doi: 10.12011/SETP2020-0098 中图法分类号: O212.1; P333.6 文献标志码: A

广义 Pareto 分布变点检测重现期预测模型及其应用

胡 尧[1,4], 谌业文[2], 王旭琴[3], 刘 伟[1]

(1. 贵州大学 数学与统计学院, 贵阳 550025; 2. 中山大学 数学学院, 广州 510000; 3. 厦门大学 经济学院, 厦门 361005; 4. 贵州大学 贵州省公共大数据重点实验室, 贵阳 550025)

摘 要 广义 Pareto 分布 (generalized Pareto distribution, GPD) 变点是指其超出量发生质变的点, 具体表现为一个或多个参数的变化. 本文将极端暴雨数据的三参数 GPD 变点检测问题表示为假设检验问题, 通过极大似然比检验统计量解决. 尽管不可能得到检验统计量的精确分布, 但通过证明 GPD 变点的极限性质和检验统计量的渐近收敛定理, 可得到它的极限分布. 同时, 利用 GPD 变点检测模型, 对深圳 55 年的月最大暴雨数据进行了分析. 结果发现 20 世纪 80 年代是深圳气象学的一个变点, 这与最小超阈值 (narrowest over threshold, NOT) 方法所得结果一致. 除此之外, 本文方法的优势在于它统一了变点检测前后的分析框架. 在此框架下, 通过对变点的分析, 发现 GPD 变点之前的极值指数为负. 结合降雨量与重现期的关系可得, 重现期小于百年的暴雨强度较之前有所减弱. 该点之后, 极值指数为正, 百年以上重现期的暴雨灾害程度较之前严重且极端降雨现象较之前频繁. 实证结果表明, GPD 变点模型能够较好地捕获传统 GPD 模型所不能捕捉的内在规律, 较好地弥补了传统 GPD 模型的不足.

关键词 广义 Pareto 分布 (GPD) 变点; 似然比检验统计量; 变点检测; 重现期; 最小超阈值 (NOT) 方法; 预测模型

Generalized Pareto distribution change point detection model and its application in modeling return period prediction

HU Yao[1,4], CHEN Yewen[2], WANG Xuqin[3], LIU Wei[1]

(1. School of Mathematics and Statistics, Guizhou University, Guiyang 550025, China; 2. School of Mathematics, Sun Yat-sen University, Guangzhou 510000, China; 3. The School of Economics, Xiamen University, Xiamen 361005, China; 4. Guizhou Provincial Key Laboratory of Public Big Data, Guizhou University, Guiyang 550025, China)

Abstract Generalized Pareto distribution (GPD) change point refers to the point where the quantity whose the peak over threshold changes qualitatively, which shows the change of one or more GPD parameters. In this paper, the detection issue of GPD change point with three parameters for extreme rainstorm data was expressed as the hypothesis test problem which could be solved by the maximum likelihood ratio test statistic. Although it is impossible to derive an exact distribution of this test statistic, fortunately, we obtained its limiting distribution by proving the limit properties of the GPD change point and the asymptotic convergence theorem of the test statistic. Meanwhile, the detection model we developed was

收稿日期: 2020-01-20

作者简介: 胡尧 (1971–), 男, 汉, 贵州遵义人, 教授, 硕士生导师, 研究方向: 应用统计, E-mail: yhu1@gzu.edu.cn; 通信作者: 谌业文 (1989–), 男, 侗族, 贵州凯里人, 博士研究生, 研究方向: 空间统计, E-mail: chenyw20150908@163.com; 王旭琴 (1991–), 女, 汉, 山西吕梁人, 博士研究生, 研究方向: 应用统计, E-mail: xqwang_0408@163.com; 刘伟 (1994–), 男, 汉, 四川宜宾人, 硕士研究生, 研究方向: 概率论与数理统计, E-mail: 910581605@qq.com.

基金项目: 国家自然科学基金 (12161016, 11661018); 贵州省科技计划项目 (黔科合平台人才 [2017]5788 号); 贵州省数据驱动建模学习与优化创新团队 (黔科合平台人才 [2020]5016); 全国统计科学研究项目 (2014LZ46)

Foundation item: National Natural Science Foundation of China (12161016, 11661018); Science and Technology Program of Guizhou Province, China ([2017]5788); Guizhou Data Drive Modeling Learning and Optimization Innovation Team ([2020]5016); Science Foundation of National Bureau of Statistics of China (2014LZ46)

中文引用格式: 胡尧, 谌业文, 王旭琴, 等. 广义 Pareto 分布变点检测重现期预测模型及其应用 [J]. 系统工程理论与实践, 2021, 41(9): 2379–2391.

英文引用格式: Hu Y, Chen Y W, Wang X Q, et al. Generalized Pareto distribution change point detection model and its application in modeling return period prediction[J]. Systems Engineering — Theory & Practice, 2021, 41(9): 2379–2391.

used to analyze the monthly maximum precipitation data of Shenzhen with a total number of 55 years. As a result, we found the 1980s is a change point of Shenzhen meteorology, which is identical with the narrowest over threshold (NOT) method. In addition, the advantage of GPD change point method is that it ensures a consistent analytical framework before and after the detection of change point. In this framework, the analysis of GPD change point revealed that the extreme value index before that point is negative. Combining the relationship between rainfall and the return period, it might be concluded that the severity of rainstorm in the return period less than 100 years is weaker than before. After that point, the extreme value index is positive, and the degree of rainstorm disaster in the return period more than 100 years is more serious and the phenomenon of extreme rainstorm is more frequent than before. The empirical result showed that the GPD change point model is capable to capture the potential rules that the conventional GPD model impotently, which could moderately make up for the deficiency of conventional GPD model in this respect.

Keywords generalized Pareto distribution (GPD) change point; likelihood ratio test statistic; change point detection; return period; narrowest over threshold (NOT) method; prediction model

1 引言

统计学中的极值理论 (extreme value theory, EVT) 是描述和研究异常极端随机事件的技术性模型, 主要包括分块极大值模型 (block maxima model, BMM) 和超阈值 (peak over threshold, POT) 模型两类 [1]. BMM 主要对广义极值 (generalized extreme value, GEV) 分布族进行分析 [2], 该分布族因使用的是块内最大值建模, 故易造成数据浪费, 而 POT 是对大于给定阀值的数据建模, 具有充分利用数据信息的优势 [3,4]. 广义 Pareto 分布 (generalized Pareto distribution, GPD) 作为 POT 渐近分布 [1,5], 在描述金融危机 [6]、自然灾害 [7−9] 与交通拥堵 [10] 等极端事件中具有广泛应用.

特别地, GPD 作为描述极端事件的理想模型, 被广泛运用于如洪水、暴雨与风暴潮海浪等水文气象领域的量化风险分析. 为评估极端事件的影响, Kiriliouk 等 [3] 构建多变量 GPD 模型, 提出尾似然方法并应用于降雨所致滑坡风险评估. Chen 等 [2] 和 Luo 等 [4] 都通过构建 GPD 模型来估计不同重现期的海浪高度, 不仅提高了估计精度而且增强了稳健性. Mo 等 [7] 给出 GPD 极端降雨频率框架, 分析澄碧河流域降雨数据, 预测诊断突变规律. Park 等 [6] 结合经验分布与理论 GPD 之间的偏差, 提出非线性加权最小二乘估计并分析丹麦火灾数据, 数值模拟与实例验证了所提估计方法的有效性. 刘新红等 [8] 构建 GPD 与负二项混合分布模型评估地震灾害损失风险, 为震灾保险提供了一种精算模型. Wu 等 [11] 将 Weibull 分布和 GPD 结合, 提出了浅海区域浪高预测模型, 该模型对于浅海浪高数据具有良好的预测性. Ashkar 等 [12] 基于极大似然方法, 在小样本情形下改进分位数估计, 并用于河水流量预测, 取得了较好效果.

GPD 变点是指超出量发生质变之点. 自然现象中某些量变过程可视为随机过程, 该随机过程可能由一个平稳状态向另一个平稳状态转变即质变, 如全球变暖导致较低海平面向较高海平面质变. GPD 是寻找和拟合这些事件质变的最佳模型之一, 质变主要体现在 GPD 参数的改变即 GPD 参数变点. 因此, 用 GPD 变点模型研究自然灾害, 寻找突变规律, 特别是对灾害重现期的预测具有应用指导意义.

目前 GPD 变点研究不多. Bayesian 框架下, Chen 等 [13] 提出 GPD 极端气候数据变点检测方法, Renard 等 [14] 分别考察平稳、跳跃和线性趋势变化三种情形下 GPD 分布变点特征, 结果发现 GPD 变点模型分析极端事件所得结论更可靠, 进而 Meng[15] 通过结合 Bayesian 理论和期望极大化算法给出 GPD 变点区间估计, 海面温度数据验证了所提 GPD 变点分析的有效性. 在经典似然领域, Dierckx 等 [16] 利用 Pareto 与指数分布的关系, 对 Pareto 分布与 GPD 变点检测方法进行比较, 发现 GPD 变点检测更佳, Susan 等 [5] 将 GPD 形状参数变点检测转化为模型选择问题, 给出基于 Kullback-Leibler 散度的 GPD 似然比变点检测统计量, 用似然方法估计变点, 实证分析极端事件变化数据. 而在其它领域, Raimondo 等 [17] 基于小波提出 POT 变点检测模型, Safari 等 [18] 构建 Pareto 箱线图用于变点和异常数据检测. 上述研究极具理论与实践意义, 但三参数 GPD 变点问题研究相对更少 [5], 对三参数 GPD 建立有效的变点检测模型, 不仅在理论上有着重要的指导意义, 在实践中也能更精准把握极端事件变化规律, 为相关人员决策提供更可靠的参考依据.

论文首先给出极端事件数据超阈值 GPD 定义、阈值选择方法与灾害重现水平分位点估计. 其次, 提出

GPD 变点假设检验问题, 构建极大化似然比检验统计量, 研究经参数变换后的 GPD 变点相关性质和有关收敛定理证明, 为模拟检验统计量渐近临界值提供理论依据. 最后, 案例分析 GPD 变点模型在水文气象领域的应用, 比较 Baranowski 等 [19] 最小超阈值 (narrowest over threshold, NOT) 变点检测方法, 并将 GPD 变点应用于深圳降雨量重现期预测分析, 以此获得一些实用性结果, 为洪灾预防提供一种参考模型.

