Bootstrap을 이용한 강우빈도해석에서의 매개변수 추정에 대한 불확실성 해석

Uncertainty Analysis for Parameter Estimation in Rainfall Frequency Analysis using Bootstrap

서영민*·지홍기**·이순탁***

Youngmin Seo·Jee, Hong Kee·Soontak Lee

요 지

Bootstrap 기법은 통계학적 추정치의 정확도 또는 불확실성을 평가하기 위한 컴퓨터 기반 리샘플링 기법으로서 플러그인 원칙을 이용하여 요약통계치의 표준오차 및 신뢰구간을 추정하며, Bootstrap 기법 중 BCa 기법은 다른 Bootstrap 기법들에 비해 적합도 기준면에서 훨씬 우수한 결과를 나타내는 것으로 알려져 있다.

본 논문에서는 강우빈도해석에서 확률분포의 매개변수 추정에 대한 불확실성 고려한 확률강우량의 산정 및 불확실성의 영향을 평가하기 위하여 Bootstrap 기법 중 비매개변수적 BCa 기법에 기반한 불확실성을 고려한 강우빈도해석모델 구축 및 적용을 통해 홍수위험평가 및 수자원 계획 등에 있어서 불확실성 표현 및 처리기법을 제시하였다.

핵심용어 : Bootstrap, 불확실성, 확률강우량, 매개변수추정

1. 서 론

현재 홍수위험관리 실무에서는 홍수위험평가 및 의사결정시 위험도와 불확실성을 충분히 설명하고 있지 않다. 홍수위험평가를 위한 모델선정 및 실행, 매개변수의 선택 또는 홍수량 추정치와 관련된 불확실성의 고려없이 통상 확정론적으로 모델링되고 있으며, 또한 공적토론, 정책결정 및 의사결정은 보통 이러한 확정론적 모델링 결과를 근거로 이루어지고 있는 실정이다.

강우빈도해석에서 통계학적 추정의 정확성 또는 불확실성은 표본 통계치 또는 확률분포의 매개변수에 대한 표본분포(표집분포, sampling distribution)에 기초하며, 표본분포는 모집단으로부터의 많은 무작위 표본으로부터 추정된다. 이러한 통계학적 추정에서의 정확도 또는 불확실성을 평가하기 위한 측도로는 표준오차, 편의, 예측오차 및 신뢰구간 등이 있으며, 이 중 표준오차는 통계학적 정확성 또는 불확실성을 나타내는 가장 일반적인 방법이다. 표준오차에 대한 대부분의 이론적 공식들은 정규이론에 근거한 근사식으로서 전통적인 통계학적 해석에서 매개변수적 가정을 전제로 하는 주된 이유는 그것이 표준오차에 대한 공식을 유도하는데 수학적으로 매우 용이하기 때문이다. 또한 전통적인 표준오차는 표본의 크기가 매우 커짐에 따라 표본평균의 분포가 근사적으

* 정회원·영남대학교 대학원·박사수료·E-mail : elofy@nate.com
** 정회원·영남대학교 건설시스템공학부·교수·E-mail : hkjee@yu.ac.kr
*** 참여회원·영남대학교 석좌교수·공학·이학박사·E-mail : leest@yu.ac.kr

로 정규분포를 따르는 중심극한정리에 근거하기 때문에 대표본에 대해서는 우수한 추정치를 제공할 수 있으나 특히 소표본에 대해서는 그 추정의 정확도가 떨어지기 때문에 항상 중심극한정리가 유용한 것만은 아닌 단점을 가지게 된다(Efron & Tibshirani, 1994).

반면 Bootstrap 기법은 통계학적 추정치의 정확도를 평가하기 위한 컴퓨터 기반 리샘플링 기법(resampling method)으로서 플러그인 원칙(plug-in principle)을 이용하여 요약통계치의 표준오차를 추정하며, 표준오차에 대한 Bootstrap 추정치는 이론적 계산이 많이 요구되지 않기 때문에 추정치에 대한 수학적 복잡성에 상관없이 사용가능한 장점을 가진다. 이러한 Bootstrap 기법에는 Bootstrap-t 기법, 퍼센타일 기법(percentile method), BCa 기법(Bias-Corrected and Accelerated method) 및 ABC 기법(Approximate Bootstrap Confidence intervals) 등이 있으며, 이 중 BCa 기법 및 ABC 기법은 표준, Bootstrap-t 및 퍼센타일 기법보다 적합도 기준면에서 훨씬 우수한 결과를 나타내는 것으로 알려져 있다(Efron & Tibshirani, 1994).

