J. Korea Water Resour. Assoc. Vol. 49, No. 6 (2016), pp. 481-493
doi: 10.3741/JKWRA.2016.49.6.481
pISSN 1226-6280
eISSN 2287-6138

Comparison of probability distributions to analyze the number of occurrence of torrential rainfall events

Kim, Sang Ug[a]* · Kim, Hyeung Bae[a]

[a]Dept. of Civil Engineering, Kangwon National University, Chuncheon, Korea

Paper number: 16-019
Received: 12 Febuary 2016; Revised: 22 March 2016; Accepted: 22 March 2016

Abstract

The statistical analysis to the torrential rainfall data that is defined as a rainfall amount more than 80 mm/day is performed with Daegu and Busan rainfall data which is collected during 384 months. The number of occurrence of the torrential rainfall events can be simulated usually using Poisson distribution. However, the Poisson distribution can be frequently failed to simulate the statistical characteristics of the observed value when the observed data is zero-inflated. Therefore, in this study, Generalized Poisson distribution (GPD), Zero-Inflated Poisson distribution (ZIP), Zero-Inflated Generalized Poisson distribution (ZIGP), and Bayesian ZIGP model were used to resolve the zero-inflated problem in the torrential rainfall data. Especially, in Bayesian ZIGP model, a informative prior distribution was used to increase the accuracy of that model. Finally, it was suggested that POI and GPD model should be discouraged to fit the frequency of the torrential rainfall data. Also, Bayesian ZIGP model using informative prior provided the most accurate results. Additionally, it was recommended that ZIP model could be alternative choice on the practical aspect since the Bayesian approach of this study was considerably complex.

Keywords: torrential rainfall, Poisson distribution, zero-inflated, Bayesian, informative prior distribution

집중호우사상의 발생횟수 분석을 위한 확률분포의 비교

김상욱[a]* · 김형배[a]

[a]강원대학교 공과대학 토목공학과

요 지

본 연구에서는 최근 기후변화로 인한 집중호우의 발생횟수의 경향을 확률적으로 분석함에 있어 1개월 동안 80 mm/day 이상의 강우사상을 집중호우로 정의하여, 대구 및 부산 강우관측소로부터 수집된 384개월 동안의 집중호우를 분석하였다. 집중호우 월별 발생횟수와 같은 형식의 자료의 확률적 분석은 대개 Poisson 분포 (POI)가 사용되나 자료에 포함된 0자료의 과잉은 확률분포를 왜곡시키는 문제를 발생시킨다. 본 연구에서는 이 문제를 개선하기 위하여 개발된 일반화 Poisson 확률분포 (GPD), 0-과잉 Poisson 확률분포 (ZIP), 0-과잉 일반화 Poisson 확률분포 (ZIGP), Bayesian 0-과잉 일반화 Poisson 확률분포 (Bayesian ZIGP)를 집중호우 자료에 적용하고, 5개 모형의 특성을 비교분석하였으며, Bayesian ZIGP 모형의 구축에 있어서는 정보적 사전분포를 사용함으로써 모형의 정확도를 개선하였다. 분석결과 분석하고자 하는 자료에 0이 과다하게 포함되어 있는 경우 POI 및 GPD 분포는 관측결과와는 다른 결과를 제시하여 적절한 모형으로 고려되지 못함을 알 수 있었다. 5가지 모형 중 정보적 사전분포를 탑재한 Bayesian ZIGP 모형이 가장 관측 자료와 유사한 결과를 도출하였으나 모형의 구축에 수반되는 실용적인 측면을 고려하면 ZIP 모형도 충분히 사용될 수 있는 모형으로 추천되었다.

핵심용어: 집중호우, Poisson 분포, 0-과잉, 베이지안, 정보적 사전분포

***Corresponding Author.** Tel: +82-33-250-6233
E-mail: sukim70@kangwon.ac.kr (S. U. Kim)

1. 서 론

지난 이십년 동안 발생되고 있는 집중호우(torrential rainfall)의 횟수는 과거에 비해 뚜렷이 증가되고 있으며, 1980년 이래 발생된 집중호우로 인한 홍수 피해액은 전 세계적으로 약 470조 원에 이르는 것으로 보고된 바 있다 (HSBC, 2011). 특히 최근 기후변화로 인한 이상기후의 발생은 이와 같은 집중호우 의 발생 경향을 가속화시키고 있으며, 이와 같은 상황은 2002년 태풍 루사로 인해 발생된 약 870 mm/day의 강우발생이나 최근 반복되고 있는 도심지 침수의 발생 등 국내에서도 같은 양상을 보이고 있다. 높은 강우강도를 가지는 집중호우는 도심홍수(urban flood)의 발생을 비롯하여 산지에서의 돌발홍수(flash flood)와 함께 산사태(land slide) 또는 토석류(debris flow)와 같은 자연재해를 발생시킬 수 있는 위험성이 매우 높으므로 과거의 발생 현황을 과학적으로 분석함으로써 이와 같은 재해발생을 사전에 대비할 필요가 있다.

Saidi et al. (2015)은 기후변화로 인한 극치강우의 가속화 정도를 정량화하는 연구를 수행한 바 있으며, Goyal (2014)은 1901년부터 2002년까지의 강우자료를 활용하여 극치강우의 장기경향성을 분석한 바 있다. 국내에서도 Cho and Kim (2015)은 극치강우의 시간변동 특성을 고려함에 있어 적절히 활용될 수 있는 강우시간에 대한 분포를 개발한 바 있으며, Yoon et al. (2014)은 2011년 7월 발생된 집중호우를 모니터링함에 있어 위성자료를 활용하여 집중호우 관측을 위한 알고리즘 및 모니터링의 가능성을 제시한 바 있다. 위에서 제시한 연구들은 대부분 연최대시계열(annual maximum series) 자료를 일강우 또는 시강우 자료로부터 추출하여 연구에 사용하였으므로 극치강우의 경향 변화가 주된 분석결과이다.

그러나 하수관거 용량의 부족으로 인한 도시홍수나 산사태 등은 연최대강우자료 보다는 적은 강우강도를 가지는 집중호우로부터도 충분히 발생될 가능성이 있다. 이를 보다 객관화하기 위해서는 먼저 집중호우에 대한 전 세계적으로 일치된 과학적 정의가 필요한 실정이나 집중호우는 일반적으로 80 mm/day 또는 100 mm/day로 정의되어 사용되기도 하고 30 mm/hr 또는 50 mm/hr로 정의되어 사용되기도 하는 등 아직까지 국가마다 집중호우에 대한 정의가 차이가 있으며 국내에서도 통일되어 있지 못한 실정이다. 다만 우리나라 기상청에서는 집중호우를 80 mm/day 나 30 mm/hr로 정의하고 있어 본 연구에서는 집중호우의 정의로 80 mm/day 이상의 강우를 사용하였다.

강우의 발생현황을 분석하기 위해서는 주로 특정 시간 동안 발생되는 강우사상의 개수를 추계학적 (stochastic)으로 분석하는 경우가 많은 데 이렇게 특정시간 동안 발생되는 자료의 개수를 ‘count data’라고 하며 자연과학뿐만 아니라 사회과학에서도 count data의 발생을 확률통계적으로 분석하고자 하는 경우가 많다. 수자원공학 분야에서의 count data를 다룬 연구사례는 연구의 중요성에 비해서는 많은 편은 아니라고 판단되며 특히 국내 연구는 찾아보기 힘들다. Todorovic and Yevjevich (1969)는 특정시간 간격동안 발생되어진 태풍의 도착횟수를 Poisson 분포를 이용하여 분석한 바 있으며, Poisson 분포 이외에도 지수분포나 Weibull 분포, Gamma 분포 등을 적용하고 그 결과를 비교 분석된 바 있다 (Katz and Parlange, 1995; Chapman, 1998). 또한 특정 확률분포를 사용하여 count data를 분석하지 않고 다양한 형태의 Markov 연쇄모형을 이용하여 강우발생의 기작을 추계학적으로 모형화하기도 하였으며 (Jimoh and Webster, 1996), 최근 Rauf and Zeephongsekul (2014)은 강우의 지속시간별 발생현황을 분석함에 있어 비모수적 코플라(copulas) 분석을 시도하기도 하였다.

1개월 동안 발생된 80 mm/day이상의 강우사상의 발생 횟수(이하, 집중호우 월별개수)와 같은 count data를 특정 확률분포에 적합(fitting)시키는 과정에서 발생되는 문제는 특정사상이 발생되지 않는 개월의 수가 많다는 것이다. 즉 count data에 포함된 ‘0’의 개수가 너무 많은 경우 (zero-inflated)에는 이로 인해 선정된 확률분포의 확률밀도가 왜곡된다는 것이다. Singh (1963), Cohen (1960), Lambert (1992)는 Poisson 분포가 0이 과잉된 자료를 모의함에 있어 왜곡된 결과를 도출하게 됨을 제시한 바 있으며, 이를 개선하기 위하여 0-과잉 Poisson (Zero-Inflated Poisson, ZIP)모형을 제안한 바 있다. 이후 Gupta (1996)는 ZIP 모형을 일반화한 0-과잉 일반화 Poisson (Zero-Inflated Generalized Poisson, ZIGP)모형으로 확장한 바 있으며, Angers and Biswas (2003)은 ZIGP 모형의 매개변수를 추정함에 있어 기존의 최대우도법(maximum likelihood estimation method)을 사용하지 않고 Bayesian 추정을 활용한 바 있다.

