ELSEVIER

Journal of Hydrology 260 (2002) 176–193

Journal of Hydrology

www.elsevier.com/locate/jhydrol

Comparative study of flood quantiles estimation by nonparametric models

Kyung-Duk Kim[a,1], Jun-Haeng Heo[b,*]

[a]Korea Infrastructure Safety and Technology Cooperation, Kyunggi-Do, South Korea
[b]Department of Civil Engineering, Yonsei University, Seoul, 120-749, South Korea

Received 11 May 2001; revised 27 November 2001; accepted 17 December 2001

Abstract

There are two basic approaches for estimating flood quantiles: a parametric and a nonparametric method. In this study, the comparisons of parametric and nonparametric models for annual maximum flood data of Goan gauging station in Korea were performed based on Monte Carlo simulation. In order to consider uncertainties that can arise from model and data errors, kernel density estimation for fitting the sampling distributions was chosen to determine safety factors (SFs) that depend on the probability model used to fit the real data. The relative biases of Sheater and Jones plug-in (SJ) are the smallest in most cases among seven bandwidth selectors applied. The relative root mean square errors (RRMSEs) of the Gumbel (GUM) are smaller than those of any other models regardless of parent models considered. When the Weibull-2 is assumed as a parent model, the RRMSEs of kernel density estimation are relatively small, while those of kernel density estimation are much bigger than those of parametric methods for other parent models. However, the RRMSEs of kernel density estimation within interpolation range are much smaller than those for extrapolation range in comparison with those of parametric methods. Among the applied distributions, the GUM model has the smallest SFs for all parent models, and the general extreme value model has the largest values for all parent models considered. © 2002 Elsevier Science B.V. All rights reserved.

Keywords: Flood quantiles estimation; Kernel density; Monte Carlo simulation; Comparison study; Safety factor

1. Introduction

Traditional flood frequency analyses mainly rely on the parametric methods which are based on the assumption that an observed sample comes from a population whose probability density function (PDF) is known. Hence it is important to select an appropriate PDF and to estimate parameters for a given data. If the parametric model is properly applied to a random variable *X*, we can describe the behavior of *X* explicitly. However, the best parametric distribution and parameter estimation methods are still not known despite extensive research into this subject. If we choose a parametric model that is not of appropriate form, then there is a danger of reaching an incorrect conclusion. On the other hand, the nonparametric approach does not make any assumptions regarding a parent distribution. In the event that there is neither knowledge nor experience gained through analysis of previous data sets, it may be desirable to use a nonparametric method.

It was Rosenblatt (1956) and Parzen (1962) who provided a considerable interest in kernel methodology

[*] Corresponding author. Fax: +82-2-2123-2805.
E-mail addresses: kkd@kistec.or.kr (K.-D. Kim), jhheo@yonsei.ac.kr (J.-H. Heo).
[1] Fax: +82-31-910-4179.

PII: S0022-1694(01)00613-8

Table 1
PDF and/or CDF of probability distributions and parameter validity conditions ($\Gamma(\cdot)$: gamma function)

Distribution	PDF and/or CDF and parameter validity conditions
Gamma	$f(x) = \frac{1}{\lvert\alpha\rvert\Gamma(\beta)}\left[\frac{x - x_0}{\alpha}\right]^{\beta-1}\exp\left[-\frac{x - x_0}{\alpha}\right]$ $\alpha > 0$ then $x_0 \le x < \infty$, $\alpha < 0$ then $-\infty < x \le x_0$, $\beta > 0$
GEV	$f(x) = \frac{1}{\alpha}\left[1 - \frac{\beta(x - x_0)}{\alpha}\right]^{1/\beta-1}\exp\left\{-\left[1 - \frac{\beta(x - x_0)}{\alpha}\right]^{1/\beta}\right\}$ $\beta = 0$: GEV-1 $-\infty < x < \infty$, $\beta < 0$: GEV-2 $x_0 + \alpha/\beta \le x < \infty$, $\beta > 0$: GEV-3 $-\infty < x \le x_0 + \alpha/\beta$
Gumbel	$F(x) = \exp\left\{-\exp\left[-\frac{(x - x_0)}{\alpha}\right]\right\}$, $-\infty < x < \infty$
Log-Gumbel	$F(x) = \exp\left[-\left(\frac{\theta - x_0}{x - x_0}\right)^{\beta}\right]$, $x_0 < x < \infty$, $\beta > 0$
Lognormal	$f(x) = \frac{1}{\sqrt{2\pi}(x - x_0)\sigma_y}\exp\left\{-\frac{1}{2}\left[\frac{\ln(x - x_0) - \mu_y)}{\sigma_y}\right]^2\right\}$, $x \ge x_0$
Log-Pearson type III	$f(x) = \frac{1}{\lvert\alpha\rvert\Gamma(\beta)x}\left[\frac{\ln(x) - y_0}{\alpha}\right]^{\beta-1}\exp\left[-\frac{\ln(x) - y_0}{\alpha}\right]$ $\alpha > 0$ then $\exp(y_0) \le x < \infty$, $\alpha < 0$ then $-\infty < x \le \exp(y_0)$
Weibull	$f(x) = \frac{\beta}{\alpha}\left[\frac{x - x_0}{\alpha}\right]^{\beta-1}\exp\left\{-\left[\frac{x - x_0}{\alpha}\right]^{\beta}\right\}$ $x_0 \le x < \infty$, $\alpha > 0$, $\beta > 0$

owing to the asymptotic mean squared error (MSE) and mean integrated squared error (MISE) calculations for the kernel density estimator. Since then, there has been a great deal of theoretical investigation into the kernel density estimator (Scott, 1979; Bowman, 1985; Terrell and Scott, 1985; Silverman, 1986; Terrell, 1990; Sheater, 1992; Marron and Wand, 1992). These theoretical results are, however, of asymptotic nature and it is still questionable how these methods perform well in small samples. One way to access the behavior of a small sample is through simulation experiments. Recent studies using many of the bandwidth selection methods were discussed in literature (Sheater, 1992; Marron, 1989; Jones et al., 1992; Park and Turlach, 1992). However, none of these studies give a clear answer to which bandwidth selection method is the best. It takes a lot of time to compute kernel density estimates, however, a drastic reduction of the computational time is possible through discretization methods. An intuitively easier approach to the idea of discretization is given by average shifted histogram (Scott, 1985). Härdle and Scott (1992) proposed the weighted averaging of around points (WARPing) to make calculations in kernel density estimation faster.

The applications of kernel density estimation for the estimation of flood quantiles have been investigated recently (Adamowski, 1985, 1989, 1996, 2000; Labatiuk and Adamowski, 1987; Guo, 1991, 1993; Guo et al., 1996; Kim et al., 1999; Lall et al., 1993; Moon et al., 1993). Although those papers were applied to several bandwidth selectors, new bandwidth selectors have not been applied in the areas of hydrology and water resources as of yet.

In this study, annual maximum flood data of Goan gauging station in the Han River basin in Korea are used for data applications. For the kernel density estimates, the Gaussian kernel function is employed and the bandwidth is selected based on seven data driven methods such as Silverman's rule of thumb (ROT) (Silverman, 1986), least squares cross-validation (LSCV) (Rudemo, 1982; Bowman, 1985; Stone, 1984; Hall and Marron, 1987), bandwidth factorized smoothed cross-validation (JMP) (Hall et al., 1992), smoothed cross-validation (SCV) (Hall et al., 1992), biased cross-validation (BCV) (Scott and Terrell, 1987), Park and Marron plug-in (PM) (Park and Marron, 1990), and SJ (Sheater and Jones, 1991). And a computational cost for kernel density estimation is reduced by using the WARPing method (Härdle and Scott, 1992).

The predictive abilities of the kernel density estimates are compared with those of the probability distributions through Monte Carlo simulation experiments based on the two types of parent distribution. One is the probability distribution that is selected as an appropriate model for Goan gauging station and the other one is a mixture distribution of two normal distributions. The flood quantiles are estimated based on the assumed probability distributions and kernel density estimates. In addition, the uncertainty analyses are carried out based on Monte Carlo simulation experiment and kernel density estimates in order to consider the effects of data and model errors.

2. Parametric method

The probability distributions used in this study for flood frequency analysis are as follows: the gamma, general extreme value (GEV), Gumbel (GUM), log-Gumbel, lognormal, log-Pearson type III, and Weibull distributions. Table 1 shows the PDF and/or cumulative distribution function (CDF) of each distribution, and gives the validity conditions of the parameters and the ranges of random variables for the PDF and CDF, respectively.

3. Nonparametric method

Kernel density estimations may provide a bridge between making no assumptions on formal structure (a purely nonparametric approach) and making very strong assumptions (a parametric approach). By kernel density estimation it is possible for the data to show the analyst what the pattern truly is.

3.1. Kernel density estimation

The idea of kernel estimators was introduced by Rosenblatt (1956). He proposed to smooth kernel weights on each of the observations. The merits of the kernel estimation are flexible formation and mathematical tractability. The general form of the kernel function is given by (Härdle, 1991)

$$K_h(x) = \frac{1}{h}K\left(\frac{x}{h}\right) \tag{1}$$

where h is a bandwidth and K the kernel function.

