

미계측유역 지역빈도해석에서 Kernel Regression에 의한 확률홍수량산정

○차영일¹⁾, 문영일²⁾, 성영두³⁾, 정의택³⁾

1. 서론

관심있는 대상지역에 빈도해석에 의한 확률홍수량을 산정하고자 할 때 일반적으로 관측된 홍수량자료가 부족하므로 점빈도해석을 한다는 것은 어려운 일이다. 또는 대상지역이 미계측 지역으로 관측자료가 없는 경우 점빈도해석을 직접사용하기에는 어려움이 따른다. 이런 경우 비교적 자료의 관측기간이 긴 강수량 자료를 사용하여 빈도해석을 한 후 얻어진 확률강수량의 결과를 강우-유출모형을 이용하여 확률홍수량을 산정하거나, 또는 대상지역 주변에 관측기간은 짧지만 여러관측소의 자료가 존재할 때 자료의 동질성이 확보된 지역을 이용하여 지역빈도해석을 실시함으로써 짧은 홍수량자료기간을 보완한 확률홍수량을 산정하는 것이 보편적이다. 본 연구는 후자의 경우로 동질성을 갖는 지역에서 일반적으로 폭넓게 사용되는 L-모멘트법 지역빈도분석을 실시한 후, 각 관측지점의 유역면적과 확률홍수량의 관계를 나타내는 방법에 있어서 기존의 방법인 매개변수적 인 선형 회귀식 방법과 비매개변수적 비선형 회귀식인 Kernel Regression 방법에 의한 확률홍수량을 비교하였다. 기존의 선형 회귀식 방법은 하나의 직선으로 제시되고 있기 때문에 관측자료의 특성을 잘 반영하지 못할 경우가 있으나 비매개변수적 Kernel Regression은 구하고자 하는 유역면적에 대한 확률홍수량을 주변의 관측자료의 가중치에 따라 산정하므로 관측자료의 특성을 잘 표현해 준다 할 수 있다.

2. 미계측유역의 지역홍수빈도해석

미계측유역에 대한 홍수량 산정은 L-moments법을 이용하여 지역확률홍수량을 산정 한 후 면적-홍수량 회귀식을 이용하여 구할 수 있으며 아래 그림 1과 같은 절차를 이용하여 구하게 된다. L-모멘트법은 지역빈도해석 방법으로 대표적인 프로그램으로 web site (<http://lib.stat.cmu.edu/general/lmoments>)상에 프로그램 소스가 공개되어 많이 사용되고 있는 방법이다.

2.1. L-moments

L-모멘트법은 근래에 많이 사용되고 있는 매개변수 추정방법(Hosking, 1997)으로 추정량은 아래 식 (1)~(4)와 같이 나타낼 수 있고, 확률가중모멘트법의 선형조합으로 표시되며, 전통적인 방

1) 서울시립대학교 토목공학과 박사과정

2) 서울시립대학교 토목공학과 조교수

3) 수자원공사 조사기획처

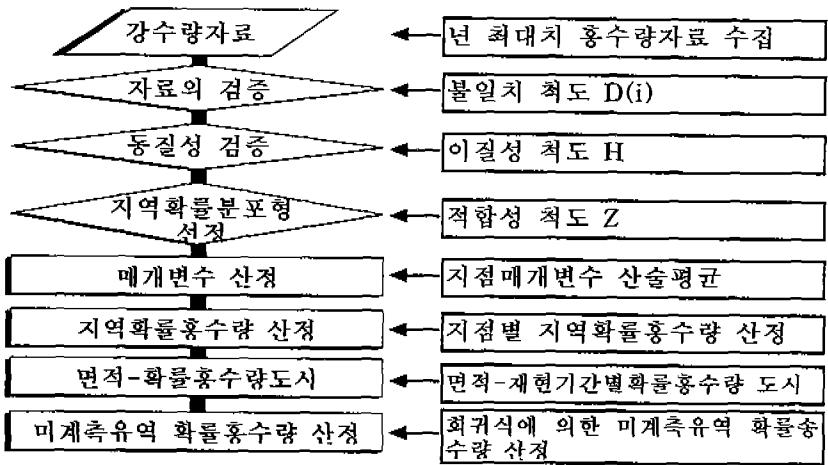


그림 2 미계측 유역에 대한 화를 홍수량 추정 절차

법인 모멘트법이나 최우도법과는 달리 매개변수의 추정에서 거의 편의 되지 않는 것으로 알려져 있다.

