韓國水資源學會論文集
第39卷 第6號 · 2006年 6月
pp. 495~502

극치값 추정에 적합한 비매개변수적 핵함수 개발

A Development of Noparamtric Kernel Function Suitable for Extreme Value

차 영 일* / 김 순 범** / 문 영 일***
Cha, Young-Il / Kim, Soon Bum / Moon, Young-Il

Abstract

The importance of the bandwidth selection has been more emphasized than the kernel function selection for nonparametric frequency analysis since the interpolation is more reliable than the extrapolation method. However, when the extrapolation method is being applied(i.e. recurrence interval more than the length of data or extreme probabilities such as 200~500 years), the selection of the kernel function is as important as the selection of the bandwidth. So far, the existing kernel functions have difficulties for extreme value estimations because the values extrapolated by kernel functions are either too small or too big. This paper suggests a Modified Cauchy kernel function that is suitable for both interpolation and extrapolation as an improvement.

keywords : nonparametric frequency analysis, extrapolation, Modified Cauchy kernel function

요 지

비매개변수적 빈도해석을 위해 제시되는 핵밀도함수 방법에서 내삽법은 외삽법보다 더 신뢰적이기 때문에 내삽법과 관련된 광역폭의 선택이 외삽 문제와 연관되는 핵함수의 선택보다 중요하다. 그러나, 재현기간이 자료구간보다 커지거나 또는 200~500년 빈도 발생과 같은 확률 값에 대한 추정을 하는 경우는 자료의 외삽이 중요한 문제이며 따라서 이에 따른 핵함수의 선택도 중요시된다. 핵함수에 따라서는 외삽에 대해 상대적으로 작거나 큰 값이 제시 될 수 있으므로 극치값 추정에는 어려운 점이 있다. 따라서 본 논문에서는 일반적으로 내삽 및 외삽에도 적합한 핵함수로 Modified Cauchy 핵함수를 제시하였다.

핵심용어 : 비매개변수적 빈도해석, 외삽, Modified Cauchy 핵함수

1. 서 론

수공구조물 계획과 설계에 있어 중요한 요소인 빈도해석은 통계학적인 관점에서 매개변수적 해석방법과 비매개변수적 해석방법의 두 가지로 나눌 수 있다. 실제 수문 관측자료는 여러 가지 복합 요인으로 인하여 발생하기 때문에 자료분포의 특징이 꼬리부분이 길게 늘어지는 경우나 첨두가 두 개(bimodal) 이상인 경우에는 단일분포형을 기초로 하는 매개변수적 빈도해석 방법으로는 많은 어려움이 있고, 또한, 자료에 적합한 최적 분포함수를 선태하기도 어렵다고 할 수 있다. 이런 경우 비매개변수적 빈도해석 방법 중 하나인 핵밀도함수(Kernel Density Function) 방법을 사용하면 이러한 문제점의 해소와 더불어 해석적으로도 좋은 결과를 얻을 수 있다고 알려 지고 있다(Lall et al., 1993; Moon et al., 1993; Adamowski, 1996). 그러나 핵함수의 선택

* 한국종합기술개발공사 과장 · 공학박사 (e-mail: ycha@kecc.co.kr)
** 서일대학 토목과 부교수 (e-mail: ksb5825@seoil.ac.kr)
*** Corresponding Author · 서울시립대학교 토목공학과 부교수 (e-mail: ymoon@uos.ac.kr)

보다는 광역폭의 선택만으로도 재현기간이 자료범위보다 작거나 100년 내외의 내삽에 해당하는 경우에는 좋은 결과를 보였고 핵함수의 선택이 민감한 결과를 보이지는 않았다. 이와 달리 500년 내외의 외삽의 경우는 광역폭의 선택과 함께 Cauchy 핵함수를 적용함으로써 좋은 결과를 보였다. 하지만 댐 위험도 분석을 위한 Monte Carlo Simulation에 따른 50년 또는 100년 자료를 100,000번 이상 모의하는 경우와 같이 극치에 가까운 외삽의 경우, 큰 재현기간이나 또는 가능최대수문량에 가까운 추정치가 적용되며, 이런 경우 기존의 핵함수는 단점이 된다. 따라서 본 논문은 이러한 단점을 보완하여 극치값을 고려한 내삽과 외삽에 적합한 새로운 핵함수를 개발하는데 목적을 두고자 한다. 연구의 범위와 방법은 다음 Fig. 1과 같다.

