*The Korean Journal of Applied Statistics* (2014) DOI: http://dx.doi.org/10.5351/KJAS.2014.27.6.1069
**27**(6), 1069–1076

# A Bayesian Prediction of the Generalized Pareto Model

Pan Huh[a] · Joong Kweon Sohn[a,1]

[a]Department of Statistics, Kyungpook National University

(Received August 29, 2014; Revised October 22, 2014; Accepted October 22, 2014)

**Abstract**

Rainfall weather patterns have changed due to global warming and sudden heavy rainfalls have become more frequent. Economic loss due to heavy rainfall has increased. We study the generalized Pareto distribution for modelling rainfall in Seoul based on data from 1973 to 2008. We use several priors including Jeffrey's noninformative prior and Gibbs sampling method to derive Bayesian posterior predictive distributions. The probability of heavy rainfall has increased over the last ten years based on estimated posterior predictive distribution.

Keywords: Generalized Pareto distribution, Markov Chain Monte Carlo method, prior distribution, Bayesian predictive posterior distribution, rainfall model.

## 1. 서론

파레토 분포는 Pareto (1897)가 소득에 대한 분포로 처음 제안하였다. Pickands (1975)는 주어진 한계를 넘어가는 경우에 대한 모형으로써 일반화 파레토분포를 제안하였고 이에 대해 Davison (1984), Castillo와 Hadi (1997), Castillo와 Daoudib (2008) 등에 의해 연구되었다. 특히 Castillo 와 Hadi (1997)는 모수와 분위 수에 대한 추정 방법으로 적률방법(method of moments), 확률가중적률방법(probability weighted moments),기초백분위 방법(elemental percentile method)을 사용하였다. 이들 추정량들은 계산이 쉽고 잘 정의되는 것으로 알려졌다. 또 Van과 Witter (1986)은 척도모수와 형상모수에 대해 최우 추정법을 사용하였다. 이 일반화 파레토분포는 많은 분야에서 적용되어왔다. 예를 든다면 도시 인구 문제의 연구에서도 중요한 역할을 하였고, 보험 위험이나 비즈니스 실패 등에 모형으로 많이 사용되었다. 특히 일반화 파레토본포의 중요한 응용분야가 강수모형(rainfall model)이다. Fitzgerald (1989)는 확률가중적률방법을 사용하여 강수의 높은 공간변동에서 기인되는 큰 표준 오차를 줄이는 문제를 연구하였다. 또 Li 등 (2005)는 호주의 남서쪽 5 군데에 대한 폭우모형을 세우고 분석을 하였고 특히 겨울철 폭우가 변화되는 시점에 대한 연구한 결과 1965년을 기점으로 가장 큰 변화가 있었음을 밝혔다.

특히 폭우의 경우 홍수나 산사태, 집 등의 자산 손실 등 많은 피해를 가져오고 있다. 기상청에서는 6시간에 70mm 이상 혹은 12시간에 110mm 이상인 경우 폭우 경보를 내린다. 한 연구에 의하면 2002년부터 2011년까지 10년간 6월부터 9월까지의 여름철 강우와 태풍에 의한 피해를 집계한 결과 약 19조

This research was supported by Kyungpook National University Research Fund, 2013.
[1]Corresponding author: Department of Statistics, Kyungpook National University, Daegu 702-701, Korea. E-mail : jsohn@knu.ac.kr

4000억 여원의 피해가 발생한 것으로 드러났다 (National Emergency Agency, 2011). 이 집계에 따르면 한 해 평균 여름철 폭풍 피해로 평균 2조 원의 피해로 이 금액은 국민총생산의 0.2% 정도나 된다. 따라서 폭우에 대한 예측이 매우 중요한 문제로 대두되었다. 이 논문에서는 기상청의 기준에 따라 70mm 이상인 경우를 한계로 한 강수모형으로 일반화 파레토모형을 사용하였다.

