第34卷 第4期 气 象 学 报 Vol. 34, No. 4
1964 年 11 月 ACTA METEOROLOGICA SINICA November, 1964

統計学长期天气預报方法的若干研究(一)

逐步迴归技术的应用*

史久恩 瞿栋根 孙山泽 崔玉璽**
(中央气象局气象科学研究所) (北京大学数学力学系) (中国科学院計算技术研究所)

提 要

本文采用了逐步迴归技术，可以在一大堆預报因子中选择出迴归方程中的主要变量，組成一个最后的迴归方程——預报方程。

为了具体說明逐步迴归方法应用于长期預报的計算情况，将 1932—1962 年华北五站（北京、天津、保定、石家庄、营口）7，8 月份平均降水总量的資料作为因变量 y，以表征太阳輻射影响的因子、前期环流影响的因子和其它若干气象要素等作为自变量进行分析研究。

对我国若干重点地区的月、季降水的迴归分析指出，各地区与降水有关的預报因子是不完全相同的。就 1963 年夏季和 6—10 月份月的降水长期預报进行检查。各重点地区 42 次預报的結果說明，用这个方法所作的降水趋势預报尚好，比偶然性預报約高 10%。

最后，作者就迴归分析的改进作了簡略的討論，并提出今后除了用于长期預报的研究和业务工作外，对中、短期和专业天气預报以及气候学方面等的研究工作也将会是一个有用工具。

一、引 言

統計方法在气象上的应用范围是十分广泛的。从对大气环流的研究和对大气过程的描述，一直到人工控制的設計和試驗以及天气預报的制作等方面，都有不少的工作[1,2]。

从近代的天气預报发展来看，概率統計方法的应用是这方面发展的途径之一。通过数学工具，可以从大量定性的經驗关系中来确定若干最为显著的函数关系。这样一方面可以建立定量的关系，另方面又可对已找到的函数关系进行物理了解并得出物理解释。

迴归分析方法多年来已得到广泛的应用。我国涂长望先生[3]早在 1937 年就初步应用了这种分析方法，他从世界天气的观点出发，进行了中国天气与大气浪动及其在中国夏季旱涝长期預报方面的研究。

近几年来由于計算技术的迅速发展，利用高速电子計算机来作包含大量气象变量的綫性迴归方程已成为現实，并在此基础上发展了一种新的統計計算技术：即从一大堆要素中选择迴归方程中主要变量的方法[4,5]，进而作天气分析、預报[6,7]。

本文的主要目的是：1)在长期天气預报中引进这种方法；2)就应用的情况举例并进行討論。

* 本文 1964 年 3 月 10 日收到。
** 应显勋、孙爱芬同志曾参加部分工作。

二、逐步迴归分析

綫性迴归分析的数学模型为:

$$y = x_n = b_0 + b_1x_1 + \cdots + b_{n-1}x_{n-1}. \tag{1}$$

其中 $y(=x_n)$ 是因变量亦即预报量, $x_1, \cdots, x_{n-1}$ 是自变量亦即预报因子。 $b_0, b_1, \cdots, b_{n-1}$ 为迴归系数,給定了 $(x_1, \cdots, x_{n-1}, x_n)$ 的 m 組观测值 $(x_{1t}, \cdots, x_{nt})$, $t = 1, 2, \cdots, m$, 可用最小二乘方法求得这些迴归系数.

逐步迴归分析的原理与上述的一般迴归分析是相类似的, 所不同的是对一般迴归而言,即一次建立包括所有自变量的迴归方程. 但在(1)式中的每一个自变量不能都对因变量 y 的方差有显著的貢献, 而逐步迴归分析可以得到只包含那些对 y 有显著貢献的自变量.

逐步迴归是从建立只有一个自变量的迴归方程开始, 然后逐步加入其它的自变量求得新的方程. 因此在逐步迴归分析的計算中,将得到一系列的迴归方程: 即

$$\left.\begin{aligned} y = x_n &= b_0^{(1)} + b_1^{(1)}x_1, \\ y = x_n &= b_0^{(2)} + b_1^{(2)}x_1 + b_2^{(2)}x_2, \\ y = x_n &= b_0^{(3)} + b_1^{(3)}x_1 + b_2^{(3)}x_2 + b_3^{(3)}x_3, \\ &\vdots \\ y = x_n &= b_0^{(l)} + b_1^{(l)}x_1 + \cdots + b_l^{(l)}x_l, \\ &\vdots \end{aligned}\right\} \tag{2}$$

在(2)式,迴归系数的上标 "(l)" 表示第 l 步所得到的迴归方程. 在各步迴归方程中有相同下标的迴归系数,其数值是会变的甚至可以相差得比較多.

