（$k=1,2,\cdots\cdots n$）按数量的大小进行排列和分组，见表1。

T_{M_1}表示1月份最高温度，单位是℃。R_6表示次年6月份降水量，单位是毫米。

表1

序号	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
年	1970	1963	1969	1966	1965	1960	1961	1968	1964	1971	1959	1958	1972	1967	1962
预报因子T_{M1}	-6.8	-5.4	-5.1	-4.0	-3.4	-2.4	-2.2	-1.7	0.0	0.5	1.0	1.2	2.1	2.3	3.2
组号	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5
预报对象R_6	30.2	56.0	25.2	67.3	64.1	35.8	40.1	25.3	26.6	85.5	169.9	119.2	86.3	27.7	35.9

分组的数目要看样本资料的多少而定，在分组时尽量注意使每个组内包含的样本数基本相等。

按照因子的分组，则它所对应的预报对象也就被分了组，见表2。

表2　（单位 mm）

组号 / 组内样本号	1	2	3	4	5
1	30.2	67.3	40.1	85.5	86.3
2	56.0	64.1	25.3	169.9	27.7
3	25.2	35.8	26.6	119.2	35.9

为了下面叙述的方便，把方差分析的有关公式作一简单说明：

总平方和 $S=\sum_{i=1}^{a}\sum_{j=1}^{b}(y_{ij}-\bar{y})^2$

自由度为 ab-1

组间平方和 $S_1=b\sum_{i=1}^{a}(\bar{y}_i-\bar{y})^2$

自由度为 a-1

组内平方和 $S_2=\sum_{i=1}^{a}\sum_{j=1}^{b}(y_{ij}-\bar{y}_i)^2$

自由度为 ab-a

i：1—a 表示组号，j：1—b 表示组内样本号。表2中 a=5，b=3。

$\bar{y}$表示总平均值，$\bar{y}_i$表示第i组的平均值。

如令 $P=\frac{1}{ab}\left(\sum_{i=1}^{a}\sum_{j=1}^{b}y_{ij}\right)^2$

$$Q=\frac{1}{b}\sum_{i=1}^{a}\left(\sum_{j=1}^{b}y_{ij}\right)^2$$

$$R=\sum_{i=1}^{a}\sum_{j=1}^{b}y_{ij}^2$$

则 $S_1=Q-P$，　$S_2=R-Q$

$S=S_1+S_2=R-P$

平均组间平方和 $\overline{S_1}$ 为 S_1 除以相应的自由度

$$\overline{S_1}=\frac{S_1}{a-1}$$

平均组内平方和 $\overline{S_2}$ 为 S_2 除以相应的自由度

$$\overline{S_2}=\frac{S_2}{ab-a}$$

比值 $F=\frac{\overline{S_1}}{\overline{S_2}}$

以上的公式适用于每组的样本数都相等，但有时各组的样本数不一定相等，此时公式要作适当修改，其表达式如下：

$$P=\frac{1}{\sum_{i=1}^{a}b_i}\left(\sum_{i=1}^{a}\sum_{j=1}^{b_i}y_{ij}\right)^2$$

$$Q=\sum_{i=1}^{a}\frac{1}{b_i}\left(\sum_{j=1}^{b_i}y_{ij}\right)^2$$

$$R=\sum_{i=1}^{a}\sum_{j=1}^{b_i}y_{ij}^2$$

S的自由度为 $\sum_{i=1}^{a}b_i-1$

S_1的自由度为 a-1

S_2的自由度为 $\sum_{i=1}^{a}b_i-a$

其它完全一样，不再赘述。

从上述公式可以看出，当参加分析的数据和分组数目确定以后，R、P、a、b都是常数，只有Q是随着分组情况的变化而变化。

对表2而言

$S_1=17198$，　$S_2=6829.4$。

$a=5$，　$b=3$，　$F=6.295$。

从表2来看，R_6的分组是由T_{M_1}决定的，如果F值大，则说明同组的T_{M_1}所对应的预报对象的数值离散程度小，而不同组的T_{M_1}所对应的预报对象的数值离散程度大，也就是说预报对象的数值是受它的前期因子所处在那一组而定的，这样我们就可以说因子对预报对象的影响是显著的。因此，需通过F检验来判断，如表2；$F=6.295$查F分布表，分子自由度为4，分母自由度为10的F为5.99（1%），这说明T_{M_1}对R_6有显著影响是可信的。

