

基于启发式分割算法的水文变异分析研究*

陈广才, 谢 平

(武汉大学水资源与水电工程科学国家重点实验室, 湖北 武汉 430072)

摘 要: 通过构建理想样本探讨了启发式分割算法在水文变异诊断中的可行性, 比较了其与传统变异检验方法在检验性能上的差异, 同时结合珠江三角洲河网三水站 1900 - 2004 年年最大洪峰流量序列检验其实际应用效果。研究表明该方法能较好排除虚假变异点干扰, 较准确识别变异点数目及其发生时间, 检验性能优于传统方法, 适用于水文变异分析, 特别是对复杂非线性非平稳水文序列, 实际应用效果较好。通过变异分析发现三水站年最大洪峰流量序列在 1993 年发生跳跃变异, 且为均值向上跳跃, 表明区域防洪形势将更加趋于严峻。

关键词: 水文变异; 启发式分割算法; 时间序列; 三水站; 年最大洪峰流量序列

中图分类号: P333.9 **文献标识码:** A **文章编号:** 0529-6579 (2008) 05-0122-04

气候变化和人类活动会导致区域水文要素发生变异并出现新的特征, 认识这一特征有助于掌握水文演变规律。水文变异通常可以采用统计检验^[1]的方法加以识别和判断, 如滑动 T 检验法、滑动 F 检验法、游程检验、秩和检验等, 这些方法多基于线性和平稳检验假定。但由于水文过程的非线性现象比较普遍, 水文序列多呈现出非线性; 同时由于气候变化和剧烈人类活动影响, 水文序列统计特性往往随时间发生变化, 序列呈现非平稳性。传统统计检验方法在检验这一类非线性非平稳水文序列时往往发生偏差(如可能检验到许多虚假变异点等)。2001 年 Bernaola - Galván^[2] 在分析人类心率非平稳数据中的尺度不变性时, 提出了启发式分割算法(又称 BG 算法), 用于处理非线性非平稳序列的均值变异点。该方法基于滑动 t 检验的思想, 并对其作了相应的改变, 不仅能有效地检测得到原信号中存在的均值变异点, 而且大大增强了其处理非线性、非平稳信号的能力。2004 年 Fukuda^[3] 通过具有不同统计特性的理想样本, 进一步分析和验证了该方法。与传统的分析方法相比, 启发式分割算法能将非平稳序列分割为多个具有不同均值的平稳子序列(均值段), 各子序列表征了不同的物理背景, 分解得到的各均值段的尺度具有可变性, 不受方法本身的限制, 解决了以往的检测方法基于平稳和线性过程的问题; 由于分割时采用多次迭代的一分为二的算法, 大大减少了计算量, 实用性较好; 白噪声和尖峰噪声对该方法的影响较小^[4,5]。启发式分割算法目前已经在医学、气象等领域有了

成功的应用, 但尚未应用于水文变异诊断。本文将简要介绍启发式分割算法, 通过构建理想样本探讨其在水文变异检验中应用的可行性, 并与传统检验方法检验性能进行对比, 最后结合受剧烈人类活动影响的珠江三角洲河网三水站 1900 - 2004 年年最大洪峰流量序列资料检验其实际应用效果。

1 启发式分割算法原理

对于一个由 N 个点构成的时间序列 $x(t)$, 从左到右分别计算每个点左边部分和右边部分的平均值 $\mu_1(i)$ 和 $\mu_2(i)$ 及标准偏差 $s_1(i)$ 和 $s_2(i)$, 则 i 点的合并偏差 $s_D(i)$ 为:

$$s_D(i) = \left(\frac{(N_1 - 1) \times s_1(i)^2 + (N_2 - 1) \times s_2(i)^2}{N_1 + N_2 - 2} \right)^{1/2} \times \left(\frac{1}{N_1} + \frac{1}{N_2} \right)^{1/2} \quad (1)$$