由于每一步加入一个自变量, 故有 k 个可能的自变量时, 可以組成包含有 1 个自变量, 2 个自变量,……, k 个自变量的許多迴归方程,其組成总数为

$$\sum_{r=1}^{k} c_k^r = \sum_{r=1}^{k} \frac{k!}{r!(k-r)!} = 2^k - 1 \text{个}.$$

今有 $k = n - 1$, 因此,即使自变量的个数 $n - 1$ 为一个不太大的数目, 其可能組成的方程总数是很大的;而逐步迴归技术,即可在这些迴归方程中选出一个方程, 其中各个自变量对因变量的方差貢献是最大的. 現将这种方法的具体步驟簡述如下:

引进一个 $n \times n$ 矩陣 A, $A = (a_{ij})$, (3)

其中

$$a_{ij} = \frac{\sum_{t=1}^{m} (x_{it} - \bar{x}_i)(x_{jt} - \bar{x}_j)}{\sqrt{\sum_{t=1}^{m} (x_{it} - \bar{x}_i)^2 \sum_{t=1}^{m} (x_{jt} - \bar{x}_j)^2}} = \frac{s_{ij}}{\sigma_i \sigma_j}, \quad i, j = 1, 2, \cdots, n. \tag{4}$$

而

$$s_{ij} = \sum_{t=1}^{m} (x_{it} - \bar{x}_i)(x_{jt} - \bar{x}_j), \tag{5}$$

$$\sigma_i = \sqrt{\sum_{t=1}^{m} (x_{it} - \bar{x}_i)^2}, \tag{6}$$

$$\sigma_j = \sqrt{\sum_{t=1}^{m} (x_{jt} - \bar{x}_j)^2}. \tag{7}$$

这里的 A 矩陣，实际上就是一个相关矩陣；而矩陣中的元素 A_{ij} 即一般常用的相关系数.

进行逐步迴归分析即先选择一个自变量 x_1，使其对因变量 x_n 貢献的方差为最大，亦即使：

$$Q_1^{(2)} \geqslant Q_i^{(2)}. \tag{8}$$

对一切 $i = 1, 2, \cdots, n-1$ 成立

其中

$$Q_1^{(2)} = s_{nn} - \min_{(b_0, b_1)} \sum_{t=1}^{m} [x_{nt} - (b_0 + b_1 x_{1t})]^2, \tag{9}$$

$$Q_i^{(2)} = s_{nn} - \min_{(b_0, b_i)} \sum_{t=1}^{m} [x_{nt} - (b_0 + b_i x_{it})]^2. \tag{10}$$

于是建立迴归方程：

$$y_t = x_{nt} = b_0 + b_1 x_{1t}. \tag{11}$$

为了求迴归系数 b_0, b_1 使 $\sum_{t=1}^{m} [x_{nt} - (b_0 + b_1 x_{1t})]^2$ 达最小值，可分别对 b_0, b_1 求偏微商，即得正規方程为：

$$\left.\begin{aligned} \sum_{t=1}^{m} [x_{nt} - (b_0 + b_1 x_{1t})] &= 0, \\ \sum_{t=1}^{m} [x_{nt} - (b_0 + b_1 x_{1t})] x_{1t} &= 0. \end{aligned}\right\} \tag{12}$$

解得：

$$b_1 = \frac{s_{1n}}{s_{11}} = \frac{a_{1n}}{a_{11}} \cdot \frac{\sigma_n}{\sigma_1},$$

$$b_0 = \bar{x}_n - b_1 \bar{x}_1.$$

其中 $\bar{x}_i$ 为第 i 个变量对时間的平均值[1].

接着对原假設 $b_i = 0$ 进行检驗，引入統計量：

$$T = \frac{\dfrac{Q^{(l)} - Q^{(l+1)}}{Q^{(l+1)}}}{m - l - 1} = \frac{Q_i^{(l+1)}(m - l - 1)}{Q^{(l+1)}}. \tag{13}$$

式中 m 即为实測值的組数，l 为第 i 个变量，x_i 尚未加入迴归方程的系数数目，$l+1$ 为 x_i 加入迴归方程后的系数数目. 其中：

$$Q^{(l+1)} = \sum_{t=1}^{m} [x_{nt} - (b_0 + b_1 x_{1t} + \cdots + b_{l+1} x_{l+1,t})]^2 \text{ 为最小}, \tag{14}$$

$$Q^{(l)} = \sum_{t=1}^{m} [x_{nt} - (b'_0 + b'_1 x_{1t} + \cdots + b'_{i-1} x_{i-1,t} + b'_{i+1} x_{i+1,t} + b'_{l+1} x_{l+1,t})]^2 \text{ 为最小}. \tag{15}$$

令

$$Q_i^{(l+1)} = Q^{(l)} - Q^{(l+1)}, \tag{16}$$

1) 这里 $i = 1$.

其中 $$l+1 \leqslant n-1.$$

假定因变量 x_{nt} 是一組来自某一正态总体的独立样本[1)]，則統計量 T 遵从自由度为(1，$m-l-1$)的 F-分布. 因此給定显著水平 α 后，可确定一个临界值 $F_{\alpha}^{(1,m-l-1)}$ 并由实測值計算出統計量 T 的实現值 F^*. 当 $F^* > F_{\alpha}^{(1,m-l-1)}$ 时，拒絕原假設 $b_i = 0$. 今 b_0, b_1 已得，于是代入公式(14)、(16)得:

$$Q^{(2)} = \min_{(b_0,b_1)} \sum_{t=1}^{m} [x_{nt} - (b_0 + b_1 x_{1t})]^2 = s_{nn} - \frac{s_{1n}s_{n1}}{s_{11}} = \sigma_n^2 \left(a_{nn} - \frac{a_{1n}a_{n1}}{a_{11}} \right), \tag{17}$$

$$Q_1^{(2)} = s_{rn} - Q^{(2)} = \frac{s_{1n}s_{n1}}{s_{11}} = \sigma_n^2 \frac{a_{1n}a_{n1}}{a_{11}}. \tag{18}$$

