지역가중다항식을 이용한 빈도해석에 관한 연구

A Study of Frequency Analysis by using Locally Weighted Polynomial Method

문영일* / 정민수 / 최병규*** / 유승연****

Moon, Young-Il / Jeong, Min Su, / Choi, Byung Gyu / You, Seung Yeon

요 지

유량자료를 이용한 매개변수적 빈도해석 방법은 주관적인 분포형 선정문제를 안고 있다. 이러한 분포형 선택문제는 수문자료의 오랜 축척에 따른 통계적 분석을 통해 하나의 확률분포형을 선택할 수 있는 경우 극복될 수 있을 것이다. 그러나, 일반적으로 수문자료의 관측 기간이 짧아 하나의 분포형을 선택하는데 어려움을 갖고 있다. 반면에, 지역가중다항식을 이용한 빈도해석의 경우 단일분포형 선택문제가 아닌 자료로부터 매개변수를 선택하고 추정함으로서 White noise를 제거 또는 감소하며 자연계의 이질적, 다중변수적 그리고 시공간적 특성을 잘 반영할 수 있는 것으로 알려져 있다.

따라서 본 연구에서는 단일 주관분포 선택문제가 아닌 자료로부터 매개변수의 선택·추정이 이루어지는 지역가중다항식을 이용한 빈도해석을 수행하였다. 분석에는 서울강우자료로 매개변수적 빈도해석을 수행하는 경우 Gumbel, GEV(Type I Extreme Value) 그리고 LN2 (Log-Normal 2) 등의 분포형을 적용하여 지역가중다항 추정자의 산출 결과와 비교·검토하였다. 또한 각각의 방법을 적용해 이중첨두(bimodal) 분포형에 대한 모형의 적합성을 도시적으로 비교·산정하였다.

핵심용어 : 지역가중다항식, 빈도해석

1. 서 론

일반적으로 매개변수적 빈도해석은 주관적인 분포형 선택문제가 있다. 미국은 Log-Pearson Type Ⅲ라는 단일분포형을 적용하고 있고 국내의 경우도 “한국확률강우량도 산정(건설교통부, 2000)”에서 전국의 모든 지속시간에 걸쳐 Gumbel 확률분포형을 적용한 사례도 있다. 하지만 미국의 경우와 같이 오랜 기간 수문자료에 대한 축적에 따른 통계적 분석을 통해 하나의 확률분포형을 선택하였다고 보기는 어렵다. 또한 Gumbel이라는 분포형이 모든 지속시간과 모든 강우관측지점에 적합한 확률분포형인지에 대한 의문과 이를 극복하기 위한 다양한 연구가 시도되어 왔다.

이러한 문제제기에 따른 방법적 선택 중 하나가 단일분포형 선택문제가 아닌 자료로부터 매개변수의 선택과 추정을 수행하고 자료특성으로부터 백색잡음(white noise)을 제거 또는 감소하는 방법인 비매개변수적 방법을 이용한 빈도해석이라 할 수 있다. 비매개변수적 방법은 원자료에 근접된 회귀모형을 할 수 있고 자연계의 이질적, 다중변수적 그리고 시공간적 특성을 잘 표현하는 것으로 알려져 있다. 반면에 비매개변수적 빈도해석을 수행하는 경우 실제 분포형을 반영하는데

