GEV 분포와 역사 자료 이용 알고리즘 EMA의 접목

Incorporation of Historical Data into GEV Distribution with EMA

성장현*, 김영오**

Jang Hyun Sung, Young-Oh Kim

요 지

재현기간이 수백년 이상인 이상홍수의 초과확률을 추정하기 위해서는 재현기간 이상의 홍수자료를 이용해 내삽(interpolation)을 해야 하지만 현재 우리나라의 체계적(systematic) 관측자료 기간은 이에 훨씬 미치지 못한다. 따라서, 역사 자료(historical data)를 이용해 자료 길이를 확장하는 방법, 홍수자료에 비해 비교적 긴 강우자료와 유출 모형에 의한 합성자료를 이용하는 방법 등이 사용되어 왔다. 본 연구에서는 역사 자료와 체계적 관측자료를 효율적으로 결합할 수 있는 EMA(Expected Moment Algorithm) 기법을 연구하였다.

EMA는 Cohn 등(1997)에 의해 제안된 방법으로 미국의 공식 분포인 LP3(Log-Pearson type 3) 분포를 대상으로 반복 계산을 통해 매개변수를 추정하는 기법으로서 본 연구에서는 LP3 분포 대신에 최근 국내 홍수빈도해석 시 많이 쓰이고 있는 GEV(Generalized Extreme Value) 분포를 대상으로 EMA 절차를 이론적으로 유도하였다.

핵심용어: 이상홍수 초과확률, EMA, 체계적 관측자료, 역사 자료, GEV

1. 서 론

우리나라는 체계적인 홍수자료의 수집이 시작된 지 불과 수 십년 밖에 안되었기 때문에 이러한 자료로 수 백년 이상의 재현기간을 갖는 이상홍수를 산정하기란 무리이다. 자료 부족의 문제는 비단 우리 뿐만 아니라 다른 여러 나라에서도 겪고 있고 이를 해결하기 위해 수많은 연구가 진행되어 왔다. 대표적으로 지역빈도해석(regional frequency analysis)은 유사한 통계 특성을 가지고 있는 각 지점들을 공간적으로 모아서 시간적 제약을 보완하는 방법이다. 하지만 지역빈도해석을 이용하더라도 수 백년 이상의 홍수량을 산정하는데는 많은 불확실성이 따르게 된다. 이미 선진국에서는 이런 지역빈도해석과 더불어 역사 자료의 이용에 열중하여 왔다. 체계적인 관측 이전에도 규칙적이지는 않으나 홍수나 강우는 다양한 형태로 기록되어 왔고 이러한 자료가 비교적 정량적이면 충분히 체계적인 관측자료와 결합이 가능할 수 있다. 다행히 우리나라도 몇 종류의 역사 자료를 가지고 있다. 특히, 서울 지점의 역사 강우자료는 당시 독보적인 관측기구인 측우기로 1700년대 후반부터 관측되어 온 바 양과 질을 모두 갖춘 우수한 자료라 할 수 있겠다. 본 연구에서는 이러한 역사 강우자료를 이용해 이상홍수의 초과확률을 산정할 수 있는 방법론 중 하나인 EMA 기법을 GEV 분포에 맞게 유도해 보았다.

2. 국내 역사자료 현황

최근 들어 역사 홍수자료에 대한 관심이 고조되고 있다. 김현준 (2000)은 조선시대 홍수 사상의 조사를 시작으로 역사 강우, 홍수자료에 대한 데이터 베이스를 구축하기 시작하였고 2007년에는 조선왕조실록과 증보문헌비고에 기록된 자료를 대상으로 통계분석을 실시한 바 있다. 역사 홍수자료 구축 노력에도 불구하고 아쉽게도 대부분의 역사 홍수자료는 정량적이기보다는 정성적이다. 예를 들어 “큰 물이 왔다” 등으로 홍수가 기록되어 있어 발생한 홍수의 정량적인 양을 추정하기란 어려운 일이다. 이에 조한범 등(2007)은 정량화를

* 서울대학교 건설환경공학부 박사과정 · E-mail: kon26@snu.ac.kr

** 서울대학교 건설환경공학부 부교수 · E-mail: yokim05@snu.ac.kr

위한 노력으로 기록 항목을 정해서 홍수를 구분한 바 있다. 역사 자료 중에서 정량적인 값으로 표현되어 있는 대표적인 자료는 측우기 관측자료이다.

