유전자 알고리즘을 이용한 Gumbel 분포의 도시위치공식 유도

Derivation of Plotting Position Formula Using Genetic Algorithm for Gumbel Distribution

김수영*, 신홍준, 고연우***, 허준행****

Sooyoung Kim, Hongjoon Shin, YounWoo Kho, Jun-Haeng Heo

요 지

확률도시위치는 주로 도시적 해석을 통한 연최대홍수량 또는 연최대강우량의 초과확률의 추정치 산정에 사용되며 빈도해석을 통해 선정된 적정 확률분포형과 표본자료의 개략적인 적합도를 도시적으로 파악할 수 있도록 해주기 때문에 오래 전부터 널리 이용되어 왔다. 본 연구에서는 Gumbel 분포에 적합한 도시위치공식을 새롭게 추정하기 위해 Gumbel 분포의 order statistic과 확률가중모멘트를 이용하여 다양한 표본크기에 대한 도시위치공식의 기본식을 유도하였고, 최적화 기법 중 하나인 유전자 알고리즘을 이용하여 유도된 도시위치공식의 매개변수를 추정하였다. 또한 본 연구에서 추정된 도시위치공식과 기존에 널리 사용되고 있는 도시위치공식의 정확도를 비교하기 위해 reduced variate 간의 오차를 계산하여 비교·검토하였다. 그 결과, 금회 추정된 도시위치공식은 높은 순위에서는 기존의 도시위치공식에 비해 더 정확도가 높은 것으로 나타났고, 표본크기에 대한 순위를 모두 고려할 경우에는 기존의 도시위치공식에 비해 정확도가 높은 것으로 나타나 Gumbel 분포에 대해서 높은 정확도를 보이는 것으로 나타났다.

핵심용어 : 도시위치공식, 유전자 알고리즘, 확률가중모멘트, Gumbel 분포, Order statistic

1. 서론

확률도시위치(probability plotting position)는 연최대홍수량 자료나 연최대강우량 자료의 도시적 해석에 이용되는데, 주로 도시적 해석을 통한 연최대홍수량 또는 연최대강우량의 초과확률의 추정에 사용된다. 또한 확률도시위치는 빈도해석을 통해 선정된 적정 확률분포형과 표본자료의 개략적인 적합도를 도시적으로 파악할 수 있도록 해주며, 주어진 확률분포형에 대한 비모수적 평균을 추정할 수 있도록 하는 역할을 수행하여 오래 전부터 수문학뿐만 아니라 수자원 분야에서 널리 이용되어 왔다. 현재 일반적으로 사용되고 있는 도시위치공식은 1980년대 이전에 연구된 것으로, Cunnane(1978)은 비편의(unbiased) 도시위치 $E(y_i)$를 reduced variate의 모집단으로부터의 표본 중 i번째 order statistic의 평균으로 정의하고 도시위치공식의 일반 형태를 정의한 바 있다. 또한 Arnell 등(1986)은 GEV 분포에 대한 order statistic과 확률가중모멘트(PWM)를 이용한 도시위치공식을 유도하였고, 같은 이론적 근거를 이용하여 In-na와 Nguyen(1989), De(2000) 등이 GEV 분포와 Gumbel 분포에 대한 도시위치공식을 유도한 바 있다.

본 연구에서는 우리나라의 강우빈도해석에 널리 적용되고 있는 Gumbel 분포에 적합한 도시위치공식을 새롭게 추정하고, 새롭게 추정된 도시위치공식과 Gumbel 분포에 적용되어오던 기존의 도시위치공식과의 정확도 비교를 통해 새롭게 추정된 도시위치공식의 적용성을 평가하였다. 이를 위해 Gumbel 분포의 order

* 정회원, 연세대학교 대학원 토목공학과 박사과정, E-mail : sykim79@yonsei.ac.kr
** 정회원, 연세대학교 대학원 토목공학과 박사과정, E-mail : sinong@yonsei.ac..kr
*** 정회원, 연세대학교 대학원 토목공학과 박사과정, E-mail : ywkho@gsconst.co.kr
**** 정회원, 연세대학교 사회환경시스템공학부 토목환경공학과 교수, E-mail : jhheo@yonsei.ac.kr

statistic과 확률가중모멘트를 이용하여 다양한 표본크기에 대한 도시위치공식의 기본식을 유도하였고, 최적화 기법 중 하나인 유전자 알고리즘(genetic algorithm)을 이용하여 유도된 도시위치공식의 매개변수를 추정하였다. 또한 유전자 알고리즘을 이용하여 추정된 도시위치공식과 기존에 널리 사용되고 있는 도시위치공식의 정확도를 비교하기 위해 reduced variate 간의 평균제곱근오차와 절대오차를 산정하여 정확도를 비교하고, 정확도 분석 결과를 토대로 Gumbel 분포에 대해 새롭게 추정된 도시위치공식의 정확성을 비교·검토 하였다.

