改进的有序聚类分析法提取时间序列转折点

陈远中[1]，陆宝宏[2,3]，张育德[2]，周笑笑[2]

(1.深圳市水务工程建设管理中心，广东 深圳 518048；2.河海大学水文水资源学院，江苏 南京 210098；
3.水文水资源及水利工程国家重点实验室，江苏 南京 210098)

摘 要：对有序聚类分析法进行改进，使其更加适用于序列转折点或突变点的提取。分别将传统和改进的有序聚类方法应用在长江下游区域年平均气温系列的转折点提取中，对两种方法提取的结果与滑动平均的结果进行比较，发现改进的方法更接近于实际。对提取点分别采用秩和检验法、游程检验法进行检验，均通过了 $\alpha=0.05$ 的置信度检验，改进后的置信度比改进前更高。

关键词：有序类聚；转折点；时间序列；显著性检验

中图分类号：P333.9 **文献标识码**：A **文章编号**：1000-0852(2011)01-0041-04

引言

20 世纪 80 年代以来，随着人类活动的增强，对环境影响逐渐加大，水文过程也发生了相应改变，致使水文资料系列的一致性发生变化。在水工程规划、设计中，水文资料的三性审查是规划设计的基础。在水文资料系列一致性分析中，要检查是否有明显的跳跃点，是否有不合理的跳跃成分或趋势成分[1]。因此，正确提取水文系列的转折点或跳跃点对规划设计有很大影响。在诸多提取方法中，时序累计值相关法要求必须有呈相关关系的另一个无趋势变化时间序列[2]；Lee-Heghinian 分析法依据正态分布的假设[3]；有序聚类分析法在无趋势变化系列的跳跃点提取很方便，但对于有趋势变化时间序列的转折点提取就有很大的出入。在突变点的提取中还有很多的方法，但是，水文过程在适应不断改变的自然环境中产生的水文时间序列是一个有趋势的渐变过程，需要提取的是一个转折点而非突变点，很多方法都有一定的局限性。为此，对有序类聚分析做一定的改进，使其应用在有趋势成分时间序列的转折点的提取中。

1 原理说明

有序类聚分析法是一种统计的估计方法，通过统计分析推估出水文时间序列最可能的突变点，然后结合实际情况进行具体分析。其主要的分割思想是使得同类之间的离差平方和最小，而类与类之间的离差平方和最大[4]。

设可能的突变点为 τ，则突变前后的离差平方和分别为：

$$V_\tau=\sum_{i=1}^{\tau}(x_i-\bar{x}_\tau)^2 \tag{1}$$

$$V_{n-\tau}=\sum_{i=\tau+1}^{n}(x_i-\bar{x}_{n-\tau})^2 \tag{2}$$

式中：$\bar{x}_\tau$ 和 $\bar{x}_{n-\tau}$ 分别为 τ 前后两部分的均值。这样总离差的平方和为：

$$S(\tau)=V_\tau+V_{n-\tau} \tag{3}$$

那么当 $S=\min\{S_n(\tau)\}(2\le\tau\le n-1)$ 时，τ 为最优二分割，即推断为突变点。

这个方法最适应的情况是系统数据产生了系统跳跃而无趋势变化，且两段数据都是无趋势变化的时间序列，如图 1 所示。但是，对于有趋势变化的数据系列就不是很适用了，需要做相应的方法改进，如图 2 所示。

使用数据系列线性拟合的趋势线代替原方法的平均值，这样，在数据系列存在趋势变化的情况下也可以

收稿日期：2009-11-13
基金项目：国家自然科学基金项目(NSFC50379008，NSFC 50979023)
作者简介：陈远中(1968-)，男，江苏盐城人，硕士研究生，工程师，从事水资源水文工程管理。E-mail：cyz.JS@163.com
通讯作者：陆宝宏(1962-)，男，安徽天长人，副教授，研究方向为水资源规划及同位素水文学。E-mail：lubaohong@126.com