2. 비매개변수적 핵밀도함수법

2.1 일반식

Rosenblatt(1956)는 모든 자료가 발생되어진 각각의 위치에 따른 막대그래프의 box 중앙에 위치하도록 하여, 구간을 이동시킬 수 있는 이동 히스토그램을 발달시켜, 핵밀도함수 추정법을 발표했는데, 모든 실수 x에 대하여 다음 Eq.(1)과 같이 정의하였다.

$$\hat{f}_h(x) = n^{-1}\sum_{i=1}^{n} K_h(x - X_i) \tag{1}$$

여기서, $K_h(\cdot) = \frac{1}{h}K\left(\frac{\cdot}{h}\right)$ 인 핵함수이고, $X_1, X_2, \cdots, X_n$은 독립적으로 동일하게 분포된 실관측치이다. h는 n이 무한대로 갈 때 영(zero)으로 접근하는 값을 갖는 양의 광역폭(bandwidth)이다.

2.2 광역폭 선택

비매개변수적 핵밀도 함수법에서 광역폭(bandwidth) h의 선택은 매우 중요한 문제이다. 광역폭을 선택하는 방법에는 rule of thumb, least squares cross validation, biased cross validation, cross validated IMSE(Integrated Mean Squared Error), maximum likelihood, 또는 Adamowski Criterion, smoothed bootstrap, plug-in method(Sheather et al., 1991) 등이 있고, 이들로부터 최적의 h를 구할 수 있다.

h의 값은 핵함수 추정 법에 있어서 매우 중요하지만 실제적으로 정확한 추정은 쉽지 않다. 광역폭 h가 크면 편차(bias)가 크게 나타나며 너무 완만(over-smooth)한 밀도함수의 추정과 정보의 손실을 가져온다.

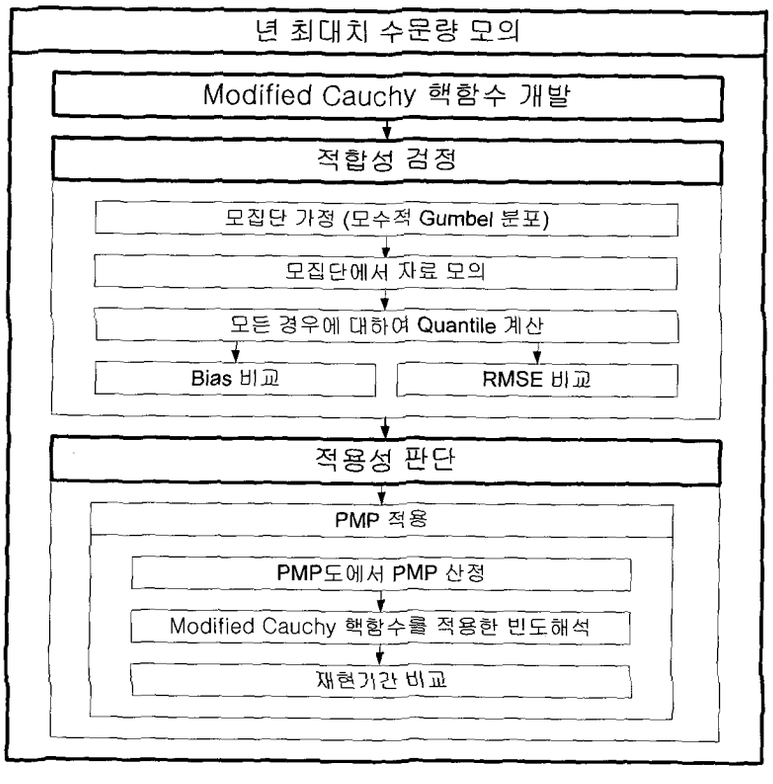


Fig. 1. Procedure of extreme probabilities using Modified Cauchy Kernel Function

반면에 광역폭 h의 선택이 작은 경우에는 분산(variance)이 커지고 거친(rough) 추정치를 나타낸다(Adamowski and Labatiuk, 1987).

2.3 핵함수

연속함수 형태인 핵함수(kernel function)를 사용하는 경우 대표적인 핵밀도함수는 다음의 Table 1과 같다.

이와 같은 핵함수(kernel function)들은 일반적으로 다음의 조건을 만족한다.

$$\int K(t)dt = 1 \tag{2}$$

여기서, $t = \frac{x - X_i}{h}$ 이고, 이 때 x는 임의의 점이고, X_i는 실 관측된 자료이다. Table 1의 그래프에서 보이는 바와 같이 핵함수들은 t=0에서 최대치를 갖고 연속이며 그리고 방정식의 형태가 대칭적이다. 즉, Eq.(2)와 같이 핵함수의 면적은 1이고 기대값은 영$\left(\int t\,K(t)\,dx = 0\right)$과 유한한 분산$\left(\int t^2 K(t)dt = \text{constant}\right)$을 갖는 특징이 있다. 그러나 때로는 이 특성들을 만족하지 않는 핵함수가 사용되어질 수도 있다.