检驗假設 $b_1 = 0$，即計算

$$F^* = \frac{Q_1^{(2)}(m-2)}{Q^{(2)}} = \frac{Q_1^{(2)}(m-2)}{s_{nn} - Q_1^{(2)}} = \frac{V_1^{(2)}(m-2)}{a_{nn} - V_1^{(2)}}. \tag{19}$$

其中

$$V_1^{(2)} = \frac{a_{1n}a_{n1}}{a_{11}}, \qquad V_1^{(2)}(m-2) = \frac{Q_1^{(2)}(m-2)}{\sigma_n^2}.$$

当 $\dfrac{V_1^{(2)}(m-2)}{a_{nn} - V_1^{(2)}} > F_{\alpha}^{(1,m-2)}$ 时，即拒絕原假設 $b_1 = 0$. $F_{\alpha}^{(1,m-2)}$ 为在显著水平 α 下，自由度为 $(1, m-2)$ 的 F-分布的临界值.

实际进行計算时并不需要对 b_0, b_1 求偏微商及解正規方程等过程，只要用矩陣 A 的元素就可求出全部所需要的量，并在估計迴归系数之前就先进行假設检驗.

当逐步进行自变量的选择时，可能遇到以下两种情况:

1. 如某一自变量 x_i 加到迴归方程中，对因变量 y 的方差貢献已达到某一个 F_1^* 水准的显著性，則可在原来的迴归方程中加入这一自变量;

2. 如在迴归方程中某一自变量 x_j 对因变量 y 的方差貢献未达某一个 F_2^* 水准的显著性，則从迴归方程中剔除这一自变量.

表 1 、表 2 分别列出选中的自变量 x_i 并組成新的迴归方程或从原来的迴归方程中剔除自变量 x_j 所需計算的統計量.

表 1 自变量 x_i 选入迴归方程的統計量

变差来源	平方和	自由度	均方	統計量的实現值[注]
将要被选入的方差貢献最显著的变量(x_i)	$V_i^{(l+1)}$	1	$V_i^{(l+1)}$	$F_1^* = \dfrac{V_i^{(l+1)}(m-l-1)}{V^{(l+1)}}$
加入变量后的剩余方差	$V^{(l)} - V_i^{(l+1)} = V^{(l+1)}$	$m-l-1$	$\dfrac{V^{(l)} - V_i^{(l+1)}}{m-l-1}$	

[注] F_1^* 要与 F_1 比較，F_1 为在显著水平 α 下加入自变量的 F-分布的临界值，其自由度为 $(1, m-l-1)$.

1) 在实际应用中，对"因变量 x_{nt} 是来自某一正态总体的样本"这一假定最好加以驗証. 但在天气預报問題中尤其是长期預报，因历史資料年代較短往往不能滿足統計学的"随机抽样"和样本大小的要求，因此这一假定往往也不作驗証.

对气象問題来說，由于各个自变量并不是相互独立的，所以在逐步回归方程中当加入一个或几个新变量时，可能使原来已选入回归方程中的自变量对 y 的方差貢献变得不显著了，这时原来的自变量将被剔除. 因此在逐步回归的每一步中，只有那些"显著"的自变量才能包含到新的回归方程中.

表 2　自变量 x_j 自回归方程剔除的統計量

变差来源	平方和	自由度	均方	統計量的实現值[注]
对应于回归方程中方差貢献最小的变量(x_j)	$V_j^{(l)}$	1	$V_j^{(l)}$	$F_2^* = \dfrac{V_j^{(l)}(m-l)}{V^{(l)}}$
在回归方程中其它变量的方差貢献	$1 - V_j^{(l)} - V^{(l)}$	$m-l-1$	$\dfrac{1 - V_j^{(l)} - V^{(l)}}{m-l-1}$	
剩余方差	$V^{(l)}$	$m-l$	$\dfrac{V^{(l)}}{m-l}$	

[注] F_2^* 要与 F_2 比較，F_2 为在指定的显著水平 α 下，自由度为$(1, m-l)$的 F-分布的临界值.

在逐步求取回归方程时，各步回归系数的計算公式如下：

$$b_k = a_{kn}^{(h)} \frac{\sigma_n}{\sigma_k}. \tag{20}$$

上式 $a_{kn}^{(h)}$ 为对应第 h 步回归方程中所选择的自变量 x_k，亦即从原来的相关矩陣 $A^{(1)}$ 經过 $h-1$ 步的运算后所得到新矩陣 $A^{(h)}$ 中的一个元素. $a_{kn}^{(h)}$ 与前一步矩陣 $A^{(h-1)}$ 中元素的关系为：

$$a_{kn}^{(h)} = \frac{a_{kn}^{(h-1)}}{a_{kk}^{(h-1)}}.$$

至于 $A^{(h)}$ 中其它元素与 $A^{(h-1)}$ 中元素的关系为：

$$a_{ij}^{(h)} = \begin{cases} \dfrac{a_{ij}^{(h-1)} a_{kk}^{(h-1)} - a_{ik}^{(h-1)} a_{kj}^{(h-1)}}{a_{kk}^{(h-1)}}, & i \neq k, \quad j \neq k. \\ \dfrac{a_{kj}^{(h-1)}}{a_{kk}^{(h-1)}}, & i = k, \quad j \neq k. \\ -\dfrac{a_{ik}^{(h-1)}}{a_{kk}^{(h-1)}}, & i \neq k, \quad j = k. \\ \dfrac{1}{a_{kk}^{(h-1)}}, & i = k, \quad j = k. \end{cases} \tag{21}$$

