doi:10.11676/qxxb2013.075 气象学报

逐对剔除的相关系数检验方法及应用*

徐寒列[1,2] 李建平[1] 冯 娟[1]
XU Hanlie[1,2] LI Jianping[1] FENG Juan[1]

1. 中国科学院大气物理研究所大气科学和地球流体力学数值模拟国家重点实验室,北京,100029
2. 中国科学院大学,北京,100049
1. *State Key Laboratory of Numerical Modeling for Atmospheric Sciences and Geophysical Fluid Dynamics, Institute of Atmospheric Physics. Chinese Academy of Sciences, Beijing* 100029, *China*
2. *University of Chinese Academy of Sciences, Beijing* 100049, *China*

2012-12-11 收稿,2013-05-23 改回.

徐寒列,李建平,冯娟. 2013. 逐对剔除的相关系数检验方法及应用. 气象学报,71(5):901-912
Xu Hanlie, Li Jianping, Feng Juan. 2013. The pair-wise deletion correlation coefficient testing method and its applications. *Acta Meteorologica Sinica*, 71(5): 901-912

Abstract To quantitatively investigate the stability and authenticity of the statistical correlation coefficient, we propose a pair-wise deletion correlation coefficient (PWDCC) testing method which is based on χ^2 test, through pair-wise deleting elements of a data sample and constructing a correlation coefficient series. The analysis using both ideal data sample and real climate data confirms the validity and feasibility of the PWDCC. The results show that, this method can objectively and quantitatively determine whether there exist extreme values which lead to large variation of the sample correlation coefficient. Furthermore, the method is simple and has clear physical significance, and it also overcomes some weakness of the traditional methods (the scatter diagram, the slide correlation method, etc).

Key words Correlation coefficient test, Extreme value, χ^2 test

摘 要 为了定量考察样本相关系数的真实性和稳定性,基于 χ^2 检验方法,通过逐对剔除样本中的元素并构造相关系数组的方式,提出了逐对剔除的相关系数检验方法,并用理想数据样本和真实气候数据验证了该方法的正确性和可行性。结果表明,此方法可以客观、定量地找出样本中影响相关系数真实性和稳定性的极端值,这些极端值的存在会使样本的相关系数发生较大变化。本方法的检验过程简单、物理意义清楚,并且,检验过程被定量化,克服了传统方法(散点图、滑动相关等)的一些不足。

关键词 相关系数检验,极端值,χ^2 检验
中图法分类号 P468.0

1 引 言

线性相关分析是气候统计中常用的分析方法之一,用于考察两个物理量之间的线性关系。线性相关系数与一元线性回归的斜率均可用于表示两个物理量之间的线性相关程度。由于现有气候资料的长度一般只有几十年,相对于气候系统这个较长时间尺度的总体来说只是一个样本,于是在求得两个气候变量的相关系数之后,通常会对计算所得的相关系数进行显著性检验(t 检验),从而考察样本的相

* 资助课题:国家重点基础研究发展计划项目(2010CB950400)和国家自然科学基金项目(41030961、41205046)。
作者简介:徐寒列,主要从事北大西洋涛动与东亚气候相关研究。E-mail: xuhanlie@mail.iap.ac.cn
通讯作者:李建平,主要从事非线性气候动力学及可预报性、季风和海气相互作用、环状模动力学及其影响等方面的研究。E-mail: ljp@lasg.iap.ac.cn

关系数是否可以代表总体相关系数。但是,经常会忽略一个重要的细节,即极端事件的发生可能会对气候变量产生影响,进而影响气候变量间的相关性,例如 Xiao 等(2011)研究发现 1991 年皮纳图博火山的一次大规模爆发可能导致了 20 世纪 90 年代平流层的年代际变冷。因此,在对相关系数进行检验时需要考虑到极端值的影响。

事实上,许多研究人员早就注意到了极端值,并且对极端值理论进行了系统的研究(Fisher, et al, 1928;Balkema, et al, 1974)。Burry(1975)指出,极端值是非典型、不常出现的观测值,它的出现对两组样本的相关系数和线性回归的斜率都会有很大的影响。有时只要存在一对极端值就能够改变样本的相关系数和线性回归的斜率,使原本不显著的相关系数变得显著,或者使显著相关变得不显著。在这种情况下,样本相关系数很可能是虚假的或者不稳定的,不能代表总体相关性的真实情况。近年来,极端天气气候事件的研究已经成为气象学的重要研究领域,中国许多学者从不同的角度分析了极端天气气候事件,研究内容主要集中在极端事件阈值的选取及极端事件的变化特征(任福民等,1998;Yan, et al, 2002;侯威等,2011;黄琰等,2011;李庆祥等,2011)、极端事件的气候影响及预测(翟盘茂等,2003;封国林等,2009)以及极端事件与大气环流的关系(江志红等,2009;孙建奇等,2011;李娟等,2012;尹姗等,2012)等方面,然而在极端事件的发生对于气候变量之间相关性的影响方面的研究则相对较少。在真实的气候现象中,极端事件的影响不容忽视。例如,在讨论南半球环状模(SAM)对澳大利亚西南部冬季降水的可能影响时,一些学者指出澳大利亚西南部冬季降水的变化与南半球环状模的变化有关(Ansell, et al, 2000;Cai, et al, 2005;Li, et al, 2005);然而,Feng 等(2010b)的进一步研究发现,南半球环状模与澳大利亚西南部冬季降水的显著关系是由于发生在 1964 年的极端气候事件造成的。他们研究发现,1964 年南半球环状模指数处于近百年的极端负位相,相对应的当年澳大利亚西南部降水异常偏多,当去掉 1964 年的数据之后,两者的相关不再显著。可见,极端值的存在对两组样本的统计关系存在着重要影响,因此,在气候统计中即使相关系数通过了显著性检验(t 检验),也需要对其真实性和稳定性做进一步的分析。

在气候统计中,考察样本的相关系数是否受到极端值的影响或相关系数在分析时段内是否稳定时,常用的方法是分析数据的散点图或对样本进行滑动相关分析(Bell, 1977; 林学椿,1978)。这两种方法各有利弊,分析样本散点图的方法简单直观,但是,判断标准较为主观;而滑动相关分析方法虽然存在客观的判断标准,但是,极端值的存在可能使滑动相关结果出现明显的年代际变化(Feng, et al, 2010a),很可能导致错误的结论。因此,如何准确、定量地检验相关系数的真实性和稳定性,并定量地确定可能存在的影响相关系数的极端值的个数和位置,还有待于进一步的研究。

针对上述问题,本研究基于 χ^2 检验方法,通过逐对剔除构造相关系数组的方式,提出一种简便、客观的分析方法——逐对剔除的相关系数检验方法,剔除样本中影响相关系数的极端值(如果存在),从而客观定量地检验样本线性相关系数的真实性和稳定性,使计算得到的样本相关系数更加真实可靠。

2 数据和方法

所用的数据有:计算机产生的随机数、南半球环状模指数和澳大利亚西南部冬季降水指数,其中南半球环状模指数选用 40°S 与 70°S 标准化的纬向平均逐月海平面气压之差(Nan, et al, 2003),该指数可以很好地表征南半球中、高纬度气压反向变化的特征,被广泛地用于气候分析中(李建平,2005;南素兰等,2005a, 2005b;Wu, et al, 2009;Feng, et al, 2010a,2012;李晓峰等,2009,2010;Blunden, et al, 2011;李建平等, 2011;郑菲等,2012;Sun, et al, 2012),澳大利亚西南部冬季降水指数为 Feng 等(2010a)所选取的指数,时间长度均为 1948—2007 年。