3. Modified Cauchy 핵함수

Table 1에서 제시된 바와 같이 일반적인 핵함수로는 Gaussian, Rectangular, Epanechnikov, Cauchy 등이 많이 사용되고 있다. 하지만 이러한 핵함수들은 꼬리가 얇거나 경계가 있어도 재현기간이 자료의 범위 이내이거나 100년 내외인 내삽에 적용하는 경우 민감한 결과를 가져오지는 않았다. 반면에 내삽과 달리 외삽의 경우에는 적용하는데 문제점을 나타내기 때문에 이를 보완하기 위해 꼬리가 두꺼운 Cauchy 핵함수를 적용하며, 이런 경우 외삽 문제를 극복할 수 있지만 극치로 발생되는 가능최대강수량(PMP)이나 가능최대홍수량(PMF) 등에서는 quantile 값이 크게 발산되는 단점이 있다.

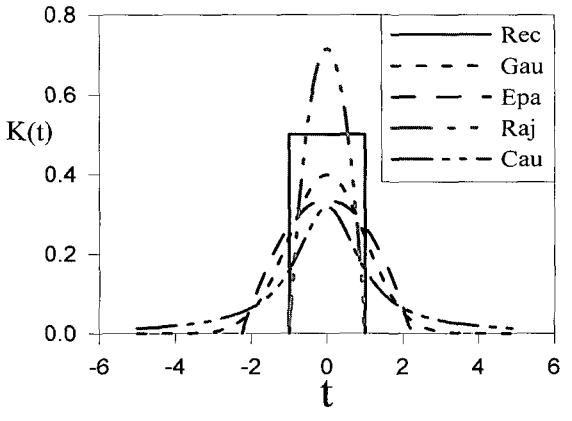
따라서 본 연구에서는 Gaussian 핵함수와 Cauchy 핵함수의 단점들을 보완한 새로운 핵함수를 개발하여 Modified Cauchy 핵함수로 명하였다. Eqs. (3) and (4)는 각각 Modified Cauchy 핵함수의 확률밀도함수와 누가분포함수이다. Modified Cauchy 핵함수는 t-distribution에서 유도한 함수로서 위에서 언급한 핵함수의 조건들과 확률밀도함수의 조건들을 모두 만족하는 것으로 분석되었다.

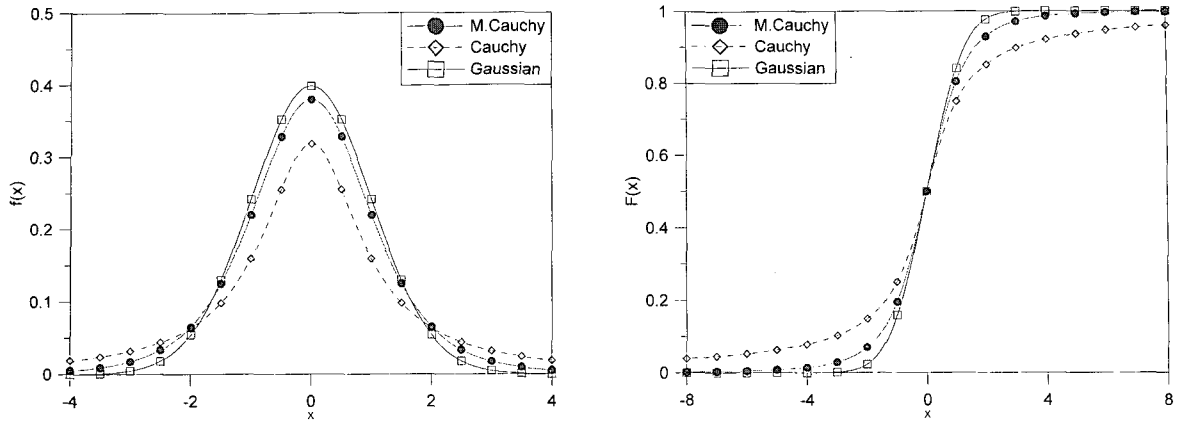
$$k(x) = \frac{8}{3\sqrt{5}\,\pi(1 + x^2/5)^3} \tag{3}$$

$$K(x) = \frac{1}{2} + \frac{5x + \frac{3x^3}{5}}{3\sqrt{5}\,\pi\left(1 + \frac{x^2}{5}\right)^2} + \frac{1}{\pi} tan^{-1}\left(\frac{x}{\sqrt{5}}\right) \tag{4}$$

다음의 Figs. 1(a) and 2(b)는 여러 가지 핵함수들 중에서 Gaussian, Cauchy 그리고 Modified Cauchy 핵함수들에 대한 각각의 PDF(확률밀도함수)와 CDF(누가분포함수)를 비교 도시한 것이다. Fig. 2(a)에서 제시된 여러 핵함수에 대한 PDF에서는 Gaussian 핵함수는 꼬리 부분이 상대적으로 얇은 반면에 Cauchy 핵함수는