$$\lambda_1 = E(X) \quad (1)$$

$$\lambda_2 = 1/2E(X_{(2,2)} - X_{(1,2)}) \quad (2)$$

$$\lambda_3 = 1/3E(X_{(3,3)} - 2X_{(2,3)} + X_{(1,3)}) \quad (3)$$

$$\lambda_4 = 1/4E(X_{(4,4)} - 3X_{(3,4)} + 3X_{(2,4)} - X_{(1,4)}) \quad (4)$$

여기서, $E(\cdot)$ 는 기대값이고, $X(i,n)$ 은 크기가 n 인 표본 중에서 i 번째로 작은 통계량이다. L-모멘트 비는 다음 식 (5)~(7)과 같다.

$$\tau_2 = \lambda_2/\lambda_1 \Rightarrow L-CV(L-Coefficient of variation) \quad (5)$$

$$\tau_3 = \lambda_3/\lambda_2 \Rightarrow L-skewness \quad (6)$$

$$\tau_4 = \lambda_4/\lambda_2 \Rightarrow L-kurtosis \quad (7)$$

2.2. 동질성을 갖는 지역의 구분

동질성을 갖는 자료군의 선택은 다음과 같은 두 가지 절차를 거쳐 선택되어질 수 있다. 첫째, 불일치척도(Discordancy Measure)에 의해 일관성이 없는 자료군이나 이상자료군을 제거한 후, 둘째 이질성척도(Heterogeneity Measure)에 의해 소유역의 소속된 자료군의 수문학적 동질성을 갖는지를 평가하여 자료군을 구분한다.

불일치척도 $D(i)$ 는 다음 식 (8)과 같이 정의한다.

$$D(i) = \frac{1}{3} (U_i - \bar{U})^T S^{-1} (U_i - \bar{U}) \quad (8)$$

여기서, U_i 는 $[t_1^i, t_2^i, t_3^i]^T$ 이며 순서대로 지점 i 의 L-moments비인 L-변동계수, L-왜도, L-첨도인 벡터이고, \bar{U} 는 소유역내 지점들 U_i 의 산술평균 벡터이고, S 는 표본자료의 공분산벡터이다. 지점의 계산된 $D(i)$ 값이 과다하게 크게 되면 소유역내의 자료군에서 일치도가 떨어질 수 있다는 것

을 의미하기 때문에 오류가 있는지 확인할 필요가 있는데; $D(i)$ 에 대한 절대적인 기준은 없지만, Hosking과 Wallis (1997)는 3.0을 기준으로 제안한 바가 있다.

불일치척도에 의해 이상자료군을 제거한 후 소속된 지점의 자료계열의 수문학적 동질성의 파악을 위해 다음 식(9)와 같이 이질성척도 H 를 사용한다(Hosking 등, 1997).

$$H = \frac{V - \mu_v}{\sigma_v} \quad (9)$$

$$V = \frac{\sum_{i=1}^N n_i (t^i - \bar{t})^2}{\sum_{i=1}^N n_i} \quad (10)$$

여기서, 표본자료 집단의 평균 L-moments인 1 , t^R , t_3^R , t_4^R 를 Kappa 분포형에 적합시킨 다음, 충분한 횟수의 모의발생을 실시하여 V 를 계산한 후 평균값과 표준편차를 계산하여 H 를 구한다. Hosking과 Wallis(1997)는 $H < 1$ 이면 동질성 수용 가능 지역, $1 \leq H < 2$ 이면 이질성 가능 지역, $H > 2$ 이면 이질성 지역으로 구분하였다.

