第52卷 第2期　　气 象 学 报　　Vol. 52, No. 2
1994年5月　　ACTA METEOROLOGICA SINICA　　May, 1994

北京冬小麦产量预报的一种建模方案*

王春乙　　潘亚茹　　白月明　　温民

（中国气象科学研究院，北京，100081）

提　　要

针对农业产量气象预测的各类统计模式中所存在的某些不足，如趋势产量外推的稳定性，预报因子的物理及生物学意义、相关阶段性、多元共线性，回归系数稳定性和剩余产量的处理等，提出了不同的解决办法。在北京冬小麦产量预报试验中收到较好的效果。

关键词：主分量分析，冬小麦，岭回归，ARMA模型。

1　引言

近年来，我国农业气象工作者在农作物产量的农业气象统计预测方面做了大量的工作，提出了适合于我国不同区域、不同时效的农作物产量统计预报模式，多数模式已投入了业务使用。

影响统计模式的预测精度因素是多方面的：如趋势产量外推的稳定性，预报因子的物理及生物学意义、相关阶段性、多元共线性和因子覆盖度，回归系数稳定性及剩余产量的处理等。本文针对上述诸问题，提出了不同的解决方案，建立了性能稳定的北京冬小麦产量业务预报的统计模型。

2　社会产量的分离及剩余产量的定义

本文将实际社会产量[1)](y_t)分离成两部分：它们是趋势产量(f_t)和气象产量(p_t)，表达式如下。

$$y_t = f_t + p_t \tag{1}$$

由于趋势产量和气象产量在实际拟合过程中，均有一定的偏差，因而本文定义趋势产量和气象产量的拟合偏差之和为剩余偏差产量(w_t)，简称为剩余产量，表达式为：

$$w_t = y_t - \hat{f}_t - \hat{p}_t$$

式中，$\hat{f}_t$ 和 $\hat{p}_t$ 分别为趋势产量和气象产量的拟合值。

3　产量预报模式的建立

* 1992年5月20日收到初稿，1992年10月8日收到最后修改稿。

1)　本文采用的冬小麦单产资料来源于国家统计局(1962—1988年)

3.1 趋势产量模式的建立

本文采用正交多项式方法组建趋势产量方程，表达式为：

$$\hat{f}_t = 170.965 + 8.184t + 0.084t^2 + 0.008t^3 + 0.005t^4 \quad (2)$$

式中，t 为年序。方程的复相关系数为 0.9570，剩余方差为 27.924，模拟结果详见图 1。

3.2 气象产量模式的建立

本文建立气象产量预报模式所用的预报因子有两类，第一类是大气环流因子，它们是副高、东亚槽、极涡、冷空气次数及欧亚和亚洲纬、经向环流指数等 14 个大气环流特征值，本文分别采用上述各特征值的月值资料，构成了大气环流因子群，即总共有 14×8 个待挑选的大气环流因子。第二类是气象产量本身的周期分量因子。