Table 1. Some kernel functions

Kernel	K(t)	Shape
Rectangular	$\frac{1}{2}$ for $\lvert t\rvert < 1$	
Gaussian	$\frac{1}{\sqrt{2\pi}}\exp(\frac{-t^2}{2})$	
Epanechnikov	$\frac{3}{4}(1-\frac{1}{5}t^2)/\sqrt{5}$ $\lvert t\rvert < \sqrt{5}$	
Rajagopalan	$\frac{3h}{(1-4h^2)}(1-t^2)$ $\lvert t\rvert \le 1$	
Cauchy	$\frac{1}{\pi(1+t^2)}$	



(a) Probability Density Function(PDF) (b) Cumulative Distribution function(CDF)

Fig. 2. PDFs and CDFs of some kernel functions

꼬리부분이 다소 두껍게 나타나고 있음을 알 수 있다. 또한, Fig. 2(b)의 경우는 Gaussian 핵함수의 CDF는 작은 Quantile 값에서 다른 핵함수들보다 더 빠르게 1에 수렴됨을 보이며 이와 달리 Cauchy 핵함수는 보다 느린 수렴성을 보임을 알 수 있다. 이러한 특성을 갖는 각각의 핵함수를 댐 위험도 분석을 위한 극치강수량에 적용하는 경우 Gaussian 핵함수는 PMP 보다도 작게 추정되어 댐 위험도 측면에서 안전측보다 작은 추정이 이루어지는 반면에 Cauchy 핵함수는 PMP보다 크게 추정되는 단점이 있다. 그러나 Modified Cauchy 핵함수는 Fig. 2(a)와 같이 PDF에서 첨두와 꼬리가 Gaussian 핵함수와 Cauchy 핵함수 사이에 존재하고, Fig. 1(b)에서는 적당한 수렴값을 보이므로 이러한 특성을 이용해 제시된 다른 두 핵함수가 갖는 단점들을 개선 보완 할 수 있다.

4. 결 과

4.1 모의기법에 의한 비교

핵함수들의 적합성을 비교하기 위해 Monte Carlo모의에 의한 방법을 사용하여 각 모형별로 누가분포함수 확률값에 대한 Quantile 값의 Bias와 Root Mean Square Error(RMSE)를 비교하였다.

$$\text{Bias} = \sum \frac{\hat{x}(F) - x(F)}{N} \tag{5}$$

$$\text{SE} = \left[\sum \frac{\{\hat{x}(F) - x(F)\}^2}{N} \right]^{1/2} \tag{6}$$

여기서, $\hat{x}(F)$는 누가확률F에 대한 추정 Quantile값, $x(F)$는 모집단의 Quantile값이다.

모의할 때 기준이 될 모집단은 『1999년도 수자원관리기법개발연구조사 보고서』(건설교통부, 2000)에서 한국의 강수량자료를 기준으로 했을 때 가장 많은 지점에서 적합판정을 받은 Gumbel 분포형으로 가정 하였고, 모집단에 대하여 매개변수적 Gumbel (Gumbel-Gumbel), 비매개변수적 고정 Gaussian (FK-SJ-GA), Epanechnikov (FK-SJ-EP), Cauchy (FK-SJ-CA), Modified Cauchy (FK-SJ-MC) 핵함수와, 비매개변수적 변동 Gaussian (VK-LSCV-GA), Epanechnikov (VK-LSCV-EP), Cauchy (VK-LSCV-CA), Modified Cauchy (VK-LSCV-MC) 핵함수를 각각 50년의 자료와 100년의 자료를 100,000번 모의하여 비교하였다. 광역폭이 일정한 고정 핵함수법의 광역폭 선택은 Sheather and Jones(1991)의 plug-in method(SJ)를 사용하였고, 자료에 따라 광역폭이 변하는 변동 핵함수법은 least squares cross validation(LSCV) 방법(Moon 등, 1993)을 적용하였다. 모의에 의한 비교 과정은 다음 순서와 같다.

