DINGJING SHUIWEN SHUIZIYUAN WENJI

丁晶水文水资源文集

丁晶 著

四川出版集团·四川科学技术出版社

丁晶水文水资源文集

《丁晶水文水资源文集》编委会

四川出版集团·四川科学技术出版社

图书在版编目(CIP)数据

丁晶水文水资源文集/丁晶著. -成都:四川科学技术出版社,2006.1

ISBN 7-5364-5859-2

Ⅰ.丁... Ⅱ.丁... Ⅲ.水文学-文集 Ⅳ.P33-53

中国版本图书馆 CIP 数据核字(2005)第 126153 号

丁晶水文水资源文集

编 著 者　丁　晶 等
责任编辑　郑　尧　陈敦和
封面设计　霍运熙
责任校对　尧汝英
责任出版　邓一羽
出版发行　四川出版集团·四川科学技术出版社
成都盐道街 3 号　邮政编码 610012
成品尺寸　280mm×203mm
印张 16　字数 400 千　插页 2
印　　刷　四川五洲彩印有限责任公司
版　　次　2006 年 1 月成都第一版
印　　次　2006 年 1 月成都第一次印刷
印　　数　1-550 册
定　　价　60.00 元
ISBN 7-5364-5859-2/TV·25

《丁晶水文水资源文集》

编委会

主　任　袁　鹏

副主任　王文圣

委　员　（按姓氏笔画排序）

王文圣　付　军　刘国东　朱　兵　李贤彬
张学成　张　翔　安雪松　赵永龙　赵太想
周激流　姚　建　金菊良　侯　玉　袁　鹏
黄伟军　陈　曜　缪　韧

编辑工作小组

王文圣　张欣莉　覃光华
邓红霞　李眉眉　袁　鹏

前　言

丁晶先生是我国知名的水文水资源学者，从事水利工作近50年，硕果累累，为中国水文水资源事业的发展作出了较大的贡献，在国内外水文水资源界具有一定的知名度。

丁晶先生生于1935年10月，江苏泰兴人。1958年毕业于河海大学陆地水文专业，毕业后在成都工学院、成都科技大学、四川联合大学、四川大学任教，其中1981年至1983年在美国科罗拉多州大学作高级访问学者。长期从事水文水资源领域的教学和科研。曾任成都科技大学水利系副主任、水文水资源研究室主任，四川联合大学第一届学位评定委员会委员。1989年被国务院学位委员会批准为"水文学及水资源"学科博士生导师。从1992年起享受国务院政府特殊津贴。现为国际水文计划中国国家委员会委员，全国高校水利学科教学委员会副主任，四川省首批学术和技术带头人，《水科学进展》编委，《四川大学学报》(工程科学版)编委。

丁晶先生1990年开始招收博士研究生，迄今已培养了硕士生13位、博士生21位、博士后3位，并培养了大量的本科生，可谓桃李满天下，他的大部分学生现正奋战在全国各地水利领域。

丁晶先生在科学研究上取得了优异成绩，获得省部级科技进步奖7项，合作出版著作5部，先后参编教科书6部，编著的《随机水文学》得到了国内水文水资源界的赞誉，合编的《水文分析计算》仍是目前水文水资源专业的重要教材。在国内外发表学术论文达200篇之多(包括各种排名)，被EI收集的论文就有50多篇。

今值丁先生70岁华诞，为反映丁先生的学术观点和学术思想，为弘扬丁先生勇于水文水资源实践和为水利事业的献身精神，特从200多篇学术论文中选集了有代表性的40篇汇编成册出版。

祝愿丁先生学术又一春！

祝愿丁先生健康长寿！

《丁晶水文水资源文集》编委会

2005年7月于四川大学

内容简介

四川省学术带头人、四川大学博士生导师丁晶教授从事水文水资源领域的教学、科研和设计近50年，在国内外发表了大量的论文，现从中精选40篇有代表性的论文，汇编成文集出版。文集集中反映了丁晶教授的基本学术思想和主要创新成果。该文集可供水文水资源领域广大的教学科研人员、工程技术人员和研究生等参考，也可供环境、地理、地质、气象和生态等学科领域的有关人员参考。

