서울시립대학교 도시과학연구원
도시과학논총 제26권 2000년 12월
The Journal of Urban Sciences
The University of Seoul
Vol. 26, No. 1, December 2000

경계핵함수를 이용한 서울의 확률강수량 산정

문영일·차영일
(서울시립대학교 토목공학과 조교수·서울시립대학교 토목공학과 박사과정)

Probability Precipitation Estimate of Seoul by using Boundary Kernel Function Method

Moon, Young-Il·Cha, Young-Il

ABSTRACT

A frequency analysis based on hydrological data such as gaged flood and precipitation amount is one of the most important element for designing and planning an economical hydrosystem.

Nonparametric boundary density estimation methods have been quite recently introduced as frequency analysis. Probability precipitation of Seoul site was estimated and compared by using parametric and boundary kernel frequency analysis methods.

For parametric frequency analyses, probability precipitation was estimated by using the method of moment (MOM), maximum likelihood (ML), probability weighted moments (PWM), and L-moments. Distributions used were Gamma, generalized extrem value, Gumbel, log-Gumbel, lognormal, log-Pearson type Ⅲ, Weibull, exponential, generalized logistic, generalized Pareto, normal, Pearson type III, and Wakeby distributions.

For boundary kernel analyses, a variable kernel density function estimation was used. Interior kernel density function considered was Epanechnikov function. Methods of optimal bandwidth h selection used was Muller's Mean Squared Error method.

Key Word : Boundary Kernel Function, Interior Kernel, Probability Precipitation, Epanechnikov Function, Muller's Mean Squared Error

I. 서 론

임의의 관측지점에서 홍수량 및 강수량 빈도해석 방법에 대한 우리나라 대부분의 연구는 관측자료를 특정 확률밀도함수를 가진 모집단으로부터 나왔다는 가정에 근거한 통계학적인 관점에 집중해왔다. 지금까지 많은 홍수량 및 강수량 빈도해석에 대한 방법들이 발표되었고, 많은 수문학자들이 이러한 분포형에 대한 최적의 매개변수 추정방법 (모멘트법, 최우도법, 확률가중모멘트법, L-모멘트법 등)과 추정된 분포형에 대한 최선의 분포형 선정 방법(χ^2 검정, Kolmogrov-Smirnov 검정, Cramer Von Mises 검정, Probability Plot Correlation Coefficient 검정 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 부합된다 하더라도 빈도해석에서의 홍수량의 크기에는 상당량의 불확실성이 존재한다. 관측 지점에서 연 최대홍수량은 여러 가지 (융설, 강우유출, 계절성 폭우 등)원인을 나타내는데, 이는 통계학적으로 이질적인 모집단 또는 복합 분포형 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 추정값을 조화시키려는 시도를 하였다. 이는 홍수가 복합 과정들로부터 발생된다라는 것에 좋은 증거가 된다할지라도 이러한 과정을 실무에서 쉽게 실행되기는 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,…, n)와 임의의 도시위치공식(plotting position formula)(p_i, i=1, 2,…, n)을 사용하여 단일 자료분포형뿐만 아니라 혼합된 분포형에서도 일관성있게 적용할 수 있는 경계 핵 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 ($0.9 < p < 1$)에 대한 외삽은 경계 핵 함수를 사용하여 빈도분석을 수행하였다.

II. 경계 핵 함수에 의한 추정식

2.1. 내부 핵(Interior Kernel)함수에 의한 빈도해석

경계 핵 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,…, n)를 내림차순으로 배열하면 순차적 연 최대계열치가 구성되고 p_i(i=1, 2,…, n)는 표준도시공식을 이용하여 경험상으로 추정되는 확률값으로 간주하여 그래프상에 대응하는 도시위치를 갖게 할 수 있다. 여기서 표준도시공식으로 Adamowski (1981) 공식을 사용하면 다음 식(1)과 같다.

$$p_i = \frac{i-0.25}{n+0.5} \quad (1)$$

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

따라서 추정되는 빈도해석 함수의 분위값(quantile) $\hat{x}(p)$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.

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

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

2.2. 핵함수(Kernel Function)

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 유한 분산값($\int t^2 K(t)dt = constant$)을 가진 대칭적 확률밀도이다. 위에서 사용한 그림1의 Epanechnikov 핵함수는 다음 식 (3)과 같다.

$$K(t)=0.75(1-t^2) \quad (3)$$

여기서 $|t| < 1$.