使用的方法主要有费希尔 Z 变换、t 检验和 χ^2 检验。

由于逐对剔除的相关系数检验方法的基础是对正态总体方差的检验方法——χ^2 检验,而当两个总体的相关系数 $\rho\neq 0$ 时,总体中任意两组样本的相关系数不服从正态分布,且总体相关系数越大,样本相关系数的分布越偏离正态分布。因此,在进行检验之前,需先对相关系数进行处理,使其服从正态分布。Fisher(1915)提出的费希尔 Z 变换可解决这个问题。对于样本相关系数 r_i,费希尔 Z 变换可以表示为

$$z_i = \frac{1}{2}\ln\frac{1+r_i}{1-r_i} \tag{1}$$

其中,z_i 为经过费希尔 Z 变换后的结果。不服从正

态分布的相关系数 r_i 经过费希尔 Z 变换后得到的 z_i 服从正态分布。

使用计算机随机产生的数据对费希尔 Z 变换的可靠性进行分析。由计算机随机产生两组相互独立且服从正态分布的数据,其样本量均为 100,相关系数 $\rho_1=0.5$,将这两组数据看作总体,且认为两组数据一一对应。随机抽取总体中 40 对数据作为样本,求其相关系数,并重复抽样 1000 次,得到相关系数样本($r_1,r_2,\cdots,r_{1000}$),相关系数 r_i 为随机变量。对 $r_i(i=1,\cdots,1000)$进行费希尔 Z 变换得到 $z_i(i=1,\cdots,1000)$。同样地,随机产生两组相互独立且服从正态分布、样本量为 100,但相关系数 $\rho_2=-0.5$ 的数据作为总体,进行同样的处理。分位数-分位数图(简称 Q-Q 图)是检验随机变量是否服从正态分布的常用方法(宗序平等,2010)。Q-Q 图以样本的分位数和按照正态分布计算的相应分位点作为坐标,把样本表现为直角坐标系中的散点。如果资料服从正态分布,则样本点应该呈现一条直线。图 1 为两次试验中相关系数样本进行费希尔 Z 变换之前和之后的 Q-Q 图结果,可见,当总体相关系数分别为 0.5 和 -0.5 时,$r_i(i=1,\cdots,1000)$与标准线并不一致,不服从正态分布(图 1a、c),而经过费希尔 Z 变换之后,$z_i(i=1,\cdots,1000)$与正态分布期望值具有明显的线性关系,接近标准线(图 1b、d)。因此,费希尔 Z 变换可以使不服从正态分布的随机变量接近正态分布。为了进一步验证费希尔 Z 变换的作用,对上述数据进行了单样本 Kolmogorov-Smirnov 拟合优度检验(简称 K-S 检验)。K-S 检验是检验样本是否来自某一特定分布的方法,它以样本数据的累计频数分布与特定理论分布比较,若两者的差距很小,则认为该样本取自某特定的分布,此处对数据进行正态分布检验。表 1 分别给出了相关系数为 0.5 和 -0.5 的样本数据经过费希尔 Z 变化前后的 K-S 检验结果与对应的置信度。在经过费希尔 Z 变换后,相关系数正态分布的置信度明显提高,这也进一步证明了费希尔 Z 变换的作用。

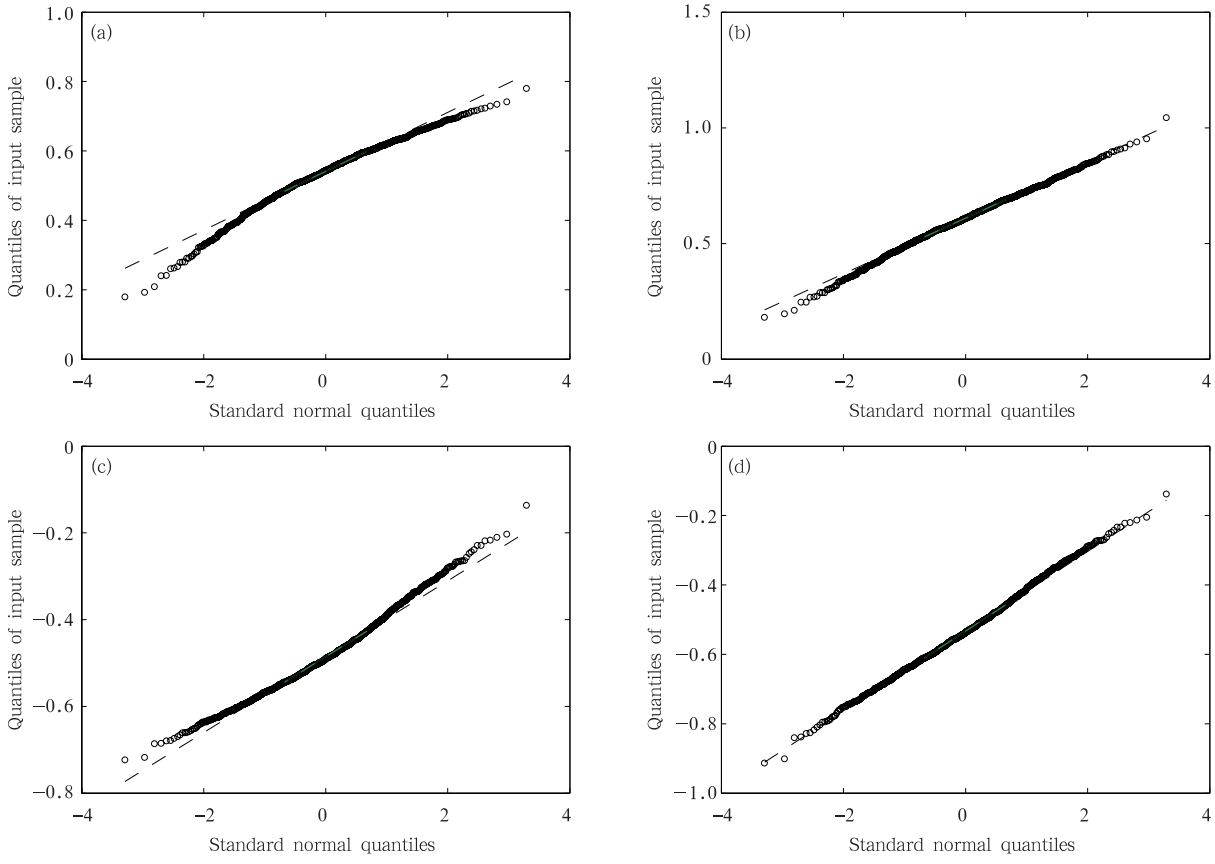

图 1 相关系数样本在费希尔 Z 变换之前(a、c)和之后(b、d)的正态分布 Q-Q 图

(a、b. 总体相关系数为 0.5, c、d. 总体相关系数为 -0.5)

Fig. 1 Normal distribution Q-Q diagrams of the correlation coefficient sample before (a, c) and after (b, d) the Fisher Z transformation ((a) and (b) represent the case with the general correlation coefficient of 0.5, and (c) and (d) are as in (a) and (b) but for the general correlation coefficient of -0.5)

表 1 相关系数样本在费希尔 Z 变换前后的单样本 K-S 检验结果以及对应的置信度

Table 1 The results of the single-sample K-S test and the corresponding confidence levels of the correlation coefficient sample before and after the Fisher Z transformation