2.3. 적정 확률분포형 선정

동질성을 갖는 한 지역 내에 있는 여러 지점들에 가장 적합한 확률분포형을 선정하기 위한 적합성 척도 Z 는 다음 식 (11)과 같다.

$$Z^{\text{DIST}} = \frac{t_4^{\text{DIST}} - \bar{t}_4 + \beta_4}{\sigma_4} \quad (11)$$

여기서, β_4 는 L-kurtosis의 편이, σ_4 는 \bar{t}_4 의 표준편차이고 다음 식(12)와 식(13)과 같다.

$$\beta_4 = \frac{1}{N_{\text{sim}}} \sum_{m=1}^{N_{\text{sim}}} (\bar{t}_4^{[m]} - \bar{t}_4) \quad (12)$$

$$\sigma_4 = \left[\frac{1}{(N_{\text{sim}} - 1)} \left\{ \sum_{m=1}^{N_{\text{sim}}} (\bar{t}_4^{[m]} - \bar{t}_4)^2 - N_{\text{sim}} \beta_4^2 \right\} \right]^{1/2} \quad (13)$$

여기서, 적합성 척도의 임계값은 1.64로 $|Z^{\text{DIST}}| \leq 1.64$ 와 같다(Hosking과 Wallis, 1997).

3. 비매개변수적 Kernel Regression

비매개변수적 회귀분석은 매개변수를 통하지 않고 자료로부터 알고자 하는 지점의 추정값을 추정한다. 이러한 비매개변수적 회귀분석은 주어진 자료의 특성으로부터 잡음(noise)을 제거 또는 감소시킬 수 있으며, 이로 인하여 자료의 해석에 있어 보다 원자료에 근접하는 회귀모형을 구할 수 있다는 장점을 지닌다. 비매개변수적 회귀분석은 기본 회귀분석이 해석하기 어려운 자연계의 이질적이고, 다중변수, 시간과 공간적인 변수를 지니게 되는 자료들에 대한 유용한 해석방법이라고 할 수 있다(Wand 등, 1995). 미매개변수적 Kernel Regression은 관측된 일련의 자료들로부터 Kernel함수를 가중치로 사용하기 때문에 매개변수의 추정을 필요로 하지 않는다. 따라서 사용되는 회귀식의 목적에 따라 Kernel 함수를 선택하고 주어진 자료의 구조에 따라 광역폭(h)이 결정됨으

로써 비대개변수적 회귀모형이 산정된다. 가장 기본적인 형태를 나타내면 다음 식(14)과 같다(Eubank, 1988).

$$f(x) = \frac{\sum_{i=1}^n K\left(\frac{(x-x_i)}{h}\right)y_i}{\sum_{j=1}^n K\left(\frac{(x-x_j)}{h}\right)} \quad (14)$$

여기서, n 은 자료개수, (x_i, y_i) 는 주어진 관측자료이며, x 는 추정하고자하는 값, $K(\cdot)$ 는 Kernel 함수, h 는 광역폭(bandwidth)으로 Kernel regression에 있어서 규모계수(scale factor)로 작용하며 GCV(Generalized Cross-Validation)방법으로 구할 수 있다(조성진, 2000).

위 식을 연속적인 값으로 나타내기 위해 발전시키면 다음과 같다.

$$\hat{f}(x) = \frac{1}{n} \sum_{j=1}^n \int_{s_{j-1}}^{s_j} K\left(\frac{(x-s)}{h}\right) ds Y_j \quad (15)$$

여기서, $s_j = \frac{(x_{j+1} + x_j)}{2}$ 이다.

식(15)은 식(14)와 같은 개념이며, 식(14)에 비하여 상대적으로 bias가 작아지는 경향이 있다. 특히 이 식은 공간적으로 비 균일한 자료에 있어서 가장 적합한 추정값을 제시해 주는 것으로 알려져 있다.