而每一步回归方程的常数 b_0 是按照下式計算：

$$b_0 = \bar{y} - \Sigma b_i \bar{x}_i. \tag{22}$$

三、降水长期預报中的应用

1. 华北地区降水长期趋势預报的一个实例

1) 資料来源　根据 1932—1962 年华北地区五站(北京、天津、保定、石家庄、营口)

的 7、8 月份降水总量作为因变量y[1]。而选降水資料前期的太阳黑子沃尔夫指数，国际地磁 C_i 指数，苏联房根盖姆的大型环流型距平天数，上海和西安、济南的气压差，以及我国几个主要地区降水量等作为自变量——預报因子。我們挑选这些因子事前是經过一番物理原因的考虑，这里不能一一縷述。表 3 列出了这些可能被选中的預报因子。

表 3 用作逐步迴归中篩选的預报因子

自变量序号	可能的預报因子名称	与預报量 y 相距的时間
x_1	太阳黑子的沃尔夫指数的年平均值	前 4 年
x_2	太阳黑子的沃尔夫指数的年平均值	前 5 年
x_3	太阳黑子的沃尔夫指数的年平均值	前 6 年
x_4	太阳黑子的沃尔夫指数的年平均值	前 7 年
x_5	太阳黑子的沃尔夫指数的年平均值	前 8 年
x_6	太阳黑子的沃尔夫指数的年平均值	前 9 年
x_7	太阳黑子的沃尔夫指数的年平均值	前10年
x_8	太阳黑子的沃尔夫指数的季平均值之差数	前 1 年夏季与春季之差
x_9	国际地磁活动的 C_i 指数的年平均值	前 2 年
x_{10}	国际地磁活动的 C_i 指数的年平均值	前 3 年
x_{11}	国际地磁活动的 C_i 指数的年平均值	前 5 年
x_{12}	国际地磁活动的 C_i 指数的年平均值	前 7 年
x_{13}	国际地磁活动的 C_i 指数的年平均值	前 9 年
x_{14}	房根盖姆的W型天数距平值	前 1 年 3 月＋4 月
x_{15}	房根盖姆的W型天数距平值	前 1 年 5 月＋6 月
x_{16}	房根盖姆的W型天数距平值	前 1 年 7 月＋8 月
x_{17}	房根盖姆的E型天数距平值	前 1 年 3 月＋4 月
x_{18}	房根盖姆的E型天数距平值	前 1 年 5 月＋6 月
x_{19}	房根盖姆的E型天数距平值	前 1 年 7 月＋8 月
x_{20}	上海与西安的气压差	前 1 年 4 月
x_{21}	上海与济南的气压差	前 1 年 7 月
x_{22}	东北地区三站(沈阳、长春、哈尔滨) 5—8 月降水	前 1 年
x_{23}	华北地区五站(北京、天津、保定、石家庄、营口) 7＋8 月降水	前 1 年
x_{24}	长江上游三站(成都、宜宾、重庆) 5—8 月降水	前 1 年
x_{25}	长江中游三站(宜昌、汉口、重庆) 5—8 月降水	前 1 年
x_{26}	长江中下游五站(上海、南京、蕪湖、九江、汉口) 3 月降水	前 1 年
x_{27}	长江中下游五站(上海、南京、蕪湖、九江、汉口) 4 月降水	前 1 年
x_{28}	长江中下游五站(上海、南京、蕪湖、九江、汉口) 5 月降水	前 1 年
x_{29}	长江中下游五站(上海、南京、蕪湖、九江、汉口) 6 月降水	前 1 年
x_{30}	华南三站(广州、汕头、廈門) 3—5 月降水	前 1 年

2) 逐步选择的情况　对上述自变量就因变量进行步选，根据不同的显著水平 α，可以有不同的 F 值[2]。本文仅举 $F_1 = F_2 = 2.5$ 及 $F_1 = F_2 = 1.5$ 的情况为例。

开始时，第一步引入的自变量是 x_{27}，經計算后得 $F_1^* > F_1$，$F_2^* > F_2$；于是 x_{27} 进入迴归方程，組成第一个中間方程。即：

$$y_i^{(1)} = b_0^{(1)} + b_{27}^{(1)} x_{27i}.$$

1) 根据 χ^2 检驗說明，华北地区五站 7、8 月份降水总量实际頻数与正态分布理論頻数間的差异是极不显著的，故可認为該 y_i 資料是屬于正态分布的。

2) 在同一自由度下，F 值决定于显著水平 α。

表 4　华北五站 7，8 月降水的步选結果 (a) $F_1 = F_2 = 2.5$，(b) $F_1 = F_2 = 1.5$.

