大 气 科 学
1993 年增刊 SCIENTIA ATMOSPHERICA SINICA Sup., 1993

试用样本加权回归方法预测峰值

俞 善 贤
(浙江省气象科学研究所,杭州 310021)

提 要

本文采用样本加权回归方法对峰值的预测作了试验. 通过对预测点相邻区域中的样本加大其在建模中的作用,有可能克服因子作用的时空可变性和非线性的影响,提高该预测点的预测精度. 对浙江省和九个市、地早稻亩产大丰、大歉年景预测进行了对比试验,与逐步回归相比,平均报对率提高了 18%.

关键词: 加权回归;峰值预测;时空可变性.

一、引 言

众所周知,线性回归模型对峰值的预测效果往往不能令人满意,而在实际预报业务中,十分重视峰值的预测,迫待提高其预测能力,但对该问题的研究国内外尚不多见. 对此,作者认为: 在峰值预测统计方法的处理过程中应具有某些特殊性. 通过分析,发现常用回归方法在预测峰值时存在一些问题,而有些可以通过样本加权回归方法获得解决.

本文提出的对峰值预测的处理思想和相应的处理过程,还不成熟,存在一些有待进一步探索和研究的问题,作者希望获得批评指教,并愿与对此感兴趣的同志共同探讨.

二、峰值预测中的几个问题

所谓峰值是指预报对象明显大于或小于平均值的值,在整个样本集中占小数. 为提高峰值的预测,应十分有效地处理好以下几方面的关系:

1. 整体拟合与局部预测的关系

一个回归预测模型,其本质是一个总体平均的结果,为了达到整体上的平衡,往往是以牺牲局部区域"利益"作为代价的. 峰值样本占小数,必然受多数样本的影响. 而每一次预测是对某一预测点或某一局部区域而言的. 整体拟合好坏不能完全反映局部预测的好坏.

这里也为我们提供了一种思路: 即在每一次预测时,改变建模的目标函数,适当降低某些样本的作用和拟合精度,借以提高局部区域(预测点相邻区域)内的样本作用和拟合精度,从而达到提高该次预测准确性的目的.

1992 年元月 4 日收到,1993 年 4 月 20 日收到修改稿.

2. 因子作用的时空可变性和入选因子的关系

因子作用的时间可变性已被人们所接受，解决该问题通常采用多层递阶方法，它考虑了因子作用(回归系数)的时序变化．作者认为：因子作用的大小还同因子所在的区域有关，即因子作用的空间可变性．预报因子往往表现为在某一区域内预测能力强，而在另一区域则无预测能力，这也符合事物作用的机理．例如，研究某作物产量与雨量关系时，当雨量值在一定的范围内，影响通常不大，决定产量丰歉的是其它因素，只有当雨量值特多或特少时，它才起决定的作用，这样的作用因子一般很难选入回归方程中，因为计算它们的相关系数时，往往是不高的．而选入回归模型中的因子和作用的大小，也仅仅反映的是一种总体平均状态而已，不能有效地反映预测点发生变化后因子作用的大小，也就是说没有充分利用预测点已给定的信息．如果针对给定的预测点，适当选取预报因子，是可以克服因子作用的空间可变性的．

3. 非线性与样本选取的关系

峰值的产生与非线性的作用有密切关系．而在实际建模过程中，一般采用线性函数．如果预报因子与预报对象间满足线性关系，那么样本越多，建立的回归方程越合理；如果是一种非线性关系，情况就不然．为了直观，以一元回归为例说明如下：

图 1 中 x^* 为预测点，用全部样本得回归方程 A，用 x^* 的若干邻域点，得回归方程 B，显然，在预测点 x^* 用 B 比 A 更合理．实际问题中，最大的困难是多维空间中的函数形式和图像不易了解．但是我们可以用局部的线性关系来处理整体的非线性关系．具体做法是提高 x^* 邻域内样本的权重，减少这个邻域外样本的权重，或删除这些样本（权重为零），这样做有可能解决一大类非线性的影响．

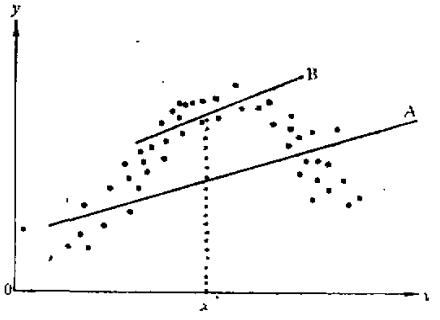

图 1　邻域点和全部样本点建立直线方程示意图

当然影响峰值预测的因素很多，但上述的几个问题具有一定的普遍性，而且可以通过样本加权回归方法的处理在较大程度上得到改善．

三、样本加权回归法的处理过程

1. 因子适用性的判别

因子作用大小的空间可变性问题，在具体做法上如何解决呢？本文分两个步骤来实现，首先，判别某一因子对当次预报是否适用，即判别该因子在 x^* 邻域点所对应的预报对象是否一致，如果意见差别很大，就删除该因子，如果意见较一致就参加建模．其次，对参加建模的因子，它们作用的大小由样本加权回归法来确定．

