第50卷 第4期 气 象 学 报 Vol. 50, No. 4
1992 年 11 月 ACTA METEOROLOGICA SINICA Nov., 1992

最优化因子处理及加权多重回归模型*

汤志成 孙 涵

(江苏省气象科学研究所,南京,210008)

因子的优劣是回归分析的关键。故在建立回归方程时，一般除对预报因子进行直线相关普查外,还要进行非线性相关普查。如将原因子 x 用 x^{-1}、$x^{1/2}$、x^2、e^x、$\ln x$ 等函数形式进行变换。但由于这些函数形式有限,故不一定能找到最优的表达形式;为此冯耀煌等[1]又给出了 x^a 和 e^{ax} 两种通式，其中 a 为待定参数,由最优化方法求得,从而大大提高了因子的相关系数。可是以上各种函数形式均反映因子与预报量之间为单调关系,有时并不适当。如太阳辐射、温度、降水等与产量的关系,既不是越多越好，也不是越少越好。若采用二次三项式变换,因子普查时又极其不便,且表达形式仍不够丰富。为此本文设计了因子线性和非线性化处理的一个变换通式，并根据最优化原理给出了待定参数的计算方法。

1. 因子的最优化处理

由于对因子 x 的非线性化处理是在有限区间内进行的,且是寻找离散序列的统计规律,故 e^{ax},$\ln x$ 等在一定区间内的曲线与 x^a 的某一段曲线是可能相似的。试算结果也证明 $\ln x$,e^{ax},ax^2+bx+c 与预报量 y 的相关系数均小于或接近于 x^a 与 y 的最优相关系数。因此,在实际工作中可以只考虑计算简单的幂函数形式。对非单调的单峰(谷)型因子，可以用 $|x-b|$ 变换成新的单调因子。由于幂函数在1附近有许多特殊性质,故因子 x 的线性和非线性化(包括单调的和非单调的单峰(谷)型)处理可归纳为一种通用变换形式。

$$x'=(|x-b|/B+0.5)^a \tag{1}$$

式中 a，b 为待定参数,且 $x_{min}\leqslant b\leqslant x_{max}$,$B=\max(x_{max}-b,b-x_{min})$。因子经式(1)变换后,$x'$ 与 y 必为单调关系,且$(|x-b|/B+0.5)$的值在区间$[0.5,1.5]$内变化。从式(1)本身来说 a 可取任意实数，b 可取 x 域内的任意值;但实际选因子时，a，b 只在一定区域内取值才有实用意义。当 $b=x_{min}$ 时,表示原因子 x 与 y 为单调关系，如千粒重与产量的关系;当 $b=x_{min}$ 且 $a=1$ 时，表示 x 与 y 为线性关系。当 $x_{min}<b<x_{max}$ 时，表示 x 与 y 为单峰(谷)型关系,如降水、温度等因子。对于单峰(谷)型关系的因子，为了避免 x 在最低或最高值附近出现的个别样本的偶然误差影响，b 的取值以 $x_{min}+(x_{max}-x_{min})/4\leqslant b\leqslant x_{max}-(x_{max}-x_{min})/4$ 为宜。至于 a 值，根据我们的实际工作经验，一般在 $(-10,-1/10)$ 和 $(1/10,10)$ 两个区间内取值效果较好。

待定参量 a，b 可根据最优化法求出。令目标函数为

$$f(a,b)=1-R^2=\min \tag{2}$$

R 为 a，b 取一定值时 x' 与 y 的相关系数。由于这不同于一般函数形式的多元寻优，而是对相关系数寻优,故应用二维寻优的变量转换法思路将其分解为一元问题逐步处理。计算步骤：

1) 令 b 分别为 b_1,b_2,b_3,且 $b_1<b_2<b_3$,用外推内插法[3]分别对 a 寻优,得 f_1,f_2,f_3;

2) 由有关 f 值对 b 寻优得 b_i;

3) 求 b_i 下的最优 a 得 f_i;

