Kernel Density Function을 이용한 비매개함수적 홍수빈도 해석방법

문 영 일[1)]

1. 서론

강우량 및 홍수량 같은 자료의 빈도해석에는 크게 매개변수적 빈도해석방법(parametric frequency analysis)과 비매개변수적 빈도해석방법(non-parametric frequency analysis)이 있다. 우리 나라에서는 지금까지 빈도해석에 매개변수적 빈도해석방법을 사용하고 있다. 주로 사용되는 분포들은 Lognormal, Gamma, GVE, Wakely, Gumbel, Log-pearson Type III 등이다. 또한, 많은 종류의 매개변수 추정방법등이 소개 되었다. 즉, 모멘트법, 최우도법, 최소자승법, 도시법, 확률가중 모멘트법, L-모멘트법 등이다. 그러나, 이러한 매개변수적 해석방법의 어려운 점은 대표적으로 다음과 같다. (1)분포함수의 객관적 선택, (2)매개변수의 신뢰도 (특히, 짧은 기록의 자료와 왜곡된 자료), (3)여러가지 원인으로 인한 복합분포(mixed distribution)의 해석의 어려움이다.

기존의 매개변수적 빈도해석의 가장 어려움 중에 하나가 특정 확률분포형을 가정하는데 있다고 볼 수 있다. 한 지역에서 여러개의 확률분포형이 적합도 판정(χ^2, Kolmogorov-Smirnov, 또는 Cramer von Mises 검정 등)을 받고, 채택지 점수에 의한 한 지역의 최적 분포형을 선택했다 하더라도 몇년후에도 같은 분포형이 선택된다는 보장이 없다. 특히, 자료 중에 "badly behaved" data 가 있는 경우 적정한 분포형을 선정하는데 많은 어려움이 따른다. 이런 경우 비매개변수적 빈도해석 방법중의 하나인 핵밀도함수(Kernel Density Function)방법을 사용하면 상당히 좋은 결과를 얻을 수 있다는 연구논문이 10여년 전부터 많이 발표(Lall 등, 1993; Moon 등, 1993; Moon과 Lall, 1994; Adamowski, 1996) 되었다. 핵밀도함수 해석방법은 어떤 분포형의 가정이 필요 없이 데이터 자체에서 분포형을 유도할 수 있다. 특히 우리 나라와 같이 관측자료 기간이 짧은 경우에는 더욱 핵밀도함수 해석방법이 좋은 결과를 줄 수 있을 것이다. 또한 우리가 어떤 지역에서 강우자료에 대한 최적 분포형을 선정했다 하더라도 몇년후 또는 몇십년후 보완된 자료를 이용하게 되면 지금 우리가 선정한 분포형과는 다른 확률분포형이 적합하다고 판단될 수도 있는 것이다. 이와 같은 분포형 선정의 어려움을 해소하기 위한 하나의 방법으로, 또는 보다 적합한 분포형을 선정하기 위한 수단으로 핵밀도함수(Kernel Density Function)에 의한 해석이 도움을 줄 것이다.

2. 비매개변수적 밀도함수 추정

비매개변수적으로 자료의 확률밀도함수 f(x)를 판단하기 위해 사용될 수 있는 방법에는 여러 가지가 있다. 여기에는 핵밀도함수(kernel density function)방법, orthogonal series, k^{th} nearest neighbor, maximum likelihood 이나 막대그래프 등을 포함한다. 여기에서, 이론적으로 가장 잘 알려 졌고 가장 발달된 것은 핵밀도함수 추정법이다. 먼저 핵밀도함수 추정법의 기본 개념인 막대그래프에 대해 살펴보겠다.

(1)막대그래프(Histogram)

가장 오래되고 가장 널리 사용되는 확률밀도함수 추정법은 막대그래프이다. 막대그래프는 시작점(x_0)이 주어진 상태에서, 막대그래프의 구간간격은 양과 음의 정수 m과 양의정수 h를 사용하여 구간 $[x_0+mh, x_0+(m+1)h]$로 정의된다. 이 때 막대그래프(Histogram)의 밀도함수는 다음과 같이 정의된다.