2 广义 Pareto 分布

水文气象领域, 强对流天气如降雨量达到某值会导致水位骤变的自然现象称为极端天气事件, 简称极端事件 [20]. 风暴潮强降雨就是其中较为典型的案例, 这种极端事件可作为统计学中的随机事件来研究, 其中比较重要的是对其重现期方面的分析, 重现期是指相应水位的风暴潮或某强降雨发生概率的倒数 [21−24].

2.1 GPD 定义

由极值定理 [1] 可知, 只要分块样本容量足够大, 分块最大值就渐近服从 GEV 分布. 但与已选择的分块最大值相比, 若分块内尚有其它极端事件未被选择, 那必定会造成数据浪费, 从而导致估计结果存在较大偏差. 基于此, 专家学者们转向研究超阈值模型, 选定阀值之后利用大于阈值的数据进行统计建模. 下面首先给出超出量分布定义.

定义 1 设独立同分布样本 $X_1,\cdots,X_n$ 服从分布函数 $F(x)$, x^* 为 $F(x)$ 的右支撑端点, 选定阈值 $u<x^*$. 若 $X_i>u$, 即 X_i 超阈值 u, 称 $Y_i=X_i-u$ 为超出量.

用 $F_u(y)$ 表示随机变量 $Y=X-u|X>u$ 的分布函数, 简称超出量分布, 则

$$F_u(y)=P(X-u\le y|X>u),y>0. \tag{1}$$

当 u 充分大, $F_u(y)$ 渐近服从 GPD[1], 其定义形式如下.

定义 2 若随机变量 X 有分布函数

$$G(x;u,\sigma,\gamma)=\begin{cases}1-\left(1+\gamma\dfrac{x-u}{\sigma}\right)^{-\frac{1}{\gamma}}, & \gamma\neq 0,\\ 1-\exp\left(-\dfrac{x-u}{\sigma}\right), & \gamma=0,\end{cases}\qquad x\ge u,1+\frac{\gamma}{\sigma}(x-u)>0. \tag{2}$$

则称 X 服从三参数 GPD 分布, 记为 $X\sim G(x;u,\sigma,\gamma)$, 密度函数为 $g(x;u,\sigma,\gamma)$. 其中 $u,\gamma\in R$ 和 $\sigma>0$ 分别为位置、形状和尺度参数, 记为 $\mathbf{\Phi}=(u,\sigma,\gamma)$.

2.2 阀值选择

阀值 u 的选择除了考虑样本数据信息能被充分利用外, 还得权衡渐近性 [1], 当选择阀值过小超出量不能收敛于 GPD, 或阀值过大浪费样本数据导致模型不稳健, 故两者都必须重视. 在此引进一种比较合理可行的阀值选择方法.

给定充分大的阈值 u_0, 对任意 $u>u_0$, 有超额均值函数 [1]

$$e(u)=\frac{\sigma+\gamma u}{1-\gamma}. \tag{3}$$

由此看出超额均值函数 $e(u)$ 与阀值 u 理论上具有线性关系. $e(u)$ 估计值为

$$\hat{e}(u)=\frac{\sum_{i=1}^{n}(X_i-u)I_{(u,+\infty)}(X_i)}{\sum_{i=1}^{n}I_{(u,+\infty)}(X_i)}=\frac{1}{N_u}\sum_{j=1}^{N_u}(X_j'-u).$$

式中, $N_u=\{i|X_i>u\}^{\#}$ 为样本 $X_1,X_2,\cdots,X_n$ 中超过阀值 u 的个数. $X_1',X_2',\cdots,X_{N_u}'$ 为 $X_1,X_2,\cdots,X_n$ 中大于给定阀值的样本观测值. 故实际操作中, 将点

$$\left\{\left(u,\frac{1}{N_u}\Sigma_{j=1}^{N_u}(X_j'-u)\right):u<X_{(n)}\right\}$$

描绘于二维坐标系中, 若 $\hat{e}(u)$ 关于 u $(u>u_0)$ 近似线性变化, 那么 u_0 可作为一个合适的阀值. 实际中通过确定 k 值, 进而选取顺序统计量 $X_{(n-k)}$ 作为阀值 u_0, 有关统计推断也将基于 $X_{(n-k+1)},\cdots,X_{(n)}$ 给出.

此外, 由于 u_0 与不同阀值 u 所对应的尺度参数 σ 关于 u 线性变化, 但形状参数 γ 并没有改变, 故可选择一个比较平稳的形状参数及修正后的尺度参数估计值对应的 u 作为选择结果 [1], 具体见 4.2 节案例实现.

2.3 GPD 分位点估计

利用条件公式推导变换 (1) 式得

$$F(x) = F(u) + [1 - F(u)]F_u(y). \tag{4}$$

将 $F_u(y)$ 用 $G(x; u, \sigma, \gamma)$ 代入得

$$F(x) = F(u) + [1 - F(u)]G(x; u, \sigma, \gamma). \tag{5}$$

上式 $F(u)$ 用经验函数 $\hat{F}(u) = 1 - \frac{N_u}{n}$ 估计, $X_1', X_2', \cdots, X_{N_u}'$ 对$G(x; u, \sigma, \gamma)$ 进行参数估计得 $G(x; \hat{u}, \hat{\sigma}, \hat{\gamma})$. 故

$$\hat{F}(x) = \begin{cases} 1 - \dfrac{N_u}{n}\left(1 + \hat{\gamma}\dfrac{x - \hat{u}}{\hat{\sigma}}\right)^{-\frac{1}{\hat{\gamma}}}, & \gamma \neq 0, \\ 1 - \dfrac{N_u}{n}\exp\left(-\dfrac{x - \hat{u}}{\hat{\sigma}}\right), & \gamma = 0, \end{cases} \quad x \geq u, 1 + \frac{\gamma}{\sigma}(x - u) > 0. \tag{6}$$

相应的 $F(x)$ 尾概率估计为

$$\hat{\bar{F}}(x) = \begin{cases} \dfrac{N_u}{n}\left(1 + \hat{\gamma}\dfrac{x - \hat{u}}{\hat{\sigma}}\right)^{-\frac{1}{\hat{\gamma}}}, & \gamma \neq 0, \\ \dfrac{N_u}{n}\exp\left(-\dfrac{x - \hat{u}}{\hat{\sigma}}\right), & \gamma = 0, \end{cases} \quad x \geq u, 1 + \frac{\gamma}{\sigma}(x - u) > 0. \tag{7}$$

令 $p = F(x)$, 则 p 分位数点 x_p 的估计值为

$$\hat{x}_p = \begin{cases} \hat{u} + \dfrac{\hat{\sigma}}{\hat{\gamma}}\left[\left(\dfrac{N_u}{n - np}\right)^{\hat{\gamma}} - 1\right], & \gamma \neq 0, \\ \hat{u} + \hat{\sigma}\ln\dfrac{N_u}{n - np}, & \gamma = 0, \end{cases} \quad x \geq u, 1 + \frac{\gamma}{\sigma}(x - u) > 0. \tag{8}$$

重现期与重现水平密切相关. 记极端事件重现期为 $(1 - p)^{-1}$, 对应的重现期水平 (简称重现水平) 记为 x_p. 对于降雨事件而言, x_p 也表示重现期为 $(1 - p)^{-1}$ 时的降雨量, 后面案例分析研究对象为月最大降雨量, 解读为特定月份发生超过降雨量水平为 x_p 的强降雨事件概率为 $1 - p$, 记 n_T 为每年的数据容量, T 为总年数, 则 (7) 式有

$$\frac{1}{Tn_T} = \begin{cases} \dfrac{N_u}{n}\left(1 + \hat{\gamma}\dfrac{x - \hat{u}}{\hat{\sigma}}\right)^{-\frac{1}{\hat{\gamma}}}, & \gamma \neq 0, \\ \dfrac{N_u}{n}\exp\left(-\dfrac{x - \hat{u}}{\hat{\sigma}}\right), & \gamma = 0. \end{cases} \tag{9}$$

那么 T 年重现水平可通过下式估计

$$\hat{x}_p = \begin{cases} \hat{u} + \dfrac{\hat{\sigma}}{\hat{\gamma}}\left[\left(\dfrac{Tn_T N_u}{n}\right)^{\hat{\gamma}} - 1\right], & \gamma \neq 0, \\ \hat{u} + \hat{\sigma}\ln\dfrac{Tn_T N_u}{n}, & \gamma = 0. \end{cases} \tag{10}$$

3 GPD 变点检测模型

3.1 GPD 变点假设检验问题

类似 Jarušková[25] 考虑检验问题

$$\begin{aligned} &H_0 : X_i \sim G(x; u_0, \sigma_0, \gamma_0), i = 1, \cdots, n \quad \text{vs.} \quad H_1 : \text{存在 } m \in \{n_0, n_0 + 1, n_0 + 2, \cdots, n - n_0\} \\ &\qquad \text{使 } X_i \sim G(x; u_1, \sigma_1, \gamma_1), i = 1, \cdots, m \text{ 和 } X_i \sim G(x; u_2, \sigma_2, \gamma_2), i = m + 1, \cdots, n. \end{aligned} \tag{11}$$

变点前 $\boldsymbol{\Phi}_1 = (u_1, \sigma_1, \gamma_1)$ 和后 $\boldsymbol{\Phi}_2 = (u_2, \sigma_2, \gamma_2)$ 未知不等. 构造极大化似然比检验统计量 [26]

$$Z_n = \sqrt{\max_{n_0 \leq m \leq n - n_0}(-2\ln\Lambda_m)}, \quad \text{式中} \quad \Lambda_m = \frac{\sup_{\boldsymbol{\Phi}_0}\prod_{i=1}^{n} g(X_i; \boldsymbol{\Phi}_0)}{\sup_{\boldsymbol{\Phi}_1}\prod_{i=1}^{m} g(X_i; \boldsymbol{\Phi}_1)\sup_{\boldsymbol{\Phi}_2}\prod_{i=m+1}^{n} g(X_i; \boldsymbol{\Phi}_2)}.$$