위에서 제시한 바와 같이 0자료가 과잉된 count data의 확률통계적인 분석을 위해 통계학 분야를 비롯하여 의학, 경영 및 경제학, 자연과학 등에서는 많은 연구가 진행되었

으나, 수자원공학과 관련되어서는 이와 유사한 연구가 거의 진행된 바 없다. 따라서 본 연구에서는 우리나라 낙동강 유역의 강우관측소를 대상으로 1개월 동안 발생된 집중호우의 개수 5개의 확률분포인 Poisson 분포, 일반화 Poisson 분포(Generalized Poisson Distribution, GPD), ZIP 분포, ZIGP 분포, Bayesian ZIGP 분포를 이용하여 분석하였으며, 5개 분포의 적용 성능을 상호 비교함으로써 분포별 장단점을 파악하고자 하였다.

2. ZIP 모형, GPD 모형 및 ZIGP 모형

자연 또는 사회 속에서 발생되는 count data를 추계학적으로 모형화하기 위해서 일반적으로 Poisson 분포가 사용되고 있다 (Mullahy, 1986). 그러나 Bohning et al. (1999)는 count data에 포함된 0의 개수가 과잉되는 경우는 Poisson 분포의 적합도가 매우 낮아져 적절한 확률산정이 어렵게 됨을 명확히 제시한 바 있으며, 이후 Johnson et al. (1992)는 이와 같은 0 과잉 문제를 해결하기 위하여 Eq. (1)과 같이 확률변수의 값이 0인 경우와 0이 아닌 경우를 특정 가중치 ω를 이용하여 분할하여 사용할 수 있는 개념을 제안하였다.

$$P(X=0)=\omega+(1-\omega)P(X=0) \text{ and}$$
$$P(X=j)=(1-\omega)P(X=j),\ j=1,2,3,... \quad (1)$$

여기서, ω는 0과 1사이의 가중치로써 자료로부터 추정되어야 하는 값이며, $P(X)$에 이산형인(discrete) Poisson 확률분포인 $P(X)=\exp(-\lambda)\lambda^x/x!$가 사용될 때, Eq. (2)와 같은 모형을 0-과잉 Poisson (Zero-Inflated Poisson, ZIP) 모형이라 한다. Eq. (2)에 포함된 ω와 λ는 적률추정법 (method of moment)를 활용한 Eq. (3)으로부터 간단하게 추정값을 산정할 수 있으며, s^2과 $\overline{X}$는 표본의 분산과 평균이다.

$$P(X=0)=\omega+(1-\omega)\exp(-\lambda) \quad \text{and}$$
$$P(X=j)=(1-\omega)\frac{\lambda^x}{x!e^{\lambda}} \quad j=1,2,3,... \quad (2)$$

$$\hat{\omega}=\frac{s^2-\overline{X}}{s^2+\overline{X}^2-\overline{X}} \quad \text{and} \quad \hat{\lambda}=\frac{s^2+\overline{X}^2-\overline{X}}{\overline{X}} \quad (3)$$

Consul and Jain (1973)은 이항분포(binomial distribution), 음이항분포(negative binomial distribution), Poisson 분포를 모수의 추정값에 따라 범용적으로 사용가능한 Eq. (4)의 이변수 일반화 Poisson 분포(Generalized Poisson Distribution, GPD)를 제안하였으며, 적률추정법을 활용한 2개의 모수에 대한 추정식은 Eq. (5)와 같다.

$$P(X)=\frac{1}{x!}\lambda(\lambda+\alpha x)^{x-1}e^{-\lambda-\alpha x},\quad x=0,1,2,3,... \quad (4)$$

$$\hat{\alpha}=1\pm\sqrt{\overline{X}/s^2} \quad \text{and} \quad \hat{\lambda}=(1-\alpha)\overline{X} \quad (5)$$

여기서, 표본으로부터 추정되는 $\hat{\alpha}$의 범위는 $0\le\hat{\alpha}<1$이며, $\hat{\lambda}$는 $\hat{\lambda}>0$인 범위를 가진다. GPD 모형은 Poisson 분포를 포함하는 범용적인 이산형 분포로 자료가 가지고 있는 이산적인 특성을 특정한 1개의 확률분포로 나타내기보다는 일반화하여 나타낼 수 있다는 장점을 가지고 있으나 이 확률분포 역시 0이 과잉된 경우 Poisson 분포와 같이 확률의 산정에 심각한 오류가 발생된다. 따라서 GPD 모형도 Eq. (1)을 사용하여 0-과잉 문제를 해결할 수 있다. 즉 Eq. (1)의 $P(X)$에 GPD를 대입함으로써 0-과잉 문제를 해결한 Eq. (6)과 같은 0-과잉 일반화 Poisson (Zero-Inflated Generalized Poisson) 모형을 구성할 수 있다.

$$P(X=0)=\omega+(1-\omega)\exp(-\lambda) \quad \text{and}$$
$$P(X=j)=(1-\omega)\frac{1}{x!}\lambda(\lambda+\alpha x)^{x-1}e^{-\lambda-\alpha x},\quad x=1,2,3,... \quad (6)$$

여기서, $\hat{\alpha}$의 범위는 $0\le\hat{\alpha}<1$으로써 ZIGP 모형의 형태를 결정하는 모수이고, $\hat{\alpha}$가 0이 되면 ZIGP 모형은 ZIP 모형과 같게 된다. Angers and Biswas (2003)는 ZIGP 모형에 포함되어있는 3개의 모수 ω, α, λ는 최대우도추정법을 이용한 바 있으나 유도과정이 길어 세부내용은 생략하고 그 추정식의 결과 제시하면 Eq. (7)과 같다.

$$L(\omega,\alpha,\lambda)=\left[\omega+(1-\omega)e^{-\lambda}\right]^{n_0}\prod_{i\ge1}\left[(1-\omega)\frac{(1+\alpha i)^{i-1}}{i!}\frac{(\lambda e^{-\alpha\lambda})^i}{e^{\lambda}}\right]$$
$$=\left[\omega+(1-\omega)e^{-\lambda}\right]^{n_0}(1-\omega)^{n-n_0}\lambda^{n\overline{N}}e^{-(n-n_0)\lambda}e^{-n\overline{N}\alpha\lambda}\prod_{i\ge1}\frac{(1+\alpha i)^{(i-1)n_i}}{(i!)^{n_i}} \quad (7)$$

여기서, n_0는 표본 n에 포함된 자료 중 $x=0$에 해당되는 자료의 개수이고 $\sum n_i=n-n_0$로부터 $\sum in_i=n\overline{N}$으로 정의된다. Eq. (7)을 3개의 모수에 대해 각각 미분하고 0이 될 때의 값을 수치적으로 반복하여 계산하면 최종적으로 3개 모수의 추정값을 구할 수 있다.

3. Bayesian ZIGP 모형의 구축

3.1 Bayesian ZIGP 모형과 Metropolis-Hastings 알고리즘

앞서 제시한 Poisson, GPD, ZIP, 및 ZIGP 모형에 포함되어진 모수의 추정은 적률추정법이나 최대우도추정법과 같은 frequentist적인 방법에 근거하여 모수를 추정할 수도 있으며, Bayesian 방법을 활용하여 관련 모수를 추정할 수도 있다. 두 가지의 접근방식 중 어떤 방법이 절대적으로 우수하다고 결론내리기에는 어려움이 있으나 일반적으로 모수의 수가 많은 경우, 우도함수(likelihood function)가 복잡하여 모수의 전체 대상영역에서의 최대값을 효과적으로 탐색하기 어려운 경우, 우도함수에 미분불가능한 점들이 존재하는 경우에는 frequentist적인 추정방법보다는 Bayesian 추정방법을 활용하는 편이 우수한 결과를 도출할 수 있으며, 불확실성의 감소 측면에서도 보다 효율적인 추정방법이라는 연구결과가 제시된 바 있다(Lee and Kim 2008; K6im and Lee 2010; Kim et al. 2013; Lee et al. 2014).