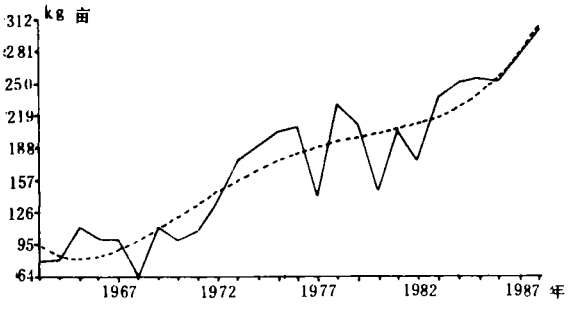


图 1 趋势产量拟合图

（——为实际社会产量，……为趋势产量）

3.2.1 第一类因子的粗选与处理

1）相关普查：利用上述 112 个大气环流因子同气象产量进行相关普查，粗选 18 个大气环流因子（见表 1）。

表 1 粗选的大气环流因子

因子	因子名称	相关系数	因子	因子名称	相关系数
x_1	上年 9 月副高西伸脊点	−0.3992	x_{10}	上年 11 月欧亚纬向环流指数	0.5206
x_2	上年 11 月副高脊线平均位置	−0.3879	x_{11}	当年 1 月欧亚纬向环流指数	0.3818
x_3	上年 12 月副高脊线平均位置	−0.4659	x_{12}	上年 11 月欧亚经向环流指数	−0.5086
x_4	当年 1 月冷空气次数	−0.5196	x_{13}	当年 2 月欧亚经向环流指数	0.4707
x_5	上年 12 月东亚槽平均位置	0.4775	x_{14}	上年 10 月亚洲纬向环流指数	0.4760
x_6	当年 3 月极涡位置	0.3851	x_{15}	上年 11 月亚洲纬向环流指数	0.5306
x_7	当年 4 月极涡位置	0.4034	x_{16}	当年 1 月亚洲纬向环流指数	0.4024
x_8	上年 12 月极涡强度	−0.4491	x_{17}	上年 11 月亚洲经向环流指数	−0.5921
x_9	上年 9 月欧亚纬向环流指数	0.4919	x_{18}	当年 2 月亚洲经向环流指数	0.3853

2）选取生物学意义显著的因子：利用上述 18 个环流因子逐个同上年 9 月至当年 4 月冬小麦生育期内的旬平均气温（T）、旬降水量（r）进行相关分析，选取两者相关显著的因子（见表 2）。

3）确定粗选因子的相关阶段性：在计算气象产量同大气环流因子的相关时，往往会发现有的因子后半段相关性不如前半段，这样的因子预报能力将有所下降。为了提高预报精度和增强因子的可靠性，笔者将预报因子和气象产量序列分成前后两段，分别计算相关性，前段标准为通过 0.05 水平，后段标准为通过 0.01 水平。

经过上述三步筛选，选取的 8 个大气环流因子，依次为 $x_1, x_2, x_7, x_8, x_{12}, x_{13}, x_{15}, x_{16}$。

4）消除因子间的多元共线性，提高因子覆盖度：因选取的因子偏多，且彼此之间又存

在着多元共线性。为了解决这个问题,本文采用主分量分析方法,对选取的因子进行浓缩处理,吸取所有因子的有效信息[2,3]。选取主分量组合因子作为新的预报因子。

表 2　大气环流因子与 T 和 r 的相关系数

大气环流因子	同旬平均气温或降水量的相关性
x_1	$R_{11中r}=0.4847$　$R_{9上T}=-0.4422$　$R_{9中T}=-0.3813$　$R_{11下T}=-0.3883$
x_2	$R_{10中r}=-0.5636$　$R_{10下r}=-0.6069$　$R_{12上r}=-0.4785$　$R_{9中T}=-0.7828$　$R_{10上T}=-0.3831$
x_6	$R_{10上r}=0.4752$　$R_{10下r}=-0.3939$　$R_{12下r}=0.3986$　$R_{2上r}=0.4489$　$R_{10中T}=-0.4131$
x_7	$R_{4中r}=-0.3989$　$R_{3下T}=-0.4790$
x_8	$R_{12中r}=-0.3980$　$R_{11上T}=-0.4394$
x_{10}	$R_{11下r}=-0.4393$　$R_{11上T}=0.4771$　$R_{12上T}=0.5364$
x_{11}	$R_{9上r}=-0.4093$　$R_{11中T}=0.4363$　$R_{1上T}=0.4509$　$R_{1下T}=0.5103$　$R_{2上T}=0.5194$
x_{12}	$R_{9上r}=0.4517$　$R_{11下T}=-0.5790$　$R_{12上T}=-0.4813$　$R_{1中T}=-0.3915$
x_{13}	$R_{9上r}=-0.3871$　$R_{12上r}=0.6043$　$R_{2中r}=0.4030$　$R_{9中T}=0.4589$
x_{15}	$R_{11下r}=-0.4182$　$R_{10上T}=0.4858$　$R_{11上T}=0.5218$　$R_{12上T}=0.6719$
x_{16}	$R_{11中T}=0.5094$　$R_{1下T}=0.6290$　$R_{1上T}=0.4472$　$R_{2上T}=0.4034$
x_{17}	$R_{11下T}=-0.4878$　$R_{12上T}=-0.4156$
x_{18}	$R_{10上T}=0.4939$　$R_{12上T}=0.4234$

