Generalized Logistic 분포형의 최우도법을 이용한 확률홍수량의 근사적 분산

Asymptotic Variance of Flood Quantiles from the Generalized Logistic Distribution using the Method of Maximum Likelihood

신홍준*, 허준행**, 김영일***

Hongjoon Shin, Jun-Haeng Heo, Youngil Kim

요 지

최근 영국의 Institute of Hydrology에서는 Generalized logistic (GL) 분포형을 홍수빈도해석시 GEV 분포형을 대체하는 분포형으로 추천한 바 있으며, 그로 인해 GL 분포형의 사용이 증가하고 있는 추세이다. 하지만 아직 그 사용빈도에 반하여 분포형 자체의 특성, 그 중에서도 확률홍수량의 근사적 분산에 관한 연구는 거의 이루어지지 않았다. 따라서 본 연구에서는 최우도법을 이용하여 GL 분포형의 확률홍수량에 대한 근사적 분산에 관한 연구를 수행하였으며, 이를 표본 크기, 재현기간, 매개변수들의 함수로 나타내었다. 또한 확률홍수량의 근사적 분산의 적용성을 검토하기 위해 Monte Carlo 모의실험을 수행하였으며, 모의실험은 형상매개변수(β)가 ±0.5이면 gamma function으로 인하여 표본 크기에 관계없이 분산값이 무한대에 가까워지므로 형상매개변수의 범위는 $-0.5 \le \beta \le +0.5$로 제한하였다. 모의결과 최우도법에 의해 계산된 분산식은 형상매개변수 $-0.25 \le \beta \le +0.5$의 범위에서 비교적 잘 맞는 것을 확인할 수 있었으며, 기존에 알려진 대로 표본크기가 크면 클수록 정확해지는 것을 알 수 있다. 또한 표본크기가 작은 경우 형상매개변수 전 범위에서 정확도가 떨어지는 것을 확인할 수 있으며, 최우도법의 경우 표본크기가 작은 경우를 제외하고 $-0.25 \le \beta \le +0.5$ 범위에서 quantile 산정시 quantile이 약간 과다추정되는 경향이 있는 것을 알 수 있으며, 이는 분산이 과다추정되는 결과를 초래하며 이로 인해 해석해보다 약간씩 큰 값을 나타내는 것으로 판단되었다.

핵심용어 : 근사적 분산, 확률홍수량, generalized logistic 분포형

1. 서 론

추정값의 오차는 크게 표본 자료의 부족으로 인한 오차와 부적절한 분포형의 선택으로 인한 오차로 나눌 수 있다. 그러므로 특정 재현기간에 대한 quantile의 추정값은 분포형의 정확도를 나타내는 척도가 존재하지 않는다면 추정값 자체로는 큰 의미를 갖는다고 할 수 없다. 위에서 언급한 오차 중 본 연구에서 사용되는 표준오차는 표본 자료의 부족으로 인한 오차와 관련이 있다고 할 수 있는데, 추정값의 표준오차는 일반적으로 매개변수 추정방법에 다르며 각 매개변수 추정방법은 각기 다른 표준오차를 나타내게 되어 가장 효율적인 방법은 quantile에 대한 표준오차 추정값 중 가장 작은 값을 가지는 매개변수 추정방법이 될 것이다 (Rao와 Hamed, 2000). 현재까지 표준오차와 신뢰구간과 관련된 연구는 많이 있어왔다. 신뢰구간과 표준오차와 관련된 초창기의 연구는 Yevjevich (1964)에 의해 정리된 바 있으며, Nash와 Amorocho (1966)는 normal 분포형과 EV1 분포형에 대해서 특정재현기간과 관련된 추정값을 표본크기와 매개변수들의 함수로 유도한 바 있다.

* 정회원, 연세대학교 대학원 토목공학과 박사과정, E-mail : sinong@yonsei.ac.kr

** 정회원, 연세대학교 사회환경시스템공학부 토목환경공학과 교수, E-mail : jhheo@yonsei.ac.kr

*** 정회원, 연세대학교 대학원 토목공학과 석사과정, E-mail : clearblue01@yonsei.ac.kr

또한 Bobee (1973)는 위험도 측정과 관련하여 신뢰구간을 유도하였으며, Heo 등 (2001)은 모멘트법, 최우도법, 확률가중모멘트법에 의한 3변수 Weibull 분포형의 quantile에 대한 신뢰구간을 유도하였다.