4. 적용 및 결과

4.1. 적용대상유역

이번 연구에 적용된 유역은 네팔의 차멜리야(Chameliya) 지역이며 분석 지점의 유역 현황은 다음 표 1과 같다. 표 1에서는 보유하고 있는 지점 자료의 수와 고도가 3000m이하인 유역면적을 나타내고 있다. 고도가 3000m이상인 곳은 얼음으로 존재하기 때문에 유역면적에서 제외 시켰다. 또한 지점 중 자료의 크기가 6개로 너무 작은 두 지점인 Rimna와 Kalimati는 해석에서 제외 시켰다. 추정하고자하는 대상지점은 발전소를 건설하기 위한 지점으로 미계측 지점이며 전체유역면적은 914km², 고도가 3000m이하인 유역면적은 409km²이다.

표 1 차멜리야 지역의 유역 현황

지점명	자료크기	유역면적(km ²)	지점명	자료크기	유역면적(km ²)
Karkale	25	450	Jamu	32	6129
Patan	19	188	Chisapani	32	20076
Nagma	15	1870	Bargadha	20	3000
Diware	25	150	Navagaon	30	1980
Asara	34	4829	Tigra	21	683
Benighat	33	6785	Kalimati	6	696
Banga	32	5265	Bagasoti	20	3380
Rimna	6	6720	Jalkundi	32	5142
발전소(미 계측)	0	409			

4.2. 지역빈도해석 결과

4.2.1. 불일치척도(Discordancy Measure)와 이질성척도(Heterogeneity Measure)

지역빈도해석 대상 유역의 자료들에 대한 이상치의 여부를 판단하기 위해 각 유역별로 표본 L-모멘트비를 산정하여 불일치척도 $D(i)$ 를 구하였으나, 대상 유역 내 지점자료들에 대한 $D(i)$ 는 모두 3.0보다 작은 값을 가지므로 자료는 양호하다고 할 수 있다. 즉, 이산도 $D(i)$ 가 3보다 큰 경우 배제하는 것이 합리적이다(Hosking, 1997). 또한 각 유역에 대하여 이질성척도 분석에서 Patan과 Bargadha 유역을 제외한 자료를 분석한 결과 H_1, H_2, H_3 모두 Hosking과 Wallis(1997)의 기준을 만족하였다.

4.2.2. 적정분포형선정 및 지역빈도해석결과

본 연구에서는 구분된 동질성의 유역에 대하여 Generalized Logistic, Generalized Extreme Value(GEV), Log-normal, Pearson type-III, Generalized Pareto 등의 확률분포형을 적용하였다. 적합도 검정(Z)의 결과 중 상대적으로 Z값이 가장 작은 Generalized Logistic분포를 최적분포형으로 선정하였다.

4.3. 유역면적-확률홍수량에 의한 미계측유역의 확률홍수량 산정

4.3.1. 지점별 확률강홍수량 산정

적합도 검정의 결과에서 채택된 분포형인 Generalized Logistic분포형을 기준으로 각 유역의 100년 빈도 확률홍수량을 산정하였고, 결과는 표 2와 같다.

표 2 각 유역에 대한 100년 빈도 확률홍수량

지점명	자료크기	유역면적	100년 빈도 확률홍수량	지점명	자료크기	유역면적	100년 빈도 확률홍수량
Karkale	25	450	623.5625	Jamu	32	6129	8691.086
Nagma	15	1870	810.8638	Chisapani	32	20076	24259.2
Diware	25	150	492.1333	Nayagaon	30	1980	1391.063
Asara	34	4829	5957.403	Tigra	21	683	1436.473
Benighat	33	6785	9129.798	Bagasoti	20	3380	3723.383
Banga	32	5265	7492.151	Jalkundi	32	5142	5907.57

4.3.2. 미계측유역의 확률강수량 산정

미계측유역인 발전소 지점에 대한 확률홍수량을 산정하기 위하여 매개변수적 회귀식과 비매개변수적 Kernel Regression을 적용하여 얻은 결과는 그림 2와 그림 3과 같다.