* 정회원·서울시립대학교 공과대학 토목공학과 부교수·E-mail : ymoon@uos.ac.kr

** 정회원·서울시립대학교 공과대학 토목공학과 박사과정·E-mail : jminsoo05@uos.ac.kr

*** 정회원·삼안기술공사 수력부 전무·E-mail : bkchoi@samaneng.com

**** 정회원·서울시립대학교 공과대학 토목공학과 석사과정·E-mail : dpzh08@uos.ac.kr

손실이 발생될 수 있으며, 자료에 의존성으로 인해 외삽에 대한 반영이 이루어지지 않을 수 있고 분포함수가 과대평활하는 경우가 발생될 수 있다. Adamowski(1989)는 외삽문제를 나타내는 변이적인 광역폭 밀도 추정자에 대한 연구를 하였는데, Moon과 Lall(1994)에 의해 그 연구결과는 분위수들에 대한 회귀 추정자를 기초로 하는 비매개변수적 핵함수로 발전된 바 있다. 비매개변수적 Kernel 핵함수 방법과 지역가중다항식 방법은 함수를 결정하고 광역폭(h) 선정을 통해 자료로부터 매개변수 선택과 추정을 수행함에 있어 유사하다고 할 수 있다. 하지만 Kernel 방법은 Gaussian, Epanechnikov, Cauchy 등 다양한 핵함수들 중에 선택하여 광역폭 선정에 따른 가중치를 적용하는 반면 지역가중다항식의 경우는 이러한 핵함수와 달리 광역폭 선정에 따른 낮은 차수의 다항식을 이용하여 국부적인 가중치를 적용하므로 함수 선택에 차이를 갖는다.

본 연구에서는 지역가중다항추정자를 이용한 빈도해석 수행과 외삽문제에 따라 Adamwski의 도시공식을 적용하였다. 또한 대상유역은 서울 강우자료를 이용한 빈도해석을 수행하였는데 이 경우 일반적으로 Gumbel 분포형, GEV 분포형 LN2 분포형 등을 이용한 매개변수적 빈도해석을 수행하여 지역가중다항식을 이용한 빈도해석과 비교분석하였다. 또한 각각의 방법을 적용해 이중첨두(bimodal) 분포형에 대한 모형의 적합성을 도시적으로 비교・산정하였다.

2. 지역가중다항 추정자 (Local Polynomial Estimator)

2.1 일반식과 추정절차

지역가중다항 추정을 위한 경험적 분위함수에 대한 일반적인 모델을 고려하면 다음과 같다.

$$Y_i = \mu(X_i) + \varepsilon_i \tag{1}$$

여기에서, $\mu(\bullet)$는 비선형 함수이고, $X_1 \in [0,1]$이며 ε_1는 평균 0을 갖는 잔차이다. 어떤 자료계열의 N년 수문추정을 고려한다면 관심대상은 추정자 $\mu(X_T)$ 즉, $X_T=1-1/T$이며 $\mu(X_T)$는 지역가중다항 회귀식을 이용하여 추정된다. $\mu(X_T)$는 일반적인 함수로 연속적이고 (p-1) 미분을 가지므로 테일러급수에 따라 차수 p인 지역 다항식을 사용하여 $\mu(X_T)$를 근사하는 것은 합리적이다. 지역가중다항식의 지역(Local)은 X_T의 근방에 있는 근사를 말하며 이웃 항의 수인 k는 목적 회귀함수의 평활과 residual process(ε_1)의 본질적 특성에 의존한다. 지역다항추정(LOCFIT)은 Loader(1999)에 의해 연구 발전되었다. 다음은 LOCFIT 방법의 추정절차로 절차는 반복적으로 수행된다.

1. 경계 조건에 근접한 근방점들(X_T), (k(=αn))의 선정
   - α의 범위는 0~1이며 α가 1이면 전 자료점들을 이웃항으로 가짐
2. Bisquare, Tricubic 등의 가중치 적용함수의 결정과 각각의 k 자료쌍의 거리가중치 적용
3. 경계치가 k항인 Smoothing 범위에서 $\mu(X_T)$는 차수 p인 다항식에 근사하며 a_0, a_1 그리고 a_2는 WLS(Weighted Least Squares) Min[F(x)]를 이용

$$\mu(X_T) = a_0 + a_1 + a_2(X)^2$$
$$\sum_{i=1}^{k} W_i(X_T)(Y_i - \mu(X_i))^2 \tag{2}$$