The kernel density estimates are given by averaging over these kernel functions in the observations (Härdle, 1991)

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{2}$$

where n is the number of observed data and $\hat{f}_h(x)$ is the kernel density estimates.

3.2. Bandwidth selectors

The practical implementation of the kernel density estimator requires the selection of the bandwidth h. One strategy for selecting the density estimates is to begin with a large bandwidth and to decrease the amount of smoothing until the fluctuations start to appear. This approach is viable but there are also many cases where it is very beneficial to have the bandwidth automatically selected from the data. A method that uses the data $X_1,\ldots,X_n$ to estimate a bandwidth $\hat{h}$ is called a bandwidth selector. In this study, seven bandwidth selectors such as ROT, LSCV, JMP, SCV, BCV, PM, and SJ are employed.

The optimal bandwidth for a kernel density estimate is typically calculated on the bias of an estimate for the integrated squared error (ISE),

$$\mathrm{ISE}(h) = \int [\hat{f}_h(x) - f(x)]^2 \, \mathrm{d}x \tag{3}$$

And its expected value, the MISE is given by

$$\begin{aligned}\mathrm{MISE} &= E\left(\int [\hat{f}_h(x) - f(x)]^2 \, \mathrm{d}x\right) \\ &= \int E(\hat{f}_h(x) - E[\hat{f}_h(x)] + E[\hat{f}_h(x)] - f(x))^2 \, \mathrm{d}x \\ &= \int E(\hat{f}_h(x) - E[\hat{f}_h(x)])^2 + (E[\hat{f}_h(x)] - f(x))^2 \, \mathrm{d}x \\ &= \int \mathrm{Var}\, \hat{f}_h(x)\mathrm{d}x + \int \mathrm{bias}^2\, \hat{f}_h(x)\mathrm{d}x\end{aligned} \tag{4}$$

where the first integral is integrated variance (IV) and the second integral is integrated squared bias (IB). IV and IB are given by (Turlach, 1993),

$$\mathrm{IV}(h) = (nh)^{-1}R(K)f(x) + O(n^{-1}h^2) \tag{5}$$

$$\mathrm{IB}(h) = \frac{h^4}{4}\mu_2^2(K)R(f^{(2)}) + O(h^8) \tag{6}$$

where $R(L) = \int L^2(x)\mathrm{d}x$, $\mu_j(L) = \int x^j L(x)\mathrm{d}x$, and $f^{(j)}$ is the jth derivative of f. For a Gaussian kernel, $R(K) = \int K^2(x)\mathrm{d}x = 1/2\sqrt{\pi}$ and $\mu_2(K) = \int x^2 K(x)\mathrm{d}x = 1$.

The asymptotic mean integrated square error (AMISE) based on the Taylor expansion of f is given by

$$\mathrm{AMISE}(h) = \frac{1}{nh}R(K) + h^4\left(\frac{\mu_2(K)}{2!}\right)^2 R(f^{(2)}) \tag{7}$$

All the above optimal bandwidths depend on the unknown density f or derivatives of f.

3.2.1. Rule of thumb

The ROT was proposed to minimize the AMISE (Silverman, 1986). The best trade-off between asymptotic variance and bias is given by

$$h_\infty = \left(\frac{R(K)}{\mu_2^2(K)R(f^{(2)})}\right)^{1/5} n^{-1/5} \tag{8}$$

where h_∞ is the minimizer of the AMISE and $R(f^{(2)})$ is the only unknown. Assuming the unknown distribution to be normal with parameter μ and σ, the estimate of h_∞ for a Gaussian kernel is given by (Härdle, 1991)

$$h_{\mathrm{ROT}} = 1.06\hat{\sigma}n^{-1/5} \tag{9}$$

The advantage of ROT is that it provides a very practical method of bandwidth selection while the disadvantage is that the bandwidth is wrong if the population is not normally distributed.

3.2.2. Cross-validation methods

3.2.2.1. Least squares cross-validation. The LSCV function is defined by (Rudemo, 1982; Bowman, 1985; Stone, 1984; Hall and Marron, 1987))

$$\mathrm{LSCV}(h) = R(\hat{f}_h(x)) - 2\sum_{i=1}^{n} \hat{f}_{h,i}(x_i) \tag{10}$$

where $\hat{f}_{h,i}(x)$ is the density estimate obtained by all data points except for the ith observation. The LSCV function can be viewed as an estimator of $\mathrm{ISE}(h) - R(f)$ (Turlach, 1993).

3.2.2.2. Bandwidth factorized cross-validation (Jones, Marron, and Park CV, JMP). The starting point for JMP is the representation $\mathrm{MISE}(h) = \mathrm{IV}(h) + \mathrm{IB}(h)$. As far $(nh)^{-1}R(K)$ has been proved to be a good estimator of $\mathrm{IV}(h)$, the main task of JMP is to search for a 'good' estimator of $\mathrm{IB}(h)$. The integrated bias can be written as

$$\mathrm{IB}(h) = \int (K_h * f - f)^2(x)\mathrm{d}x \tag{11}$$

where $*$ denotes the convolution of two functions K_h and f

$$(K_h * f)(x) = \int K_h(x-u)f(u)\mathrm{d}u = \int K_h(u)f(x-u)\mathrm{d}u \tag{12}$$

Introducing K_0 the Dirac function $K_0(u) = I(u=0)$, $\mathrm{IB}(h)$ is estimated by (Turlach, 1993)

$$\mathrm{IB}(h) = n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}(K_h * K_h - 2K_h + K_0) * K_g * K_g(x_i - x_j) \tag{13}$$

The JMP function is obtained as

$$\mathrm{JMP}(h) = (nh)^{-1}R(K) + \mathrm{IB}(h). \tag{14}$$

Table 2
The estimated parameters based on PWM

Probability distribution	Location parameter	Scale parameter	Shape parameter
Gamma-2	0.0	4209.9	3.1
Gamma-3	788.9	4532.4	2.7
GEV	9606.0	5495.1	−0.1
Gumbel	9747.5	5808.1	
Log-Gumbel-2	0.0	8977.1	2.6
Lognormal-2	0.0	9.4	0.6
Lognormal-3	−4396.3	9.7	0.4
Log-Pearson type III	16.0	−0.1	132.6
Weibull-2	0.0	14 759.5	1.9
Weibull-3	2415.9	11 800.3	1.5

3.2.2.3. Smoothed cross-validation. This method was introduced by Hall et al. (1992). In Eq. (13), different kernel L and bandwidth g are used. By deleting all diagonal terms (i.e. where $i = j$) and using $n \approx n - 1$, this yields the following criterion

$$\mathrm{SCV}(h) = \frac{1}{nh}R(K) + \frac{1}{n(n-1)}\sum_{i \neq j}(K_h * K_h - 2K_h + K_0) * L_g * L_g(x_i - x_j) \tag{15}$$

Because of using the high order kernel L in Eq. (15), this method is only superior to other methods for very large sample sizes.

3.2.2.4. Biased cross-validation. This technique is based on AMISE (Scott and Terrell, 1987). To estimate $R(f^{(2)})$ in Eq. (7), Scott and Terrell (1987) gave a formula for the expectation of $R(f_h^{(2)})$ when the kernel K and the density f are at least twice continuously differentiable

$$R(\hat{f}^{(2)}) = R\left(\hat{f}_h^{(2)}\right) - \frac{1}{nh^5}R(K^{(2)}) \tag{16}$$

This leads to

$$\mathrm{BCV}(h) = \frac{1}{nh}R(K) + h^4\left(\frac{\mu_2(K)}{2!}\right)^2\left(R\left(\hat{f}_h^{(2)}\right) - \frac{1}{nh^5}R(K^{(2)})\right) \tag{17}$$

For the skewed distributions, BCV tends to give a too big bandwidth as shown by the simulations in Scott and Terrell (1987).

3.2.3. Plug-in methods

Plug-in methods target the AMISE as the distance to be minimized. h_∞ is a function of $R(f^{(2)})$, which is estimated through a sequence of bandwidths $h_1, h_2, \ldots$ The first bandwidth h_1 serves initially to estimate the density $\hat{f}_{h_1}(x)$, which is used to compute $\hat{R}(f^{(2)}) = R(\hat{f}_{h_1}(x))$; subsequently, $\hat{R}(f^{(2)})$ is plugged into Eq. (8) to calculate the second bandwidth h_2 in the sequence. Thus, a new estimate of $R(f^{(2)})$ is obtained as $\hat{R}(f^{(2)}) = R(\hat{f}_{h_2}(x))$, and again plugged into Eq. (8) to derive h_3, and so on. The process iterates until convergence of the bandwidths is reached.