其中, $\boldsymbol{\Phi}_0$ 为参数真值. 由于样本量较小对 Λ_m 估计的可靠性难以保证, 故假定变点的位置 m 取值在 $[n_0, n - n_0]$ 区间, n_0 为较小的正整数. 显然, 由于 $\Lambda_m \leq 1$, 若真实变点 m^* 发生改变, 将导致更小的 Λ_{m^*}, 从而越大的 Z_n 越趋于拒绝 H_0, 即变点存在的可能性更大. 此外

$$\max_{n_0 \leq m \leq n - n_0}(-2\ln\Lambda_m) = \max_{n_0 \leq m \leq n - n_0}(2(L_1(\hat{\boldsymbol{\Phi}}_1) + L_2(\hat{\boldsymbol{\Phi}}_2) - L_n(\hat{\boldsymbol{\Phi}}))),$$

其中, $\hat{\boldsymbol{\Phi}}_1 = \hat{\boldsymbol{\Phi}}_1(X_1, \cdots, X_m)$, $\hat{\boldsymbol{\Phi}}_2 = \hat{\boldsymbol{\Phi}}_2(X_{m+1}, \cdots, X_n)$, $\hat{\boldsymbol{\Phi}} = \hat{\boldsymbol{\Phi}}(X_1, \cdots, X_n)$, L 为相应对数似然函数.

注意到解决检验问题 (11) 需推导 Z_n 在 H_0 下的精确分布, 但这极难研究, 故转而寻找其渐近分布并以此获得给定检验水平下的临界值. 而指数分布族下, 已有较为成熟的结论 [26]. 但对 GPD 而言, 由于分布定义在 $\{x: x \geq u, 1+\gamma(x-u)\sigma^{-1} > 0\}$ 上并不满足经典似然条件, 同时也不满足 Csörgö等 [26] 中 C.4 和 C.5 一些连续条件, 从而已有结论不能直接获得, 故对 GPD 做参数变换 [27].

1) 当 $\gamma < 0$ 时, 令 $\theta' = u - \sigma\gamma^{-1}$, $\beta = (-\gamma\sigma^{-1})^{-\gamma^{-1}}$ 和 $\alpha = -\gamma^{-1}$, 则 $\tilde{g}_1(x;\theta',\alpha,\beta) = \alpha\beta(\theta'-x)^{\alpha-1}I((\theta'-\beta^{-\alpha^{-1}}) \leq x \leq \theta')$, $I(\cdot)$ 为示性函数. 再令 $Y = -X$, $\theta = -\theta'$, 得

$$g_1(x;\theta,\alpha,\beta) = \alpha\beta(x-\theta)^{\alpha-1}I(\theta \leq x \leq (\theta+\beta^{-\alpha^{-1}})). \tag{12}$$

2) 当 $\gamma > 0$ 时, 令 $\theta = u - \sigma\gamma^{-1}$, $\beta = (\gamma\sigma^{-1})^{-\gamma^{-1}}$ 和 $\alpha = \gamma^{-1}$, 则

$$g_2(x;\theta,\alpha,\beta) = \alpha\beta(x-\theta)^{-\alpha-1}I((\theta+\beta^{\alpha^{-1}}) \leq x < \infty). \tag{13}$$

由函数 g_1 和 g_2 的一般形式不难得知, $\theta \in R, (\alpha,\beta) \in \Psi \subseteq R^2$, (θ,α,β) 的极大似然估计 $(\hat{\theta},\hat{\alpha},\hat{\beta})$, 满足 Smith[28] 极限定理条件, 故关于函数的似然方程存在序列 $(\hat{\theta}_k,\hat{\alpha}_k,\hat{\beta}_k)$, 使得如下成立.

$$\hat{\theta}_k - \theta_0 = O_p(\sqrt{k}), \quad \hat{\alpha}_k - \alpha_0 = O_p(\sqrt{k}), \quad \hat{\beta}_k - \beta_0 = O_p(\sqrt{k}).$$

式中, $A_k = O_p(a_k)$ 表示随机序列 A_k 和相关常数列 a_k 比值的随机有界, 即对于任意小的 $\epsilon > 0$, 存在 $K, \delta > 0$, 当 $k > K$ 时, 使得概率 $P(|\frac{A_k}{a_k}| > \delta) < \epsilon$ 成立. 下文中 $O_p(a_k)$ 简写为 $O(a_k)$.

3.2 GPD 变点检测模型

由于形状参数 γ 对于 GPD 是一个非常重要的量, 以下分别研究 $\gamma < 0$ 和 $\gamma > 0$ 情形. 为方便研究, 重构 GPD 参数变换后 g_1 和 g_2 的检验问题. 首先讨论形状参数小于 0 时的 GPD 变点检测理论问题.

关于 (12) 式, 考虑检验问题

$$\begin{aligned} &H_0: X_i \sim g_1(x;\theta_0,\alpha_0,\beta_0), i = 1,\cdots,n \quad \text{vs.} \quad H_1: \text{存在 } \tau \in \{n_0, n_0+1, n_0+2, \cdots, n-n_0\} \\ &\qquad \text{使 } X_i \sim g_1(x;\theta_1,\alpha_1,\beta_1), i = 1,\cdots,\tau \ \text{ 和 } \ X_i \sim g_1(x;\theta_2,\alpha_2,\beta_2), i = \tau+1,\cdots,n. \end{aligned} \tag{14}$$

仍记变点前 $\boldsymbol{\Phi}_1 = (\theta_1,\alpha_1,\beta_1)$ 和变点后 $\boldsymbol{\Phi}_2 = (\theta_2,\alpha_2,\beta_2)$, 参数未知且不等. 构造统计量

$$Z_n = \sqrt{\max_{n_0 \leq \tau \leq n-n_0}(-2\ln\Lambda_\tau)}.$$

为得到 Z_n 的渐近分布, 下面给出 $g_1(x;\theta,\alpha,\beta)$ 的主要结果, 更多细节见文献 [29].

$$L_k(\boldsymbol{\Phi}) = \sum_{i=1}^{k}\ln g_1(X_i;\boldsymbol{\Phi}) = k\ln\alpha + k\ln\beta + (\alpha-1)\sum_{i=1}^{k}\ln(X_i-\theta).$$

由 Smith[28] 可知, 当 $\alpha > 2$ 时, 存在 $\hat{\boldsymbol{\Phi}}_k = (\hat{\theta}_k,\hat{\alpha}_k,\hat{\beta}_k)$ 使 $\frac{\partial L_k(\hat{\boldsymbol{\Phi}}_k)}{\partial\boldsymbol{\Phi}} = 0$.

定义信息矩阵 $\boldsymbol{M} = \begin{pmatrix} m_{\theta\theta} & m_{\theta\alpha} & m_{\theta\beta} \\ m_{\alpha\theta} & m_{\alpha\alpha} & m_{\alpha\beta} \\ m_{\beta\theta} & m_{\beta\alpha} & m_{\beta\beta} \end{pmatrix}$, 其中,

$$m_{\theta\theta} = -E\left(\frac{\partial^2\ln g_1(X_i;\boldsymbol{\Phi})}{\partial\theta^2}\right), \quad m_{\alpha\alpha} = -E\left(\frac{\partial^2\ln g_1(X_i;\boldsymbol{\Phi})}{\partial\alpha^2}\right), \quad m_{\beta\beta} = -E\left(\frac{\partial^2\ln g_1(X_i;\boldsymbol{\Phi})}{\partial\beta^2}\right),$$

$$m_{\theta\alpha} = m_{\alpha\theta} = -E\left(\frac{\partial^2\ln g_1(X_i;\boldsymbol{\Phi})}{\partial\theta\partial\alpha}\right), m_{\theta\beta} = m_{\beta\theta} = -E\left(\frac{\partial^2\ln g_1(X_i;\boldsymbol{\Phi})}{\partial\theta\partial\beta}\right), m_{\alpha\beta} = m_{\beta\alpha} = -E\left(\frac{\partial^2\ln g_1(X_i;\boldsymbol{\Phi})}{\partial\alpha\partial\beta}\right).$$

对 $\delta > 0$ 和任意 $\{\delta_k > 0\}$, 满足 $\delta_k k^{\alpha^{-1}+\delta} \to 0$. 若 $\boldsymbol{\Phi}_0 = (\theta_0,\alpha_0,\beta_0)$, 设 $I_{\delta_k} = \{\hat{\theta} \in R, \hat{\alpha} > 2, \hat{\beta} > 0: |\hat{\theta}-\theta_0| < \delta_k, |\hat{\alpha}-\alpha_0| < \delta_k, |\hat{\beta}-\beta_0| < \delta_k\}$, $0 < r < 1-2\alpha^{-1}$, 有 (注: 其他参数情形类似, 见文献 [27, 29])

$$\lim_{k\to\infty} k^r\left(\sup_{I_{\delta_k}}\frac{1}{k}\left|\frac{\partial^2 L_k(\hat{\boldsymbol{\Phi}})}{\partial\theta\partial\alpha} - \frac{\partial^2 L_k(\boldsymbol{\Phi}_0)}{\partial\theta\partial\alpha}\right|\right) = 0 \ \text{ a.s.}, \lim_{k\to\infty} k^r\left(\sup_{I_{\delta_k}}\frac{1}{k}\left|\frac{\partial^2 L_k(\hat{\boldsymbol{\Phi}})}{\partial\theta^2} - \frac{\partial^2 L_k(\boldsymbol{\Phi}_0)}{\partial\theta^2}\right|\right) = 0 \ \text{ a.s..}$$

式中, a.s. 是 almost surely 简称, 表示事件 A 发生的概率为 1, 即 $P(A) = 1$, 记 A a.s., 以下类似.

