论备选因子集中的因子构成及选取

王双一

（总参气象局气象室）

提　要

本文从预报对象残差和预报因子残差之间的相关性出发，论证了备选因子集中的因子应该由两部分组成：一部分为与预报对象之间的相关性较好而因子之间的相关性较差或适中的因子，另一部分为与预报对象之间的相关性较差而与主因子之间的相关性很差或很好的因子；提出了选取备选因子集中因子的一般思路——配合主因子法。

一、前　言

在利用统计方法制作气象要素预报时，大都涉及到选取预报因子和建立备选因子集的问题。选取预报因子的一般思路为：在分析预报对象产生的物理过程的基础上，以预报经验为线索，通过普查对比、统计历史资料来寻找可能的预报因子。经过这种途径所选出的一批因子就构成了为建立预报模式所准备的备选因子集。众所周知，并不需要将所有的因子都作为备选因子集中的因子。通常认为，备选因子集中的因子应该满足两个条件：①因子和预报对象之间的相关性好且稳定；②因子之间的配合要好。因子间的配合好可以理解为：各因子对预报对象都具有较好的预报指示性，且这种预报指示性可以相互补充、相互修正，以期所建立的预报模式能有较好的预报准确率。一般具有相关性好的因子对预报对象的预报指示性往往是重复的，缺乏因子间的相互配合，这就是说备选因子集中各因子之间应该具有较好的独立性。

我们在建立预报模式的实践中知道，在各种气象要素之间或各种物理量之间，或多或少地存在着不同程度的相关性，要使备选因子集中的各因子在与预报对象之间相关性好的同时，又要满足相互之间的独立性好是非常困难的。在实际工作中，我们更多地考虑了相关因子之间的相互配合问题。本文从预报对象残差和预报因子残差之间的相关性出发，从理论上初步探讨了相关因子之间对预报指示性的相互配合作用，指出：对预报对象的预报指示性相互配合好的因子，不仅因子之间的独立性好，而且在一定条件下，相关好的因子之间对预报对象的预报指示性也具有较好的相互配合。在所给的实例中，说明了备选因子集中各因子的选取方法及这种因子间相关好的因子对提高模式预报准确率的作用。

二、预报对象残差和预报因子残差之间的相关性

在用逐步回归方法建立预报方程时，当第一步选入与预报对象相关性最好的因子 X_1 后，以后各步所选的都是因子残差与预报对象残差之间的相关性最好的因子，多元线性回归方程中因子的逐个选入过程，就是因子残差对预报对象残差的逐步回归订正的过程。所以，因子残差和预报对象残差之间的相关性的好坏，直接关系到所建方程的好

坏。

设预报对象Y和因子X_1及X_2，Y和X_2关于X_1的一元回归方程为：

$$\begin{cases}\hat{Y}=B_0+BX_1\\ \hat{X}_2=A_0+AX_1\end{cases}\tag{1}$$

X_2的残差X'_2和预报对象残差Y'分别为：

$$\begin{cases}X'_2=X_2-\hat{X}_2=X_2-A_0-AX_1\\ Y'=Y-\hat{Y}=Y-B_0-BX_1\end{cases}\tag{2}$$

（一）预报对象残差和预报因子残差之间的线性相关系数$R_{x'_2y'}$

记：$\overline{Y}=\frac{1}{n}\sum_{i=1}^{n}y_i$

$$\overline{X}_1=\frac{1}{n}\sum_{i=1}^{n}x_{1i}$$

$$\overline{X}_2=\frac{1}{n}\sum_{i=1}^{n}x_{2i}$$

$$S_{YY}=\sum_{i=1}^{n}(y_i-\overline{Y})^2$$

$$S_{x_1x_1}=\sum_{i=1}^{n}(x_{1i}-\overline{X}_1)^2$$

$$S_{x_2x_2}=\sum_{i=1}^{n}(x_{2i}-\overline{X}_2)^2$$

$$S_{x_1y}=\sum_{i=1}^{n}(x_{1i}-\overline{X}_1)(y_i-\overline{Y})$$

$$S_{x_2y}=\sum_{i=1}^{n}(x_{2i}-\overline{X}_2)(y_i-\overline{Y})$$

$$S_{x_1x_2}=\sum_{i=1}^{n}(x_{1i}-\overline{X}_1)(x_{2i}-\overline{X}_2)$$

其中n为样本数。

在(1)、(2)两式中，

$$\begin{cases}B=S_{x_1y}/S_{x_1x_1}\\ B_0=\overline{Y}-B\overline{X}_1\\ A=S_{x_1x_2}/S_{x_1x_1}\\ A_0=\overline{X}_2-A\overline{X}_1\end{cases}\tag{3}$$