	相关系数为 0.5		相关系数为 -0.5	
	变换前	变换后	变换前	变换后
K-S 检验结果	0.041	0.018	0.035	0.018
对应置信度	0.073	0.895	0.165	0.914

χ^2 检验在天气预报、工农业生产统计中应用较为广泛,通常用于对单个正态总体的方差进行假设检验(盛骤等,2001)。当需要判断总体方差 σ^2 是否等于 σ_0^2(σ_0^2 为已知常数)时,原假设和备择假设分别为

$$\begin{aligned} H_0&:\sigma^2 = \sigma_0^2 \\ H_1&:\sigma^2 \neq \sigma_0^2 \end{aligned} \tag{2}$$

由于样本方差 s^2 是总体方差 σ^2 的无偏估计,取

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} \tag{3}$$

作为检验统计量,给定显著性水平 α,当 $\chi^2_{1-\alpha/2}(n-1) \leqslant \chi^2 \leqslant \chi^2_{\alpha/2}(n-1)$ 时,则接受原假设,认为总体的波动性没有发生较大的变化;否则,拒绝原假设,认为总体的波动性发生了较大的变化,拒绝域为 $\chi^2 < \chi^2_{1-\alpha/2}(n-1)$ 或 $\chi^2 > \chi^2_{\alpha/2}(n-1)$。

文中相关系数的统计检验方法为 t 检验。在下面的介绍中,如无特殊说明,所用符号都采用统计学中的常用记法。例如,对任意随机变量 x,其平均值表示为 $\bar{x}$,标准差表示为 s_x,方差表示为 s_x^2。

3 逐对剔除的相关系数检验方法

本文所提出的逐对剔除的相关系数检验方法的基本思路如下:

假设 $x_1, x_2, \cdots, x_n$ 是来自总体 $X \sim N(\mu_1, \sigma_1^2)$ 的样本,$y_1, y_2, \cdots, y_n$ 来自总体 $Y \sim N(\mu_2, \sigma_2^2)$ 的样本,其中 μ_1、σ_1^2、μ_2、σ_2^2 均为未知数,$x_1, x_2, \cdots, x_n$ 和 $y_1, y_2, \cdots, y_n$ 一一对应,且相关系数为 r。依次去掉样本中的第 $i(i=1,2,\cdots,n)$ 对数值 (x_i, y_i),余下样本的相关系数记为 r_i,从而得到相关系数组 $(r_1, r_2, \cdots, r_n)$。如果相关系数 r 受到某对极端值 (x_j, y_j) 的影响,则去掉 (x_j, y_j) 之后,余下样本的相关系数 r_j 会与原相关系数 r 相差较大,即 r_j 可能为相关系数组 $(r_1, r_2, \cdots, r_n)$ 中的离群值。根据 Grubbs (1969)的定义,离群值是一个显著偏离它所在样本中其他成员的数据。在统计学上,离群值是在数值上与其他数据相差很大的观察值(Barnett, et al, 1994)。因此,可以通过考察相关系数组中数据的波动性的变化情况来判断相关系数组中是否存在离群值。这里采用 χ^2 检验法对数组的波动性是否发生了显著变化进行检验。依次构造检验统计量 $\chi_i^2 (i=1,2,\cdots,n)$,给定显著性水平,如果 χ_j^2 存在于拒绝域中,则认为 s_{r_j} 与相关系数组的方差 s_r 的差别是显著的,数据的波动性发生了显著变化,对应的 r_j 为相关系数组中的离群值,原始样本中的 (x_j, y_j) 为影响相关系数真实性的极端值。

根据以上基本思路,逐对剔除的相关系数检验方法的具体计算过程如下(为了方便起见,以下记样本为 $X=(x_1, x_2, \cdots, x_n)$,$Y=(y_1, y_2, \cdots, y_n)$):

(1)计算样本相关系数

样本 X 和 Y 的相关系数为

$$r = r(X,Y) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \cdot \sum_{i=1}^{n}(y_i-\bar{y})^2}} \tag{4}$$

(2)逐对剔除并构造相关系数组

在 $x_1, x_2, \cdots, x_n$ 和 $y_1, y_2, \cdots, y_n$ 中,分别去掉第 $i(i=1,2,\cdots,n)$ 对数据 (x_i, y_i),将余下的 $(n-1)$ 对样本分别记为 $X_i = X \backslash x_i$ 和 $Y_i = Y \backslash y_i$,其相关系数为 $r_i = r(X_i, Y_i)$,样本的相关系数组记为 $(r_1, r_2, \cdots, r_n)$。

对 r_i 是否服从正态分布进行检验(K-S 检验或 χ^2 检验),若不服从正态分布,则对 $(r_1, r_2, \cdots, r_n)$ 中的元素分别进行费希尔 Z 变换,使之正态化,得到 $(rf_1, rf_2, \cdots, rf_n)$,再对数组 $(rf_1, rf_2, \cdots, rf_n)$ 或 $(r_1, r_2, \cdots, r_n)$(若 r_i 服从正态分布)标准化(方差为 1,方便后续的计算),得到 $(rz_1, rz_2, \cdots, rz_n)$,此时,相关系数组的方差变为 $s_r = 1$。

(3)构造方差数组

在 $(rz_1, rz_2, \cdots, rz_n)$ 中,依次去掉第 $i(i=1,2,\cdots,n)$ 个值 rz_i,求余下 $(n-1)$ 个值的方差 s_{r_i},从而得到方差数组 $(s_{r_1}, s_{r_2}, \cdots, s_{r_n})$。

(4)构造 χ^2 统计量,考察方差数组的波动性是否发生变化

逐个考察方差数组中 s_{r_i} 与 $s_r=1$ 的差异是否显著。原假设和备择假设分别为

$$H_0: s_{r_i} = s_r \qquad H_1: s_{r_i} \neq s_r \tag{5}$$

构造统计量 $\chi_i^2 = \dfrac{(n-1)s_{r_i}}{s_r}$。给定显著性水平 α,依次对方差数组中的每个值进行 χ^2 检验,如果某个 s_{r_i} 所对应的检验统计量 χ_i^2 处于拒绝域中,则拒绝原假设,认为在相关系数组 $(rz_1, rz_2, \cdots, rz_n)$ 中去掉 rz_i 时,相关系数组波动性发生了显著变化,即原始样本中数据对 (x_i, y_i) 对样本相关系数 r 的影响比较显著;若所有检验统计量 χ_i^2 都存在于接受域中,则可认为相关系数组的波动性没有受到任何值的影响,即相关系数组中无离群值存在,样本相关系数 r 真实稳定,可以代表总体相关系数。

(5)如果在第 4 步中,某个 χ_i^2 处于拒绝域中,去掉原始样本中 (x_i, y_i),得到两组新样本 $X_i = X \backslash x_i$ 和 $Y_i = Y \backslash y_i$。反复执行上述过程中 1—5 步,直至相关系数组中不存在离群值。当样本中去掉全部影响相关系数的极端值后,计算得到的相关系数可被认为是代表了两组样本的相关性的真实情况。图 2 为计算步骤的流程图。

4 方法验证

通过在理想情况和实际气候数据中应用逐对剔除的相关系数检验方法,以验证此方法的有效性。极端值对相关系数的影响可能存在两种情况:(1)由于极端值的存在,使原本具有较高相关性的数据变得相关性较低,从而相关系数未能通过显著性检验;(2)由于极端值的存在,使原本相关性较低的数据变得相关性较高,从而相关系数可能通过显著性检验。在实际的气候统计中,两种情况均可能存在。因此,下面采用计算机随机产生的理想数据和真实气候数据分别进行试验,来验证此检验方法的正确性和可行性。