定理 1 对任意 r, 使得 $0 < r < 1-2\alpha^{-1}$, 有

$$\lim_{k\to\infty}\frac{k^r}{\sqrt{k\ln\ln k}}\left(\begin{pmatrix} \partial L_k(\boldsymbol{\Phi}_0)/\partial\theta \\ \partial L_k(\boldsymbol{\Phi}_0)/\partial\alpha \\ \partial L_k(\boldsymbol{\Phi}_0)/\partial\beta \end{pmatrix} - k\boldsymbol{M}\begin{pmatrix} \hat{\theta}_k-\theta_0 \\ \hat{\alpha}_k-\alpha_0 \\ \hat{\beta}_k-\beta_0 \end{pmatrix}\right) = \boldsymbol{0} \quad \text{a.s.} \tag{15}$$

证明 令 $\tilde{\delta}_k = \ln\ln\ln k\sqrt{\ln\ln k/k}$, 且 $\tilde{\delta}_k k^{1/\alpha+\delta} \to 0$, 极大似然估计 $\hat{\boldsymbol{\Phi}}_k = (\hat{\theta}_k,\hat{\alpha}_k,\hat{\beta}_k)$, 则 Taylor 展开

$$\frac{\partial L_k(\hat{\boldsymbol{\Phi}}_k)}{\partial\theta}=\frac{\partial L_k(\boldsymbol{\Phi}_0)}{\partial\theta}+\frac{\partial^2 L_k(\tilde{\boldsymbol{\Phi}}_k)}{\partial\theta^2}(\hat{\theta}_k-\theta_0)+\frac{\partial^2 L_k(\tilde{\boldsymbol{\Phi}}_k)}{\partial\theta\partial\alpha}(\hat{\alpha}_k-\alpha_0)+\frac{\partial^2 L_k(\tilde{\boldsymbol{\Phi}}_k)}{\partial\theta\partial\beta_0}(\hat{\beta}_k-\beta_0). \tag{16}$$

式中, $\tilde{\boldsymbol{\Phi}}_k=(\tilde{\theta}_k,\tilde{\alpha}_k,\tilde{\beta}_k)$, $|\tilde{\theta}_k-\theta_0|<|\hat{\theta}_k-\theta_0|<\tilde{\delta}_k$, $|\tilde{\alpha}_k-\alpha_0|<|\hat{\alpha}_k-\alpha_0|<\tilde{\delta}_k$, $|\tilde{\beta}_k-\beta_0|<|\hat{\beta}_k-\beta_0|<\tilde{\delta}_k$. 进一步整理 (16) 式得

$$\begin{aligned}
&\frac{1}{\sqrt{k\ln\ln k}}\left(\frac{\partial L_k(\boldsymbol{\Phi}_0)}{\partial\theta}-k(\hat{\theta}_k-\theta_0)m_{\theta\theta}-k(\hat{\alpha}_k-\alpha_0)m_{\theta\alpha}-k(\hat{\beta}_k-\beta_0)m_{\theta\beta}\right)\\
&\quad=-\frac{1}{\sqrt{k\ln\ln k}}\left(\frac{\partial^2 L_k(\tilde{\boldsymbol{\Phi}}_k)}{\partial\theta^2}-\frac{\partial^2 L_k(\boldsymbol{\Phi}_0)}{\partial\theta^2}\right)(\hat{\theta}_k-\theta_0)-\frac{1}{\sqrt{k\ln\ln k}}\left(\frac{\partial^2 L_k(\tilde{\boldsymbol{\Phi}}_k)}{\partial\theta\partial\alpha}-\frac{\partial^2 L_k(\boldsymbol{\Phi}_0)}{\partial\theta\partial\alpha}\right)(\hat{\alpha}_k-\alpha_0)-\\
&\quad\frac{1}{\sqrt{k\ln\ln k}}\left(\frac{\partial^2 L_k(\tilde{\boldsymbol{\Phi}}_k)}{\partial\theta\partial\beta}-\frac{\partial^2 L_k(\boldsymbol{\Phi}_0)}{\partial\theta\partial\beta}\right)(\hat{\beta}_k-\beta_0)-\frac{1}{\sqrt{k\ln\ln k}}\left(\frac{\partial^2 L_k(\boldsymbol{\Phi}_0)}{\partial\theta^2}+km_{\theta\theta}\right)(\hat{\theta}_k-\theta_0)-\\
&\quad\frac{1}{\sqrt{k\ln\ln k}}\left(\frac{\partial^2 L_k(\boldsymbol{\Phi}_0)}{\partial\theta\partial\alpha}+km_{\theta\alpha}\right)(\hat{\alpha}_k-\alpha_0)-\frac{1}{\sqrt{k\ln\ln k}}\left(\frac{\partial^2 L_k(\boldsymbol{\Phi}_0)}{\partial\theta\partial\beta}+km_{\theta\beta}\right)(\hat{\beta}_k-\beta_0).
\end{aligned}\tag{17}$$

又 $|\tilde{\theta}_k-\theta_0|<\ln\ln\ln k\sqrt{\frac{\ln\ln k}{k}}$, $|\tilde{\alpha}_k-\alpha_0|<\ln\ln\ln k\sqrt{\frac{\ln\ln k}{k}}$, $|\tilde{\beta}_k-\beta_0|<\ln\ln\ln k\sqrt{\frac{\ln\ln k}{k}}$. 有

$$\lim_{k\to\infty}\frac{k^r}{\sqrt{k\ln\ln k}}\left[\frac{\partial L_k(\boldsymbol{\Phi}_0)}{\partial\theta}-k(\hat{\theta}_k-\theta_0)m_{\theta\theta}-k(\hat{\alpha}_k-\alpha_0)m_{\theta\alpha}-k(\hat{\beta}_k-\beta_0)m_{\theta\beta}\right]=0\quad\text{a.s.}$$

针对 α,β 类似可证, 详细证明见文献 [29].

推论 1 由定理 1 得似然估计序列 $\hat{\boldsymbol{\Phi}}_k$ 满足

$$\sup_k\frac{\sqrt{k}}{\sqrt{\ln\ln k}}|\hat{\theta}_k-\theta|=O(1)\quad\text{a.s.},\quad\sup_k\frac{\sqrt{k}}{\sqrt{\ln\ln k}}|\hat{\alpha}_k-\alpha|=O(1)\quad\text{a.s.},\quad\sup_k\frac{\sqrt{k}}{\sqrt{\ln\ln k}}|\hat{\beta}_k-\beta|=O(1)\quad\text{a.s.}.$$

式中, $A_k=O(1)$ 表示随机序列 A_k, 对任意小 $\epsilon>0$, 存在 $K,\delta>0$, 当 $k>K$ 时, 使得 $P(|A_k|>\delta)<\epsilon$ 成立.

证明 下证 β 情形. 密度 $g_1(x;\theta,\alpha,\beta)$ 的似然函数 $L_k(X;\boldsymbol{\Phi})$, 存在极大似然估计序列 $\hat{\boldsymbol{\Phi}}_k$, 使得 $\frac{\partial L_k(\hat{\boldsymbol{\Phi}}_k)}{\partial\beta}=0$. 对于函数 $L_k(\theta+\delta_k x,\alpha+\delta_k y,\beta+\delta_k z)$, 令 $(x,y,z)=(0,0,1)$, 利用 Taylor 展开得

$$\frac{\partial L_k(\hat{\boldsymbol{\Phi}}_k)}{\partial\beta}=\frac{\partial L_k(\boldsymbol{\Phi}_0)}{\partial\beta}+\frac{\partial^2 L_k(\tilde{\boldsymbol{\Phi}}_k)}{\partial\beta^2}(\hat{\beta}_k-\beta),$$

其中, $|\tilde{\beta}_k-\beta|<|\hat{\beta}_k-\beta|$. 而 $\frac{\partial L_k(\hat{\Phi}_k)}{\partial\beta}=0$, 对上式进行简单变换得

$$0=\frac{\partial L_k(\hat{\boldsymbol{\Phi}}_k)}{\partial\beta}=\frac{1}{k}\frac{\partial L_k(\boldsymbol{\Phi}_0)}{\partial\beta}-m_{\beta\beta}(\hat{\beta}_k-\beta)+\frac{1}{k}\left(\frac{\partial^2 L_k(\tilde{\boldsymbol{\Phi}}_k)}{\partial\beta^2}-\frac{\partial^2 L_k(\boldsymbol{\Phi}_0)}{\partial\beta^2}\right)(\hat{\beta}_k-\beta)+\frac{1}{k}\left(\frac{\partial^2 L_k(\boldsymbol{\Phi}_0)}{\partial\beta^2}+km_{\beta\beta}\right)(\hat{\beta}_k-\beta).$$

当 $k\to\infty$ 时,

$$\frac{1}{k}\left(\frac{\partial^2 L_k(\tilde{\boldsymbol{\Phi}}_k)}{\partial\beta^2}-\frac{\partial^2 L_k(\boldsymbol{\Phi}_0)}{\partial\beta^2}\right)(\hat{\beta}_k-\beta)\to 0,\frac{1}{k}\left(\frac{\partial^2 L_k(\boldsymbol{\Phi}_0)}{\partial\beta^2}+km_{\beta\beta}\right)(\hat{\beta}_k-\beta)\to 0.$$

故 $\lim\limits_{k\to\infty}\left(\frac{1}{k}\frac{\partial L_k(\boldsymbol{\Phi}_0)}{\partial\beta}-m_{\beta\beta}(\hat{\beta}_k-\beta)\right)=0$, 所以 $\limsup\limits_{k\to\infty}\frac{\sqrt{km_{\beta\beta}}|\hat{\beta}_k-\beta|}{\sqrt{2\ln\ln k}}=1$ a.s., 由文献 [27] 性质 3.2 命题得证. 同理可证前两式.