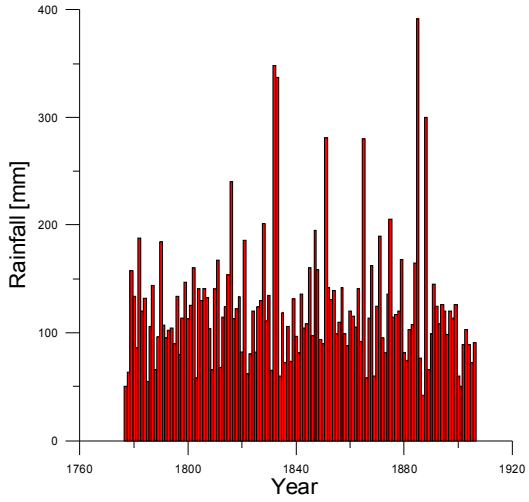

그림 1. 측우기로 관측된 연최대 일강우량 (서울: 1777 ~ 1907년)

측우기는 조선 세종 이후부터 말기에 이르기까지 강우량을 측정하기 위하여 쓰인 기구로 1441년 세종 23년 8월에 호조가 측우기를 설치할 것을 건의하여, 다음해 5월에는 측우에 관한 제도를 새로 제정하고 측우기를 만들어 서울과 각 도의 군현에 설치되었다. 측우기는 안지름은 7인치(14.7cm), 높이 약 1.5척의 원통으로 되어 있고, 측우기에 괸 물의 깊이는 자[尺]로 측정하여 현대적인 우량계가 도입되기 이전인 1907년까지 정량적인 강우량 관측에 사용되었다(심태현과 임규호, 2007). 이러한 측우기 자료에 대한 관심과 연구도 꾸준하여 특히 전종갑은 1997년에 승정원 일기와 일성록을 이용해 측우기 관측 강우량 자료집을 작성한 바 있다. 그림 1은 이렇게 복원된 서울지점의 1777년부터 1907년까지 130년간의 시계열도이다.

2. 기존 연구: LP3-EMA

EMA는 Cohn 등(1997)에 의해 제안된 방법으로 미국의 공식 분포인 LP3 분포를 대상으로 반복 계산을 통해 매개변수를 추정하는 절차이다. EMA는 다양한 종류의 자료(체계적인 자료, 역사적 자료, 고수문 자료)의 결합에 유용하다. 미국의 공식 분포인 LP3 분포에 적용된 EMA의 일반적인 절차는 그림 2와 같다.

(1) 알고 있는 값들로부터 표본 모멘트(평균, 표준편차, 왜도계수)를 구한다.

⬇

(2) (1)에서 구한 모멘트를 이용해 LP3 분포의 위치, 규모, 형상 매개변수를 계산한다.

⬇

(3) 알고 있는 값과 그렇지 않은 값을 포함한 완전한 자료로부터 표본 모멘트를 구한다.

⬇

(4) (3)에서 구한 모멘트를 이용해 LP3 분포의 매개변수를 계산한다.

⬇

(5) (2)에서 계산한 매개변수와 (4)에서 계산한 매개변수가 수렴될 때까지 (3), (4) 과정을 반복한다.

그림 2. EMA 기법의 일반적인 절차

각 단계를 좀 더 자세히 살펴보면 우선, 1 단계에서는 알고 있는 값들로부터 표본 모멘트($\hat{\mu}, \hat{\sigma}^2, \hat{\gamma}$)를 구한다. 2 단계에서는 1 단계에서 구한 표본 모멘트로부터 LP3 분포의 매개변수($\hat{\tau}, \hat{\alpha}, \hat{\beta}$) 초기값을 계산하게 되며, 이때의 매개변수 추정식은 아래 식 1 ~ 3과 같다.