2. Gumbel 분포에 대한 도시위치공식의 유도

2.1 Order statistic과 PWM을 이용한 도시위치공식의 유도

Arnell 등(1986)에 따르면 오름차순으로 정열된 표본크기 n의 관측자료 $y_1 \le y_2 \le y_3 \le \cdots \le y_n$의 m번째 ordered statistic의 확률밀도함수(PDF) $g(y_m)$은 다음과 같다.

$$g(y_m)=\frac{n!}{(m-1)!(n-m)!}\{F(y_m)\}^{m-1}\{1-F(y_m)\}^{n-m}f(y_m) \tag{1}$$

여기서, F와 f는 표본으로부터 유도된 누가분포함수와 확률밀도함수를 나타낸다. m번째 관측자료 y_m의 평균은 $E(y_m)$과 같으며 식 (2)와 같이 나타낼 수 있다.

$$E(y_m)=\frac{n!}{(m-1)!(n-m)!}\int_{-\infty}^{\infty} y_m\{F(y_m)\}^{m-1}\{1-F(y_m)\}^{n-m}f(y_m)\,dy_m \tag{2}$$

만약 m번째 관측자료 y_m에 누가분포함수 $F_m=F(y_m)$라 하면, $y_m=F^{-1}(F_m)$이고 $dF_m=f(y_m)dy_m$이다. $F(\cdot)$이 누가분포함수이므로 식 (2)는 다음과 같이 나타낼 수 있다.

$$E(y_m)=\frac{n!}{(m-1)!(n-m)!}\int_0^1 F^{-1}(F_m)F_m^{m-1}\{1-F_m\}^{n-m}dF_m \tag{3}$$

또한 Gumbel 분포의 누가분포함수는 다음과 같다.

$$F(x)=\exp[-\exp\{-(x-x_o)/\alpha\}] \tag{4}$$

여기에서 x_o는 위치 매개변수이고, α는 규모 매개변수이다. Gumbel 분포의 m번째 reduced variate는 식 (5)와 같고, 이를 식 (3)에 대입하여 정리하면 식 (6)과 같다.

$$y_m=F^{-1}(F_m)=-\ln\{-\ln(F_m)\} \tag{5}$$

$$E(y_m)=\frac{n!}{(m-1)!(n-m)!}\int_0^1 -\ln\{-\ln(F_m)\}F_m^{m-1}\{1-F_m\}^{n-m}dF_m \tag{6}$$

Gumbel 분포의 도시위치공식을 유도하기 위해 식 (6)에 확률가중모멘트를 적용하게 되는데, Greenwood 등(1979)이 제시한 확률가중모멘트는 식 (7)과 같이 나타낼 수 있다.

$$M_{p,r,s}=\int_0^1 x^p F^r(1-F)^s dF \tag{7}$$

$n=r+s+1$ 및 $p=1$에 대해 식 (6) 및 (7)을 결합하여 m번째 reduced variate y_m에 대한 평균을 나타내면 식(8)과 같고, 이를 다시 정리하면 식 (9)로 나타낼 수 있다.

$$E(y_m)=\frac{n!}{(m-1)!(n-m)!}\sum_{s=0}^{n-m}\binom{n-m}{s}(-1)^s M_{1,m+s-1,0} \tag{8}$$

$$E(y_m)=\frac{n!}{(m-1)!(n-m)!}\sum_{s=0}^{n-m}\binom{n-m}{s}(-1)^s[\gamma+\ln(m+s)](m+s)^{-1} \tag{9}$$

여기서, γ=0.5772로 Euler의 상수를 나타낸다.

식 (9)를 가장 큰 ordered statistic($m=n$)에 대해 나타내면 식 (10)과 같으며, Gumbel 분포에 대한 도시위치공식의 매개변수 추정 과정에 이를 이용하게 된다.

$$E(y_n)=\gamma+\ln n \tag{10}$$

Gumbel 분포에 대한 도시위치 P_n은 식 (9)를 Gumbel 분포의 누가분포함수에 대입하여 구할 수 있으며, 이는 식 (11)과 같다.

$$P_n=F\{E(y_n)\} \tag{11}$$

Gumbel 분포에 적합한 도시위치공식의 유도를 위해 본 연구에서 적용한 도시위치공식의 일반식은 식 (12)와 같이 나타낼 수 있다.

$$P_m=\frac{m+a}{n+b} \tag{12}$$

여기서, a와 b는 도시위치공식의 매개변수이다.