2.3. 대역폭(Bandwidth)

대역폭 h는 경계 핵함수법에 의한 빈도해석시 중요한 인자로서 추정되는 회귀식함수의 완만함과 거칠기를 결정한다. 작은 대역폭은 임의의 점에서 $\hat{x}(p)$ 산정에 적은 수의 관측자료가 고려되고, 보다 큰 대역폭에서는 상대적으로 많은 수의 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 다음 식 (4)와 같이 제시하였다.

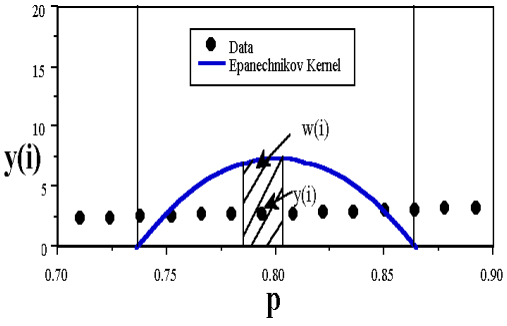

그림 1. x(0.8)에서의 핵함수 회귀추정 (h=0.064)

$$MSE(\hat{x}(p))=E[\hat{x}(p)-x(p)]^2 \sim \frac{\sigma^2}{nh}\int_{-1}^{q}(K_x(q,t))^2dt + \frac{1}{4}h^4\{x''(p)\}^2\left\{\int_{-1}^{q}K_x(q,t)t^2dt\right\}^2 \quad (4)$$

식 (4)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(4)에서 평균제곱오차(MSE)를 최소화시키는 최적의 대역폭은 다음 식 (5)와 같이 제시하였다.

$$h=\left\{\frac{1.5}{n}\frac{c_1}{c_2}\frac{\sigma^2}{\int_0^1\{x''(p)\}dt}\right\}^{0.2} \quad (5)$$

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

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

2.4. 경계 핵(Boundary Kernel) 함수에 의한 빈도해석

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 Epanechnikov 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.

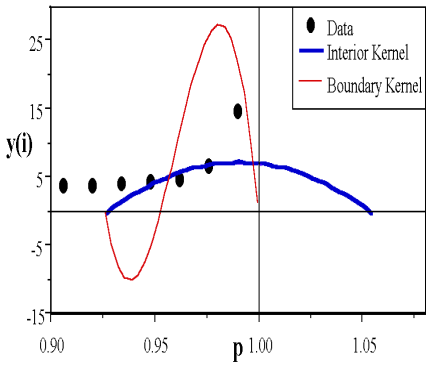

그림 2. x(0.99)에서의 핵함수 회귀추정(h=0.064)

Muller(1991)는 내부 핵함수에 대응하는 경계 핵함수를 식(4)의 평균제곱오차(MSE) 의 최적화에 따라 제시하였다. 여기에서 Epanechnikov 핵함수를 사용하여 내부 핵함수와 경계 핵함수를 살펴보면 각각 다음 식(6), 식(7)과 같다.

내부 핵함수 :

$$K(t)=0.75(1-t^2),\ \text{여기서}\ |t|<1 \quad (6)$$

경계 핵함수:

$$K(q,t)=6(1+t)(q-t)\frac{1}{(1+q)^3}\times\left\{1+5\left(\frac{1-q}{1+q}\right)^2+10\frac{1-q}{(1+q)^2}t\right\} \quad (7)$$

여기서 $t=(p-p_i)/h$ 이며, $K(q,t)$는 오른쪽 경계면 구간 [1-h, 1]에서 사용되는 경계 핵함수이며 $q=(1-p)/h$ 이다. 반면에 왼쪽 경계면 구간 [0, h]에서는 $q=p/h$ 이다. 이 때 q=1이면 경계 핵함수 식(7)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.

Ⅲ. 적용 및 분석결과

기상청 서울관측소의 연 총 강수량인 연 강수량 자료(1908년~1998년)를 사용하여 재현기간별로 확률강수량을 경계 핵밀도함수에 의한 회귀식 추정과 함께 기존의 매개변수적 방법에 의한 빈도해석을 수행하여 추정하였다. 여러 매개변수적 방법과 경계 핵함수 방법에 의한 추정값과 검정결과가 표1~표6에 정리되어 있다. 또한 서울지점의 연 강수량, 연 중 월 강수량이 최대인 월 최대치 계열 및 연 중 일 강수량이 최대인 일 최대치 계열의 확률밀도 함수와 누가분포함수의 결과는 그림3~그림8에 정리되어 있다.

