Boundary Kernel 함수를 이용한 빈도해석

○문영일[1)]

1. 서 론

임의의 관측지점에서 홍수량 및 강수량 빈도해석 방법에 대한 우리나라 대부분의 연구는 관측자료를 특정 확률밀도함수(PDF)를 가진 모집단으로부터 나왔다는 가정에 근거한 통계학적인 관점에 집중해왔다. 지금까지 많은 홍수량 및 강우량 빈도해석에 대한 방법들이 발표되었고 많은 수문학자들이 최선의 분포형 선정 방법 (χ^2 검정, Kolmogrov-Smirnov 검정, Cramer Von Mises 검정, 편차검정 등)과 이러한 분포형에 대한 최적의 매개변수 추정방법 (모멘트법, 최우도법, 확률가중모멘트법, L-모멘트법 등)에 대한 연구를 계속해 왔다. 그러나 항상 어떤 단일 확률밀도함수가 관측된 자료에 대한 최선의 선택이 될 수는 없다. 이러한 전통적인 매개변수적인 추정방법은 가정된 매개변수적 확률밀도함수가 관측자료의 통계학적 특성에 잘 맞는다할 지라도 빈도해석에서의 홍수량의 크기에는 상당량의 불확실성이 존재한다. 관측 지점에서 연 최대홍수량은 여러 가지 (융설, 강우유출, 계절성 폭우 등)원인을 나타내는데, 이는 통계학적으로 이질적인 모집단 또는 복합분포형 등을 초래하게 된다. 일반적으로 n=20~90 정도의 짧은 기록들로부터의 임의 또는 미지의 모집단의 한정된 혼합을 동일하게 보는 것은 그리 논리적이라고 볼 수 없고 일반적으로 수행되지도 않는다. Webb과 Betancourt (1992)는 폭우의 형태를 분류하고 분류한 폭우를 근간으로 사상을 분리시켜, 각각의 폭우형태에 따라 매개변수적 확률밀도함수를 적용시켜 결과적으로 추정 값을 조화시키려는 시도를 하였다. 이는 홍수가 복합 과정들로부터 발생된다라는 것에 좋은 증거가 된다 할 지라도 이러한 과정을 실무에서 쉽게 실행되기 어렵다. 복합의 후미거동은 가장 두꺼운 꼬리(tail)를 가지는 혼합된 분포형에 대응하는 꼬리거동에 의해 그리고 각각의 요소에 대응되는 사건들의 상대비에 의해 규정되어진다. 그러므로 복합화된 상황에서도 제한된 관측자료로부터 이론적으로 명확하고 재현기간 100년 이상의 사상에 대한 외삽에 대해서도 타당한 답을 줄 수 있는 방법이 필요하다.

따라서 본 연구에서는 관측자료 (y_i, $i=1,2,\ldots,n$)와 임의의 도시위치공식(plotting position formula) (p_i, $i=1,2,\ldots,n$)을 사용하여 단일 자료 분포형뿐만 아니라 혼합된 분포형에서도 일관성 있게 적용할 수 있는 Boundary Kernel 밀도함수를 이용한 빈도해석 방법을 제시하였다. 이 해석방법은 기본적인 확률밀도함수에 대한 가정이 없다는 점에서 또는 꼬리의 거동이 뚜렷하게 만들어지고 추정량들이 데이터의 경험적 빈도 값을 핵함수로 완화(smoothing)시킨다는 점에서는 완전하게 비매개변수적 기법이다. 지금까지 많은 비매개변수적 해석방법이 홍수량 및 갈수량의 산정에 적용(Lall 등, 1993; Moon 등, 1993; Moon과 Lall, 1994; Moon, 1996; Adamowski, 1996; Guo, 1996; Moon, 2000)되고 있다. 이러한 방법은 임의의 분위값 (quantile)에 대한 초과확률을 구하는 방식을 취한 핵밀도함수(Kernel Density Function)방법이나 본 연구에서는 반대로 임의의 초과확률에 대한 분위값을 구하는 방식을 취하였으며 일반적으로 관심이 되는 상위 빈도값

1) 서울시립대학교 토목공학과 조교수

($0.9<p<1$)에 대한 외삽은 Boundary Kernel 함수를 사용하여 빈도분석을 하였다.

2. Boundary Kernel 함수에 의한 추정식

Boundary Kernel 함수에 의한 빈도해석 추정식의 구성은 관측자료의 경험적 발생확률 값과 핵함수의 핵완화(kernel smoothing)에 근거를 두고 있다. 경험적 발생확률은 임의의 표준도시공식을 통하여 구할 수 있다. 만약 관측자료 $y_i(i=1,2,\ldots.n)$를 내림차순으로 배열하면 순차적 연 최대계열치가 구성되고 $p_i(i=1,2,\ldots n)$는 표준도시공식을 이용하여 경험상으로 추정되는 확률값으로 간주하여 그래프 상에 대응하는 도시위치를 갖게 할 수 있다. 여기서 표준도시공식으로 Adamowski (1981) 공식을 사용하면 아래 식(1)과 같다.

$$p_i = \frac{i-0.25}{n+0.5} \tag{1}$$

경험적 빈도해석의 기본이 되는 함수는 각각의 p_i에 대응되는 관측값 y_i로 결정지어질 수 있다.