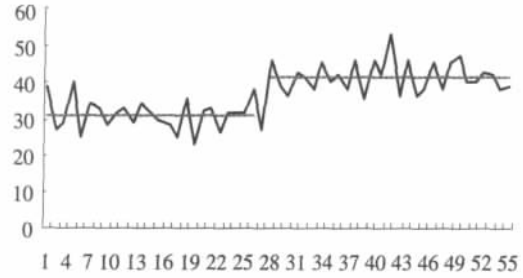

图 1 有突变的系列

Fig.1 The time series with change

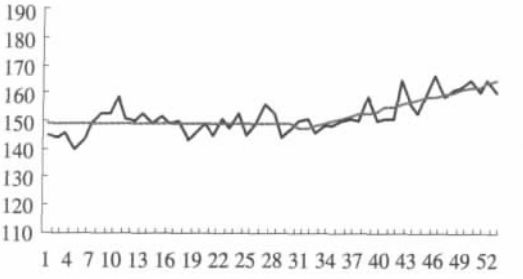

图 2 有趋势变化的系列

Fig.2 The time series with trend

使用。其意义在于,转折点前后两段数据到拟合曲线的距离平方和求和最小。

设拟合直线的方程为 $y=kx+b$,点到直线的距离如图 3 所示。

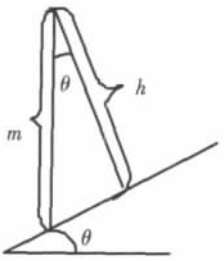

图 3 点到直线的距离

Fig.3 The distance between point and fitted line

由图 3 可知,$h=m/\sqrt{k^2+1}$

式中:k 为斜率;$m=x_i-y_i$。

转折点前后的离差平方和分别表示为:

$$VV_\tau=\frac{1}{k_\tau^2+1}\sum_{l=1}^{\tau}(x_l-y_{\tau,l})^2;VV_{n-\tau}=\frac{1}{k_{n-\tau}^2+1}\sum_{l=\tau+1}^{n}(x_l-y_{(n-\tau)l})^2$$

式中:$y_{nl}=k_\tau\times l+b_\tau$;$y_{(n-\tau)l}=k_{n-\tau}(l-\tau)+b_{n-\tau}$

$y_{nl}=k_\tau\times l+b_\tau$ 是转折前的拟合曲线,根据最小二乘法可得:

$$\begin{pmatrix}\tau & \frac{\tau(\tau+1)}{2}\\ \frac{\tau(\tau+1)}{2} & \sum_{i=1}^{\tau}i^2\end{pmatrix}\begin{pmatrix}b_\tau\\ k_\tau\end{pmatrix}=\begin{pmatrix}\sum_{i=1}^{\tau}x_i\\ \sum_{i=1}^{\tau}i\times x_i\end{pmatrix}$$

求解

$$k_l=\frac{2\sum_{i=1}^{\tau}i\times x_i-(\tau+1)\sum_{i=1}^{\tau}x_i}{2\sum_{i=1}^{\tau}i^2-\frac{\tau(\tau+1)^2}{2}}\qquad b_\tau=\frac{\sum_{i=1}^{\tau}x_i-\tau(\tau+1)k_\tau/2}{\tau}$$

同理:

$$k_{n-\tau}=\frac{2\sum_{i=\tau+1}^{n}(i-\tau)x_i-(n-\tau+1)\sum_{i=\tau+1}^{n}x_i}{2\sum_{i=\tau+1}^{n}(i-\tau)^2-\frac{(n-\tau)(n-\tau+1)^2}{2}}$$

$$b_{n-\tau}=\frac{\sum_{i=\tau+1}^{n}x_i-(n-\tau)(n-\tau+1)k_{n-\tau}/2}{n-t}$$

令

$$M(\tau)=VV_\tau+VV_{n-\tau}\qquad(4)$$

当 $M=\min\{M_n(\tau)\}(3\le\tau\le n-2)$ 时,τ 为最优二分割点,即推断为转折点。

下面证明改进后的方法在转折点求得的距离平方和更小更合理。

设实测数据点为 $y_i=f(x_i)\quad(i=0,1,\cdots m)$

使用函数 $y=S^*(x)$ 对其拟合,误差表示为 $\delta_i=S^*(x_i)-y_i$,$\delta=(\delta_0,\delta_1,\cdots\delta_m)^T$

设 $\varphi_0(x),\varphi_1(x),\cdots\varphi_n(x)$ 是 $C[a,b]$ 上线性无关的函数簇,在 $\varphi=span\{\varphi_0(x),\varphi_1(x),\cdots\varphi_n(x)\}$ 中找一函数 $S^*(x)$,使误差平方和[5]最小

$$\|\bar{\delta}\|_2^2=\sum_{i=0}^{m}\delta_i^2=\sum_{i=0}^{m}[S^*(x_i)-y_i]^2=\min_{S(x)\in\varphi}\sum_{i=0}^{m}[S(x_i)-y_i]^2\qquad(5)$$

式中:$S(x)=a_0\varphi_0(x)+a_1\varphi_1(x)+\cdots+a_n\varphi_n(x)\quad(n<m)$

有唯一解 $S^*(x)=a_0\varphi^*_0(x)+a_1\varphi^*_1(x)+\cdots+a_n\varphi^*_n(x)$

可以证明,对于任何形式的 $S(x)$ 都有

$$\sum_{i=0}^{m}[S^*(x_i)-f(x_i)]^2\le\sum_{i=0}^{m}[S(x_i)-f(x_i)]^2\qquad(6)$$

改进的方法中 $S(x)=a_0+a_1x=S^*(x)$,原方法中的 $\bar{x}$ 可表示为 $S(x)=b_0$,假设 x_k 是系列 x_i 的转折点,则由式(4)和式(6)可得:

$$M(k)=\min\{M_n(\tau)\}=(1+k^2_\tau)VV_\tau+(1+k^2_{n-\tau})VV_{n-\tau}<V_\tau+V_{n-\tau}=S(k)\qquad(7)$$

