

洪水频率分析的非参数密度变换模型研究

董 洁, 谢悦波, 颜秉龙

(河海大学水资源环境学院, 江苏 南京 210098)

摘要 以非参数统计理论为基础, 把变换理论和非参数统计理论结合起来, 建立了洪水频率分析的非参数密度变换模型. 通过变换, 改进了非参数核密度估计方法, 避开了窗宽选择问题, 提高了小样本的估计精度.

关键词 非参数变换 洪水频率分析 核估计 窗宽

中图分类号 P333.9 **文献标识码** A **文章编号** 1000-1980(2005)04-0378-04

计算设计值的方法主要有参数法和非参数法 2 种. 参数法数据对假设的模型不具稳健性, 数据即使轻微偏离假设模型, 也可能产生估计误差. 非参数法小样本时估计精度低, 常依靠大样本保证精度^[1]. 本文建立了洪水频率分析的非参数密度变换模型. 该模型避开了窗宽选择问题, 可达到提高小样本估计精度的目的.

1 密度函数估计

核估计是常用的非参数密度估计^[2]. 小样本时核估计性能较差. 其原因是, 估计的偏差和方差对窗宽选择的要求矛盾比较突出, 即核估计的偏差与窗宽成正比, 而方差与窗宽的平方成反比. 另外, 核估计的偏差还和待估计函数的图形有关. 因此, 可考虑一种新的核估计——变换核估计^[3]. 变换核估计是一种迭代过程, 它包括样本变换、新样本估计和函数反变换. 迭代过程中新样本逐步逼近“最佳样本”, 最佳样本对应的最佳函数有较高的估计精度.

设 $X_1, X_2, \dots, X_n$ 是一个来自密度函数为 $f(x)$ 的未知总体的独立同分布随机样本, 则核估计是

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^n k\left(\frac{x - x_i}{h}\right) \quad (1)$$

核函数 $k(\cdot)$ 起窗的作用, 窗宽 h 决定了平滑点数, n 为样本容量. 通常用积分均方误差

$$MISE_x(\hat{f}) = E \int [\hat{f}(x) - f(x)]^2 dx = \int \{E\hat{f}(x) - f(x)\}^2 + \text{var}\hat{f}(x) dx \quad (2)$$

衡量密度估计性能^[4], 所以积分均方误差 $MISE$ 可以对偏差 $\int \{E\hat{f}(x) - f(x)\}^2 dx$ 和方差 $\int [\text{Var}\hat{f}(x)] dx$ 进行综合度量. 要使 $MISE$ 较小, 必须偏差和方差都较小.

对原样本 X_i 进行变换得到新样本 $U_i: U_i = T(X_i) (i = 1, 2, \dots, n)$. 若变换 T 是连续且一一对应函数, 则 U_i 仍是独立同分布样本, 可以先对 U_i 进行密度估计. 设 U_i 的密度函数为 $g(u)$, 其估计为 $\hat{g}(u)$, 则反变换

$$\hat{f}(x) = |T'(x)| \hat{g}[T(x)] \quad (3)$$

即可作为 X_i 的密度估计.

对于未变换核估计, 可以找到最佳密度函数 $f_{\text{opt}}(x)$. 用变分法求积分均方误差 $MISE$ 渐进式的极小值, 即可求得最佳变换函数

$$T(x) = 2\cos\left\{\frac{1}{3}\arccos[1 - 2F(x)] + \frac{2}{3}\pi\right\} \quad (4)$$

变换核估计的迭代算法步骤如下:

a. 变换核估计的初始估计:

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^n k\left(\frac{x-x_i}{h}\right) \quad \hat{F}_1(x) = \int_{-\infty}^x \hat{f}(u) du \tag{5}$$

b. 变换核估计的迭代算法：

$$\left\{ \begin{aligned} T_l(x) &= 2\cos\left\{\frac{1}{3}\arccos[1-2\hat{F}_l(x)] + 240^\circ\right\} \\ |T'_l(x)| &= \hat{f}_l(x)/g[T_l(x)] \\ Y_{l,i} &= T_l(x_i) \\ \hat{g}_l(y) &= \frac{1}{nh} \sum k\left(\frac{y-Y_{l,i}}{h}\right) \\ \hat{f}_{l+1}(x) &= |T'_l(x)| \hat{g}_l[T_l(x)] \\ \hat{F}_{l+1} &= \int_{-\infty}^x \hat{f}_{l+1}(x) dx \end{aligned} \right. \tag{6}$$

迭代次数 $l=1,2,\cdots$.

c. 迭代门限的确定：

$$d(g_l, g) = \int [\hat{g}_l(x) - g(y)]^2 dy \tag{7}$$

门限大小与最佳样本的估计精度一致.