Generalized logistic (GL) 분포형은 영국의 Flood Estimation Handbook (Institute of Hydrology, 1999)에서 영국의 홍수량 자료에 대해서 사용이 추천된 바 있는 분포형이다. 이 분포형의 특징은 형상매개변수가 음수일 경우 추정되는 quantile의 상한계가 존재하지 않는다는 점이다. 즉, 홍수빈도해석에서 상한계를 가진다는 것은 빈도해석 결과로 나타나는 추정값이 관측값의 최대값에 근접해진다는 것으로 빈도해석에 있어서 단점 중의 하나로 지적되는 것이다. 이러한 이유로 GL 분포형의 적용이 점차 증가하고 있는 추세라고 할 수 있으나, 분포형의 정확도를 나타내는 신뢰구간에 대한 연구는 아직 미흡한 실정이다. 따라서 본 연구에서는 최우도법(method of maximum likelihood)에 의한 GL 분포형의 quantile 추정 과정을 정리하였으며, quantile 추정값의 근사적인 분산을 표본크기, 재현기간, 매개변수들의 함수로 나타내었다. 또한 Monte Carlo 모의 실험을 수행하여 이렇게 유도된 quantile에 대한 신뢰구간의 적용성을 살펴보았다.

2. 모 형

GL 분포형은 2변수 logistic 분포형의 일반화된 모형이며 kappa 분포형의 특수한 모형이라고 할 수 있다. Logistic 분포형으로부터의 일반화 과정은 기존의 문헌들에서 정의되었던 것들과는 다른 것으로, Ahmad 등 (1988)의 log-logistic 분포형을 재매개변수화한 형태라고 할 수 있다. 또한 GL 분포형의 명칭은 generalized Pareto 분포형이나 generalized extreme value 분포형과 같이 분포형의 특징을 반영하기 위해 붙이진 것으로 (Hosking과 Wallis, 1997), GL 분포형의 누가분포함수와 확률밀도함수는 각각 다음과 같다.

$$F(x) = \left[1 + \left\{1 - \frac{\beta}{\alpha}(x - x_0)\right\}^{1/\beta}\right]^{-1} \tag{1}$$

$$f(x) = \frac{1}{\alpha}\left\{1 - \frac{\beta}{\alpha}(x - x_0)\right\}^{1/\beta - 1}\left[1 + \left\{1 - \frac{\beta}{\alpha}(x - x_0)\right\}^{1/\beta}\right]^{-2} \tag{2}$$

여기서, x_0는 위치매개변수, α는 크기매개변수, β는 형상매개변수이며, GL 분포형의 범위는 $\beta < 0$인 경우, $x_0 + \alpha/\beta \le x \le \infty$, $\beta > 0$인 경우, $-\infty < x \le x_0 + \alpha/\beta$이며 $\beta = 0$인 경우 GL 분포형은 2변수 logistic 분포형이 된다.

3. 최우도법에 의한 근사적 분산

최우도법에 의한 quantile 추정값의 분산은 다음의 분산, 공분산을 이용하여 구할 수 있다.

$$Var(x_0) = 3\alpha^2(-S_1S_3 + S_4\beta^2 + S_1S_3S_4\beta^2 - S_5^2 - 1)/D \tag{3}$$

$$Var(\alpha) = 3\alpha^2\beta^2(S_1^2S_3g_2 - 2S_1S_5g_2 + S_1S_4g_2\beta^2 - S_5^2 + 2S_1S_2S_5 - S_1^2S_2^2)/D \tag{4}$$

$$Var(\beta) = 3\beta^4S_1(g_2 + S_1S_3g_2 - S_1S_2^2)/D \tag{5}$$

$$Cov(x_0, \alpha) = -3\alpha^2\beta(-S_1S_2S_5 + S_1S_2S_4\beta^2 - S_5^2 + S_1S_3S_5 + S_5 - S_1S_2)/D \tag{6}$$

$$Cov(x_0, \beta) = -3\alpha\beta^2S_5(-S_1S_2 + 1 + S_1S_3)/D \tag{7}$$

$$Cov(\alpha, \beta) = 3\alpha\beta^3S_1(-S_5g_2 + S_1S_3g_2 + g_2 + S_2S_5 - S_1S_2^2)/D \tag{8}$$