Bayesian 모수 추정 절차는 Bayes의 정리를 연속 확률밀도함수에 대해 적용한 Eq. (8)을 이용하여 궁극적으로 특정 표본 $x_1, x_2, ..., x_n$이 발생되었다는 조건 하에서 발생 가능한 특정 모수 $\boldsymbol{\theta}$를 탐색하는 과정이라 할 수 있다. 즉 특정표본이 발생되었다는 조건을 활용하기 때문에 표본의 추가적인 발생 등에 따르는 모수의 변화를 합리적으로 추정절차에 반영할 수 있다는 장점이 있으며, 이로 인하여 모수의 불확실성을 frequentist적인 적률추정법이나 최대우도추정법보다 현실적으로 반영할 수 있다.

$$\pi(\boldsymbol{\theta}|x_1, x_2, ..., x_n) = \frac{f(x_1|\boldsymbol{\theta}) \cdots f(x_n|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int_{\theta} f(x_1|\boldsymbol{\theta}) \cdots f(x_n|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}} \quad (8)$$

여기서, 좌변의 $\pi(\boldsymbol{\theta}|x_1, x_2, ..., x_n)$는 사후분포(posterior distribution), 우변 분자의 $\pi(\boldsymbol{\theta})$는 사전분포(prior distribution)라 명명되며, 우변의 분모는 상수로서 주변분포(marginal distribution)이고, 우변의 분자의 $f(x_1|\boldsymbol{\theta}) \cdots f(x_n|\boldsymbol{\theta})$는 발생할 수 있는 모든 가능성을 고려한 우도함수(likelihood function)이다. 본 연구에서 활용된 모형인 Bayesian ZIGP 모형은 Eq. (1)의 우변에 사용된 우도함수에 ZIGP 모형의 우도함수인 Eq. (7)을 사용한 것으로 추정되어야 할 모수, $\boldsymbol{\theta} = [\alpha, \lambda, \omega]$이다.

Bayesian 추정에서 가장 중요한 인자는 우변 분자에 포함된 사전분포의 구성으로 사전분포는 모형을 구성하는 사람 또는 이와 관련된 특정 자료의 주관적 또는 경험적 정보를 제공하는 역할을 제공한다. 그러나 과거 컴퓨터의 하드웨어 및 소프트웨어가 발전하지 못했던 시절의 주관적 또는 경험적 자료를 기반으로 작성된 사전분포는 공액사전분포(conjugated prior distribution)라는 해석적(analytical)으로 계산될 수 있는 몇몇의 사전분포를 제외하고서는 계산자체가 불가능하였다. 따라서 이 시기 Bayesian 추정을 이용한 계산은 대부분 사전분포를 인위적으로 조절하여 우도함수와의 관계를 통해 Eq. (8)이 해석적으로 산정 가능한 경우에 대해서만 적용되었고, 이는 Bayesian 추정방법이 가지는 사전분포의 원래적 의미와 모순되는 경우가 많아 많은 비판을 받으며 실제로 발생되는 자연과학이나 사회과학의 현상을 설명하기에는 매우 부족한 방법으로 평가되었었다.

그러나 20세기에 들어서면서 하드웨어의 발달과 함께 각종 탐색용 알고리즘과 같은 소프트웨어가 발달되면서 Bayesian 추정방법이 가지고 있던 계산에 대한 어려움이 급작스럽게 해소되었고 이로 인해 통계학(특히 계산통계학)을 비롯하여 의학, 생물학, 물리학 등의 자연과학과 공학의 분야에서도 상당히 많은 연구가 이미 진행되었다. Malakoff (1999)가 236년 이전의 구식 접근방식이 이제야 돌아와 의학부터 낚시까지에 이르는 많은 실생활의 자료를 분석하고 있다는 학술기사를 발표한 이래 Bayesian 추정방법에 대한 연구는 더욱 많이 활성화되어 최근에는 수자원공학 분야에서도 이 방법을 활용한 다양한 연구 및 응용사례가 진행된 바 있다(Kavetski et al. 2006; Seidou et al. 2006; Lee and Kim 2008; Kim and Lee 2010; Chung and Kim 2013; Kim et al. 2013; Lee et al. 2014).

본 논문의 다음 절에서 설명할 사전분포의 구성에 있어서도 경험적 자료를 이용하여 상당히 복잡한 사전분포가 구성되기 때문에 Eq. (1)로부터 해석적으로 특정 모수 $\boldsymbol{\theta}$를 직접 추출하는 것은 수학적으로는 불가능하다. 그러므로 본 연구에서는 유도된 사후분포로부터 모수를 추정하기 위하여 Bayesian MCMC (Markov Chain Monte Carlo)방법을 사용하였다. Bayesian MCMC 기법이란 Markov 연쇄(Markov chain)를 이용하여 모수간의 관계를 구성하고 이를 상당히 큰 수만큼 반복하는 몬테카를로 적분(Monte Carlo integration)을 이용하여 최종적으로 모수의 통계적 특성을 산정하는 방법이다. 특히 기존의 MCMC 기법을 Bayesian 추정방법으로 유도한 사후분포에 적용할 때 이를 특별히 Bayesian MCMC 기법이라 명명하고 있으며, 이 기법에 따

른 여러 가지의 알고리즘이 있다. 여러 가지 Bayesian 계산 방법 중에서 가장 활발히 사용되고 있는 알고리즘은 Metropolis-Hastings 알고리즘(Metropolis et al, .1953)으로 최근 개선되어져 많은 연구에서 활발히 이용되고 있다. Metropolis-Hastings 알고리즘의 이론적 내용에 대해서는 이미 많은 논문에서 다룬바 있으므로 본 논문에서는 생략하고 알고리즘의 구성에 있어 가장 중요한 채택확률 산정식인 Eq. (9)만을 나타내었다.

$$\rho(\boldsymbol{\theta}_j, \boldsymbol{\theta}_{j+1}) = \min\left[\frac{q(\boldsymbol{\theta}_j|\boldsymbol{\theta}_{j+1})\pi(\boldsymbol{\theta}_{j+1}|\boldsymbol{D})}{q(\boldsymbol{\theta}_{j+1}|\boldsymbol{\theta}_j)\pi(\boldsymbol{\theta}_j|\boldsymbol{D})}, 1 \right] \quad (9)$$

여기서 $q(\bullet)$를 제안분포, $\pi(\bullet)$는 사전분포, $\boldsymbol{D}$는 추출된 특정 표본벡터, ρ는 채택확률(acceptance probability)라고 하며, 채택확률을 이용한 알고리즘의 작동절차는 다음과 같다.

Step 1) $j=0$에서의 임의의 $\boldsymbol{\theta}_0$를 선정한다.

Step 2) 제안분포로부터 제안 모수 $\boldsymbol{\theta}_*$를 무작위적으로 d

Step 3) Eq. (5)로부터 채택확률을 계산한다.

Step 4) 0과 1사이의 균일분포로부터 무작위수 u를 생성한다.

Step 5) 만약 $u < \rho$이면, $\boldsymbol{\theta}_*$를 $\boldsymbol{\theta}_{j+1}$로 교체하고 반대의 경우에는 $\boldsymbol{\theta}_j$를 $\boldsymbol{\theta}_{j+1}$로 교체한다.

(즉 $u > \rho$인 경우에는 제안분포로부터 생성된 $\boldsymbol{\theta}_*$를 사용하지 않는다.)

Step 6) j를 1 증가시키고 Step 2)로 돌아간다.

Step 7) 충분히 큰 수만큼 위 과정을 반복한다.

Step 8) 발생된 모수를 이용하여 평균값을 계산한다.

3.2 Bayesian ZIGP 모형 구성을 위한 경험적 사전분포의 구축

앞의 3.1절에서 언급한 바와 같이 Bayesian 방법을 이용한 모수 추정 절차에서 가장 핵심이 되는 사항은 사전분포의 구성방법 및 절차이다. 따라서 사전분포를 구성하기 위한 다양한 방법들이 제시되고 있으며 적절한 사전분포의 구성은 절대적으로 좋은 방법이 존재하기 보다는 모형화하는 상황, 표본추출의 상황, 가지고 있는 정보의 상황 등에 영향을 받는다. 이론적으로만 보았을 때, 사전분포는 '정보적 사전분포(informative prior distribution)'과 '무정보적 사전분포(non-informative prior distribution)'의 두 가지 형태로 구분된다. 정보적 사전분포는 다양한 자료를 객관적으로 분석하고, 분석자의 주관적 견해를 통해 채택되는 사전분포이다. 이와 달리 무정보적 사전분포는 수학적으로 입증된 분포를 채택하는 경우로 앞서 언급한 공액사전분포(conjugate prior distribution)를 주로 사용하게 된다.

어떤 방식의 사전분포를 선정할 지를 결정하기 위해서는 가지고 있는 정보의 특성을 먼저 알아볼 필요가 있다. 만약 추정되어져야 할 모수에 대한 정보가 사전에 상당히 밝혀져 있거나 모수의 추정에 사용할 수 있는 자료의 길이가 충분히 길면, 사전분포가 사후분포에 미치는 영향은 미미해 계산이 간단해 질 수 있는 공액사전분포를 그대로 사용한다 해도 사실상 최종결과에는 큰 영향을 미치지 않는다. 그러나 그렇지 못한 경우에는 사전분포가 사후분포에 미치는 영향이 증대되므로, 무정보적 사전분포보다는 정보적 사전분포를 사용하는 것이 최종 결과에 많은 영향을 미치게 되므로 어떤 사전분포를 채택할 지에 관한 문제는 중요한 문제로 대두된다.