3.1. 확률강수량 산정

일반적으로 매개변수에 의한 확률강수량의 추정은 Chi-square 검정 또는 Kolmogrov-Smirnov 검정을 통과한 분포형이더라도 연 강수량에서 작게는 모멘트법의 경우 4%, 크게는 확률가중모멘트법에서 20% 이상의 차이를 보이고 있으며, 같은 분포형일지라도 매개변수 추정방법에 따라 10%정도의 차이를 보이고 있다. 경계핵밀도함수에 의한 재현기간 100년 연 강수량은 2,368mm이나, 반면에 매개변수적인 방법에 의한 서울의 100년 빈도 연 강수량은 선택된 분포형과 매개변수 산정방법에 따라 2,193~2,756mm사이에서 값이 다르게 나타났다. 즉 모멘트법(MOM)을 사용한 매개변수 산정시 100년 빈도 년 확률강수량은 2,293~2,386mm, 최우도법(ML)은 2,193~2,504mm, 확률가중모멘트법(PWM)은 2,284~2,756mm, L-모멘트법은 2,215~2,442mm이다.

또한 월 최대강수량의 경우도 연 강수량과 마찬가지로 선택된 분포형과 매개변수 산정방법에 따라 1,012~1,435mm의 차이를 보이고 있으나 경계핵 밀도함수에 의한 재현기간 100년의 월최대 강수량은 1,349mm로 매개변수적 방법의 범위 안에서 적당하고 일관된 값을 제시해주고 있다.

일최대 강수량의 매개변수적 빈도해석의 경우 단 한 개의 분포형만이 채택되었고, 그 추정값 또한 경험적 도시공식에 의한 값보다 너무 작은 값을 추정했기 때문에 선택하는데 있어 많은 어려움을 보이고 있으나, 경계핵 밀도 함수에 의한 재현기간 100년의 일 최대 강수량은 350mm로 경험적 도시공식에 의한 값에 비교적 잘 적합한 것으로 보인다.

표4~표6의 적합도 검정결과에서는 Chi-square 검정이 Kolmogrov-Smirnov 검정보다 기각력이 좋은 것으로 평가된다.

표 1. 100년 빈도 연 강수량(mm)

분포형	L-M	PWM	ML	MOM	BK
GAM2	2215				
GAM3		2309		2304	
GEV	2330	2351	2257	2296	
GUM	2401	2401	2504	2385	
LGU2		2755			
LN2		2284	2318	2306	2368
LN3	2230	2330	2252	2292	
WBU3			2193		
WKB5	2330				
PE3	2309				
GLOG	2442				
최대	2442	2755	2504	2385	2368
최소	2215	2284	2193	2292	

표 2. 100년 빈도 월 최대강수량(mm)

분포형	L-M	PWM	ML	MOM	BK
GAM2	1025	1025	1012	1085	
GAM3			1048		
GEV	1224	1226	1144	1175	
GUM	1072	1072	1031	1114	
LGU2		1435			
LGU3		1224	1144	1176	
LN2		1072	1123	1159	1349
LN3	1197	1197	1096		
WKB5	1264				
G PAR	1098				
PE3	1150				
EXP	1225				
G LOG	1267				
최대	1267	1435	1144	1176	1349
최소	1025	1025	1012	1085	

표 3. 100년 빈도 일 최대강수량(mm)

분포형	L-M	PWM	ML	MOM	BK
GAM2		309	308	319	350
최대		309	308	319	350
최소		309	308	319	

표 4. 연 강수량의 적합도 검정(MOM)

분포형	CHI-SQUARE TEST		K-S TEST	
	X^2	TABLE	K-S	TABLE
Gamma2	12.52	9.49	0.10	0.144
Gamma3	5.62	7.81	0.08	0.144
GEV	5.13	7.81	0.09	0.144
Gumbel	4.64	9.49	0.07	0.144
Log-Gumbel2				
Log-Gumbel3				
Log-Normal2	5.13	9.49	0.09	0.144
Log-Normal3	5.13	7.81	0.09	0.144
Log-PearsonIII				
Weibull2	18.32	9.49	0.15	0.144
Weibull3				

표 5. 월 최대강수량의 적합도 검정(MOM)