推论 2 对任意 r, 使得 $0<r<1-2\alpha^{-1}$, 有

$$\lim_{k\to\infty}k^r\left[\frac{1}{k}\begin{pmatrix}\frac{\partial L_k(\boldsymbol{\Phi}_0)}{\partial\theta}\\ \frac{\partial L_k(\boldsymbol{\Phi}_0)}{\partial\alpha}\\ \frac{\partial L_k(\boldsymbol{\Phi}_0)}{\partial\beta}\end{pmatrix}^{\mathrm{T}}\boldsymbol{M}^{-1}\begin{pmatrix}\frac{\partial L_k(\boldsymbol{\Phi}_0)}{\partial\theta}\\ \frac{\partial L_k(\boldsymbol{\Phi}_0)}{\partial\alpha}\\ \frac{\partial L_k(\boldsymbol{\Phi}_0)}{\partial\beta}\end{pmatrix}-k\begin{pmatrix}\hat{\theta}_k-\theta\\ \hat{\alpha}_k-\alpha\\ \hat{\beta}_k-\beta\end{pmatrix}^{\mathrm{T}}\boldsymbol{M}\begin{pmatrix}\hat{\theta}_k-\theta\\ \hat{\alpha}_k-\alpha\\ \hat{\beta}_k-\beta\end{pmatrix}\right]=0\quad\text{a.s.}\tag{18}$$

证明 $\boldsymbol{M}$ 正定, 必存在矩阵 $\boldsymbol{Q}$ 使 $\boldsymbol{M}=\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{Q}$, 要证 (18) 式, 只需证 $\lim_{k\to\infty}\boldsymbol{q}_k\boldsymbol{q}_k^{\mathrm{T}}=0$. 其中,

$$\boldsymbol{q}_k=\sqrt{\ln\ln k}\left[\frac{1}{\sqrt{k}}\left(\frac{\partial L_k(\boldsymbol{\Phi}_0)}{\partial\theta},\frac{\partial L_k(\boldsymbol{\Phi}_0)}{\partial\alpha},\frac{\partial L_k(\boldsymbol{\Phi}_0)}{\partial\beta}\right)\boldsymbol{Q}^{-1}+\sqrt{k}(\hat{\theta}_k-\theta,\hat{\alpha}_k-\alpha,\hat{\beta}_k-\beta)\boldsymbol{Q}^{\mathrm{T}}\right],$$

$$\boldsymbol{q}_k^{\mathrm{T}}=\frac{k^r}{\sqrt{\ln\ln k}}\left[\frac{1}{\sqrt{k}}(\boldsymbol{Q}^{\mathrm{T}})^{-1}\begin{pmatrix}\frac{\partial L_k(\boldsymbol{\Phi}_0)}{\partial\theta}\\ \frac{\partial L_k(\boldsymbol{\Phi}_0)}{\partial\alpha}\\ \frac{\partial L_k(\boldsymbol{\Phi}_0)}{\partial\beta}\end{pmatrix}-\sqrt{k}\boldsymbol{Q}\begin{pmatrix}\hat{\theta}_k-\theta\\ \hat{\alpha}_k-\alpha\\ \hat{\beta}_k-\beta\end{pmatrix}\right].$$

当 $k\to\infty$, $\frac{\boldsymbol{q}_k}{\sqrt{\ln\ln k}}=O(\sqrt{\ln\ln k})$ a.s., 由 (15) 式可知 $\lim\limits_{k\to\infty}\boldsymbol{q}_k^{\mathrm{T}}=\boldsymbol{0}$ a.s.

定理 2　对于任意 r, 使 $0<r<1-2\alpha^{-1}$, 当 $k\to\infty$ 时, 有

$$k^r\left[2\left(L_k(\hat{\boldsymbol{\Phi}}_k)-L_k(\boldsymbol{\Phi}_0)\right)-\frac{1}{k}\left(\frac{\partial L_k(\boldsymbol{\Phi}_0)}{\partial\theta},\frac{\partial L_k(\boldsymbol{\Phi}_0)}{\partial\alpha},\frac{\partial L_k(\boldsymbol{\Phi}_0)}{\partial\beta}\right)\boldsymbol{M}^{-1}\begin{pmatrix}\frac{\partial L_k(\boldsymbol{\Phi}_0)}{\partial\theta}\\ \frac{\partial L_k(\boldsymbol{\Phi}_0)}{\partial\alpha}\\ \frac{\partial L_k(\boldsymbol{\Phi}_0)}{\partial\beta}\end{pmatrix}\right]\to 0\quad \text{a.s.}$$

类似 Rencová[30], 利用 Taylor 展开方法与推论 2, 定理即证.

定理 3　检验问题 (14) 在 H_0 下, 对任意 $t,\alpha>2$ (即 $-0.5<\gamma<0$), 有

$$\lim_{n\to\infty}P\left(A(\ln n)\sqrt{\max_{1\leqslant\tau\leqslant n-1}(2\ln\Lambda_\tau)}\leqslant t+D_3(\ln n)\right)=\exp\{-2\mathrm{e}^{-t}\}.$$

式中, $A(x)=\sqrt{2\ln x}$, $D_d(x)=2\ln x+(0.5d)\ln\ln x-\ln\Gamma(0.5d)$, $d=3$ 表示参数维度, 下同.

证明　利用定理 2 可证

$$\left|\max_{1\leqslant k\leqslant n}\left[2\left(L_k(\hat{\boldsymbol{\Phi}}_k)-L_k(\boldsymbol{\Phi}_0)\right)\right]-\max_{1\leqslant k\leqslant n}\frac{1}{k}\left(\frac{\partial L_k(\boldsymbol{\Phi}_0)}{\partial\theta},\frac{\partial L_k(\boldsymbol{\Phi}_0)}{\partial\alpha},\frac{\partial L_k(\boldsymbol{\Phi}_0)}{\partial\beta}\right)\boldsymbol{M}^{-1}\begin{pmatrix}\frac{\partial L_k(\boldsymbol{\Phi}_0)}{\partial\theta}\\ \frac{\partial L_k(\boldsymbol{\Phi}_0)}{\partial\alpha}\\ \frac{\partial L_k(\boldsymbol{\Phi}_0)}{\partial\beta}\end{pmatrix}\right|=o_p(\ln\ln n).$$

式中, $A_k=o_p(a_k)$ 表示序列 A_k 与常数列 a_k 比值的极限依概率收敛于 0, 即对任意小 $\epsilon>0$, 使 $\lim\limits_{k\to\infty}P(|\frac{A_k}{a_k}|<\epsilon)=1$ 成立. 同时, 注意到变点 τ 是 k 的某取值, 由此容易得到定理结论, 具体过程类似 Csörgö[26] 证明.

对于 $\gamma>0$ 的 GPD 变点检测问题, 根据 (13) 式 $g_2(x;\theta,\alpha,\beta)$ 定义, 易知其对数似然函数各阶偏导不存在间断点, 利用 $\gamma<0$ 方法, 也有定理 3 结果. 此外, 当 $\gamma=0$ 时, GPD 退化为指数分布, 上述结论依然成立[26]. 从而, 下述定理成立.

定理 4　检验问题 (11), H_0 下, 对于任意 t, $\gamma>-0.5$ 有

$$\lim_{n\to\infty}P\left(A(\ln n)\sqrt{\max_{1\leqslant m\leqslant n-1}(2\ln\Lambda_m)}\leqslant t+D_3(\ln n)\right)=\exp\{-2\mathrm{e}^{-t}\}.$$

式中, $A(x)=\sqrt{2\ln x}$, $D_d(x)=2\ln x+(0.5d)\ln\ln x-\ln\Gamma(0.5d)$.

3.3　GPD 变点检测模型算法实现步骤

利用 GPD 变点检测模型实现重现期预测的具体步骤如下.

Step 1　输入 n 个观测值, 并初始化常数 n_0, 论文案例取 $n_0=5$.

Step 2　给定检验水平 α, 依据定理 4 模拟获得单参数和两参数临界值点 $x_\alpha^{(1)},x_\alpha^{(2)}$ (注: 一般情况 $x_\alpha^{(2)}>x_\alpha^{(1)}$).

Step 3　从 n_0 到 $n-n_0$ 循环计算 3.1 节中的 $\Lambda_{n_0},\Lambda_{n_0+1},\cdots,\Lambda_m,\cdots,\Lambda_{n-n_0}$.

Step 4　计算 z_m 并判断 z_m 落点区域:

Step 4.1　若 $x_\alpha^{(1)}\leq z_m<x_\alpha^{(2)}$, 则存在单参数变点;

Step 4.2　若 $z_m\geq x_\alpha^{(2)}$, 则存在两参数变点;

Step 4.3　若 $z_m<x_\alpha^{(1)}$, 则无法拒绝原假设, 即未检测到变点;

Step 5　若 Step 4.1 或 Step 4.2 成立, 基于二分法, 更新检测数据的边界 n_0 和 $n-n_0$, 重复 Step 3∼Step 5.

在 GPD 变点检测模型中, 根据上述步骤检测出变点, 基于变点预测判断可能发生的异常情况.

4　案例分析 ——GPD 变点方法在水文气象中的应用

深圳市地处广东省南部, 由于深受季风的影响, 夏季盛行偏东南风, 时有季风低压和热带气流光顾, 高温多雨. 易受暴雨风暴潮等类似自然灾害侵袭, 因此对该地区历年各月最高降雨量进行 GPD 变点建模, 预测给定重现期下未来可能出现的极端降雨水平, 并给出更为合理的估计是非常必要的 [31].

表 1　深圳市降雨量的基本统计量及其正态性检验结果

n	最大值	最小值	均值	标准差	中位数	偏度	峰度	W	p-value
660	385.80	0.00	64.62	65.46	46.10	1.65	3.24	0.84	$<2.2\times10^{-16}$

4.1　数据资料

数据源于深圳 1960 年 01 月至 2014 年 12 月气象数据, 分析对象为月最大 24 小时降雨量 (1975 年、1977 年和 1980 年数据缺失, 分别采用上一年数据进行填补, 单位: mm/24 h). 从表 1 可看出, 偏度 1.65 大于 0, 表

明深圳市降雨量数据具有右偏性, 峰度值 3.24 大于标准正态分布下峰度值 3, 表明具有尖峰特点. 由 Shapiro-Wilk 统计量 p 值小于 2.2×10^{-16} 正态性检验结果, 说明该数据并不服从正态分布. 降雨量时序如图 1 和图 2 所示.