由(3)式容易证明：

$$\begin{cases}\overline{Y'}=\frac{1}{n}\sum_{i=1}^{n}y'_i=0\\ \overline{X'_2}=\frac{1}{n}\sum_{i=1}^{n}x'_{2i}=0\end{cases}\tag{4}$$

记：$$S_{x'_2y'}=\sum_{i=1}^{n}(y'_i-\overline{Y'})(x'_{2i}-\overline{X'_2})=\sum_{i=1}^{n}y'_ix'_{2i}$$

$$=\sum_{i=1}^{n}\left[x_{2i}-\overline{X}_2-A(x_{1i}-\overline{X}_1)\right]\left[y_i-\overline{Y}-B(x_{1i}-\overline{X}_1)\right]$$

$$=S_{x_2y}-AS_{x_1y}-BS_{x_1x_2}+ABS_{x_1x_1}$$

$$S_{x'_2x'_2}=\sum_{i=1}^{n}x'^2_{2i}=\sum_{i=1}^{n}\left[x_{2i}-\overline{X}_2-A(x_{1i}-\overline{X}_1)\right]^2$$

$$=S_{x_2x_2}-2AS_{x_1x_2}+A^2S_{x_1x_1}$$

$$S_{Y'Y'}=\sum_{i=1}^{n}y'^2_i=\sum_{i=1}^{n}\left[y_i-\overline{Y}-B(x_{1i}-\overline{X}_1)\right]^2$$

$$=S_{YY}-2BS_{x_1y}+B^2S_{x_1x_1}$$

$$R^2_{X'_2Y'}=S^2_{X'_2Y'}/(S_{x'_2x'_2}S_{Y'Y'})$$

$$=\frac{(S_{x_2y}-AS_{x_1y}-BS_{x_1x_2}+ABS_{x_1x_1})^2}{(S_{x_2x_2}-2AS_{x_1x_2}+A^2S_{x_1x_1})(S_{YY}-2BS_{x_1y}+B^2S_{x_1x_1})}$$

将(3)代入上式得：

$$R^2_{X'_2Y'}=\frac{(S_{x_2y}-S_{x_1x_2}S_{x_1y}/S_{x_1x_1}-S_{x_1y}S_{x_1x_2}/S_{x_1x_1}+S_{x_1x_2}S_{x_1y}/S_{x_1x_1})^2}{(S_{x_2x_2}-2S^2_{x_1x_2}/S_{x_1x_1}+S^2_{x_1x_2}/S_{x_1x_1})(S_{YY}-2S^2_{x_1y}/S_{x_1x_1}+S^2_{x_1y}/S_{x_1x_1})}$$

$$= \frac{(R_{x_2Y} - R_{x_1x_2}R_{x_1Y})^2}{(1-R^2_{X_1Y})(1-R^2_{X_1X_2})} \qquad (5)$$

在(5)式中，$R_{x_1Y} = S_{x_1Y}/(S_{x_1x_1}S_{YY})^{\frac{1}{2}}$ 表示 X_1 和 Y 之间的相关系数；

$R_{x_2Y} = S_{x_2Y}/(S_{x_2x_2}S_{YY})^{\frac{1}{2}}$，表示 X_2 和 Y 之间的相关系数；

$R_{x_1x_2} = S_{x_1x_2}/(S_{x_1x_1}S_{x_2x_2})^{\frac{1}{2}}$，表示 X_1 和 X_2 之间的相关系数。

记：$a = R_{X_1Y}$，$b = R_{X_2Y}$，$x = R_{X_1X_2}$，显然有：$-1 \leqslant a, b, x \leqslant 1$。

