Bayesian MCMC를 이용한 저수량 점 빈도분석: Ⅰ. 사전분포의 적용성 비교

At-site Low Flow Frequency Analysis Using Bayesian MCMC: Ⅰ. Comparative study for construction of Prior distribution

김상욱*, 이길성, 박경신*****

Sang Ug Kim, Kil Seong Lee, Kyungshin Park

요 지

저수분석(low flow analysis)은 수자원공학에서 중요한 분야 중 하나이며, 특히 저수량 빈도분석(low flow frequency analysis)의 결과는 저수(貯水)용량의 설계, 물 수급계획, 오염원의 배치 및 관개와 생태계의 보존을 위한 수량과 수질의 관리에 중요하게 사용된다. 그러므로 본 연구에서는 저수량 빈도분석을 위한 점 빈도분석을 수행하였으며, 특히 빈도분석에 있어서의 불확실성을 탐색하기 위하여 Bayesian 방법을 적용하고 그 결과를 기존에 사용되던 불확실성 탐색방법과 비교하였다. 본 논문의 Ⅰ편에서는 Bayesian 방법 중 사전분포(prior distribution)와 우도함수(likelihood function)의 복잡성에 상관없이 계산이 가능한 Bayesian MCMC(Bayesian Markov Chain Monte Carlo) 방법과 Metropolis-Hastings 알고리즘을 사용하기 위한 여러 과정의 이론적 배경과 Bayesian 방법에서 가장 중요한 요소인 사전분포를 구축하고 이를 비교 및 평가하였다. 고려된 사전분포는 자료에 기반하지 않은 사전분포와 자료에 기반한 사전분포로써 두 사전분포를 이용하여 Metropolis-Hastings 알고리즘을 수행하고 그 결과를 비교하여 저수량 빈도분석에 합리적인 사전분포를 선정하였다. 또한 알고리즘의 수행과정에서 필요한 제안분포(proposal distribution)를 적용하여 그에 따른 알고리즘의 효율성을 채택률(acceptance rate)을 산정하여 검증해 보았다. 사전분포의 분석 결과, 자료에 기반한 사전분포가 자료에 기반하지 않은 사전분포보다 정확성 및 불확실성의 표현에 있어서 우수한 결과를 제시하는 것을 확인할 수 있었고, 채택률을 이용한 알고리즘의 효용성 역시 기존 연구자들이 제시하였던 만족스러운 범위를 가지는 것을 알 수 있었다. 최종적으로 선정된 사전분포는 본 연구의 Ⅱ편에서 Bayesian MCMC 방법의 사전분포로 이용되었으며, 그 결과를 기존 불확실성의 추정방법의 하나인 2차 근사식을 이용한 최우추정(maximum likelihood estimation)방법의 결과와 비교하였다.

핵심용어 : 저수량 점 빈도분석, 불확실성, Bayesian MCMC, Metropolis-Hastings 알고리즘, 사전분포, 최우추정방법, 2차 근사법

1. 서 론

저수분석(Low flow analysis)은 수공구조물의 설계, 하천환경의 보전 및 생활·공업·농업용수의 안전한 취수를 위한 최소 유량의 보장, 오염원의 배치 등 수량과 수질의 관리에 중요하게 사용된다. 그러나 수자원 장기종합계획, 댐 건설 장기계획 등의 국내 주요 중장기 계획은 빈도분석의 확정적인(Deterministic) 값만을 이용하여 수립되고 있으며, 빈도분석결과의 불확실성을 반영한 확률적인(Probabilistic) 값이 이용되는 계획은 찾아보기 힘들다. 이는 불확실성에 대한 인식 부족과 함께 불확실성에 대한 계산방법이 현실을 제대로 반영하지 못함으로써 빈도분석 결과의 정확성에 대한 신뢰도가 낮은 것에 기인한다고 할 수 있다. 불확실성을 나타내기 위해서 일종의 근사식을 사용하여 대략적인 신뢰구간을 산정하고 이로부터 분석결과의 불확실성을 표현하고자 하는 연구가 진행된 바 있다. 그러나 근사식을 사용한 신뢰구간 산정방법은 확률 분포함수