2.2 유전자 알고리즘을 이용한 도시위치공식의 매개변수 추정

본 연구에서는 Gumbel 분포에 대한 도시위치공식의 매개변수를 추정하기 위해 유전자 알고리즘을 적용하였다. 유전자 알고리즘은 자연 선택(natural selection)과 유전자의 응용을 기반으로 하는 기법으로(Goldberg, 1989), 1960년대에 Holland(1975)와 그의 동료들에 의해서 창안되었으며 다윈이 제안한 '적자생존(survival of the fittest)'을 컴퓨터 기법화한 최적화 방법이다. 유전자 알고리즘은 수 십 년간 발전되어 오면서 수 없이 많은 형태로 변형되었지만, 1) 해(chromosome)의 집합인 군(population)을 기본단위로 하고, 2) 교배(crossover), 돌연변이(mutation), 전치(inversion) 등의 유전자연산자(genetic operator)를 사용하며, 3) 룰렛 휠 선택법(roulette wheel selection), 토너먼트 선택법(tournament selection) 등의 선택법을 통해서 다음 세대에 더 좋은 해를 넘겨주는 방법을 통칭한다. 유전자 알고리즘은 초기에 무작위로 초기 모집단을 형성하고 이들을 부모세대로 하여 선택, 교배, 돌연변이 등의 연산과정을 거쳐 부모세대보다 진화한 새로운 자식세대를 생성하게 되며, 적합도를 평가하여 적합한 개체를 생성시킨다.

본 연구에서는 유전자 알고리즘의 한 종류인 Real-coded Genetic Algorithm(Deb and Beyer, 2001; Beyer and Deb, 2001)을 적용하였고, 유전자 알고리즘의 목적함수(objective function)는 식 (12)의 좌변과 우변간의 평균제곱근오차(Root Mean Square Error)로 설정하였다. 유전자 알고리즘의 초기조건 중 모집단(population) 수는 5,000개, 전체 세대(generation)수는 1,000세대, 교배(crossover) 확률은 0.8, 돌연변이(mutation) 확률은 0.01로 설정하였고, 총 1,000회에 걸쳐 수행하였다. Seed number는 0.1부터 1.0까지 0.1씩 증가시키면서 총 10개의 seed number에 대해 수행하여 seed number에 따른 도시위치공식의 매개변수 추정치의 경향성을 살펴보았다. 입력자료로 사용된 표본크기는 n=2(1)30(5)100(10)200(50)500으로 표본크기 $2\leq n\leq 30$의 범위에서는 각각의 order별 reduced variate까지 고려할 수 있도록 하였으며, $n>30$의 범위에서는 최대 표본크기 n에 대한 reduced variate를 식 (10)을 이용하여 산정하여 고려하였다.

2.3 Gumbel 분포에 대한 도시위치공식의 추정

Gumbel 분포에 대한 도시위치공식의 매개변수를 추정하기 위해 유전자 알고리즘을 적용한 결과, 최소 평균제곱근오차 0.02466에 대해 적합된 a=-0.30, b=0.22가 추정되었다. Seed number에 따른 매개변수 추정치를 살펴보면 모든 seed number에 대해 매개변수가 동일한 결과를 보여, 도시위치공식의 매개변수 추정에 있어 seed number에 따른 영향이 크지 않은 것으로 나타났다. 최종적으로 유전자 알고리즘을 이용하여 추정된 Gumbel 분포에 대한 도시위치공식은 다음과 같다.