2.2 광역폭의 선택과 추론

지역가중다항 추정자의 주요변수는 이웃 항(k)과 다항식차수(p)이며 k는 광역폭(h)와 관련된다. h가 작으면 자료특성이 불충분해서 잡음과 분산이 증가된다. 반면에 h가 커지면 추정이 원활하지 않고 편의 증가로 평균함수인 $\mu(X_T)$의 중요 특징들이 왜곡될 수 있다. 따라서 h는 편의와 분산의 절충점에서 선택되어야 한다. 또한, p는 고차다항식의 목적함수 $\mu(X)$ 추정에 정밀근사치를 제공하여 편의가 작아지는 반면 계수의 수가 많아지고 결과의 변동성이 커질 수 있다. 따라서 h와 p의 선택문제에 따른 적합도 선정방법이 제시가 필요한데 본 논문에서는 예측오차와 같은 장래예측을 표현에 적합한 GCV(Generalized Cross Validation) 추정법을 이용하였고 그 식은 다음과 같다.

$$GCV(\alpha, p) = n\frac{\sum_{i=1}^{n}(Y_i - \hat{\mu}(X_i))^2}{\left(1 - \sum_{i=1}^{n} h_{ii}\right)^2} \tag{3}$$

여기서, n은 샘플 크기이고, $Y_i - \hat{\mu}(X_i)$는 잔차이며 h_{ii}는 H(hat matrix)의 대각 항으로 표준선형회귀절차에 따른 $X(X^T)X^{-1}X^T$로 추정된다. 또한, GCV의 지역적 적용은 LGCV(Locall GCV)로 국부적 예측잔차의 분산을 얻는데 사용되며, 큰 자료군으로부터 적지 않은 추정치들을 나타내는 경우로 비선형 시계열모델 예측을 위해 지역회귀식이 사용되는 경우이다.식은 다음과 같다.

$$LGCV_1(\hat{f}) = \frac{e_i^T W_l e_l}{\left(\frac{k - d'}{k}\right)^2} \tag{4}$$

2.3 Adamowski의 확률도시공식

일반적으로 다양한 확률도시공식이 제안되고 있지만 대부분의 경우 다소 임의적이라고 할 수 으며 사용성이 많은 Weibull의 확률도시공식(i/(n+ 1))의 경우도 편의성과 보수적인 결과를 보이는 것으로 알려졌다. 이와 달리 Adamowski(1981)가 제안한 도시공식은 MSE (Mean Square Error)에 기초를 둔 것으로 확률도시공식이 큰 값에서 실제 초과확률값에 근접함을 보였고 Pearson Type Ⅲ 분포형에 맞는 적절한 확률도시공식이 없음을 증명하였다. 다음은 기록된 연최대 자료를 경험적 분위함수로 나타내는 식으로 순차적 계열의 자료 한 쌍((X_i, Y_i), $i = 1,2,...,n$)으로 정의될 수 있다. 여기서, x_i = (i-0.25)/(N+ 0.50)이고, Y_i는 연최대 자료순위이다.

3. 모형의 적용

3.1 지역가중다항 추정자를 이용한 복잡한 양상을 갖는 함수의 추정

모형을 이용한 빈도해석에 적용에 앞서 지역가중다항 추정자와 매개변수적인 방법을 이용해 다양하게 발생될 수 있는 자료들에 대한 추정을 비교분석하였다. 분석에는 정규분포 N(0,1), N(1,2)와 N(0,1), N(1,3)을 갖는 이중첨두(bimodal) 분포형이 사용되었다. 다음은 각 추정결과의 도시적 표현이며 지역가중다항 추정자의 광역폭(h) 0.15인 경우가 가장 잘 표현되었다.