이 논문에서는 $u$를 한계로 한 강우량 $X$의 분포를 $F$라고 하고 $X$에 대한 예측분포를 구하고자 한다. 따라서 일반화 파레토 분포 $F$는 다음과 같이 정의된다.

$$F(x) = P(X - u \le x | X < u), \quad 0 \ge x \ge x_{\max} - u. \tag{1.1}$$

여기서 $x_{\max}$는 $F$의 오른쪽 끝 값이며, $x = X - u$는 초과량이 된다. Pickands (1975)는 $u$가 매우 큰 경우 조건부 분포 $F$는 다음과 같이 잘 근사됨을 보였는데

$$F_u(x) \approx F(x|\xi, \sigma), \quad u \to \infty, \tag{1.2}$$

여기서 $F(x|\xi,\sigma)$는 형상모수 $\xi$와 척도모수 $\sigma$인 일반화 파레토 분포(GPD$(\xi,\sigma)$)이며 다음과 같이 정의되어진다.

$$F(x|\xi,\sigma) = \begin{cases} 1 - \left(1 - \dfrac{x\xi}{\sigma}\right)^{\frac{1}{\xi}}, & \left(1 - \dfrac{x\xi}{\sigma}\right) > 0,\ x > 0,\ \text{for } \xi \le 0, \\ 1 - \exp\left(-\dfrac{x}{\sigma}\right), & \xi = 0,\ 0 \le x \le \dfrac{\sigma}{\xi},\ \text{for } \xi > 0. \end{cases} \tag{1.3}$$

또 GPD$(\xi,\sigma)$의 확률밀도함수 $f(x|\xi,\sigma)$는 다음과 같다.

$$f(x|\xi,\sigma) = \begin{cases} \dfrac{1}{\sigma}\left(1 - \dfrac{x\xi}{\sigma}\right)^{\frac{1-\xi}{\xi}}, & \dfrac{1 - x\xi}{\sigma} > 0, \\ \dfrac{1}{\sigma}\exp\left(-\dfrac{x}{\sigma}\right), & \xi = 0. \end{cases} \tag{1.4}$$

GPD$(\xi,\sigma)$는 다음의 몇몇 중요 성질을 가지고 있다.

1. $\xi$=0이면 GPD$(\xi,\sigma)$는 평균이 $\sigma$인 지수분포가 된다.
2. $\xi$=1이면 GPD$(\xi,\sigma)$는 항등분포 $U(0,\sigma)$가 된다.
3. $\xi \le 0$이면 GPD$(\xi,\sigma)$는 모수가 $\xi$인 파레토분포가 된다.

## 2. 사전분포에 대한 사후예측분포

이 논문에서 우리가 고려하는 일반화 파레토 모형의 분포는 다음과 같다.

$$f\left(x|\xi,\sigma\right) = \frac{1}{\sigma}\left(1 - \frac{x\xi}{\sigma}\right)^{\frac{1-\xi}{\xi}}, \quad \frac{1 - x\xi}{\sigma} > 0. \tag{2.1}$$

$\underline{X} = (X_1, X_2, \ldots, X_n)$을 $n$개의 랜덤표본이라 하고 그 관찰치를 $\underline{x} = (x_1, x_2, \ldots, x_n)$라고 하자. 일반적으로 모수 $\underline{\theta}$에 대한 사전분포를 $\pi(\underline{\theta})$, 사후 분포를 $\pi(\underline{\theta}|\underline{x})$라 하면 사후 예측분포 $f(y|\underline{x})$는 다음과 같이 얻어진다.

$$\pi\left(\underline{\theta}|\underline{x}\right) = \frac{f\left(\underline{x}|\underline{\theta}\right)\pi\left(\underline{\theta}\right)}{\int f\left(\underline{x}|\underline{\theta}\right)\pi\left(\underline{\theta}\right)d\underline{\theta}}, \tag{2.2}$$

$$f\left(y|\underline{x}\right) = \int f\left(y|\underline{\theta}\right)\pi\left(\underline{\theta}|\underline{x}\right)d\underline{\theta}. \tag{2.3}$$

여기서 일반적으로 식 (2.2)와 식 (2.3)에서 적분이 닫힌 형태로 얻어지지 않기 때문에 마코프 체인 몬테 칼로(Markov Chain Monte Carlo) (Gelfand 등 (1990) 참조)에 기초한 깁스 샘플링 방법을 응용한다. 이 방법은 이미 널리 알려져 있기 때문에 소개하는 것은 생략하기로 한다.