目 录

论水文计算中的假相关

1 引言

相关分析广泛地应用于水文计算,但是,只有正确使用这种分析技术才能得出有意义的结果。所谓“正确”,涉及的面十分广泛,本文仅就避免出现假相关这一点加以论述。

假相关一般指两个变量之间表现出的一种虚假的相关关系,而实际并不存在统计关系。早在1897年,皮尔逊[1]就对假相关进行过研究,但是直到1965年,才由本森[2]明确指出水文和水力学中的假相关现象。随后,叶菲耶维奇[3]在其《水文概率和统计》一书中作了比较系统的论述。1977年赫安[4]在《水文学的统计方法》一书中进一步论述了这一课题。我国金光炎工程师[5]在1981年也发表文章指出水文中的假相关现象。最近,肯奈[6]进一步分析了这一现象,并警告要当心自身假相关。尽管对假相关的研究已有很长时间,但迄今对这一问题的认识还有些模糊不清,特别是在我国水文计算中尚未引起足够的重视,误用假相关的现象仍然存在。因此,系统地论述这一课题,澄清一些模糊概念,指出水文中可能出现的各种假相关现象,找出避免的方法,无疑是非常必要的。

为叙述和分析方便,本文尝试将通常遇到的假相关分为隐假相关和显假相关。前者系指间接引起的假相关;后者系指明显的假相关。

2 隐假相关

隐假相关的含义和一般所说的自身假相关相同,是由于包含相同的变量造成的假相关。

设有两个变量 x 和 y,相互间完全独立。现构成两个新变量 A 和 B。

第一种情况:$A=x+y,B=x$;第二种情况:$A=x\cdot y,B=x$;第三种情况:$A=y/x,B=x$。

以后将会看到:尽管 x,y 之间相互完全独立,即 $r_{x,y}=0$,但是,A 和 B 却显出一定的相关性,即 $r_{A,B}\neq 0$。这种 A 和 B 间出现的相关关系常叫作假相关。实际上,这种提法并不严格,就 A 和 B 两个变量而言,$r_{A.B}$ 确是度量两者之间线性相互关系的客观指标,并非假的。之所谓“假”,仅仅是相对于 x,y 两变量的关系而言。下面结合水文的实例分别讨论这三种情况。

2.1 第一种情况($A=x+y,B=x$ 型)

对于变量 A 和 B,其表征两者线性相关关系的定量指标——相关系数为[6]:

$$r_{A,B}=\frac{n\sum A\cdot B-\sum A\sum B}{\left[n\sum A^{2}-(\sum A)^{2}\right]^{1/2}\left[n\sum B^{2}-(\sum B)^{2}\right]^{1/2}} \tag{1}$$

将 $A=x+y,B=x$ 代入(1)得:

$$r_{A,B}=\frac{1+r_{x,y}\left(\frac{S_y}{S_x}\right)}{\left[1+\left(\frac{S_y}{S_x}\right)^{2}+2r_{x,y}\left(\frac{S_y}{S_x}\right)\right]^{1/2}} \tag{2}$$

式中，$r_{x,y}$——x 和 y 间的相关系数；S_x——x 的标准差；S_y——y 的标准差。

若 $r_{x,y}=0$

则

$$r_{A,B}=\frac{1}{\left[1+\left(\frac{S_y}{S_x}\right)^2\right]^{1/2}} \tag{3}$$

式(3)可写成另一种形式

$$r_{A,B}=\frac{1}{\left[1+\left(\frac{C_y\bar{y}}{C_x\bar{x}}\right)^2\right]^{1/2}} \tag{4}$$

式中，C_y，$\bar{y}$——y 的变差系数和平均值；C_x、$\bar{x}$——x 的变差系数和平均值。

如果变量 x 和 y 的均值和变差系数相等，那么，即使 $r_{x,y}=0$，新变量 A 和 B 的相关系数由式(4)计算得 $r_{A,B}=0.707$。这种现象便看作是假相关，其假的程度与原始变量 x 和 y 的统计特性紧密相关。图 1 给出了 $r_{A,B}$随和 C_y/C_x变化的情况。显然，在一定的情况下，C_y/C_x愈小，$r_{A,B}$愈大。其原因在于 A 和 B 的变化主要由 x 的变化所制约。