* 정회원・서울대학교 BK21 SIR 사업단 박사후 연구원・E-mail : plethor1@snu.ac.kr
** 정회원・서울대학교 공과대학 건설환경공학부 교수・E-mail : kilselee7@snu.ac.kr
*** 서울대학교 공과대학 건설환경공학부 석사과정・E-mail : diploma@snu.ac.kr

의 모수를 산정함에 있어서 정상성(Normality), 선형성(Linearity) 등의 가정이 필요하므로, 불확실성을 산정함에 있어서 비현실적인 값을 산정하거나 과대 추정되는 경우가 있는 것으로 알려져 있다(Reis Jr. and Stedinger, 2005). 근사식을 사용한 모수의 불확실성 추정방법을 대신하여 Bayesian 접근방법을 사용한 모수 및 불확실성의 추정이 수행될 수 있다. Bayesian 방법은 근사식을 사용하기 위한 가정 조건이 필요하지 않기 때문에 특히 불확실성을 표현하는 데 있어서 근사식을 사용한 방법보다 우월할 수 있다. Bayesian 방법을 사용하기 위해서는 사전분포(Prior distribution)로부터 사후분포(Posterior distribution)를 산정해야 하는데, 이를 계산하기 위해서는 해석적으로 산정되기 어려운 적분항들이 포함되어 진다. 그러므로 과거 초기 연구단계에서는 공액사전분포(Conjugated prior distribution)을 적용하여 사후분포를 쉽게 해석적으로 구하여 모수 및 불확실성을 나타내고자 하는 연구가 진행된 바 있다(Vicens et al., 1975; Wood and Rodriguez-Iturbe, 1975a, b). 그러나 공액 사전분포를 사용한 Bayesian 접근 방법의 수자원 공학으로의 적용은 자료를 합리적으로 표현할 수 있는 사전분포를 사용하게 되지 못하는 경우가 많아 Bayesian 방법을 사용하기 위한 원래 목적을 벗어나게 될 수 있다는 논쟁으로 그 간 활발히 적용되지 못하다가 최근 들어 발전한 계산 능력과 전체 탐색법 개념의 알고리즘의 개발로 인하여 90년대 후반부터 다시 빈도분석 분야, 강우-유출모형의 보정 분야 등에 활발히 사용되고 있다. 그러나 위에서 제시한 저수분석의 중요성에도 불구하고 Bayesian 방법을 사용하여 저수량 빈도분석을 수행하기 위한 일련의 과정들은 연구사례가 상대적으로 적다.

2. Bayesian MCMC와 Metropolis-Hastings 알고리즘

베이즈의 정리를 연속 확률밀도함수(Probability density function)로 나타내면 베이즈의 정리는 식 (1)과 같이 표현될 수 있다.

$$\pi(\boldsymbol{\theta}|\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_n}) = \frac{f(\boldsymbol{x_1}|\boldsymbol{\theta}) \cdots f(\boldsymbol{x_n}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int_{\theta} f(\boldsymbol{x_1}|\boldsymbol{\theta}) \cdots f(\boldsymbol{x_n}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\theta} \quad (1)$$

식 (1)에서 좌변의 $\pi(\boldsymbol{\theta}|\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_n})$는 사후분포(Posterior distribution), 우변 분자의 $\pi(\boldsymbol{\theta})$는 사전분포(Prior distribution)라 명명되며, 우변의 분모는 상수로서 주변분포(Marginal distribution)이고, 우변의 분자의 $f(\boldsymbol{x_1}|\boldsymbol{\theta}) \cdots f(\boldsymbol{x_n}|\boldsymbol{\theta})$는 발생할 수 있는 모든 가능성을 고려한 우도함수(Likelihood function)이다. 그러므로 식(1)로부터 사후분포는 우도함수와 사전분포의 곱에 비례하게 됨을 알 수 있다. 분석하고자 하는 자료를 나타낼 수 있는 확률밀도함수가 결정되면 이로부터 우도함수를 유도할 수 있고, 적절한 사전분포를 부여함으로써 사후분포로부터 확률밀도함수의 모수를 추출하고 모수의 불확실성을 탐색할 수 있다.