### 2.1. 제프리의 비정보 사전분포의 경우

$(X_1, X_2, \ldots, X_n)$을 식 (2.1)과 같은 일반화 파레토 모형에서 얻어지는 $n$개의 랜덤표본이라고 하고, $(x_1, x_2, \ldots, x_n)$을 관찰된 값이라고 하자. 여기서 형상모수 $\xi$와 척도모수 $\sigma$는 각각 독립이라고 가정한다. 이 경우 제프리의 비정보 사전분포(Jeffrey's noninformative prior distribution)는 다음과 같이 얻어진다.

$$\begin{aligned}\pi(\xi, \sigma) &= \pi(\xi)\pi(\sigma) \\ &\propto \sigma^{-1}(1+\xi)^{-1}(1+2\xi)^{-\frac{1}{2}}, \quad \xi > -0.5.\end{aligned} \tag{2.4}$$

여기서 사후분포 $\phi(\xi, \sigma|\underline{x})$는 닫혀진 함수로는 구할 수 없으나 커널은 다음과 같이 얻어진다.

$$\pi(\xi, \sigma|\underline{x}) \propto \sigma^{-(n+1)}(1+\xi)^{-1}(1+2\xi)^{-\frac{1}{2}} \prod_{i=1}^{n} \left(\frac{1-x_i\xi}{\sigma}\right)^{\frac{1-\xi}{\xi}} \tag{2.5}$$

깁스 샘플링 방법을 사용하기 위해서는 다음과 같이 조건부 분포를 사용한다. 즉,

$$\begin{aligned}\pi(\xi|\sigma, \underline{x}) &= \pi(\xi)L(\xi, \sigma|\underline{x}) \\ &\propto (1+\xi)^{-1}(1+2\xi)^{-\frac{1}{2}}\sigma^{-n} \prod_{i=1}^{n} \left(\frac{1-x_i\xi}{\sigma}\right)^{\frac{1-\xi}{\xi}}\end{aligned} \tag{2.6}$$

와

$$\begin{aligned}\pi(\sigma|\xi, \underline{x}) &= \pi(\sigma)L(\xi, \sigma|\underline{x}) \\ &\propto \sigma^{-(n+1)} \prod_{i=1}^{n} \left(\frac{1-x_i\xi}{\sigma}\right)^{\frac{1-\xi}{\xi}}\end{aligned} \tag{2.7}$$

이다. 또한 사후 예측분포는 다음과 같은 모양이 된다.

$$f(y|\underline{x}) = \int \delta^{-(n+2)}(1+\xi)^{-1}(1+2\xi)^{-\frac{1}{2}} \left(1+\frac{\xi y}{\sigma}\right)^{-\frac{1}{\xi}-1} \prod_{i=1}^{n} \left(1-\frac{\xi x_i}{\sigma}\right)^{-\frac{1}{\xi}-1} d\xi d\delta \tag{2.8}$$

### 2.2. 특정한 사전분포의 경우

여기서 우리는 형상모수와 척도모수에 대해 흔히 사용되는 사전분포를 다음과 같이 가정하고자 한다. 즉 형상모수 $\xi$에 대한 사전분포는 모수가 $a(>0)$와 $b(>0)$인 감마분포 $G(a, b)$이며, 척도모수 $\sigma$에 대한 사전분포는 모수가 각각 $\alpha(>0)$와 $\beta(>0)$인 역감마분포 $\mathrm{IG}(\alpha, \beta)$라고 하자. 그러면 사후분포 $\pi(\xi, \sigma|\underline{x})$는 다음과 같다.

$$\begin{aligned}\pi(\xi, \sigma|\underline{x}) &= \frac{\beta^{\alpha}\xi^{-n+a-1}\sigma^{-n-\alpha-1}\exp\left[-\frac{\xi}{b}-\frac{\beta}{\sigma}+\left(\frac{1}{\xi}-1\right)\sum_{i=1}^{n}\ln\left(1-\frac{x_i}{\sigma}\right)\right]}{b\Gamma(a)\Gamma(\alpha)} \\ &\propto \xi^{-n+a-1}\sigma^{-n-\alpha-1}\exp\left[-\frac{\xi}{b}-\frac{\beta}{\sigma}+\left(\frac{1}{\xi}-1\right)\sum_{i=1}^{n}\ln\left(1-\frac{x_i}{\sigma}\right)\right]\end{aligned} \tag{2.9}$$