---

1) 서울시립대학교 토목공학과 조교수

$$f(x) = \frac{1}{nh}(x와\ 같은\ 구간에\ 있는\ 자료의\ 수) \quad (1)$$

막대그래프는 가장 오래된 확률밀도함수 추정법이고, 자료의 시각적인 그래픽 표시를 위한 고전적인 비매개변수적 확률밀도함수 추정법이다. 막대그래프는 이해하기가 쉽고 수작업에 의해서도 쉽게 계산될 수 있는 반면에, 계급구간이 변화는 점에서 불연속적이고 구간간격과 그래프의 시작점의 선택에 따라 확률밀도함수의 모양이 달라진다. 그러므로, 막대그래프는 IMSE(Integrated Mean Squared Error)의 관점에서는 상당히 비능률적이다. 확률변수가 주어 졌을 때, IMSE는 막대그래프의 구간간격이 최적으로 선택되어 졌을 때 $n^{2/3}$에 비례하게 점근선적으로 적어지게 된다고 알려지고 있다(Devroye와 Gyorfi, 1985). 그러나, 막대그래프를 만들 때의 발생되는 문제는 아무리 이상적인 동일한 구간간격을 사용할지라도, 다른 시작점의 위치 선택에 따라 같은 자료에서도 매우 다른 확률밀도함수의 결과를 초래한다는 것이다. 이런 고전적인 막대그래프의 저지르기 쉬운 잘못은 많은 연구자들에게 연속적이며 변화하는 히스토그램을 연구하는 동기를 주었다.

Rosenblatt(1956)는 모든 자료가 발생되어진 각각의 위치에 막대그래프의 박스중앙을 위치하도록 하여 구간을 이동시킬 수 있는 이동 히스토그램을 발달시켰다. 만약 자료가 에너지원에 비유된다면, 자료는 발생 위치에서 최고 값을 갖고 발생위치로부터 멀어질수록 급속히 감소한다. 이 이동할 수 있는 히스토그램은 핵밀도함수 추정법으로서 알려져 왔다.

(2)핵밀도함수 추정법(Kernel Density Function Estimator)

Rosenblatt(1956)은 핵밀도함수 추정법을 발표했는데, 모든 실수 x에 대하여 다음과 같이 정의하였다.

$$f(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K(\frac{x-X_i}{h}) \quad (2)$$

여기서 x는 임의의 추정점이며 $X_1, X_2, \cdots, X_n$ 은 독립적으로 동일하게 분포된 실관측치이다. K(.)는 핵함수이고 h는 n이 무한대로 갈 때 0으로 접근하는 값을 갖는 양의 bandwidth이다. Parzen(1962)은 Rosenblatt의 핵밀도함수 추정법의 특성을 연구하여 일반화 시켰다. 그는 이 추정법이 Rosenblatt-Parzen 추정법으로 불려질 정도로 보편화 시켰다. 또한 Silverman(1986)은 핵밀도함수 추정법의 기본 개념을 잘 보여 주었다. 확률밀도 함수의 정의로부터 임의의 변수 x는 다음과 같은 밀도함수 f(x)를 갖는다.

$$f(x) = \lim_{h\to 0}\frac{1}{2h}P(x-h<X<x+h) \quad (3)$$

주어진 h에 대하여 P(x-h<x<x+h)의 값은 구간[x-h, x+h] 사이의 있는 자료의 비율에 의해 계산되어질 수 있다. 그래서 본래 추정식은 작은 숫자 h의 선택으로 다음과 같이 얻을 수 있다.

$$\hat{f}(x) = \frac{1}{2hn}(구간[x-h,\ x+h]\ 사이에\ 있는\ 자료의\ 수) \quad (4)$$

이 추정식을 보다 명료하게 표현하기 위해서 가중치 함수 w(x)를 다음과 같이 정의하자.

$$w(x) = \frac{1}{2} \text{ if } |x| < 1 \quad (5)$$
$$0 \text{ if otherwise}$$

이 때 추정식은 이해가 쉽게 다음과 같이 표현할 수 있다.