따라서 본 논문에서는 홍수위험평가 및 수자원 계획 수립시 가장 기본이 되는 강우빈도해석에서 확률분포의 매개변수 추정에 대한 불확실성 고려한 확률강우량의 산정 및 불확실성의 영향을 평가하기 위하여 Bootstrap 기법 중 BCa 기법에 기반한 불확실성을 고려한 강우빈도해석모델 구축 및 적용을 통해 홍수위험평가 및 수자원 계획 등에 있어서 불확실성 표현 및 처리기법을 제시하였다.

2. 비매개변수적 BCa 기법(Nonparametric BCa Method)

BCa 기법은 퍼센타일 기법을 개선한 기법으로서 BCa 신뢰구간(BCa interval)의 양 끝점은 Bootstrap 분포의 퍼센타일로 주어지며, 퍼센타일은 가속인자(acceleration) $\hat{a}$와 편의수정인자(bias-correction) $\hat{z}_0$에 의해 결정된다.

BCa 기법에서 포함확률(coverage) 1-2α에 대한 BCa 신뢰구간은 식 (1)과 같다.

$$(\hat{\theta}_{lo},\ \hat{\theta}_{up}) = (\hat{\theta}^{*(\alpha_1)},\ \hat{\theta}^{*(\alpha_2)}) \tag{1}$$

$$\alpha_1 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z^{(\alpha)}}{1-\hat{a}\left(\hat{z}_0 + z^{(\alpha)}\right)}\right), \qquad \alpha_2 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z^{(1-\alpha)}}{1-\hat{a}\left(\hat{z}_0 + z^{(1-\alpha)}\right)}\right) \tag{2}$$

여기서, α는 유의수준, $\Phi(\bullet)$는 누가표준정규분포함수, $z^{(\alpha)}$는 표준정규분포의 100α번째 퍼센타일이고 $\hat{\theta}_{lo}$ 및 $\hat{\theta}_{up}$는 $\hat{\theta}$의 하한 및 상한이다. 만약 식 (2)로부터 $\hat{a}=0$, $\hat{z}_0=0$이면, $\alpha_1=\Phi(z^{(\alpha)})=\alpha$, $\alpha_2=\Phi(z^{(1-\alpha)})=1-\alpha$가 되어 BCa 신뢰구간과 퍼센타일 신뢰구간은 동일한 결과를 나타내게 되며, BCa 기법은 $\hat{a}$와 $\hat{z}_0$의 값에 의해 표준 및 퍼센타일 신뢰구간의 결점을 보완하게 된다.

$\hat{z}_0$는 원추정치 $\hat{\theta}$보다 작은 Bootstrap 추정치 $\hat{\theta}^*$의 비율로서 식 (3)과 같이 나타낼 수 있다.

$$\hat{z}_0 = \Phi^{-1}\left(\frac{\#\{\hat{\theta}^*(b) < \hat{\theta}\}}{B}\right) \tag{3}$$

여기서, $\Phi^{-1}(\bullet)$는 표준누가정규분포의 역함수이고 B는 Bootstrap 표본 발생수이다.

$\hat{a}$는 매개변수의 참값 θ에 대한 추정치 $\hat{\theta}$의 표준오차의 변화율을 나타내며, $\hat{\theta}=s(\boldsymbol{X})$에 대한

Jackknife 값을 이용한다. $\boldsymbol{X}_{(i)}$를 i번째 값 x_i가 제거된 원표본(original sample)이라고 하면 n 세트(n : 원자료수)의 Jackknife 표본을 이용한 추정치는 $\hat{\theta}_{(i)}=s(\boldsymbol{X}_{(i)})$이고 n 세트의 Jackknife 표본에 대한 추정치의 평균을 $\hat{\theta}_{(\cdot)}$라고 하면 $\hat{a}$는 식 (4)와 같이 나타낼 수 있다.

$$\hat{a}=\frac{\sum_{i=1}^{n}\left(\hat{\theta}_{(\cdot)}-\hat{\theta}_{(i)}\right)^3}{6\left\{\sum_{i=1}^{n}\left(\hat{\theta}_{(\cdot)}-\hat{\theta}_{(i)}\right)^2\right\}^{3/2}} \tag{4}$$