在转折点上 $M(k)<S(k)$,可认为改进后的方法距离平方和更小。而 $S(k)\ge\min\{S(\tau)\}$ 说明有序类聚的

分析方法求取的转折点可能不是实际的转折点,与实际情况有所偏差。综上所述改进后的方法优于改进前。

2　方法使用

对于长江流域下游区的气温变化如图 4 所示。由图可以很明显地看出温度系列的后面有着明显的增长趋势,所以数据系列应该分段讨论。用后一系列的数据来推求本时段内的温度变化率更接近实际情况。因此正确地划分两个系列对于增温率的推求影响很大。

下面分别用有序类聚分析法和改进后的方法推求气温系列的转折点,并使用滑动平均法处理数据,然后目估判断其转折点的范围,用此作为参照结果比较两种方法推求的结果的优劣。

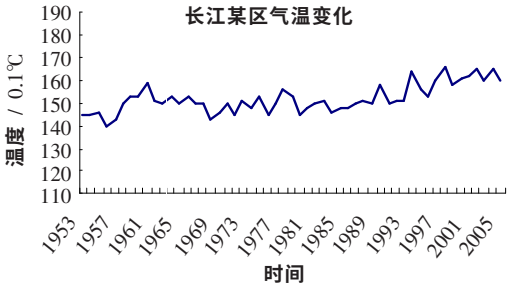

图 4　气温时间系列

Fig.4 The temperature time series

使用有序类聚分析法对数据系列进行处理,计算出其相应的 S 值,S 值的变化如图 5 求得其最小值为 S=802.69,其转折点发生的年份为 1993 年。

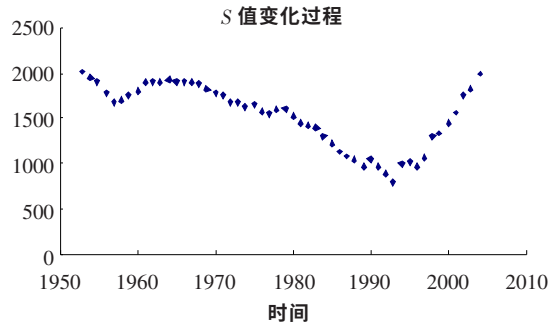

图 5　有序类聚分析的 S 值变化

Fig.5 The change of S value of sequential cluster

使用改进的方法对气温数据进行处理,其 M 值的变化如图 6,求得其最小的 M 值为 M=710.63,所求的转折点发生的年份为 1986 年。

对原始数列使用滑动年数 n=5 的滑动平均处理,结果见图 7。可见在 1984、1985、1986 年的滑动平均温度都为 14.86℃,1987 年后温度开始逐渐增大。可以认为转折点在 1985 年附近,由滑动平均的趋势也可以

图 6　改进后的方法的 M 值变化

Fig.6 The change of M value of improved method

图 7　滑动平均处理的温度系列变化

Fig.7 The change of temperature's moving average time series

看出 1986 年后有着明显的增长趋势。所以认为改进的方法比原方法提取的转折点更具有代表性、准确性。同时改进的方法 M 值在转折点也明显小于原始方法的 S 值也进一步表明了这一点。

3　显著性检验

对于跳跃点的显著性检验使用秩和检验法[6]和游程检验法:

秩和检验法的统计量为:$U=\dfrac{W-\frac{n_1(n_1+n_2+1)}{2}}{\sqrt{\frac{n_1n_2(n_1+n_2+1)}{12}}}$,$W$ 为较小容量样本的秩和,n_1 为较小样本的容量。U 服从正态分布,检验结果如表 1 所示。

表 1　秩和检验结果

Table 1 The results of rank tests

使用方法	W	U	α	$U_{\alpha/2}$	检验结果
有序类聚分析法	118	−4.82	0.05	1.96	$\|U\|>U_{\alpha/2}$ 跳跃点显著
改进后的方法	316	−4.9	0.05	1.96	$\|U\|>U_{\alpha/2}$ 跳跃点显著

游程检验法使用的统计量为:$U=\dfrac{k-(1+\frac{2n_1n_2}{n})}{\sqrt{\frac{2n_1n_2(2n_1n_2-n)}{n^2(n-1)}}}$,本文使用游程个数检验法,所以 k