$$f(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}W(\frac{x-X_i}{h}) \quad (6)$$

이것은 식 (2)와 같은 표현으로, 각 관측치에 폭 2h, 높이 $(2nh)^{-1}$의 막대를 각 관측치에 놓고, 그 들의 합계에 의해 추정 식이 만들어지는 것을 보여준다. 여기서, 가중함수 w(x) 대신 연속함수 형태인 핵함수(kernel function)를 사용하면 핵밀도함수의 모양은 그림 (1)과 같다. 이 핵함수(kernel function)는 다음조건을 만족한다.

$$\int K(t)dt = 1 \tag{7}$$

여기서 $t = \frac{x - X_i}{h}$

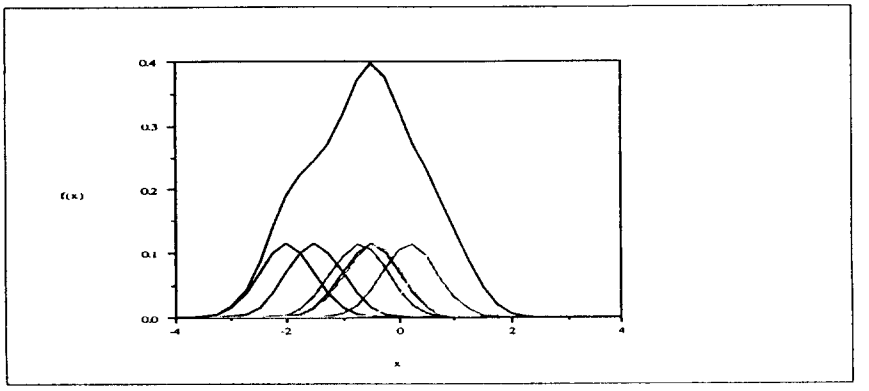

그림 1. 각각의 핵함수에 의한 핵밀도함수 추정(bandwidth 0.5 사용).

이 때 x는 임의의 점이고, X_i 는 실 관측된 자료이다. 핵함수는 대개 x=0에서 최대치를 가지며, 연속적이며, 대칭인 방정식의 형태를 가진다. 즉, 핵함수의 면적은 1이고($\int K(t)dt = 1$) 기대값 0 ($\int tK(t) = 1$)과 유한한 분산값($\int t^2 K(t)dt = constant$)을 갖는다. 그러나, 때로는 이 특성들을 만족하지 않는 핵함수가 사용되어질 수도 있다. 또한, 수렴 추정치에 있어 n이 무한대에 접근함에 따라 nh는 0에 가까워지는 경향이 있다.

핵밀도함수 추정 방법을 실제로 적용할 때 핵함수 K(t)와 bandwidth h를 선택하는 것이 필요하다. 핵함수의 선택은 Epanchnikov(1969)에 의해 고려되었는데 그는 평균자승오차(mean square error, MSE)에 의하면 포물선 형태의 핵함수가 거의 최적의 결과를 나타낸다는 것을 보여 주었다. 그러나, 많은 연구에서 핵함수의 선택은 실제 생각되어진 만큼 중요한 문제가 아니라고 주장되어 졌다. 선택된 함수의 효율성(선택된 핵함수의 MSE/최적의 핵함수의 MSE)은 주어진 핵함수의 특성을 만족하면 거의 1에 가깝다. 몇 가지 유용한 핵함수들이 표 1에 주어졌다. 그러나, 각각 다른 핵함수들은 사용목적에 따라 점검되어야 한다. 예를 들면 밀도 함수의 연속성과 미분가능이 필요하다면 제한된 구간을 갖는 핵함수보다는 무한한 구간을 갖는 핵함수를 선택하는 것이 좋을 것이다.