유역면적-확률홍수량에 대한 매개변수적 회귀식은 그림 2의 식과 같고, 매개변수적 회귀식에 의한 미계측유역인 발전소지점의 100년 빈도 확률홍수량은 $670 \text{ m}^2/\text{s}$ 이며, 비매개변수적 Kernel Regression의 100년 빈도 확률홍수량은 $857 \text{ m}^2/\text{s}$ 으로 산정되었다. 선형회귀식에 의한 확률홍수량이 작게 산정된 이유는 최대 유역면적인 Chisapani지점의 확률홍수량 값이 작게 반영된 결과로 판단된다. 또한 결정계수는 각각 $R^2 = 0.867$, $R^2 = 0.988$ 로써 Kernel Regression의 값이 더 크다.

며 자료의 분산을 잘 표현하는 것으로 볼 수 있다.

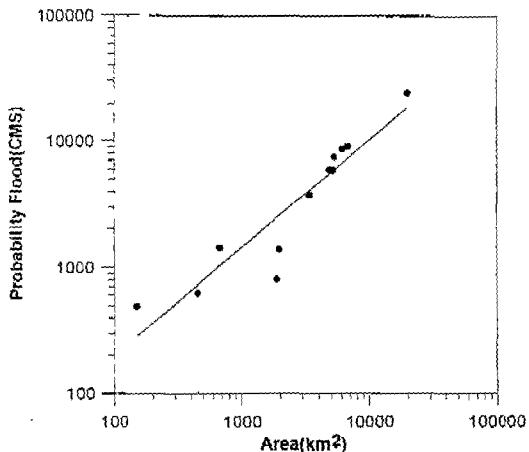


그림 3 매개변수적 회귀식($R^2 = 0.867$)

$$(회귀식 : Q_{100} = 3.987 \times A^{0.852})$$

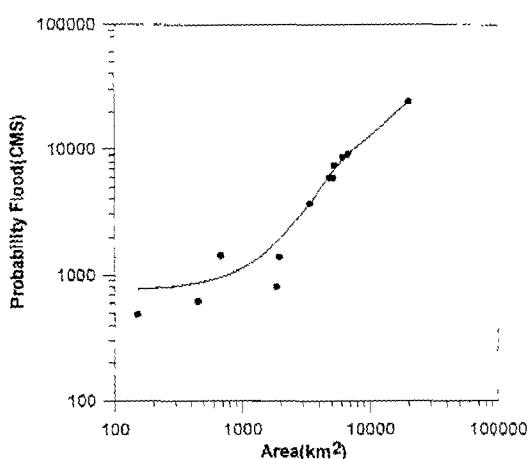


그림 4 비매개변수적 Kernel Regression

$$(R^2 = 0.988)$$

5. 결론

본 연구에서는 네팔의 차멜리야(Chameliya) 지역의 미계측유역에 대한 확률홍수량을 산정하는 방법으로 L-moments법을 이용하여 불일치척도(Discordancy Measure)와 이질성척도(Heterogeneity Measure) 검정과 적정분포형선정을 통한 지역별도 해석을 실시한 후, 산정한 확률홍수량값을 유역면적-확률홍수량의 그래프에 도시하여 선형 회귀식과 Kernel Regression을 적용한 후 미계측유역의 면적에 대한 확률홍수량 값을 비교하였다. Kernel Regression 방법은 자료 자체의 특성을 이용하여 회귀식을 구하기 때문에 결정계수(98.7%)가 기존의 선형 회귀식(86.7%)보다는 크게 나와 자료의 분산을 잘 표현하는 것으로 볼 수 있다.

참고문헌

- 조성진. 2001. 매개변수적 방법과 비매개변수적 방법의 수위-유량곡선 비교연구, 서울시립대학교, 석사논문.
- Eubank, R.L. 1988. Spline Smoothing and Nonparametric Regression. DEKKER, New York.
- Hosking, J.R.M. and Wallis, J.R. 1997. Regional frequency analysis : An approach based on L-moments, Cambridge University Press.
- Tand, M.P. and Jones, M.C., 1995. Kernel Smoothing. CHAPMAN & HALL, New York.