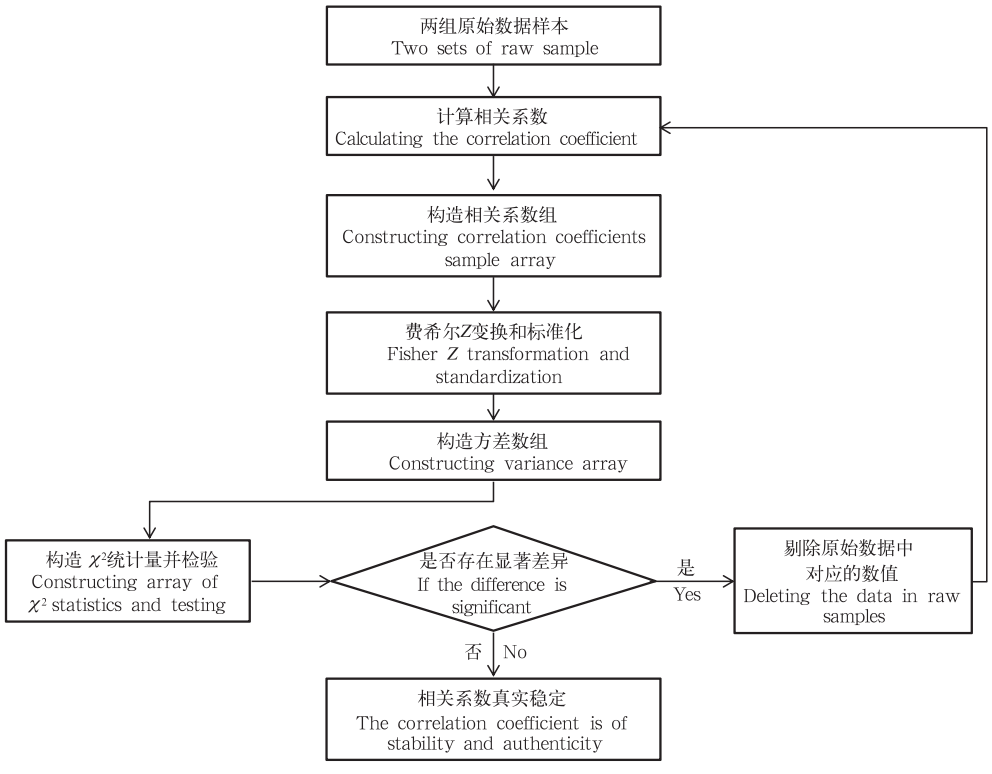

图 2 逐对剔除的相关系数检验方法的计算步骤流程

Fig. 2 Flow chart of the pair-wise deletion correlation coefficient testing method

4.1 理想数据验证

由计算机随机产生相关系数为某确定值的两组样本数据,且这两组样本分别来自两个相互独立且服从正态分布的总体。再人为地加入一对可对相关系数产生较大影响的数据(认为此数据为样本中的极端值),组成试验样本。理想试验分成两组:试验1,原始样本相关系数比较显著,通过人为加入一对数据,使样本的相关系数变得不显著;试验2,原始样本的相关系数不显著,加入一对数据之后,使样本具有显著相关。

在试验1中,由计算机随机产生样本量为49,相关系数 $r_{real\_1}=0.61$ 的两组数据 $X=(x_1,x_2,\cdots,x_{49})$ 和 $Y=(y_1,y_2,\cdots,y_{49})$ 作为试验样本,并且,认为 X 和 Y 一一对应。人为加入一对极端值,使得样本的相关系数显著下降。试验2与试验1类似,但随机产生的样本相关系数 $r_{real\_2}=0.18$。人为加入一对极端值,使样本的相关系数显著上升。因此,作为试验样本的样本量 $n=50$,相关系数分别为 $r_{ex\_1}=0.27$(没有通过0.1的显著性水平的统计检验)和 $r_{ex\_2}=0.34$(通过了0.1的显著性水平的统计检验)。试验1和试验2的样本数据散点分布见图3。

图4给出根据试验1的样本数据所构造的相关系数组、方差数组以及对应的 χ^2 统计量,由图4a可见,当去掉第50对数据时,样本的相关系数明显升高,达到0.6左右,而当样本中包含第50对数据时,相关系数均在0.3左右,并没有通过0.1的显著性水平的统计检验。因此,第50对数据的存在把原始样本的相关系数降低至0.3左右,使样本的相关系数并不能反映真实的相关情况。对此相关系数组求其对应的方差数组,从方差数组(图4b)变化中可以发现,当去掉相关系数组中的第50个数据时,样本方差变化很大,而当包含第50个数据时,样本的方差均在1左右,与相关系数组的方差 s_r 相近。图4c为根据方差数组计算的 χ^2 统计量,只有当 $i=50$ 时,所对应的 χ^2_{50} 落入信度为0.01的统计检验的拒绝域中,因此拒绝原假设,即去掉相关系数组中第50个数据时,相关系数组的波动性发生了显著的变化。可以认为相关系数组中的第50个数据为离群值,对应的原始样本中的第50对数据 (x_{50},y_{50}) 是影响样本相关系数的极端值。

去掉原始数据样本中的第50对数据 (x_{50},y_{50}) 之后,重新计算得到的相关系数组(图4d)、方差数组(图4e)和对应的 χ^2 检验(图4f)结果可见,在去掉 (x_{50},y_{50}) 之后的新样本数据所构造的相关系数组中,虽然相关系数组存在一定的波动,但相关系数都在0.6上下,数据之间不存在较大的差异;在方差数组(图4 e)中,当相关系数组中去掉某些值之后,余下数据的方差也发生了一定的变化,当去掉第5个或第14个值时,方差约为0.8,当去掉第2个或第26个数据时,方差约为0.9,方差的变化相对较大,但所对应的 χ^2 统计量均在接受域中(图4f),表明当去掉这些数据之后,相关系数组的方差变化并不显著。以上说明相关系数组中不再存在离群值,检验过程停止。两组原始数据在去掉 (x_{50},y_{50}) 之后的相关系数为0.61,反映了真实相关情况。

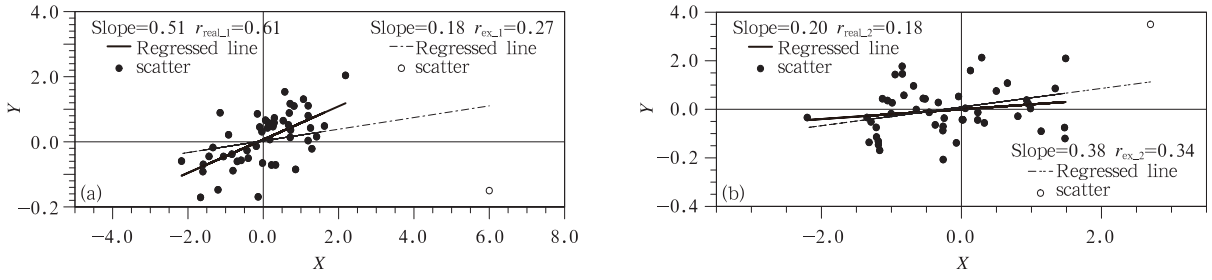

图3 两组理想试验中随机样本数据 X 和 Y 的散点分布

(a. 试验1, b. 试验2;实心圆点为由计算机随机产生的数据,实线为利用最小二乘法拟合的这两组样本的线性关系,空心圆点表示人为加入的数据,点线表示加入这对数据之后所构成的试验样本的线性关系)

