多因变量矩阵回归预报方法

严华生 张永昆* 曹 杰 谢应齐
(云南大学地球科学系，昆明 650091)

尤卫红
(云南省气象台，昆明 650032)

摘 要 本文针对一个因变量多元回归在气象预报应用中的不足和局限，引入多因变量多元回归，讨论了应用中的公共预报因子问题和因子筛选问题，并对多因变量回归与单因变量回归、场展开预报、逐步回归、逐段回归等方法进行了比较，探讨了其特点及应用效果。

关键词 多因变量 公共预报因子 矩阵回归 统计预报

1 引言

在以往的气象统计预报中，一般是采用多个预报因子对一个预报对象建立回归方程进行预报，这是很不够的，我们常常遇到要同时预报相关联的多个预报对象的问题。例如，空间距离相隔很近的甲乙两地，属同一天气气候区域，天气气候变化密切相关。如果用同样的大气环流和海温资料作两地的长期预报，若用一个因变量的回归分析方法分别对甲乙两地建立各自的回归方程作预报，则各个回归方程入选的预报因子就可能互不相同，预报结果可能完全不同、毫无关联，它既没有考虑到预报对象之间的联系，也很难用天气气候学原理来解释，更无法从统计上谈联合估计的好坏。本文就是针对这一问题来研究多因变量矩阵回归方法在气象预报中的应用。

在多因变量预报中，若存在某个预报因子，对所考虑的每一个预报对象，都具有预报作用，我们就把它称为公共预报因子。多因变量矩阵回归，就是对同一天气气候区域内的多个预报对象，寻找这样的公共预报因子来建立矩阵回归方程作预报。

2 数学方法

设有 m 个自变量，p 个因变量，建立多因变量多自变量回归模型，当对资料进行标准化处理后，写成矩阵回归方程的形式为

$$\underset{n\times p}{\hat{Y}} = \underset{n\times m}{X}\ \underset{m\times p}{B} \tag{1}$$

式中 n 表示样本数，下标表示资料矩阵的行列数，具体方法见文献[1]。

1995-09-22 收到，1996-11-18 收到再改稿

* 现在云南省玉溪地区气象局工作

在气象统计预报中，备选因子成千上万，如何在计算条件容许的情况下，从如此众多的备选因子中，挑选出具有明显预报物理意义、历史拟合相关显著、外推预报效果稳定的优秀预报因子来建立回归方程，一直是气象学家和统计学家在不断探讨的问题，我们这里是按如下残差筛选方案来进行的：

第一步，设有 m 个备选预报因子，对其中每一个因子 x_i，$i=1,\cdots,m$，按文献[1]计算全相关系数，从中找出最大者，记为 x_1，作显著性检验，若显著，就引入该因子，建立多因变量一元回归方程，记为 $\underset{n\times p}{\hat{Y}^{(1)}} = \underset{n\times 1}{X}\ \underset{1\times p}{B}$；残差记为：$Y^{(1)} = Y - \hat{Y}^{(1)}$。

第二步，对 $Y^{(1)}$，又与每一个 x_i，计算全相关系数，从中找出最大者，记为 x_2，作显著性检验，若显著，就引入该因子，用 x_1, x_2 与 Y 建立多因变量二元回归方程，记为 $\underset{n\times p}{\hat{Y}^{(2)}} = \underset{n\times 2}{X}\ \underset{2\times p}{B}$；残差记为 $Y^{(2)} = Y - \hat{Y}^{(2)}$。

继续进行这一因子引入过程，不妨设已到第 k 步，对 $Y^{(k)}$ 与每一个 x_i，计算全相关系数，从中找出最大者，记为 x_{k+1}，作显著性检验，若显著，就引入该因子，建立多因变量 $k+1$ 元回归方程，直到没有显著影响因子引入为止。

3 应用实例

3.1 资料及建模

取云南省空间均匀分布的 21 个站 5～10 月雨量，为便于各站点的雨量及回归系数相互具有可比性，并排除异常大小的极值点的过强影响干扰，我们把雨量转化为秩序统计量作为预报对象，逐月建立雨量预报矩阵回归方程。用北半球 100 hPa、500 hPa 高度场和西北太平洋海温场这三个场 1～3 月的各月平均网格点资料作为备选因子集，共 3450 个因子，用 1956～1993 年作为样本资料建模，1994 年作为外推预报检验，1995 年进行业务预报试验。建模时随着因子的逐个引入，矩阵回归方程的总相关系数变化如图 1 所示。