① 모집단으로 가정한 Gumbel 분포에서 50년, 100년 자료 모의
② 모의된 자료로부터 각 모형에 대하여 CDF값, P = 0.95, 0.98, 0.99, 0.995, 0.998, 0.999에 해당하는 Quantile 값 계산
③ ①~②을 100,000번 반복
④ 모의된 자료의 Quantile 값과 모집단의 Quantile 값에 대하여 Bias 와 RMSE를 산정

Fig. 3에서 Fig. 10은 모의 결과의 Bias와 RMSE를 보여주고 있다. 가로축은 누가분포함수값(F(x))이고, 세

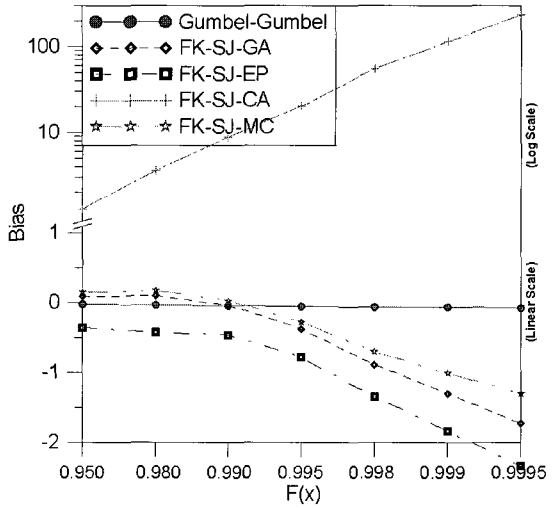


Fig. 3. Bias of fixed kernel estimator using 50 years data

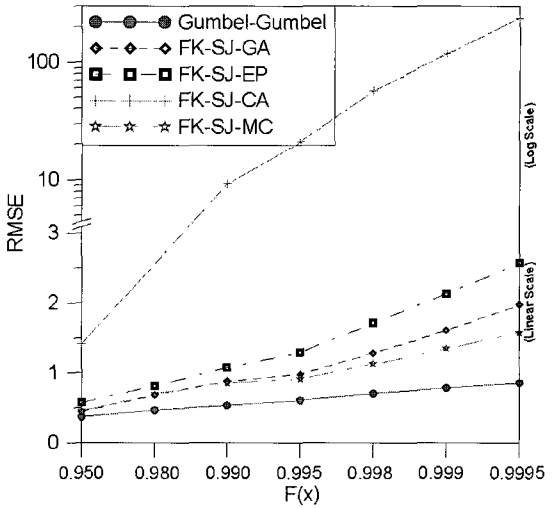


Fig. 4. RMSE of fixed kernel estimator using 50 years data

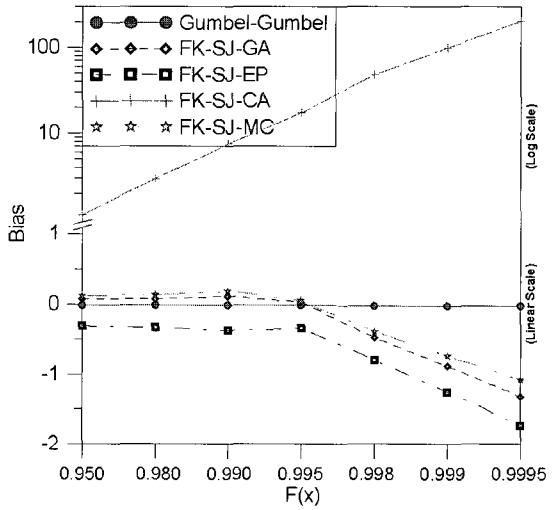


Fig. 5. Bias of fixed kernel estimator using 100 years data

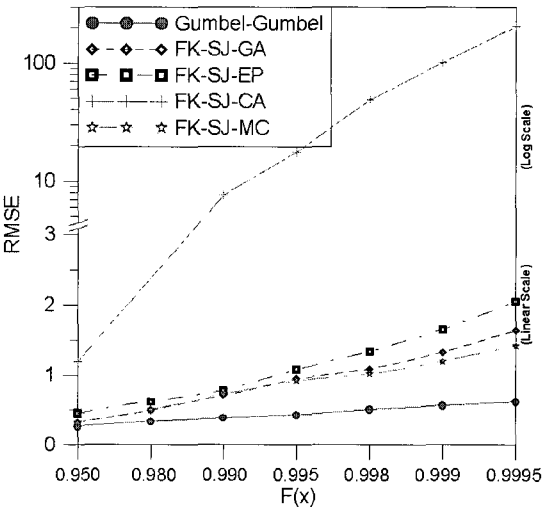


Fig. 6. RMSE of fixed kernel estimator using 100 years data

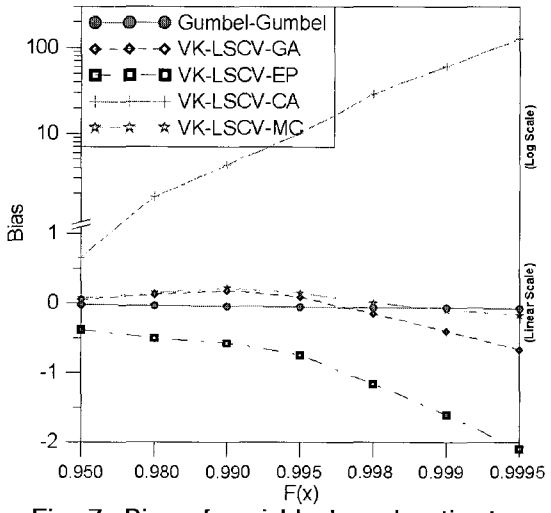


Fig. 7. Bias of variable kernel estimator using 50 years data