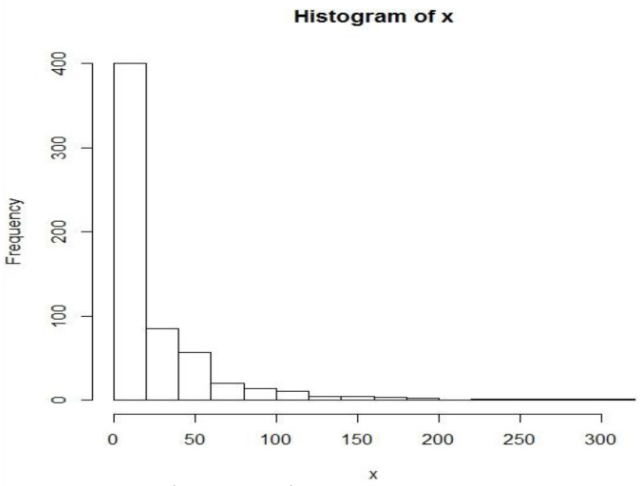



**Figure 3.1.** The pivotal legislator ($149^{\text{th}}$) in the $18^{\text{th}}$ national assembly: Posterior probabilities

또한 $\sigma$와 $\underline{x}$가 주어졌을 때 $\xi$의 조건부 사후분포와 $\xi$와 $\underline{x}$가 주어졌을 때 $\sigma$의 조건부 사후분포는 다음과 같다.

$$\begin{aligned}\pi(\xi|\sigma,\underline{x}) &= \pi(\xi)L(\underline{x}|\xi,\sigma)\\ &= \frac{\xi^{-n+a-1}e^{-\frac{\xi}{b}}\sigma^{-n}\prod_{i=1}^{n}\left(1-\frac{x_i}{\sigma}\right)^{\frac{1}{\xi}-1}}{b^a\Gamma(a)}\\ &\propto \xi^{-n+a-1}e^{-\frac{\xi}{b}}\sigma^{-n}\prod_{i=1}^{n}\left(1-\frac{x_i}{\sigma}\right)^{\frac{1}{\xi}-1}\end{aligned} \tag{2.10}$$

와

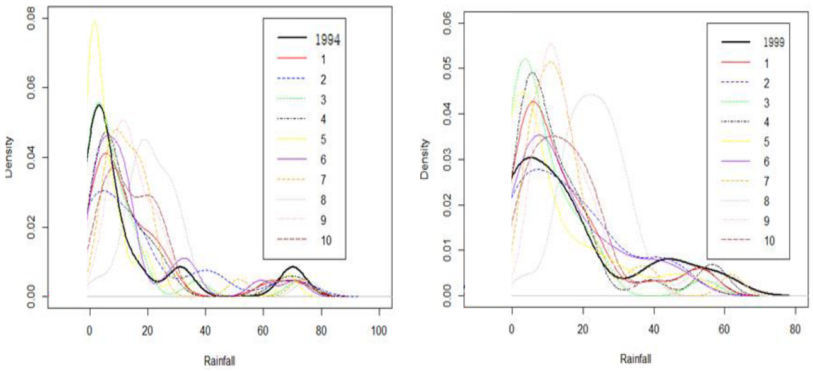
$$\begin{aligned}\pi(\sigma|\xi,\underline{x}) &= \pi(\sigma)L(\underline{x}|\xi,\sigma)\\ &= \frac{\beta^\alpha\sigma^{-n-\alpha-1}e^{-\frac{\beta}{\sigma}}\xi^{-n}\prod_{i=1}^{n}\left(1-\frac{x_i}{\sigma}\right)^{\frac{1}{\xi}-1}}{\Gamma(\alpha)}\\ &\propto \sigma^{-n-\alpha-1}e^{-\frac{\beta}{\sigma}}\xi^{-n}\prod_{i=1}^{n}\left(1-\frac{x_i}{\sigma}\right)^{\frac{1}{\xi}-1}\end{aligned} \tag{2.11}$$

이다. 또한 사후 예측분포는 다음과 같다.