표 1. 여러 핵함수의 종류와 식

Kernel	K(t)
Rectangular	1/2 , 여기서 $\|t\| < 1$
Gaussian	$\frac{1}{\sqrt{2\pi}} \exp(-\frac{t^2}{2})$
Epanechnikov	$\frac{3}{4}(1 - \frac{1}{5}t^2)/\sqrt{5}$, 여기서 $\|t\| < \sqrt{5}$
Cauchy	$\frac{1}{\pi(1+t^2)}$

실제로 적절한 핵함수의 선택의 문제는 Cline(1989)에 의해서 연구되었다. 그는 자료 갯수의

함수로 최적의 핵함수를 유도했다. 언급한 것과 같이 핵함수의 선택은 중요하지 않지만, bandwidth h의 선택은 매우 다른 문제이다. h의 값은 핵함수 추정법에 있어서 매우 중요하지만 실제로는 정확하게 구하기는 쉽지 않다. 너무 큰 h는 큰 편차(bias)와 너무 완만(oversmooth)한 밀도함수의 추정과 정보의 손실을 가져온다. 반면에, 너무 작은 h는 큰 분산(variance)과 거친(rough) 추정치를 나타낸다. 그러나, 주어진 자료에서 최적의 h를 구하는 방법은 있다. 그 중에서 대표적인 것이 IMSE, 최우도법, 최소자승법, 또는 Adamowski Criterion 등으로 이들 값이 최소가 되는 점에서 최적의 h를 구할 수 있다.

Dodge (1986)가 밝혔듯이 아래에 같이 최적 h에(IMSE에 기본을 둔) 대한 표현은 명확하다.

$$h = C\ (f(x)/f''(x))^{\frac{1}{5}}\ n^{\frac{-1}{5}} \tag{8}$$

이 최적 h는 표본의 크기에 의존되고 f''(x)에 반비례한다. f''(x)는 x에 따라 변하기 때문에 h(x)는 오히려 포괄적인 h가 사용되어야 한다. 이 때 h가 변화 없이 일정한 값을 가지면 이를 고정 핵밀도 추정법 (fixed kernel density estimator)이라 한다. 그러나, 고정 핵밀도추정법은 군데군데 길게 늘어진 자료분포에 적용될 경우 결점을 갖는다. 왜냐하면, 일정한 bandwidth, h를 전체자료에 적용될 경우 추정치의 꼬리(tail) 부분에서 자료가 드물게 존재할 경우에 추정치의 값이 불연속 또는 거친 밀도함수의 모양을 갖기 때문이다. 이런 고정 핵밀도추정법의 단점을 보안한 것이 변동 핵밀도함수 추정법이다.

(3)변동핵 밀도함수 추정법(Variable Kernel Density Estimator)

Breiman 등(1977)은 고정 핵밀도함수 추정법의 특성에다 자료의 지역적인 밀도를 고려하는 k^{th} nearest neighbor 방법을 결합한 변동 핵밀도함수 추정법을 제안했다. 변동 핵밀도함수 추정법은 고정 핵밀도함수 추정법과 유사한 방식으로 만들어졌다. 하지만 자료가 발생된 위치에 놓여지는 핵함수의 폭이 자료의 밀도에 따라 변한다.

K(X)를 핵함수라 하고 k를 양의 정수로 놓고, $d_{j,k}$를 한 개의 자료 x_j에서 그 나머지 자료 (n-1)개 중에서 k번째로 가까운 지점에 있는 자료까지의 거리라 하자. 그러면, 변동 핵밀도함수 추정법은 다음과 같이 정의된다.

$$f(x) = \frac{1}{nh} \int \frac{1}{d_{j,k}} K\left(\frac{x - X}{hd_{j,k}}\right) dx \tag{9}$$