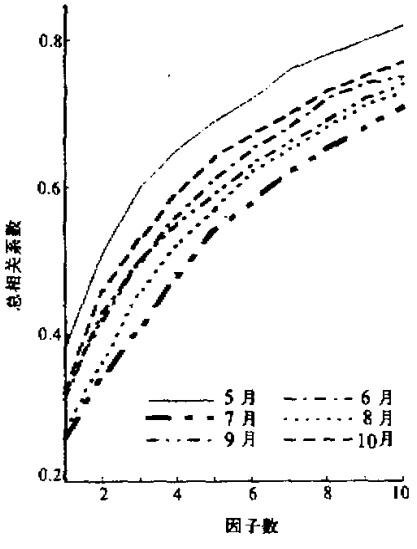

图 1 引入因子数与模型拟合总相关系数变化曲线图

如图所示，随着因子的逐渐引入，总相关在不断提高，但随着引入因子数量的逐渐增多，总相关提高得越来越慢，尤其到后几个因子时，多入选一个因子，总相关只提高 0.01～0.02，说明此备选预报因子集只能提供这么高的拟合预报能力，再往后继续入选因子已无必要。

顺便指出，我们同时用原始雨量建立回归方程作比较，得出用秩序统计量建立的回归方程，其总相关和复相关都要高一些。这说明在雨量预报中，对降雨预报对象作一些预处理，以在一定程度上消除一些随机偶然性，增大可预报性，是很有必要的。

3.2 外推预报检验

根据预报对象秩序统计量得到经验分布函数，以拐点为界，将预报对象划分为 3 个等级范围如表 1 所示。外推预报就是根据这 3 个等级来评定预报准确率的。凡预报和实况均为同一等级即判为正确，否则判为错误，分点 9 和 29 可两边靠。对每个月的预报，记录报对的站数，除以总站数，即得该月的预报准确率。对 1994 年 5～10 月外推预报检验结果如表 2 所示。

表 1　预报对象等级划分

级别	偏少	正常	偏多
秩序统计量	1～9	9～29	29～38

表 2　云南省 21 个台站 1994 年外推预报检验结果

	5 月	6 月	7 月	8 月	9 月	10 月
报对站数	14	14	14	19	14	12
报错站数	7	7	7	2	7	9
总评效果（%）	67	67	67	90	67	57

由表可见，外推预报准确率基本稳定在 67%以上，把预报和实况值绘在地理底图上进行分析可看到，预报结果连片性很好，往往报错的站点大都是成片偏多或偏少中的奇异孤立点。云南地处低纬高原山地，地形复杂，常有中小尺度局地单点大雨、暴雨出现，在月雨量空间分布图上就表现为这种奇异点。对这种奇异点，长期预报是很难报出来的。这也说明，在长期预报中，大范围区域预报的重点在于气象要素的空间分布，个别奇异点的预报需要结合当地气象及地理特点制作。

我们于 1995 年投入业务预报试验，经 1994 年 5～10 月和 1995 年 5～7 月共 189 站次预报与实况对比检验，总准确率为 69%。根据我们对 1978～1991 年云南省气象台所用经验长期天气预报的检验，降雨预报还处于随机水平[2]。1993 年我们应用浑沌理论和相空间方法，对昆明近百年月降雨量可预报性的估算得出，对 5～10 月雨量距平符号预报准确率可达到 65%[3]。由此可见，应用多因变量矩阵回归方法大范围预报效果较好，对云南省降雨长期预报水平有一定的提高，且持续稳定。

3.3 入选公共预报因子的物理意义初探

我们把所入选的预报因子分别点绘于相应的 100 hPa、500 hPa 冬季多年平均环流及西北太平洋洋流分布图上，见图 2、3 所示。

由图可见，所入选的公共预报因子均有规律地分布在大气超长波的槽脊位置及孟加拉湾、青藏高原、西太平洋洋流位置上，在 100 hPa 和 500 hPa 上对应一定的层次空间立体结构，类似于驻波形式；在西太平洋上，入选公共预报因子最多的是黑潮暖洋流区；说明冬季冷热源、大气活动中心、超长波及黑潮暖洋流的异常是夏季云南降水异常的主要影响因素，它的物理意义是很明显的，对此我们专门作过讨论[4]。以上分析提示我们，用多因变量回归方法较能选出具有一定物理意义的公共预报因子。