经初步探索性分析, 可得如下结论:

1) 深圳全年降雨量主要集中在 04–09 月份, 约占当年 80% 以上;

2) 1981 年至 1992 年之间出现月最大降雨量大于 (或等于) 150 mm 的暴雨现象明显低于其他年份, 若仍用常规 GPD 模型进行重现期预测 [32], 可能导致较大偏差;

3) 如果变点存在, 变点位置极有可能处于图 2 中所标出的两条黑色线段之间的区域.

因此, 应用基于 GPD 模型变点探寻深圳市历年降雨周期特征及分析未来给定重现期下的降雨量.

4.2 阈值选择与变点检测

根据 2.2 节可知, 过 $(x_{(n-k)}, k^{-1}\sum_{i=1}^{k}(x_{(n-k+i)}-x_{(n-k)}))$ 绘制阈值 u 的超额均值函数 $e(u)$ 如图 3 所示. 从图判断阈值 $u\in[50,200)$ 都是合适的, 但还不能得到一个相对较小的选择范围, 因此, 利用 2.2 节选择方法再次进行阈值探索. 绘制形状和修正尺度参数 [1] 估计值随阈值变化如图 4 所示. 图 4 左图表示形状参数估计值与阈值关系, 右图则表示经修正后的尺度参数估计值与阈值关系. 考虑到变点检测过程的可操作性与结果可靠性, 结合图 4 看出, 阈值 $u=50$ 处较为合适, 下面进行变点存在性检验及位置诊断.

使用 NOT 方法 [19] 检测变点存在性. NOT 方法基本思想: 数据长度 L 内随机生成 B 个子区间 $(s_b,e_b]$ $(b=1,\cdots,B)$, 确定阀值 $u_L=K\sqrt{(2\ln L)}$, K 为常数, 计算统计量

$$C^m_{(s_b,e_b]}(\boldsymbol{x})=|<\boldsymbol{x},\phi^m_{(s_b,e_b]}>|.$$

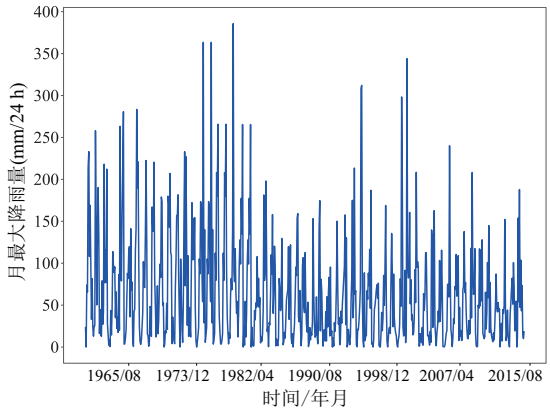


图 1 深圳市历年月最大降雨量条形图

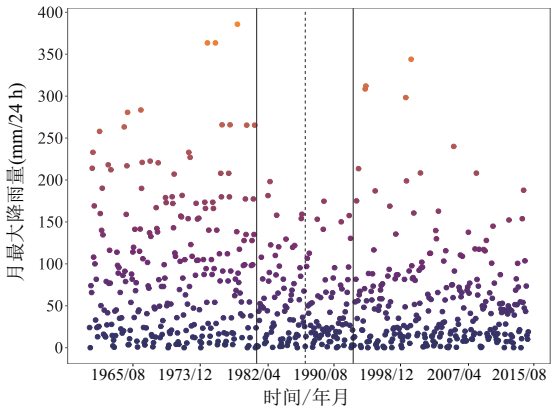


图 2 深圳市历年月最大降雨量散点图

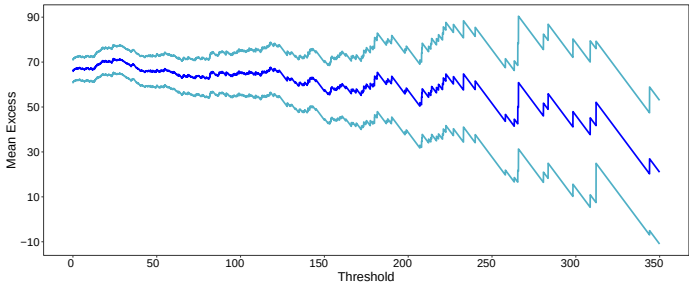


图 3 超额均值函数曲线图

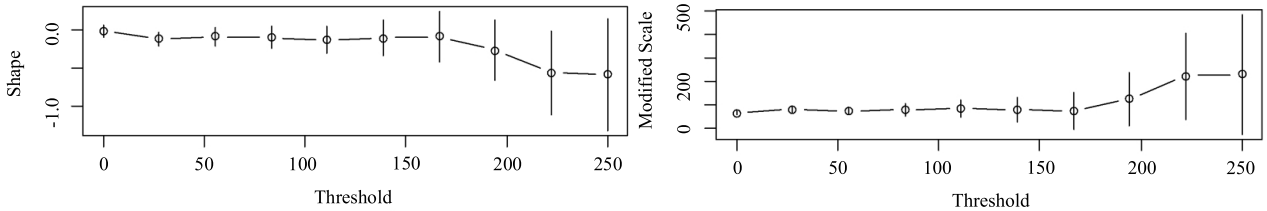


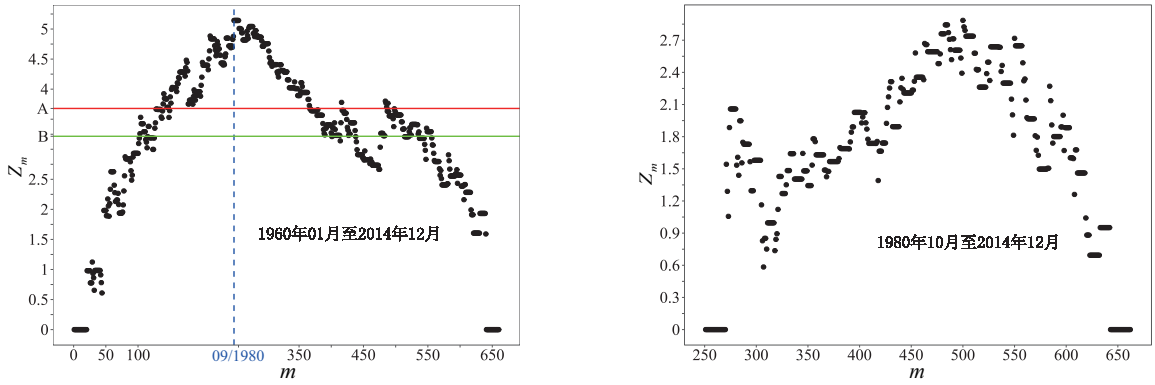
图 4 形状参数估计值和修正后尺度参数估计值与阈值关系图

式中, $<\cdot,\cdot>$ 为内积, $m=s_b+1,\cdots,e_b, \boldsymbol{x}=(x_{s_b+1},\cdots,x_{e_b})^{\mathrm{T}}$, $\phi^m_{(s_b,e_b]}=(\phi^m_{(s_b,e_b]}(s_b+1),\cdots,\phi^m_{(s_b,e_b]}(e_b))^{\mathrm{T}}$,

$$\phi^m_{(s_b,e_b]}(t)=\begin{cases}\dfrac{(e_b-m)^{\frac{1}{2}}}{[(e_b-s_b)(m-s_b)]^{\frac{1}{2}}}, & t=s_b+1,\cdots,m,\\ \dfrac{(m-s_b)^{\frac{1}{2}}}{[(e_b-s_b)(e_b-m)]^{\frac{1}{2}}}, & t=m+1,\cdots,e_b.\end{cases}$$

当 $\max_{s_b<m\le e_b} C^m_{(s_b,e_b]}(\boldsymbol{x})>u_L$ 时 m 可能为变点, 并令 $m^b=\arg\max_{s_b<m\le e_b} C^m_{(s_b,e_b]}(\boldsymbol{x})$, 得可能变点集 $S=S\bigcup m^b$, 用 SIC 准则筛选 S, 最终确定变点集 S^*. 本数据案例长度 $L=660$, 取 $B=1000$ 和 $K=1$, NOT 变点检测方法最终确定变点集 $S^*=\{249\}$, 即数据为单变点, 变点位置 $m^*=249$ (对应 1980 年 09 月).

按本文构建 GPD 变点模型, 计算不同 m 值所对应统计量 Z_m 并绘制图 5 关系图. 图 5(a) 中直线 A 表示两参数 GPD 变点下的临界值参考线 (3.679), 直线 B 表示单参数 GPD 变点下的临界值参考线 (3.216). 经计算得 $Z_n=5.143$ 大于 3.679 (直线 A 临界值), 拒绝原假设, 故变点存在且 m^* 也是 249, 即对应的也是 1980 年 09 月. 此外, 为验证数据是否仍存在其它变点, 运用 "二分法", 继续运用论文 GPD 变点检测方法进行探索. 结果发现, 1960 年 01 月至 1980 年 09 月与 1980 年 10 月至 2014 年 12 月均不存在其他变点. 图 5(b) 表示后者的检测结果, 图 5(b) 也显示, 并没有超过临界值 Z_m 的值出现. 检测结果显示, 数据仅存在单变点 (1980 年 09 月).



