典型相关系数及其在短期气候预测中的应用*

张礼平 杨志勇 陈正洪
(武汉中心气象台，武汉 430074) (武汉工业大学，武汉 430070) (武汉暴雨研究所，武汉 430074)

摘 要 借助典型相关系数，对场与场的关系进行分析，并由短期气候预测理论与实践以及线性方程组理论，提出了多因子场预测未来要素场的新方法。

关键词：广义相关系数；典型相关分析；最小二乘法

1 引言

在近代短期气候预测中，场的分析和预测所占比重越来越大，首先是由于用户希望得到一个场的预测，而不是孤立的单站预测；其次就是人们逐渐认识到局地因素造成的小尺度扰动使得单站要素变化随机性大，规律不易掌握，而大范围场的变化规律随机性小，可预报性大于单点。

我们也曾做过场的分析和预报，即将预报对象场中的每一点与因子场中的每一点进行相关分析，选取通过检验相关关系好的格点，然后通过回归分析对单站要素进行预测，最后组合成一预报场。另一种做法是用典型相关分析提取的典型变量作为因子，对预报场单点进行回归分析，严格地说这仍是单站预测，因为都是独立的单站分析和预测，没有考虑预测场中点与点之间的相互联系，很可能造成地理位置相邻的站点要素预测值差异很大，这显然与天气学原理相悖。还有一种做法是用因子场的全部典型变量，由回归分析预测预报对象场对应典型变量，最后恢复为预报对象场，由于选用了关系并不好的典型变量，这将直接影响其预报精度。

本文将借助典型相关系数，对场与场间的关系进行统计分析，探寻因子场预报要素场的新方法。

2 基本原理

一般来说，两个随机变量关系的密切程度是用相关系数来度量的，相关系数绝对值大的就认为关系密切。对于x、y两个场，可分别视为两组随机变量，分别假设为p和q个变量，观测样本为n，经方差标准化处理，可用矩阵表示为

1999-01-14 收到，1999-05-24 收到再改稿
* 本工作得到"九五"攻关项目 96-908-05-04-07 资助

$$X=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ x_{p1} & x_{p2} & \cdots & x_{pn} \end{bmatrix}, \qquad Y=\begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1n} \\ y_{21} & y_{22} & \cdots & y_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ y_{q1} & y_{q2} & \cdots & y_{qn} \end{bmatrix}.$$

若Lyy^{-1}、Lxx^{-1}存在，则定义$Myx=Lyy^{-1}LyxLxx^{-1}Lxy$为$x$和$y$的线性关联阵，其中$Lyy=YY^T$、$Lxx=XX^T$、$Lyx=YX^T$、$Lxy=XY^T$。称矩阵$Myx$的秩为$x$、$y$的相关秩，记为$r$，$Myx$的全部非零特征根按大小用$\lambda_1$、$\lambda_2$、…、$\lambda_r$表示。$x$、$y$关系的密切程度可用广义相关系数来度量，广义相关系数有5种定义，为了便于建立预测模式，我们取$\rho_{xy}^{(3)}=\max\limits_{1\leqslant i\leqslant r}\lambda_i$[1]。$\sqrt{\max\limits_{1\leqslant i\leqslant r}\lambda_i}$即为两组变量间最大的典型相关系数，也即两组变量线性组合后可能达到的最大相关。由于典型相关分析中典型相关系数总是对应着x、y两组变量特定的线性变换（组合），所以，$\rho_{xy}^{(3)}$对应的两个线性变换，使得x、y两组变量分别变换为新变量（典型变量）ξ_1和η_1，典型相关系数也即为ξ_1和η_1一元统计相关系数。由于它是最大的典型相关系数，所以一般都能通过信度检验。既然ξ_1和η_1具有两场最大可能的相关关系，x组为具有实况值的因子场，这就提示我们可由ξ_1和η_1的一元回归分析来预测η_1。由于原变量都已方差标准化处理，使得平均值为0，线性变换后平均值仍为0，且典型变量的方差为1，由回归分析理论可知，ξ_1和η_1的回归方程中常数项为0，回归系数即为$\sqrt{\rho_{xy}^{(3)}}$，这样就容易由x实况值对η_1进行估计。设x的变换矩阵为$\underset{1\times p}{a}$，y的变换矩阵为$\underset{1\times q}{b}$，则预测公式为

$$\underset{1\times q}{b}\,y=\eta_1=\sqrt{\rho_{xy}^{(3)}}\,\xi_1=\sqrt{\lambda_1}\,\underset{1\times p}{a}\,x. \tag{1}$$

然而由于$\eta_1=by$，考虑到$y=(y_1,y_2,\cdots,y_q)^T$，y仍无法惟一确定，这在数学上称为约束条件不够。为了解出y，要求构造更多的约束。另一方面，从短期气候预测的理论与实践可知，短期气候的变化过程与短期天气过程的重要区别是：前者影响因子多，后者因子少。很显然，某月、季降水（或气温）场并非仅由前期一二个因子场的变化引起，要根据多个因子场的变化对要素场的未来变化进行尽可能全面的描述，否则得到的极可能是片面的预测结果。