* 本文于1989年11月14日收到,1990年12月1日收到最后修改稿。

4) 重复 2)、3)步计算，直至满足式(2)，其 a，b 值即为所求。

经此变换寻优的因子普查过程后获得的因子，是一批与产量相关系数最高的因子。对这些因子再经稳定性检验和独立性检验等方法进一步筛选，供建模之用。当然也可从稳定的因子中，选取物理意义明确的部分因子使用有偏估计等现代回归分析方法建立预报模型。

2. 加权多重回归模型

多重回归模型常应用于多时效的动态产量预报[2-4]。其数学模型为

阶乘型
$$\hat{Y}_j=\hat{Y}_t\prod_{i=1}^{j}\hat{Y}_{wi} \tag{3}$$

累加型
$$\hat{Y}_j=\hat{Y}_t+\sum_{i=1}^{j}\hat{Y}_{wi} \tag{4}$$

式中 $\hat{Y}_j$ 为第 j 次产量预测值，$\hat{Y}_t$ 为趋势产量值，$\hat{Y}_{wi}$ 为第 i 生育阶段气象产量的剩余预测值[5]。这种方法的特点是可以随生育期的进展逐步订正前期预报。但事实上作物不同生育阶段对产量形成的影响是不同的，在模式中应考虑各生育阶段对产量影响的权重，同时还要考虑各生育阶段预报子模式的模拟效果。若各生育阶段与产量的关系可用相关系数 r_i 表示，则各生育阶段对产量的权重就可表示为相关系数的函数。各子模式的模拟效果也可用其复相关系数 R_i 表示。于是式(3)、(4)可改为加权多重回归模型

阶乘型
$$\hat{Y}_j=\hat{Y}_t\left(\prod_{i=1}^{j}\hat{Y}_{wi}^{\mu_i^{(j)}}\right)^{1/j} \tag{5}$$

累加型
$$\hat{Y}_j=\hat{Y}_t+\sum_{i=1}^{j}\mu_i^{(j)}Y_{wi} \tag{6}$$

式中 $\hat{Y}_j$ 为 Y 的第 j 次预报值，$\hat{Y}_{wi}$ 为第 i 生育阶段 Y_w 的估计值，是由 p 个气象因子组成的子模式

$$\hat{Y}_{wi}=f_i(x_p)\qquad i=1,\cdots,j$$

求得，$\mu_i^{(j)}$ 为第 j 次预报时 $\hat{Y}_{wi}$ 对 $\hat{Y}_t$ 的权重，其中

阶乘型
$$\mu_i^{(j)}=r_iR_i\Big/\left(\prod_{k=1}^{j}r_kR_k\right)^{1/j}\qquad \begin{matrix} i=1,\cdots,j \\ r_i>0,R_i>0 \end{matrix} \tag{7}$$

累加型
$$\mu_i^{(j)}=r_iR_i\Big/\sum_{k=1}^{j}r_kR_k\qquad \begin{matrix} i=1,\cdots,j \\ r_i>0,R_i>0 \end{matrix} \tag{8}$$

它包含了各生育阶段对产量形成的贡献及各子模式拟合效果这两个方面的共同作用。

对于使用逐步回归或其它现代回归分析方法建立第 j 个生育阶段的子模式 $\hat{Y}_{wj}=f_j(x_p)$ 时 $\hat{Y}_{wj}$ 的样本序列为

阶乘型
$$Y_{wj}=\left(Y_w^j\Big/\prod_{i=1}^{j-1}\hat{Y}_{wi}^{\lambda_i^{(j)}}\right)^{1/\lambda_j^{(j)}} \tag{9}$$

累加型
$$Y_{wj}=\left(Y_w-\sum_{i=1}^{j-1}\lambda_i^{(j)}\hat{Y}_{wi}\right)/\lambda_j^{(j)} \tag{10}$$