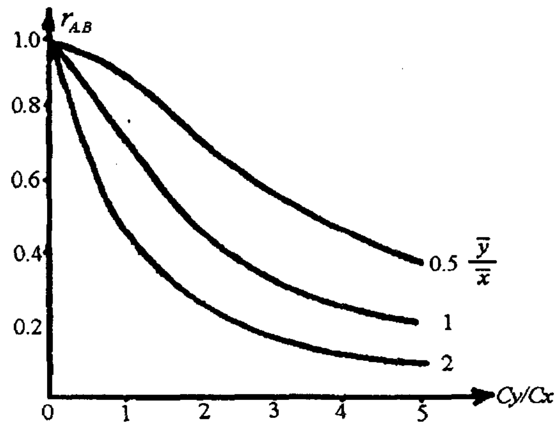

图 1 $r_{A,B}\sim\bar{y}/\bar{x}\sim C_y/C_x$ 图

由式(4)知，$r_{A,B}$取决于原始变量 x 和 y 的一阶矩和二阶矩的统计特性，x 和 y 的三阶矩特性（偏态系数 Cs）是否影响 $r_{A,B}$呢？即当原始变量呈偏态时，是否仍可用式(4)计算？为了回答这一问题，作了一系列统计试验，其结果列于表 1。对比由式(4)和模拟系列计算的数值，考虑到模拟序列的抽样误差，可以认为两者无显著的差异，这就是说式(4)适用于偏态系列。

A 和 B 之所以相关，其根本原因是由于两者均包含有相同的变量 x。A 和 B 之间的关系是客观的，但是利用 A 和 B 的这种相关信息，由 B（即 x）估计出 A，然后由 A 减去 x 而估计出 y，这就意味着 x 中包含了 y 的信息，即认为 x 和 y 之间有关系，实际上，两者之间完全无关。因此所谓“假”，就是由于 x 估计出与之完全无关的的 y。

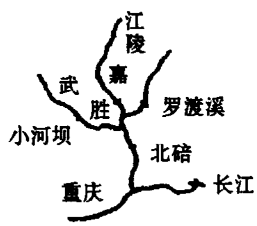

图 2 测站示意图

水文计算中出现这种情况的例子屡见不鲜，现以北碚站年经流量为例说明。如图 2 所示，x 为上游三站年径流量之和，y 为区间径流。A 为北碚站的年经流量，$A=x+y$，$B=x$。由实测的 28 年资料得 $r_{A,B}=0.99$，表明 A 和 B 之间确有关系。若直接计算和的相关系数，则 $r_{x,y}$仅为 0.32，关系非常微弱，由 x 难以估计 y。如通过 $r_{A,B}$来估计 y，显然其结果只是假象。

表1 统计试验成果表

总体参数值							$A=x+y,B=x$ 型		$A=x\cdot y,B=x$ 型		$A=y/x,B=x$ 型	
$\bar{x}$	$\bar{y}$	C_x	C_y	Cs_x/C_x	Cs_y/C_y	$r_{x,y}$	$r_{A,B}$ 模拟	$r_{A,B}$ 式(4)	$r_{A,B}$ 模拟	$r_{A,B}$ 式(6)	$r_{A,B}$ 模拟	$r_{A,B}$ 式(8)
10	10	0.48	0.48	4	4	0	0.694	0.707	0.674	0.707	-0.535	-0.707
10	10	0.35	0.80	4	4	0	0.438	0.401	0.413	0.401	-0.336	-0.401
10	10	0.80	0.35	4	4	0	0.896	0.916	0.872	0.916	-0.626	-0.916
10	1 000	0.48	0.35	4	4	0	0.002	0.01	0.674	0.707	-0.535	-0.707
1 000	10	0.48	0.48	4	4	0	1.00	1.00	0.674	0.707	-0.535	-0.707
10	10	0.48	0.48	2	4	0	0.727	0.707	0.694	0.707	-0.610	-0.707
10	10	0.48	0.48	4	2	0	0.707	0.707	0.667	0.707	-0.521	-0.707
10	10	0.48	0.48	2	2	0	0.696	0.707	0.694	0.707	-0.604	-0.707
1 000	1 000	0.48	0.48	2	2	0	0.685	0.707	0.694	0.707	-0.604	-0.707
1 000	1 000	0.35	0.35	2	2	0	0.698	0.707	0.706	0.707	-0.651	-0.707
10	1 000	0.01	0.90	4	4	0	0.018	0.001				
1 000	10	0.90	0.01	4	4	0	1.00	1.00				