본 연구에서 다루고자 하는 집중호우 월별개수는 자료의 길이가 짧고 강우관측소의 개수도 제한적이므로 무정보적 사전분포를 그대로 사용하기에는 무리가 있다고 판단되었으며 따라서 정보적 사전분포를 구축하여 Bayesian MCMC 기법을 수행하였다. 정보적 사전분포를 구축하는 세부적인 기법은 다시 2-단계 Bayes 기법(Kaplan, 1985), 경험적 Bayes 기법(Berger, 1985; Carlin and Louis, 1996; Maritz and Lwin, 1989), 최대 엔트로피 기법(Bertucio and Julius, 1990; Wheeler, 1993)으로 구분된다. 본 연구에서는 여러 가지의 기법 중 경험적 Bayes 기법에 따라 사전분포를 구축하였는데, 이 기법은 에르고딕(ergodic) 가정 하에 분석하고자 하는 지점의 특성을 나타내는 모수를 추정하기 위하여 인근 지역의 자료가 통계적 동질성(homogeneous)이 있다는 가정을 사용한다. 즉 분석대상 지점 인근의 자료에 대한 확률밀도함수를 선정하고 선정된 확률밀도함수의 모수를 각각 최우추정법을 사용하여 추정한 이후, 추정된 모수의 값들을 다시 특정 확률밀도함수를 따르는 것으로 간주하여 이 확률밀도함수를 사전분포로 선정한다. 따라서 경험적 Bayes 기법에 의한 정보적 사전분포를 구축하기 위해서는 먼저 분석하고자하는 강우관측소들을 동질성있는 그룹들로 구성할 필요가 있다. 먼저 낙동강 유역에 위치하고 있는 도시 중에서 집중호우의 발생으로 인해 많은 자산 또는 인명의 피해가 발생될 수 있는 대도시에 위치하고 있는 대구와 부산 강우관측소에 대한 집중호우 월별개수를 분석하기로 하였으며, 이 두 관측소에 대한 경험적 정보를 수집하기 위하여 낙동강 유역 내 10개의 강우관측소를 추가로 선정하였다.

Fig. 1은 분석대상 강우관측소인 대구 및 부산 강우관측소와 사전분포 구축을 위해 선정된 10개의 추가적인 강우관측소를 나타낸 것이고, Table 1은 선정된 강우관측소에 대한 기초적인 정보를 나타낸 것이다. Fig. 1에서 ① 대구, ② 부산, ③ 영주, ④ 문경, ⑤ 의성, ⑥ 구미, ⑦ 영천, ⑧ 거창, ⑨ 합천, ⑩ 밀양, ⑪ 포항, ⑫ 울산 강우 관측소를 나타내며, Table 1에서 384개월 (32년, 1983년~2014년) 중 80 mm/day 이상의 집중호우가 월 1회 이상 발생했던 개월의 수를 보면 영천이 가장 적은 56개였고, 거창이 가장 많은 129개였음을 알 수 있다. 따라서 집중호우가 발생하지 않았던 즉 count data가 0인 횟수는 최소가 255개로 약 66%를 차지하고 최대가 328개로 약 85%를 차지하고 있음을 알 수 있다. 특히 본 연구에서의 분석대상 지점인 부산은 0의 개수가 275개, 대구는 309개로 전체 384개의 자료 중 약 71.6 % 및 80.5 %를 차지하고 있어 0-과잉 모형을 고려할 필요가 있는 것으로 잠정적으로 판단되었다. 특히 Table 2에는 집중호우의 발생횟수별 월의 개수를 나타내었는데, 부산, 영주, 문경, 구미, 거창, 밀양, 포항의 7개 관측소에서는 1개월간 6번의 집중호우가 발생했던 횟수가 1회였던 것으로 나타나 집중호우의 발생빈도가 타 관측소보다 상대적으로 높은 것을 알 수 있었다.

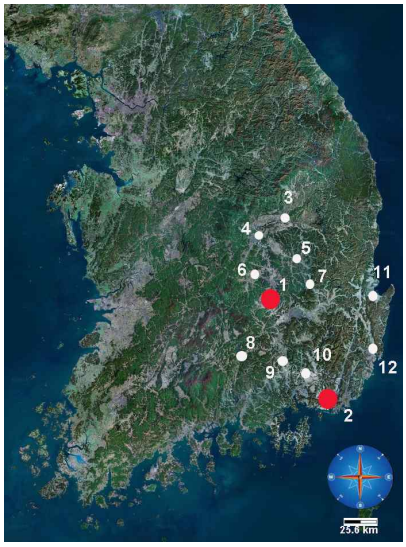


Fig. 1. Selected 12 rainfall gauges

Table 1. Information of the selected 12 rainfall gauges

Rainfall gauges	Latitude	Longitude	Elevation (EL,m)	Annual mean rainfall (32 years, mm)	The number of torrential rainfall (384 months)	Number of zeros (384 months)
Daegu	35.885	128.619	64.1	1,087	75	309
Busan	35.105	129.032	69.6	1,344	109	275
Yeongju	36.872	128.517	210.8	1,125	89	295
Mungyeong	36.627	128.149	170.6	1,095	87	297
Uiseong	36.356	128.689	81.8	1,039	73	311
Gumi	36.131	128.321	48.9	1,104	73	311
Yeongcheon	35.977	128.951	93.6	1,074	56	328
Geochang	35.671	127.911	221.0	1,334	129	255
Hapcheon	35.565	128.170	33.1	1,317	124	260
Milyang	35.491	128.744	11.2	1,249	117	267
Pohang	36.033	129.380	2.3	1,274	103	281
Ulsan	35.560	129.320	34.6	1,287	107	277

Table 2. Frequency table of the occurrence of torrential rainfalls

Rainfall gauges	0	1	2	3	4	5	6	Total
Daegu	309	53	18	2	2	0	0	384
Busan	275	67	32	5	4	0	1	384
Yeongju	295	60	25	1	1	1	1	384
Mungyeong	297	57	22	3	2	2	1	384
Uiseong	324	42	14	3	1	0	0	384
Gumi	311	50	15	4	2	1	1	384
Yeongcheon	328	49	6	1	0	0	0	384
Geochang	255	80	35	8	4	1	1	384
Hapcheon	260	87	32	4	1	0	0	384
Milyang	267	81	27	4	3	1	1	384
Pohang	281	69	29	2	2	0	1	384
Ulsan	277	78	24	2	2	1	0	384

낙동강 유역에서 추가로 선정된 10개의 강우관측소를 이용하여 대구와 부산 강우관측소에 대한 정보적 사전분포를 구축하기 위해서는 먼저 10개의 강우관측소를 대구와 부산의 통계적 특성과 가까운 강우관측소끼리 그룹을 구성해야 한다. 본 연구에서는 통계적 동질성을 가지는 두 개의 그룹을 구성하기 위하여 12개 관측소에 대한 연평균강우량과 1회 이상 집중호우 발생횟수를 변수로 선정하고 K-means 알고리즘을 이용한 군집분석(cluster analysis)를 수행하였다. Fig. 2는 12개 관측소를 대상으로 수행한 군집분석 결과를 나타낸다. 분석결과 파랑색 원으로 표시된 대구, 영주, 문경, 의성, 구미, 영천 강우관측소(그룹 1)와 빨강색 사각형으로 표시된 부산, 거창, 합천, 밀양, 포항, 울산 강우관측소(그룹 2)로 구분되어 대구와 동질한 5개의 강우관측소 및 부산과 동질한 5개의 강우관측소를 선정할 수 있었다.

앞서 수행한 군집분석의 결과를 이용하여 그룹 1과 그룹 2에 대한 각각의 사전분포를 구축하였다. 본 연구에서 사용하고자 하는 Bayesian ZIGP 모형에 포함되어진 모수 $\boldsymbol{\theta} = [\alpha, \lambda, \omega]$이므로 그룹 1 및 2에 대한 α, λ, ω를 나타낼 수 있는 각각의 사전분포 3개가 필요하다. 이를 위하여 먼저 식 (7)에서 제안된 ZIGP 모형의 우도함수에 최대우도함수법을 적용하여 그룹 A, B로 구분된 각각의 강우관측소에 대한 α, λ, ω의 추정치 $\hat{\alpha}, \hat{\lambda}, \hat{\omega}$를 산정하여 Table 3에 나타내었다.

경험적 Bayes 기법은 표 3에서 추정된 α, λ, ω를 각각 분리하여 개별 모수의 추정치를 가장 잘 나타내는 확률분포와 그 때의 모수를 사용하는 데 이를 초모수(hyper-parameter)라 한다. 예를 들어 그룹 1의 α에 대한 초모수를 산정하기 위해서는 먼저 대구, 영주, 문경, 의성, 구미, 영천 강우관측소의 6개 지점에 대한 추정치 $\hat{\alpha}$ (Table 3에서 0.023992부터 0.052796까지 6개)에 대한 적정 확률분포를 찾고 그 때의 모수를 최대우도법으로 찾으면 된다. 이같은 과정을 수행함에 있어 적정 확률분포를 찾기 위해서 본 연구에서는 χ^2-테스트, Kolmogorov-Smirnov 테스트, Cramer Von 테스트 및 Probability Plot Correlation Coefficient (PPCC) 테스트를 그룹 1과 2에 대해 수행하였으며 그 결과를 Table 4와 Table 5에 나타내었다. 먼저 그룹 1에 대한 분석결과인 Table 4에서 α는 2모수 Log-normal 분포, λ와 ω는 2모수 Weibull 분포가 가장 적절한 것으로 나타났으며, 그룹 2에 대한 분석결과인 Table 5에서 α와 λ는 2모수 Weibull 분포, ω는 2모수 Log-normal 분포가 가장 적절한 것으로 나타났다.