그러나 Bayesian 방법을 이용하여 사후분포를 계산하는 것은 쉬운 작업이 아니다. 특히 식 (1)의 우변의 분모의 적분은 결정된 확률밀도함수에 따라 적분이 가능할 수도 있으나, 대부분의 확률밀도함수는 수학적으로 적분하기 어려운 경우가 많다. 그러므로 Bayesian 방법의 초기 연구단계에서는 위와 같은 Bayesian 방법의 적용에 필요한 계산을 위하여 공액사전분포(Conjugated prior distribution)를 이용하는 경우가 많았다. 그러나 최근 계산능력의 하드웨어부분의 발전과 주변 확률분포함수의 적분이 필요없는 Metropolis-Hastings 알고리즘, Gibbs sampling 알고리즘, 주표집(Importance sampling) 알고리즘과 같은 Bayesian MCMC 방법에 입각한 Bayesian 계산방법의 소프트웨어부분의 발전으로 인하여 공액사전분포를 사용하지 않아도 식 (1)의 계산이 가능하게 됨으로서 서론에 언급한 연구사례와 같이 최근 들어 수자원공학 분야에서도 다시 활발히 적용되고 있는 실정이다. Bayesian MCMC방법이란 마코프 연쇄(Markov chain)과 몬테카를로 적분(Monte carlo integration)을 이용하여 사후분포로부터 모수를 추출하고 통계적 특성치를 계산하는 방법이다. 즉, 마코프 연쇄를 이용하여 모수간의 관계를 구성하고 이를 상당히 큰 수만큼 반복하는 몬테카를로 적분기법을 이용하여 최종적으로 모수의 통계적 특성을 산정하는 방법이다. 여러 가지 Bayesian 계산 방법 중에서 가장 활발히 사용되고 있는 알고리즘은 Metropolis-Hastings 알고리즘으로 기본적인 개념은 Metropolis et al.(1953)에 의하여 만들어져 최근 들어 활발히 이용되고 있다.

3. 사전분포의 구축방법

3.1 확률밀도함수의 선정 및 우도함수

본 연구에서 사용되는 확률밀도함수는 Nathan and McMahon(1990), Önöz and Bayazit(2001)이 저수량 빈도분석을 위해 사용한 바 있는 2모수 Weibull 분포로서 이에 대한 우도함수는 식(2)와 같다.

$$L(\mathbf{x}\,|\alpha,\beta)=\left(\frac{\alpha}{\beta}\right)^{n}\prod_{i=1}^{n}\left(\frac{x_i}{\beta}\right)^{\alpha-1}\exp\left[-\sum_{i=1}^{n}\left(\frac{x_i}{\beta}\right)^{\alpha}\right] \tag{2}$$

여기서, α는 형상모수(Shape parameter)이고 β는 축척모수(Scale parameter)이다.

3.2 자료에 기반하지 않은 사전분포: Prior Ⅰ

자료에 기반하지 않은 사전분포 중 대표적인 사전분포는 무정보적 사전분포(Non-informative prior distribution)로써 이는 추정하고자 하는 모수에 대한 과거 경험에 대한 정보나 사용자의 주관이 전무한 경우에 사용된다. 무정보적 사전분포로는 각각의 모수에 대하여 다음과 같은 균일분포만을 적용하고 두 균일분포가 서로 통계적으로 독립적이란 가정 하에 최종적으로 $\pi(\alpha,\ \beta)$를 유도하였다.