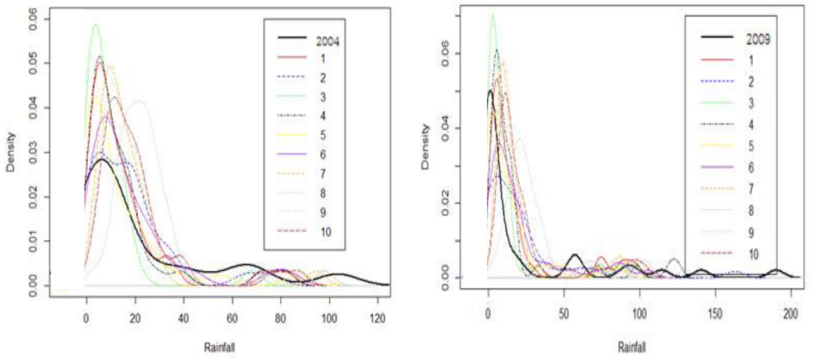
$$\begin{aligned}f(y|\underline{x}) &= \int f(y|\xi,\sigma)\pi(\xi,\sigma|\underline{x})d\xi d\sigma\\ &= \int \frac{\beta^\alpha\xi^{-n+a-2}\sigma^{-n-\alpha-2}\left(1-\frac{y}{\sigma}\right)^{\frac{1}{\xi}-1}\exp\left[-\frac{\xi}{b}-\frac{\beta}{\sigma}+\left(\frac{1}{\xi}-1\right)\sum_{i=1}^{n}\ln\left(1-\frac{x_i}{\sigma}\right)\right]}{b^a\Gamma(a)\Gamma(\alpha)}d\xi d\sigma.\end{aligned} \tag{2.12}$$

**Table 3.1.** Values of parameters for several prior distributions

| 사전분포 | 모수 | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
|---|---|---|---|---|---|---|---|---|---|---|
| $G(a,b)$ | a | 0.1 | 0.1 | 0.1 | 3 | 3 | 3 | 0.5 | 0.5 | 0.5 |
| | b | 3 | 3 | 3 | 0.1 | 0.1 | 0.1 | 0.5 | 0.5 | 0.5 |
| $\mathrm{IG}(\alpha,\beta)$ | $\alpha$ | 0.1 | 3 | 0.5 | 0.1 | 3 | 0.5 | 0.1 | 3 | 0.5 |
| | $\beta$ | 3 | 0.1 | 0.5 | 3 | 0.1 | 0.5 | 3 | 0.1 | 0.5 |




(a) (b)



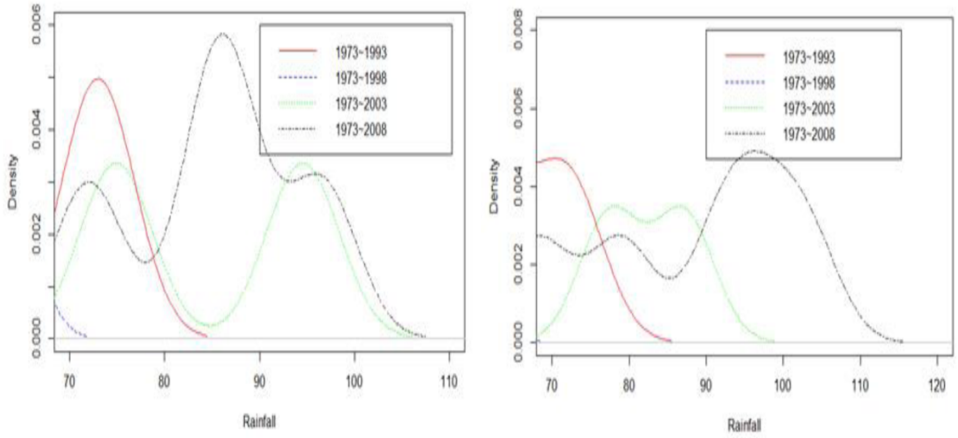

(c) (d)

**Figure 3.2.** Bayesian posterior predictive distributions with different priors and observed rainfalls (a) based on the data from 1973 to 1993 and rainfall at 1994 (b) based on the data from 1973 to 1998 and rainfall at 1999 (c) based on the data from 1973 to 2003 and rainfall at 2004 (d) based on the data from 1973 to 2008 and rainfall at 2009