3. 모델구축 및 적용

본 연구에서는 강우빈도해석에서 확률분포의 매개변수 추정에 대한 불확실성을 고려한 확률강우량의 산정 및 불확실성의 영향을 평가하기 위하여 비매개변수적 BCa 기법에 기반한 강우빈도해석 모델을 구축하였다. 확률분포의 매개변수 추정기법으로는 L-moment 법(Rao & Hamed, 2000)을 이용하였으며, 적합도 검정을 위해 Chi-Square 및 Kolmogorov-Smirnov 검정을 적용하였다. 모델의 적용성을 평가하기 위하여 위천 유역 내 의성(기상청, 위도 36°21', 경도 128°41', 표고 EL.81.1m) 지점의 1973~2007년(35년) 동안의 시우량 자료를 수집하였으며, 먼저 시우량 자료의 이상치 및 결측치 등을 보정하고 지속시간별 최대강우량을 산정하였다. 다음으로 산정된 지속시간별 최대강우량으로부터 10,000세트의 Bootstrap 표본을 발생시킨 후 각 Bootstrap 표본에 대하여 L-moment 법을 이용하여 각 확률분포의 매개변수를 추정하였다. 그리고 추정된 매개변수로부터 각 확률분포형에 대한 확률강우량을 추정하였으며, 매개변수의 적합성 체크 및 각 확률분포에 대한 적합도 검정을 통해 1차적으로 최적 확률분포형을 필터링하였다. 그 결과 확률분포형의 적합도 순위는 General Logistic 분포(GLO), General Extreme 분포(GEV) 및 Weibull 분포의 순으로 분석되었다. 그리고 1차적으로 선정된 확률분포형들을 대상으로 확률강우량에 대한 Bootstrap 추정치들로부터 표준오차 및 변동계수를 산정하였으며, Bootstrap 추정치의 변동성이 가장 낮은 순으로 확률분포형을 2차적으로 필터링하여 대표확률분포형을 선정하였다. 선정된 각 확률분포형에 대한 지속시간 24시간에서의 재현기간별 Bootstrap 추정치에 대한 요약통계치 및 BCa 신뢰구간은 Table 1~3과 같고 재현기간 증가에 따른 변동계수 및 표준오차의 변화는 Fig. 1과 같이 분석되었다. 또한 재현기간 증가에 따른 BCa 신뢰구간의 변화를 Fig. 2와 같이 도식적으로 비교하였다.

Table 1. Statistics and Estimation of Quantiles and BCa Intervals for GEV (duration=24hr)

Return(yr)	Quantile(mm)	MIN(mm)	MEAN(mm)	MAX(mm)	SE(mm)	CV	α_1	α_2	LB(mm)	UB(mm)
20	174.06	150.0	192.2	238.5	19.14	0.100	0.035	0.945	154.6	222.9
30	189.24	153.7	207.5	259.2	23.61	0.114	0.027	0.934	159.7	246.8
50	209.54	156.0	227.6	285.7	30.38	0.133	0.014	0.908	161.9	274.8
80	229.48	157.6	247.0	323.0	37.88	0.153	0.018	0.918	166.4	305.6
100	239.41	158.2	256.7	343.5	41.92	0.163	0.018	0.917	167.6	321.1
200	272.28	159.7	288.3	415.8	56.68	0.197	0.022	0.926	171.2	372.2
500	320.98	161.0	334.9	535.8	82.23	0.246	0.020	0.922	175.6	445.3

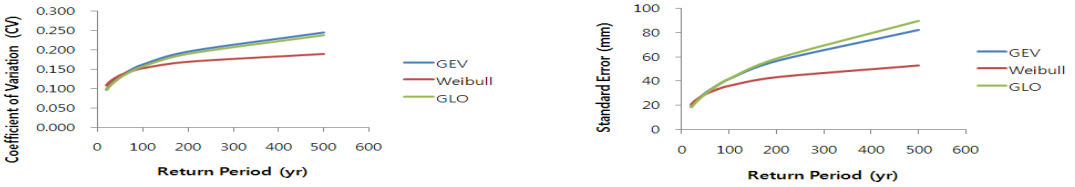
※ Return : Return period, MIN : Minimum qunatile, MAX : Maximum qunatile, MEAN : Mean quantile, SE : Standard Error, CV : Coefficient of variation, LB : Lower Bound, UB : Upper Bound

Table 2. Statistics and Estimation of Quantiles and BCa Intervals for Weibull (duration=24hr)