设有预报因子 x_i 和预报对象 $y_i(i=1,2,\cdots,n)$ 以及当次预报因子值 x_i^*，其中

$$X_i=\begin{bmatrix}x_{i_1}\\x_{i_2}\\\vdots\\x_{i_n}\end{bmatrix}\qquad Y=\begin{bmatrix}y_1\\y_2\\\vdots\\y_n\end{bmatrix}$$

找出 x_i^* 的m个邻域点，记 $x_{i_1}^*,x_{i_2}^*\cdots x_{i_m}^*$ 及相应的 $y_{i_1}^*,y_{i_2}^*\cdots y_{i_m}^*$，本文取 $m=[n/4]$. 计算

$$\bar{y}^*=\frac{\sum_{k=1}^{m}y_{i_k}}{m} \tag{1}$$

$$\hat{S}=\frac{\sqrt{\sum_{k=1}^{m}(y_{i_k}^*-\bar{y}^*)^2}}{m} \tag{2}$$

给一判别 y_i 离散度的置信值 SL. 如果 $\hat{S}>SL$，则说明 $y_{i_k}^*(k=1,2,\cdots,m)$ 之间差别较大，意见不统一，该因子对当次预报无分辨能力，予以删除. 如 $\hat{S}\leqslant SL$，则说明 $y_{i_k}^*(k=1,2,\cdots,m)$ 之间差别较小，该因子对当次预报有一定的分辨能力，参加建模. 关于 SL 取值，应结合实际例子中 y 的标准差 $\hat{S}_y$ 来确定，并根据实际效果加以调整. 实例中的 SL 取 $\hat{S}_y$ 的 1/5—2/5.

2. 样本权重的确定

样本权重确定的基本原则是：与预测点 $x^*=(x_1^*,x_2^*\cdots x_p^*)^T$ 相似程度大的样本权重也大. 通常采用欧氏距离来计算相似系数，但是由于样本大小与维数之间的关系不相匹配，研究表明通常效果不佳[3].

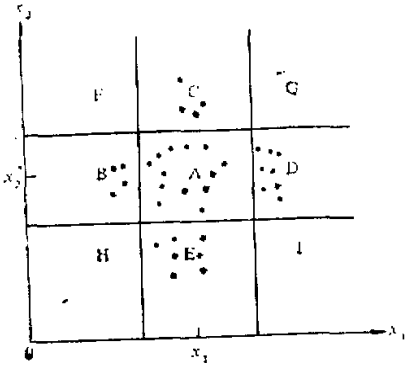

图 2 各区域内权重确定示意图

本文采用某一历史样本各因子与预测点的各分量相邻的次数的累加值作为权重. 为了直观，以二个因子为例作图 2 说明如下：

图 2 区域 A 中，x_1,x_2 与 x_1^*,x_2^* 均为邻域点，权重取值为 2；

区域 B、C、D、E 中，x_1,x_2 与 x_1^*，x_2^* 有一因子为邻域点，权重值为 1；

区域 F、G、H、I 中，x_1,x_2 与 x_1^*，x_2^* 均不为相邻点，权重取值为 0.

因子适用判别和权重确定的具体算法如下：

(1) 置初值 SL，$W_i=0,i=1,2\cdots n$；

(2) 对 x_i，找出 x_i^* 的邻域点，并用(1)、(2)式计算出$\hat{S}$；

(3) 若 $\hat{S}>SL$，则删除 x_i；若 $\hat{S}\leqslant SL$，则 $W_{i_1},W_{i_2}\cdots W_{i_m}$ 累加 1；

(4) 重复(2)、(3),直到所有因子处理完.

3. 样本加权的回归算法

对于参加建模的因子,一般还较多. 虽然它们有一定的分辨能力,但由于受错综复杂的因素影响,有时,甚至可以出现完全相反的预报意见,而且它们作用的大小也是不平等的,所以必须进行整体平衡,这个平衡过程用回归方法来实现.

文献[2]中给出了样本加权的回归算法. 但是没有象逐步回归一样具有自动选取预报因子的功能. 本文在常用的逐步回归基础上,按 W_i 值的大小,把原来一个样本看作 W_i 个样本来处理. 具体计算时只要在求平均、方差之类的计算时,乘上权重即可,这里不再细述.