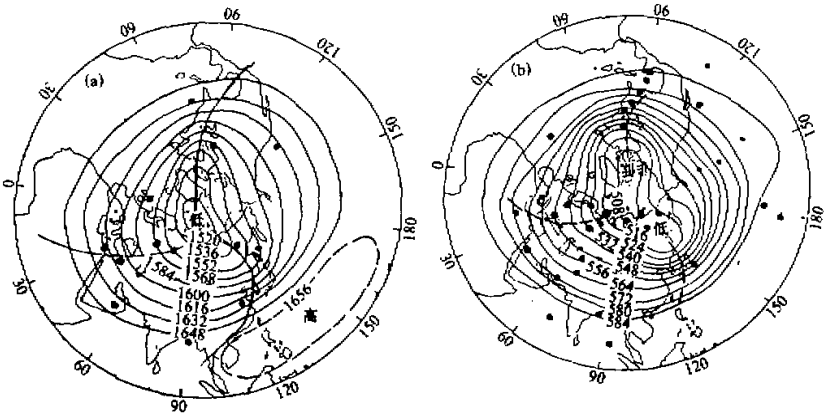

图2 100 hPa (a)、500 hPa (b)多年平均图上预报因子分布

4 讨论

4.1 单因变量与多因变量回归预报方法比较

单因变量相关回归方法筛选预报因子，只经过一个预报对象检验，随机偶然性大；而多因变量相关回归方法筛选公共预报因子，经过多个预报对象的统计检验，因而其可靠性、稳定性相对较大，易选出具有一定物理意义的预报因子。

图3 北太平洋海区预报因子分布

单因变量相关回归预报方法仅对单个预报对象进行分析预报，受个别波动及局地误差影响较大，不能反映具有较大尺度的群体变化规律，而多因变量相关回归方法正是弥补了这个不足之处，在区域范围分布及多要素预报中比单因变量相关回归方法有优越性。所以，在气象统计预报中，发展多因变量相关回归预报方法是有重要意义的。

4.2 场展开预报与多因变量预报方法比较

周家斌曾研究过用正交函数展开预报对象场，提取典型特征量作要素场水平分布预报问题[5]。若把要素场的每一个点都看作一个变量，则场的预报可转化为多因变量预报。作者对这两种预报方法进行了比较，有如下看法：

(1) 从预报角度看，当把预报对象场展开，提取前 k 个特征量来作预报时，丢失了部分原始信息，且先作出特征量预报，再根据所预报的特征量去预报要素场的水平分布值，这种预报的预报，会在一定程度上增大预报误差，因此不如用多因变量回归直接

作预报效果好。

（2）从分析角度看，场展开具有特定的几何图像，对场的分布规律描述较好，这一点比多因变量回归方法优越。

（3）当因变量个数多于样本数或者因变量间相关较大，致使行列式L_{11}趋近于零时，对多因变量回归模型的显著性检验或求全相关、总相关系数公式就会产生较大误差，甚至计算失败，此时可采用多个因变量的平均复相关或平均单相关来代替总相关及全相关系数进行分析，当然也可采用场预报方法。即使采用场预报方法，在建立前k个特征量的回归方程中，也可采用多因变量回归方法。

4.3　因子筛选问题

以往常用逐步回归或是逐段回归方案来筛选因子，现对这两种方案与本文提出的筛选因子方案进行对比分析如下：

用逐步回归筛选因子，要计算$\underset{(m+p)\times(m+p)}{L}$数据矩阵，而用本文采用的方案引入因子，则只要对每个因子x_i，$i=1,\cdots,m$，计算全相关系数并找最大。以本文例子，若用逐步回归来筛选因子，要多计算一个$\underset{(3450+21)\times(3450+21)}{L}$数据矩阵，这显然是难以进行的。因此，当备选因子成千上万时，更适宜用本文提出的方案来引入因子，其减少的计算量是很可观的。实际上，只要从数学上比较一下两种引入因子的计算过程就可发现，这两种方法引入因子的依据和计算过程是等价的。黄嘉佑[6]对此曾作过讨论。若在本文引入因子建立的 k 因子回归方程中，再加一个从 k 因子回归方程中剔出一个因子的因子剔出过程，实际上就是一个计算量比逐步回归小得多的完整的因子引入、剔出筛选过程。