Return(yr)	Quantile(mm)	MIN(mm)	MEAN(mm)	MAX(mm)	SE(mm)	CV	α_1	α_2	LB(mm)	UB(mm)
20	175.38	143.1	191.3	231.2	20.99	0.110	0.038	0.949	153.1	224.7
30	187.80	147.2	203.6	248.8	24.71	0.121	0.034	0.944	157.3	245.1
50	203.08	151.8	218.5	275.6	29.59	0.135	0.031	0.940	162.2	265.4
80	216.81	155.6	231.8	300.8	34.23	0.148	0.033	0.943	166.4	288.1
100	223.24	157.3	238.0	312.8	36.48	0.153	0.030	0.939	168.3	297.6
200	242.85	162.2	256.9	350.2	43.60	0.170	0.030	0.939	173.9	324.5
500	268.08	168.0	280.9	400.1	53.29	0.190	0.027	0.935	180.4	360.5

Table 3. Statistics and Estimation of Quantiles and BCa Intervals for GLO (duration=24hr)

Return(yr)	Quantile(mm)	MIN(mm)	MEAN(mm)	MAX(mm)	SE(mm)	CV	α_1	α_2	LB(mm)	UB(mm)
20	187.80	140.5	189.9	227.2	18.53	0.098	0.019	0.919	149.2	214.1
30	204.11	144.6	206.6	250.4	22.91	0.111	0.022	0.925	157.1	239.2
50	226.60	149.4	229.8	282.5	29.77	0.130	0.021	0.924	166.6	267.6
80	249.49	153.6	253.7	324.2	37.68	0.149	0.014	0.908	172.0	300.5
100	261.18	155.4	266.0	346.7	42.05	0.158	0.014	0.908	175.6	320.7
200	301.29	161.0	308.7	431.2	58.74	0.190	0.014	0.908	186.1	387.9
500	364.56	167.7	377.8	580.8	89.87	0.238	0.021	0.924	198.5	504.9



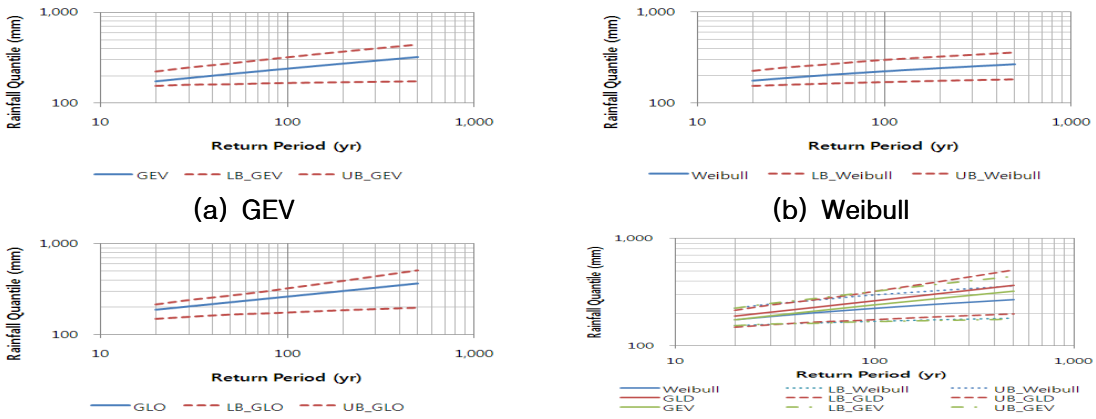
(a) Coefficient of Variation (CV) (b) Standard Error (SE)

Fig. 3. Coefficienct of Variation(CV) and Standard Error(SE)

먼저 Table 1~3 및 Fig. 3으로부터 각 확률분포형의 Bootstrap 추정치에 대한 분석결과를 살펴보면 재현기간이 증가함에 따라 표준오차 및 변동계수가 증가하는 경향을 나타내고 있음을 알 수 있다. 특히 GLO 및 GEV 분포의 경우 표준오차는 각각 18.53~89.87, 19.14~82.23, 변동계수는 각각 0.098~0.238, 0.100~0.246의 변동성을 나타내고 있고 Weibull 분포의 경우 표준오차 및 변동계수는 각각 20.99~53.29, 0.110~0.190의 변동성을 나타내고 있어 Weibull 분포로부터의 Bootstrap 추정치가 세 확률분포형 중 가장 낮은 변동성을 나타냄을 확인할 수 있으며, Bootstrap 추정치의 변동성 측면에서는 가장 낮은 변동성을 나타내는 Weibull 분포가 가장 적합한 확률분포형이라고 할 수 있다. 그러나 매개변수의 적합성 및 확률분포형의 적합도 검정 결과는 GLO, GEV 및 Weibull 분포의 순으로 적합도 순위가 분석되었기 때문에 적합도 검정 및 변동성 분석결과를 종합적으로 판단하여 대표확률분포형을 선정해야 할 것으로 판단된다. 또한 변동성 분석결과에서 각 확률분포형간 변동성의 차이는 재현기간이 증가할수록(특히 100년 이상) 그 차이가 더욱 커지는 것을 확인할 수 있으며, Weibull 분포의 경우 재현기간의 증가에 따른 변동성의 증가율이 다른 확률분포형에 비해 낮게 나타나는 결과로부터 재현기간 100년 이상의 확률강우량 추정시 확률분포형 선정에 따른 확률강우량의 불확실성에 대한 영향이 매우 큼을 확인할 수 있다.