분포형	CHI-SQUARE TEST		K-S TEST	
	X^2	TABLE	K-S	TABLE
Gamma2	9.34	9.49	0.10	0.144
Gamma3				
GEV	4.00	7.81	0.06	0.144
Gumbel	5.29	9.49	0.08	0.144
Log-Gumbel2				
Log-Gumbel3	4.00	7.81	0.06	0.144
Log-Normal2	4.32	9.49	0.07	0.144
Log-Normal3				
Log-PearsonIII	1.62	7.81	1.00	0.144
Weibull2	12.69	9.49	0.12	0.144
Weibull3				

표 6. 일 최대강수량의 적합도 검정(MOM)

분포형	CHI-SQUARE TEST		K-S TEST	
	X^2	TABLE	K-S	TABLE
Gamma2	9.21	9.49	0.11	0.144
Gamma3				
GEV	13.66	7.81	0.09	0.144
Gumbel	13.98	9.49	0.10	0.144
Log-Gumbel2				0.144
Log-Gumbel3	13.66	7.81	0.09	0.144
Log-Normal2	11.40	9.49	0.09	0.144
Log-Normal3	11.40	7.81	0.09	0.144
Log-PearsonIII				
Weibull2	10.60	9.49	0.15	0.144
Weibull3				0.144

3.2. 강수량자료계열의 확률밀도함수 (PDF)

그림3~그림5는 서울지점의 강수량자료에 대하여 연 계열, 월 최대치 계열 및 일 최대치 계열에 대한 확률밀도함수(PDF)의 결과를 보여주고 있다. 그래프에서 x축의 단위는 mm인 강수량이고, y축은 확률밀도함수 값 f(x)를 나타낸다. 그림에서 일 최대 강수량의 확률밀도함수그래프가 multimodal의 경향을 가장 뚜렷하게 보여주고 있고, 적합도검정 결과에서도 높은 기각률을 보이고 있다.

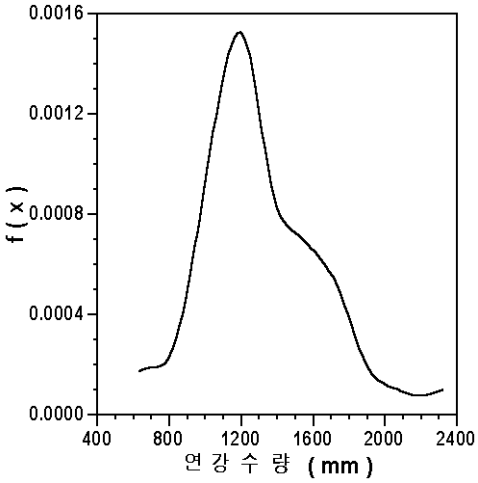

그림 3. 연 강수량의 확률밀도함수

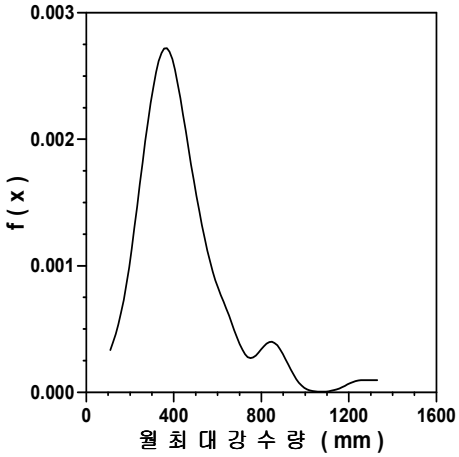

그림 4. 월 최대강수량의 확률밀도함수

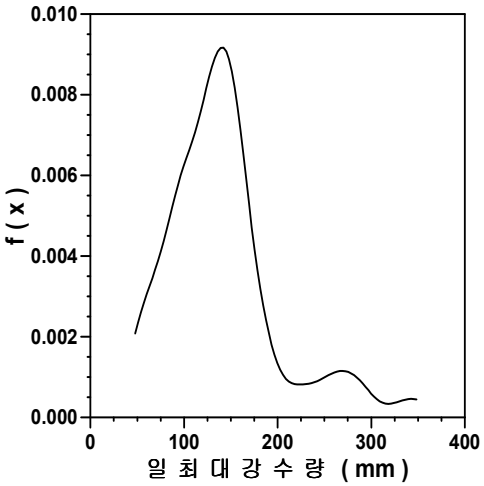

그림 5. 일 최대강수량의 확률밀도함수

3.3. 강수량자료계열의 누가분포함수 (CDF)

그림6~그림8은 서울지점의 강수량자료에 대하여 연 계열, 월 최대치 계열 및 일 최대치 계열에 대한 누가분포함수의 그래프들이다. 관측치는 Adamowski 도시공식에 의해 도시되었고, 경계핵 함수(Boundary Kernel Function)방법과 검정을 통과한 다수의 매개변수적 방법인 모멘트법에 의해 추정된 누가분포함수(CDF)를 보여주고 있다. 그래프에서 x축은 대수(log scale)를 취한 재현기간(년)이고, y축은 강수량(mm)이다.