注:$R_{11中r}$　表示 x_1 同 11 月中旬降水量的相关系数

$R_{9上T}$　表示 x_1 同 9 月上旬平均气温的相关系数

$R^{n}_{0.05}=0.38$　$R^{n}_{0.01}=0.49$　$n=26$

利用上述 8 个环流因子构成原始矩阵 f_{ij}, $i=1,2,\cdots\cdots H$, $j=1,2,\cdots\cdots N$。H 为因子数,N 为样本数。对 f 矩阵进行中心化处理,得距平矩阵 f_{ij}。

$$f_{ij}=f_{ij}-\overline{f_i} \tag{3}$$

$$\overline{f_i}=\sum_{j=1}^{N}f_{ij}/N \tag{4}$$

$$S=f'\cdot f'^{T} \tag{5}$$

式中,S 为协方差矩阵,f'^{T} 为 f' 的转置矩阵。利用雅可比算法求其 S 矩阵的特征向量矩阵 U,进而求得主分量矩阵 X

$$X=U\cdot f' \tag{6}$$

利用主分量组合因子同气象产量再进行相关分析,得知,第 1,2,3 主分量同气象产量的相关系数分别为 0.5392、0.5328、0.4191,表述式依次为:

$$X_1=-0.096x_1-0.003x_2+0.992x_7-0.03x_8-0.009x_{12}+0.049x_{13}+0.055x_{15}+0.013x_{16} \tag{7}$$

$$X_2=-0.062x_1-0.009x_2-0.029x_7-0.028x_8-0.06x_{12}+0.068x_{13}+0.095x_{15}+0.989x_{16} \tag{8}$$

$$X_3=-0.181x_1-0.015x_2-0.074x_7-0.031x_8-0.20x_{12}+0.052x_{13}+0.95x_{15}-0.122x_{16} \tag{9}$$

其余主分量相关不显著。第 1,2,3 特征值的总方差贡献为 97.3%,说明了前三个主分量组合因子集中了初始 8 个环流因子有用信息的 97.3%。

3.2.2 第二类因子的粗选与处理

本文在选取预报因子时附加了气象产量的均值周期因子，这样处理会使方程更加稳定，从而达到提高模式精度之目的。均值周期因子的生成如下[4,5]：

设气象产量序列为 $P_{(t)}=\{P_{(1)},P_{(2)},\cdots\cdots P_{(N)}\}$ (10)

式中，N 为样本数。$P_{(t)}$ 的算术平均值为：

$$\overline{P}_{(t)}=\frac{1}{N}\sum_{t=1}^{N}P_{(t)} \tag{11}$$

拓广(11)式，定义(10)式的均值周期分量因子为：

$$\overline{P}_{l(i)}=\frac{1}{N_l}\sum_{j=0}^{N_l-1}P(i+jl)\qquad i=1,2,\cdots\cdots,l\qquad 1\leqslant l\leqslant m \tag{12}$$

式中，N_l 为满足 $N_l\leqslant\left[\frac{N}{l}\right]$ 的最大整数，M 取 $\left[\frac{N}{2}\right]$。

对 $\overline{P}_l$ 作周期性延拓有：

$$f_{l(i)}=\overline{P}_{l(i)}\qquad t=i\bmod(l),t=1,2,\cdots\cdots,N \tag{13}$$

这里 Mod 表示同余，由此构造出 $P_{l(i)}$ 的外延矩阵，

$$F=(fij)_{N,M} \tag{14}$$

$$fij\equiv f_l(t) \tag{15}$$

即均值周期因子矩阵。

利用生成的均值周期分量因子同气象产量进行相关分析，选取的因子为气象产量的6年(x_1')、7年(x_2')、……、13年(x_8')周期分量因子，相关系数分别为：0.5152，0.5200，0.4784，0.8222，0.6609，0.6488，0.8012，0.6734。同样利用主分量和相关分析，得知第1主分量因子同气象产量的相关系数为0.9584，表达式为：

$$X_1'=0.233x_1'+0.159x_2'+0.138x_3'+0.511x_4'+0.351x_5'+0.351x_6'+0.522x_7'+0.511x_8' \tag{16}$$