$$P_m=\frac{m-0.30}{n+0.22} \tag{13}$$

3. 도시위치공식의 비교 및 검토

본 연구에서 새롭게 추정된 도시위치공식의 정확도를 분석하기 위해 Gumbel 분포에 대해 기존에 주로 사

용되거나 개발된 Gringorten(1963), Cunnane(1978), De(2000)에 의한 도시위치공식을 금회 추정된 도시위치공식과 비교·검토하였다. NERC(1975)에 수록되어 있는 Gumbel 분포에 의한 표본크기 50까지의 reduced variate와 금회 추정된 도시위치공식 및 기존의 도시위치공식을 이용하여 산정되는 reduced variate 간의 평균제곱근오차(Root Mean Square Error, RMSE)와 절대오차(Absolute Error, AE)를 표본크기 n에 대해 order별로 계산하여 나타내면 표 1과 같다. 표 1에 의하면 순위가 작을 때, 금회 추정된 도시위치공식에 의해 추정된 평균제곱근오차와 절대오차는 Gringorten(1963)과 Cunnane(1978)의 도시위치공식에 의한 평균제곱근오차와 절대오차에 비해 상대적으로 큰 것으로 나타났으나, 순위가 커질수록 금회 추정된 도시위치공식에 의해 추정된 평균제곱근오차와 절대오차는 다른 도시위치공식에 의한 값보다 상대적으로 작은 것으로 나타났다. 금회 추정된 도시위치공식은 낮은 순위에 대해서는 기존의 도시위치공식에 비해 오차가 상대적으로 크게 나타나긴 하나, 높은 순위에서는 반대로 오차가 더 작은 것으로 나타남에 따라 큰 재현기간에 대해 보다 정확한 도시위치를 추정하게 하는 것으로 판단된다.

표본크기별 순위를 모두 고려할 수 있도록 표본크기에 대해 평균한 평균제곱근오차와 절대오차를 정리하면 표 2와 같다. 표 2에 의하면, 금회 추정된 도시위치공식의 평균제곱근오차는 다른 도시위치공식의 의한 값에 비해 작은 것으로 나타났으며, 절대오차의 경우도 Cunnane(1978), De(2000)보다 작은 것으로 나타났다. Gringorten(1963)과 비교하였을 때는 상대적으로 큰 오차를 보이는 것으로 분석되었으나, 두 식 간의 차이는 소수점 셋째짜리로 그 차이가 미미한 것으로 나타났다. 결과적으로, 금회 추정된 도시위치공식이 기존의 도시위치공식에 비해 상대적으로 정확하게 Gumbel 분포의 도시위치를 나타내고 있는 것으로 분석되었다.

표 1. 도시위치공식에 따른 순위별 reduced variate간의 오차

순위	n=10								n=30							
	금회 추정		Gringorten (1963)		Cunnane (1978)		De(2000)		금회 추정		Gringorten (1963)		Cunnane (1978)		De(2000)	
	RMSE	AE	RMSE	AE	RMSE	AE	RMSE	AE	RMSE	AE	RMSE	AE	RMSE	AE	RMSE	AE
1	0.854	0.924	0.793	0.891	0.810	0.900	0.860	0.928	1.269	1.126	1.214	1.102	1.230	1.109	1.276	1.129
2	0.474	0.688	0.449	0.670	0.455	0.674	0.476	0.690	0.928	0.963	0.904	0.951	0.910	0.954	0.930	0.965
3	0.268	0.518	0.254	0.504	0.257	0.506	0.269	0.519	0.756	0.869	0.740	0.860	0.744	0.863	0.757	0.870
4	0.133	0.365	0.125	0.353	0.126	0.355	0.133	0.365	0.608	0.780	0.597	0.773	0.600	0.774	0.609	0.781
⋮				⋮								⋮				
8	0.129	0.360	0.139	0.373	0.142	0.376	0.133	0.365	0.288	0.536	0.283	0.532	0.283	0.532	0.288	0.537
9	0.455	0.675	0.478	0.692	0.490	0.700	0.468	0.684	0.237	0.487	0.233	0.483	0.233	0.483	0.237	0.487
10	1.527	1.236	1.621	1.273	1.691	1.300	1.603	1.266	0.191	0.437	0.187	0.432	0.187	0.433	0.191	0.437
⋮				⋮								⋮				
26	-	-	-	-	-	-	-	-	0.445	0.667	0.453	0.673	0.457	0.676	0.450	0.671
27	-	-	-	-	-	-	-	-	0.673	0.821	0.685	0.827	0.691	0.831	0.681	0.825
28	-	-	-	-	-	-	-	-	1.036	1.018	1.053	1.026	1.065	1.032	1.048	1.024
29	-	-	-	-	-	-	-	-	1.717	1.310	1.752	1.324	1.779	1.334	1.746	1.321
30	-	-	-	-	-	-	-	-	3.449	1.857	3.577	1.891	3.687	1.920	3.568	1.889