$$\hat{\alpha} = \frac{4}{\hat{\gamma}^2} \qquad \text{<식 1>}$$

$$\hat{\beta} = sign(\hat{\gamma})(\frac{\hat{\sigma}^2}{\hat{\alpha}})^2 \qquad \text{<식 2>}$$

$$\hat{\tau} = \hat{\mu} - \hat{\alpha}\hat{\beta} \qquad \text{<식 3>}$$

3 단계에서는 완전한 자료(complete data set)으로부터 새로운 표본 모멘트를 구하게 되는데 여기서 말하는 완전한 자료란 기록된 일정 크기(즉, 임계값) 이상의 역사적 홍수자료와 체계적인 관측자료의 조합을 말한다. 이 때 표본 모멘트는 식 4 ~ 6으로 구할 수 있다.

$$\hat{\mu}_{i+1} = \frac{\sum X^{<}_{S} + \sum X^{>} + N^{<}_{H} E[X^{<}_{H}]}{N} \qquad \text{<식 4>}$$

$$\hat{\sigma}^2_{i+1} = \{ c_2 (\sum (X^{<}_{S} - \hat{\mu}_{i+1})^2 + \sum (X^{>} - \hat{\mu}_{i+1})^2) + N^{<}_{H} E[(X^{<}_{H} - \hat{\mu}_{i+1})^2] \}/N \qquad \text{<식 5>}$$

$$c_2 = \frac{N^{<}_{S} + N^{>}}{N^{<}_{S} + N^{>} - 1}$$

$$\gamma_{i+1} = \{ c_3 (\sum (X^{<}_{S} - \hat{\mu}_{i+1})^3 + \sum (X^{>} - \hat{\mu}_{i+1})^3) + N^{<}_{H} E[(X^{<}_{H} - \hat{\mu}_{i+1})^3] \}/N\hat{\sigma}^3_{i+1} \qquad \text{<식 6>}$$

$$c_3 = \frac{(N^{<}_{S} + N^{>})^2}{(N^{<}_{S} + N^{>} - 1)(N^{<}_{S} + N^{>} - 2)}$$

여기서, C_1, C_2, C_3은 편이 수정량(bias correction factor)이다. $\{X^{>}_{S}\}$는 체계적인 관측자료에서 임계값 Y 보다 큰 관측 유량을 의미하며, $\{X^{>}_{H}\}$는 역사적 자료에서 임계값 Y 보다 큰 관측값을 나타낸다. 또한 $\{X^{<}_{S}\}$는 체계적인 관측자료에서 임계값보다 작은 값을 의미하며 $\{X^{<}_{H}\}$는 역사적 관측 자료에서 임계값보다 작은 값으로 역사적 홍수자료의 특성상 정확한 값을 알 수 없어도 된다. 즉, 옛 문헌에 일정 크기 이상의 홍수만 기록된 경우, 문헌을 통해 $\{X^{>}_{H}\}$의 크기와 개수는 명확하게 알 수 있으나 $\{X^{<}_{H}\}$를 정량적으로 알기란 어려울 수 있다. 과거값에서 임계값보다 작은 값을 구하기는 어렵다는 문제를 해결하기 위해 EMA에서는 $X^{<}_{H}$에 기대값을 취하며 식 7과 같은 조건부 모멘트 산정식을 이용해 구할 수 있다.

$$E[X^{<}_{H} | \alpha, \beta, \tau] = E[X | X < Y, | \alpha, \beta, \tau] = \tau + \beta \frac{\Gamma(\frac{Y-\tau}{\beta}, \alpha + 1)}{\Gamma(\frac{Y-\tau}{\beta}, \alpha)} \qquad \text{<식 7>}$$

$$\Gamma(y, \alpha) = \int_0^y t^{\alpha - 1} \exp(-t) dt$$

여기서, 매개변수는 현 단계의 매개변수인 $\hat{\alpha}_{i+1}, \hat{\beta}_{i+1}, \hat{\tau}_{i+1}$ 이고 $Y = \log(T)$이다. 이렇게 구한 표본 모멘트를 이용해 LP3 분포형의 매개변수를 추정하는 과정이 4 단계이다. 5 단계에서는 3, 4 단계의 구한 매개변수가 2단계 매개변수값과 수렴될 때까지 반복한다.