$$\therefore \pi(\alpha,\beta)=\pi(\alpha)\pi(\beta)=\frac{1}{\beta} \tag{3}$$

3.3 자료에 기반한 사전분포: Prior Ⅱ

자료에 기반한 사전분포는 추정하고자 하는 모수의 과거 자료나 사용자의 경험에 기반한 주관적인 판단에 의해 구축될 수 있다. 그러나 사용자의 경험에 기반한 주관적 사전분포는 사용자의 경험을 자료를 이용하여 표현하는데 있어서 정량화가 힘들어 어려운 부분이 있다. 본 연구에서는 자료에 기반한 사전분포를 구축하기 위하여 에르고딕(Ergodic)가정을 이용하여 낙동강 유역의 13개 지점의 7Q자료를 이용하여 진동에서의 자료에 기반한 사전분포를 구축하였다. 그림 1에서 진동 지점을 제외한 나머지 13개 지점에서의 통계적 특성치가 진동 지점에서의 13년간의 통계적 특성치와 일치하게 된다는 가정을 수립할 수 있고, 이로부터 진동지점에 대한 자료에 기반한 사전분포를 구축할 수 있다. 필요한 사전분포는 형상모수 α와 축척모수 β에 대하여 구축되어져야 하므로, 먼저 진동지점을 제외한 13개 지점의 7Q 유량을 2모수 Weibull 분포에 적합시킨 후, 13개의 α 추정치와 β 추정치를 각각 모수 λ를 가지는 지수분포와 형상모수 b, 축척모수 a를 가지는 2모수 Weibull 분포를 이용하여 다시 적합시켰다. 위와 같은 과정을 통하여 본 연구에서 사용되어지는 2모수 Weibull 분포의 형상모수 α와 축척모수 β에 대한 사전분포의 구축을 나타내면 다음 식과 같다.

Figure 1. Concept for Prior Ⅱ

$$\pi(\alpha)=\exp(-\lambda\alpha) \tag{4}$$

$$\pi(\beta)=\frac{b}{a}\left(\frac{\beta}{a}\right)^{b-1}\exp\left[-\left(\frac{\beta}{a}\right)^{b}\right] \tag{5}$$

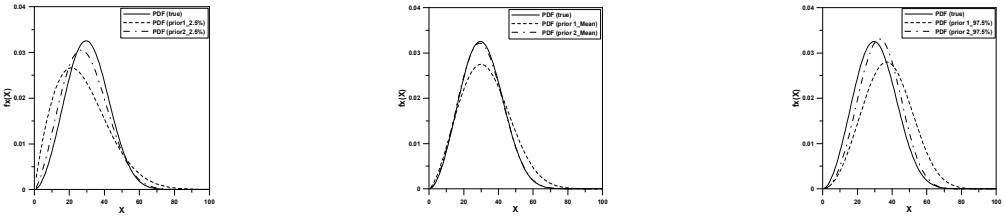
$$\therefore \pi(\alpha,\ \beta)=\frac{b}{a}\left(\frac{\beta}{a}\right)^{b-1}\exp\left[-\left(\left(\frac{\beta}{a}\right)^{b}+\lambda\alpha\right)\right] \tag{6}$$

여기서 λ, a, b는 위에서 각각 추정된 0.98, 27.15, 1.92이다.

4. Prior Ⅰ과 Prior Ⅱ의 적합성 비교

본 연구에서는 사전분포의 적합성에 대한 통계적 실험을 위하여 진동지점의 36개 7Q값에 대해 최우추정법을 이용하여 추정된 $\alpha=2.8371$과 $\beta=34.4203$을 참값(True parameter)로 사용하였다. 아래 그림 2를 보면 평균값에 있어서 Prior Ⅱ가 Prior Ⅰ보다 참값을 추정하는 데 있어서 보다 정확하게 추정하였음을 알 수 있다. 또한 불확실성을 표현할 수 있는 97.5 %와 2.5 %의 차이는 α가 Prior Ⅰ, Prior Ⅱ의 경우 각각 1.1061, 0.6613 이고 β가 각각 10.882와 4.1224로써 Prior Ⅱ를 사용한 경우 불확실성 측면에서 많은 감소가 있었음을 알 수 있다. 그러므로, Bayesian MCMC방법을 이용하여 저수량 점 빈도분석을 수행하고자 하는 경우, 자료에 기반하지 않은 무정보적 사전분포보다는 자료에 기반한 사전분포를 사용하는 것이 Bayesian MCMC를 수행하는 데 있어서 합리적일 수 있다는

결론을 얻을 수 있었으며, 본 연구의 Ⅱ편에서는 자료에 기반한 사전분포를 이용한 Bayesian MCMC 방법과 MLE 2차 근사방법의 추정결과를 비교함으로써 저수량 빈도분석에 있어서 불확실성을 보다 정확히 분석하는 연구를 수행하였다.