(a)　　(b)

图 5　深圳市月最大降雨量数据下的 Z_m 统计量随 m 变化图

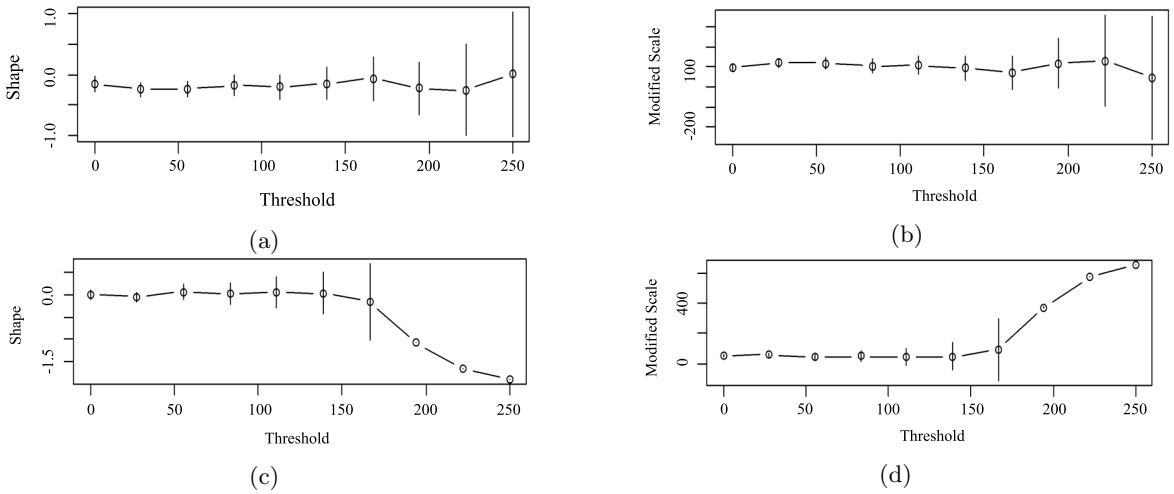
4.3　方法比较和变点诊断

论文 GPD 模型的变点位置检测结果尽管与文献 [27] 略有差异, 但差别不大, 都在探索性范围 04–09 月内, 且变点数目均为单变点, 并且 GPD 模型与较新的 NOT 方法检测结果一致. 使用模型虽受缺失填补随机等因素的影响, 但结合实际分析, 符合数据客观情况. GPD 变点模型研究方法的优势在于不仅能检测变点, 还能对检测结果进行进一步深入分析. 下面将以 GPD 模型检测到的变点作为分界点, 对原始数据分组拟合 GPD 模型和预测性建模分析.

4.4　参数估计和模型检验

重复 4.2 节阈值选择, 对分组数据确定阈值 (如图 6 所示). 图 6 上面的两图表示 1960 年 01 月至 1980 年 09 月数据所得结果, 而图 6 下面的两图则表示 1980 年 10 月至 2014 年 12 月数据所得结果. 由图 6 看出, 参数值在阈值 51 mm 至 140 mm 之间走势类似直线变化, 故确定阈值在 51 mm 处较为合适. 利用极大似然方法分别得完全样本 (1960 年 01 月至 2014 年 12 月)、组 1 (1960 年 01 月至 1980 年 09 月) 和组 2 (1980 年 10 月至 2014 年 12 月) 下参数估计及拟合优度检验 (cramer-von mises 统计量 W^2) 结果见表 2. 其中, 标准差 (Sd) 和 95% 置信区间上限 LCL 和下限 UCL, 均由极大似然估计的渐近正态性质并结合基于样本的 Fisher 信息矩阵计算所得 [1].

表 2 中 Cramer-von Mises 检验 p 值都大于 0.05, 表明在此检验水平下均无充分证据显示这些样本不是来自 GPD. 为了更直观看到 GPD 的拟合效果, 分别给出三种分组下的数据拟合结果, 图 7(a)~(c) 分别表示完全样本、组 1 和组 2 拟合效果图. 其中实线表示理论概率值, 散点表示经验分布值. 从图 7 所示可知, 三种分组下的 GPD 拟合结果均较好.

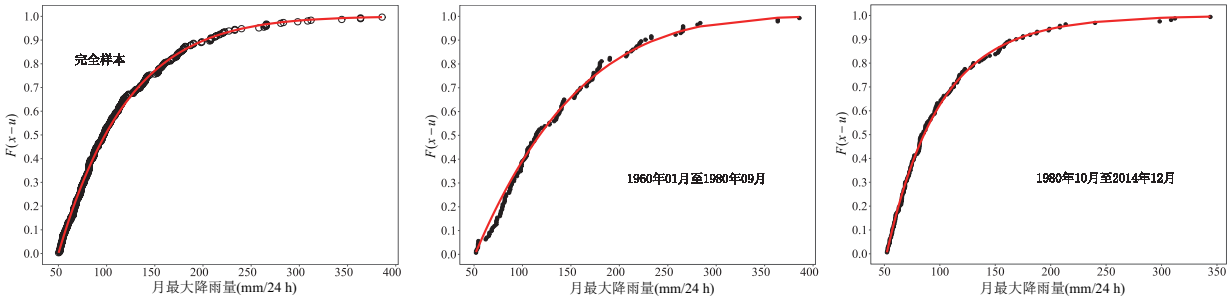


(a) (b) (c) (d)

图 6 变点前后形状参数估计值和修正后尺度参数估计值与阈值关系图

表 2 三种分组下 GPD 参数估计及模型检验

分组	估计参数	Estimation (Sd)	LCL (95%)	UCL (95%)	W^2	p-value
1960.01	$\hat{\mu}$	52.20(14.17)	22.43	77.97		
至	$\hat{\sigma}$	71.80(5.75)	60.54	83.07	0.05	0.47
2014.12	$\hat{\gamma}$	−0.08(0.06)	−0.19	0.03		
1960.01	$\hat{\mu}$	52.00(15.07)	22.47	81.53		
至	$\hat{\sigma}$	103.90(10.86)	82.61	125.19	0.09	0.55
1980.09	$\hat{\gamma}$	−0.23(0.07)	−0.36	−0.11		
1980.10	$\hat{\mu}$	52.30(17.35)	18.29	86.31		
至	$\hat{\sigma}$	47.78(5.58)	36.86	58.71	0.02	0.39
2014.12	$\hat{\gamma}$	0.05(0.09)	−0.12	0.22		



(a) (b) (c)

图 7 深圳市历年月最大降雨量拟合效果图

4.5 模型预测

从表 2 的参数估计结果看到, 完全样本与组 1 对应的形状参数小于 0, 而组 2 样本下的形状参数大于 0. 可初步断定, 20 世纪 80 年代后的数据尾部较 20 世纪 60 年代至 80 年代的数据尾部更厚. 为验证此结论, 作出三种分组下的概率密度拟合效果如图 8 所示. 内嵌图清晰表明 20 世纪 80 年代以来, 450 mm 以上极端暴雨现象较之前概率高.

为了验证结论的可靠性, 避免一次估计所带来的误差, 分别对组 1 和组 2 考虑不同阈值 (每次选取顺序统计量 $x_{(n-k)}$ 作为阈值) 下形状参数极大似然估计的变化情况, 所得结果如图 9 所示. 从图 9 看出, 基于 1980 年 10 月至 2014 年 12 月数据的形状参数估计结果基本上均比 1960 年 01 月至 1980 年 09 月的大, 后者基本都在 0 以下. 利用 (9) 式可得关于重现期与降雨量关系以及三种分组下的对比情况, 如图 10 与图 11 所示.

图 10(a)(b) 分别表示完全样本和 1960 年 01 月至 1980 年 09 月数据重现期与降雨量的关系图. 图 11

则表示 1980 年 10 月至 2014 年 12 月重现期与重现期下的降雨量即重现水平的关系图. 从图 10 和图 11 可以看出, 图 11 比图 10(b) 拟合效果好. 图 10(a) 和图 10(b) 降雨量随着重现期的增加, 增长速度有放缓趋势, 对于这种趋势, 图 10(b) 更加明显, 而图 11 所示则有加快趋势.

图 12 表示三种分组下, 降雨量与重现期关系图. 由图 12 可看到, 在基于 1980 年 10 月至 2014 年 12 月的预测下, 遭遇一百年以下重现期的暴雨灾害, 所带来的降雨量要低于另外两种情形, 而遭遇一百年以上重现期的暴雨灾害, 结论则相反. 结合表 2 尺度参数估计结果, 组 1 (103.90) 和完全样本 (71.80) 的值大于组 2 (47.78) 的值, 大致表明 20 世纪 80 年代以来, 极端暴雨现象更加频繁, 应该警惕更加极端气候事件发生.

此外, 图 12 中完全样本下重现期与降雨量关系曲线正好处于另外两种情形的中间位置, 更加表明 20 世纪 80 年代以来的暴雨气候现象规律发生了改变, 应该引起人们的重视.

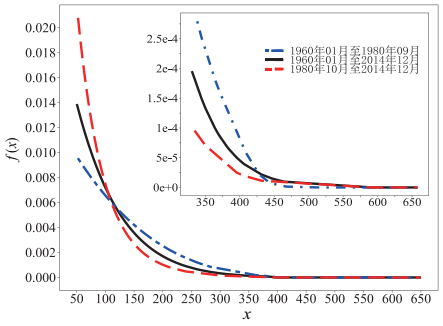


图 8　三种分组下概率密度拟合图

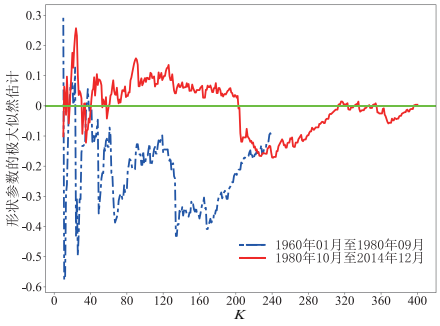
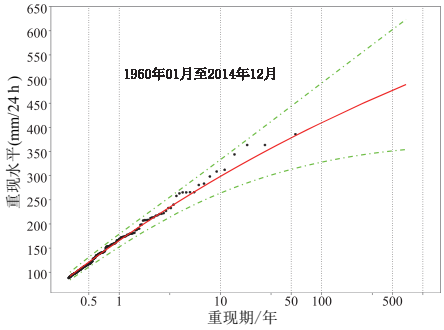
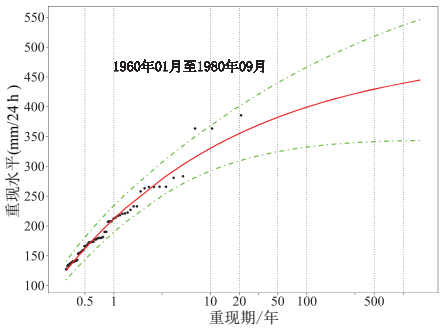