它既包含了 Y_w 的全部信息，又包含了前阶段预报的剩余信息。其中 $\lambda_i^{(j)}$ 为第 j 次预报时第 i 阶段对产量的影响程度

阶乘型
$$\lambda_i^{(j)}=r_i\Big/\left(\prod_{k=1}^{j}r_k\right)^{1/j}\qquad i=1,\cdots,j \tag{11}$$

累加型 $$\lambda_i^{(j)}=r_i\Big/\sum_{k=1}^{j}r_k \qquad i=1,\cdots,j \tag{12}$$

并分别满足$\prod_{i=1}^{j}\lambda_i^{(j)}=1$(阶乘型) 和$\sum_{i=1}^{j}\lambda_i^{(j)}=1$(累加型)。

3. 应用与效果

将南京市冬小麦全生育期分为穗数、粒数、粒重三个形成阶段，根据田间试验资料求得穗数、粒数、千粒重与产量的相关系数分别为 $r_1=0.9041, r_2=0.6837, r_3=0.7985$。由式(11)、(12)分别求得阶乘型和累加型的各生育阶段对产量的影响程度 $\lambda_i^{(j)}$ 如表 1

由式(9)、(10)分别求得阶乘型和累加型三个阶段带权的气象产量序列，再通过逐步回归方法筛选经变换寻优后的因子分别建立阶乘型和累加型的三个阶段 $\hat{Y}_{wi}$ 子模式

阶乘型

$$\hat{Y}_{w1}=0.785931+0.169482\,x'_{11}-0.0291501\,x'_{12}+0.0191746\,x'_{13}-0.0116776\,x'_{14}-0.0029333\,x'_{15}-0.00617437\,x'_{16}+0.0547559\,x'_{17}$$

$$\hat{Y}_{w2}=0.939011+0.0681732\,x'_{21}-0.0962506\,x'_{22}+0.00621146\,x'_{23}+0.00733432\,x'_{24}$$

$$\hat{Y}_{w3}=0.76037+0.051103\,x'_{31}-0.0079986\,x'_{3}+0.146699\,x'_{33}$$

累加型

$$\hat{Y}_{w1}=0.214069-0.002933\,x'_{11}+0.0191746\,x'_{12}-0.00617437\,x'_{13}-0.011676\,x'_{14}-0.02915\,x'_{15}+0.169482\,x'_{16}+0.0547559\,x'_{17}$$

$$\hat{Y}_{w2}=0.593495+0.0667873\,x'_{21}-0.00662491\,x'_{22}-0.0995498\,x'_{23}-0.0756103\,x'_{24}$$

$$\hat{Y}_{w3}=-0.0578269+0.0633142\,x'_{31}-0.0119493\,x'_{32}-0.0142608\,x'_{33}$$

式中 x'_{ij} 均为由式(1)变换经优化处理后的气象因子。根据上述各子模式的复相关系数 R_i 由式(7)、(8)分别求出阶乘型和累加型的权重 $\mu_i^{(j)}$ 如表 2。

表 1 各生育阶段对产量形成的影响程度

阶乘型				累加型			
i	$\lambda_i^{(1)}$	$\lambda_i^{(2)}$	$\lambda_i^{(3)}$	i	$\lambda_i^{(1)}$	$\lambda_i^{(2)}$	$\lambda_i^{(3)}$
1	1	1.1500	1.1440	1	1	0.5694	0.3789
2		0.8696	0.8651	2		0.4306	0.2865
3			1.0104	3			0.3346
Π	1	1	1	Σ	1	1	1

表 2 各次预报模式中不同生育阶段的权重$\mu_i^{(j)}$

阶乘型				累加型			
i	$\mu_i^{(1)}$	$\mu_i^{(2)}$	$\mu_i^{(3)}$	i	$\mu_i^{(1)}$	$\mu_i^{(2)}$	$\mu_i^{(3)}$
	1	1.192250	1.265010	1	1	0.586422	0.415135
		0.838747	0.889927	2		0.413578	0.292789
			0.888287	3			0.292058
	1	1	1	Σ	1	1	1