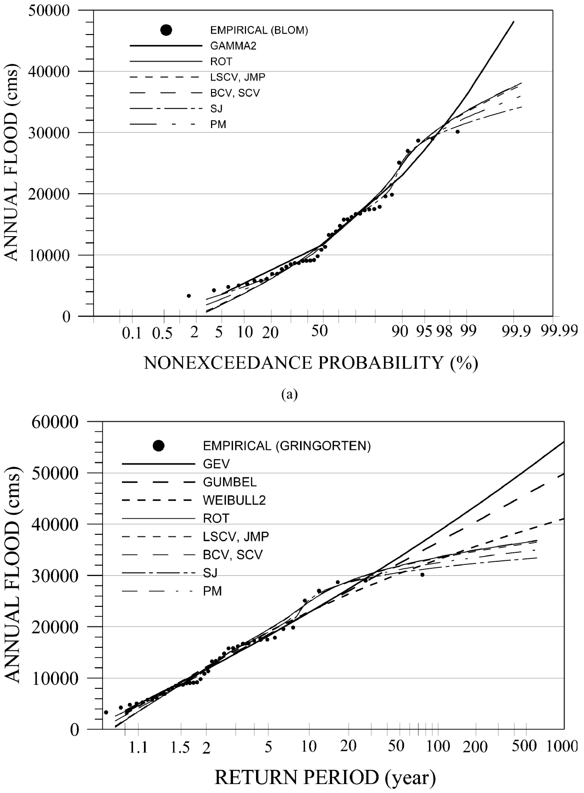
3.2.3.1. Park and Marron plug-in. Supposing bandwidth g to estimate $R(f^{(2)})$ in MISE, then the second derivative of $\hat{f}_g$ can be computed as

$$\hat{f}_g^{(2)}(x) = \frac{1}{ng^3}\sum_{i=1}^{n}K^{(2)}\left(\frac{x - X_i}{g}\right) \tag{18}$$

Of course, this will yield a bandwidth choice problem as well. However, we can now use a ROT bandwidth in this first stage. Park and Marron (1990) showed that the bandwidth can be taken as $g = C(\varphi_{\hat{\sigma}})h^{10/13}$, where $C(\varphi_{\hat{\sigma}})$ is a constant calculated from the normal density $\varphi_{\hat{\sigma}}(x) = \varphi(x/\hat{\sigma})/\hat{\sigma}$ for every x, and a power constant 10/13 results from the optimal rate for the second derivatives. As usual, $\hat{\sigma}$ denotes the standard deviation estimated from the data. A further problem occurs due to the bias of $R(\hat{f}_g^{(2)})$ which can be

Table 3
The results of goodness of fit tests (significance level: 0.05) (COM: computed value, TAB: tabulated value, RST: result)

Distributions	Chi-square			K–S			CVM			PPCC		
	COM	TAB	RST	COM	TAB	RST	COM	TAB	RST	COM	TAB	RST
Gamma-2	4.90	7.81	OK	0.11	0.18	OK	0.06	0.46	OK	0.984	0.978	OK
Gamma-3	4.90	5.99	OK	0.10	0.18	OK	0.06	0.46	OK	0.985	0.978	OK
GEV	4.86	5.99	OK	0.11	0.18	OK	0.07	0.46	OK	0.985	0.951	OK
Gumbel	7.43	7.81	OK	0.12	0.18	OK	0.07	0.46	OK	0.983	0.961	OK
Log-Gumbel-2	9.71	7.81	NG	0.15	0.18	OK	0.26	0.46	OK			
Lognormal-2	6.86	7.81	OK	0.12	0.18	OK	0.07	0.46	OK	0.959	0.973	NG
Lognormal-3	5.14	5.99	OK	0.11	0.18	OK	0.07	0.46	OK	0.959	0.973	NG
Log-Pearson type III	8.24	5.99	NG	0.10	0.18	OK	0.06	0.46	OK			
Weibull-2	7.71	7.81	OK	0.12	0.18	OK	0.07	0.46	OK	0.985	0.962	OK
Weibull-3	4.29	5.99	OK	0.09	0.18	OK	0.05	0.46	OK	0.987	0.962	OK



(a)

(b)

Fig. 1. Empirical and fitted frequency curves for annual maximum flood on (a) normal probability paper and (b) GUM probability paper.

overcome by using a bias-corrected estimate

$$R(\hat{f}^{(2)}) = R\left(\hat{f}_g^{(2)}\right) - \frac{1}{ng^5} R(K^{(2)}) \tag{19}$$

The performance of PM in the simulation studies is usually quite good. A disadvantage is that for small bandwidths, the estimator $R(\hat{f}^{(2)})$ may give negative results.

3.2.3.2. Sheater and Jones plug-in. Sheater and Jones (1991) reconsidered the problem of estimating $R(\hat{f}^{(2)})$. They used the same idea as Park and Marron (1990) but with bandwidth

$$g \propto \frac{R(f^{(2)})}{R(f^{(3)})} h^{5/7} \tag{20}$$

in which $R(f^{(2)})$ and $R(f^{(3)})$ can be estimated by $R(\hat{f}_{g_1}^{(2)})$ and $R(\hat{f}_{g_2}^{(3)})$ and both bandwidths g_1, g_2 are determined by asymptotic optimal values and only in this step the normal PDF is used as a reference probability model. This improves the PM but is not the best achievable rate yet.

4. Applications to flood data

For data applications, annual maximum flood data of Goan gauging station in the Han River Basin in Korea are selected. In addition, 1000 data samples with 100 random numbers which come from a mixture of two normal distributions are used for assessing the applicability of parametric

Table 4
The RBIAS for sample size $N = 100$

Parent model	Return period (years)	Parametric method					Smoothing technique						
		GAM2	GEV	GUM	LN3	WBU2	ROT	LSCV	BCV	SCV	JMP	PM	SJ
GAM2	20	−0.00162	−0.00447	−0.00958	−0.00445	−0.03125	0.02032	0.01929	0.02028	0.02044	0.01865	0.01617	0.01202
	50	0.00364	0.02135	0.00282	0.01576	−0.05769	0.01686	0.01585	0.01681	0.01694	0.01532	0.01273	0.00892
	100	0.00299	0.03607	0.00955	0.02592	−0.08036.	0.01794	0.01710	0.01789	0.01801	0.01659	0.01444	0.01102
	200	0.00344	0.05936	0.01800	0.04359	−0.10042	−0.00025	−0.00118	−0.00030	−0.00017	−0.00175	−0.00410	−0.00793
GEV	20	−0.00463	−0.00114	−0.01233	0.00015	−0.03427	0.01729	0.01656	0.01730	0.01751	0.01602	0.01443	0.01069
	50	−0.03740	−0.00734	−0.03664	−0.01321	−0.09674	0.00946	0.00888	0.00940	0.00961	0.00844	0.00720	0.00451
	100	−0.05255	0.00625	−0.04505	−0.00693	−0.13170	0.01005	0.00944	0.01001	0.01018	0.00922	0.00770	0.00457
	200	−0.07497	0.00704	−0.06016	−0.01316	−0.17117	−0.02060	−0.02113	−0.02064	−0.02047	−0.02154	−0.02265	−0.02527
GUM	20	0.00311	−0.00315	−0.00360	−0.00369	−0.02719	0.02419	0.02346	0.02428	0.02447	0.02277	0.02186	0.01750
	50	−0.00260	0.00275	−0.00100	−0.00171	−0.06442	0.02506	0.02450	0.02510	0.02526	0.02391	0.02321	0.01988
	100	−0.01163	0.00913	−0.00276	0.00119	−0.09458	0.02731	0.02684	0.02737	0.02747	0.02636	0.02592	0.02280
	200	−0.01930	0.01419	−0.00228	0.00339	−0.12177	0.00409	0.00349	0.00416	0.00430	0.00295	0.00227	−0.00118
LN3	20	−0.00439	−0.00172	−0.01237	−0.00041	−0.03392	0.02494	0.02405	0.02492	0.02514	0.02350	0.02146	0.01755
	50	−0.02365	0.00772	−0.02363	0.00158	−0.08356	0.02523	0.02449	0.02516	0.02535	0.02406	0.02221	0.01918
	100	−0.04276	0.01076	−0.03555	−0.00181	−0.12260	0.03141	0.03083	0.03134	0.03150	0.03051	0.02889	0.02596
	200	−0.05180	0.03120	−0.03819	0.01040	−0.14985	0.00704	0.00642	0.00701	0.00716	0.00597	0.00443	0.00163
WBU2	20	0.03156	0.00467	0.02305	0.00229	0.00105	0.02785	0.02692	0.02816	0.02810	0.02575	0.02545	0.01953
	50	0.06282	0.01884	0.06235	0.01804	−0.00227	0.02751	0.02656	0.02782	0.02775	0.02529	0.02496	0.01884
	100	0.09350	0.04210	0.10076	0.04369	0.00262	0.02704	0.02612	0.02732	0.02727	0.02488	0.02468	0.01891
	200	0.11843	0.05406	0.13508	0.06052	0.00251	0.01654	0.01549	0.01686	0.01682	0.01409	0.01390	0.00709