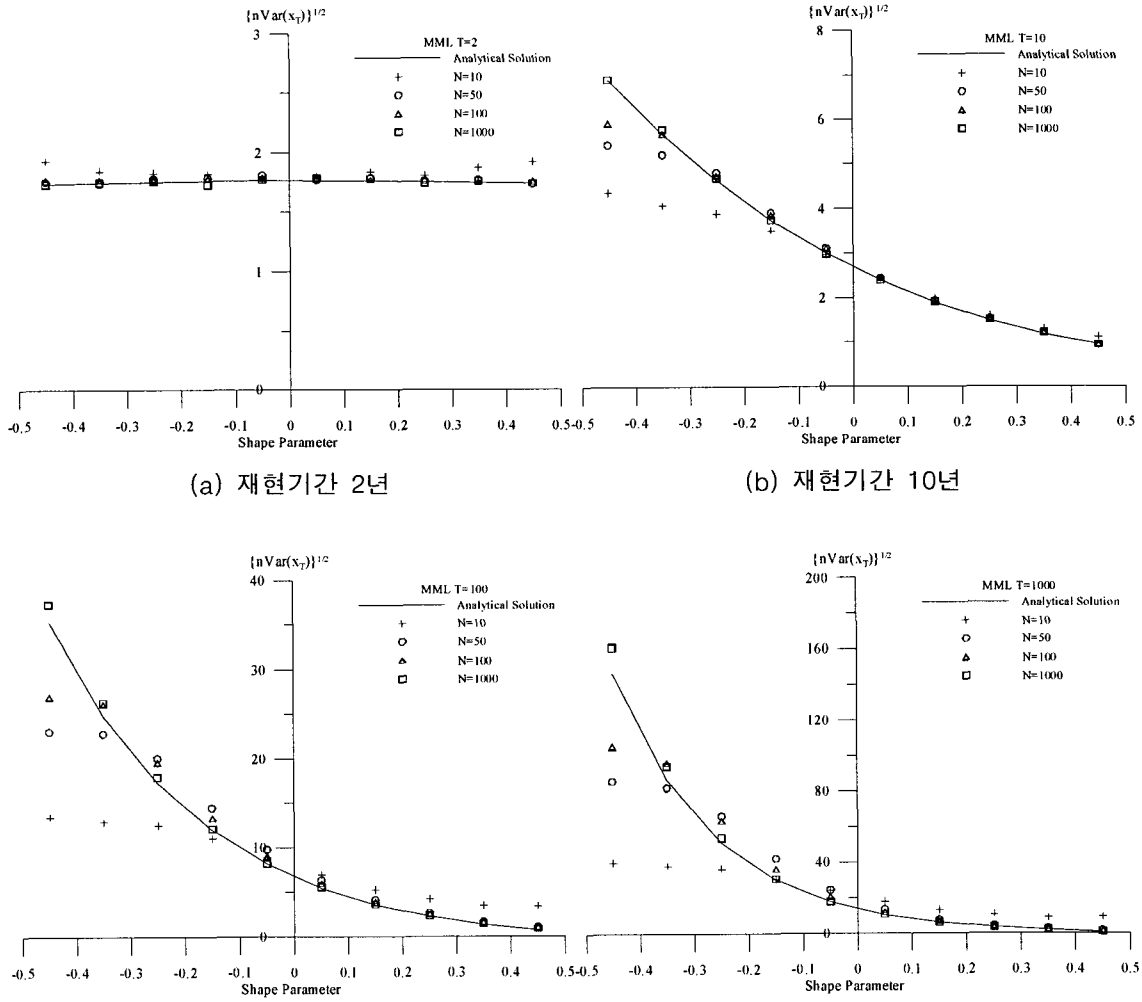
여기서, $S_1 = 1 - \beta^2$, $S_2 = g_2 - g_1$, $S_3 = g_2 - 2g_1$, $S_4 = 1 + \frac{1}{\beta^2} + \frac{\pi^2}{3}$,

$S_5 = g_1\left\{1 - \frac{\beta(1-\beta^2)}{\psi(1-\beta) - \psi(\beta)}\right\}$,

$$D = N\left\{(S_2^2 - S_3g_2)S_1^2 + (g_2 + S_1S_3g_2 - S_1S_2^2)S_1S_4\beta^2 + (2S_2 - g_2 - \frac{1}{S_1} - S_3)S_1S_5^2 - S_1g_2\right\}.$$