(a) 2.5% (b) Mean (c) 97.5%

Figure 5. Comparison with PDFs of Weibull distribution:

5. 결론

본 연구에서는 Bayesian MCMC 방법을 이용하여 저수량 점 빈도분석에서 추정될 수 있는 불확실성의 표현을 위한 여러 가지 이론적 배경을 서술하였다. 또한 Bayesian MCMC 방법을 적용하는 데 있어서 가장 중요한 요소인 사전분포를 선정함에 있어서 자료에 기반하지 않은 사전분포와 에르고딕(Ergodic)가정을 이용한 자료에 기반한 사전분포를 구축하고, 두 가지 사전분포를 통계적 실험을 통하여 비교함으로써 자료에 기반한 사전분포가 평균값의 추정과 불확실성 측면에서 보다 나은 결과를 돌출함을 확인하였다. 일반적으로 Bayesian 방법을 적용하는 경우에 계산을 간편하게 하기 위하여 공액사전분포를 사용하는 등 과거의 경험을 합리적으로 나타내지 못하는 사전분포를 선정하여 사용하는 것은 최종 결과가 현실을 제대로 반영하지 못 할 요소가 많으므로 가능하면 자료에 기반한 사전분포를 구축하여 사용하는 것이 합리적이라는 결론을 얻을 수 있었다. 본 연구의 Ⅱ편에서는 Ⅰ편에서 제시한 이론들과 선정된 사전분포를 사용하여 낙동강 유역에서의 저수량 점 빈도분석을 수행하였다. 또한 Ⅰ편에서 상세히 서술하지 못한 저수량 자료에 대한 한계성 부분과 적용방법을 추가적으로 설명하였다.

감사의 글

본 연구는 21세기 프런티어 연구개발 사업인 수자원의 지속적 확보기술개발 사업단(과제번호 1-7-3)의 서울대학교 공학연구소를 통한 연구비 지원(30 %)과 서울대학교 BK21 안전하고 지속가능한 사회기반건설사업단의 연구비 지원(70 %)에 의해 수행되었습니다. 연구비 지원에 심심한 감사의 뜻을 표합니다.

참고문헌

Metropolis, N., Rosenbluth, A.W., Teller A.H., and Teller E. (1953). "Equations of state calculations by fast computing machines." *Journal of Chemical Physics*, Vol. 21, pp. 1087-1092.

Nathan, R.J., and McMahon, T.A. (1990). "Practical aspects of low flow frequency analysis." *Water Resources Research*, Vol. 26, No. 9, pp. 2135-2141.

Önöz, B., and Bayazit, M. (2001). "Power distribution for low streamflows." *Journal of Hydrologic Engineering*, Vol. 6, No. 5, pp. 429-435.

Reis Jr., D.S., and Stedinger, J.R. (2005). "Bayesian MCMC flood frequency analysis with historical information." *Journal of Hydrology*, Vol. 313, pp. 97-116.

Vicens, G.J., Rodriguez-Iturbe, I., and Schaake Jr, J.C. (1975). "A Bayesian framework for the use of regional information in hydrology." *Water ResourcesResearch*, Vol. 11, No. 3, pp. 405-414.

Wood, E.F., and Rodriguez-Iturbe, I. (1975a). "Bayesian inference and decision making for extreme hydrologic events." *Water Resources Research*, Vol. 11, No. 4, pp. 533-542.

Wood, E.F., and Rodriguez-Iturbe, I. (1975b). A Bayesian approach to analyze uncertainty among flood frequency models. *Water Resources Research*, Vol. 11, No. 6, pp. 839-843.