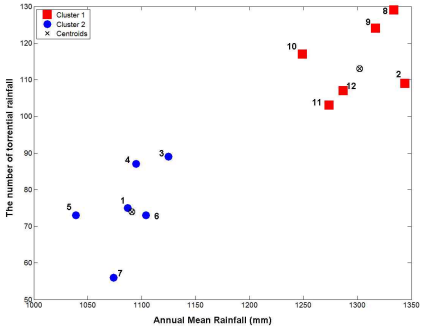


Fig. 2. Classification of homogeneous two groups

Table 3. Estimates of ZIGP model to construct prior distribution

Group 1	$\hat{\alpha}$	$\hat{\lambda}$	$\hat{\omega}$
Daegu	0.023992	0.675029	0.614373
Yeongju	0.066584	0.897875	0.626508
Mungyeong	0.117776	1.140179	0.688344
Uiseong	0.038740	0.674795	0.678778
Gumi	0.127569	1.129153	0.737419
Yeongcheon	0.052796	0.257250	0.395497
Group 2	$\hat{\alpha}$	$\hat{\lambda}$	$\hat{\omega}$
Busan	0.066026	0.999810	0.579964
Geochang	0.057089	1.050627	0.501002
Hapcheon	0.073083	0.574802	0.261814
Milyang	0.049425	0.940706	0.529185
Pohang	0.032599	0.819537	0.534272
Ulsan	0.136015	0.693241	0.461623

Table 4. Selection of probability distribution for calculation of hyper-parameter: Group 1

Distribution	χ^2 test	Decision	K-S test	Decision	CVM test	Decision	PPCC test cal.	PPCC test Table	Decision
				α					
NOR	5.67	rejected	0.25	accepted	0.18	accepted	0.91	0.92	rejected
LN(2)	***3.00***	***accepted***	***0.20***	***accepted***	***0.13***	***accepted***	***0.93***	***0.92***	***accepted***
GEV	5.75	rejected	0.25	accepted	0.18	accepted	0.92	0.92	rejected
WBU(2)	3.00	accepted	0.25	accepted	0.17	accepted	0.64	0.65	rejected
				λ					
NOR	3.33	accepted	0.25	accepted	0.15	accepted	0.90	0.92	rejected
LN(2)	1.67	accepted	0.25	accepted	0.21	accepted	0.90	0.92	rejected
GEV	6.08	rejected	0.23	accepted	0.15	accepted	0.92	0.00	rejected
WBU(2)	***3.33***	***accepted***	***0.25***	***accepted***	***0.17***	***accepted***	***0.89***	***0.65***	***accepted***
				ω					
NOR	1.00	accepted	0.25	accepted	0.18	accepted	0.89	0.92	rejected
LN(2)	1.00	accepted	0.25	accepted	0.23	accepted	0.89	0.93	rejected
GEV	6.00	rejected	0.24	accepted	0.16	accepted	0.91	0.00	rejected
WBU(2)	***1.00***	***accepted***	***0.26***	***accepted***	***0.17***	***accepted***	***0.89***	***0.65***	***accepted***

Table 5. Selection of probability distribution for calculation of hyper-parameter: Group 2

Distribution	χ^2 test	Decision	K-S test	Decision	CVM test	Decision	PPCC test		Decision
							cal.	Table	
				α					
NOR	1.00	accepted	0.20	accepted	0.16	accepted	0.90	0.92	rejected
LN(2)	1.33	accepted	0.23	accepted	0.13	accepted	0.90	0.93	rejected
GEV	7.00	rejected	0.17	accepted	0.13	accepted	0.90	0.91	rejected
WBU(2)	***2.00***	***accepted***	***0.16***	***accepted***	***0.11***	***accepted***	***0.91***	***0.65***	***accepted***
				λ					
NOR	1.67	accepted	0.20	accepted	0.09	accepted	0.90	0.92	rejected
LN(2)	1.67	accepted	0.19	accepted	0.08	accepted	0.91	0.93	rejected
GEV	8.42	rejected	0.18	accepted	0.08	accepted	0.90	0.91	rejected
WBU(2)	***1.67***	***accepted***	***0.23***	***accepted***	***0.11***	***accepted***	***0.95***	***0.65***	***accepted***
				ω					
NOR	6.00	rejected	0.21	accepted	0.19	accepted	0.85	0.92	rejected
LN(2)	***2.12***	***accepted***	***0.23***	***accepted***	***0.23***	***accepted***	***0.93***	***0.92***	***accepted***
GEV	14.08	rejected	0.27	accepted	0.12	accepted	0.92	0.93	rejected
WBU(2)	-	rejected	-	rejected	-	rejected	-	-	rejected

Table 6. Results of hyper-parameters in Bayesian ZIGP model

Group 1			
Parameters	Selected distribution	Hyper-parameter	Hyper-parameter
α	LN 2	$\hat{\mu}_{\alpha}^{1}$= -2.653	$\hat{\sigma}_{\alpha}^{1}$= 0.542
λ	WBU 2	$\hat{\kappa}_{\lambda}^{1}$= 0.842	$\hat{\beta}_{\lambda}^{1}$= 2.080
ω	WBU 2	$\hat{\kappa}_{\omega}^{1}$= 0.651	$\hat{\beta}_{\omega}^{1}$= 5.139
Group 2			
Parameters	Selected distribution	Hyper-parameter	Hyper-parameter
α	WBU 2	$\hat{k}_{\alpha}^{2}$= 0.086	$\hat{\beta}_{\alpha}^{2}$= 1.724
λ	WBU 2	$\hat{\kappa}_{\lambda}^{2}$= 0.859	$\hat{\beta}_{\lambda}^{2}$= 6.634
ω	LN 2	$\hat{\mu}_{\omega}^{2}$= -0.731	$\hat{\sigma}_{\omega}^{2}$= 0.166

Table 4와 Table 5를 통해 결정된 적정 확률분포의 선정과정에서는 최대우도법을 이용하여 해당 분포의 모수를 추정하게 되며 이 값이 초모수로 사용되는데, 각각의 그룹에 대한 결과를 정리하여 Table 6에 나타내었다. 예를 들어 그룹 1에 대한 Bayesian ZIGP model의 모수 α는 2모수 Log-normal 분포를 따르며, 이 때 Bayesian ZIGP 모형의 초모수로 사용된 Log-normal 분포의 평균 $\hat{\mu}_{\alpha}^{1}$=-2.653이고 편차 $\hat{\sigma}_{\alpha}^{1}$=0.542로 사용될 수 있음을 알 수 있다. 그러므로 앞의 과정을 통해서 최종적으로 결정된 경험적 Bayes 기법에 의한 그룹 1 및 그룹 2의 사전분포는 각각 식 (10)과 식 (11)로 나타낼 수 있으며, 이 사전분포를 Bayesian 모수추정 기법을 나타내는 식 (8)에 적용하면 그룹 1과 그룹 2에 대한 각각의 최종적인 모수를 추정할 수 있다. 단, 최종 사전분포를 구축함에 있어 각각의 모수 α, λ, ω는 서로 통계적으로 독립이라는 가정이 사용되었다.

그룹 1:

$$\pi(\alpha,\lambda,\omega) = LN2\,(\hat{\mu}_{\alpha}^{1}, \hat{\sigma}_{\alpha}^{1}) \times WBU2\,(\hat{\kappa}_{\lambda}^{1}, \hat{\beta}_{\lambda}^{1}) \times WBU2\,(\hat{\kappa}_{\omega}^{1}, \hat{\beta}_{\omega}^{1}) \quad (10)$$

그룹 2:

$$\pi(\alpha,\lambda,\omega) = WBU2\,(\hat{\kappa}_{\alpha}^{2}, \hat{\beta}_{\alpha}^{2}) \times WBU2\,(\hat{\kappa}_{\lambda}^{2}, \hat{\beta}_{\lambda}^{2}) \times LN2\,(\hat{\mu}_{\omega}^{2}, \hat{\sigma}_{\omega}^{2}) \quad (11)$$

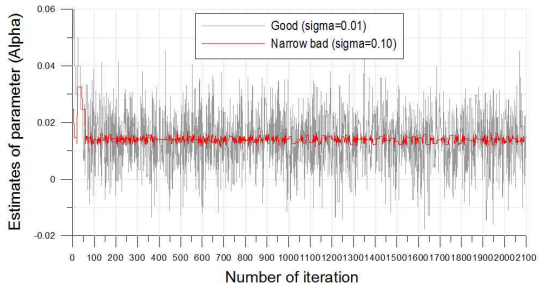
4. Metropolis-Hastings 알고리즘의 수행 및 수렴 판정

Metropolis-Hastings 알고리즘은 특정 확률분포로부터 수학적으로나 이론적으로 직접적 표본의 추출이 어려운 경우에도 효율적이고 안정적으로 무작위적인 표본을 추출하는 대표적인 MCMC (Markov Chain Monter Carlo)기반의 알고리즘으로 최근 Bayesian 사후분포로부터 난수를 추출함에 있어 가장 많이 사용되고 있다 (Bate and Campbell, 2011; Chib and Greenberg, 1995; Marshall et al., 2004). 이 알고리즘을 이용하여 안정적인 무작위 표본을 추출하기 위해서는 앞서 제시한 식 (9)인 채택확률을 산정하기 위한 제안분포 $q(\bullet)$의 선정 및 제안분포의 모수 설정에 매우 신중해야 하며, 안정적인 표

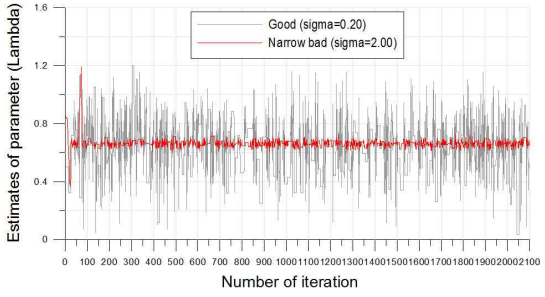
본의 추출에 대하여 반드시 그 결과를 확인할 필요가 있다.