Fig. 3 Scatter plots of the two data samples X and Y used in the ideal test 1 (a) and test 2 (b)

(the solid dots are randomly generated by computer and the solid line is their corresponding linear fitting, the blank dot is the extreme value which is artificially inserted, and the dotted line is the linear fitting that includes the blank dot)

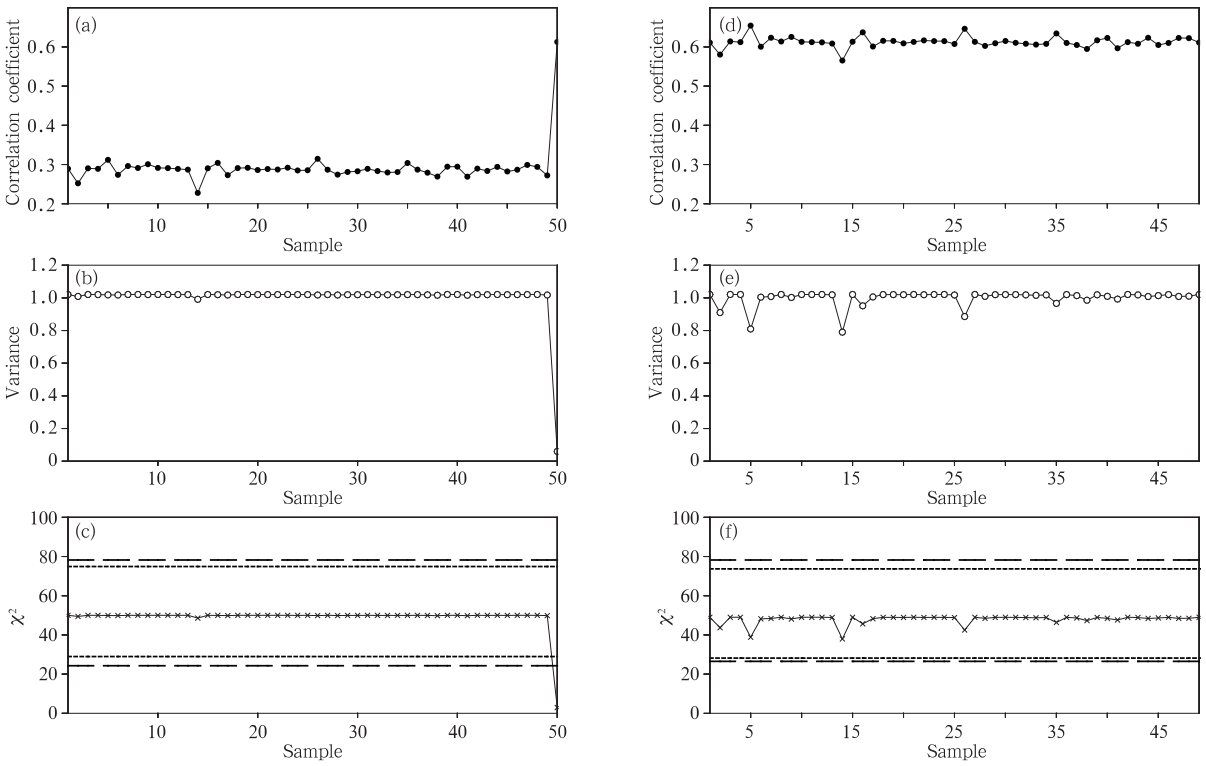

图 4 理想试验 1 的数据计算得到的相关系数组(a、d)、方差数组(b、e)以及对应的 χ^2 统计量(c、f)
(a、b、c. 原始样本计算的结果,d、e、f. 去掉原始样本中的极端值之后的结果;(c)和(f)中的短虚线
和长虚线分别表示自由度为$(n-1)$的 χ^2 分布中 0.02 和 0.01 显著性水平的统计检验的阈值)

Fig. 4 Correlation coefficient sample array (a, d), the variances sample array (b, e) and the array of χ^2 (c, f) which are calculated based on the data in the ideal test 1
(The left column is based on the raw data samples, and the right column is as the left column but based on the samples after excluding the extreme values. In (c) and (f), the short and long dashes represent the thresholds at the significant levels of 0.02 and 0.01 for the χ^2 distribution with a freedom of $(n-1)$, respectively)

试验 2 的过程与试验 1 相同,试验结果见图 5。由图 5a 可知,当去掉第 1 对数据后,余下样本的相关系数明显降低到 0.2 以下,而包含第 1 对数据时相关系数均为 0.3 以上。从方差数组的情况可见,去掉相关系数组中第 1 个数据之后,方差发生了明显的变化,而包含第 1 个数据时,方差并没有发生明显的变化。计算对应的 χ^2 统计量的结果(图 5c)可知,当 $i=1$ 时,所对应的 χ_1^2 落入拒绝域中,因此,相关系数组中 r_1 为相关系数组中的离群值,对应的原始样本中的第 1 对数据(x_1, y_1)就是影响样本相关系数的极端值。去掉原始数据中的(x_1, y_1)之后重新进行检验,结果表明,去掉(x_1, y_1)之后新的数据样本的相关系数组(图 5d)和方差数组(图 5e)中存在一定的波动,但 χ^2 检验的结果(图 5f)显示,方差数组中各数据与相关系数组的方差并没有显著的差异,可以认为此时的数据样本中不存在影响相关系数结果的极端值,检验过程停止。因此,原始样本的相关系数 $r=0.34$ 受到极端值的影响,并不能真实地表征样本以及样本所对应的总体的真实相关情况,只有当去掉极端值之后所计算的相关系数才是真实可靠的。

图 6a 和 b 分别直观地表现出了理想试验 1 和理想试验 2 的情况。由图 6a 可见,理想试验 1 的两组原始样本的相关系数为 0.27,逐对剔除并构造的相关系数组中,大部分相关系数均为 0.3 左右,只有一个相关系数的值为 0.6,远远偏离了其他相关系数值。通过前面的结果(图 6a)可知,0.6 对应于去掉(x_{50}, y_{50})之后所得到的相关系数,而原始数据去掉第 50 对数据之后的新样本数据所构造的相关系数组中,相关系数都在 0.6 左右,无明显离群值存

在。因此,原始样本的相关系数 $r=0.27$ 受到极端值的影响,并不能真实的表征样本以及样本所对应的总体的真实相关情况,只有当去掉极端值之后所计算的相关系数才是真实可靠的。由图 6b 可见,理想试验 2 的两组样本的相关系数为 0.34,当依次去掉一对数据之后,得到的相关系数组中明显存在离群值(相关系数小于 0.2),而在去掉第 1 对数据之后,得到的相关系数组样本均未通过 0.1 的显著性水平的统计检验。