其余主分量相关不显著。这一主分量因子集中了原始8个周期分量因子有用信息的82%。

经过上述几步的筛选与处理，最后选取了4个主分量因子，作为组建模式的预报因子。本文采用岭回归分析方法组建气象产量的回归预测模型。

岭回归系数的估计式为[6]：

$$\widehat{\beta}_k=(X'X+kI)^{-1}X'Y \tag{17}$$

式中，X 为自变量矩阵，X' 为 X 的转置，Y 为因变量矩阵，I 为单位矩阵，k 为给定的常数。据 Hoerl 等人的研究[7]，k 值满足下式关系：

$$0<k<\sigma^2/\max\ r_i^2 \tag{18}$$

则由式(17)定义的岭回归系数便更接近于理论回归系数值。其中，$r=p^*\beta$，P^* 为正交矩阵，由 $X'X=P^*\wedge P$ 得出，$\wedge$ 为对角矩阵。为了方便，将 $\sigma^2/\max\ r_i^2$ 记为 k_l，k_l 的确定详见文献[7]。在实际计算过程中，取 $k=0,0.05k_l,0.10k_l,0.15k_l,\cdots\cdots,k_l$，选取方差最小者。

计算结果表明，当 $k=0.35$ 时，剩余标准差最小，则对应于 $k=0.35$ 时的岭回归预测模型为：

$$\widehat{P}_{(t)}=-2.9696+0.178X_1+0.1008X_2+0.0566X_3+0.622X_1' \tag{19}$$

方程的复相关系数为 0.9736,剩余标准差为 4.1984。模拟结果详见图 2。

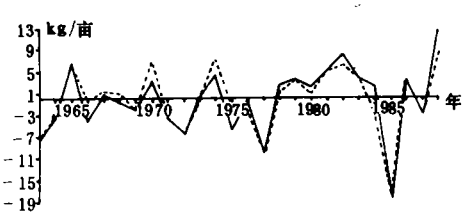


图 2　气象产量拟合图

(—— 气象产量,…… 模拟的气象产量)

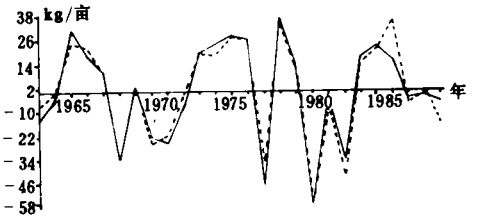


图 3　剩余产量拟合图

(—— 剩余产量,…… 模拟的剩余产量)

3.3 剩余产量模型的建立

这里假定剩余产量近似零均值,服从正态分布的平稳线性时间序列,这样剩余产量可用线性的 ARMA(p,q) 模型来模拟[8,9]。

$$w_t - \phi\omega_{t-1} - \phi_2 w_{t-2} \cdots\cdots - \phi_p w_{t-p} = a_t - Q_1 a_{t-1} - Q_2 a_{t-2} \cdots\cdots - Q_q a_{t-q} \tag{20}$$

式中,$\phi_1,\phi_2,\cdots\cdots,\phi_p$ 和 $Q_1,Q_2,\cdots\cdots,Q_q$ 分别为自回归和滑动平均参数,$a_t,a_{t-1},\cdots\cdots a_{t-q}$ 为相互独立的白噪声。

3.3.1 模式识别　　所谓模式识别就是对模型中 p,q 的阶数做出合理的判据。这儿采用对一阶预报误差(a_t) 的高斯白噪声检验方法加以确定[9],统计式为:

$$\begin{aligned} Q_m &= n\sum_{k=1}^{m}\rho_k^2(n,a) \\ \rho_k(n,a) &= r_k(n,a)/r_0(n,a) \\ r_k(n,a) &= \frac{1}{n}\sum_{t=1}^{n-k}a_t a_{t+k} \\ k &= 1,2,\cdots\cdots,m \end{aligned} \tag{21}$$

这里,$\rho_k(n,a)$ 为后延 k 时,a_t 的自相关函数,它是自由度为($m-p-g$) 的 χ^2 变量($m=\frac{n}{10}+p+q$). 如果 χ^2 变量的概率大于 Q_m 小于 α,p 和 q 的阶数便升高,直到 χ^2 变量的概率大于 α 为止。这里的 α 为 x^2 变量的检验水平,在程序中设置,本文 α 为 0.01 水平。