注:(1)模拟系列相关系数是由模拟系列直接按相关系数的定义公式计算而得。
(2)模拟系列的随机模拟方法采用一般的舍选法。

2.2 第二种情况($A=x\cdot y\quad B=x$ 型)

对于变量 A 和 B 的相关系数 $r_{A,B}$,在原始变量和的变差系数很小时,可以下式计算

$$r_{A,B}=\frac{r_{x,y}C_y+C_x}{[C_y^2+C_x^2+2r_{x,y}C_xC_y]^{1/2}} \tag{5}$$

式中,符号意义同前。若 $r_{x,y}=0$。

则

$$r_{A,B}=\frac{1}{[1+(C_y/C_x)^2]^{1/2}} \tag{6}$$

由式(6)知,在 $C_x=C_y$时,$r_{A,B}=0.707$。这表明,即使 x 和 y 毫无关系,A 和 B 之间的线性相关关系也达到0.707,和第一种情况一样,这也是由于包含共同变量而引起的假相关。

式(6)是近似公式,在 C_x和 C_y较小时适用性较好,那么当 C_x和 C_y较大时,近似程度如何?另外在原始变量 x 和 y 具有偏态特性时,该公式的适用性又如何?对这些问题,通过统计试验作了研究,其结果如表1所示。总的来说,式(6)误差在5%以下,从适用的角度来看可以接受。

第二种情况的假相关在水文计算中可以找到很多例子。例如,为了展延含沙量(y),先建立流量(x)和输沙率(A)的关系,然后由流量展延输沙率,最后由输沙率转换成含沙量。这是典型的 $A=x\cdot y,B=x$ 型的假相关。就流量和输沙率而言,两者之间的关系并非假的。黄河流域某站有24年的观测资料,年径流量和输沙率之间的相关系数为0.80,而年径流量和含沙量之间的相关系数则为 -0.10。这清楚地表明,含沙量绝对不可能由年径流量来展延。但是,用上面所说的方法,通过展延得到的输沙率转换成含沙量,其展延成果无疑是虚假的。

2.3 第三种情况($A=y/x,B=x$ 型)

这种情况下 $r_{A,B}$计算式,在原始变量 x 和 y 变差很小时,可表示为[6]

$$r_{A,B}=\frac{r_{x,y}C_y-C_x}{[C_y^2+C_x^2-2r_{x,y}C_xC_y]^{1/2}} \tag{7}$$

若 $r_{x,y}=0$，则
$$r_{A,B}=\frac{-1}{[1+(C_y/C_x)^2]^{1/2}} \tag{8}$$

由式(8)知，当 $C_x=C_y$ 时，$r_{A,B}=-0.707$。

式(8)和式(6)一样是近似的，统计试验的结果(见表1)表明，在一般情况下，误差较大，随变差系数的减小而减小，而随偏态系数的加大而加大，实际应用时，应予以注意。

属于 $A=y/x, B=x$ 型的假相关在水文计算中也时常碰到。例如，设降雨量为 y，降雨历时为 x，则降雨强度 $A=y/x$ 和 x 降雨历时 $B=x$ 的关系便为这种类型的假相关。在 y 和 x 确是无关的情况下，若企图通过 A 和 B 的关系由历时 x 推估降雨量 y，则结果很不可靠。又如设径流量为 y，流域面积为 x，则径流模数 $A=y/x$ 和流域面积 $B=x$ 的关系也隶属于这一类型的假相关。只要 x 和 y 之间无关系，通过 A 和 B，间接由 x 估计 y 便是假相关的结果。

上面就隐假相关分别讨论了三种情况。还有很多其它情况，例如 $A=x-y, B=x$；$A=y\pm z, B=x\pm z$；$A=yz, B=xz$；$A=z/y, B=z/x$ 等类型。这些类型在水文计算中不很常见，若工作中遇此情况，可按类似上述的方法进行分析。为节省篇幅，在此不再赘述。