表 4a

	引入变量	剔除变量	F_1^*	F_2^*	y 的标准誤差	中　間　过　程	多重相关系数 R	Sb_{27}	Sb_{12}	Sb_9	Sb_{22}	Sb_{30}
第一步	x_{27}		5.17	3×10^5	106.63	$y_t^{(1)} = 254.65 + 0.929x_{27}$	0.33	0.416				
第二步	x_{12}		3.43	4.99	102.60	$y_t^{(2)} = 2.45 + 392.29x_{12} + 0.823x_{27}$	0.43	0.405	215.18			
第三步	x_9		2.62	3.32	99.91	$y_t^{(3)} = -281.09 + 317.57x_9 + 494.66x_{12} + 0.754x_{27}$	0.47	0.396	219.22	199.76		
第四步	x_{22}		3.47	2.53	95.83	$y_t^{(4)} = -144.15 + 373.43x_9 + 507.69x_{12} - 0.355x_{22} + 0.584x_{27}$	0.54	0.391	210.39	194.02	0.194	
第五步		x_{27}	5.33	2.22	97.98	$y_t^{(5)} = -113.69 + 414.63x_9 + 561.65x_{12} - 0.424x_{22}$	0.51		211.90	196.34	0.193	
第六步	x_{30}		6.00	4.46	90.30	$y_t^{(6)} = 25.54 + 347.88x_9 + 580.61x_{12} - 0.581x_{22} - 0.044x_{30}$	0.61		195.44	183.06	0.189	0.018

y 的标准誤差(原始的)=113.49

表 4b

	引入变量	剔除变量	F_1^*	F_2^*	y 的标准誤差	中　間　过　程	多重相关系数 R	Sb_{27}	Sb_{12}	Sb_9	Sb_{22}	Sb_{30}	Sb_6	Sb_{11}	Sb_4	Sb_{13}	Sb_8	Sb_{18}
第一步	x_{27}		5.17	3×10^5	106.63	$y_t^{(1)} = 254.65 + 0.929x_{27}$	0.33	0.416										
第二步	x_{12}		3.43	4.99	102.60	$y_t^{(2)} = 2.45 + 392.29x_{12} + 0.823x_{27}$	0.43	0.406	215.18									
第三步	x_9		2.62	3.32	99.91	$y_t^{(3)} = -281.09 + 317.57x_9 + 494.66x_{12} + 0.754x_{27}$	0.47	0.396	219.22	199.76								
第四步	x_{22}		3.47	2.53	95.83	$y_t^{(4)} = -144.15 + 373.43x_9 + 507.69x_{12} - 0.355x_{22} + 0.584x_{27}$	0.54	0.391	210.39	194.02	0.194							
第五步	x_{30}		5.33	2.22	89.03	$y_t^{(5)} = -8.85 + 318.11x_9 + 534.65x_{12} - 0.514x_{22} + 0.484x_{27} - 0.041x_{30}$	0.62	0.366	195.83	181.91	0.193	0.018						
第六步	x_6		1.78	1.74	87.80	$y_t^{(6)} = -43.17 + 0.836x_6 + 501.79x_9 + 327.11x_{12} - 0.523x_{22} + 0.456x_{27} - 0.034x_{30}$	0.63	0.362	250.04	227.90	0.191	0.019	0.639					
第七步	x_{11}		3.26	1.58	84.16	$y_t^{(7)} = 209.22 + 1.247x_6 + 546.44x_9 - 336.29x_{11} + 200.16x_{12} - 0.514x_{22} + 0.441x_{27} - 0.027x_{30}$	0.67	0.347	250.19	219.90	0.183	0.018	0.656	190.26				
第八步		x_{12}	3.31	0.64	83.52	$y_t^{(8)} = 316.51 + 1.598x_6 + 588.94x_9 - 379.99x_{11} - 0.510x_{22} + 0.465x_{27} - 0.023x_{30}$	0.68	0.343		211.77	0.181	0.018	0.483	180.88				
第九步	x_4		3.89	1.75	79.14	$y_t^{(9)} = 418.83 + 0.856x_4 + 1.315x_6 + 557.69x_9 - 553.33x_{11} - 0.505x_{22} + 0.538x_{27} - 0.031x_{30}$	0.72	0.327		201.33	0.172	0.017	0.480	193.48	0.443			
第十步	x_{18}		4.13	2.70	74.50	$y_t^{(10)} = 125.97 + 1.71x_4 + 0.603x_6 + 518.21x_9 - 605.49x_{11} + 452.55x_{18} - 0.440x_{22} + 0.718x_{27} - 0.037x_{30}$	0.75	0.321		190.56	0.165	0.016	0.577	184.02	0.600	227.59		
第十一步		x_6	3.21	1.10	74.66	$y_t^{(11)} = 101.47 + 2.10x_4 + 408.86x_9 - 614.96x_{11} + 600.34x_{18} - 0.411x_{22} + 0.825x_{27} - 0.043x_{30}$	0.75	0.305		159.70	0.163	0.016		184.18	0.475	178.87		
第十二步	x_8		4.58	6.32	69.71	$y_t^{(12)} = 47.17 - 1.319x_8 + 2.801x_4 + 621.61x_9 - 415.13x_{11} + 451.01x_{18} - 0.547x_{22} + 0.587x_{27} - 0.046x_{30}$	0.79	0.306		180.43	0.166	0.015		196.67	0.557	181.60	0.630	
第十三步	x_{16}		2.08	3.67	68.19	$y_t^{(13)} = 32.87 - 1.366x_8 + 2.809x_4 + 617.76x_9 - 357.60x_{11} + 465.29x_{18} + 2.030x_{16} - 0.635x_{22} + 0.436x_{27} - 0.044x_{30}$	0.80	0.318		177.49	0.174	0.014		196.67	0.544	177.94	0.617	1.439

y 的标准誤差(原始的)=113.49

其中 $b_0^{(1)} = 254.65$（毫米），$b_{27}^{(1)} = 0.929$. 这时算出 y 的标准誤差为 106.63，系数 $b_{27}^{(1)}$ 的标准誤差 $Sb_{27}^{(1)} = 0.416$，自变量 x_{27} 与 y 的相关系数 $R^{(1)} = 0.33$.