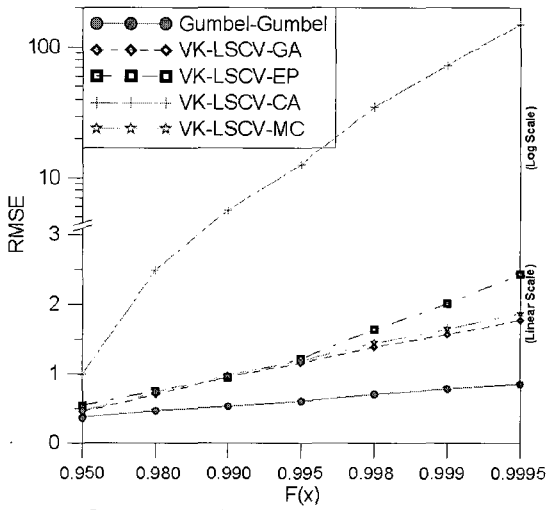


Fig. 8. RMSE of variable kernel estimator using 50 years data

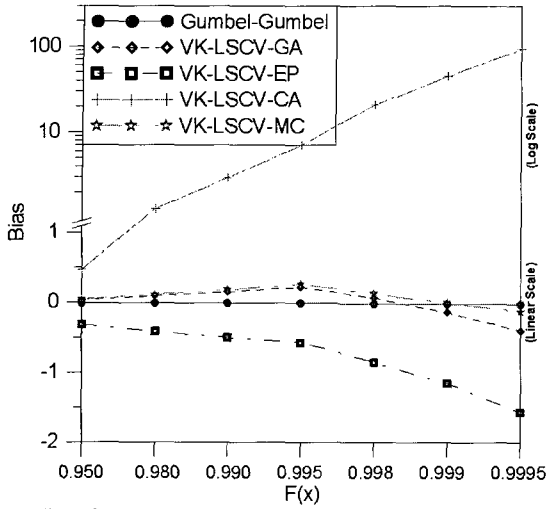


Fig. 9. Bias of variable kernel estimator using 100 years data

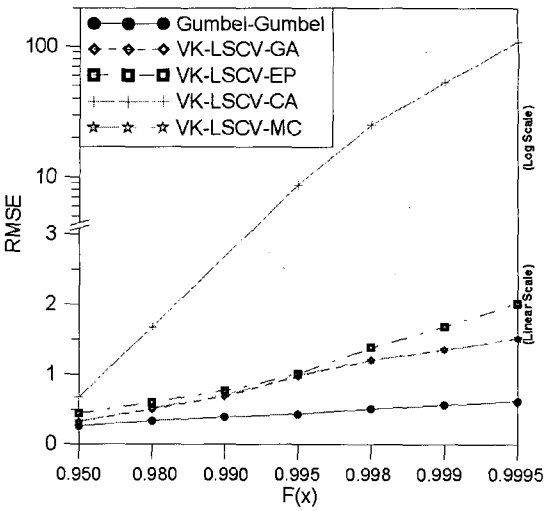


Fig. 10. RMSE of variable kernel estimator using 100 years data

로축은 Bias 값과 RMSE 값이며 Cauchy 핵함수의 Bias 값과 RMSE 값이 너무 크기 때문에 축 가운데를 기준으로 하부는 Linear Scale, 상부는 Log Scale로 표시하였다. Gumbel 분포형을 모집단으로 가정하여 모의한 결과 모집단으로 가정한 Gumbel이 가장 작은 Bias 값과 RMSE 값을 보였고, 자료의 크기에 상관없이 변동핵밀도함수법에서도 50년 자료에 대한 RMSE 값만 제외한 모든 결과에서 F(x) 값이 커질수록 비매개변수적 Modified Cauchy 핵함수의 Bias 값과 RMSE 값이 더 작은 결과를 보였다. 고정핵밀도함수법에서는 자료의 개수에 상관없이 50개나 100개 모두에서 Modified Cauchy 핵함수가 가장 좋은 결과를 보였고, 핵함수의 특성상 Cauchy 핵함수에 의한 추정값은 F(x) 값이 커질수록 Bias 값과 RMSE 값이 크게 증가하는 단점을 보였다.