图 9　不同 k 值的 GPD 形状参数估计结果



(a)　完全样本的降雨量与重现期关系



(b)　组 1 的降雨量与重现期关系

图 10　降雨量与重现期关系图

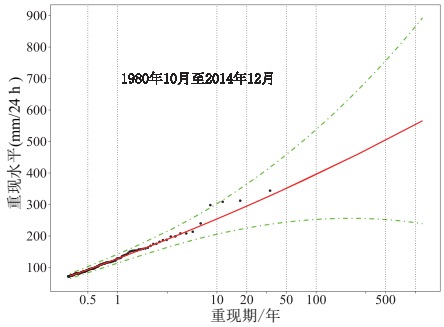


图 11　组 2 降雨量与重现期关系

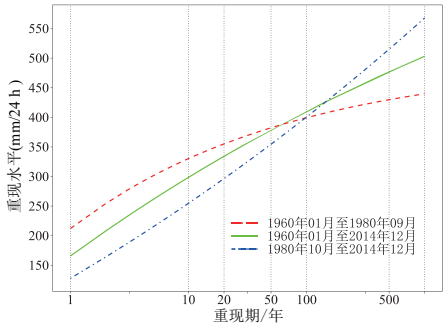


图 12　三种分组下降雨量与重现期关系

5　结论

针对极端事件风暴潮数据流, 提出了 GPD 变点检验问题与检测预测模型. 构造了检验三参数 GPD 变点检测与诊断的极大化似然比检验统计量. 通过推导得 GPD 似然函数不连续点性质, 得出 GPD 随机变量收敛条件和证明顺序统计量相关收敛结论. 在此基础上, 对似然函数及其估计的相关极限性质进行论证. 通过与 NOT 方法比较, 结果表明, 所构建的检验统计量具有较好功效. 实证分析结果表明, 无论有无变点存在, 基于变点方法的 GPD 预测模型具有传统 GPD 所不具备的优势.

论文还有待作如下深入研究: 混合 GPD 模型, 就 GPD 模型而言, 通过利用论文提出的变点方法, 进而构造混合 GPD 模型是可行的探讨方向, 将变点方法与混合 GPD 结合进行实际数据拟合, 对于 GPD 的分位点估计或许更加精准; 含参变量的 GPD 变点模型, 就 GPD 参数而言, 可进一步研究参数含协变量的 GPD 变点模型, 从而更加完善 GPD 处理实际问题的手段.

参考文献

[1] Coles S, Bawa J, Trenner L, et al. An introduction to statistical modeling of extreme values[M]. London: Springer, 2001.

[2] Chen B Y, Zhang K Y, Wang L P, et al. Generalized extreme value-Pareto distribution function and its applications in ocean engineering[J]. China Ocean Engineering, 2019, 33(2): 127–136.

[3] Kiriliouk A, Rootzén H, Segers J, et al. Peaks over thresholds modeling with multivariate generalized Pareto distributions[J]. Technometrics, 2019, 61(1): 123–135.

[4] Luo Y, Zhu L S. A new model of peaks over threshold for multivariate extremes[J]. China Ocean Engineering, 2014, 28(6): 761–776.

[5] Susan M, Waititu A G, Mwita P N, et al. Consistency of the φ-divergence based change point estimator[J]. Open Journal of Statistics, 2020, 10(5): 832–849.

[6] Park M H, Kim J H T. Estimating extreme tail risk measures with generalized Pareto distribution[J]. Computational Statistics & Data Analysis, 2016, 98: 91–104.

[7] Mo C X, Ruan Y L, He J Q, et al. Frequency analysis of precipitation extremes under climate change[J]. International Journal of Climatology, 2019, 39(12): 1373–1387.

[8] 刘新红, 孟生旺, 李政宵. 地震损失风险的 Copula 混合分布模型及其应用 [J]. 系统工程理论与实践, 2019, 39(7): 1855–1866.
Liu X H, Meng S W, Li Z X. Copula-mixed distribution model and its application in modeling earthquake loss in China[J]. Systems Engineering — Theory & Practice, 2019, 39(7): 1855–1866.

[9] 崔玫意, 张玉虎, 陈秋华. Box-Cox 正态分布及其在降雨极值分析中的应用 [J]. 数理统计与管理, 2017, 36(1): 8–17.
Cui M Y, Zhang Y H, Chen Q H. Box-Cox normal distribution and its application in rainfall extreme value[J]. Journal of Applied Statistics and Management, 2017, 36(1): 8–17.

[10] 胡立伟, 杨锦青, 何越人, 等. 城市交通拥塞辐射模型及其对路网服务能力损伤研究 [J]. 中国公路学报, 2019, 32(3): 145–154.
Hu L W, Yang J Q, He Y R, et al. Urban traffic congestion radiation model and damage caused to service capacity of road network[J]. China Journal of Highway and Transport, 2019, 32(3): 145–154.

[11] Wu Y, Randell D, Christou M, et al. On the distribution of wave height in shallow water[J]. Coastal Engineering, 2016, 111: 39–49.

[12] Ashkar F, Adlouni E S. Adjusting for small-sample non-normality of design event estimators under a generalized Pareto distribution[J]. Journal of Hydrology, 2015, 530: 384–391.

[13] Chen S, Li Y X, Kim J, et al. Bayesian change point analysis for extreme daily precipitation[J]. International Journal of Climatology, 2017(37): 3123–3137.

[14] Renard B, Lang M, Bois P. Statistical analysis of extreme events in a nonstationary context via a Bayesian framework: Case study with peak-over-threshold data[J]. Stochastic Environmental Research & Risk Assessment, 2006, 21(2): 97–112.

[15] Meng R. Growth curve analysis and change-points detection in extremes[D]. Jeddah: King Abdullah University of Science and Technology, Thuwal, Kingdom of Saudi Arabia, 2016.

[16] Dierckx G, Teugels J L. Change point analysis of extreme values[J]. Environmetrics, 2010, 21(7–8): 661–686.

[17] Raimondo M, Tajvidi N. A peaks over threshold model for change-point detection by wavelets[J]. Statistica Sinica, 2004: 395–412.

[18] Safari M A M, Masseran N, Ibrahim K. Outliers detection for Pareto distributed data[C]// Proceedings of the University Kebangsaan Malaysia, Faculty of Science and Technology 2017 Postgraduate Colloquium, 2017.

[19] Baranowski R, Chen Y N, Fryzlewicz P. Narrowest-over-threshold detection of multiple change points and change-point-like features[J]. Journal of the Royal Statistical Society, Statistical Methodology, Series B, 2019, 81(3): 649–672.

[20] 李鑫鑫, 桑燕芳, 谢平, 等. 基于离散小波分解的水文随机过程平稳性检验方法 [J]. 系统工程理论与实践, 2018, 38(7): 1897–1904.
Li X X, Sang Y F, Xie P, et al. A method for testing the stationarity of stochastic hydrological process based on discrete wavelet transform[J]. Systems Engineering — Theory & Practice, 2018, 38(7): 1897–1904.

[21] 叶文, 王红端, 许新谊, 等. 不同实际尺度选样法的洪水频率分析 [J]. 系统工程理论与实践, 2017, 37(2): 535–544.
Ye W, Wang H R, Xu X Y, et al. Flood frequency analysis based on different time scales sampling method[J].

Systems Engineering — Theory & Practice, 2017, 37(2): 535–544.
[22] 高峰, 朱川林, 李昊, 等. 海口湾海洋水动力自然灾害评估分析 [J]. 中国海洋大学学报, 2019, 49(9): 130–138.
Gao F, Zhu C L, Li H, et al. Analysis on risk assessment of marine hydrodynamics for Haikou bay, Hainan[J]. Periodical of Ocean University of China, 2019, 49(9): 130–138.
[23] 高超, 汪丽, 陈财, 等. 海平面上升风险中国大陆沿海地区人口与经济暴露度 [J]. 地理学报, 2019, 74(8): 1590–1604.
Gao C, Wang L, Chen C, et al. Population and economic risk exposure in coastal region of China under sea level rise[J]. Acta Geographica Sinica, 2019, 74(8): 1590–1604.
[24] 刘家宏, 李泽锦, 梅超, 等. 基于 TELEMAC — 2D 的不同设计暴雨下厦门岛城市内涝特征分析 [J]. 科学通报, 2019, 64: 2055–2066.
Liu J H, Li Z J, Mei C, et al. Urban flood analysis for different design storm hyetographs in Xiamen island based on TELEMAC — 2D[J]. China Science Bulletin, 2019, 64: 2055–2066.
[25] Jarušková D. Maximum log-likelihood ratio test for a change in three parameter Weibull distribution[J]. Journal of Statistical Planning and Inference, 2007, 137(6): 1805–1815.
[26] Csörgö M, Horváth L. Limit theorems in change-point analysis[M]. New York: John Wiley & Sons Inc, 1997.
[27] 谌业文. 广义 Pareto 分布变点问题及其应用研究 [D]. 贵阳: 贵州大学, 2016.
Chen Y W. The research on change-point problem for generalized Pareto distribution and its application[D]. Guiyang: Guizhou University, 2016.
[28] Smith R L. Maximum likelihood estimation in a class of nonregular cases[J]. Biometrika, 1985, 72(1): 67–90.
[29] 胡尧, 谌业文. 广义 Pareto 分布变点检测似然比模型 [J]. 应用数学学报, 2021, 44(4): 553–573.
Hu Y, Chen Y W. A likelihood ratio model for change point detection of generalized Pareto distribution[J]. ACTA Mathematicae Applicatae Sinica, 2021, 44(4): 553–573.
[30] Rencová M. Change-point detection in temperature series[D]. Prague: Czech Technical University, 2009.
[31] 蔡志杰. 深圳洪灾风险评估与洪灾损失预测 [J]. 数学建模及其应用, 2014, 3(4): 66–70.
Cai Z J. Risk assessment and loss prediction of flood disaster in Shenzhen[J]. Mathematical Modeling and Its Applications, 2014, 3(4): 66–70.
[32] 谌业文, 刘圣达, 王旭琴. 洪灾损失的概率统计预测模型 [C]// 2014 年 “深圳杯” 数学建模夏令营论文集. 北京: 高等教育出版社, 2014.
Chen Y W, Liu S D, Wang X Q. Probability statistical prediction model of flood losses[C]// 2014 “Shenzhen Cup” summer camp of mathematical modeling. Beijing: Higher Education Press, 2014.