由(5)得：

$$R^2_{X'_2Y'} = (b-ax)^2/[(1-a^2)(1-x^2)] \qquad (6)$$

$$R_{x'_2Y'} = \pm \left| b-ax \right| /[(1-a^2)(1-x^2)]^{\frac{1}{2}} \qquad (7)$$

（二）$R_{x'_2Y'}$ 随 a、b 及 x 的变化规律

在建立模式时，设 X_1 是首先被选入的因子，也就是说，在备选因子集中，X_1 与 Y 的相关性最好。在选入 X_1 后，根据各因子残差和预报对象残差之间的相关性，选入残差相关性最好的因子 X_2，即 $|a| > |b|$。

1. $R^2_{X'_2Y'}$ 关于 x 的极值点

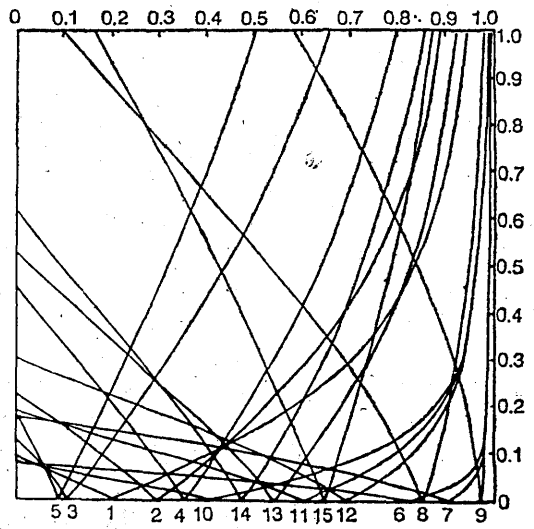
由 $(R^2_{X'_2Y'})'_x = 0$ 解得 $R^2_{X'_2Y'}$ 关于 x 的极值点为：

$$\begin{cases} x_1 = \dfrac{b}{a} \\ x_2 = \dfrac{a}{b} \end{cases}$$

由于 $|a| > |b|$，故极值点 $x_2 = b/a$ 是不符合 x 的定义域的，是没有意义的。所以 $x_1 = b/a$ 是 $R^2_{X'_2Y'}$ 在 x 定义域上的唯一极小值点，且有 $R^2_{x'_2Y'}|_{x=x_1} = 0$，即 X'_2 和 Y^2 在 $x = b/a$ 处的相关系数为零。

2. $R_{x'_2Y'}$ 随 a、b 及 x 的变化规律

在(7)式中，设 c、a 同号，且 $|b| = |c|$，当 a、b 同号时，有 $+R_{x'_2Y'}(x) = |c-ax| /[(1-a^2)(1-x^2)]^{\frac{1}{2}}$；当 a、b 异号时，有 $-R_{x'_2Y'}(x) = |c+ax| /[(1-a^2)(1-x^2)]^{\frac{1}{2}}$，所以有 $+R_{x'_2Y'}(x) = -R_{x'_2Y'}(-x)$，即 $R_{x'_2Y'}$ 在 a、b 同号和异号两种情况下，随 x 的变化成轴对称，在此仅考虑 a、b 同号且 $|a| > |b|$ 的情况下，$R_{x'_2Y'} = |b-ax| /\sqrt{(1-a^2)(1-x^2)}$ 随 a、b、x 的变化规律。附图中仅给出在几种不同的 a、b 取值下，$R_{x'_2Y'}$ 随 $x(0 \leqslant x < 1)$ 的变化规律。



附图　几种不同的a、b取值(见表1)下Rx'_2Y'随x(0≤x<1)的变化曲线

横坐标为x，纵坐标为Rx'_2Y'，1，2，3，……15为曲线序号

由附图可以看出，$R_{x'_2Y'}$ 随 a、b、x 的变化规律是：

1) 当 a 与 b 较接近时，X_1 和 X_2 之间的相关性越差即独立性越好时，$|R_{x'_2Y'}|$ 越大；

2) 当 a 比 b 大得多时，X_1 和 X_2 之间的相关性很差或很好，都有利于 $|R_{x'_2Y'}|$ 具有较大值，特别是当 X_1 和 X_2 之间的相关性很好时，尤能获得 $|R_{x'_2Y'}|$ 的大值；当 x