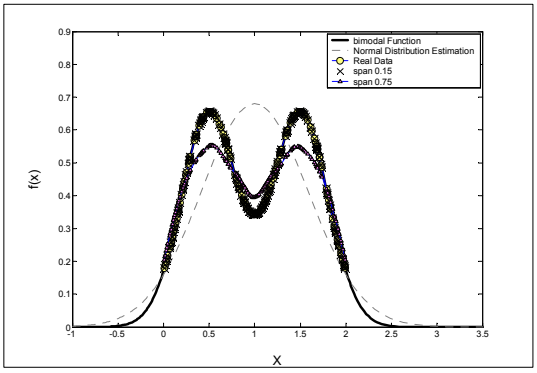

Fig. 1. bimodal N(0, 1)과 N(1, 2) (h : 0.15)

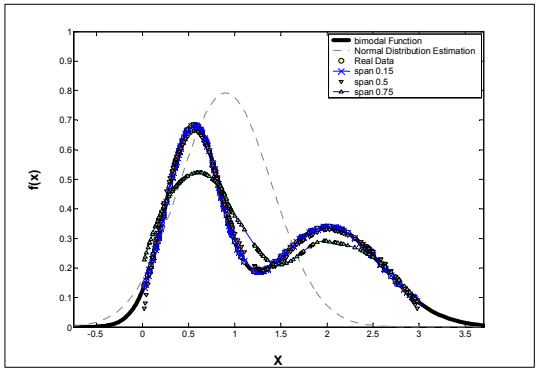

Fig. 2. bimodal N(0, 1)과 N(1, 3) (h : 0.15)

3.2 모형의 적용

본 연구에서는 일반적인 매개변수적 빈도해석과 지역가중추정자를 이용한 빈도해석을 수행하여 그 결과를 비교·분석하였다. 대상은 서울지역 강우자료를 이용하였으며 매개변수적 방법에 사용된 분포형은 각각 GUM, GEV 그리고 LN2 분포형이다. 경험적 분위함수(Empirical CDF)는 Adamwoski 도시공식을 이용하였다. 다음 그림은 방법별 추정 결과를 나타내는 것으로 매개변수적 방법으로 제시된 것은 재현기간 100년에 따른 확률가중모멘트 방법을 이용한 추정결과이다.

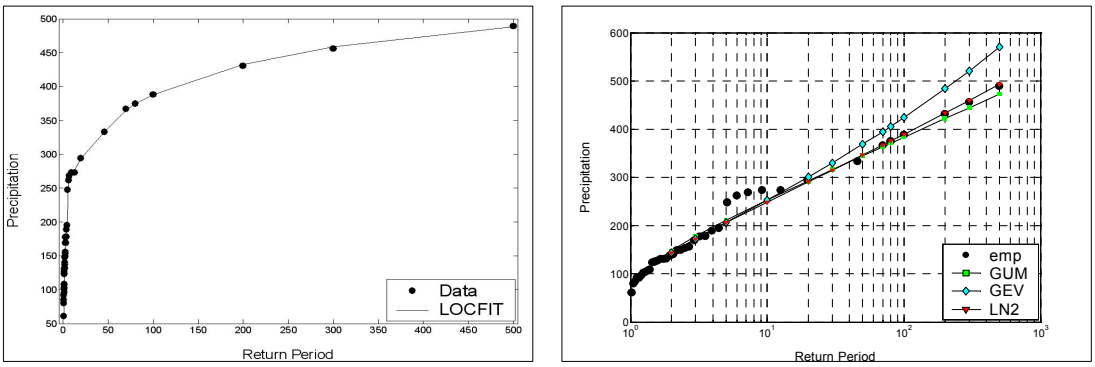

Fig. 3. 지역가중다항 추정자와 매개변수적 확률가중모멘트 방법을 이용한 빈도해석 결과

4. 결과

이상과 같이 서울지역 연 최대치 일강우자료에 대한 방법별 빈도해석 결과는 다음 표로 정리될 수 있으며, 빈도 100년, 200년, 300년 그리고 500년에 대해 매개변수적 방법과 지역가중 다항식을 이용해 얻은 결과를 Empirical CDF와 각각 비교 산정하였다.