Table 5
The RRMSE for sample size $N = 100$

Parent model	Return period (year)	Parametric method					Smoothing technique						
		GAM2	GEV	GUM	LN3	WBU2	ROT	LSCV	BCV	SCV	JMP	PM	SJ
GAM2	20	0.07092	0.07389	0.06731	0.07522	0.07708	0.08533	0.08507	0.08518	0.08517	0.08516	0.08408	0.08450
	50	0.07831	0.09897	0.07130	0.09557	0.09564	0.10566	0.10568	0.10555	0.10555	0.10587	0.10548	0.10662
	100	0.08057	0.12326	0.07195	0.11229	0.11182	0.12965	0.12976	0.12956	0.12963	0.12989	0.12965	0.13023
	200	0.08511	0.15672	0.07662	0.13501	0.12897	0.15618	0.15652	0.15614	0.15608	0.15688	0.15746	0.15960
GEV	20	0.07716	0.08173	0.07271	0.08400	0.08411	0.09585	0.09568	0.09574	0.09584	0.09583	0.09511	0.09570
	50	0.09129	0.11003	0.08343	0.10803	0.12648	0.12494	0.12501	0.12490	0.12492	0.12526	0.12513	0.12627
	100	0.09996	0.12951	0.08772	0.12051	0.15511	0.16144	0.16125	0.16148	0.16146	0.16173	0.16087	0.16054
	200	0.11230	0.15391	0.09443	0.13573	0.18877	0.20225	0.20255	0.20227	0.20219	0.20285	0.20331	0.20507
GUM	20	0.07002	0.07426	0.06567	0.07569	0.07470	0.08925	0.08900	0.08906	0.08922	0.08904	0.08799	0.08805
	50	0.07627	0.09752	0.06925	0.09465	0.09840	0.11513	0.11516	0.11501	0.11509	0.11532	0.11498	0.11594
	100	0.08069	0.12018	0.07029	0.11075	0.12208	0.14376	0.14388	0.14371	0.14374	0.14406	0.14400	0.14441
	200	0.08482	0.14065	0.07236	0.12371	0.14488	0.17157	0.17181	0.17148	0.17148	0.17216	0.17218	0.17405
LN3	20	0.07447	0.07901	0.06987	0.08142	0.08169	0.09703	0.09680	0.09689	0.09697	0.09693	0.09590	0.09625
	50	0.08531	0.10575	0.07738	0.10307	0.11587	0.13034	0.13036	0.13025	0.13031	0.13054	0.13030	0.13116
	100	0.09465	0.12562	0.08290	0.11636	0.14718	0.16315	0.16322	0.16308	0.16316	0.16343	0.16306	0.16273
	200	0.10019	0.15347	0.08372	0.13269	0.17072	0.19236	0.19261	0.19234	0.19232	0.19293	0.19347	0.19507
WBU2	20	0.07001	0.06328	0.06356	0.06328	0.06198	0.07118	0.07071	0.07120	0.07122	0.07048	0.06959	0.06852
	50	0.09311	0.07969	0.08855	0.07695	0.06711	0.08449	0.08422	0.08448	0.08452	0.08425	0.08354	0.08365
	100	0.12042	0.10206	0.12134	0.09709	0.07338	0.09779	0.09767	0.09776	0.09782	0.09784	0.09727	0.09797
	200	0.14189	0.12661	0.15127	0.11863	0.07478	0.10978	0.10984	0.10970	0.10976	0.11027	0.10968	0.11162

and kernel density estimation with respect to nonunimodal density.

4.1. Parametric methods

For the parametric method, the probability distributions as shown in Table 1 are assumed to be the underlying distributions for a given flood data. The parameters of each model are estimated based on the method of probability weighted moments (PWM) and then the validity conditions of the estimated parameters are checked. Table 2 shows the estimated parameters which are obtained based on PWM. The log-Gumbel-3 distribution is discarded because the estimated parameters violate the parameter validity conditions.

For the distributions which meet with parameter validity conditions, the goodness of fit tests such as chi-square, Kolmogorov–Smirnov (K–S), Cramer von Mises (CVM), and probability plot correlation coefficient (PPCC) tests are performed to select an appropriate probability distribution for a given data. Table 3 shows the results of goodness of fit tests at significance level 0.05. And PPCC test is also performed for the gamma, GEV, GUM, lognormal, and Weibull distributions. As shown in Table 3, the null hypothesis of good fit can be rejected for the log-Gumbel-2, and log-Pearson type III distributions based on chi-square test and for the lognormal-2 and lognormal-3 (LN3) distributions based on PPCC test, respectively.

Finally, the gamma, GUM, GEV, and Weibull distributions are selected as the appropriate distributions for annual maximum flood of Goan station on the basis of the parameter validity conditions and the goodness of fit tests.

4.2. Kernel density estimation

Gaussian kernel density function among several kernel density functions is employed because of many times differentiability. The bandwidth selectors are estimated based on seven data driven methods such as ROT, LSCV, JMP, SCV, BCV, PM, and SJ. Highly computational cost for the density estimation can be reduced by the WARPing method (Härdle et al., 1995). For flood frequency estimation, the kernel estimate $F_n(x)$ of the CDF $F(x)$ is interested, which is defined as

$$F_n(x) = \int_{-\infty}^{x} \sum_{i=1}^{n} \frac{1}{nh} K\left(\frac{t - x_i}{h}\right) \mathrm{d}t = \frac{1}{h} \sum_{i=1}^{n} K^*\left(\frac{x - x_i}{h}\right) \tag{21}$$

and

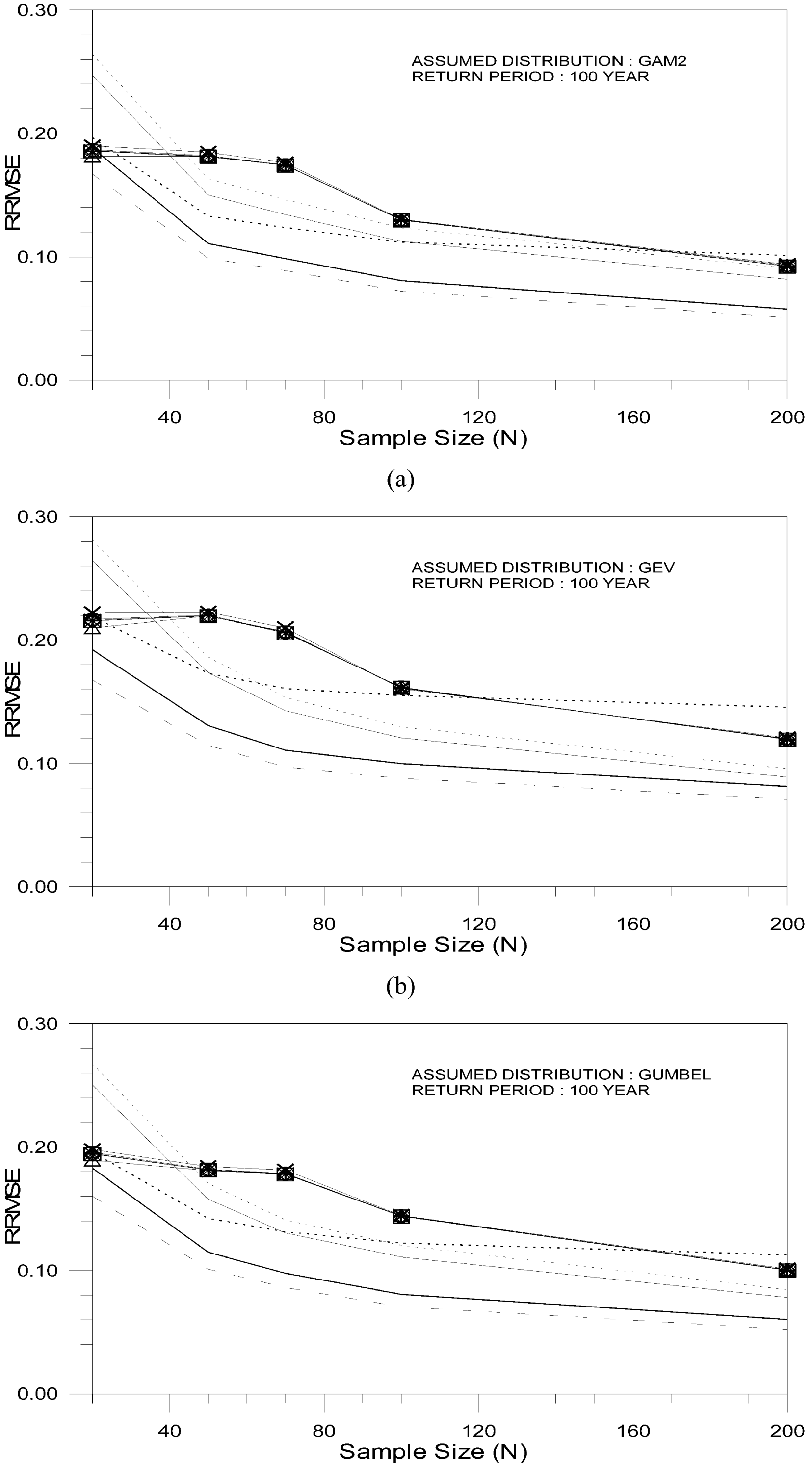
$$K^*(t) = \int_{-\infty}^{t} K(u)\mathrm{d}u = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} \exp\left(-\frac{1}{2}u^2\right)\mathrm{d}u \tag{22}$$

A quantile estimator can be defined through the inverse of the distribution function $F_n(x)$ (Nadaraya, 1964; Azzalini, 1981) or through kernel averaging of the sample quantile function. There are extensive studies on the kernel estimation of quantiles within the range of the data (Sheater and Marron, 1990; Falk, 1984; Yang, 1985).