于是再繼續进行第二步、第三步、……，具体的中間过程列于表（4a, b）. 其中值得注意的有下列几点：

（1）从（表 4）中可以观察到在逐步迴归中剔除变量的特点. 例如（表 4a）中第一步选入的自变量为 x_{27}，当加入自变量 x_{12}，x_9 及 x_{22} 以后，它的作用已經显得不重要了；于是，在新增自变量 x_{30} 之前被剔除. 这一点又可从（表 4a）的第四、五步进行考察，可見多重相关系数的数值变化不大. 由第四步中間方程（包含有 4 个自变量）的多重相关系数为 0.54，到第五步的中間方程（只有 3 个自变量）的多重相关系数仅仅降低了 0.03 而为 0.51. 这就說明了剔除一个自变量 x_{27} 对因变量 y 的拟合程度的影响很小.

（2）从（表 4）的步选过程中可以看到，由于每进行一步計算时有自变量的选入或剔除，因而使每一步中間方程中的迴归系数都不断变化. 就一般情况而言，因变量的标准誤差 S_y 和各自变量所对应的迴归系数的标准誤差 S_{b_i} 是随着逐步进行选择的計算而減少的，但多重相关系数則随之而逐步增大.

（3）在作检驗时，当所取的 F 水平不同，則組成的預报方程——最后一个迴归方程就可能不同. 往往在預报方程中所包含的自变量数目可以有多有少，而且被剔除的自变量序数也不尽相同. 例如从（表 4）可見，当 $F_1 = F_2 = 2.5$ 时，最后只有 4 个自变量（即 x_9、x_{12}、x_{22} 及 x_{30}）包含在預报方程中. 而当降低 F 水平时，令 $F_1 = F_2 = 1.5$ 时，則預报方程中包含了 9 个自变量（即 x_3、x_4、x_9、x_{11}、x_{18}、x_{22}、x_{27} 及 x_{30}）. 同时由于給定的 F 水平不同，因而使預报方程中計算出的多重相关系数 R 也有所差别. 当 $F = 2.5$ 时，$R = 0.61$；当 $F = 1.5$ 时，$R = 0.80$.

3）預报方程及計算結果　經过逐步处理以后，华北地区五站（北京、天津、保定、石家庄、宫口）的 7、8 月份平均降水总量資料在指定的 F 水平下（$F = 1.5$ 及 $F = 2.5$）得到了两个預报方程，它們分别为：

（1）在 $F_1 = F_2 = 2.5$ 时

$$y_t = 25.54 + 347.88x_{9t} + 580.61x_{12t} - 0.581x_{22t} - 0.044x_{30t},$$

（2）在 $F_1 = F_2 = 1.5$ 时

$$y_t = 32.87 - 1.366x_{3t} + 2.809x_{4t} + 617.76x_{9t} - 357.60x_{11t} + \\ + 465.29x_{13t} + 2.030x_{18t} - 0.635x_{22t} + 0.436x_{27t} - 0.044x_{30t}.$$

从上列方程可見，当 F 水平降低时，迴归方程中的自变量迅速地由 4 个增加为 9 个. 这說明只要在作检驗时降低信度要求，就可以将那些关系比較次要的因子也包括在預报方程中. 但是当检驗的要求过低时，往往使迴归方程对原資料的拟合情况很好，而在預报时这些次要因子与預报量之間的关系就不稳定了.

根据以上两个方程，将方程中所对应的預报因子代入并計算出 y 值，其計算結果如（图 1）所示. 图中所取的横坐标是計算数值，縱坐标是实測值. 从图上可以大致看出，“●”的点子要比“○”的点子更接近于計算和实况相一致的直綫. 这一点是很明显的，因为当 $F = 1.5$ 时从迴归方程所求得的多重相关系数要比 $F = 2.5$ 时从另一个迴归方程所得到的多重相关系数的数值大. 这就說明了为什么“●”点要比“○”点拟合得好一些. 但

从預报的角度来进行具体考察，就 1963 年的預測情况，在(图 1)中可以看出"⊙"点不如"◎"点[1]. 这似乎又提供了一个事实，即預报因子的选取还是需要有一定的 F 水平要求. 如只是单純地依賴增加自变量而組成包含大量自变量的多重迴归方程，在表面上虽可使得从迴归方程所計算出的数值与实际数值的逼近較好，但在对长期天气的物理关系和物理过程缺乏了解的情况下，貿然地将迴归方程作为預报方程来使用，必然会导致錯誤的結論并作出不正确的預报.