## 3. 강수데이터를 이용한 모형 분석

### 3.1. 강수 데이터와 모형의 조건

여기서 사용되는 강수데이터는 1973년 부터 2011년까지 매년 7월의 강수데이터이며 70mm를 폭우의 한계값으로 정했다. 여기서 얻어진 데이터의 히스토그램은 Figure 3.1과 같다. 여기서 각각의 사전분




(a) (b)

**Figure 3.3.** Tail parts of four Bayesian posterior predictive densities with data in Figure 2 (a) Case of $G(0.5, 0.5)$ and $IG(0.1, 3)$ (b) Case of Jeffrey's noninformative prior

포는 모수의 값에 따라 분포의 모양이 다양하게 변화하는 것으로 나타나며 이는 사전정보의 결정요소가 된다. 그러나 아직은 어떤 경우가 적합한지 알려진 바가 없어 다양한 값을 선택해서 결과를 보기로 했다 (Table 3.1). 또 시간에 따른 강수패턴의 변화를 보다 세밀히 보기 위해 1973년부터 20년인 1993년을 기점으로 매 5년씩 더한 4가지 경우로 나누어 보았다. 즉 1973년부터 1993년, 1973년부터 1998년, 1973년부터 2003년, 1973년부터 2008년으로 나누어서도 고려해보았다. 사후예측분포를 얻기 위해서 깁스 샘플링 방법을 사용하되 초기 3,000번은 버리고 10,000번을 반복했으며, 각각의 데이터 조합으로 사후예측분포를 그려보았다 (Figure 3.2). 그 결과는 다음과 같다. 각각의 그림은 1994년, 1999년, 2004년. 2009년에 대한 사후예측분포와 실제 관측값의 분포를 그렸으며 10번은 제프리의 무정보 사전분포의 경우이다. 또 주어진 데이터 조합에서 각각의 사전분포들에 대한 사후예측분포의 꼬리부분을 확대해 보았다 (Figure 3.3).

Figure 3.2로는 각각의 사전분포에 따른 사후예측분포를 비교하기에는 어려워서 우리는 70mm 이상의 강우에 대해 실제 관측된 값과 각각의 사후예측분포에서 구한 확률을 구해 비교해 보았다 (Table 3.2).

이상의 결과를 두고 본다면 어떤 사전분포든 비교적 잘 예측하고 있음을 알 수 있으나, $a = b$인 경우와 $\alpha$가 비교적 작은 값을 가지며 $\beta$ 의 값이 $\alpha$ 보다 큰 경우와 제프리의 비정보 사전분포의 경우가 상대적으로 정확한 것을 알 수 있다. 또한 각각의 사전분포에 대한 사후예측분포들에 대해 조사해 본 결과 지난 15년간 100mm 이상의 폭우가 내릴 확률이 점차 증가하고 있음을 볼 수 있었다.

## 4. 결론

본 논문에서는 관측된 데이터를 기초로 일반화 파레토 분포를 활용하여 강수모형을 통해 사후예측분포를 구해보았다. 형상모수의 경우 사전분포의 모수에 대해 $a = b$인 경우와 척도모수의 경우 $\alpha$가 비교적 작은 값을 가지며 $\beta$의 값이 $\alpha$ 보다 큰 경우와 전반적으로 제프리의 비정보 사전분포가 잘 예측하고 있어 특정 사전분포보다는 그 예측이 좀 더 정확함을 보였다. 이는 강수패턴이 변화함에 따라 사전분포의 모수의 특정 값에 의존해서 예측분포를 구하는 경우보다 무정보사전분포의 경우가 더 데이터에 의존하

**Table 3.2.** Comparisons between several Bayesian posterior predictive probabilities and observed relative frequencies

| 예측된 연도 | | 1994 | 1999 | 2004 | 2009 |
|---|---|---|---|---|---|
| 관측된 상대 확률 | | 0.0323 | 0 | 0.0645 | 0.1293 |
| 사후예측분포로부터 추정된 확률 | (1) | 0.0332 | *** | 0.0556 | 0.1038 |
| | (2) | 0.0455 | **** | 0.0335 | 0.0976 |
| | (3) | 0.0319 | **** | 0.0426 | 0.0806 |
| | (4) | 0.0449 | ** | 0.0545 | 0.1022 |
| | (5) | 0.0156 | **** | 0.0322 | 0.0534 |
| | (6) | 0.0278 | *** | 0.0598 | 0.1028 |
| | (7) | 0.0373 | 0.0002 | 0.0612 | 0.1027 |
| | (8) | 0.0425 | **** | 0.0449 | 0.0835 |
| | (9) | 0.0370 | 0.0005 | 0.0565 | 0.1052 |
| | (10) | 0.0331 | *** | 0.0639 | 0.1197 |