当预报因子选定后，我们用下述方法来估计预报对象y的值。

当因子X_K的数值处在第i组时，则它对对象y的估计值为$y_{k_i}^*$（$i=1.2\cdots\cdots a$）。

用最小平方逼近的方法确定 y_{k1}^*

$$\sum_{i=1}^{a}\sum_{j=1}^{b}(y_{ij}-y_{k1}^*)^2=\text{最小}$$

令 $\sum_{i=1}^{a}\sum_{j=1}^{b}(y_{ij}-y_{k1}^*)^2=M$

显然，要使M取极小值必须满足下述条件

$$\frac{\partial M}{\partial y_{k1}^*}=0,\ \frac{\partial M}{\partial y_{k2}^*}=0,\ \cdots\cdots\ \frac{\partial M}{\partial y_{kA}^*}=0。$$

由此求得，

$$\begin{cases} y_{k1}^*=\sum_{j=1}^{b}y_{1j}/b \\ y_{k2}^*=\sum_{j=1}^{b}y_{2j}/b \\ \vdots \\ y_{KA}^*=\sum_{j=1}^{b}y_{aj}/b \end{cases}$$

从最后一式看出，因子 x_k 的数值处在第 i 组时，y_{ki}^* 就是该组内预报对象的平均值。例如表 2，

$y_{k1}^*=37.1$，　$y_{k2}^*=55.7$，

$y_{k3}^*=30.6$，　$y_{k4}^*=124.9$

$y_{k5}^*=49.9$。

显然由 y_{ki}^* 组成的第k个因子的估计序列不能完全逼近预报对象原序列（见表 3）。

再用残差序列为预报对象，重复上述过程。如此往复数次（实例中为 5 次）则使最后残差序列达到误差许可范围（实例中最大误差 3 mm）

表 3　　（T_{M1} 的组号是前一年的）

年	1959	1960	1961	1962	1963	1964	1965	1966	1967	1968	1969	1970	1971	1972	1973
原序列	119.2	169.9	35.8	40.1	35.9	56.0	26.6	64.1	67.3	27.7	25.3	25.2	30.2	85.5	86.3
T_{M1} 组号	4	4	2	3	5	1	3	2	2	5	3	1	1	4	5
T_{M1} 估计序列	124.9	124.9	55.7	30.6	49.9	37.1	30.6	55.7	55.7	49.9	30.6	37.1	37.1	124.9	49.9
残差序列	-5.7	45.0	-19.9	9.5	-14.0	18.9	-4.0	8.4	11.6	-22.2	-5.3	-11.9	-6.9	-39.4	36.4

下面就预报 6 月份降水量为实例说明本方法。

为了减少计算量，以 6 月降水距平为原序列（按多年平均60毫米计算）。表 4 中 x_1、x_2、x_3、x_4、x_5、x_6……x_{28}、x_{29}、x_{30} 分别代表所选的30个因子，其中x_6为 1 月最高气温，每栏中的数字为各组的样本号（分 5 组）。$y^{(1)}$为降水距平序列，$y^{(2)}$、$y^{(3)}$、$y^{(4)}$、$y^{(5)}$、$y^{(6)}$ 为残差序列。。