其中, N_1 、 N_2 分别为 i 点左边和右边部分的点数。 i 点左右两部份均值的差异可以用 t 检验统计值 $T(i)$ 来量化表示:

$$T(i) = \left| \frac{\mu_1(i) - \mu_2(i)}{s_D(i)} \right| \quad (2)$$

对 $x(t)$ 中的每一个点重复上述计算过程, 得到与 $x(t)$ 一一对应的检验统计值序列 $T(t)$ 。 T 越大, 表示该点左右两部份的均值相差越大。计算 $T(t)$ 中最大值 $T_{\max}$ 的统计显著性 $P(T_{\max})$:

$$P(T_{\max}) = \text{Prob}(T \leq T_{\max}) \quad (3)$$

$P(T_{\max})$ 表示在随机过程中取到 T 值小于等于 $T_{\max}$

* 收稿日期: 2008 - 03 - 28

基金项目: 国家自然科学基金资助项目 (50579052)

作者简介: 陈广才 (1981 年生), 男, 博士研究生; E-mail: cgc07@163.com

的概率。一般情况下 $T_{\max}$ 可近似表示:

$$P(T_{\max}) \approx (1 - I_{v/(v+T_{\max}^2)}(\delta v, \delta))^{\eta} \quad (4)$$

由蒙特卡洛模拟可以得到: $\eta = 4.19 \ln N - 11.54$, $\delta = 0.4$, N 是时间序列 $x(t)$ 的长度, $v = N - 2$, $I_x(a, b)$ 为不完全 β 函数。设定一个临界值 P_0 , 如果 $P(T_{\max}) \geq P_0$, 则于该点将 $x(t)$ 分割成两段均值有一定差异的子序列, 否则不分割。对新得到的两个子序列分别重复上述操作, 如果子序列有 $P(T_{\max}) \geq P_0$, 并且子序列与其左、右相邻的子序列间均值的差异程度均满足上述条件, 则对子序列再进行分割, 否则不分割。如此重复直至所有的子序列都不可分割为止。为确保统计的有效性, 当子序列的长度小于等于 l_0 (l_0 为最小分割尺度) 时不再对其进行分割。通过上述操作, 可以将原序列分割为若干不同均值的子序列, 分割点即为均值变异点。

一般而言, l_0 的取值不小于 25, P_0 可取 0.5 ~ 0.95^[4]。通过改变 l_0 和 P_0 , 可以实现对序列不同尺度的变异检测。

2 数值实验

2.1 理想样本的生成

由于水文要素多服从 P-III 型分布, 为了验证启发式分割算法在水文要素变异检验中的可行性, 随机生成满足 P-III 型分布的一组均值不同的序列。令 $x_t = f(x_0, C_v, C_s)$, x_t 为服从 P-III 型分布, 且统计参数为 x_0 , C_v 和 C_s 的随机数。本文假设 x_t 为

$$x_t = \begin{cases} f(1500, 0.5, 1.0) & t \leq 800 \\ f(3000, 0.5, 1.0) & 800 < t \leq 2800 \\ f(1000, 0.5, 1.0) & 2800 < t \leq 5000 \\ f(2000, 0.5, 1.0) & 5000 < t \leq 8000 \end{cases} \quad (5)$$

设置理论变异点分别为 800, 2800 和 5000, 采用启发式分割算法对该理想样本进行检验 (如图 1 所示), 取 $l_0 = 25$, $P_0 = 0.95$ 。

图 1 (a) 为随机生成的理想样本。由图 1 (b) 可知第一次检验到在 $t = 2800$ 时, T 达到最大值, 且 $P(T_{\max}) = 1$, 因 $P(T_{\max}) > P_0$, 故 2800 为变异点, 将原序列分割成两段, 继续检验可得图 1 (c) 和图 1 (d)。由图 1 (c) 和图 1 (d) 可以看出第二次检验到在 $t = 805$ 和 $t = 5000$ 时, T 达到最大值, 且均有 $P(T_{\max}) = 1$, 因 $P(T_{\max}) > P_0$, 故变异点为 805 和 5000, 原序列被分割为四段。分别对四段继续进行变异检验。以 $x_1 - x_{805}$ 段为例, 检验结果如图 1 (e) 所示。由图 1 (e) 可以看出, 在 $t = 389$ 时, T 达到最大值。而 $P(T_{\max}) =$