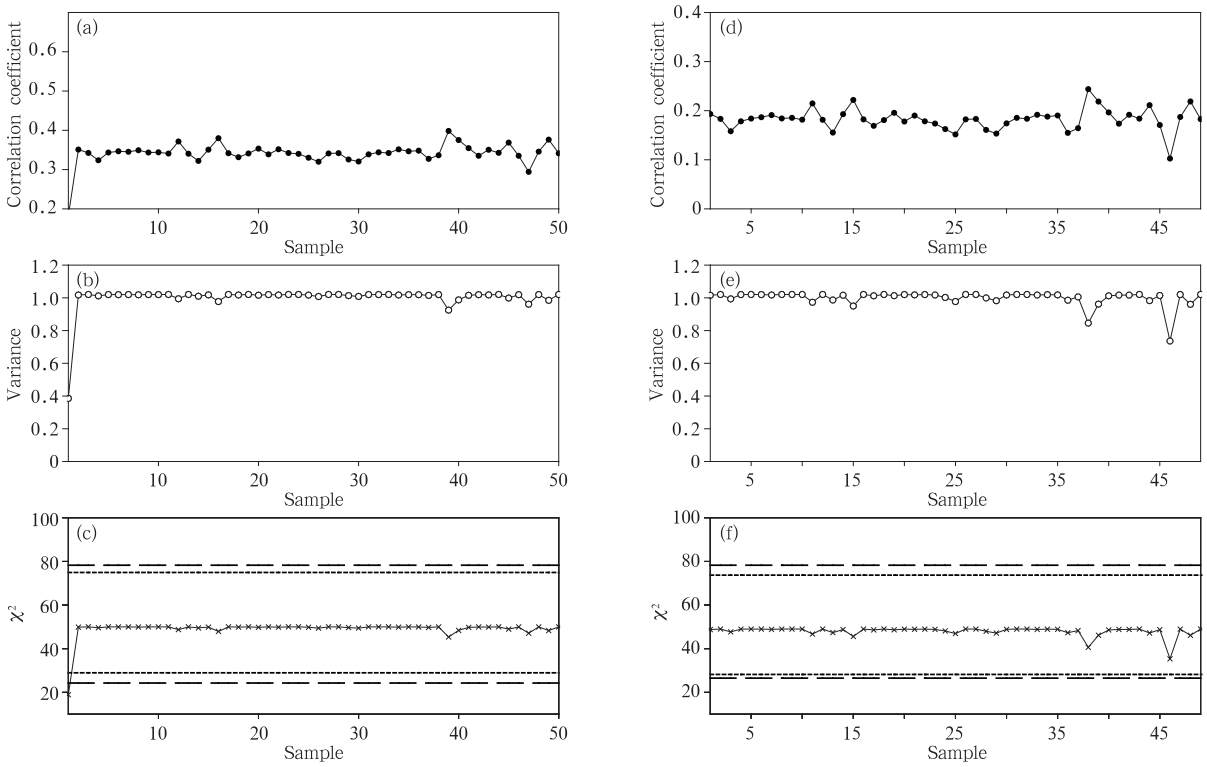

图 5 同图 4,但为理想试验 2 的结果

Fig. 5 As in Fig. 4 but for the ideal test 2

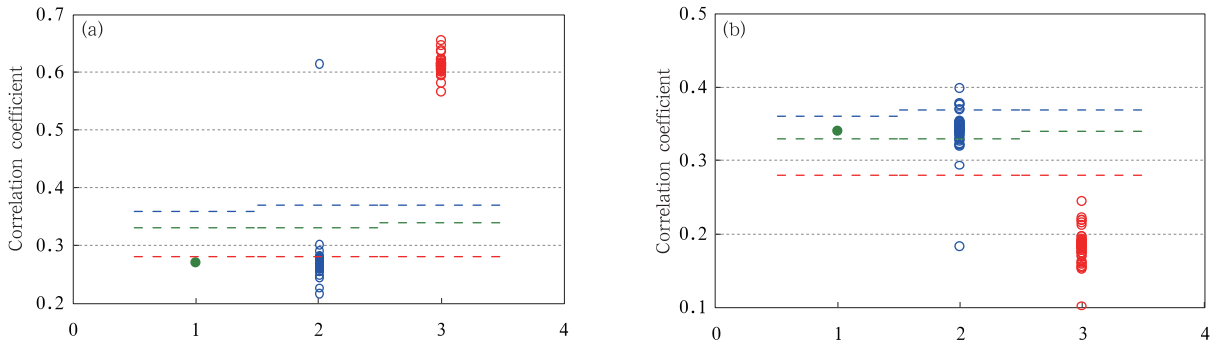

图 6 理想试验 1(a)和理想试验 2(b)的数据样本的相关系数以及构造的相关系数组

(其中横坐标 1 对应原始数据样本的相关系数,横坐标 2 对应原始数据样本的相关系数组,横坐标 3 对应于原始样本中去掉极端值后的新数据样本的相关系数;图中虚线由下至上分别表示相关系数的 0.05、0.02 和 0.01 的显著性水平的统计检验的阈值)

Fig. 6 Raw correlation coefficients and the constructed correlation coefficient samples in the ideal test 1 (a) and ideal test 2 (b)

(In the abscissa "1" is for the raw correlation coefficient, "2" is for the correlation coefficients array when the pairs of values from the raw sample are sequentially deleted, "3" is for the correlation coefficients array of the new sample in which the extreme values are excluded; the dashed lines from the bottom to up indicate the thresholds of the significance levels of 0.05, 0.02 and 0.01, respectively)

在以上两组理想试验的情况中,对于随机产生的两组来自正态分布的样本,当加入一对影响相关系数的数据之后,逐对剔除的相关系数检验方法可以准确、客观地找出相关系数组中的离群值和原始样本中影响相关系数的极端值,试验结果证实了此方法在理想情形下的正确性和可行性。

4.2 真实气候数据验证

在理想情形下,已经验证了逐对剔除的相关系数检验方法的正确性和可行性,下面将此方法用于真实的气候数据。试验数据为 1948—2007 年共 60 年的南半球冬季(6—8 月)季节平均的南半球环状模指数和澳大利亚西南部降水指数(Feng, et al, 2010a),从冬季南半球环状模指数和澳大利亚西南部降水指数的散点分布(图 7)可见,两者的相关系数达到了 -0.41,通过了 0.05 显著性水平的统计检验。Feng 等(2010b)的分析认为 1964 年的极端事件使两者的关系变得显著,去掉 1964 年的数据之后相关系数为 -0.27,并不显著。利用逐对剔除的相关系数检验方法对 Feng 等(2010b)的结果做进一步的验证。

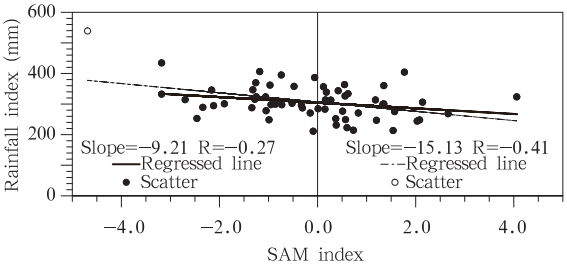

图 7 南半球冬季(6—8 月)南半球环状模指数和澳大利亚西南部降水指数(单位:mm)的散点分布

(空心圆点表示 1964 年所对应的值;实线表示样本的线性相关趋势,点线表示去掉 1964 年之后,样本的线性相关趋势)

Fig. 7 Scatter plot of the austral winter (June − August) SAM index and the Southwest Western Australia rainfall index

(The blank dot is the value in 1964, the solid line is the linear fit that includes the blank dot, and the dotted line is the linear fit after excluding the blank dot)