여기서 h는 양의 bandwidth이다. 자료의 분포가 적은 낮은 밀도지역에서 $d_{j,k}$ 는 커지고 변동 핵함수의 모양은 넓게 퍼지게 된다. 자료의 분포가 많은 밀도 높은 지역에서는 그 반대 현상이 일어나, 변동핵함수는 좁게 밀착된 형태를 가질 것이다. 그러나 고정 핵밀도함수 추정법은 f(x)의 크기의 변동에 적절히 대응할 수가 없다. 예를 들어, 꼬리 부분에 만일 하나의 자료 x_k 만을 포함하는 낮은 f(x)의 지역이 있다면, 꼬리 부분에서의 핵밀도함수의 추정값은 $x = x_k$ 에서 지역적인 최고치를 갖게될 것이고 나머지 지역에서는 아주 낮을 것이다. 따라서 고정 핵밀도함수 추정법의 문제점은 자료의 분포 형태에 대응하지 못한다는 것이다. Moon과 Lall(1994)은 변동 핵밀도함수 추정법이 밀도함수의 꼬리 부분인 최빈값 추정이나 또는 자료가 비대칭 분포일 때 장점이 많다는 것을 보여 주었다. 변동 핵밀도함수 추정법의 일관성과 수렴성은 Devroye과 Gyorfi(1985)에 의해 설명되어졌다.

3. 적용과 결론

변동 핵밀도함수 추정(VK-C-AC)과 여러 매개밀도함수 추정에 의한 결과를 비교하였다. 매개밀도함수로는 2 Lognormal, 3 Lognormal, Type I Extremal, Gamma 와 Log Pearson Type III를 사용하였다. 변동 핵밀도함수 추정에는 Cauchy(C) 핵함수와 Adamowski Criterion(AC)를 사용하여 누가분포함수(CDF)에 대하여 비교하였다. 특히 자료가 bimodal의 모습을 가질 때, 비매개변수적 변동 핵밀도함수추정의 CDF와 실측자료의 empirical CDF는 상당히 근접한 모습을 보여 주고 있으나, 대부분 매개변수적 밀도함수의 CDF는 다른 거동의 모습을 보여 준다. 이 경우와 같이 bimodal의 경우에는 일반적인 테

스트 방법으론 밀도함수를 각각의 밀도함수로 분리 하기는 어려움이 있다. 그러나, 비매개함수적 밀도함수 추정법인 변동 핵밀도함수 추정의 결과는 자료 자체의 모습을 표현하여 주므로, 자료의 밀도함수의 모양과 상관없이 일관성 있는 결과를 주는 장점이 있다. 그림 2에서는 미국 애리조나주의 투산에 있는 Santa Cruz 강의 연 최대홍수량에 대한 빈도해석의 결과를 보여준다. 여러가지 매개함수적 빈도해석에 의하면 재현기간 100년에 대한 홍수량은 구간 572-2,780 m^3/sec의 분포를 보이며, Webb and Betancourt(1992)이 여러가지 테스트에 의하여 밀도함수를 원인에 따라 3가지로 나누어 해석한 결과 재현기간 100년의 크기로 1,050 m^3/sec를 얻었다. 그러나, 변동핵에 의한 추정법은 일관된 방법에 의해 1,094 m^3/sec를 얻어 Webb and Betancourt(1992)에 결과와 비슷한 값을 보여준다. 그림 3의 경우는 실측 자료의 밀도함수가 기존의 매개변수적 밀도함수의 모양과 다른 경우로, 미국 유타 주의 Beaver 강의 연 최대홍수자료이다. 이 경우도 매개함수적 빈도해석의 결과는 실측자료와 상당히 다른 거동의 모습을 보여주는 반면, 변동 핵밀도함수 추정의 결과 실측자료의 거동과 비슷하다.

기존의 매개변수적 빈도해석은 자료의 형태에 따라 특정 확률분표형 선정에 어려움이 있다. 이런 어려움을 해소하기 위한 하나의 방법으로, 또는 보다 적합한 분포형을 선정하기 위한 수단으로 핵밀도함수(Kernel Density Function)에 의한 빈도해석이 수공구조물의 적정한 크기를 정하는데 도움을 줄 것이다. 더욱이, 우리나라와 같이 관측자료 기간이 짧은 지역의 빈도해석은 매개변수에 의한 결과에 대하여 불확실 할 수 있으며 또한 현장의 실무자가 적정한 분포형을 구분하여 사용하기엔 어려움이 따를 수 있는 매개변수적 방법보다는 항상 일관성을 가지고 사용 가능한 비배개변수적 해석방법이 더욱 유용하게 현장 실무자에게 사용되어질 수 있을 것이다.