변동핵밀도함수법에서도 50년 자료에 대한 RMSE만 제외하고는 모든 결과에서 Modified Cauchy 핵함수법이 가장 좋은 결과를 보였다. 결과적으로 고정핵밀도함수법과 변동핵밀도함수법 모두에서 Modified Cauchy가 가장 좋은 결과를 보였고, Gaussian, Epanechnikov, Cauchy 핵함수 순으로 좋은 결과를 보였다.

4.2 가능최대강수량(PMP) 비교

실제 적용성의 비교를 위해 춘천지점의 지속기간 24시간 강수량을 이용하여 각각의 방법으로 재현기간별 확률강수량을 산정한 후, 춘천소재의 소양강댐 유역의 24시간 PMP와 재현기간을 비교하였다. 춘천지점의 소양강댐 유역(유역면적 2,703㎢)에 대해 최근에 건설교통부(2000)에서 제시한 [한국 가능최대강수량도]를 사용하여 소양강댐 유역의 DAD곡선을 작성하고 PMP를 구하였다. Fig. 11.은 소양강댐 유역의 DAD 곡선을 나타내는 것이고, 소양강댐 유역의 24시간 PMP는 573mm로 계산 되었다. Fig. 12.는 매개변수적 Gumbel (Gumbel), 비매개변수적 고정 Gaussian (FK-SJ-GA), Epanechnikov (FK-SJ-EP), Cauchy (FK-SJ-CA), Modified Cauchy (FK-SJ-MC) 핵함수로, Fig. 13 은 매개변수적 Gumbel (Gumbel) 비매개변수적 변동 Gaussian (VK-LSCV-GA), Epanechnikov (VK-LSCV-EP), Cauchy (VK-LSCV-CA), Modified Cauchy (VK-LSCV-MC) 핵함수로 빈도해석을 실시한 결과이다.

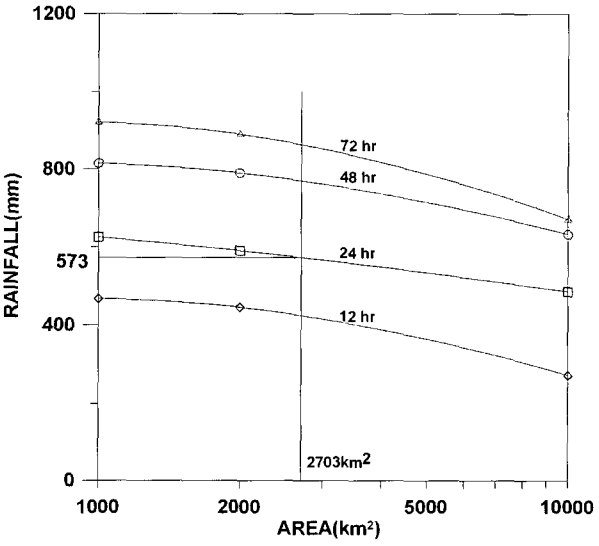


Fig. 11. DAD of Soyang river dam basin

Fig. 12에서 비매개변수적 고정 Modified Cauchy 핵함수를 적용한 방법은 PMP 값을 $10^5 \sim 10^6$년 정도의 재현기간에 보이고 있는 반면에, 비매개변수적 고정