由上两方面的原因，促使我们努力寻找多个因子场，对未来要素场进行预测。设已选取m个因子场，由（1）式可估计出η_1、η_2、…、η_m。这里$i(i=1,2,\cdots,m)$对应第i因子场。设第i因子场对应的预测公式为

$$\underset{1\times q}{b_i}\,y=\eta_i=\sqrt{\lambda_i}\,\xi_i=\sqrt{\lambda_i}\,\underset{1\times p}{a_i}\,x, \tag{2}$$

或

$$\underset{1\times q}{b_i}\,y=\sqrt{\lambda_i}\,\underset{1\times p}{a_i}\,x. \tag{3}$$

设

$$\underset{1\times q}{b_i} = (b_{i1} \quad b_{i2} \quad \cdots \quad b_{iq}),$$

$$\underset{1\times p}{a_i} = (a_{i1} \quad a_{i2} \quad \cdots \quad a_{ip}),$$

则（3）式也可表示为

$$b_{i1}y_1 + b_{i2}y_2 + \cdots + b_{iq}y_q = \sqrt{\lambda_i}\,(a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{ip}x_p).$$

可构造线性方程组

$$\begin{cases} b_{11}y_1 + b_{12}y_2 + \cdots + b_{1q}y_q = \sqrt{\lambda_1}\,(a_{11}x_1 + a_{12}x_2 + \cdots + a_{1p}x_p), \\ b_{21}y_1 + b_{22}y_2 + \cdots + b_{2q}y_q = \sqrt{\lambda_2}\,(a_{21}x_1 + a_{22}x_2 + \cdots + a_{2p}x_p), \\ \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots, \\ b_{m1}y_1 + b_{m2}y_2 + \cdots + b_{mq}y_q = \sqrt{\lambda_m}\,(a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mp}x_p). \end{cases} \tag{4}$$

用矩阵表示为

$$\underset{m\times q}{B}\ \underset{q\times 1}{y} = \underset{m\times 1}{\eta},$$

或

$$By - \eta = 0, \tag{5}$$

这里 B 为由预测对象场 y 与第 i 因子场典型相关分析的第 1 典型变量变换行矩阵组成的 $m\times q$ 矩阵，η 为与第 i 因子场对应由（2）式估计的 η_i 组成的 $m\times 1$ 列矩阵。

由线性方程组理论，$m = q$ 时，y 即为线性方程组的解；$m > q$ 时，即在方程个数多于变量个数时，（4）式可能无解，也即不一定能找到一组 y_1，y_2，…，y_q，使得

$$f(y_1, y_2, \cdots, y_q) = (By - \eta)^T (By - \eta) = 0. \tag{6}$$

那么，我们设法寻找 y_1^0，y_2^0，…，y_q^0，使得 $f(y_1^0, y_2^0, \cdots, y_q^0)$ 为最小，这样的 y_1^0，y_2^0，…，y_q^0 就称为方程组（4）的最小二乘法解，其几何解释是：要找到这样的向量 y，使线性组合向量 By 到向量 η 的距离最短。如果 $m = q$ 时，这个距离就为 0，可见最小二乘法解是一般线性方程组解的推广，而线性方程组的解是最小二乘法解的一个特例。

由线性代数理论，最小二乘法解要满足

$$B^T(By - \eta) = 0,$$

或

$$B^T By = B^T \eta,$$

这是一个线性方程组，系数矩阵为 $B^T B$，常数项为 $B^T \eta$。解出的 y_1^0，y_2^0，…，y_q^0，即为要素场 q 个点的预测值。这样就找到了由典型相关系数预测要素场的方法。为了更好地理解典型相关系数，下面我们将它与自然正交函数（EOF）做一比较。

典型相关分析（CCA）与 EOF 都是对变量组实施线性变换，前者是对两组变量同

时实施不同的变换，使得变换后的一对新变量（典型变量）具有最大的线性相关，且方差为1；后者是对一组变量实施线性变换，使得新变量（主分量）具有最大的方差，且保持变量组总方差不变。可见，两者尽管都是线性变换，但目的和作用是显然不同的：CCA的典型变量，建立了两组变量的联系，用广义相关系数可以定量度量两组变量关系的密切程度，其广义相关系数平方根即为这对典型变量的相关系数，同时也是可用于由因子场典型变量预测预报场典型变量的回归方程系数，为从因子场提取和浓缩主要信息预测预报要素场提供了可能；而EOF是自身变量组内的线性变换，可起到浓缩场的信息、突出场的主要特征、以较少变量代替多变量的作用，所以EOF本身并不是预报方法而是一种分析方法。

3 实例

按照上面的思路，我们对湖北省武汉、郧县、老河口、恩施、宜昌、荆州、咸宁、黄石、麻城、随州10站1998年6～8月总降水量场进行预报试验。因子场为1～4月北半球500 hPa高度、海平面气压共8个场（均为576空间点）。