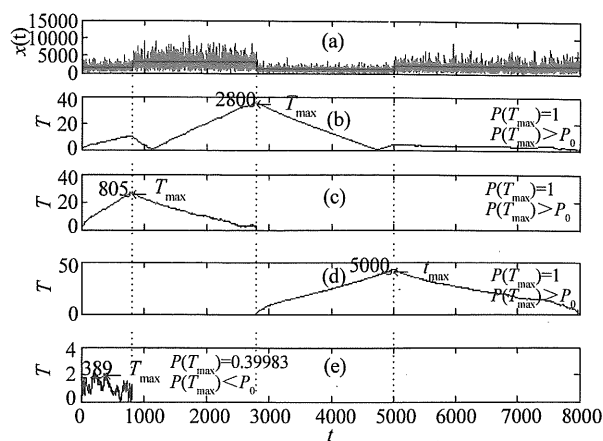


图 1 理想样本分割检验图

Fig. 1 Segmentation of ideal sample

0.399 83, $P(T_{\max}) < P_0$, 变异不显著, 故不需要再分割该段样本。在 $x_{806} - x_{2800}$, $x_{2801} - x_{5000}$ 和 $x_{5001} - x_{8000}$ 也有类似结论, 无需继续分割样本。

综上所述, 启发式分割算法在对理想样本的检验中成功地找到了 3 个变异点, 分别为 805, 2800 和 5000, 与理想变异点位置基本相同。产生的微弱偏差是在生成理想样本时的随机误差造成的。由此可见, 启发式分割算法可以应用于水文要素的均值变异诊断中。

2.2 参数对检验结果的影响

改变参数 l_0 和 P_0 的取值, 重新对理想样本进行变异诊断。取 $P_0 = 0.95$, l_0 分别取 50, 100, 500, 1 000 和 1 500, 检验结果见表 1。

表 1 不同 l_0 对检验结果影响

Tab. 1 Influence of different l_0 on the change point analysis results

l_0	50	100	500	1 000	1500
变	805	805	805	805	805
异	2800	2800	2800	2800	2800
点	5000	5000	5000	5000	5000
变异点个数	3	3	3	3	3

由表 1 可以看出, 不同 l_0 检测结果均相同。类似的取 $l_0 = 25$, P_0 分别取 0.75, 0.8, 0.9, 变异点检验结果也完全相同, 这进一步说明启发式分割算法变异检验的结果比较稳定。

2.3 与传统检验方法比较

为了便于说明, 本文仅选用理想样本前 800 点进行变异检验。由理想样本的设计可知, 该段不存在变异点, 称该段为无变异理想样本。选用常用的 10 种水文变异识别方法对该段进行变异检验, 分别是有序聚类法、Lee-Heghinan 法、滑动秩和检验

法、滑动游程检验法、滑动 F 检验法、滑动 T 检验法、最优信息二分割模型、R/S 法、Brown-Forsythe、Mann-Kendall 法。显著性水平取为 0.05, 计算结果如表 2 所示。

表 2 传统检验方法对无变异理想样本的检验结果

Tab. 2 Results of various traditional change point analysis methods in checking change point of ideal sample

方法名称	检验变异点	方法名称	检验变异点
有序聚类法	389	滑动 T 检验法	389
Lee-Heghinan 法	799	最优信息二分割模型	257
滑动秩和检验法	14	R/S 法	574
滑动游程检验法	18	Brown-Forsythe	无变异点
滑动 F 检验法	无变异点	Mann-Kendall	175

由表 2 可以看出, 上述 10 种方法中有 8 种方法找到了变异点, 且 389 点 2 次被检验为变异点, 显然传统检验方法带来了虚假变异点。而启发式分割算法在检验该段序列时, 没检验出变异点, 检验性能较好。