为游程的个数，n_1，n_2 分别为两个系列的样本容量，U 服从正态分布，检验结果如表 2 所示。

表 2 游程检验结果

Table 2 The results of run-length tests

使用方法	k	U	α	$U_{\alpha/2}$	检验结果
有序类聚分析法	10	-4.01	0.05	1.96	$\|U\|>U_{\alpha/2}$ 跳跃点显著
改进后的方法	10	-4.7	0.05	1.96	$\|U\|>U_{\alpha/2}$ 跳跃点显著

由上面检验结果可见，两种方法求得的转折点经过检验都是显著的跳跃点，但是明显地改进后的 $U_{后}>U_{前}$。说明改进后的方法的置信度比改进前的有所提高。

4 结语

对于改进后的方法在转折点的提取上结果优于原方法，并且置信度相对更大，不过复杂度和计算量增加。基于当前计算机技术的发展水平，计算复杂度的增加并没有带来本质上的计算麻烦。计算程序与原方法的程序并没有很大的差别，所以总的来讲改进的方法还是可取的。两种方法的比较见表 3。

本文对有序类聚分析方法的做了一定的改进，同时使用在长江某流域的气温变化的转折点提取中，结果表明改进后的方法提取的结果比原方法更接近实际并且置信区间度比原来更高。所以此方法可以广泛用在数据系列存在趋势变化的转折点或跳跃点提取中。并且改进前的方法是的改进后的方法在 $\theta=0$ 时的一种特殊情况。改进后的方法包含了原方法的情况，是使用范围在原基础上的扩大。

表 3 两方法的优缺点比较

Table 3 Comparison of the advantages and disadvantages between two methods

	原方法	改进后
优点	计算比较简单，原理易于理解	精度比较高，更符合数据的实际转折情况
缺点	实用范围比较窄，容易出现较大偏差	计算的复杂度增加

参考文献：

[1] 刘攀，郭生练，肖义，等. 水文时间序列趋势和跳跃分析的再抽样方法研究[J].水文，2007，27(2):49-53. (LIU Pan, GUO Shenglian, XIAO Yi, et al. Trend and change analysis of hydrological time series on resample methods [J]. Journal of China Hydrology, 2007, 27(2):49-53. (in Chinese))

[2] 史卫东. 水文资料系列的一致性分析[J].甘肃水利水电技术，2001,37 (1): 22-25, 42.(SHI Weidong. Hydrology data series consistence analysis [J]. Gansu Water Resource Technology, 2001,37 (1):22 -25, 42. (in Chinese))

[3] 汪丽娜，陈晓宏，李粤安，等. 水文时间序列突变点分析的启发式分割方法[J].人民长江，2009，40(9):15-17 (WANG Lina, CHEN Xiaohong, LI Aoan, et al. Heuristic segmentation method for change-point analysis of hydrological time series [J]. Yangtze River, 2009, 40 (9):15-17. (in Chinese))

[4] 王文圣，金菊良，李跃清，等. 水文水资源随机模拟技术[M].成都：四川大学出版社，2003. (WANG Wensheng, JIN Juliang, LI Yueqing, et al. Hydrology and Water Resources Stochastic Simulation Technique [M]. Chengdu: Sichuan University Press, 2003.(in Chinese))

[5] 李庆扬，王能超，易大义. 数值分析[M].北京：清华大学出版社，2001. (LI Qingyang, WANG Nengchao, YI Dayi. Numerical Analysis [M]. Beijing: Tsinghua University Press,2001. (in Chinese))

[6] 杨喜寿，杨洪昌. 气候事件序列变点的推断[J].大气科学，1996，20(1): 47-53. (YANG Xishou, YANG Hongchang. Inference on change point in a climate time series [J]. Scientia Atmospherica Sinica, 1996, 20(1):47-53. (in Chinese))

Improvement of Sequential Cluster Analysis Method for Extracting Turning Point of Time Series

CHEN Yuanzhong[1], LU Baohong[2,3], ZHANG Yude[2], ZHOU Xiaoxiao[2]

(1. *Shenzhen Water Engineering Construction Management Center, Shenzhen* 518048, *China*;
2.*College of Hydrology and Water Resources, Hohai University, Nanjing* 210098,*China* ;
3. *State Key Laboratory of Hydrology-Water Resources and Hydraulic Engineering, Nanjing* 210098, *China*)

Abstract: Because the sequential cluster analysis method can not deal with turning point with change trend. In order to overcome the shortcomings and obtain more accurate turning points for a time series with changed trend, the conventional sequential cluster analysis approach was improved. The modified method was used to extract the turning point of annual average temperature time series of a sub-basin in Yangtze River Basin. By comparing the performance between the modified and conventional approaches, it can be found that the improved method is more close to practice. Furthermore, the obtained tuning points from the two method all passed the confidence test using rank tests and run-length tests withα=0.05, but confidence degree from the improved method is higher than the conventional approach.

Key words: sequential cluster; turning point; time series; confidence test