由表 2 即可获得三个阶段的加权重多重回归预报模型：

阶乘型

$$\hat{Y}_1=\hat{Y}_t\cdot\hat{Y}_{w1}$$

$$\hat{Y}_2=\hat{Y}_t(\hat{Y}_{w1}^{1.19225}\cdot\hat{Y}_{w2}^{0.838747})^{1/2}$$

$$\hat{Y}_3=\hat{Y}_t(\hat{Y}_{w1}^{1.26501}\cdot\hat{Y}_{w2}^{0.889927}\cdot\hat{Y}_{w3}^{0.88828})^{1/3}$$

累加型(此处为相对气象产量形式)

$$\hat{Y}_1=\hat{Y}_t(1+\hat{Y}_{w1})$$

$$\hat{Y}_2=\hat{Y}_t(1+0.586422\,\hat{Y}_{w1}+0.413578\,\hat{Y}_{w2})$$

$$\hat{Y}_3=\hat{Y}_t(1+0.415153\,\hat{Y}_{w1}+0.292789\,\hat{Y}_{w2}+0.292058\,\hat{Y}_{w3})$$

经计算试报 1988 年的产量误差情况如表 3，完全可以满足业务使用需要。

表 3　各阶段产量预报值及其误差

模型	阶乘型		累加型	
项目	预报值(kg)	误差(%)	预报值(kg)	误差(%)
一阶段	221.8	3.14	221.8	3.14
二阶段	226.4	1.22	226.2	1.14
三阶段	227.7	0.55	227.5	0.65

表 4　因子优化处理与不处理的效果比较

生育阶段	显著性相关的因子个数				预报模式的复相关系数		预报的相对误差%	
	$r>r_{0.05}$		$r>r_{0.01}$					
	优化	未优化	优化	未优化	优化	未优化	优化	未优化
1	32	7	19	1	0.96	0.71	3.14	4.32
2	16	1	9	0	0.97	0.78	1.22	5.81
3	19	6	12	0	0.97	0.80	0.55	4.86
栏	1		2		3		4	

4. 小　　结

1. 经优化处理后的因子与产量的关系较为合理；而一般用光温水等因子的原始序列与产量进行相关普查，只能找出显著性相关的单调关系的因子，实际上不太合理。如降水、温度等对产量的贡献既不是越多越好，也不是越少越好。本文的因子优化处理可有效地解决这一问题。

2. 从表 4 上清楚地看出：因子优化处理后，大幅度地增加了显著相关的因子个数（1 栏）；明显地提高了单因子的相关显著性（2 栏）；提高了预报方程的复相关系数（3 栏）；降低了预报误差（4 栏）。无庸置疑，因子优化处理的效果极为显著。此为累加型模型情况，阶乘型亦然。

3. 加权多重回归考虑了各生育阶段对产量影响的权重，使分段预报模型更为合理。

参 考 文 献

[1] 冯耀煌等，论最优预报因子与最优预报方程，气象学报，47，1，52—60，1989。
[2] 宛公展等，华北地区冬小麦产量预报模式的初步研究，气象学报，42，3，356—362，1984。
[3] 范鸣玉等，最优化技术基础，111—142，清华大学出版社，1982。
[4] 汤志成等，江苏双季早稻产量预报的累加型模式，江苏农业学报，2，3，12—19，1986。
[5] 刘树泽等，作物产量预报方法，114，气象出版社，1987。

THE OPTIMAL TRANSFORMATION OF PREDICTOR AND WEIGHTED MODEL OF MULTI-REGRESSION

Tang Zhicheng　　Sun Han

(*Jiangsu Research Institute of Meteorological Science, Nanjing*, 210008)

Abstract

By discussing the relationship between predictor and predictand (i.e. yield) in yield prediction, the basic form of linear and nonlinear transformation of predictor was summed up. The calculation method of parameter in the form was qiven out, based on the principles of optimization method, and at the same time the optimal predictors were selected. In addtion, the weighted model of multi-regression for yield prediction was established, based on the stage of plant.