표 2. 도시위치공식별 Reduced variate간의 오차

표본크기	금회 추정		Gringorten(1963)		Cunnane(1978)		De(2000)	
	RMSE	AE	RMSE	AE	RMSE	AE	RMSE	AE
5	0.554	0.476	0.554	0.471	0.565	0.480	0.564	0.483
10	0.625	0.517	0.626	0.515	0.635	0.520	0.633	0.522
15	0.654	0.532	0.655	0.530	0.662	0.535	0.660	0.535
20	0.668	0.539	0.670	0.538	0.676	0.542	0.674	0.542
25	0.679	0.543	0.681	0.543	0.686	0.546	0.684	0.546
30	0.684	0.546	0.686	0.546	0.691	0.548	0.689	0.549
35	0.689	0.549	0.691	0.548	0.695	0.551	0.693	0.551
40	0.694	0.552	0.695	0.551	0.700	0.554	0.698	0.554
45	0.697	0.553	0.698	0.552	0.702	0.554	0.701	0.554
50	0.699	0.553	0.700	0.552	0.704	0.554	0.702	0.554

4. 결론

본 연구에서는 Gumbel 분포에 적합한 도시위치공식을 새롭게 추정하기 위해 Gumbel 분포의 order statistic과 확률가중모멘트를 이용하여 다양한 표본크기에 대한 도시위치공식의 기본식을 유도하였고, 도시위치공식의 매개변수를 추정하기 위해 유전자 알고리즘을 이용하였다. 또한 금회 추정된 도시위치공식의 정확도를 평가하기 위해 기존에 이용되고 있는 도시위치공식과 금회 추정된 도시위치공식에 의한 reduced variate 간의 오차를 산정하여 정확도를 비교하였다. 본 연구의 결과는 다음과 같다.

(1) Gumbel 분포에 대한 도시위치공식은 $P_m=(m-0.30)/(n+0.22)$와 같이 추정되었다.

(2) 금회 추정된 도시위치공식은 상대적으로 낮은 순위에서는 기존의 도시위치공식에 비해 정확도가 낮은 것으로 나타났으나, 상대적으로 높은 순위에서는 기존의 도시위치공식에 비해 더 정확도가 높은 것으로 나타났다.

(3) 표본크기에 대한 순위를 모두 고려한 결과, 금회 추정된 도시위치공식이 기존의 도시위치공식에 비해 정확도가 높은 것으로 나타났다.

감사의 글

본 연구는 국토해양부가 출연하고 한국건설교통기술평가원에서 위탁시행 한 2003년도 건설기술혁신사업(03산학연C01-01)에 의한 도시홍수재해관리기술연구사업단의 연구성과입니다.

참고문헌

1. Arnell, N. W., Beran, M., and Hosking, J. R. M. (1986). "Unbiased Plotting Positions for the General Extreme Value Distribution", Journal of Hydrology, Vol.86, pp.59-69.
2. Beyer, H.-G. and Deb, K. (2001). "On self-adaptive features in real-parameter evolutionary algorithms", IEEE Transactions on Evolutionary Computation, Vol.5, No.3, pp.250-270.
3. Cunnane, C. (1978). "Unbiased plotting positions - A review", Journal of Hydrology, Vol. 37, No. 3/4, pp. 205-222.
4. De, M. (2000). "A New Unbiased Plotting Position Formula for Gumbel Distribution", Stochastic Environmental Research and Risk Assessment, Vol.14, pp.1-7.
5. Deb, K. and Beyer, H.-G. (2001). "Self-adaptive genetic algorithms with simulated binary crossover", Evolutionary Computation Journal, Vol.9, No.2, pp.197-221.
6. Goldberg, D. E. (1989). Genetic algorithms in search, optimization & machine learning. Addison Wesley, Massachusetts.

7. Greenwood, J. A., Landwehr, J. M., Matalas, N. C., and Wallis, J. R. (1979). "Probability Weighted Moments : Definition and Relation to Parameters of Several Distributions Expressible in Inverse Form", Water Resources Research, Vol.15, No.5, pp.1049-1054.
8. Gringorten, I. I. (1963). "A plotting rule for extreme probability paper", Journal of Geophysical Research, Vol.68, No.3, pp.813-814.
9. In-na, N. and Nguyen, V-T-V. (1989). "An Unbiased Plotting Position Formula for the Generalized Extreme Value Distribution", Journal of Hydrology, Vol.106, pp.193-209.
10. Holland, J. H. (1975). Adaptation in natural and artificial systems. University of Michigan Press.
11. Natural Environment Research Council (1975). Flood Studies Report, Vol.1, NERC, London.