Fig. 1(a) shows the plots of the frequency curves from the Blom's plotting position formula (Stedinger et al., 1993), and the fitted ones obtained from the gamma-2 (GAM2) and seven bandwidth selectors on the normal probability paper while the same plots from the Gringorten's formula (Stedinger et al., 1993), GEV, GUM, Weibull-2 (WBU2), and the bandwidth selectors are displayed on the GUM probability paper as shown in Fig. 1(b). As shown in Fig. 1, the fitted frequency curves for the GAM2, GEV, GUM, and WBU2 models seem to overestimate the flood quantiles beyond 50 years return period (nonexceedance probability: 0.98) because the fitted curves pass over the maximum flood data. However, the fitted curves for the kernel smoothing follow the empirical one fairly well even though these curves slightly pass over the maximum value.

4.3. Sampling properties of statistical estimates

The simulation experiments are designed to investigate how well a distribution can estimate the quantiles for a given return period when the population distribution is different from the assumed distribution. For this purpose, two cases of simulation experiments are performed. For first one, the selected probability model is assumed as a parent model. And in the second one, an arbitrary probability model such as a mixture of two

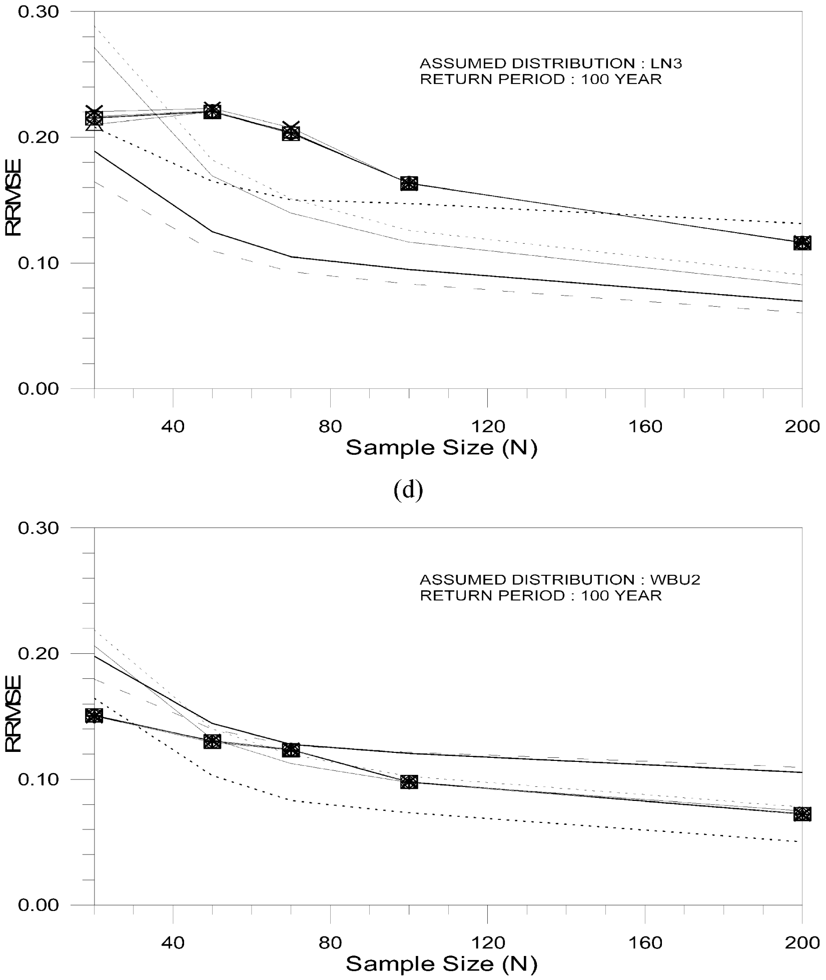


(a)

(b)

(c)

Fig. 2. The RRMSE of 100-years flood quantiles as a function of sample size (parent distributions: (a) GAM2, (b) GEV, (c) GUM, (d) LN3, and (e) WBU2).



(d)

(e)

Fig. 2. (*continued*)

normal distributions is assumed as a parent model to investigate the performances in the case of bimodal model.

4.3.1. Unimodal distributions

In the first case, the GAM2, GEV, GUM, LN3, and WBU2 distributions are assumed as a parent model, respectively. The estimated parameters of each model obtained from the annual maximum flood of Goan station are assumed as the parameter values of a parent distribution. Then, 1000 flood data sets are generated based on each assumed parent distribution with various sample sizes and return periods. For the generated data sets, the parameters of the GAM2, GEV, GUM, LN3, and WBU2 models are estimated based on PWM. Kernel density estimation is also applied. Finally, the estimated flood quantiles are obtained from the estimated parameters of each data set for a given method, and then the relative biases (RBIAS) and relative root mean square error (RRMSE) can be calculated from both estimated flood quantiles and population ones. RBIAS and RRMSE are used as the indicators to test the predictive ability of the models. RBIAS and RRMSE are

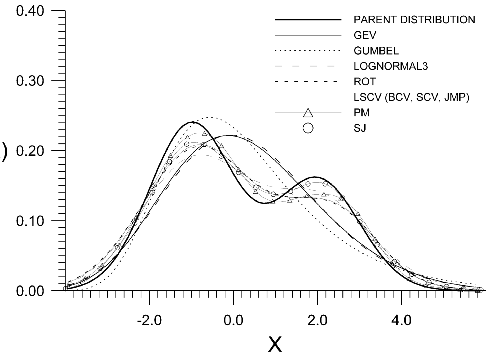


Fig. 3. The PDF of the mixture distribution and the fitted PDFs.

given by

$$\mathrm{RBIAS} = \frac{1}{M}\sum_{i=1}^{M}\left(\frac{\hat{Q}_{i,T} - Q_T}{Q_T}\right) \tag{23}$$

$$\mathrm{RRMSE} = \left[\frac{1}{M}\sum_{i=1}^{M}\left(\frac{\hat{Q}_{i,T} - Q_T}{Q_T}\right)^2\right]^{1/2} \tag{24}$$

where $\hat{Q}_{i,T}$ and Q_T are the estimated and population flood quantiles, respectively, and M represents the number of simulations (i.e. $M = 1000$ in this case). These are then used to compare the performance of different procedures (five parametric models and seven bandwidth selectors) for the given return periods and sample sizes.

Tables 4 and 5 present the RBIASs and the RRMSEs for the specific sample size ($N = 100$) and Fig. 2 shows the RRMSEs of the 100-years flood quantiles as a function of sample size when the GAM2, GEV, GUM, LN3 and WBU2 models are assumed to be a parent distribution, respectively. The followings are the summary of the simulation results:

1. When the parent probability distribution is the same as the applied model, almost all RBIASs of the applied models are smaller than those of any other models (Table 4).
2. Among seven bandwidth selectors, the RBIASs of SJ are the smallest in most cases (Table 4).
3. As the return period increases, the RRMSEs increase (Table 5). And the RRMSEs decrease as the sample size increases (Fig. 2).
4. The RRMSEs of the GUM are smaller than those of any other models regardless of parent models. However, those of WBU2 are the smallest when the WBU2 is assumed as a parent model (Table 5 and Fig. 2).
5. When the WBU2 is assumed as a parent model, the RRMSEs of seven bandwidth selectors are relatively small, while those of seven bandwidth selectors are much bigger than those of parametric methods when the other parametric models are assumed as parent models
6. There are no big differences in the RRMSEs among the bandwidth selectors. However, ROT, LSCV, and JMP have relatively bigger RRMSEs than other bandwidth selectors (Table 5).

4.3.2. Bimodal distribution

The second case is that a mixture of two normal PDFs is assumed as a parent model as follows

$$f(x) = 0.6\varphi_1(x) + 0.4\varphi_2(x) \tag{25}$$

where $\varphi_1(x) = \varphi(x+1)$ a normal density with mean $\mu = -1$ and variance $\sigma^2 = 1$ and $\varphi_2(x) = \varphi(x-2)$ a normal density with mean $\mu = 2$ and variance $\sigma^2 = 1$.

Fig. 3 shows the PDF of the mixture distribution and the fitted PDFs for data generated from the mixture distribution. The GAM2 and WUB2 are discarded because the null hypothesis of good fit can be rejected based on a chi-square test. As shown in Fig. 3, the GEV, GUM, and LN3 do not fit the PDF of the mixture model well while the kernel density estimation represent the bimodal characteristics fairly well, especially in the cases of SJ and PM.

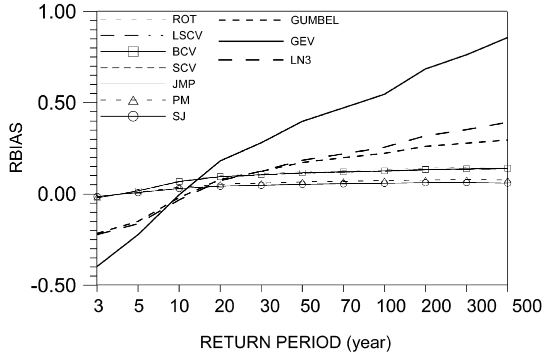


Fig. 4. The RBIASs in case of the mixture distribution as a parent model.

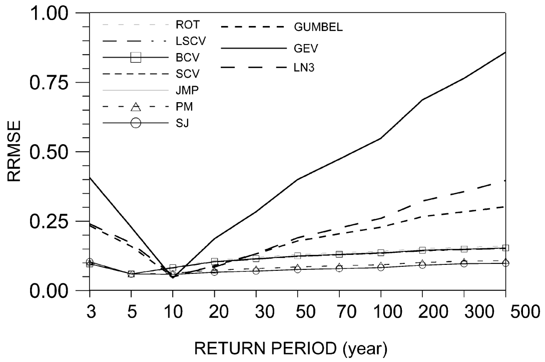


Fig. 5. The RRMSEs in case of the mixture distribution as a parent model.