本小节讨论的隐假相关，主要涉及这样一类假相关：直接关系(表面上)是真实的，而间接关系(隐藏着)是虚假的。下小节讨论比较明显的假相关——显假相关。

3 显假相关

属于这种类型的假相关，可以举出下述三种情况：

3.1 点簇假相关

图3为点簇假相关最典型的情况。尽管变量 x 和 y 没有关系，但计算出的相关系数却相当大，显然这是假的。设一对无关变量为 x_1 和 y_1，另一对无关变量为 x_2 和 y_2，如该两组变量的均值数量级相差很大，则往往引起点簇相关。

为了阐述这一点，进行了统计试验研究，结果列于表2。从该表可以明显看出，影响点簇假相关的主要因素是两组变量均值的差异。正是由于均值的显著差异性，才会出现分开的两组散乱点群，从而导致计算出的相关系数较大，造成假相关现象。

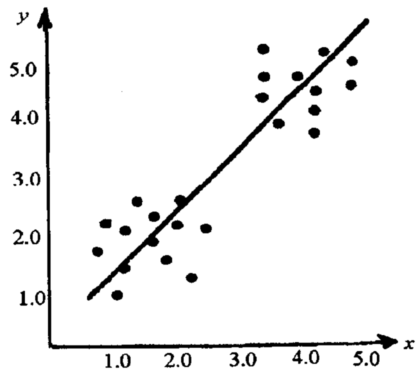

图3 点簇假相关示意图

表2 点簇假相关统计试验成果表

总体参数										统计试验参数		
$\bar{x}_1$	$\bar{y}_1$	$\bar{x}_2$	$\bar{y}_2$	C_x	C_y	Cs_x/C_x	Cs_y/C_y	r_1	R_2	$\hat{r}_1$	$\hat{r}_2$	r_0
10	10	1 000	1 000	0.48	0.48	2	4	0.0	0.0	-0.006	-0.026	0.68
10	10	1 000	1 000	0.48	0.48	4	2	0.0	0.0	0.047	0.011	0.69
10	10	1 000	1 000	0.35	0.80	4	4	0.0	0.0	0.013	0.005	0.63
10	10	1 000	1 000	0.80	0.35	4	4	0.0	0.0	0.011	0.016	0.63
1	1	1 000	1 000	0.48	0.48	4	4	0.0	0.0	0.048	-0.054	0.69
100	100	1 000	1 000	0.48	0.48	4	4	0.0	0.0	0.048	-0.054	0.64

注：(1)对于变量(x_1,y_1)和(x_2,y_2)均分别模拟容量为40的样本50组，就每组计算相关系数，然后以50组平均得和 $\hat{r}_2$；(2)合并每两组系列计算相关系数，然后取平均得 r_0。

3.2 对数变换假相关

对原始变量 x 和 y 作对数变换，是否造成假相关，不得一概而论，必须针对具体情况加以分析。现分以下几种情况论述

(1)若原始变量 x、y 严格服从对数正态分布，则其相关系数 $\rho_{x,y}$ 和 $\rho_{\ln x,\ln y}$ 之间的理论关系为[7]

$$\rho_{\ln x,\ln y}=\frac{1}{\sigma_{\ln x}\sigma_{\ln y}}\ln\left[\sqrt{(e^{\sigma_{\ln x}^2}-1)(e^{\sigma_{\ln x}^2})\rho_{xy}}+1\right] \tag{9}$$

式中，$\sigma_{\ln x}$，$\sigma_{\ln y}$ 分别为 $\ln x$，$\ln y$ 的标准差。计算表明，$\rho_{\ln x,\ln y}$ 略大于 $\rho_{x,y}$。这就是说，在原始变量取对数后，其相关系数仅少许提高。如原始变量为微弱相关或不相关，通过对数变换一般不会显著提高相关系数，即不会造成明显的假相关。