1.挑选对$y^{(1)}$影响显著的因子

$P=1.7$，$R=24029$，它们对任何因子而言均是常数。因此只要计算每个因子的Q值即可得F值。如不通过F检验则应扩大因子以供选择。

表 4

因子 / 年	x_1	x_2	x_3	x_4	x_5	x_6	……	x_{28}	x_{29}	x_{30}	$y^{(1)}$	$y^{(2)}$	$y^{(3)}$	$y^{(4)}$	$y^{(5)}$	$y^{(6)}$
1958	3	2	2	2	4	2	……	5	1	5	59	-6	5	-9	0	2
1959	3	5	5	5	4	4	……	5	5	3	110	45	12	1	0	1
1960	5	3	5	3	2	1	……	2	4	1	-24	-20	-15	-1	-2	-1
1961	1	5	3	4	3	1	……	3	2	4	-20	9	15	4	-3	-1
1962	5	4	5	2	5	5	……	1	3	4	-24	-14	-3	2	1	3
1963	5	4	4	5	1	5	……	4	3	1	-4	19	-14	-9	-2	0
1964	2	4	1	3	3	2	……	1	2	4	-33	-4	1	6	-2	0
1965	4	2	4	3	2	3	……	4	5	1	4	3	13	-1	6	1
1966	1	2	3	1	2	1	……	2	1	3	7	11	23	9	2	3
1967	1	1	4	4	5	4	……	4	4	2	-32	-22	-16	-11	-4	-3
1968	3	3	1	4	3	3	……	2	4	3	-35	-6	0	14	6	1
1969	2	1	1	2	1	2	……	3	3	2	-35	-12	-1	4	-4	-3
1970	4	5	2	1	1	3	……	3	2	5	-30	-7	5	-6	3	-2
1971	4	3	2	1	4	5	……	5	5	5	26	-39	-27	-13	-4	-2
1972	2	1	3	5	5	4	……	1	1	2	26	36	3	8	1	2

本例中 x_5 的 $F=6.295$，可以认为影响是显著的。

2.确定 x_5 的估计序列和残差序列 $y^{(2)}$，所得结果见表 5 和表 4 $y^{(2)}$ 列。

3.挑选对$y^{(2)}$影响显著的因子

这时 Q 值应根据 $y^{(2)}$ 来计算，因子 x_4 影响显著，$F=4.219$，$F(5\%)=3.48$。

4.确定 x_4 的估计序列和第二残差序列 $y^{(3)}$

见表 5 和表 4 $y^{(3)}$ 列，如此继续下去，求得对 $y^{(3)}$，$y^{(4)}$，$y^{(5)}$ 影响显著的因子分别为 x_2，x_3，x_6。

第五残差序列 $y^{(6)}$ 中最大值为 ±3mm，故已达到要求。计算表明，残差平方和随着引进因子呈指数衰减，这给迅速达到拟合程度带

来了方便。

1973年$x_5=-4.5$℃处于第一组和第二组之间，但靠近第二组，取估计值－4。$x_4=-3.8$℃处于第三组中，取估计值－5。$x_2=15$，处于第一组，取估计值－5。$x_3=-13.6$℃处于第三组，取估计值为7。$x_6=13.3$℃，处于第二组和第三组之间，但靠近第二组，取估计值－2。总估计值＝－4－5－5＋7－2＝－9。加上距平值60mm，故1974年$R_6=51$mm，实况52.8mm。

表5　（单位　mm）

年	1959	1960	1961	1962	1963	1964	1965	1966	1967	1968	1969	1970	1971	1972	1973
x_5 估计序列	65	65	－4	－29	－10	－23	－29	－4	－4	－10	－29	－23	－23	65	－10
x_4 估计序列	－11	33	－5	－6	－11	33	－5	－5	－12	－6	－6	－11	－12	－12	33
x_2 估计序列	14	11	－14	11	－5	－5	－5	14	14	－5	－14	－5	11	－14	－5
x_3 估计序列	－9	1	1	7	1	－7	8	－7	7	－7	8	8	－9	－9	7
x_6 估计序列	－2	－1	－1	－2	－2	－2	－2	5	－1	－1	5	－1	5	－2	－1
总估计序列	57	109	－23	－19	－27	－4	－33	3	4	－29	－36	－32	－28	28	24
$Y_{(1)}$	59	110	－24	－20	－24	－4	－33	4	7	－32	－35	－35	－30	26	26

应用方差分析来选择因子和作预报量的估计，计算量小，历史拟合程度高。但在分组数目的确定、出现两个相同样本数据时分组会有困难等方面还需进一步探讨和改进。