2 洪水频率分析的非参数变换模型

设 n 年的实测资料系列为 $x_1, x_2, \cdots, x_n$, 它可以看作是一个来自密度函数为 $f(x)$ 的未知总体的独立同分布随机样本的观测值. 用变换方法将样本变换为最佳样本, 再按如下方法进行洪水频率分析, 可以得到精度较高的总体密度 $f(x)$ 的估计值 $\hat{f}(x)$ [5~7]. 引进指数核函数

$$k(u) = \frac{1}{2} \lambda e^{-\lambda|u|} \quad (-\infty < u < \infty) \tag{8}$$

它可以看作是随机变量 x 的密度函数. 因为

$$\begin{aligned} Eu &= \int_{-\infty}^{\infty} uk(u) du = 0 \\ Du &= Eu^2 = \int_{-\infty}^{\infty} u^2 k(u) du = \frac{1}{2} \lambda \int_{-\infty}^{\infty} u^2 e^{-\lambda|u|} du = 2/\lambda^2 \end{aligned} \tag{9}$$

所以

$$\lambda = \sqrt{2}/\sqrt{Dx} \tag{10}$$

由核估计定义知, u 与 $\frac{x-x_i}{h_n}$ 有关, 故可用 $\sqrt{Dx} = \sigma_x$ 代替 $\sqrt{Du}$, 即 $\lambda = \sqrt{2}/\sigma_x$. 为计算方便, 取经验窗宽 $h_n = 1.0 \sim 1.2\sigma_x$. 下面讨论设计值的推求过程.

假设给定频率

$$p = \int_{x_p}^{\infty} \hat{f}(x) dx = \frac{1}{n} \sum_{i=1}^n \int_{-\infty}^{\infty} \frac{1}{h_n} k\left(\frac{x-x_i}{h_n}\right) dx = \frac{1}{n} \sum_{i=1}^n E_i(x_p) \quad (0 < p < 1) \tag{11}$$

其中

$$\begin{aligned} E_i(x_p) &= \frac{1}{h_n} \int_{-\infty}^{\infty} k\left(\frac{x-x_i}{h_n}\right) dx = \int_{\frac{x_p-x_i}{h_n}}^{\infty} \lambda e^{-\lambda|u|} du = \\ &\begin{cases} \frac{1}{2} \exp\{-\sqrt{2}(x_p-x_i)/(h_n\sigma_x)\} & (x_p \geq x_i) \\ 1 - \frac{1}{2} \exp\{\sqrt{2}(x_p-x_i)/(h_n\sigma_x)\} & (x_p < x_i) \end{cases} \end{aligned} \tag{12}$$

这样可以用迭代法求出相应频率 p 的 x_p 的估计值 $\hat{x}_p$.

3 模型稳健性分析

为了论证该模型的稳健性, 本文进行了统计试验研究. 假设 2 个不同的总体 P-Ⅲ 型分布和对数正态分

布,对不同的总体参数及不同的样本容量,比较了适线法和非参数密度变换法的稳健性.

3.1 试验设计

选择样本容量 $n = 40, 50, 60$, 取 2 种设计频率 $p_1 = 0.01, p_2 = 0.001$. 为在一定的参数范围内讨论各种估计方法的稳健性, 采用了 3 组总体参数, $C_v = 0.5, 1.0, C_s/C_v = 2, 3$.

3.2 计算方案

理论总体分别采用 P-Ⅲ 型分布和对数正态分布 LN, 认为其生成的系列来自 P-Ⅲ 型分布总体或对数正态分布总体 LN, 并分别用适线法 CFM 和非参数的密度变换法 DEM 估计设计值. 对 2 个总体, 用 2 种估计方法对 3 组参数、3 组样本共 144 种方案进行了比较分析, 计算结果见表 1 和表 2 (只列出样本容量 $n = 60$ 的结果).

3.3 参数估计稳健性评判标准

一般采用设计值对真值的均方误差 σ 和相对均方误差 δ 作为参数估计稳健性的评判标准^[8].

$$\sigma = \sqrt{\frac{\sum_{i=1}^m (\hat{x}_{pi} - x_p)^2}{m}} \quad \delta = \frac{\sigma}{x_p} \times 100\%$$

(13)

式中 $\hat{x}_{pi}$ ——估计的设计频率的设计值; x_p ——已知分布的设计频率的真值.