제안분포 $q(\bullet)$의 선정에 있어서 Chib and Greenberg (1995)는 확률분포의 특성에 따라 효율적인 표본추출이 가능할 수 있는 여러 형태의 제안분포를 제안한 바 있으며, 본 연구에서는 이 중 가장 간단한 대칭분포의 형태인 정규분포 $N(0,\sigma^2)$.를 제안분포로 사용하였다. 선정된 정규분포를 제안분포로 사용함에 있어서 결정되지 않은 모수인 표준편차 σ를 어떤 값으로 사용하는 지에 따라 Metropolis-Hastings알고리즘의 표본추출 성능이 증가되거나 저하될 수 있다. 즉 제안분포의 모수 중 σ가 너무 작게 선정되면 알고리즘 진행에 있어 후보로 제안된 많은 수의 표본이 채택되므로 표본의 변화가 발생되지 않게 되고 궁극적으로 전체 탐색 영역을 반영하지 못하는 비혼합 표본 (not mixed sample)이 추출되며, 반대로 σ가 지나치게 크면 후보로 제안된 많은 수의 표본이 탈락하여 대부분의 표본이 각각 다른 새로운 표본으로만 추출되는 비혼합 표본이 추출된다. 따라서 최적의 σ를 산정하기 위해서는 다양한 최적화 기법 (optimization technique)들을 사용하는 적응형 MCMC (adaptive MCMCM) 기법을 사용할 수 있으나 본 연구에서는 이를 수동적인 (manual) 시행오차 방법을 사용하여 안정적인 무작위 표본이 추출될 때까지 반복하여 최종적인 σ를 결정하였다. 향후에는 적응형 MCMC기법을 적용함으로써 보다 안정적인 표본을 얻을 수 있으리라 판단된다.

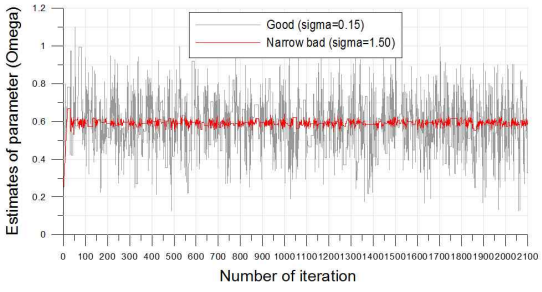
앞서 제시한 비혼합 표본추출의 문제는 알고리즘의 최종결과에 매우 큰 영향을 미치므로 MCMC기반의 모든 알고리즘은 안정적인 표본 추출에 대해 반드시 확인절차를 수행하여야 한다. 알고리즘의 수렴 (convergence) 판정은 크게 도시적인 방법 (plotting method), 채택확률값의 판정 및 특정 지표 산정의 세 가지 방법을 통해 수행될 수 있다. 먼저 도시적인 방법은 Fig. 3의 (a) ~ (c)와 같이 추출된 분포를 비교하여 추출된 표본의 안정성을 정성적으로 판단하는 방법이다. Fig. 3의 (a) ~ (c)는 대구 강우관측소 자료에 대해서 식 (8)의 Bayesian 사후분포로부터 α, λ, ω를 최종 추출한 결과이다. 각각의 모수 $\sigma_\alpha, \sigma_\lambda, \sigma_\omega$가 0.1, 2.0, 1.5로 선정된 표본추출 결과는 붉은색으로 나타내었으며, 이 경우 σ가 너무 크게 선정되어 탐색영역이 특성 표본 인근에 집중된 것을 알 수 있으며, $\sigma_\alpha, \sigma_\lambda, \sigma_\omega$가 0.01, 0.2, 0.15로 선정된 경우에는 탐색 범위도 확대되고 표본의 안정적 유지도 효율적인 것으로 보인다. 도시적인 방법은 정량적으로 알고리즘의 효율성을 판정할 수 없는 단점이 있으나 채택확률을 활용하면 이러한 문제를 일부 극복할 수 있다. Roberts et al. (1994)는 적정 채택확률로 0.45를 제안한 바 있으며, Gamerman (1997)은 이를 0.2~0.5로 제안한 바 있는데, $\sigma_\alpha, \sigma_\lambda, \sigma_\omega$가 0.1, 2.0, 1.5로 선정된 경우의 채택확률은 0.055로 매우 부적절하였던 반면, $\sigma_\alpha, \sigma_\lambda, \sigma_\omega$가 0.01, 0.2, 0.15로 선정된 경우의 채택확률은 0.385로 안정적인 표본으로 보기에 적절한 것으로 판단된다. 단 이때 사용된 표본은 2,100개 중 초기 추출 표본 100개는 제외한 나머지 2,000개만을 대상으로 모든 계산을 수행하였다.



(a)



(b)



(c)

Fig. 3. Trace plots of good mixed sample at Daegu rainfall gauge: (a) α, (b) λ, (c) ω

앞서 제시한 도시적 방법과 채택확률을 이용한 방법은 추출된 표본의 효율성을 쉽고 간단히 판정할 수 있다는 장점이 있으나 도시적인 방법은 정성적인 측면이라는 점에서 채택확률의 이용은 추출된 표본 전체를 사용함으로써 국부적인 표본 추출의 효율성을 판정할 수 없다는 점에서 단점이 있다. 이러

한 단점을 개선하기 위하여 Gelman & Rubin 지표 (Gelman and Rubin, 1992), Raftery & Lewis 지표 (Raftery and Lewis, 1992), Geweke 지표 (1992)가 활용될 수 있다. 이 지표들은 모두 모수적 또는 비모수적 통계적 가설검정 (statistical hypothesis test)을 기본적 이론으로 하여 개발된 것으로 Gelman & Rubin 지표는 1에 가까울수록, Raftery & Lewis 지표는 5를 넘지 않을수록, Geweke 지표는 95% 신뢰구간 내부에 존재하여야 안정적인 표본이 추출이 완료되었음을 의미한다. Table 7은 대구 강우관측소에 대한 각각의 $\sigma_\alpha, \sigma_\lambda, \sigma_\omega$ 에 대한 3가지 지표의 산정결과 및 최종적인 표본추출 효율성의 판정결과를 나타낸 것이다. Table 7을 보면 $\sigma_\alpha, \sigma_\lambda, \sigma_\omega$ 가 0.01, 0.2, 0.15로 산정된 경우의 모든 지표가 안정적인 표본추출 결과를 나타내는 것을 알 수 있으며, 부산 강우 관측소의 경우도 대구 강우 관측소와 동일한 방법을 사용하여 적절한 사후분포의 모수를 추출하였다.

Table 7. Results of 3 quantitative methods and decision of algorithm convergence

Cases	Gelman & Rubin	Raftery & Lewis	Geweke	Decision
σ_α =0.10	1.2584	5.84	-2.125	Poor
σ_α =0.01	1.0104	1.05	-1.654	Good
σ_λ =2.00	1.3548	4.98	-2.001	Poor
σ_λ =0.20	1.1120	2.01	-1.451	Good
σ_ω =1.50	1.1586	4.48	-2.748	Poor
σ_ω =0.15	1.0001	1.05	-1.889	Good

5. 모형별 모의결과의 비교 및 성능분석

대구 및 부산 강우관측소에서의 집중호우 월별개수를 확률적으로 모형화하기 위하여 5개의 확률분포모형인 Poisson 확률분포(POI), 일반화 Poisson 확률분포(Generalized Poisson Distribution, GPD), 0-과잉 Poisson 확률분포 (Zero-Inflated Poisson distribution, ZIP), 0-과잉 일반화 Poisson 확률분포(Zero-Inflated Generalized Poisson distribution, ZIGP), Bayesian 0-과잉 일반화 Poisson 확률분포(Bayesian Zero-Inflated Generaized Poission distribution), Bayesian ZIGP)가 사용되었다. Table 8은 POI, GPD, ZIP, ZIGP, Bayesian ZIGP 모형의 모수 추정값을 나타낸 것으로 모든 모형에서 존재하고 있는 모수 λ에 대한 추정값을 보면 대구와 부산 강우관측소 모두에서 0-과잉 모형이 사용된 ZIP, ZIGP, Bayesian ZIGP 모형이 POI, GPD보다 더 큰 값으로 추정되었음을 알 수 있다.