그림5~그림8의 그래프에서 경계핵 밀도함수법의 추정치는 관측치를 비교적 잘 표현해주고 있다. 그림8의 일 최대강수량의 누가분포함수에는 Gamma2 분포형만이 적합도 검정을 통과하여 선택하는데 도움이 될 수 있겠지만, 그 추정치가 관측치보다 너무 작아 합리적인 선택이라고 하기엔 어려움이 있다. 그러나 경계 핵 함수에 의한 추정치는 관측치를 잘 표현해 주고 있다.

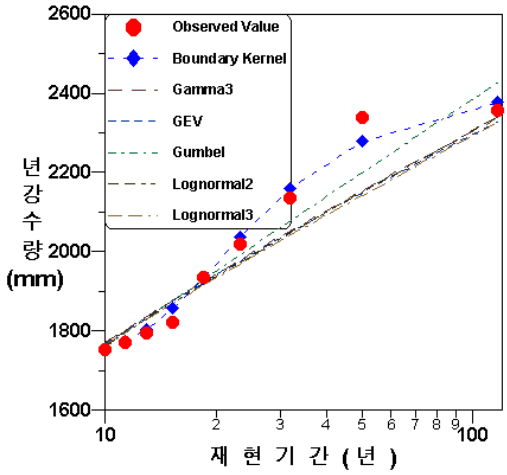

그림 6. 연 강수량의 누가분포함수(MOM)

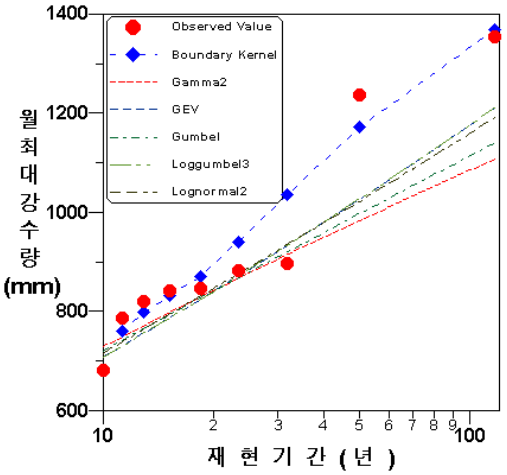

그림 7. 월 최대강수량의 누가분포함수(MOM)

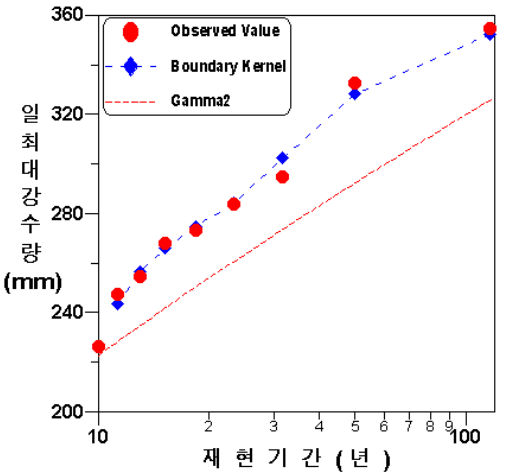

그림 8. 일 최대강수량의 누가분포함수(MOM)

Ⅳ. 결 론

홍수량 또는 강수량 자료 등의 최대치 계열에 대한 빈도해석 시 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 산정방법에 따라 추정값의 차이가 있고 분포형에 따라 100년 빈도 년 강수량의 경우 500mm이상 차이가 있어 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 추정이 근본적으로 다른 문제라는 것을 인식한다면 이러한 이론은 꼬리거동의 외삽에 대한 최상의 모형을 찾는 것으로 귀결될 수 있다.

<참고문헌>

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

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

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

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

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

6. 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.

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

8. 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.

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

10. 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.
11. 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, pp. 1-40.