为了分析设计值对真值的平均偏差程度, 一般采用设计值的绝对误差 θ 和相对误差 ϑ 作为评判标准.

$$\theta = \bar{x}_p - x_p \quad \vartheta = \frac{\theta}{x_p} \times 100\%$$

(14)

$$\bar{x}_p = \frac{1}{m} \sum_{i=1}^m \hat{x}_{pi}$$

式中 $\bar{x}_p$ ——设计频率设计值的均值; $E\hat{x}_p$ ——设计值的均值; m ——统计试验数, $m = 1\,000$ 次.

3.4 统计试验结果

统计试验结果见表 1 和表 2.

表 1 P-Ⅲ 型分布理论总体计算结果 ($p_1 = 0.001, EX = 1\,000, p_2 = 0.01, n = 60$)

Table 1 Calculated results of P-Ⅲ distribution												
C_v	C_s	方法	$E\hat{x}_{p1}$	$\sigma_{x_{p1}}$	$\delta_{x_{p1}}$	$\theta_{x_{p1}}$	$\vartheta_{x_{p1}}$	$E\hat{x}_{p2}$	$\sigma_{x_{p2}}$	$\delta_{x_{p2}}$	$\theta_{x_{p2}}$	$\vartheta_{x_{p2}}$
0.5	1.0	CFM	3 279	396.9	12.16	14	0.43	2 523	242.2	9.64	11	0.44
0.5	1.0	CFM	4 634	1 553.7	47.59	1 369	41.93	3 084	685.5	27.29	572	22.77
0.5	1.0	DEM	4 085	501.0	15.30	820	25.10	2 741	432.0	17.20	229	11.30
1.0	2.0	CFM	6 919	1 108.0	28.85	11	0.16	4 615	660.9	14.35	10	0.22
1.0	2.0	CFM	24 367	19 929.0	18.20	17 459	252.70	9 662	5 747.9	124.80	5 057	109.80
1.0	2.0	DEM	8 145	1 256.0	18.20	1 237	17.90	5 124	894.0	19.40	519	11.30
0.5	1.5	CFM	3 628	470.5	13.01	11	0.30	2 674	284.8	10.69	9	0.34
0.5	1.5	CFM	4 112	789.4	21.82	495	13.69	2 835	382.6	14.36	170	6.38
0.5	1.5	DEM	3 842	521.0	14.40	225	6.20	2 773	341.0	12.80	108	4.10

注: $C_v = 0.5, C_s = 1.0$ 时, $x_{p1} = 3\,265, x_{p2} = 2\,512$; $C_v = 1.0, C_s = 2.0$ 时, $x_{p1} = 6\,908, x_{p2} = 4\,605$; $C_v = 0.5, C_s = 1.5$ 时, $x_{p1} = 3\,617, x_{p2} = 2\,665$.

表 2 对数正态分布理论总体计算结果 ($p_1 = 0.01, EX = 1\,000, p_2 = 0.01, n = 60$)

Table 2 Calculated results of log-normal distribution												
C_v	C_s	方法	$E\hat{x}_{p1}$	$\sigma_{x_{p1}}$	$\delta_{x_{p1}}$	$\theta_{x_{p1}}$	$\vartheta_{x_{p1}}$	$E\hat{x}_{p2}$	$\sigma_{x_{p2}}$	$\delta_{x_{p2}}$	$\theta_{x_{p2}}$	$\vartheta_{x_{p2}}$
0.5	1.0	CFM	3 444	622.6	16.23	-392	-10.22	2 585	305.2	11.42	-87	-3.26
0.5	1.0	CFM	3 984	606.4	15.80	58	1.51	2 760	332.2	12.43	88	3.29
0.5	1.0	DEM	4 231	614.0	16.00	395	10.30	2 988	297.0	11.10	316	11.80
1.0	2.0	CFM	6 752	3 207.5	34.63	-2 511	-27.10	4 404	1 101.4	22.45	-501	-10.21
1.0	2.0	CFM	9 985	2 799.1	30.22	722	7.79	5 169	1 127.8	23.00	264	5.38
1.0	2.0	DEM	10 781	1 234.0	13.30	1 518	16.40	5 645	996.0	20.30	740	15.10
0.5	1.5	CFM	3 444	622.6	16.23	-392	-10.22	2 585	305.2	11.42	-87	-3.26
0.5	1.5	CFM	3 984	606.4	15.80	58	1.51	2 760	332.2	12.44	88	3.29
0.5	1.5	DEM	4 219	581.0	15.20	383	10.00	3 102	301.0	11.30	430	16.10