表1　　附图中a、b的15种取值

序号	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
a	0.40	0.60	0.80	0.80	0.90	0.10	0.20	0.80	0.90	0.20	0.30	0.40	0.70	0.80	0.90
b	0.08	0.18	0.08	0.28	0.08	0.08	0.18	0.68	0.88	0.08	0.18	0.28	0.38	0.38	0.68

落在$x_1=\frac{b}{a}$附近时，可造成$|R_{x'_2y'}|$的值很小；

3）当a和b相近且都较小时，无论X_1和X_2之间的相关性如何，都难以获得$|R_{x'_2y'}|$的较大值；

4）当a和b相近且都较大时，当X_1和X_2之间的相关性较差或适中时，均能获得较大的$|R_{x'_2y'}|$值。

必须指出的是，当将a、b、x作为独立变量，在〔-1，1〕之间任意取值时，在(7)式中有可能出现$|R_{x'_2y'}|>1$的情况，这与$|R_{x'_2y'}|\leqslant 1$的前提是不符的。出现这种情况的原因是，a、b、x之间并不是独立的，三者之间的变化是相互制约的，将a、b、x作为独立变量而任意取值时，当取了一组实际上不可能出现的a、b、x值后，就导致了$|R_{x'_2y'}|>1$的出现。

三、备选因子集中的因子构成及选取

（一）备选因子集中的因子构成

由前面的讨论可知，为了使所建立的模式具有较好的预报性能，备选因子集中的因子应该由如下两大部分构成。

1．预报因子和预报对象之间的相关性均较好，而因子之间的相关性较差或适中，由$R_{x'_2y'}$的变化规律可以看出，满足该条件的这一部分因子，保证了预报因子残差和预报对象残差之间具有较好的相关性，也即因子之间具有了较好的配合。

2．预报因子和预报对象之间的相关性较差，但与第一部分因子中的主因子（即与预报对象之间的相关性最好的因子）相关性很差或很好。在预报模式选入主因子后，满足该条件的这一部分因子保证了预报因子残差和预报对象残差之间具有较好的相关性，显示出这一部分因子与主因子之间对预报指示性的配合较好。

（二）备选因子集中因子选取的一般思路——配合主因子法

根据以上的讨论和实际因子选取的情况，这里给出了备选因子集中因子选取的一种可行方法：

首先，找出主因子。对于给定的预报对象，利用目前各种选因子的思路，找出与预报对象相关性最好的因子，即为主因子；

第二，选出备选因子集中的第一部分因子。根据主因子和预报对象之间的相关性，选出与预报对象之间的相关性同主因子与预报对象之间的相关性相近、且与主因子之间的相关性较差或适中的那一部分因子；

第三，选出备选因子集中的第二部分因子。根据主因子和预报对象之间的相关性，选出与预报对象之间的相关性比主因子与预报对象之间的相关性差得多、但与主因子之间的相关性很差或很好的那一部分因子。

由以上的主因子和另两部分因子构成了对预报对象建模式时所需的备选因子集。备选因子集中所包含的因子个数应多于模式可能包含的因子数。

四、实例

在本例中，预报对象为北京12月、1月及2月份每日的平均温度，样本为1984—1987年12月、1985—1988年1、2月，共12个月中的351个样本（不包括缺测）。利用当

天的资料制作第二天的日平均温度的预报。

（一）备选因子集中因子的选取——配合主因子法

以各日欧洲中期天气预报中心中期数值预报的500hPa及地面的网格点分析资料和24小时的预报资料、计算的各层地转风u、v分量和地转风涡度资料作为备选因子集中因子选取的对象。

1. 主因子X_1的选取。根据上述各物理量场的网格点资料，计算出各物理量与平均温度的相关系数，以与平均温度之间的相关系数最大的那个物理量作为备选因子集中的主因子X_1。