3. GEV 분포를 위한 EMA의 유도

GEV 분포에 맞게 EMA를 기법을 유도하면 다음과 같다. 위에 서술한 식 1 ~ 3과 식 7를 GEV 분포의 맞게 교체해야 한다. LP3 분포의 매개변수 추정식에 해당되는 식 1 ~ 3은 식 8 ~ 10으로 바뀌게 되고 차례로 위치, 규모, 형상 매개변수 추정식이 되겠다(Stedinger, 1993).

$$\hat{\xi} = \hat{\mu} - \frac{\hat{\alpha}}{\hat{\kappa}}[1 - \Gamma(1 + \hat{\kappa})] \qquad \text{<식 8>}$$

$$\hat{\alpha} = \frac{\hat{\sigma}|\hat{\kappa}|}{[\Gamma(1 + 2\hat{\kappa}) - \Gamma(1 + \hat{\kappa})^2]^{1/2}} \qquad \text{<식 9>}$$

$$\hat{\gamma} = \text{sign}(\hat{\kappa}) \cdot \frac{-\Gamma(1 + 3\hat{\kappa}) + 3\Gamma(1 + \hat{\kappa})\Gamma(1 + 2\hat{\kappa}) - 2[\Gamma(1 + \hat{\kappa})]^3}{[\Gamma(1 + 2\hat{\kappa}) - \Gamma(1 + \hat{\kappa})^2]^{3/2}} \qquad \text{<식 10>}$$

식 7은 GEV 분포의 경우에도 아래와 같은 조건부 확률(conditional probability)을 시작으로 유도된다.

$$f_t(x|-\infty \le X \le x_0) = \frac{f(x)}{F(x_0)} \qquad \text{<식 11>}$$

여기서, x_0는 임계값이다. 식 11과 같은 연속형 확률밀도함수는 식 12와 같이 기대값을 구할 수 있다.

$$E[x|-\infty \le X \le x_0] = \int_{-\infty}^{x} x f_t(x)\,dx = \frac{1}{F(x_0)} \int_{-\infty}^{x_0} x f(x)\,dx \qquad \text{<식 12>}$$

GEV 분포는 식 13과 같은 확률밀도함수(probability density function)를 갖으며 누가분포함수(cumulative distribution function)는 식 14와 같다.

$$f(x) = e^{-y}\left(\frac{\alpha}{\kappa} - \frac{\alpha}{\kappa} y^{\kappa} + \xi\right)$$

$$y = \left[1 - \kappa\left(\frac{x-u}{\alpha}\right)\right]^{1/\kappa} \qquad \text{<식 13>}$$

$$F(x) = \exp\left[-\left(1 - \kappa\frac{x-\xi}{\alpha}\right)^{1/k}\right] \quad \cdot \quad \kappa \neq 0$$

$$= \exp\left[-\exp\left(-\frac{x-\xi}{\alpha}\right)\right] \quad \kappa = 0 \qquad \text{<식 14>}$$

식 13과 같이 $\left[1 - \kappa\left(\frac{x-\xi}{\alpha}\right)\right]^{1/\kappa}$를 y로 치환하고 정리하면 식 15와 같다. 참고로 적분의 아래 끝은 0이 된다.

$$\int_{-\infty}^{x_0} x f(x)\,dx = \int_{0}^{y_0} e^{-y}\left(\frac{\alpha}{\kappa} - \frac{\alpha}{\kappa} y^{\kappa} + \xi\right) dy \qquad \text{<식 15>}$$

본 연구에서는 식 15의 적분을 정리해 식 12에 대입하여 조건부 1차 모멘트를 구하였다(식 16). 차례로 조건부 2차, 3차 모멘트도 아래와 같이 유도하였다.