* : $10^{-3}$ , ** : $10^{-6}$ , *** : $10^{-9}$ , **** : $< 10^{-9}$

여 데이터의 변화를 더 적극적으로 반영하기 때문에 예측분포의 예측력이 좀 더 나은 것으로 생각된다. 무엇보다 중요한 것은 1994년. 1999년, 2004년, 2009년에 대한 사후예측분포를 봤을 때 어떤 사전 분포를 사용하든 폭우의 가능성 즉 70mm 이상의 폭우가 내릴 가능성이 점차 증가하는 추세를 읽을 수가 있어 강수패턴의 변화가 있음을 읽을 수가 있었다.

## References


Castillo, E. and Hadi, A. S. (1997). Fitting the generalized Pareto distribution to data, *Journal of the American Statistical Association*, **92**, 1609–1620.

Castillo, J. and Daoudib J. (2008). Estimation of the generalized Pareto distribution, *Statistics & Probability Letters*, In Press.

Davidson, A. C. (1984). Modeling excesses over high threshold with an application, In: J. Tiago de Oliveria (ed.), *Statistical Extremes and Applications*, Reidel, Dordrech, 416–482.

Fitzgerald, D. L. (1989). Single station and regional analysis of daily rainfall extremes, *Stochastic Hydrology and Hydraulics*, **3(4)**, 281–292.

Gelfand, A. E., Hills, S. E., Racine-Poon, A. and Smith, A. F. M. (1990). Illustration of Bayesian inference in normal data models using Gibbs sampling, *Journal of the American Statistical Association*, **85**, 972–985.

Li, Y., Cai, W. and Campbell, E. P. (2005). Statistical modeling of extreme rainfall in southwest Western Australia, *Journal of Climate*, **18**.

National Emergency Agency (2011). *2011 재해연보*, 소방방재청 복구지원과.

Pareto, V. (1897). Cours d'economic politique, *Lausanee and Paris; Range and Ci.*

Pickands, J. (1975). Statistical inference using extreme order statistics, *The Annals of Statistics*,**3**, 119–131.

Van Montfort, M. A. J. and Witter, J. V. (1986). The generalized Pareto distribution applied to rainfall depths, *Hydrological Sciences Journal*, **31**, 151–162.

# 일반화 파레토 모형에서의 베이지안 예측

판허[a] · 손중권[a,1]

[a]경북대학교 통계학과

(2014년 8월 29일 접수, 2014년 10월 22일 수정, 2014년 10월 22일 채택)

**요 약**

기후 온난화의 한 현상으로 받아들여지는 집중호우로 인한 관심이 늘어난 만큼 강우량에 대한 예측 모형이 필요하다. 이러 환경 문제를 다룰 때, 모형을 설정하는 방법 중에 하나로 일반화 파레토 모형을 활용하는 연구가 이루어지고 있다. 본 논문에서는 서울특별시에 대한 1973년부터 2011년까지 매 7월 일별강우량 자료를 가지고 일반화 파레토 모형을 사용하여 강우량의 임계값(70mm) 이상의 분포가 어떻게 되는지 연구한다. 모수의 사전분포는 감마분포랑 역감마분포를 정의하고, 또는 제프리의 정보가 없는 사전분포를 두고, 깁스 표본방법을 통해 베이지안 사후예측분포를 구하고 얻어진 결과를 비교해 본다.

주요용어: 일반화 파레토 분포, 마코프 체인 몬테칼로 방법, 사전분포, 베이지안 사후 예측분포, 강수모형

이 논문은 2013학년도 경북대학교 학술연구비에 의하여 연구되었음.

[1]교신저자: (702-701) 대구광역시 북구 대학로 80, 경북대학교 통계학과. E-mail : jsohn@knu.ac.kr