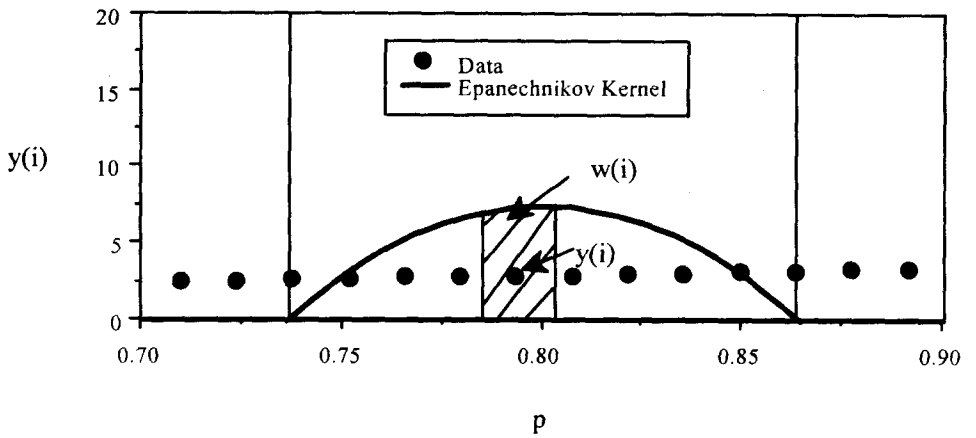

그림 1. x(0.8)에서의 핵함수 회귀추정 (h=0.064).

따라서 추정되는 빈도해석 함수의 분위값 (quantile) x(p)는 Gasser-Muller (1984)가 식(2)와 같이 제시한 핵함수를 이용한 회귀추정식으로부터 추정할 수 있다. 이는 그림1과 같이 하나의 가중함수인 핵함수를 이용하여 경험적 빈도해석 함수의 포선형(convolution)을 고려한 것이다.

$$\hat{x}(p) = \sum_{i=1}^{n} \frac{1}{h} \int_{s_{i-1}}^{s_i} y_i K\left(\frac{p-u}{h}\right) du = \sum_{i=1}^{n} \frac{1}{h} y_i \int_{s_{i-1}}^{s_i} K\left(\frac{p-u}{h}\right) du = \sum_{i=1}^{n} y_i w_i \tag{2}$$

여기서 h는 점 p와 관련된 대역폭(bandwidth), K(.)는 핵함수(kernel function)이며, $s_i=(p_i+p_{i+1})/2$ ($i=1,\ldots,n-1$), $s_0=0, s_n=1$이다. 이 때 p는 초과확률을 나타내는 구간 [0,1]에서의 임의의 값이다.

핵함수 K(t)는 대개 t=0에서 최대치를 가지고 연속적이며, 대칭인 방정식의 형태를 가진다. 또한 $\int K(t)\,dt = 1$이며 $t = (p-u)/h$에서 유한한 분산값($\int t^2 K(t)dt = \text{constant}$) 가진 대칭적 확률밀도이다. 대역폭 h는 중요한 인자로서 추정되는 회귀식함수의 완만함과 거칠기를 결정한다. 작은 대역폭은 임의의 점에서 x(p) 산정에 적은 관측자료가 고려되고, 보다 큰 대역폭에서는 상대적으로 많은 관측자료가 고려되어 유연한 회귀식을 만든다. 따라서 대역폭이 증가함에 따라 편의(bias)는 증가하고 분산은 감소한다. Muller(1991)는 최적의 광역폭 선택을 위한 평균제곱오차(Mean Squared Error)를 아래 식 (3)과 같이 제시하였다.

$$MSE(\hat{x}(p)) = E[\hat{x}(p) - x(p)]^2$$

$$\sim \frac{\sigma^2}{nh}\int_{-1}^{q}(K_x(q,t))^2dt + \frac{1}{4}h^4\{x''(p)\}^2\left\{\int_{-1}^{q}K_x(q,t)t^2dt\right\}^2 \qquad (3)$$

식 (3)에서 첫 번째 항은 분산의 추정치를 의미하며 두 번째 항은 편의의 제곱이다. 편의와 분산을 고려한 최적화된 대역폭을 찾는 방법에는 Gasser 등(1991) 에 의해 제안되었다. Gasser 등(1991)에 의해 0<p<1에 걸친 범위에 대해 식(3)에서 평균제곱오차(MSE)를 최소화시키는 최적의 대역폭은 아래의 식(4)와 같이 제시하였다.

$$h = \left\{\frac{1.5}{n}\frac{c_1}{c_2}\frac{\sigma^2}{\int_0^1\{x''(p)\}dt}\right\}^{0.2} \qquad (4)$$

여기서 $c_1 = 2\int_{-1}^{1}K_x(q,t)^2dt$, $c_2 = 4\int_{-1}^{1}K_x(q,t)t^2dt$.