4. 참고문헌

Adamowski, K. 1996. Nonparametric Estimation of Low-Flow Frequencies. Journal of Hydraulic Engineering 122(1); 46-49.

Cline. 1989. Optimal kernel estimation of densities.

Devroye and Gyorfi. 1985. Nonparametric density estimation:The L1 view. John Wiley, New York.

Dodge, Y. 1986. Some difficulties involving nonparametric estimation of a density function. Journal of Official Statistics 2(2):193-202.

Epanechnikov, V.A. 1969. Nonparametric estimation of a multidimensional probability density. Theory Probability and Applications 14:153-158.

Lall, U., Young-Il Moon, and K. Bosworth. 1993. Kernel flood frequency estimators :bandwidth selection and kernel choice. Water Resources Research 29(4):1003-1015.

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

Moon, Young-Il and Lall, U. 1994. Kernel Quantile Function Estimator for Flood Frequency Analysis. Water Resources Research.

Parzen, E. 1962. On estimation of a probability density function and mode. Ann. Math. Statist. 33:1065-1076.

Rosenblatt, M. 1956. Remarks on some nonparametric estimates of a density function. Ann. Math. Statist. 27:832-837.

Silverman, B.W. 1986. Density estimation for statistics and data analysis. Chapman and Hall, New York.

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

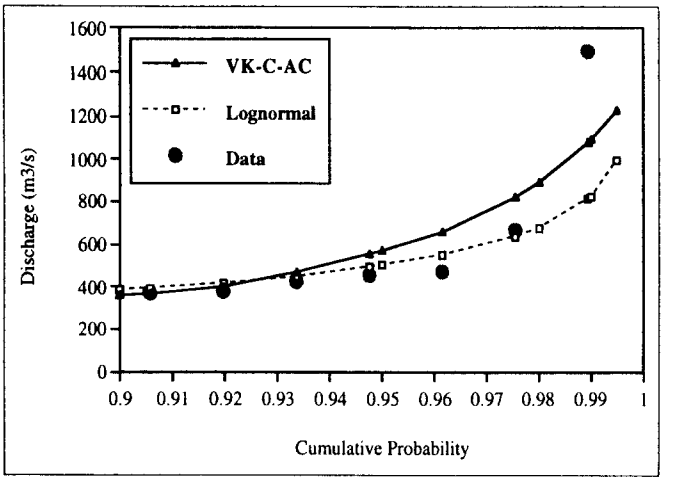

그림 2. Arizona, Santa Cruz 강의 홍수량(1914-1986)에 대한 비매개변수적 변동 핵밀도함수의 추정(VK-C-AC)과 매개변수적 Lognormal과의 비교.

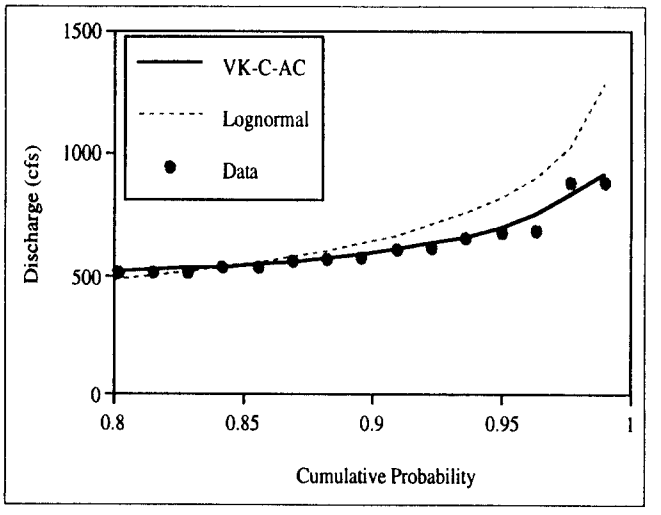

그림 3. Utah, Beaver 강의 연 최대 홍수량에 대한 비매개변수적 변동 핵밀도함수의 추정 (VK-C-AC)과 매개변수적 Lognormal과의 비교.