为了保证典型变量的稳定，典型相关分析要求分析样本n大于变量场空间点数。而500 hPa高度、海平面气压场空间点数目远远大于样本个数，且相邻空间点一般具有较大的相关，致使矩阵求逆困难。针对这种情况，采用Barnett和Preisendorfer提出的方案[2]，先对预报对象场和因子场分别进行EOF分析，将变量场投影到前几个EOF上，然后将得到的两场主分量作为两组新变量进行典型相关分析。这不仅减少了变量个数，使变量个数小于分析样本n，又使同组变量间相互正交，方便了CCA中的矩阵求逆运算，且浓缩了原场的主要信息。

对1959～1997年预报对象场方差标准化，然后进行EOF分析，用1959～1997年的平均值，标准差对1998年资料方差标准化。对1959～1998年因子场方差标准化后进行EOF分析。预报场截取了前4个主分量，可解释总方差的83%。每个因子场均截取了前11个主分量，它们都可解释本场总方差的70%以上。预报场的4个主分量分别与这8个场的11个主分量进行典型相关分析，分析样本为1959～1997年，1998年因子场主分量值用于预测。其最大典型相关系数见表1：

表1 预报场主分量与因子场主分量最大典型相关系数

因子场名	500 hPa				海平面气压			
因子场时间	1月	2月	3月	4月	1月	2月	3月	4月
最大典型相关系数	0.79	0.65	0.72	0.79	0.74	0.61	0.64	0.73

由于原变量已方差标准化，使得平均值为0，经EOF、CCA后平均值仍为0，且方差为1，因此表1中的典型相关系数，也是由因子场典型变量预测预报场对应典型变量的回归方程系数。由（2）式第i因子组实况值分别乘以对应变换矩阵元素之和，即得到实况ξ_i，回归系数乘以ξ_i即得到η_i的预报。

8个因子场，重复上述计算，可构造方程组（4），这里$m=8$、$p=11$，$q=4$。用豪

斯荷尔德变换求解线性最小二乘问题[3]，解出的 y 即为预报场主分量的预报值，由预报场空间函数将主分量反演为原场预报值，结果见表 2。

表 2　1998 年 6～8 月总雨量场预报与实况（方差标准化值）

站名	武汉	郧县	老河口	恩施	宜昌	荆州	咸宁	黄石	麻城	随州
实况	1.4319	−0.1144	0.1979	1.9991	1.1177	0.8733	1.5058	3.8367	−0.5857	0.8033
预报	0.034	0.033	−0.027	0.068	0.104	0.069	0.020	0.021	0.030	0.056

由于变量已方差标准化，输出结果不需处理即可知道雨量较常年多还是少，且可大致估计其量级。从表 2 可见，除郧县、老河口、麻城趋势预测与实况相反外，其余 7 站趋势预测均正确，其预报效果比较理想。

4　结语

场对场的预报，或多因变量对多预报量的预报，可看作是点对点、多点对单点，或单因子、多因子对单预报量的推广。由于场分布的连续性，使得每个场都具有自己的特征，单站预报没有考虑周围站点的相互联系，不具备连续性，所以单站分析预报后拼加在一起的预测场，也不能反映出场的水平分布固有的物理联系。

CCA 的典型变量，建立了因子场与预测场总体上的联系，其广义相关系数可定量地度量这种联系的密切程度，从短期气候预测的理论与实践，以及线性方程组理论，促使我们要用多个因子场预测某个要素场，同时也使问题可用最小二乘法求解。采用 BP 方案典型相关分析[2]，不仅可减少变量个数，又使同组变量间相互正交，方便了 CCA 中的数值计算，且近一步浓缩了原场的主要信息，使得大范围数值场预测小范围数值场成为可能。月、季尺度的气候变化，是众多因素共同影响的最后结果，很难单独试验某因素所起的作用。所以在做预报时，有必要考虑一切可能影响因素。本文预报方法，考虑的是多因子场对要素场这种场与场间的最主要影响关系，重点考虑大尺度变化，滤去小扰动，力图从尽可能多的方面对要素场变化进行主体描述，这种思路与天气学原理相一致。本方法为短期气候预测提出了一条新思路。

参　考　文　献

1　张尧庭，方开泰，多元统计分析引论，北京：科学出版社，1982, 305～332.

2　Barnett, T. P. and R. W. Preisendorfer, Origins and levels of monthly and seasonal forecast skill for United States surface air temperatures determined by canonical correlation analysis, *Mon. Wea. Rev.*, 1987, **115**, 1825～1850.

3　徐士良，C常用算法程序集，北京：清华大学出版社，1996, 27～29.

Canonical Correlation Coefficients and Their Applications in a Short-Range Climate Prediction

Zhang Liping

(*Wuhan Central Meteorological Observatory, Wuhan* 430074)

Yang Zhiyong

(*Wuhan University of Technology, Wuhan* 430070)

Chen Zhenghong

(*Wuhan Heavy Rain Institute, Wuhan* 430074)

Abstract The relationship between one space and others is analysed by the generalized correlative coefficients, and a new method forecasting the spaces of weather elements with a multitude of factor's space is proposed according to the theory and practice of short-range climate predictions and the theory of group of linear equations.

Key words: generalized correlative coefficient; canonical correlation; least square