일반적으로 빈도해석 수행시 상위 초과확률(0.9<p<1)에 관심을 가지게 된다. 그러나 홍수량 및 강수량 등 재현기간별 빈도해석시 주어지는 전형적인 관측자료의 크기는 20~90개의 범위를 가지게 된다. 결과적으로 $p > p_n$에 대한 빈도해석시 주어진 자료의 외삽이 필요하게 된다. 그러나 지금까지 사용한 그림1의 Epanchenique 핵함수와 같이 내부 핵(Interior Kernel) 함수를 사용하면 그림2와 같이 한정된 영역(p<1)을 벗어나기 때문에 경계 핵(Boundary Kernel) 함수가 필요하다. 경계 핵함수는 구간 [1-h, 1] 내에서 가중된 포선형의 편의를 제거하기 위해 필요하다.

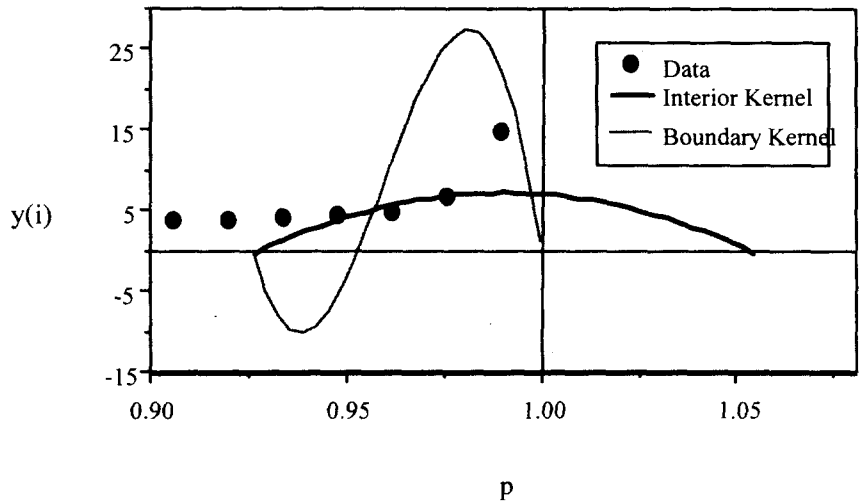

그림 2. x(0.99)에서의 핵함수 회귀추정 (h=0.064).

Muller(1991)는 내부 핵함수에 대응하는 경계 핵함수를 식(3)의 평균제곱오차(MSE) 의 최적화에 따라 제시하였다. 여기에서 Epanechnikov 핵함수를 사용하여 내부 핵함수와 경계 핵함수를 살펴보면 각각 아래의 식(5), 식(6)와 같다.

내부 핵함수 : $K(t)=0.75(1-t^2)$, 여기서 $|t|<1$ (5)

경계 핵함수: $K(q,t)=6(1+t)(q-t)\frac{1}{(1+q)^3}\left\{1+5\left(\frac{1-q}{1+q}\right)^2+10\frac{1-q}{(1+q)^2}t\right\}$ (6)

여기서 $t=\frac{(p-p_i)}{h}$ 이며, $K(q,t)$는 오른쪽 경계면 구간 [1-h, 1]에서 사용되는 경계 핵함수이며 $q=\frac{1-p}{h}$ 이다. 반면에 왼쪽 경계면 구간 [0, h]에서는 $q=\frac{p}{h}$ 이다. 이 때 q=1이면 경계 핵함수 식(6)은 내부 핵함수 Epanechnikov 핵함수와 동일하다. 여기서 홍수량 등 연 최대계열 자료에 대한 빈도해석에는 오른쪽 경계면을 고려하여야 하며, 갈수량 등 연 최소계열 자료에서는 왼쪽 경계면을 고려한 경계 핵함수를 사용하여야 한다.