Table 8. Results of estimated parameters in 5 different probability models

Distributions	POI	GPD	ZIP	ZIGP	Bayesian ZIGP
Daegu					
Estimated α	-	0.1640	-	0.0240	0.0140
Estimated λ	0.1198	0.2242	0.6990	0.6750	0.6613
Estimated ω	-	-	0.6163	0.6144	0.5981
Busan					
Estimated α	-	0.2060	-	0.0660	0.0100
Estimated λ	0.2708	0.3474	1.0238	0.9998	0.9308
Estimated ω	-	-	0.5727	0.5800	0.5301

Table 9와 Fig. 4는 5개의 모형들을 이용하여 모의한 대구 및 부산 강우관측소에서의 집중호우 월별개수와 과거 384개월 간 실제 관측된 집중호우 월별개수를 비교한 것이다. 먼저 0-과잉 현상을 처리할 수 없는 구조를 가지고 있는 POI 및 GPD모형은 두 강우관측소에서 관측값보다 훨씬 적은 0회 집중호우 월별개수를 산정하였으며, 이러한 확률밀도의 왜곡은 1회 집중호우 월별개수를 증가시키는 현상을 발생시키게 됨을 알 수 있었다. 또한 2회 집중호우 월별개수에 있어서는 두 모형 모두 관측값보다 적은 횟수를 모의하는 특징을 보였으나 3회 이상에서는 GPD모형은 오히려 큰 회수를 모의하는 등 특별한 모의특성을 확인하기는 어려웠다.

Table 9. Comparative results between 5 different probability models

Models	Observed	POI	GPD	ZIP	ZIGP	Bayesian ZIGP
Daegu						
0	309	294	296	310	311	309
1	53	79	72	51	50	52
2	18	11	14	18	18	18
3	2	1	1	4	4	4
4	2	0	1	1	1	1
5	0	0	0	0	0	0
6	0	0	0	0	0	0
Total	384	384	384	384	384	384
Busan						
0	275	248	250	279	290	275
1	67	108	99	60	52	67
2	32	24	25	31	27	31
3	5	4	8	11	10	7
4	4	0	3	3	4	3
5	0	0	1	1	1	1
6	1	0	0	0	0	0
Total	384	384	384	384	384	384

POI 및 GPD 모형이 0 및 1회 집중호우 월별개수의 모의에 실패한 반면, 0-과잉 현상을 감소시킬 수 있는 가중치가 모수로 고려되어 있는 ZIP, ZIGP, Baysian ZIGP 모형은 0 및 1회 집중호우 월별개수를 보다 적절히 모의하고 있는 것을 알 수 있었다. 먼저 Bayesian ZIP 모형이 0, 1, 2회 집중호우 월별개

수의 발생모의를 가장 정확하게 모의하고 있는 것을 알 수 있었으며, 이는 대구와 부산 강우관측소의 실제 0회 집중호우 월별개수가 상당한 차이 (309회 vs. 275회)가 있음에도 불구하고 모두 좋은 성능을 보이고 있음을 알 수 있었다. ZIP 모형은 0회 집중호우 발생개수를 조금 높게 모의하고 1회 집중호우 발생개수를 조금 낮게 모의하는 성향을 보였으며, 이러한 경향은 부산 강우관측소에서 더 확실하게 나타났다. 즉 0의 개수가 많으면 많을수록 정확한 모의성능을 나타내지만 0의 개수가 상대적으로 적어지는 경우에는 모의성능이 일부 감소됨을 알 수 있었다. 그러나 ZIP 모형의 특성은 ZIGP 모형에 비교했을 때 더욱 두드러지지는 않아 ZIP 모형이 두 번째로 좋은 성능을 보이고 있음을 알 수 있다. 또한 3회 집중호우 월별개수의 모의결과에 있어서는 0-과잉 현상을 처리할 수 있는 ZIP, ZIGP, Baysian ZIGP 모형은 관측자료의 횟수보다 높은 모의결과를 보이는 특성이 있음을 알 수 있었으나, 4회 이상의 집중호우 월별개수의 모의결과에서는 모형별 특성을 일반화하여 제시하기에는 무리가 있음을 알 수 있었다.

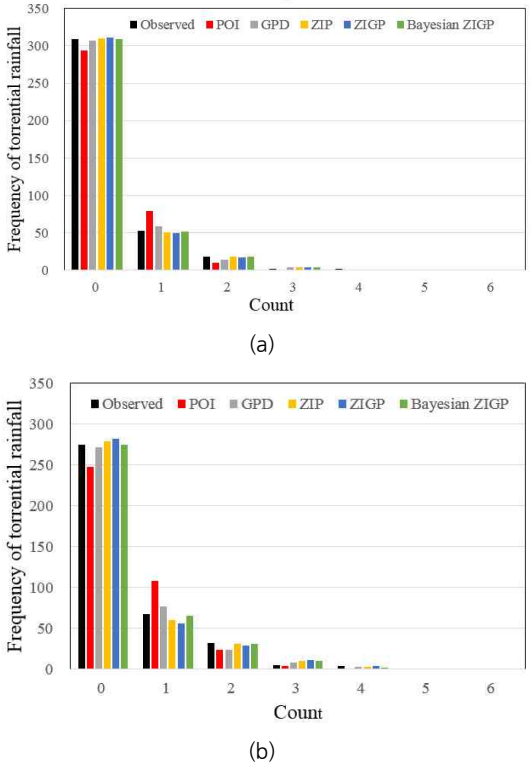


Fig. 4. Comparative results at Daegu and Busan rainfall gauges

이와 같은 모형별 결과를 종합하여 모형의 성능을 판단하면, 분석하고자 하는 자료에 포함되어 있는 0자료의 과다로 인하여 확률적 왜곡이 의심될 수 있는 경우 Bayesian ZIGP 모형을 사용하면 가장 적절한 모의결과를 얻을 수 있을 것으로 판단된다. 그러나 본 연구에서 사용된 경험적 Bayes 방법을 사용한 Bayesian ZIGP 모형은 그 과정의 구축 및 추가적인 자료의 구성에 있어서 많은 시간과 노력을 필요하게 되므로 모형의 간편성과 경제성을 고려한다면 ZIP 모형의 사용도 권고될 수 있다는 결론을 얻을 수 있었다.

6. 결 론

최근 기후변화로 인한 이상기후의 발생은 집중호우의 발생 경향을 증가시키고 있으며 집중호우의 빈번한 발생은 도심지의 침수, 산사태 발생으로 이어져 많은 인명 및 재산피해가 발생되었으며, 이와 같은 경향은 전 세계 및 국내에서 향후 지속적으로 증가될 것으로 전망되고 있어 과거 집중호우 발생 현황을 과학적으로 분석함으로써 관련 재해발생을 사전에 대비할 필요가 있다. 일반적으로 강우자료를 활용한 수자원 공학에서의 분석은 연최대시계열자료를 대상으로 하는 경우가 많으나 도심지 침수나 산사태 발생 등은 연최대 강우보다도 적은 집중호우의 발생으로도 충분히 발생가능하다. 따라서 본 연구에서는 80 mm/day 이상의 강우를 집중호우로 정의하고 1개월 동안 발생된 80 mm/day이상의 강우사상의 발생횟수를 집중호우 월별개수로 명명하고 이에 대한 확률적 분석을 수행하였다.

본 논문에서 정의된 집중호우 월별개수와 같이 특정시간 간격동안 발생된 사상의 발생횟수 자료는 count data라는 용어로 정의되어 자연과학 및 사회과학에서 중요하게 다루어지고 있으나 수자원 공학 분야에서는 상대적으로 다루어진 빈도가 낮고 특히 국내에서는 관련 연구를 찾아보기 힘들다. 이러한 count data는 대개 Poisson 분포(POI)를 이용하여 모형화되는데, 0회 발생횟수를 기록한 count data가 너무 많은 경우 확률분포의 밀도가 왜곡된다는 데 문제가 발생되며 이를 개선하기 위하여 일반화 Poisson 확률분포(Generalized Poisson Distribution, GPD), 0-과잉 Poisson 확률분포(Zero-Inflated Poisson distribution, ZIP), 0-과잉 일반화 Poisson 확률분포 (Zero-Inflated Generalized Poisson distribution, ZIGP), Bayesian 0-과잉 일반화 Poisson 확률분포(Bayesian Zero-Inflated Generaized Poission distribution), Bayesian ZIGP)가 개발된 바 있으나, 이러한 모형들이 집중호우의 발생분석을 위해 사용된 연구는 찾아보기 힘들다.