注: $C_v = 0.5, C_s = 1.0$ 时, $x_{p1} = 3\,836, x_{p2} = 2\,672$; $C_v = 1.0, C_s = 2.0$ 时, $x_{p1} = 9\,263, x_{p2} = 4\,905$; $C_v = 0.5, C_s = 1.5$ 时, $x_{p1} = 3\,836, x_{p2} = 2\,672$.

3.5 估计方法的比较分析

为了分析 P-Ⅲ 型分布和对数正态分布拟合理论总体为 P-Ⅲ 型分布和对数正态分布的稳健性以及非参

数统计法的稳健性,表 1 和表 2 分别统计了设计值与真值的均方误差、均方相对误差,设计值均值的绝对误差和相对误差.

3.5.1 理论总体为 P-Ⅲ 型分布

用 P-Ⅲ 型分布曲线拟合 P-Ⅲ 型分布总体,百年一遇和千年一遇的设计值相对误差一般不超过 1%,百年一遇的相对均方误差不超过 15%,千年一遇的相对均方误差不超过 20%;对数正态分布拟合 P-Ⅲ 型分布总体适线的稳定性较差,百年一遇和千年一遇的设计值相对误差会超过 20%,百年一遇和千年一遇的相对均方误差超过 25%;用非参数的密度变换法 DEM 估计设计值,百年一遇和千年一遇的设计值相对误差一般介于以上 2 种情况之间.

3.5.2 理论总体为对数正态分布 LN

用对数正态分布拟合对数正态分布总体,百年一遇和千年一遇的设计值相对误差一般不超过 8%,百年一遇的相对均方误差不超过 15%,千年一遇的相对均方误差不超过 20%;P-Ⅲ 型分布拟合对数正态分布总体适线的稳定性较差,求出的设计值偏小,百年一遇的相对均方误差不超过 15%,千年一遇的相对均方误差不超过 20%;用非参数的密度变换法 DEM 估计设计值,百年一遇和千年一遇的设计值相对误差一般介于以上 2 种情况之间,当样本容量由 40,50 增加到 60 时,相对误差和相对均方误差均会减小,说明样本容量增加时稳定性也会增加.

4 结 语

本文建立了洪水频率分析的密度变换模型,并用统计试验方法考察了模型的稳健性.结果表明,该模型稳健性很好.

参考文献：

[1] 陈希孺.非参数统计教程[M].上海 华东师范大学出版社,1993.80—89.
[2] KYANG J. Comparative study of flood quantities estimation by nonparametric modes[J]. Journal of Hydrology,2002,260 :176—193.
[3] 但尧,丁鹭飞. 非参数变换核估计[J]. 数理统计与应用概率,1989,4(1) :13—23.
[4] GUO Shen-lian. Nonparametric variable kernel estimation with historical floods and paleoflood information [J]. Water Resources Research, 1991,27(1) 91—98.
[5] ADAMOWSKI K. Regional analysis of annual maximum and partial duration flood data by nonparametric and L-moment methods[J]. Journal of Hydrology,2000,229 219—231.
[6] YUE S. The gumbel mixed model for flood frequency analysis [J]. Journal of Hydrology,1999,226 38—100.
[7] GOEL N K. A derived flood frequency distribution for correlated rainfall intensity and duration[J]. Journal of Hydrology,2000,228 36—67.
[8] 黄振平.水文统计[M]. 南京 河海大学出版社,2003.55—60.

Study on nonparametric transformed kernel estimate model for flood frequency analysis

DONG Jie, XIE Yue-bo, YAN Bing-long

(College of Water Resources and Environment, Hohai University, Nanjing 210098, China)

Abstract By combination of the transformation theory with nonparametric statistical theory, the nonparametric transformed kernel estimate model was developed. Without requirement for selection of window width, the present model improves the precision in small sample estimation. The comparison of the estimated result by the developed model with that by the Monte-Carlo method shows that the nonparametric transformed kernel estimate model is robust.

Key words nonparametric transformation ; flood frequency analysis ; kernel estimate ; window width