3. 결 과

기상청 서울관측소의 연 강수량 자료(1908년~1998년)를 사용하여 재현기간 100년에 해당하는 강수량을 경계 핵밀도함수에 의한 회귀식 추정과 함께 기존의 매개변수적 방법에 의한 빈도해석을 수행하였다. 일반적으로 매개변수에 의한 확률강수량의 추정은 Chi-square 또는 Kolmogrov-Smirnov 검정을 통과한 분포형이더라도 연 강수량에서 작게는 모멘트법의 경우 4%, 크게는 확률가중 모멘트법에서 20% 이상의 차이를 보이고 있으며, 같은 분포형일지라도 매개변수 추정방법에 따라 10%정도의 차이를 보이고 있다. 여러 매개변수적 분포형에 따른 결과와 경계 핵

함수에 의한 결과가 표1에 정리되어 있다. 경계 핵밀도함수에 의한 재현기간 100년 연 강수량은 2,368mm이나, 반면에 매개변수적인 방법에 의한 서울의 100년 빈도 년 강우량은 선택된 분포형과 매개변수 산정방법에 따라 2,060~2,700mm사이에서 값이 다르게 나타났다. 즉 모멘트법(MOM)을 사용한 매개변수 산정시 100년 빈도 년 확률강우량은 2293~2386mm, 최우도법(ML)은 2193~2504mm, 확률가중모멘트법(PWM)은 2284~2756mm, L-모멘트법은 2215~2442mm이다.

표 1. 서울의 100년 빈도 년 강수량(mm) (□ 최대, 최소)

분포형	L-M	PWM	ML	MOM	Boundary Kernel
Gamma2	2215				
Gamma3		2309		2304	
GEV	2331	2351	2257	2296	
Gumbel	2402	2402	2504	2386	
Log-Gumbel2		2756			
Log-Normal2		2284	2319	2306	
Log-Normal3	2231	2330	2252	2293	2368
Weibull			2193		
Wakeby	2330				
Pearson T3	2309				
G-Logistic	2445				
최대	2442	2756	2504	2386	
최소	2216	2284	2193	2293	

4. 결 론

홍수량 또는 강우량 자료 등 최대치 계열에 대한 빈도해석시 사용되는 분포형의 종류는 지금까지 다양하게 소개되어 왔고, 많은 수문학자들이 관측자료에 대한 최적의 분포형과 이에 대한 적절한 매개변수 추정 방법에 대한 연구를 수행하고 있다. 이러한 매개변수에 의한 확률강수량 또는 홍수량의 추정은 한 지점에서 여러 개의 확률분포형이 적합도 판정을 받는 경우가 일반적이고 또한 매개변수 산정 방법에 따라 추정값의 차이가 커 재현기간별 적절한 추정값을 선택하는데 어려움이 있다. 반면에 비매개변수적 빈도해석 방법은 어느 특정 분포형의 가정이 없어 분포형 선택의 어려움이 해소된다. 일반적인 비매개변수적 빈도해석 방법인 고정대역폭 핵밀도함수 추정방법에서 자료에 따라 나타나는 외삽의 문제를 경계 핵함수를 사용하여 해결할 수 있었고 경계 핵함수에 의해 추정된 값은 일반적으로 매개변수적 방법들에 의한 선택된 분포형에 따른 최대값과 최소값 사이에 나타났다.

관측자료의 복합화된 상황에서 매개변수적 방법으로 선택된 분포형이 꼬리부분이 아닌 다른 부분에서는 적당하게 도시되고 작은 분산을 가진다 할지라도 해당 분포형의 꼬리에는 상당히 부적합할 수 있다. 자료 전체에 대한 적당한 밀도함수를 추정하는 것과는 확률밀도함수의 꼬리거동의 추정은 근본적으로 다른 문제라는 것을 인식한다면 이러한 이론은 꼬리거동의 외삽에 대한 최상의 모형을 찾는 것으로 귀결될 수 있다.

5. 참고문헌

Adamowski, K. 1981. Plotting formula for flood frequency. Water Resources Bulletin 17(2), 197-202.

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

Gasser, T. and Muller, H. G. 1984. Estimating regression functions and their derivatives by the kernel method. Scandinavian Journal of Statistics 11, 171-185.

Gasser, T., Kneip, A., and Kohler, W. 1991. A flexible and fast method for automatic smoothing. Journal of the American Statistical Association 86(514), 643-652.

Guo, S. L. 1996. Nonparametric kernel estimation of low flow quantiles. J. of Hydrology, Vol. 185, pp. 335-348.

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

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

Moon, Young-Il, and U. Lall. 1994. Kernel Quantile Function Estimator for Flood Frequency Analysis. Water Resources Research, Vol. 30, No. 11, pp. 3095-3103.

Moon, Young-Il. 1996. Nonparametric flood frequency analysis. Journal of the Institute of Metropolitan Studies, Vol. 22, No. 1, pp. 231-248.

Moon, Young-Il. 2000. The Study of Parametric and Nonparametric Mixture Density Estimator for Flood Frequency Analysis. Water Engineering Research Vol.1, No.1, pp.61-73.

Webb, R. H.and Betancourt, J. L. 1992. Climatic variability and flood frequency of the Santa Cruz River, Pima county, Arizona, USGS Water-Supply Paper 2379, 1-40.