그러므로 본 연구에서는 대구 및 부산 강우관측소에서 과거 384개월 동안 관측된 집중호우 월별개수를 수집하여 정리

하고, 이를 5개의 모형(POI, GPD, ZIP, ZIGP, Bayesian ZIGP)을 이용하여 과거 집중호우의 발생현황을 모의함으로써 5개 모형의 특성을 분석하고 각각의 모형별 성능을 분석하여 가장 적절한 모형을 선정하여 제시하였다. 특히 Bayesian ZIGP 모형을 구축하고 수행함에 있어서는 경험적 Bayes 기법으로 구축된 정보적 사전분포를 구축함으로써 보다 객관적인 사전분포가 사용될 수 있도록 고려하였으며, 이를 위해 K-means 알고리즘을 이용한 군집분석을 통해 동질성이 확인된 2개위 그룹, 10개의 추가적인 강우관측소의 자료가 사용되었다.

최종적인 분석결과 분석하고자 하는 자료에 포함되어 있는 0자료가 과다한 경우 Bayesian ZIGP 모형을 사용하면 가장 적절한 모의결과를 얻을 수 있을 것으로 판단되었으며, 모형의 간편성과 경제성을 고려한다면 ZIP 모형의 사용도 충분히 권고될 수 있는 성능을 보인 것을 알 수 있었다. 그러나 일반적으로 사용되는 POI 모형과 GPD 모형은 0이 과잉된 경우를 처리할 수 있는 수학적 고려가 없으므로 관측 자료와는 매우 다른 결과를 도출할 수 있어 자료에 따라 주의 깊게 사용되어야 함을 알 수 있었다. 마지막으로 본 연구에서는 다루어지지 않았지만 불확실성을 포함한 결과에 있어서는 Bayesian ZIGP 모형이 나타낼 수 있는 불확실성의 범위가 본 연구에서 다루어진 다른 모형에 비하여 월등히 우수할 수 있어 불확실성을 포함한 연구의 추가적인 수행이 필요할 것으로 판단된다.

감사의 글

이 연구는 2014년 교육부 재원을 한국연구재단을 통한 기초연구사업(NRF-2014R1A1A2053328) 및 2015년도 강원대학교 대학회계 학술연구조성비(D1000072-01-01, 520150072)에 의해 수행되었습니다. 연구지 지원에 감사를 표합니다.

References

Abdul, R.U. and Zeephongsekul, P. (2014) "Copula based analysis of rainfall severity and duration: a case study" *Theoretical and Applied Climatology* Vol. 115, No. 12, pp. 153-166.

Angers, J.F., and Biswas, A. (2003) "A Bayesian analysis of zero-inflated generalized Poisson model." *Computational Statistics & Data Analysis* Vol. 42, No. 1-2, pp. 37-46.

Bates, B.C., and Campbell, E.P. (2001) "A Markov Chain Monte Carlo scheme for parameter estimation and inference in conceptual rainfall-runoff modeling." *Water Resources Research* Vol. 37, No. 4, pp. 937-947.

Berger, J.O. (1985) Statistical Decision Theory and Bayesian Analysis. Springer-Verlag, New York.

Bertucio, R.C., and Julius, J.A. (1990) "Analysis of CDF:Surry, Unit 1 internal events." US *Nuclear Regulatory Commission* NUREG/CR-4550, Vol. 5, pp. 2-8.

Bohning, D., Dietz, E., Schlattmann, P., Mendonca, L., and Kirchner, U. (1999) "The zero-inflated Poisson model and the decayed, missing and filled teeth index in dental epidemiology." *Journal of the Royal Statistical Society* Series A (Statistics in Society) Vol. 162, No. 2, pp. 195-209.

Carlin, B.P., and Louis, T.A. (1996) Bayes and empirical Bayes methods for data analysis. Chapman and Hall, New York.

Chapman, T. (1998) "Stochastic modelling of daily rainfall: the impacts of adjoining wet days on the distribution of rainfall amounts." *Environmental Modelling and Software* Vol. 13, No. 3-4, pp. 317-324.

Chib, S., and Greenberg, E. (1995) "Understanding the Metropolis-Hastings algorithm." *Journal of the American Statistical Association* Vol. 49, No. 4, pp. 327-335.

Chung, E.S., and Kim, S.U. (2013) "Bayesian rainfall frequency analysis with extreme value using the informative prior distribution." KSCE *Journal of Civil Engineering* Vol. 17, No. 6, pp. 1502-1514.

Cohen, Jr. A.C. (1960) "Estimating the parameters of a modified Poisson distribution." *Journal of American Statistical Association* Vol. 55, No. 289, pp. 139-143.

Consul, P.C., Jain, G.C. (1973) "A Generalization of the Poisson distribution." *Technometrics* Vol. 15, No. 4, pp. 791-799.

Gamerman, D. (1997) Markov Chain Monte Carlo-Stochastic simulation for Bayesian inference. Chapman&Hall, London UK.

Gelman, A., and Rubin, D.B. (1992) "Inference from iterative simulation using multiple sequences." *Statistical Science* Vol. 7, No. 4, pp. 457- 511.

Geweke, J. (1992) Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In Bayesian statistics 4, eds. Bernardo, J.M., Berger, J., Dawid, A.P., and Smith, A.F.M. Oxford, UK, Oxford University Press.

Goyal, M.K. (2014) "Statistical analysis of long term trends of rainfall during 1901-2002 at Assam, India." *Water Resources Management* Vol. 28, No. 6, pp. 1501-1515.

HSBC (2011) Climate investment update. HSBC Global

Research, 13 October.

Jimoh, O.D., and Webster, P. (1996) "Optimum order of Markov chain for daily rainfall in Nigeria." *Journal of Hydrology* Vol. 185, No. 1-4, pp. 45-69.

Kaplan, S. (1985) "Two-stage Poisson-type problem in probabilistic risk analysis." *Risk Analysis* Vol. 5, No. 3, pp. 227-230.

Katz, R.W., and Parlange, M.B. (1995) "Generalizations of chain-dependent processes: applications to hourly precipitation." *Water Resources Research* Vol. 31, pp. 1331-1341.

Kavetski, D., Kuczera, G., and Franks, S.W. (2006) "Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory." *Water Resources Research* 42: W03407

Kim, S.U., and Lee, K.S. (2010) "Regional low flow frequency analysis using Bayesian regression and prediction at ungauged catchment in Korea." KSCE *Journal of Civil Engineering* Vol. 14, No. 1, pp. 87-98.

Lambert, D. (1992) "Zero inflated Poisson regression with an application to defects in manufacturing." *Technometrics* Vol. 34, No. 1, pp. 1-14.

Lee, K.S., and Kim, S.U. (2008) "Identification of uncertainty in low flow frequency analysis using Bayesian MCMC method." *Hydrological Processes* Vol. 22, No. 12, pp. 1949-1964.

Lee, C.E., Kim, S.U., and Lee, S. (2014) "Time-dependent reliability analysis using Bayesian MCMC on the reduction of reservoir storage by sedimentation." *Stochastic Environmental Research Risk Assessment* Vol. 28, No. 3, pp. 639-654.

Malakoff, D. (1999) "Bayes offers a 'New way to make sense of numbers'." *Science* Vol. 286, No. 5444, pp. 1460-1464.

Maritz, J.S., and Lwin, T. (1989) "Empirical Bayes Approach to Multiparameter Estimation: With Special Regerence to Multinomial Distribution" *Ann. Inst. Statis. Math.* Vol. 41, No. 1, pp. 81-99.

Marshall, L., Nott, D., and Sharma, A. (2004) "A comparative study of Markov Chain Monte Carlo methods for conceptual rainfall-runoff modeling." *Water Resources Research* 40: W02501

Mullahy, J. (1986) "Specification and testing of some modified count data models." *Journal of Econometrics* Vol. 33, No. 3, pp. 341-365.

Raftery, A.E., and Lewis, S. (1992) How many iterations in the Gibbs sampler? In: Bernardo, J.M., Berger, J., Dawid, A.P., and Smith, A.F.M. (ed) Bayesian statistics 4, Oxford University Press, Oxford, UK, pp. 763-773.

Roberts, G.O., Gelman, A., and Gilks, W.R. (1994) Weak convergence and optimal scaling of random walk Metropolis-Hastings algorithms. Technical Report, University of Cambridge.

Saidi, H., Ciampittiello, M., Dresti, C., and Ghiglieri, G. (2015) "Assessment of trends in extreme precipitation event: A case study in Piedmont (North-West Italy)." *Water Resources Management* Vol. 29, No. 1, pp. 63-80.

Seidou, O., Ouarda, T.B.M.J., Barbet, M., Bruneau, P., and Bobee, B. (2006) "A parametric Bayesian combination of local and regional information in flood frequency analysis." *Water Resources Research* Vol. 42: W11408

Singh, S.N. (1963) "A note on inflated Poisson distribution." *Journal of Indian Statistical Association* Vol. 1, pp. 140-144.

Todorovic, P., and Yevjevich, V. (1969) Stochastic process of precipitation. Hydrology papers 35, Colorado State University, Fort Collins, Colorado

Wheeler, T.A. (1993) "Analysis of the LaSalle Unit 2 nuclear power plant: risk methods integration and evaluation program (RMIEP): parameter estimation analysis and screening human reliability analysis." US *Nuclear Regulatory Commission* NUREG/CR-4832:Vol. 5.