表3　对数转换部分统计试验成果

总体参数							样本容量	$\hat{\rho}_{x,y}$	$\hat{\rho}_{\ln x,\ln y}$
$\bar{x}$	$\bar{y}$	C_x	C_y	Cs_x/C_x	Cs_y/C_y	$\rho_{x,y}$	n		
								0.18	0.16
10	10	0.48	0.48	4	4	0	40	0.29	0.15
								0.30	0.31
10	10	0.38	0.80	4	4	0	40	0.31	0.37
								0.39	0.30
10	10	0.80	0.35	4	4	0	40	0.20	0.12
								0.15	0.23
10	1 000	0.48	0.48	4	4	0	40	0.15	0.22
								0.18	0.15
1 000	10	0.48	0.48	4	4	0	40	0.23	0.32
								0.29	0.23
1 000	1 000	0.48	0.48	2	2	0	40	0.25	0.18
								0.19	0.23
10	10	0.48	0.48	2	4	0	40	0.32	0.31
								0.18	0.20
10	10	0.48	0.48	4	2	0	40	0.33	0.35
								0.11	0.16
10	10	0.48	0.48	2	2	0	40	0.25	0.17
								0.19	0.23
1 000	1 000	0.35	0.35	2	2	0	40	0.28	0.23
								0.20	0.23
10	10	0.90	0.30	4	4	0	40	0.55	0.45
								0.35	0.41
10	10	0.90	0.10	4	4	0	40	0.20	0.18
								0.30	0.35

(2)若原始变量 x，y 互不相关且服从皮尔逊Ⅲ型分布，从这样的总体随机抽取样本，分别计算 $\hat{\rho}_{x,y}$ 和 $\hat{\rho}_{\ln x,\ln y}$。统计试验表明，$\hat{\rho}_{\ln x,\ln y}$ 和 $\hat{\rho}_{x,y}$ 相差不大，且 $\hat{\rho}_{\ln x,\ln y}$ 不一定都大于 $\hat{\rho}_{x,y}$。表3是从大量统计试验中摘出的部分结果，该表清楚地表明，在所讨论的条件下，不会造成明显的假相关。

当然，由于对原始变量取对数，比例尺缩小(数量级减少)，图上点据分散的程度看起来有

所减小，给人一种假象，好像相关关系提高了。这单纯是肉眼上的错觉，若定量计算，变换前后的相关系数相差无几。

(3)若原始变量 x,y 之间呈现出某种曲线相关关系，计算而得 $\hat{\rho}_{x,y}$ 可能较小。此时对原始变量作对数变换，计算而得的 $\hat{\rho}_{\ln x,\ln y}$ 可能相当大。对于这种情况，在我们看来不宜叫作假相关，因为原始变量之间确有关系，只不过这种关系并非线性而已。

综上所述，在一般情况下，对数变换不会导致明显的假相关。若变换后，相关系数大大提高，则原始变量可能存在着曲线相关关系。

水文计算工作中常常对变量取对数，这是无可非议的。要注意的是：若原始变量 x,y 的关系相当差（点据分散），通过对数变换缩小比例尺使点据看起来比较密切，并以肉眼定出相关线，便会造成假相关。实际工作中应杜绝这种作法。

3.3 辗转假相关

两变量 x,y 之间没有什么关系，但分别通过 x 和 y 同第三变量 z 相关，可能获得 x 和 y 之间的辗转关系。这似乎表明 x 中包含着 y 的信息，显然，这是假的。为了定量说明这种假相关的特性，我们作了辗转三次的统计试验研究，结果如表 4 所示。由该表清楚地看出：尽管 x 和 z，z 和 w，w 和 y 之间的相关系数均较高，而实际上 x 和 y 之间的相关关系却相当低。因此，就 x 和 y 两变量而言，这种辗转相关是假的。

辗转假相关的例子在水文计算中还是能见到的。例如，为了利用洪峰流量展延长历时的洪量，就可能误用辗转相关。松花江丰满站的资料表明，洪峰和历时十一日的洪量之间的相关系数仅为 0.59。若由资料建立洪峰流量和三日洪量关系，三日洪量和七日洪量关系，进而七日洪量和十一日洪量关系，这些关系均较密切，通过它们由洪峰流量推求出的十一日洪量，就是典型辗转相关的结果。

表 4　辗转假相关统计试验成果表

总体参数						r_{xz}	r_{zw}	r_{wy}	r_{xy}
E_x	E_w	E_y	C_x	C_w	C_y	辗转相关系数值			相关系数
100	100	100	0.20	0.20	0.20	0.8	0.8	0.8	0.35
100	100	100	0.20	0.20	0.20	0.9	0.9	0.9	0.43
100	100	100	0.20	0.20	0.20	0.9	0.9	0.9	0.43
100	100	100	0.25	0.25	0.25	0.9	0.9	0.9	0.43
100	100	100	0.25	0.25	0.25	0.7	0.7	0.7	0.27