3 水文应用实例

3.1 数据来源及变异检验

采用 1900–2004 年珠江三角洲河网上游控制站三水站年最大洪峰流量序列进行变异诊断。取 $l_0 = 25$, $P_0 = 0.95$ 进行检验, 结果如图 2。

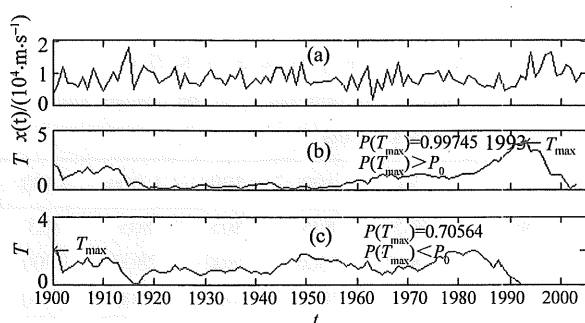


图 2 1900–2004 年三水站年最大洪峰流量序列分割检验图

Fig. 2 Segmentation of annual maximum flood peak discharge data at Shanshui station from 1900 to 2004

由图 2 (b) 可知, 1993 年发生变异, 序列分为两段, 1900–1993 和 1994–2004。第二段序列长度为 11, 小于检验序列的最小长度 l_0 , 该段无需再检验。对第一段序列长度 $l_2 = 94$, $l_2 > l_0$, 需要继续进行检验 (如图 2 (c))。由图 2 (c) 可知, 在 $t = 1900$ 时, T 达到最大值。而 $P(T_{\max}) =$

0.70564 , $P(T_{\max}) < P_0$, 变异不显著, 故不需要再分割该段样本。

取用不同的 $l_0 = 30, 50, 100$, $P_0 = 0.95$ 进行变异检验, 诊断结果均显示 1993 年发生变异, 表明在 30、50、100 年分割尺度内, 1993 年均发生变异。

综上所述, 1900–2004 年三水站年最大洪峰流量序列在 1993 年发生了变异, 且为均值向上跳跃。

3.2 物理成因分析

由于气候变化和人类活动的影响, 特别是近几十年珠江三角洲地区经济和社会的快速发展, 该地区水文情势发生了较大变化。20 世纪 80 年代中期, 该地区出现大规模采掘河床泥沙的现象, 至 90 年代初达到高潮。陈晓宏^[6]分析了珠江三角洲网河区水文与地貌特征变异及其成因, 指出该地区 1992 年前后的 10 a 期间水文情势发生了显著变异; 黄勇强^[7]指出三水站水位流量关系曲线发生变化, 1993 年后分流比和分沙比发生变化。本文采用 1990–2003 年实测资料分析三水站的大断面变化情况。在 1993 年以前, 历年断面冲淤基本平衡, 中泓河底高程变化不大; 1993 年后河床逐年大幅下切, 断面面积逐年显著增大。另据 1980–2003 年的资料分析^[8]: 三水站河段最大切深达 8.07 m, 中高水断面过水面积增大达 40%, 中低水断面过水面积增大将近 100%。

由此可见, 三水站在 1993 年河床开始下切, 过水断面面积不断增大, 水位流量关系、水沙关系、思贤窖分流比分沙比等均发生显著变化, 这是三水站年最大洪峰流量序列发生变异的物理成因之一。

4 结 论

在对理想样本进行检验研究的基础上, 初步探讨了启发式分割算法在水文要素变异诊断研究中的可行性, 主要结论如下:

(1) 启发式分割算法具有较好的检验均值变点的能力, 并可以对多变点进行识别检验。通过改变最小分割长度和临界水平, 可以实现不同尺度的变异识别。

(2) 通过设定变异点构建总长度为 8000, 服从 P-III 型分布的理想样本, 并应用启发式分割算法对理想样本进行均值变异诊断, 其检验变异点与设定变异点基本相同, 微弱差异主要由生成样本时的随机误差造成, 表明该方法在水文变异诊断中具有一定的可行性。