以往的逐段回归方案是这样的[7]：设有 Y 与X_m，从中找出相关最大者建立一元回归方程，记为$\hat{Y}_1=B_1X_1$，残差记为$Y_2=Y-\hat{Y}_1$；再对Y_2与x_m找出相关最大者，又建立一元回归方程，记为$\hat{Y}_2=B_2X_2$，残差记为$Y_3=Y_2-\hat{Y}_2$，如此继续直到没有显著因子引入为止，于是得$\hat{Y}=\hat{Y}_1+\hat{Y}_2+\cdots+\hat{Y}_k=B_1X_1+B_2X_2+\cdots+B_kX_k$，简化得逐段回归最终方程，记为

$$\hat{Y}=B_1X_1+B_2X_2+\cdots+B_kX_k \tag{2}$$

其中B_1的估计由上面所述逐段回归过程计算得出，请注意每一步都是一元回归！

若我们以这 k 个因子$x_1,\cdots,x_k$与 Y 重新进行多元回归，可得 k 元回归方程，记为

$$\hat{Y}=b_1x_1+\cdots\cdots+b_kx_k \tag{3}$$

则逐段回归（2）式中的回归系数与多元回归（3）式中的回归系数是不相同的，即$B_i\neq b_i$，（2）式的复相关系数小于（3）式，（2）式的残差大于（3）式。这就说明逐段回归方程中的回归系数不是无偏估计，其筛选过程和所建立的方程不能保证是最优筛选和最优方程。这是因为在因子引入过程中，设当已有 k 因子回归方程，再引入第 k+1 个因子进入方程中，则所有前 k 个因子的回归系数都要改变，而逐段回归则保持回归系数不变。显然，逐段回归对回归系数的估计是有偏的。而本文所提出的筛选方案正是改进了逐段回归的不足之处，在筛选过程中，当筛选到第 k 个因子时，是计算 k 元回归而不是像逐段回归那样仍然是计算一元回归，从而保证了每一步对回归系数的估计都

是无偏估计和因子的最优筛选，最终得出最优方程。

5 小结

通过本研究得出如下结论：

(1) 提出了对区域或多要素气象统计预报，最好采用筛选公共预报因子集，建立多因变量矩阵回归方程来制作，其有更客观的物理意义和较好的效果。

(2) 给出了一种因子筛选方案，该方案与逐步回归相比，计算量要小得多，与逐段回归相比，统计优良性有了明显改善。

(3) 对多因变量回归与单因变量回归、正交函数场展开回归几种方法进行了比较讨论，认为多因变量回归方法是值得进一步深入研究和推广应用的。

参 考 文 献

1 严华生、王学仁，1991，多因变量及要素场统计预报，北京：气象出版社.
2 尤卫红，1992，对云南省气象台近年来长期天气预报的检验，云南气象，第2期，30～31.
3 尤卫红、严华生，1995，月雨量的可预报性估算试验，热带气象学报，11(1)，73～79.
4 严华生等，1995，云南省5月雨量的天气气候成因分析，应用气象学报，6(1)，124～127.
5 周家斌，1990，车贝雪夫多项式及其在气象中的应用，北京：气象出版社.
6 黄嘉佑，1993，统计动力分析与预报，北京：气象出版社，228～230.
7 王宗皓、李麦村，1978，天气预报中的概率统计方法，北京：科学出版社，99～100.

Multiple Dependent Variable Matrix Regression Forecast Method

Yan Huasheng, Zhang Yongkun, Cao Jie and Xie Yingqi
(*Department of Earth Sciences, Yunnan University, Kunming* 650091)
You Weihong
(*Meteorological Observatory of Yunnan Province, Kunming* 650032)

Abstract To over come the defect and the limit that were caused by single dependent variable regression model, this paper uses the multiple dependent variable matrix regression method to solve the problem of common predictors and the problem of screening predictors . The multiple dependent variable matrix regression forecast method is compared with other statistical regression methods. The characteristic and the effect using this method are also analyzed.

Key words multiple dependent variable common prediction factors matrix regression statistical forecast