图 8同图4,分别给出了根据南半球冬季南半球

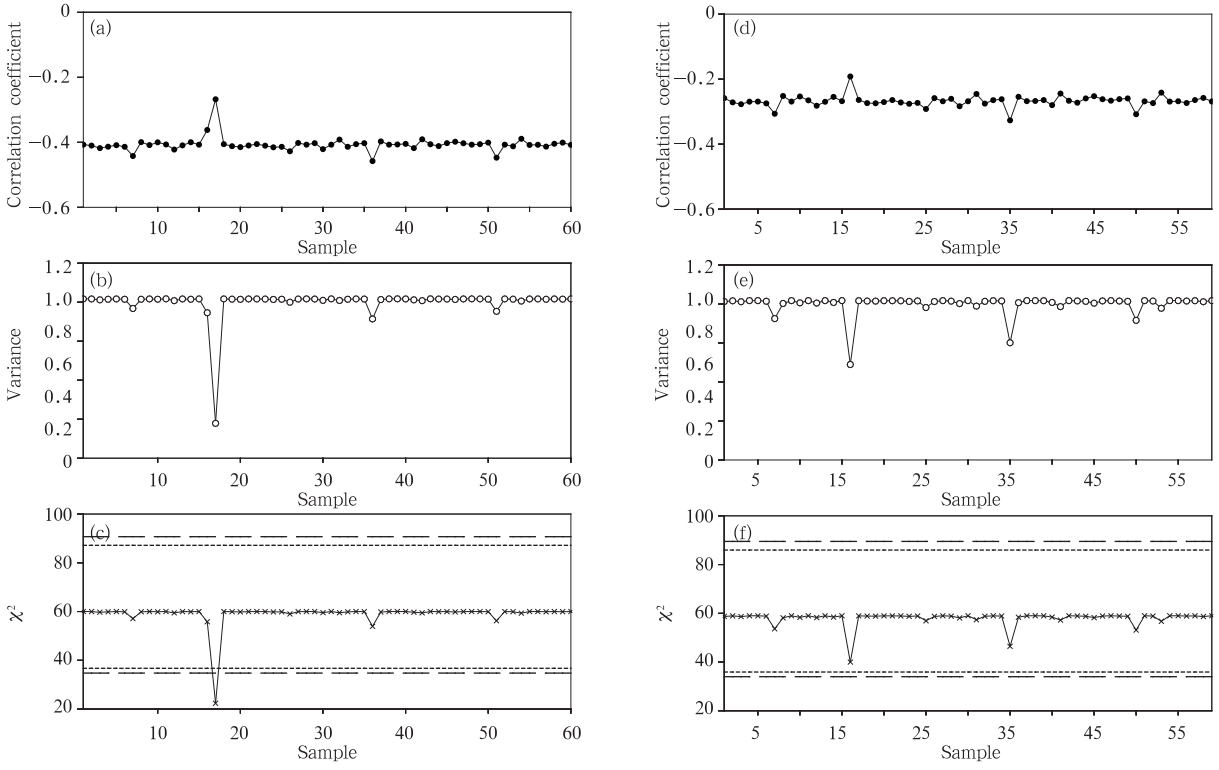

图 8 同图 4,但为南半球冬季(6—8 月)南半球环状模指数和澳大利亚西南部降水指数的试验结果

Fig. 8 As in Fig. 4 but for the results based on the austral winter (June − August) SAM index and the Southwest Western Australia rainfall index

环状模指数和澳大利亚西南部降水指数计算得到的相关系数组、方差数组和 χ^2 检验数组。由图 8 可知，在 $i=17$ 时，即去掉 1964 年的数据时，样本的相关系数明显降低到约 0.27。在方差数组中，$i=17$ 时，方差的变化相对明显。χ^2 统计量数组的结果可以证明(图 8c)，当 $i=17$，即对应于 1964 年时，所对应的 χ^2_{17} 落入拒绝域中，即去掉相关系数组中第 17 个数据时，相关系数组样本的波动性发生了显著的变化。因此，1964 年的数据为相关系数组中的离群值，对应的原始样本中的 1964 年就是影响样本相关系数的极端年。在去掉 1964 年之后的新数据样本中，重新计算得到的相关系数组、方差数组以及 χ^2 检验数组虽然数据存在一定的波动性，但是当给定 0.02 的显著性水平的统计检验，则 χ^2 统计量均存在于接受域中，可以认为相关系数组中不存在极端值，检验过程停止。图 9 更为直观的表现出数据中极端值的存在对于相关系数的影响。

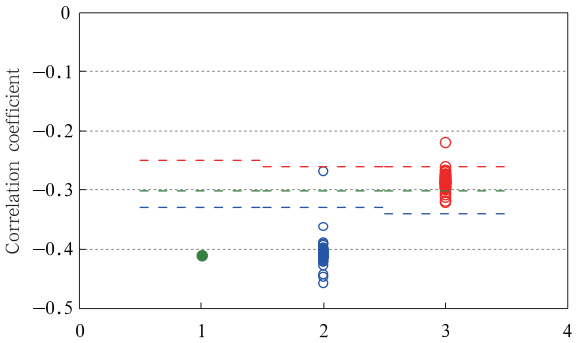

图 9 同图 5a，但为南半球冬季(6—8 月)南半球环状模指数和澳大利亚西南部降水指数的试验结果

Fig. 9 As in Fig. 5a but for the results based on the austral winter (June－August) SAM index and the Southwest Western Australia rainfall index

在本试验中，1964 年为南半球冬季南半球环状模的极端负位相年，而同期澳大利亚西南部降水为极端偏多，由于这一对极端反位相关系的存在，使南半球冬季南半球环状模指数和澳大利亚西南部降水指数的相关系数表现为显著相关。但是这种相关系数并非真实的，并不能代表南半球冬季南半球环状模指数和澳大利亚西南部降水指数的真实相关情况。只有在去掉 1964 年的数据之后，计算得到的相关系数才能代表南半球冬季南半球环状模和澳大利亚西南部降水关系的真实情况。因此，南半球冬季南半球环状模与澳大利亚西南部降水在统计上并不存在显著的相关。

5 结论和讨论

基于 χ^2 检验提出了逐对剔除的相关系数检验方法，并将其用于对气候数据分析中常用的线性相关的真实性和稳定性的分析。该方法可以客观、定量地判断气候数据样本中是否存在影响样本相关系数的极端值。本方法在分析相关系数的真实性和稳定性方面，通过逐对剔除来构造相关系数组，从而进行假设检验，检验过程简便，判断标准客观，结论相对准确，相对于传统方法(散点图法和滑动相关法)有了很大改进。

本研究采用理想数据和真实气候数据对该方法的正确性和可行性进行了验证。理想试验分为两组，第 1 组在计算机随机产生的两组显著相关的数组中加入一对影响相关系数的极端值，使原本显著相关的数据样本的相关系数变得不显著；第 2 组是在计算机随机产生的两组相关不显著的数组中加入一对极端值，使原本不显著相关的数据样本的相关系数变得显著。采用逐对剔除的相关系数检验方法分别对两类理想情况进行检验，结果显示此方法可以定量的检验出人为加入的、影响样本相关系数的极端值。当去掉极端值后，两组样本相关系数变得稳定。在真实的气候数据试验中，对南半球冬季南半球环状模指数和澳大利亚西南部降水指数的相关关系进行了分析。结果发现，1964 年是影响相关系数的极端年，当去掉 1964 年后，两者的相关系数比较稳定，这与 Feng 等(2010a)的结果一致。通过理想试验和真实气候数据验证均表明，本研究提出的逐对剔除的相关系数检验方法可以准确、客观、定量地判断两组样本相关性的真实性和稳定性。值得注意的是，线性相关方法是统计学中普遍使用的方法。因此，该方法不仅可以用于定量地判断气候数据相关系数的真实性和稳定性，而且也可以应用于其他学科，用于判断两组数据之间线性相关关系的稳定性和真实性。