注：$r_{x,y}$为容量是 40 的 50 个样本的平均值

4 小结

假相关是相关分析中出现的一种现象。它并非相关技术本身的问题，而是误用这种技术所引起的。若对两原始变量作变换或构成另一对新变量，则新变量之间的相关系数只能表明新变量本身之间的线性关系。在一般情况下，不能将新变量之间客观存在着的关系信息不加分析地用于原始变量。否则会导致假相关。

假相关可分为两大类，一是隐假相关，二是显假相关。这两类假相关在水文计算中都会遇到，在假相关情况下获得的计算结果，显然是不可靠的，所以无论何时都应避免。

为了避免误用假相关，最关键的问题在于直接研究原始变量，就是说，不作任何数学转换和运算。

参考文献

1 Pearson K . On a form of spurious correlation which may arise when indices are used in the measurement of organs. Proc R Soc London, 1987, 60.

2 Benson M A. Spurious correlation in hydraulics and hydrology. J Hydrau Div Am Sos Eng, 1965.91

3 Yevjevich V. Probablity and Statistics in Hydrology Water Resources Publication. Fort Collins Co. U . S . A . 1972

4 Hean CT, Statistical Method in Hydrology. The lowa State University Press , 1977

5 金光炎,用统计方法计算设计洪水的若干问题,水文,1981(6)

6 Kenney B C . Beware of spurious self - correlation! Water Resources Research , 1982, 18(4)

7 Stedinger J R . Estimating correlation in Multivariate stream model. Water Resources Research , 1981, 17 (1)

载于《四川水力发电》,1987 年第 4 期

洪水特大值的检测

洪水系列中的特大值和特小值称为特异值。它们对洪水频率计算的影响非常大，尤以特大值为然，因此对特异值的处理至关重要。所谓处理，一般指从系列中检出特异值以及正确的估计出重现期。本文在于讨论当观测洪水系列出自皮尔逊Ⅲ型分布总体时，如何应用统计检验法，检测其中的特大值。当然，处理特异值的各种方法最终需要同时从数学和水文的角度考虑判断[1]，但本文偏重于数学方面。直观上来看，特大值偏离其它的观测数据异常显著，以致被怀疑为来自不同的统计机制[2]。当样本出自两种分布时，其中之一称为基本分布，产生“基本一致”观测值；而另一种分布称为污染分布，产生污染值。污染值统称为特异值，而偏离在上端的，本文称为特大值。基本分布和污染分布可能是两种不同的分布，也可能是属于同一种分布但参数不同（参数平移）[4]。本文仅讨论后一种情况。

1　检验方法的基本原理

特大值的检验方法，随总体分布的不同而异。为了检验总体为皮尔逊Ⅲ型分布时的特大值，我们引用了下述两种方法。

第一种方法是建立在总体χ^2分布的基础上。它是用考克因（Cochren）提出的统计量，故叫作考克因检验。该法的基本原理简述于下：系列$x_1,x_2,\cdots,x_n$，相互独立且服从自由度为v的$\sigma_i^2\chi_v^2$分布$(i=1,2,\cdots,n)$，令$x_{(n)}=\max\{x_1,x_2,\cdots,x_n\}$，我们有假设：

$$H_0:\ \sigma_i=1,(i=1,2,\cdots,n)$$

构造统计量
$$C_n=\frac{x_{(n)}}{W} \tag{1}$$

式中，$W=\sum\limits_{n=1}^{n}x_i$

若$C_n>C_\alpha$则拒绝假设H_0，即推断$x_{(n)}$为特大值，反之则接受，C_α为显著性水平α下的临界值，是在原假设H_0条件下，由$C_n=x_{(n)}/W$的分布，$P\{C\geqslant C_\alpha\}=\alpha$而确定出来的。$C$的严格分布为已知，但相当复杂，近似的分位数值$C_\alpha$已列成表可直接查用[2]在一定的$\alpha$下，$C_\alpha$取决于样本容量$n$和自由度$v$。