Table 1. 서울지점 일 확률강우량 산정 결과 비교

	재현기간별 확률 강우량(mm)			
	100년	200년	300년	500년
LN2-MOM	387.4	429.9	455.3	487.8
LN2-MLM	382.7	424.1	448.4	480.4
LN2-PWM	389.2	433.2	459.4	493.1
LOCFIT	388.3	432.5	458.6	488.6
Emp. CDF	388.1	430.8	456.2	488.8

매개변수적 방법의 분포형 선정결과 분포형은 LN2가 선정되었으며 분포형별로 확률가중모멘트법이 가장 적합한 추정이 이루어졌다. 하지만 선택된 분포형과 매개변수에 따라 GUM-PWM과 LN2-MOM 추정도 매우 좋은 추정을 나타냈으며 또한 적합도 검정에서는 기각률이 높은 PPCC 검정에서는 기각을 나타냈다. 반면에 지역가중다항식을 적용한 경우는 관측자료의 누가분포함수에 비하여 0.2 ~ 2.4 mm/day의 편차를 보임으로서 가장 근사적인 추정이 이루어졌음을 알 수 있다.

5. 결론

본 논문에서는 지역가중추정자의 기법적 특성을 연구하였고 지역가중다항식의 적용성 검토를 위하여 이중첨두를 갖는 확률분포형의 형태에 적용과 실제 서울지역 강우자료에 적용하여 매개변수적 빈도해석 방법과 비교·분석을 수행하였다.

지역가중다항 추정자와 비교하기 위해 매개변수적 방법을 동일지점, 동일자료에 대하여 일반적인 빈도해석 절차에 따라 분석을 실시하였다. 추정결과 이중첨두 분포형은 광역폭 h에 따라 좋은 추정을 보인반면 매개변수적 방법은 추정이 이뤄지지 않았다. 서울지역 강우의 경우는 매개변수적 빈도해석 방법에서는 LN2-PWM 분포형이 가장 적합한 것으로 선정되었지만 분포형별·매개변수 추정별로 선택된 값이 큰 차이가 없어 분포형 선택문제를 나타냈다. 이와 달리 지역가중다항추정자는 0.2 ~ 2.4 mm/day의 약간의 편차를 보였지만 가장 좋은 추정치로 나타났다.

따라서 자료의 축적문제와 명확한 분포형 제시 기준이 마련되지 않는 국내 현실을 고려할 때 비매개변수적 지역가중다항식은 빈도해석에 좋은 대안으로서 제시될 수 있을 것이다.

참고문헌

1. 정민수(2005). “지역가중다항식의 "수문학적 적용성 분석에 관한 연구“, 서울시립대학교 공학석사학위논문
2. 허준행(1996). “확률가중 모멘트법을 이용한 매개변수 추정과 적용”, 한국수자원학회 학술대회지, pp.107-190.
3. Adamowski, K., 1981, Plotting formula for flood frequency, Water resources Bulletin, Vol. 17, No. 2, pp, 197~202.
4. Adamowski, K., and W. Fleuch, 1990, Nonparametric flood-frequency analysis with historical information, Journal of Hydraulic Engineering, 116(8), 1035-1047.
5. Cleveland, W. S., and S. J. Devlin, 1988, Locally Weighted Regression: An Approach to regression Analysis by Local Fitting, J. Amer. Stat. Assn., 83 (403), 596-610.
6. Cleveland, W. S., and S. J. Devlin, and E. Grosse, 1988, Regression by Local Fitting, J. Econometrics, 37, 87-114.
7. Lall, u., and M. E. Mann, 1995, The Great Salt Lake: a barometer of low frequency climatic variability, Water Resour. Res., 31(10), pp 2503-2516.
8. Lall, U., Y.-I. Moon and K. Bosworth, Kernel Flood Frequency Estimators: Bandwidth Selection and Kernel Choice, Water Resources Research, 29(4), 1003-1015.
9. Loader, C., 1999, Local regression and likelihood, Springer-Verlag, New York.
10. Moon, Y.-I., and U. Lall, 1994, A kernel Quantile Function Estimation.