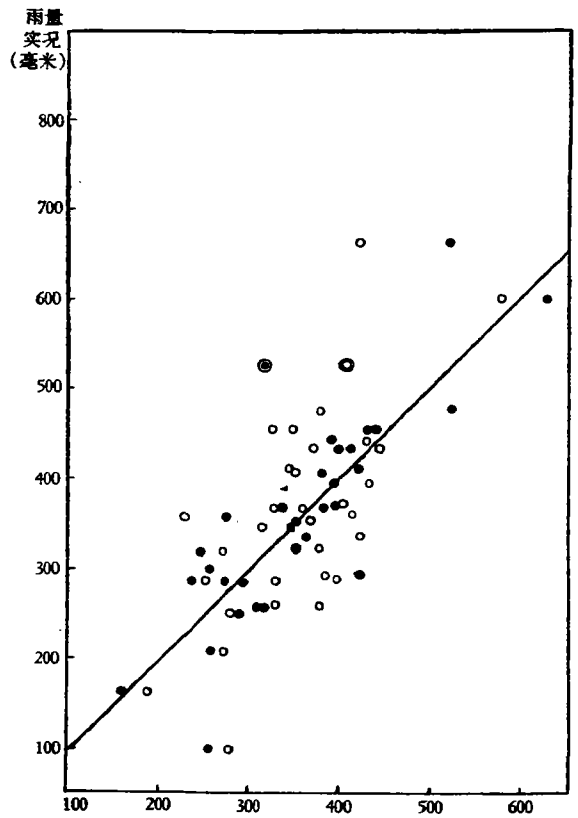

图 1　华北五站 7、8 月平均降水总量的实测值与計算值的散布图

("○"是根据 $F = 2.5$ 的方程所算出的数值与实况比較，"●"是 $F = 1.5$ 的方程所計算出的数值与实况比較，"◎"为前者之預报值与实况比較，"⊙"为后者之預报值与实况比較)

2. 試报的情况

应用上述的逐步迴归方法求得預报方程以后，試作了 1963 年夏季和 6—10 月份月的降水趋势預报检查情况. 根据各重点地区 42 次預报[2]的結果来看(表 5)，如按历史資料中降水量出現的頻次划为 5 级[3]，并允許以預报与实况可相差 1 级为趋势符合的話，

表 5　1963 年夏季和 6—10 月份月的降水趋势預报检查情况 (a) $F_1 = F_2 = 2.5$ (b) $F_1 = F_2 = 1.5$

表 5a

实况等級 \ 預报等級	I	II	III	IV	V
I	0	3	1	1	0
II	0	5	2	0	1
III	1	3	8	2	3
IV	1	2	1	3	2
V	1	1	1	0	0

表 5b

实况等級 \ 預报等級	I	II	III	IV	V
I	0	2	1	1	1
II	1	3	4	0	0
III	1	4	6	5	1
IV	1	4	1	1	2
V	1	0	1	0	1

1) 这里只討論了一个例子，从我們分析的許多例子中看出，"⊙"点大都不如"◎"点.

2) 实际上作一次預报需要有若干次計算和复合，具体过程从略.

3) 按出現頻次划分等級的規定是：設資料年份为 n，I 級为降水量最大的占 $\frac{n}{8}$ 年；II 級为降水量次多的占 $\frac{n}{4}$ 年；III 級为降水量接近正常的占 $\frac{n}{4}$ 年；IV 級为降水量偏少的占 $\frac{n}{4}$ 年；V 級为降水量最小的占 $\frac{n}{8}$ 年.

則約有 70% 的趋势可称符合，比偶然性預报約高 10%（表 6）.

表 6 偶然性預报与逐步迴归方法所作預报之比較

得分 / 誤差范圍 / 預报方法	差一級
偶然性預报	59.4%
逐步迴归 ($F=2.5$)	69.0%
逐步迴归 ($F=1.5$)	71.4%

从(表 5)可見 $F=2.5$ 时，預报与实况差一級的仅比 $F=1.5$ 的少一次. 但仔細分析（表 5）可見，对于預报和实况的等級一致的情况，則由 $F=1.5$ 的 11 次迅速提高到 $F=2.5$ 的 16 次，卽提高了 11.9%. 这說明了 F 水平适当地取得严一些，对于預报还是很重要的.

四、討 論

从逐步迴归分析的方法来看，它能較客观地且有一定理論依据地进行預报因子的选择，可在指定的 F 水平下从一批自变量中把与因变量 y 关系最主要的一些預报因子挑选出来并組成預报方程. 对于不同地区的降水分析表明，包含在預报方程中的預报因子，就降水的关系而言是不完全相同的. 根据已做的一些計算和試报情况来看，这个方法不論在天气預报、专业气象預报和气候学研究中，都将会是一个有用的工具和方法. 通过一段时間的实践和探索，我們感到还有一些方面值得在今后工作中継續深入和改进的.

1. 非綫性的問題

前面討論的是綫性的迴归方程，如果所研究的問題是非綫性的，将如何办？关于这个問題的解决并不是十分困难的. 只要我們能够从天气預报的实践經驗或根据一定的物理考虑，知道这个函数关系式是 $y=b_if_i(z_i)$，便可将一些非綫性的項加入迴归方程. 这样，使多重迴归分析也可以拟合非綫性方程. 卽設：

$$y=b_0+b_1z_1+b_2z_1^2+b_3z_1z_2+\cdots+b_{k-1}f_{k-1}(z_1,z_2,\cdots,z_{k-1}).$$

其中 z_i 为一些自变量($i=1,2,\cdots,k-1$)，作变換

令
$$x_1=z_1,$$
$$\vdots$$
$$x_{n-1}=f_{k-1}(z_1,z_2,\cdots,z_{k-1}).$$

于是这个方程便与(1)式相同了，从而也可用逐步迴归方法进行分析，寻找与因变量 y 关系最大的那些預报因子.