Fig. 4 shows the RBIASs for 1000 samples of sample size 100 generated from the mixture distribution. As shown in Fig. 4, the probability models have the largest RBIASs except the return period $T = 10$ years. Among the bandwidth selectors, the RBIASs of SJ are smaller than those of any other bandwidth selectors and PM has the next smallest RBIASs while ROT has the largest ones. This result supports the existing study which reported that ROT is suitable for a flat density but fails to detect the bimodality of density (Härdle, 1991). Similar results can be observed for the RRMSEs as shown in Fig. 5. Take into consideration that, the kernel density estimation should be used instead of the parametric models when a parent model has a bimodality of density.

5. Safety factors of flood quantiles

Uncertainty in flood quantile estimation can be caused by inaccurate measurements of the data. Also, uncertainty is caused by model errors, which consist of incorrect estimation of the population parameters owing to sampling errors, incorrect choice of the parameter estimation procedures, and incorrect choice of population density function. Simulation experiments are carried out to investigate the extent to which the estimated flood quantiles might be affected by data and model errors.

For a given parent distribution, sample size, and return period, the simulation procedures consist of the following steps:

(A) The GAM2 model is assumed as a population distribution. The parameters of GAM2 model are selected from annual maximum flood data of Goan.

Step 1. Generate a sample of size N: x_i, $i = 1, \ldots, N$.

Step 2. Estimate quantile $\hat{Q}_T$ from x_i series based on

(a) parametric methods (GAM2, GEV, GUM, LN3, WBU2 models)

(b) Kernel smoothing (ROT, LSCV, BCV, SCV, JMP, PM, and SJ).

Step 3. Repeat the procedure (Steps 1 and 2) 1000 times to obtain 1000 quantile estimates of $\hat{Q}_j$, $j = 1, \ldots, 1000$ for each method.

Step 4. Estimate the PDF of $\hat{Q}_T$ obtained from Step 3. Since it is not possible to assume a single suitable parametric form to describe the sampling distribution of $\hat{Q}_T$ values (Mkhandi et al., 1996), kernel smoothing of fitting distribution functions are chosen.

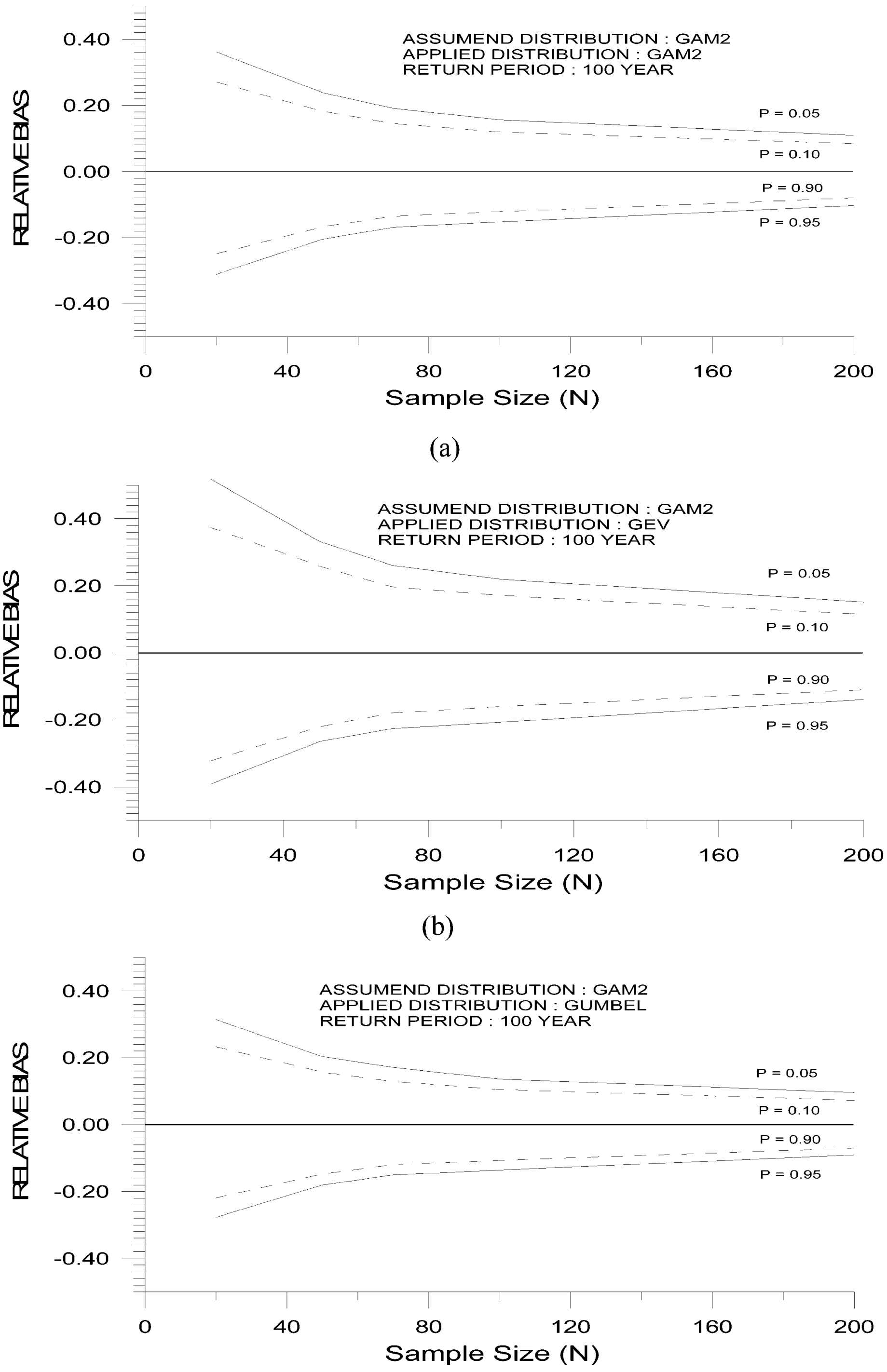
Step 5. Perform the uncertainty and risk analysis.

(B) The GEV, GUM, LN3, and WBU2 models are assumed as the population distributions, respectively, and then repeat the above procedures (Steps 1–5).

To assess the effects of data and model errors in quantile estimation, quantile values $\hat{Q}_T$ are calculated for specified probability of exceedance $P(\hat{Q}_T) = 0.05$, 0.10, 0.90, and 0.95 based on kernel density estimation. The RBIAS of $\hat{Q}_T$ is given by

$$\mathrm{RBIAS}_{P(\hat{Q}_T)} = \lfloor \hat{Q}_T - E(\hat{Q}_T) \rfloor / E(\hat{Q}_T) \tag{26}$$

The magnitude of RBIAS is the indicator of the effect of errors such as model and data errors on the quantile estimates for each sample considered. Fig. 6 presents RBIAS of 100-years quantiles as a function of sample sizes for the GAM2 model. Each quantile of 1000 samples is estimated by parametric methods and sampling distribution of the estimated quantiles is obtained by kernel density estimation. The magnitude of uncertainty decreases as the sample size increases. As shown in Fig. 6, the RBIAS band is the widest when the GEV model is applied one. However, the GUM

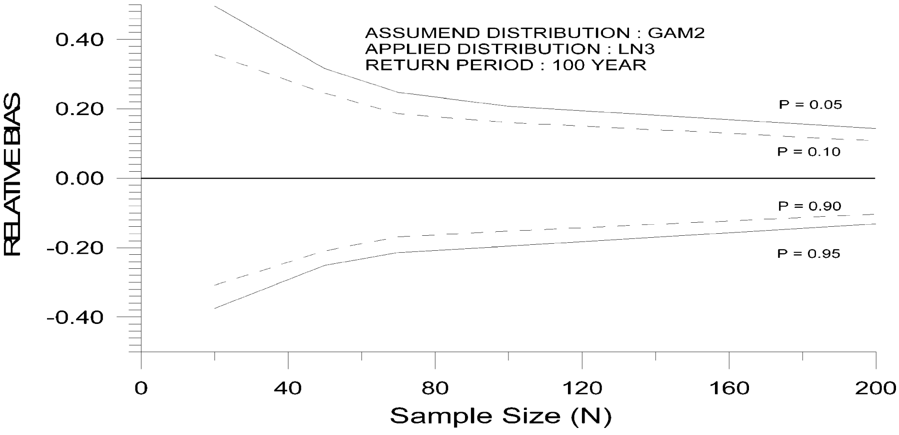


(a)

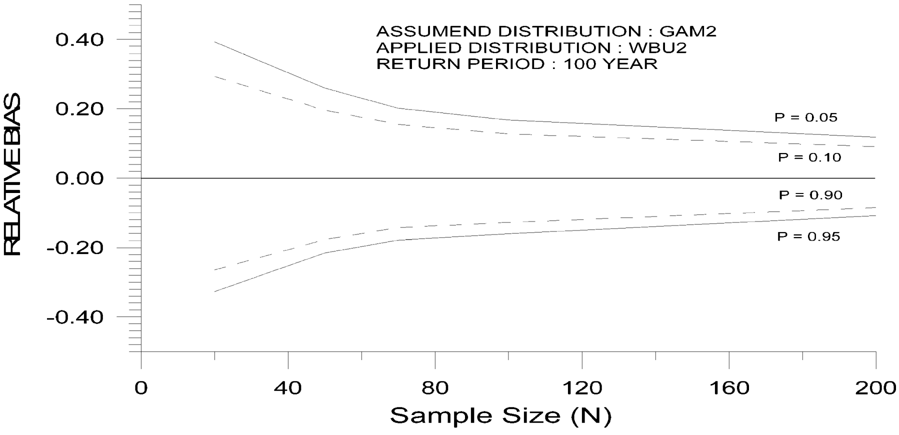
(b)

(c)

Fig. 6. The RBIASs of 100-years flood quantile estimates for the GAM2 as a parent when the applied models are (a) GAM2, (b) GEV, (c) GUM, (d) LN3, and (e) WBU2.