3.3.2 参数估计　　所谓参数估计就是对 ARMA(p,q) 模型中的自回归和滑动平均参数做出合理的判据。这里采用极大似然估计方法加以确定[8,9]。

自回归参数的估计:

$$\text{若 } p>0 \text{ ，则 } \quad A\phi_o = X \tag{22}$$

$$A_{ij} = C_{|q+i-j|} \qquad X_i = C_{q+i}$$

$$i,j = 1,2,\cdots\cdots,p$$

滑动平均参数的估计:

首先利用 w_t 的自协方差 C_k,并对它加以修正:

$$C_j' = \sum_{i=0}^{p}\sum_{k=0}^{p}\phi_{io}\phi_{ko}C_{|q+i-j|} \tag{23}$$

$$p>0, \quad \phi_{\infty} = -1 \qquad j = 0,1,2,\cdots\cdots,q$$

利用高斯牛顿算法:

$$\tau^{j+1} = \tau^j - h$$

$$T^j h = f^j$$

$$\tau = (\tau_0, \tau_1, \tau_2 \cdots\cdots, \tau_q)$$

$$f^j = \sum_{i=0}^{q-j} \tau_i \tau_{i+j} - C_j'$$

$$f^j = (f_0, f_1, f_2, \cdots\cdots, f_q)$$

$$T = \begin{pmatrix} \tau_0\ \tau_1\ \tau_2 \cdots\cdots \tau_q \\ \tau_1\ \tau_2\ \tau_3 \cdots \tau_q \\ \cdots\cdots \\ \tau_q \end{pmatrix} + \begin{pmatrix} \tau_0\ \tau_1\ \tau_2 \cdots\cdots \tau_q \\ \tau_0\ \tau_1 \cdots\cdots \tau_{q-1} \\ \cdots\cdots \\ \tau_0 \end{pmatrix}$$

初始值 $\tau_0 = \sqrt{C_0}, \tau_1 = \tau_2 = \cdots\cdots = \tau_q = 0$

当 $|f^j| < \varepsilon$ 时，$j = 0,1,2,\cdots\cdots,q$，这个过程收敛，参数的估计值可从最后一步 τ 的迭代值求出：

$$Q_j = -\tau_j/\tau_0 \qquad (24)$$

$$j = 1,2,\cdots\cdots,q$$

3.3.3 ARMA(p,q)模型的建立　　经上述各步的计算求其模型阶数 $p = 1, q = 1$，参数 $\phi_1 = 0.1361, Q_1 = 0.2058$，则式(20)为：

$$\widehat{W}_t = 0.1361 w_{t-1} + a_t - 0.2058 a_{t-1} \qquad (25)$$

拟合结果详见图3。

经上述几步分析与计算，北京地区冬小麦产量预报的统计模型为：

$$\begin{aligned} Y_t = {} & 170.965 + 8.184t + 0.084t^2 + 0.008t^3 + 0.005t^4 \\ & - 2.9696 + 0.0178X_1 + 0.1008X_2 + 0.0566X_3 + 0.622X_1' \\ & + 0.1361 w_{t-1} + a_t - 0.2058 a_{t-1} \end{aligned} \qquad (26)$$

4 预报模型的历史回代和独立样本预报检验

4.1 历史回代检验

目前，统计预报的拟合准确率一般在90%以上，因此，在历史资料的拟合中，我们假定相对误差小于5%为拟合正确。本例的计算结果为，历史拟合准确率达100%，仅1984年拟合误差偏大，为4.7%。方程的复相关系数为0.9865，剩余标准差为2.417。这说明了本文所组建的模式(26)具有较高的精度。

4.2 独立样本的预报检验

趋势产量的预报：在产量构成中，趋势产量占绝大部分，它的外推准确率是产量预报的关键。因趋势产量主要受近几年产量变化的制约和控制，因此，对它的外推应十分慎重。本文提出的处理方法为，从预报年份的前5年开始，随着年代的递增，分别建立正交多项式模式，并作出外推，取其近5年趋势产量的算术平均值，作为预报年份的趋势产量预报值。这样处理的好处在于，它考虑了近5年冬小麦产量的变化趋势，使之更加稳定。