第二种方法是建立在正态分布的基础上，其原理如下，系列$x_1,x_2,\cdots,x_n$相互独立且服从正态分布$N(\mu_i,\sigma^2)(i=1,2,\cdots,n)$，令$x_{(n)}=\max(x_1,x_2,x_3,\cdots,x_n\}$，我们有假设：

$$H_0:\ \mu_i=\mu(i=1,2,\cdots,n)$$

构造统计量
$$T_n=\frac{x_{(n)}-\bar{x}}{S} \tag{2}$$

式中，$\bar{x}=\frac{1}{n}\sum\limits_{i=1}^{n}x_i,S=\sqrt{\frac{\sum\limits_{i=1}^{n}(x_i-\bar{x})^2}{n-1}}$

若 $T_n > T_\alpha$,则拒绝原假设 H_0,即推断 $x_{(n)}$ 是特大值;反之则接受 H_0,即推断 $x_{(n)}$ 不是特大值。T_α 为显著性水平 α 下的临界值,是在原假设 H_0 的条件下,$T_n=(x_{(n)}-\bar{x})/S$ 由的分布,$P(T\geqslant T_\alpha)=\alpha$,而确定。$T_\alpha$ 在文献[3]中有专用表可供查用,但该文献难以获得,为方便使用,特摘录专用表中的常用部分列于附表。在一定的 α 下,T_α 仅取决于样本容量 n。

本文将第一种方法叫作方法一,而第二种方法叫方法二,关于它们的应用分别叙述于下。

2 方法一的应用

2.1 皮尔逊Ⅲ型分布转换为近似的 χ^2 分布

第一种方法适用于 χ^2 分布,为用于皮尔逊Ⅲ型分布,必须进行适当的转换。

皮尔逊Ⅲ型分布密度函数为:

$$f(x)=\frac{\beta^\alpha}{\Gamma(\alpha)}(x-a_0)^{\alpha-1}\mathrm{e}^{-\beta(x-a_0)} \tag{3}$$

式中:α,β,a_0 为参数

令 $y=x-a_0$

则得 $$f(x)=\frac{\beta^\alpha}{\Gamma(\alpha)}y^{\alpha-1}\mathrm{e}^{-\beta y}$$

再令 $z=2\beta y$

则得 $$f(x)=\frac{(0.5)^\alpha}{\Gamma(\alpha)}z^{\alpha-1}\mathrm{e}^{-\frac{1}{2}z}$$

若 $2\alpha=v$ 为整数,z 即为 χ_v^2 分布,一般 2α 不为整数,此时取最近整数值而得 v,在这种情况下,z 为近似的 χ_v^2 分布。方法一直接用于 z 系列检测其中的特大值。

2.2 参数估计

a_0,β 和 α 三参数由 $\bar{x}$,Cv 和 Cs 估算,计算公式如下:

$$a_0=\bar{x}(1-\frac{2Cv}{Cs}) \tag{4}$$

$$\beta=2\sqrt{x}\cdot Cv\cdot Cs \tag{5}$$

$$\alpha=4/Cs \tag{6}$$

由上可知,为了将这种方法用于皮尔逊Ⅲ型分布的系列,必须首先估计出该系列的三个参数。由于 Cs 的样本估计值很不稳定,我们将原系列 h_i 变换成新系列 x_i($\mathrm{i}=1,2,\cdots,n$)。使 x_i 具有 $Cs=2Cv$ 的统计特性,从而由 x_i 来确定 Cs。变换公式如下:

$$x_i=h_i-0.5h_{(1)} \tag{7}$$

式中,$h_{(1)}=\min\{h_1,h_2,\cdots,h_n\}$。

求得 x_i 后,即可以 $\bar{x}$,Cv 通过公式(5)和(6)算得 β、α,最后得 z 系列及其自由度值 v。

这是一种经验处理办法,其实质在于对原系列减去一个适当值,使对新系列而言,式(4)中的 a_0 等于零。这个适当值的确定是比较困难的,通过反复试验,本文暂选用 $0.5h_{(1)}$。

2.3 适用性的验证

(1)样本中无特大值的情况(非污染情况)

样本来自基本的皮尔逊Ⅲ型分布,样本未受污染,故其中无特大值。用方法一检测样本中的特大值,把犯第一类型错误的概率和显著性水平 α 的比较作为判断该方法适用性的依据之一。