(d)



(e)

Fig. 6. (*continued*)

model has the narrowest width of RBIAS band. In other words, the GEV model has the largest RBIAS whereas the GUM model has the smallest RBIAS. For other distributions such as the GAM2, GEV, LN3, and WBU2 assumed as parent models it is the same as the earlier results.

If safety factor (SF) is defined as the value of 95% confidence limit of the sampling distribution of $\hat{Q}_T(j)$, $j = 1,\ldots,1000$, it can be calculated by (Mkhandi et al., 1996)

$$\mathrm{SF} = 1.0 + \mathrm{RBIAS}(\%)_{P(\hat{Q}_T)=0.05} \qquad (27)$$

Table 6 summarizes the SF values of 100-years flood quantiles for different parent distributions, models, and sample sizes. As shown in Table 6, the SFs are the largest when the GEV is assumed as a parent model for all applied distributions except for when the LN3 is used as both parent and applied models especially for sample size $N = 20$. On the other hand, the SFs are the smallest when the WBU2 model is assumed as a parent model regardless of the applied models. Among the applied distributions, the GUM model has the smallest SFs for all parent models. In addition, the GEV model has the largest values for all parent models. It is the same as shown in Fig. 6.

Overall, the largest SFs are obtained based on the combination of the GEV as both a parent and an

Table 6
SFs of 100-years flood quantiles

Applied model	Bandwidth selector	Parent model														
		GAM2			GEV			GUM			LN3			WBU2		
		Sample size														
		20	50	100	20	50	100	20	50	100	20	50	100	20	50	100
GAM2	ROT	1.34621	1.19324	1.14215	1.38014	1.23030	1.15619	1.34556	1.21531	1.14303	1.35223	1.22060	1.15563	1.27652	1.16667	1.11915
	LSCV	1.34553	1.19337	1.14226	1.38022	1.22984	1.15629	1.34506	1.21538	1.14314	1.35241	1.22076	1.15568	1.27612	1.16677	1.11923
	BCV	1.34610	1.19337	1.14226	1.37993	1.23042	1.15629	1.34562	1.21538	1.14314	1.35158	1.22076	1.15574	1.27663	1.16677	1.11923
	SCV	1.34630	1.19337	1.14226	1.38022	1.23042	1.15629	1.34569	1.21538	1.14314	1.35222	1.22076	1.15574	1.27663	1.16677	1.11923
	JMP	1.34553	1.19277	1.14174	1.37985	1.22984	1.15580	1.34506	1.21503	1.14263	1.35158	1.21998	1.15521	1.27612	1.16628	1.11884
	PM	1.34509	1.19360	1.14259	1.38016	1.23016	1.15674	1.34512	1.21543	1.14343	1.35121	1.22056	1.15573	1.27656	1.16718	1.11936
	SJ	1.34431	1.19326	1.14233	1.38003	1.22959	1.15659	1.34475	1.21521	1.14316	1.35073	1.22017	1.15545	1.27621	1.16688	1.11900
GEV	ROT	1.45920	1.28124	1.20673	1.53408	1.34584	1.22602	1.50353	1.31524	1.21635	1.54470	1.33720	1.22219	1.35380	1.22812	1.15338
	LSCV	1.45797	1.28063	1.20683	1.53334	1.34598	1.22617	1.50275	1.31480	1.21593	1.54438	1.33672	1.22232	1.35399	1.22768	1.15350
	BCV	1.45953	1.28063	1.20683	1.53334	1.34575	1.22552	1.50275	1.31536	1.21647	1.54438	1.33733	1.22232	1.35309	1.22824	1.15350
	SCV	1.45797	1.28114	1.20683	1.53334	1.34598	1.22617	1.50275	1.31536	1.21647	1.54438	1.33733	1.22232	1.35399	1.22824	1.15350
	JMP	1.45797	1.28063	1.20634	1.53334	1.34533	1.22544	1.50275	1.31480	1.21593	1.54438	1.33672	1.22172	1.35309	1.22768	1.15290
	PM	1.45214	1.27903	1.20683	1.53036	1.34584	1.22588	1.50045	1.31532	1.21603	1.54355	1.33641	1.22249	1.35317	1.22806	1.15388
	SJ	1.44988	1.27760	1.20651	1.52965	1.34545	1.22541	1.49975	1.31502	1.21558	1.54340	1.33567	1.22208	1.35272	1.22711	1.15352
GUM	ROT	1.30112	1.16847	1.12455	1.31471	1.19747	1.13615	1.29498	1.18414	1.12283	1.29685	1.18894	1.13481	1.23669	1.14620	1.10498
	LSCV	1.30047	1.16860	1.12464	1.31418	1.19759	1.13625	1.29511	1.18413	1.12291	1.29706	1.18909	1.13490	1.23680	1.14585	1.10505
	BCV	1.30110	1.16860	1.12464	1.31485	1.19759	1.13625	1.29468	1.18424	1.12291	1.29610	1.18856	1.13490	1.23680	1.14630	1.10505
	SCV	1.30092	1.16860	1.12464	1.31485	1.19759	1.13625	1.29511	1.18424	1.12291	1.29706	1.18909	1.13490	1.23680	1.14630	1.10505
	JMP	1.30047	1.16799	1.12423	1.31418	1.19703	1.13578	1.29448	1.18379	1.12250	1.29610	1.18841	1.13445	1.23627	1.14585	1.10470
	PM	1.29994	1.16898	1.12502	1.31489	1.19769	1.13669	1.29515	1.18441	1.12336	1.29632	1.18896	1.13522	1.23679	1.14669	1.10524
	SJ	1.29939	1.16873	1.12475	1.31458	1.19741	1.13651	1.29489	1.18422	1.12315	1.29576	1.18863	1.13497	1.23655	1.14636	1.10505
LN3	ROT	1.43456	1.26341	1.19315	1.50859	1.32610	1.21258	1.47540	1.29666	1.20118	1.52028	1.31783	1.20881	1.33324	1.21188	1.14295
	LSCV	1.43332	1.26277	1.19326	1.50795	1.32622	1.21273	1.47467	1.29643	1.20078	1.51999	1.31735	1.20894	1.33339	1.21199	1.14307
	BCV	1.43489	1.26330	1.19326	1.50795	1.32622	1.21224	1.47467	1.29676	1.20117	1.51999	1.31796	1.20880	1.33276	1.21199	1.14307
	SCV	1.43332	1.26286	1.19326	1.50795	1.32622	1.21273	1.47467	1.29676	1.20129	1.51999	1.31789	1.20894	1.33339	1.21199	1.14307
	JMP	1.43332	1.26277	1.19275	1.50795	1.32563	1.21201	1.47467	1.29627	1.20078	1.51999	1.31735	1.20833	1.33269	1.21151	1.14250
	PM	1.42753	1.26088	1.19320	1.50594	1.32622	1.21245	1.47323	1.29664	1.20070	1.51921	1.31674	1.20884	1.33301	1.21226	1.14353
	SJ	1.42514	1.25944	1.19293	1.50536	1.32581	1.21201	1.47266	1.29637	1.20025	1.51905	1.31591	1.20849	1.33271	1.21176	1.14319
WBU2	ROT	1.36739	1.20489	1.15115	1.41161	1.24557	1.16528	1.36864	1.22982	1.15235	1.37929	1.23506	1.16492	1.29699	1.17656	1.12662
	LSCV	1.36667	1.20502	1.15125	1.41168	1.24511	1.16538	1.36801	1.22990	1.15247	1.37945	1.23442	1.16504	1.29710	1.17667	1.12644
	BCV	1.36667	1.20502	1.15125	1.41150	1.24569	1.16538	1.36845	1.22990	1.15247	1.37869	1.23523	1.16504	1.29710	1.17667	1.12671
	SCV	1.36717	1.20502	1.15125	1.41168	1.24569	1.16538	1.36872	1.22990	1.15247	1.37902	1.23523	1.16504	1.29710	1.17667	1.12671
	JMP	1.36667	1.20440	1.15076	1.41136	1.24511	1.16492	1.36801	1.22954	1.15194	1.37869	1.23442	1.16450	1.29657	1.17616	1.12630
	PM	1.36567	1.20502	1.15137	1.41159	1.24549	1.16600	1.36761	1.22996	1.15266	1.37816	1.23488	1.16504	1.29707	1.17726	1.12686
	SJ	1.36474	1.20461	1.15116	1.41147	1.24502	1.16585	1.36708	1.22979	1.15236	1.37778	1.23437	1.16478	1.29676	1.17693	1.12648

applying distribution for $N = 50$ and 100 and on the combination of the LN3 as a parent model and GEV as an applying model for $N = 20$. The smallest values are obtained based on the combination of the WBU2 as a parent model and the GUM as an applying model for all sample sizes considered.