한편, Table 4~6 및 Fig. 4로부터 각 확률분포형별 BCa 신뢰구간 산정결과를 살펴보면 재현기간이 증가함에 따라 BCa 신뢰구간의 크기가 증가하는 경향을 나타내고 있음을 알 수 있다. 지속시간 24시간, 재현기간 100년을 기준으로 각 확률분포형별 BCa 신뢰구간 또는 불확실성 범위를 살펴보면 GLO, GEV 및 Weibull 분포가 각각 145.09, 153.48 및 129.28로 분석되었으며, Weibull 분포의 확률강우량에 대한 불확실성 범위가 가장 작고 GLO 분포의 불확실성 범위가 가장 크게 나타남을 확인할 수 있다. 확률강우량의 불확실성 범위는 그 크기가 작을수록 신뢰성 있는 추정결과가 될 수 있고 이러한 결과를 향후 강우-유출분석 및 홍수위분석 등에 적용하여 불확실성을 전파시킬 경우 그 분석결과에 따른 불확실성의 범위는 더욱 증가할 것으로 예상되기 때문에 확률강

우량에 대한 불확실성 범위 또는 BCa 신뢰구간의 크기가 상대적으로 작은 확률분포형이 선정되어야 할 것이며, 본 연구에서는 확률분포형의 매개변수 추정기법으로 L-moment 법만을 적용하였으나 다른 추정기법과의 비교를 통해 확률강우량에 대한 불확실성 범위가 상대적으로 작게 추정되는 매개변수 추정기법을 선정해야 할 것으로 판단된다.



(a) GEV (b) Weibull

(c) GLO (d) GEV, Weibull and GLO

Fig. 4. Estimation of Quantile and BCa Interval

4. 결 과

본 연구에서는 강우빈도해석에서 확률분포의 매개변수 추정에 대한 불확실성 고려한 확률강우량의 산정 및 불확실성의 영향을 평가하기 위하여 Bootstrap 기법에 기반한 불확실성을 고려한 강우빈도해석모델 구축 및 적용을 실시하였으며, 다음과 같은 결론을 얻었다.

1) 강우빈도해석에서 확률분포의 매개변수 추정에 대한 불확실성을 분석하기 위해 Bootstrap 기법에 기반한 불확실성 평가모델을 구축하였으며, 이를 확률분포의 매개변수 추정기법인 L-moment 법과 결합하여 매개변수의 불확실성을 고려한 확률강우량을 산정모델을 제시하였다.
2) 확률분포형의 매개변수에 대한 불확실성을 고려한 Bootstrap 기법의 적용을 통해 매개변수의 적합성 및 확률분포형의 적합도 검정, Bootstrap 추정치의 표준오차 및 변동계수를 종합적으로 고려하여 대표확률분포형을 선정하는 기법을 제시하였다.
3) 재현기간별 확률강우량의 BCa 신뢰구간의 추정을 통해 매개변수의 불확실성에 따른 확률강우량의 불확실성 범위를 정량적으로 제시하는 방안을 제시하고 그 적용을 통해 확률강우량 추정에 대한 불확실성 해석에서 BCa 기법의 적용성을 확인할 수 있었다.

참 고 문 헌

1. Brandley Efron, Robert J. Tibshirani(1994), An Introduction to the Bootstrap, Chapman & Hall/CRC.
2. A. Ramachandra Rao, Khaled H. Hamed(2000), Flood Frequency Analysis, CRC Press LLC.
3. Thomas J. DiCiccio, Brandley Efron(1996), Bootstrap Confidence Intervals, Statistical Science,

Vol. 11, No. 3, pp. 189-228.
4. Brandley Efron(2003), Second Thoughts on the Bootstrap, Statistical Science, Vol. 18, No. 2, pp. 135-140.
5. Peter K. Dunn(2001), Bootstrap Confidence Intervals for Predicted Rainfall Quantiles, International Journal of Climatology, 21. pp. 89-94.