2. 选出除X_1以外的与平均温度之间的相关系数较大的那些物理因子，共选了14个，分别记为X_2，X_3…，X_{15}。

3. 选出与平均温度之间的相关系数很小但与X_1之间的相关系数很小或很大的那些因子。分别计算出X_1和平均温度与上述各物理量之间的相关系数，设某一物理量与X_1的相关系数为x，与预报对象之间的相关系数为b，而X_1和预报对象之间的相关系数为a，根据(7)式可以计算出该物理量残差和预报对象残差之间的相关系数，根据所计算的相关系数的大小，决定该物理量是否作为备选因子集中的因子之一而被选入；当没有根据(7)式计算出该物理量残差和预报对象残差之间的相关系数时，可根据$\frac{b}{a}$和x值的大小来决定，当$\frac{b}{a}$比x大得多或小得多时，说明该物理量残差和预报对象残差之间具有较好的相关性，此时，该物理量应该作为备选因子而被选入；当x值落在$\frac{b}{a}$附近时，该物理量不应该作为备选因子而选入。运用这一方法，我们共选取了这样的4个因子，分别记为X_{16}、X_{17}、X_{18}和X_{19}。

根据以上三个步骤，共选取了19个因子构成了备选因子集，其中部分因子的物理意义见表2。各因子与预报对象、主因子之间的相关系数及各因子残差和预报对象残差之间的相关系数见表3，表3还给出了各因子与预报对象之间的相关系数与主因子同预报对象之间相关系数的比值。

表2 备选因子集中部分因子的物理意义

因子	各因子所表示的物理意义
X_1	500hPa分析的位势高度格点值
X_4	地面分析的气压格点值
X_7	地面分析的地转风分量u_g格点值
X_{11}	地面分析的地转风分量v_g格点值
X_{13}	地面24小时预报的地转风分量v_g格点值
X_{16}	地面分析的气压格点值
X_{18}	500hPa分析的地转风分量v_g格点值

表3 各因子同预报对象和主因子之间的相关系数及残差之间的相关系数

类别	因子X_i	R_{X_iY}	$R_{X_iX_1}$	$R_{X'_iY'_i}$	$\frac{R_{X_iY}}{R_{X_1Y}}$
第一部分因子	X_1	0.58011	1.00000	0.00000	1.00000
	X_2	−0.45805	−0.11737	−0.48209	−0.78959
	X_3	0.51157	0.85196	0.04065	0.88185
	X_4	−0.38625	−0.23279	−0.31711	−0.66582
	X_5	−0.51638	−0.73704	−0.16140	−0.89014
	X_6	−0.55772	0.89144	−0.11004	0.96140
	X_7	−0.40338	0.02783	−0.51525	−0.69535
	X_8	−0.44597	−0.62666	−0.12993	−0.76877
	X_9	0.46292	0.73066	0.07028	0.79799
	X_{10}	0.39308	0.75452	−0.08343	0.67760
	X_{11}	0.47144	0.51655	0.24634	0.81267
	X_{12}	0.35124	0.59186	0.01206	0.60547
	X_{13}	0.35522	0.48651	0.10260	0.61233
	X_{14}	−0.51993	−0.78755	−0.12572	−0.89626
	X_{15}	−0.43306	−0.61272	−0.12061	−0.74651
第二部分因子	X_{16}	0.05490	0.54840	−0.38644	0.09464
	X_{17}	0.05239	0.48819	−0.32463	0.09031
	X_{18}	0.17101	0.54202	−0.20951	0.29479
	X_{19}	0.11583	0.39870	−0.15454	0.19967

（二）该备选因子集与预报对象的回归效果同不考虑第二部分因子时回归效果的比较

取F最小水平值为5.0，应用逐步回归方法，计算得出在该备选因子集的19个因子中，共有5个因子被选入预报对象的预报方程中，其预报方程为：

$$\hat{Y}_1 = -1133.676147 + 2.477134X_1 - 1.616685X_7 + 1.521886X_{11} - 1.795453X_{16} - 0.614396X_{18}$$