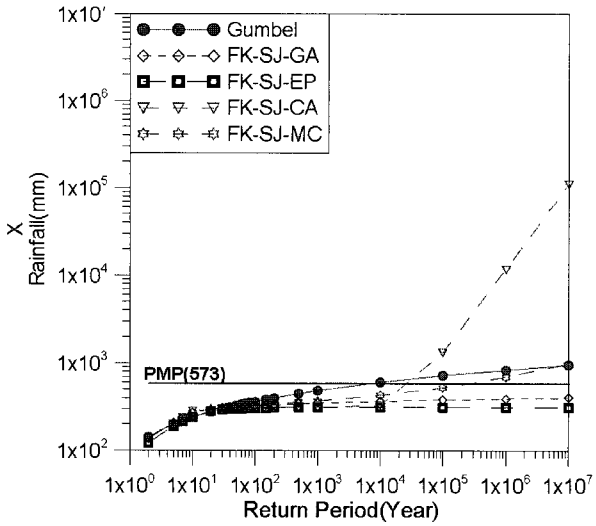


Fig. 12. PMP of Chuncheon -Fixed kernel estimator

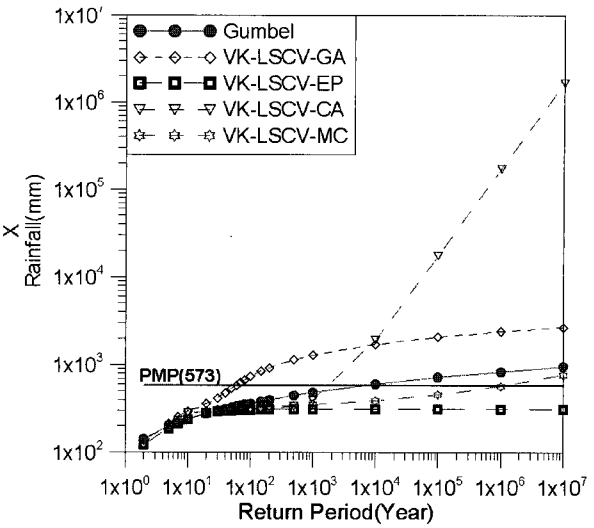


Fig. 13. PMP of Chuncheon -Variable kernel estimator

Cauchy 핵함수를 적용한 방법들은 모두 PMP와 동떨어진 매우 큰 값으로 증가하고 있다. 매개변수적 Gumbel은 10^3~10^4년의 재현기간을 보여 PMP라고 하기에는 작은 재현기간을 보이고 있고, Gaussian, Epanechnikov 핵함수는 107년 까지도 PMP에 비해 너무 작은 값을 보이고 있다. Fig. 13에서도 비매개변수적 Modified Cauchy 핵함수를 적용한 방법은 PMP 값을 10^6년 정도의 재현기간에 보이고 있는 반면에 Cauchy 핵함수는 너무 큰 확률강수량값을 제시하고, Gaussian 핵함수는 다소 큰 확률강수량값을 제시하고 있다.

5. 결 론

지금까지의 비매개변수적 수문빈도 해석은 핵함수의 선택보다는 광역폭의 선택을 더 중요시해왔고, 실질적으로 재현기간 100~200년 내의 내삽에 해당하는 해석에서 핵함수의 선택은 결과에 그리 민감하지 못하였다. 그러나 매우 큰 재현기간이나 가능최대수문량과 같이 극치에 가까운 외삽에 대한 추정치를 원할 경우나 Monte Carlo 방법에 의한 댐 위험도분석과 같이 극치값을 필요로 하는 경우에는 광역폭의 선택만큼 핵함수의 선택도 매우 중요하다. 따라서 본 논문에서는 내삽 및 외삽에 적합한 핵함수로 Modified Cauchy 핵함수를 개발하였다. Modified Cauchy 핵함수는 고정 및 변동핵 밀도함수법에서 모의 기법으로 추정한 누가분포함수 확률값에 대한 Quantile 값의 Bias와 RMSE가 비매개변수적 방법 중 가장 좋은 결과를 보였다. 실측 자료인 춘천지점의 24시간 강수량자료의 확률강수량과 춘천소재 소양강댐유역의 24시간 PMP를 비교한 결과에서도 다른 비매개변수적 방법보다 Modified Cauchy 핵함수 방법이 타당한 결과를 보였다. 기존의 핵함수들은 재현기간이 커지면 커질수록 너무 크거나 작은 Quantile 값으로 증가하여 Bias 및 RMSE 값이 크게 증가하기 때문에 외삽의 경우 사용하기에 적합하지 않은 것으로 판단된다. 따라서 앞으로의 비매개변수적 수문빈도해석은 기존 핵함수 보다는 Modified Cauchy 핵함수를 선택하여 사용하는 것이 바람직하고, 특히 외삽의 경우 더욱 권장할 수 있을 것이다.

감사의 글

본 연구의 일부는 2004년도 서일대학 학술연구비 지원에 의해 수행되었으며, 이에 감사드립니다.

참 고 문 헌

건설교통부 (2000). 1999년도 수자원관리기법개발연구조사 보고서.

Adamowski, K., and Labatiuk, C. (1987). "Estimation of flood frequencies by a non-parametric density procedure." *Hydrologic Frequency Modeling*, pp. 97~106.

Adamowski, K. (1996). "Nonparametric Estimation of Low-Flow Frequencies." *Journal of Hydraulic Engineering*, Vol. 122, No. 1, pp. 46~49.

Lall, U., Moon, Young-Il, and Bosworth, K. (1993). "Kernel flood frequency estimators: bandwidth selection and kernel choice." *Water Resources Research*, Vol. 29, No. 4, pp. 1003-1015.

Moon, Young-Il, Lall, U., and Bosworth, K. (1993). "A comparison of tail probability estimators." *Journal of Hydrology*, Vol. 151, pp. 343-363.

Rosenblatt, M. (1956). "Remarks on some non-parametric estimates of a density function." *Ann. Math. Statist.*, Vol. 27, pp. 832~837.

Sheather, S. F., and Jones, M. C. (1991). "A reliable data-based band-width selection method for kernel density estimation." *J. Roy. Statistical Soc.*; B., Vol. 53, pp. 683~690.

(논문번호:05-126/접수:2005.09.20/심사완료:2006.04.26)