这里需要特别指出的是: 上述做法涉及一理论问题,因为把一个样本看作 W_i 个来处理,这样样本间的独立性就变差,在计算选人或剔除因子的 F 检验值时,会有所不同.但在实际问题中,逐步回归的 F 检验通常也仅是用来控制入选因子个数和反映因子间相对重要程度的,而并不是严格按照 F 检验值来入选因子. 所以采用独立样本计算 F 检验值,用以控制入选因子和反映因子间相对重要性,也不妨是一种权宜之计.

四、实例试验对比

要评价一个预测方法的优劣,最好的考察是实际预报. 而本文中使用的样本较少($n=30$),不可能留下足够多的样本作试报检验,为此采用"刀切法"的思想,依次把某一个样本去掉,重新用该方法在众多因子中选择因子,建立模型,然后来预测该年的预报对象.这样做的目的是尽可能同实际预报相接近.

预报对象选为浙江省全省和九个市、地早稻气象产量预报,预报因子均为环流和海温因子,预报时效为5个月,属长期年景预报.

对大丰、大歉年用样本加权方法和逐步回归方法作了对比. 试验对比结果见表1.

报对率的评分标准为:

$$\text{报对率}=\frac{\text{报对个数}}{\text{报对个数}+\text{空报}+\text{漏报}}$$

参考中国气象局长期预报评分的有关标准,以全省早稻大丰年报对为例说明如下:气象产量 20kg 以上为丰年,允许有 5kg 的上下偏差,即预报 15kg 以上,实况为 20kg 以上评为报对;实况为 15kg 以上,预报为 20kg 以上也评为报对,其余均为报错.

空报: 实况不超过 15kg,预报为 20kg 以上.

漏报: 实况超过 20kg,预报不超过 15kg.

各地大丰、大歉的标准略有不同,但同一地区,两种方法对比均为同一标准. 两种方法选取的预报因子个数均控制在4—6个左右.

从表中可见: 报对率平均提高18%,只有金华市的报对率有所下降.

采用"刀切法"仅是考察该方法优劣的一种手段. 在实际预测中,即确定未来是大丰或大歉还是平年,同采用"刀切法"中去掉某一个样本的做法相类似. 同传统回归方法不

同的是该方法没有固定的预报方程，方程中的因子和系数大小随预测因子的变动而变化.

表 1 早稻产量大丰、大歉预报对比

预报区域	全省	杭州	嘉兴	宁波	绍兴	温州	金华	丽水	台州	舟山	平均
逐步回归报对率	0.38	0.18	0.33	0.36	0.26	0.38	0.67	0.40	0.38	0.38	0.37
样本加权报对率	0.64	0.47	0.50	0.75	0.44	0.63	0.50	0.50	0.41	0.60	0.55

五、讨 论

(1) 本文总的思路是针对预测点找出邻域点相邻的样本，作加权处理，目的是突出这些样本的作用，提高预测点(局部)的预测能力，借以改善因子的空间可变性和非线性的影响. 从实际试验对比也说明了基本思路是正确的，预测峰值的能力有所提高. 对正常年的预测同逐步回归基本相当. 这可能是正常年样本占多数，加权和不加差别不大，应该说这是正常的.

(2) 加权回归方法，对每一次预测均需重建方程，计算量稍大一些——尤其在评价该方法对历史样本预测时，但从目前省级拥有的计算手段来看，也是不成问题的.

(3) 对样本权重的确定，是一个值得探讨的问题. 应该承认本文采用的具体确定权重的方法，还不成熟，是值得试验对比的，例如可结合相似预报的某些方法，来确定相似系数，然后根据相似系数的大小确定相应的权重.

(4) 对样本加权后，选取预报因子的算法，本文曾采用 F 检验法，对存在的问题文中已作了说明，如果问题较严重，也可以采用一些非参数的方法，例如 PRESS 准则的算法，该算法不涉及统计检验的问题. 考虑到逐步回归算法已被广大气象统计预报工作者所掌握，所以本文采用该算法.

参 考 文 献

[1] 成 平、李国英，1986，投影寻踪——一类新兴的统计方法，应用概率统计，第二卷、第三期.
[2] 陈希孺、王松桂，1984，近代实用回归分析，254—256，广西人民出版社.

Forecasting Peak Value by the Method of Sample Weighted Regression

Yu shanxian

(*Zhejiang Meteorological Science Institute, Hangzhou* 310021)

Abstract

In this article a test is made to forecast the peak value with the sample weighted regression method. The effect of time and space variability and nonlinearity may be overcome by adding more samples the region near the forecast point while establishing model, so as to improve the forecast accuracy. A comparing test has been made to forecast the bumper harvest and poor harvest of early rice per mu yield in Zhejiang Province and nine cities and regions, respectively compared with the successive regression method,' the average accuracy has increased by 18 per cent.

Key words: Weighted regression; Forecasting peak value; Nonlinear.