2. 抽 样 問 題

迴归分析所用的資料并不要求是一个不間断的时間序列，而对这些資料的得来最好

是一个随机抽样．在气象問題中，常常不能滿足这样的要求；尤其是长期預报的研究，只能利用在給定时期內有限的观测資料作为样本，通过分析从而推論母体的有关参数．由于这个問題的客观存在，有时会在最小二乘方法的概念下对样本实測值求得了平方最优的結果．但如根据所得統計关系对新的、另一个样本作預报时，这些关系就不一定能够稳定．因此，在使用分析的結論时需要特别慎重．

3. 选择适当的統計判据問題

对一給定的样本而言，似乎采用F水平較低和自变量数目較多的迴归方程可以拟合得好一些．但当具体制作預报时，这就产生了一个問題：对一个新的样本，原来已求得的統計关系是否仍能适合，其关系是否稳定．从統計学的假設检驗理論和具体計算，試报均說明了这一点．即F水平不能太低，否則所得的迴归方程的稳定性就差；但在另一方面F水平也不能取得过高，否則預报方程中一个自变量也沒有，这就无法进行預报．

4. 自变量的选择

选择合适的自变量是一个十分重要的問題．从純数学理論上講，可以訊为能够无限地处理自变量，然后从其中寻找出最好的預报因子．但在实际問題中，由于工作量的大小等等原因，不可能实現理論上的假想．因此如何最經济而又最有效地将关系最好、最密切的預报因子选入預报方程是值得深入探索的．同时为了进一步寻找和分析一些統計关系的物理原因，在統計方法中考虑一些物理量，看来还是很需要的．

5. 气象資料的正态分布問題

关于所用气象資料是否呈正态分布是一个比較重要的問題．以降水量为例，年雨量呈正态分布已由徐尔灝先生[8]作过論証．对月(或季)降水量的分布問題，我們曾选择了四个重点地区进行了一些分析、检驗．我們发現了在夏半年內的月(季)降水都属于正态分布，而冬半年內的降水則大半不属于正态分布；同时发現月(季)降水量是否呈正态分布与資料年份的长短有关．一般記录年代越长，其分布就越趋于正态．这一点与徐尔灝先生[8]所得的結論比較一致．此外，在对降水量資料进行检驗时注意到区域的降水量比单站的正态分布要标准一些，季的降水分布比月的降水分布的正态性要好些．由于目前在这方面所进行的分析、驗証还很初步，对于不同地区究竟需要年代多长的資料，以及对全国各个地区那些月份的气象資料是属于正态分布等等問題，暫时还不能进行全面的討論，但值得在今后工作中进行研究．

致謝：本文得到楊鑒初先生的鼓励和指导，张尧廷、潘純修等同志的热情帮助，刘鈡玲同志的繪图，在此一并致以深切的感謝．

参　考　文　献

[1] Malone, T. F., Compendium of Meteorology, 1951.
[2] White, R. M., Statistical methods, Transaction of American Geophysics Union, 1960.
[3] 涂长望，中国近代科学論著丛刊——气象学 (1919—1949)，科学出版社，1955，369—422.
[4] Williams, E. J., Regression Analysis, 1959.

[5] Ralston, A., Mathematical Methods for Digital Computers, 1960.
[6] Arakawa, H., Prediction of Movements and Surface Pressures of Typhoon Centers in the Far East by Statistical Methods, Technical Report of the Japan Meteorological Agency, No. 14.
[7] Martin, F. L. et al., Statistical Prediction Methods for North American Anticyclones, *Journal of Applied Meteorology*. 1963.
[8] 徐尔灏，論年雨量之常态性．气象学报 **21**(1950) 1—4 期．

STUDIES ON LONG-RANGE FORECASTING BY STATISTICAL METHODS: PART I——APPLICATION OF STEPWISE MULTIPLE REGRESSION TECHNIQUE

SZE KIU-UNG CHU TUNG-KEN
(Research Institute of Meteorology, Central Meteorological Service)
SUN SHAN-TSE
(Department of Mathematics & Mechinics, Peking University)
TSUI YU-HSI
(Institute of Computing Techniques, Academia Sinica)

ABSTRACT

In this paper, a stepwise multiple regression procedure is used to screen from a large number of predictors that are most significantly related to a particular predictand. If one or more predictors are not statistically significant, they may be eliminated from the equation.

The stepwise technique is here used to solve the problem of monthly and seasonal forecasting of precipitation over China. As an example, from the mid-summer data of 31 years (1932—1962), the summary of the calculation procedure of monthly precipitation over North China is illustrated. The analyses indicates that the monthly precipitation in each tract is related to different predictors, such as solar radiation indices, parameters of general circulation, and other factors for the periods prior to rainfall.

The linear regression equations (for predicting the monthly precipitation over several regions of China) based on these correlations are tested on an independent sample including 42 cases. The results are compared with those generated by chance. It is found that the prediction scores are higher than above verification standard.

Finally, a brief discussion is given concerning the improvement of regression analysis and suggestions are made for further statistical work on that problem.