同样取F最小水平值为5.0，不考虑备选因子集中第二部分的4个因子，即由X_1，X_2，…，X_{15}构成备选因子集，利用逐步回归的方法，计算得出该15个因子中共有4个因子被选入到预报对象的预报方程中，该预报方程为：

$$\hat{Y}_2 = -1022.377686 + 2.233325X_1 - 1.617561X_4 - 3.137505X_7 - 1.756386X_{13}$$

上述两个方程的回归效果见表4。

由$\hat{Y}_1$方程和$\hat{Y}_2$方程分别对1988年12月和1989年1月及2月的88天（不包括缺测）进行了实际预报，各方程的预报效果见表4。

表4　　各方程的拟合效果和预报效果之比较

方　程	最低F水平值	入选方程因子数	复相关系数	拟合平均误差	预报平均误差
方程$\hat{Y}_1$	5.0	5	0.752980	1.727(℃)	1.830(℃)
方程$\hat{Y}_2$	5.0	4	0.741787	1.806(℃)	2.025(℃)

由表4可见，预报方程$\hat{Y}_1$的回归效果明显优于预报方程$\hat{Y}_2$的回归效果，而且实际预报效果也是前者明显优于后者。由$\hat{Y}_1$方程和$\hat{Y}_2$方程中所包含的因子差别可以看出，$\hat{Y}_1$方程中包含有X_{16}和X_{18}这两个第二部分因子，而$\hat{Y}_1$方程中只包含有第一部分的因子，这说明所选的第二部分因子虽然与预报对象之间的相关性较差，但与同预报对象之间具有较好相关性的第一部分因子之间、对预报对象的预报指示性具有较好的相互配合，第二部分因子对于提高回归方程拟合率和方程的预报准确率都是非常必要的，且是不可缺少的。必须指出，备选因子中的第二部分因子，在往常的因子选取中往往是不可能被选入到备选因子集中的，这正是以往在选取备选因子集中的因子时所存在的弊病之一。

五、讨论

通过上述的讨论和实例说明，得出了备选因子集中的因子应该由两部分组成，这两部分因子与预报对象之间的相关性是不同的，但这两部分因子对预报对象的预报指示性却构成了较好的配合。我们在建立预报模式时，根据现有的可供选取因子的资料，要使预报模式具有尽可能好的预报准确率，仅考虑选取与预报对象之间相关性好的因子是不够的，并且这部分因子中的某些因子对预报指示性并不表现出好的配合，应该更加注意选取与预报对象之间的相关性不很好、但残差之间的相关性却较好的第二部分因子。

根据文中所提出的“配合主因子”的因子选取方法，由表3可以看出，尽管X_3、X_9、X_{10}和X_{12}因子与预报对象之间的相关性都很好，但其残差和预报对象残差之间的相关性都很差，这说明，这些因子可以不被选入到备选因子集中，在实际建立方程的过程中，这些因子也没有被选入到方程中。当然这些因子被选为第一部分因子，并不会影响到所建预报模式的预报质量。

根据上面的讨论，可以看出，对于第一部分因子的选取，为了方便起见，我们可以

不用配合主因子法，而可以将所有的与预报对象之间的相关性较好的因子均作为第一部分因子；但对于第二部分因子的选取，我们不能将所有与预报对象之间相关性不好的因子都选入，导致备选因子集显得过于庞大，应该使用配合主因子法选出尽可能多的第二部分因子。

制作统计预报模式时，选取好第一部分因子后，是否能够利用配合主因子法选取出更多更好的第二部分因子，这是影响我们所建立的模式能否具有较好预报质量的关键。

The constitution of the candidate predictors and the method of selecting predictors

Wang Shuangyi
(Weather Service, Meteorological Bureau, Headquarters of the General Staff PLA)

Abstract

Discussions of the correlation between the residuals of predictand and predictor suggested that the candidate predictors consist of two different parts of Predictors.Predictors of the first part are these that the correlation between the predictand and predictor is better, while the correlation between one and another of the predictors is not good.Predictors of the second part are these that the correlation between the predictand and predictor is very bad while the correlation between the main predictor and the predictors is best or worst. And a method of selecting predictors which are made up of the candidate predictors is put forward and called the Method suited to the main predictor.