The SFs in Table 6 may be used to decide the 100-years design flood of Goan station incorporating the uncertainties caused by the model and data errors for a given sample size. In general, the SFs decrease as sample size increases. These values are very similar to each other regardless of kernel density estimation.

6. Conclusion

It is very important for a hydrologist to estimate flood quantiles corresponding to a given probability of occurrence. Current methods of flood frequency analysis are mainly based on the assumption that the sample of flood observations comes from known PDF. However, in the hydrological context, the population distribution function is not known exactly. Thus, alternative kernel density estimation was investigated and compared with the parametric methods for annual maximum flood data. It is also important to consider data and model errors of estimated flood quantiles. In order to consider such errors, the SFs were derived based on simulation experiments. Several conclusions obtained from this study are as follows:

1. When the parent probability distribution is the same as the applied model, almost all RBIASs of the applied models are smaller than those of any other models. Among seven bandwidth selectors, the RBIASs of SJ are the smallest in most cases.
2. As the return period increases, the RRMSEs increase while the RRMSEs decrease as the sample size increases. The RRMSEs of the GUM are smaller than those of any other models regardless of parent models applied. However, those of the WBU2 are the smallest when the WBU2 is assumed as a parent model.
3. When the WBU2 is assumed as a parent model, the RRMSEs of kernel density estimation are relatively small, while those of kernel density estimation are much bigger than those of parametric methods for the other parametric models assumed as parent models. Especially, the RRMSEs of kernel density estimation within interpolation range are much smaller than those for extrapolation range in comparison with those of parametric methods.
4. Kernel density estimation turns out to be superior to the parametric methods when a parent model has a bimodality of density because kernel density estimation can detect a bimodality of density while the parametric method cannot.
5. In general, the estimated SFs decrease as a sample size increases. Among the applied distributions, the GUM model has the smallest SFs for all parent models applied. And the GEV model has the largest values for all parent models.

Taking into consideration all the results of this study, kernel density estimation can be suggested to estimate the flood quantiles within interpolating ranges. Especially, at the moment of initial data application, the kernel density estimation is preferable to the parametric method. The SFs can give reliable flood quantiles in the design and evaluation of hydraulic structures for the assumed probability distributions.

Acknowledgements

The research leading to this paper has been sponsored by the Korea Research Foundation and Internal Research Fund of Yonsei University in Korea. Acknowledgements are also due to three anonymous reviewers who gave suggestions that improved the paper.

References

Adamowski, K., 1985. Nonparametric kernel estimation of flood frequency. Water Resour. Res. 21 (11), 1585–1590.

Adamowski, K., 1989. A Monte Carlo comparison of parametric and nonparametric estimation of flood frequencies. J. Hydrol. 108, 295–308.

Adamowski, K., 1996. A Monte Carlo comparison of parametric and nonparametric estimation of flood frequencies. J. Hydrol. 108, 295–308.

Adamowski, K., 2000. Regional analysis of annual maximum and partial duration flood data by nonparametric and L-moment methods. J. Hydrol. 229, 219–231.

Azzalini, A., 1981. A note on the estimation of a distribution function and quantiles by a kernel method. Biometrika 68, 326–328.

Bowman, A., 1985. A comparative study of some kernel-based nonparametric density estimators. J. Stat. Comp. Simul. 21, 313–327.

Falk, M., 1984. Relative deficiency of kernel type estimators of quantiles. Ann. Stat. 12, 261–268.

Guo, S.L., 1991. Nonparametric variable kernel estimation with historical floods and paleoflood information. Water Resour. Res. 27 (1), 91–98.

Guo, S.L., 1993. Parametric and nonparametric mixture density estimation with historical flood and paleoflood information. IAHR Publ. 212, 277–286.

Guo, S.L., Kachroo, R.K., Mngodo, R.J., 1996. Nonparametric kernel estimation of low flow quantiles. J. Hydrol. 185, 335–348.

Hall, P., Marron, J.S., 1987. Extent to which least-squares cross-validation minimizes integrated squared in nonparametric density estimation. Probab. Theor. Relat. Fields 74, 567–581.

Hall, P., Marron, J.S., Park, B.U., 1992. Smoothed cross-validation. Probab. Theor. Relat. Fields 92, 1–20.

Härdle, W., 1991. Smoothing Technique with Implementation in S. Springer, New York.

Härdle, W., Scott, D.W., 1992. Smoothing by weighted averaging of rounded points. Comput. Stat. 7, 97–128.

Härdle, W.W., Klinke, S., Turlach, B.A., 1995. XploRe: An Interactive Statistical Computing Environment. Springer, Berlin.

Jones, M.C., Marron, J.S., Sheater, S.J., 1992. Progress in data-based bandwidth selection for kernel density estimation. Working Paper Series, 92-014. Australian Graduate School of Management, University of New South Wales.

Kim, K.D., Heo, J.H., Cho, W.C., 1999. Risk analysis of flood quantiles based on smoothing techniques. ASCE's International Water Resources Engineering Conference, FD-02, Seattle, Washington.

Labatiuk, C., Adamowski, K., 1987. Application of nonparametric density estimation to computation of flood magnitude/frequency. Stochastic Hydrology. Reidel, Dordrecht pp. 161–180.

Lall, U., Moon, Y.I., Bosworth, K., 1993. Kernel flood frequency estimators: bandwidth selection and kernel choice. Water Resour. Res. 29 (4), 1003–1015.

Marron, J.S., 1989. Comments on a data based bandwidth selector. Comput. Stat. Data Anal. 8, 155–170.

Marron, J.S., Wand, M.P., 1992. Exact mean integrated squared errors. Ann. Stat. 20, 712–736.

Mkhandi, S.H., Kachroo, R.K., Guo, S.L., 1996. Uncertainty analysis of flood quantile estimates with reference to Tanzania. J. Hydrol. 185, 317–333.

Moon, Y.I., Lall, U., Bosworth, K., 1993. A comparison of tail probability estimators. J. Hydrol. 151, 343–363.

Nadaraya, E.A., 1964. Some new estimates for distribution functions. Theor. Probab. Appl. 15, 497–500 England translation.

Park, B.U., Marron, J.S., 1990. Comparison of data driven bandwidth selectors. J. Am. Stat. Assoc. 85, 66–72.

Park, B.U., Turlach, B.A., 1992. Practical performance of several data driven bandwidth selectors. Comput. Stat. 7, 251–270.

Parzen, E., 1962. On estimation of a probability density function and mode. Ann. Math. Stat. 33, 1065–1076.

Rosenblatt, M., 1956. Remark on some nonparametric estimates of a density function. Ann. Math. Stat. 27, 832–837.

Rudemo, M., 1982. Empirical choice of histograms and kernel density estimators. Scand. J. Stat. 9, 65–78.

Scott, D.W., 1979. On optimal and data-based histograms. Biometrika 66, 605–610.

Scott, D.W., 1985. Averaged shifted histograms: effective nonparametric density estimators in several dimensions. Ann. Stat. 13, 1024–1040.

Scott, D.W., Terrell, G.R., 1987. Biased and unbiased cross-validation in density estimation. J. Am. Stat. Assoc. 82, 1131–1146.

Sheater, S.J., 1992. The performance of six popular bandwidth selection methods on some real data sets (with discussion). Comput. Stat. 7, 271–281.

Sheater, S.J., Jones, M.C., 1991. A reliable data-based bandwidth selection method for kernel density estimation. J. R. Stat. Soc. Ser. B 53, 683–690.

Sheater, S.J., Marron, J.S., 1990. Kernel quantile estimators. J. Am. Stat. Assoc. 85, 410–416.

Silverman, B.W., 1986. Density Estimation for Statistics and Data Analysis. Chapman & Hall, London.

Stedinger, J.R., Vogel, R.M., Foufoula-Georgiou, E., 1993. Frequency analysis of extreme events. In: Maidment, D.R. (Ed.). Handbook of Hydrology. McGraw-Hill, New York, pp. 18–25 Chapter 18.

Stone, C.J., 1984. An asymptotically optimal window selection rule for kernel density estimates. Ann. Stat. 12, 1285–1297.

Terrell, G.R., 1990. The maximal smoothing principle in density estimation. J. Am. Stat. Assoc. 85, 470–477.

Terrell, G.R., Scott, D.W., 1985. Oversmoothed density estimates. J. Am. Stat. Assoc. 80, 209–214.

Turlach, B.A., 1993. Bandwidth Selection in Kernel Density Estimation: A Review. Discussion Paper 9317, Institut de Statistique, Universite Catholique de Louvain, Louvain-la-Neuve, Belgium.

Yang, S.S., 1985. A smooth nonparametric estimator of a quantile function. J. Am. Stat. Assoc. 80, 1004–1011.