$$E[X|X<Y,|\xi,\alpha,\kappa]=(1+\frac{\alpha}{\kappa})(1-e^{y_0})+\Gamma(\kappa+1,y_0)(-\frac{\alpha}{\kappa}) \quad \text{<식 16>}$$

$$E[X^2|X<Y,|\xi,\alpha,\kappa]=(\xi+\frac{\alpha}{\kappa})^2(1-e^{y_0})-2\frac{\alpha}{\kappa}(\xi+\frac{\alpha}{\kappa})\Gamma(\kappa+1,y_0)$$
$$+(\frac{\alpha}{\kappa})^2\Gamma(2\kappa+1,y_0) \quad \text{<식 17>}$$

$$E[X^3|X<Y,|\xi,\alpha,\kappa]=(\xi+\frac{\alpha}{\kappa})^3(1-e^{y_0})-3\frac{\alpha}{\kappa}(\xi+\frac{\alpha}{\kappa})^2\Gamma(\kappa+1,y_0)$$
$$+3(\frac{\alpha}{\kappa})^2(u+\frac{\alpha}{\kappa})\Gamma(2\kappa+1,y_0)-(\frac{\alpha}{\kappa})^2\Gamma(3\kappa+1,y_0) \quad \text{<식 18>}$$

윗 식은 $E[X^p|X<Y,|\xi,\alpha,\kappa]$를 구하기 위해서 임계값 x_0로부터 구한 y_0와 현 매개변수 ξ,α,κ 만을 필요로 한다. 이는 식 7과 같이 4개의 입력값으로 기대값을 구할 수 있는 형태이다.

GEV-EMA의 절차는 LP3-EMA와 같고, 다만 LP3 분포에서만 유효하였던 각 식을 위와 같이 GEV 분포에 맞게 유도된 식으로 교체하면 된다.

4. 결론 및 향후 연구

빈도해석 선진국격인 미국에서는 이상홍수의 초과확률을 통계적으로 보다 강건하게 추정하기 위해서 이미 LP3 분포를 대상으로 EMA 추정법을 제시하였다. 공식 분포로 LP3를 삼고있는 미국에서는 홍수보험 등의 이유로 망설이고 있지만 공식 분포를 GEV 분포로 대체하자는 목소리가 커지고 있다. 이러한 때에 본 연구에서는 역사 자료 이용 알고리즘인 EMA를 GEV 분포에 맞게 유도하였다. 비록 역사 자료 이용에 대한 관심은 늦었지만 미국의 사례를 교훈 삼는다면 시행 착오를 줄여 늦은 시작을 만회할 수 있을 것이라 기대된다. 궁극적으로 이런 노력은 우수한 이상홍수 초과확률 산정 기법의 밑바탕이 될 것이다.

향후에 역사 강우자료를 강우-유출 모형으로 역사 홍수자료로 변환해 본 알고리즘을 적용하거나 직접 역사 홍수자료에 적용할 예정이다.

감사의 글

본 연구는 건설교통부 한국건설교통기술평가원의 이상기후대비시설기준강화 연구단에 의해 수행되는 2005 건설기술기반구축사업(05-기반구축-D03-01)에 의해 지원되었습니다.

참고문헌

1. 김현준 (2000) “조선왕조실록에 수록된 홍수기록 조사.” 한국수자원학회논문집 , 1226-6280 , 33(S1), pp. 365-370.
2. 심태현, 임규호 (2007). “측우기 관측 자료에서 나타난 서울 강수 시계열의 특징.” 2007년 한국기상학회 봄학술대회 논문집, pp. 500-501.
3. 전종갑, 문병권 (1997). “측우기 강우량 자료의 복원과 분석.” 한국기상학회지, 33(4), pp. 691-707.
4. 조한범, 김현준, 노성진, 장철희 (2007). “역사 문헌을 통한 극한홍수 데이터베이스 구축.” 2007년 한국수자원학회 학술발표회 논문집, pp. 741-745.
5. Cohn, T. A., Lane, W. L., and Baier, W. G. (1997). “An algorithm for computing moments-based flood quantile estimates when historical flood information is available." *Water Resources Research*, 33(9), pp. 2089-2096.
6. Stedinger, J. R., Vogel, R. M., and Foufoula-Georgiou, E. (1993). “Chapter 18, Frequency analysis of extreme events,” Handbook of Hydrology, edited by Maidment, D. R., McGraw-Hill.