4. Monte Carlo 모의실험

Monte Carlo 모의는 근사적으로 유도된 수식의 수치적인 평가를 위해 많이 사용되는 방법으로, 모의를 통해 가정된 성질을 가지는 여러 가지 표본자료를 발생시키고, 발생된 자료가 설정된 자료 길이와 반복 횟수로부터 생성된 통계적인 성질을 가지는 모집단이라고 가정한다. 본 연구에서 모의실험은 최우도법에 의한 quantile의 근사적 분산의 적용성을 검토하기 위해 수행되었다. 이를 위해 위치매개변수 (location parameter)와 규모매개변수 (scale parameter)는 각각 $x_0 = 0$, $\alpha = 1$로 고정시켰으며, 형상매개변수 (shape parameter) β는 -0.45, -0.35, -0.30, -0.25, -0.20, -0.15, -0.10, -0.05, +0.05, +0.10, +0.15, +0.20, +0.25, +0.30, +0.35, +0.40, +0.45로 변화시켰다. 가정된 매개변수에 대해서 표본크기 N = 10, 50, 100, 1000에 대한 10,000개의 자료들을 발생시켰으며, 각각의 발생된 자료를 이용하여 매개변수들을 추정하고 추정된 매개변수들을 이용하여 quantile과 그 분산을 계산하였다. 다음의 그림 1은 재현기간 2년, 10년, 100년, 1000년에 대하여 최우도법에 의한 해석해와 모의값의 비교를 나타낸 그림이다.



(a) 재현기간 2년 (b) 재현기간 10년

(c) 재현기간 100년 (d) 재현기간 1000년

그림 1. 최우도법에 의한 해석해와 모의값의 비교

모의결과 최우도법에 의해 계산된 분산식은 형상매개변수 $-0.25 \leq \beta \leq +0.5$의 범위에서 비교적 잘 맞는 것을 확인할 수 있었으며, 기존에 알려진 대로 표본크기가 크면 클수록 정확해지는 것을 알 수 있다. 또한 표본크기가 작은 경우 형상매개변수 전 범위에서 정확도가 떨어지는 것을 확인할 수 있으며, 최우도법에 의한 결과의 경우 표본크기가 작은 경우를 제외하고 $-0.25 \leq \beta \leq +0.5$ 범위에서 quantile 산정시 quantile이 약간 과다추정되는 경향이 있는 것을 알 수 있으며, 이는 분산이 과다추정되는 결과를 초래하며 이로 인해 해석해보다 약간씩 큰 값을 나타내었다.

5. 결 론

본 논문에서는 모멘트법, 최우도법, 확률가중모멘트법에 기초하여 3변수 generalized logistic (GL) 분포형에 대한 신뢰구간 결정에 관한 연구를 수행하였다. 본 연구에서는 GL 분포형에 대한 quantile 추정값의 근사적인 분산을 표본 크기, 재현기간, 매개변수 등의 함수로 나타내었으며, 이렇게 유도된 분산은 모집단의 quantile에 대한 신뢰구간 추정에 사용될 수 있다. 또한 최우도법에 의해 유도된 신뢰구간의 적용성을 파악하기 위해 모의실험을 수행하였으며, 모의결과 최우도법에 의해 계산된 분산식은 형상매개변수 $-0.25 \leq \beta \leq +0.5$의 범위에서 비교적 잘 맞는 것을 확인할 수 있었으며, 기존에 알려진 대로 표본크기가 크면 클수록 정확해지는 것을 알 수 있다. 또한 표본크기가 작은 경우 형상매개변수 전 범위에서 정확도가 떨어지는 것을 확인할 수 있으며, 최우도법의 경우 표본크기가 작은 경우를 제외하고 $-0.25 \leq \beta \leq +0.5$ 범위에서 quantile 산정시 quantile이 약간 과다추정되는 경향이 있는 것을 알 수 있으며, 이는 분산이 과다추정되는 결과를 초래하며 이로 인해 해석해보다 약간씩 큰 값을 나타내는 것으로 판단되었다.

참 고 문 헌

1. Ahmad, M.I., Sinclair, C.D., and Werritty, A. (1988). Log-logistic flood frequency analysis with historical information, Journal of Hydrology, 98:205-224.
2. Bobee, B. (1973). Sample error of T-year events computed by fitting a Pearson type-3 distribution, Water Resources Research, 9(5):1264-1270.
3. Heo, J.H., Salas, J.D., and Kim, K.D. (2001). Estimation of confidence intervals of quantiles for the Weibull distribution, Stochastic Environmental Research and Risk Assessment, 15:284-309.
4. Hosking, J.R.M. and Wallis, J.R. (1997). Regional Frequency Analysis, New York, Cambridge University Press.
5. Institute of Hydrology (1999). Flood Estimation Handbook, UK, Institute of Hydrology.
6. Kite, G.W. (1988). Frequency and risk analysis in hydrology, Colorado, Water Resources Publications.
7. Nash, J.E. and Amorocho, J. (1966). The accuracy of the prediction of floods of high return period, Water Resources Research, 2(2):191-198.
8. Rao, A.R. and Hamed, K.H. (2000). Flood Frequency Analysis, Florida, CRC Press.
9. Yevjevich, V.M. (1964). Handbook of Applied Hydrology, New York, McGraw-Hill.