气象产量的预报：利用已确定的8个环流因子独立样本值，代入式(7)、(8)、(9)，确定

其第 1,2,3 主分量组合因子;再利用 8 个均值周期分量因子独立样本值,代入式(16),确定其第 1 主分量因子。然后将 X_1,X_2,X_3,X_1' 代入式(19),便可得到气象产量的预报值。

剩余产量的预报:对于 ARMA 模型,设 $\hat{W}_{t(l)}$ 为时刻 l 及以前的观测序列 w_t, w_{t+1}……,对未来时刻 $t+l$ 的所作的近似平稳最小方差预报值,则有:

$$\hat{W}_{t+l}=\sum_{j=0}^{\infty}G_{l+j}a_{t-j}$$

预报误差 $e_{t(l)}$ 的方差为:

$$V_{ar}[e_t(l)]=\sigma_a^2[1+G_1^2+G_2^2+\cdots\cdots G_{l-1}^2] \tag{27}$$

式中,$G_1,G_2,\cdots\cdots,G_l,\cdots\cdots$,是格林函数。根据新的观测值,可由下式计算:

$$\hat{W}_{t+1(l)}=\hat{W}_{t(l+1)}+G_la_{t+1} \tag{28}$$

修正求得预测值[8,9]。式中,

$$a_{t+1}=\hat{W}_{t+1}-\hat{W}_{t(l)} \tag{29}$$

经上述分析和计算,得到北京 1989,1990 两年冬小麦产量预报分别为 321.1kg/ 亩和 345.7kg/ 亩,相对误差分别为 2.4% 和 3.7%,预报完全准确。

5　结语

1. 使用主分量分析方法,可以对选取的多个预报因子进行浓缩处理,提取所有预报因子的有效信息,提高因子的覆盖度,并消除了因子间的多元共线性,是一种有效的分析方法。

2. 附加预报量的均值周期因子具有较高的相关系数,它可以起到提高模型的历史拟合和预测精度,使方程更加稳定可靠。

3. 对剩余产量的模拟处理,还是首次尝试,它对提高产量的预报精度具有一定的补偿作用。

4. 本文在因子筛选、因子物理及生物学意义的确定、回归模型组建、趋势产量外推及剩余产量的模拟和预报等处理方面所采用的一些分析计算方法,均收到一定的效果。这说明了,这些处理和分析过程在某种程度上具有一定的普遍意义。

参考文献

[1] 王春乙,潘亚茹. 我国北方主要产麦区冬小麦产量海温业务预报模型. 数学的实践与认识,1989,(1):19—25.

[2] 王春乙,潘亚茹. 我国北方冬小麦总产年型的宏观回归预测模型. 科学通报,1989,34(21):1680.

[3] 王春乙,潘亚茹. 用带周期分量的逐步回归模型预测冬小麦产量的试验. 气象,1989,15(8):22—25.

[4] 王春乙,潘亚茹. 一种预测高温危害天气的统计模型. 应用气象学报,1990,1(2):217—221.

[5] 王春乙,潘亚茹. 试用多元分段回归进行作物产量年景预测的试验. 中国农业气象,1992,13(3):40—44.

[6] 王春乙,潘亚茹. 石家庄地区干热风年型指标分析及统计预报模型. 气象学报,1991,49(1):104—107.

[7] Hoerl A E,Kennard R W. Ridge regression applications to nonorthogonal problems. Technometrics, 1970, 12(4):456—463.

[8] 王春乙,潘亚茹. 时间序列的 ARMA 模型在干旱长期预测中的应用. 中国农业气象,1989,10(1):58—61.

[9] Davis J M . The use of time series analysis techniques in forecasting meteorological drought. Monthly

Weather Review, 1974,102:176—180.

THE STATISTICS MODEL OF OPERATIONAL PREDICTION FOR WINTER WHEAT PRODUCTION IN BEIJING

Wang Chunyi Pan Yaru Bai Yueming Wen Min

(*Chinese Academy of Meteorological Sciences Beijing*, 100081)

Abstract

Some techniques are put forward to overcome the disadvantages in various statistical models of meteorological predication of agricultural production, which are tested through the predication of winter wheat in Beijing and the results are proved to be better than usual.

Key words: Principal components analysis, Winter wheat, Ridge regression, ARMA model.