(3) 通过三水站 1900 - 2004 年年最大洪峰流量序列的实际验证, 发现启发式分割算法的实际应用效果较好。不同分割尺度检验均表明三水站年最大洪峰流量序列在 1993 年发生变异, 变异前后均值分别为 $8\ 104.68$ 和 $11\ 540.28\text{m}^3/\text{s}$, 两者相差 $3\ 435.6\text{m}^3/\text{s}$, 且为均值向上跳跃, 表明区域防洪形势将更加趋于严峻。

参考文献:

- [1] 丁晶, 邓育仁. 随机水文学 [M]. 成都: 成都科技大学出版社, 1988.
DING Jing, DENG Yuren. Random Hydrology [M]. Chengdu: Chengdu University of Science and Technology Press, 1988.
- [2] BERNOLA - GALVAN P, IVANOV P Ch, NUNES AMARAL L A, et al. Scale invariance in the nonstationarity of human heart rate [J]. Physical Review Letters, 2001, 87(16): 160815
- [3] FUKUDA K, EUGENE STANLEY H, NUNES AMARAL L A. Heuristic segmentation of a nonstationary time series [J]. Physical Review E, 2004, 69: 021108
- [4] 龚志强. 基于非线性时间序列分析方法的气候突变检测研究 [D]. 江苏扬州: 扬州大学, 2006.
GONG Zhiqiang. Research on the climate mutation detection based on the nonlinear time series analysis method [D]. Yangzhou: Yangzhou University, 2006.
- [5] 龚志强, 封国林, 万仕全等. 基于启发式分割算法检测华北和全球气候变化的特征 [J]. 物理学报. 2006, 55(1): 477 - 484.
GONG Zhiqiang, FENG Guoling, WAN Shiquan, et al. Analysis of features of climate change of Huabei area and the global climate change based on heuristic segmentation algorithm. Acta Physica Sinica, 2006, 55(1): 477 - 484.
- [6] 陈晓宏, 陈永勤. 珠江三角洲网河区水文与地貌特征变异及其成因 [J]. 地理学报. 2002, 57(4): 429 - 436.
CHEN Xiaohong, CHEN Yongqin. Hydrological change and its causes in the river network of the Pearl River Delta [J]. Acta Geographica Sinica, 2002, 57(4): 429 - 436.
- [7] 黄勇强. 浅析河床下切与北江三水站水文要素的变化 [J]. 广东水利电力职业技术学院学报. 2004, 2(4): 25 - 27.
HUANG Yongqiang. River-bed deepened with the change of hydrologic element at Sanshui station [J]. Journal of Guangdong Technical College of Water Resource and Electric Engineering, 2004, 2(4): 25 - 27.
- [8] 谢绍平. 西江中、下游河床下切变化及洪水预报改进研究 [D]. 武汉: 武汉大学, 2004.
XIE Shaoping. The change of the mid-downstream bed-downcutting in Xijiang river and the improved research of flood forecasting [D]. Wuhan: Wuhan University, 2004.

Change Point Analysis of Hydrologic Time Series Based on Heuristic Segmentation Method

CHEN Guang-cai, XIE Ping

(State Key Laboratory of Water Resources and Hydropower Engineering Science,
Wuhan University, Wuhan 430072, China)

Abstract: The feasibility of hydrologic variation analysis with heuristic segmentation was discussed with constructed ideal samples in this paper. And its performance was compared with traditional change point analysis methods. The annual maximum flood peak discharge data from 1900 to 2004 of Shanshui station in river network of pearl river delta was used to test the practical effect in change point analysis. The results show that the heuristic segmentation method can eliminate interference of illusive change point, identify the change point and its occurring time. It is suitable to deal with change point in hydrologic time series, especially for nonlinear nonstationary time series. The practical application effect is satisfactory. The change point in Shanshui station is 1993. The mean of the series jumped up after 1993, and it goes against the flood defense situation.

Key words: hydrologic variation; heuristic segmentation method; time series; Shanshui station; annual maximum flood peak discharge