以上讨论的是在样本中存在一对极端值的情况下，对相关系数的真实性和稳定性的检验。如果样本中存在两对或者两对以上数值非常相近或者相同的极端值，可能对样本相关系数造成影响，这种情况该如何进行检验呢？可以根据逐对剔除的相关系数检验方法的思路，进行逐多对剔除的检验。比如，样本量为 n 的两组数据中，假设存在 m 对影响相关系

数的极端值,那么可以进行逐 m 对的剔除以达到对样本相关系数真实性和稳定性进行定量化检验的目的。

线性相关是气象统计中常用的相关分析方法,但其易受到极端值和离群值的不良影响。在非参数统计方法中,斯皮尔曼秩相关(Spearman's Rank correlation)对离群值和极端值不敏感,并且,适用于资料不是正态分布或总体分布未知的情况。虽然斯皮尔曼秩相关系数可以考察数据中是否存在离群值和极端值,但其结果并不能定量地检验出样本中可能存在的离群值或极端值的个数以及其所在位置,而本研究提出的逐对剔除的相关系数检验方法则可以较好地解决这个问题。该方法的提出基于皮尔逊线性相关,但其检验思想仍然可以应用于其他的相关系数的检验中,以考察相关系数的真实性和稳定性,并定量地找出样本中可能存在的离群值和极端值。

参考文献

封国林,杨杰,万仕全等. 2009. 温度破纪录事件预测理论研究. 气象学报,67(1):61-74

黄琰,封国林,董文杰. 2011. 近50年中国气温、降水极值分区的时空变化特征. 气象学报,69(1):125-136

侯威,章大全,周云等. 2011. 一种确定极端事件阈值的新方法:随机重排去趋势波动分析方法. 物理学报,60(10):790-804

江志红,杨金虎,张强. 2009. 春季印度洋SSTA对夏季中国西北东部极端降水事件的影响研究. 热带气象学报,25(6):641-648

李建平. 2005. 海气耦合涛动与中国气候变化//秦大河. 中国气候与环境演变(上卷). 北京:气象出版社,324-333

李建平,吴国雄,胡敦欣. 2011. 亚印太交汇区海气相互作用及其对我国短期气候的影响(上卷). 北京:气象出版社,516pp

李娟,董文杰,严中伟. 2012. 中国东部1960—2008年夏季极端温度与极端降水的变化及其环流背景. 科学通报,57(8):641-646

李庆祥,黄嘉佑. 2011. 对我国极端高温事件阈值的探讨. 应用气象学报,22(2):138-144

李晓峰,李建平. 2009. 南、北半球环状模月内活动的主要时间尺度. 大气科学,33(2):215-231

李晓峰,李建平. 2010. 月内尺度南半球环状模对应的大气环流异常传播特征. 大气科学,34(6):1099-1113

林学椿. 1978. 统计天气预报中相关系数的不稳定性问题. 大气科学,2(1):55-63

南素兰,李建平. 2005a. 春季南半球环状模与长江流域夏季降水的关系I:基本事实. 气象学报,63(6):837-846

南素兰,李建平. 2005b. 春季南半球环状模与长江流域夏季降水的关系II:印度洋、南海海温的"海洋桥"作用. 气象学报,63(6):847-856

任福民,翟盘茂. 1998. 1951—1990年中国极端气温变化分析. 大气科学,22(2):217-227

盛骤,谢式千,潘承毅. 2001. 概率论与数理统计(第三版). 北京:高等教育出版社,225-231

孙建奇,王会军,袁薇. 2011. 我国极端高温事件的年代际变化及其与大气环流的联系. 气候与环境研究,16(2):199-208

宗序平,姚玉兰. 2010. 利用QQ图与PP图快速检验数据的统计分布. 统计与决策,(20):151-152

尹姗,冯娟,李建平. 2012. 前冬北半球环状模对春季中国东部北方地区极端低温的影响. 气象学报,71(1):96-108

翟盘茂,潘晓华. 2003. 中国北方近50年温度和降水极端事件变化. 地理学报,58(增刊):1-10

郑菲,李建平. 2012. 前冬南半球环状模对春季华南降水的影响及其机理. 地球物理学报,55(11):3542-3557

Ansell T, Reason C, Smith I, et al. 2000. Evidence for decadal variability in southern Australian rainfall and relationships with regional pressure and sea surface temperature. Int J Climatol, 20(10):1113-1129

Balkema A, Laurens de Haan. 1974. Residual life time at great age. Annals of Probability, 2(5):792-804

Barnett V, Lewis T. 1994. Outliers in Statistical Data. 3rd ed. New York: John Wiley & Sons,604pp

Bell G T. 1977. Changes in sign of the relationship between sunspots and pressure, rainfall and the monsoons. Weather, 32(1):26-32

Blunden J, Arndt D S, Baringer M O. 2011. State of the Climate in 2010. Bull Amer Meteor Soc, 92(6):S1-S236

Burry K V. 1975. Statistical Methods in Applied Science. New York: John Wiley & Sons

Cai W, Shi G, Li Y. 2005. Multidecadal fluctuations of winter rainfall over southwest Western Australia simulated in the CSIRO Mark 3coupled model. Geophys Res Lett, 32(12):L12701, doi:10.1029/2005 GL022712

Feng J, Li J P, Li Y. 2010a. A monsoon-like southwest Australian circulation and its relation with rainfall in Southwest Western Australia. J Climate, 23(6):1334-1353

Feng J, Li J P, Li Y. 2010b. Is there a relationship between the SAM and Southwest Western Australian winter rainfall? J Climate, 23(22):6082-6089

Feng J, Li J P, Xu H L. 2012. Increased summer rainfall in northwest Australia linked to southern Indian Ocean climate variability. J Geophys Res, 118(2):467-480, doi:10.1029/2012JD018323

Fisher R A. 1915. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. Biometrika, 10(4):507-521

Fisher R A, Tippett L H C. 1928. Limiting forms of the frequency distribution of the largest or smallest member of a sample. Proc Cambridge Phil Soc, 24(2):180-190

Grubbs F E. 1969. Procedures for detecting outlying observations in samples. Technometrics, 11(1):1-21

Li Y, Cai W, Campbell E. 2005. Statistical modeling of extreme rainfall in Southwest Western Australia. J Climate, 18(6):852-863

Nan S L, Li J P. 2003. The relationship between the summer precipitation in the Yangtze River valley and the boreal spring Southern Hemisphere annular mode. Geophys Res Lett, 30(24):2266, doi:10.1029/2003 GL018381

Sun C, Li J P. 2012. Space-time spectral analysis of the Southern Hemisphere daily 500-hPa geopotential height. Mon Wea Rev, 140(12): 3844-3856, doi:10.1175/MWR-D-12-00019.1

Wu Z W, Li J P, Wang B, et al. 2009. Can the Southern Hemisphere annular mode affect China winter monsoon? J Geophys Res, 114:D11107, doi:10.1029/2008JD011501

Xiao D, Li J P. 2011. Mechanism of stratospheric decadal abrupt cooling in the early 1990s as influenced by the Pinatubo eruption. Chinese Sci Bull, 56(8): 772-780, doi:10.1007/s11434-010-4287-9

Yan Z W, Jones P D, Davies T D, et al. 2002. Trends of extreme temperatures in Europe and China based on daily observations. Climatic Change, 53(1-3): 355-392