Journal of the Korean Data & Information Science Society
2013, **24**(4), 857–865

http://dx.doi.org/10.7465/jkdi.2013.24.4.857
한국데이터정보과학회지

일반화 극단 분포를 이용한 강우량 예측

성용규[1] · 손중권[2]

[12]경북대학교 통계학과

접수 2013년 4월 21일, 수정 2013년 7월 5일, 게재확정 2013년 7월 18일

요 약

집중 호우로 인한 피해가 증가하면서 다양한 기법들을 이용하여 강우량 예측에 대한 관심이 높아졌다. 최근에는 극단분포를 활용하여 강우량을 예측하려는 시도가 늘고 있다. 본 연구에서는 일반화 극단 분포를 활용하여 실제 서울시의 1973년부터 2010년까지 7월달의 사후예측분포를 생성하고, 수치적인 계산을 위해서 MCMC (Markov chain Monte Carlo)알고리즘을 활용하였다. 이 연구를 통해서 사후예측분포의 점추정값들을 비교하였고 2011년 7월달의 자료와 비교해 봤을 때 집중 호우의 확률이 증가한 것을 알 수 있었다.

주요용어: 마르코프 연쇄 몬테칼로, 베이지안, 일반화 극단분포.

1. 서론

소방방재청에서 발간한 재해연보 (2011)을 분석한 결과에 따르면 2011년 전체 재산 피해금액 중 집중호우로 인한 피해금액이 5,276억원으로 전체 피해액의 66.4%이다. 이 것은 태풍으로 인한 피해액인 2,183억의 두배가 넘는 금액이다. 이런 집중호우는 산사태, 도로 및 제방유실, 건물붕괴, 통신두절, 정전 등으로 이어지므로 강우량 예측모형을 통해 피해에 대한 대비책이 필요하다.

환경문제에서 극단값을 모형화하는 것은 일반적인 연구방법 중에 하나이다. 극단분포 (extreme value distribtuion)를 활용한 강우량에 대한 선행연구들이 이루어져 오고 있다. 몇 가지 대표적인 연구들로, Coles와 Tawn (1996)은 영국 남서지역의 54년간의 일별 강우량 자료를 가지고 베이지안 추정을 통해 일반화 극단분포 (generalized extreme value distribution)에 대한 사후예측분포를 구했는데, 모수에 대한 사전분포를 사용하지 않고 분위수를 활용하여 사전분포를 정의하고, 30mm와 40mm 이상을 분계점으로 극단값으로 정했다. 또 다른 논문으로 Coles와 Pericchi (2003)는 베네수엘라의 40년간의 일별 강우량 자료를 가지고 일반화 극단분포의 사후예측분포를 구했는데, 모수에 대한 사전분포를 각각 독립인 정규 분포로 가정하였으며, 10mm를 분계점으로 극단값으로 정했다.

본 연구의 2절에서는 일반화 극단분포에 대해서 설명하고, 사후 베이지안 예측분포에 대해 연구한다. 그 후 일반화 극단분포로부터의 사후 베이지안 예측분포에서 모평균에 대한 추정량을 연구한다. 3절에서는 소개된 방법을 이용하여 기상청으로부터 제공된 1973년부터 2011년도까지의 서울시 일별강수량 자료에 대해 분석한다. 마지막으로 4절에서는 본 연구 결과의 요약과 결론을 제시하고자 한다. 또한 본 연구를 통해 국지성 집중 강우가 빈번해지는 요즈음 기후변화에 따른 대책 수립에서 도움이 될 수 있는 가능성을 탐색해 볼 수 있다.

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

2. 연구 자료 및 방법

Figure 2.1은 기상청에서 제공한 1973년 1월부터 2011년 12월까지 서울시 일별강수량 자료이다. 총 14,244개의 자료로 이루어져 있으며, 이 중 61%인 8,689개가 결측치이며, 자료가 없거나 장애/결측된 기간의 관측자료이므로, 비나 눈이 오지 않는 날로 가정한다. Figure 2.2은 사례 연구에서 사용된 자료의 분포이다. 서울시 강우량 자료에서 매년 7월 최대값을 극단 값으로 정의했다. 자료의 개수는 2011년을 제외한 38개이며, 평균은 114.22mm이다. 시간이 지남에 따라 강우량이 증가하는 것을 확인할 수 있다. 2011년 자료는 3절의 마지막에서 분포의 예측에 관해서 다룰 때 사용하겠다.

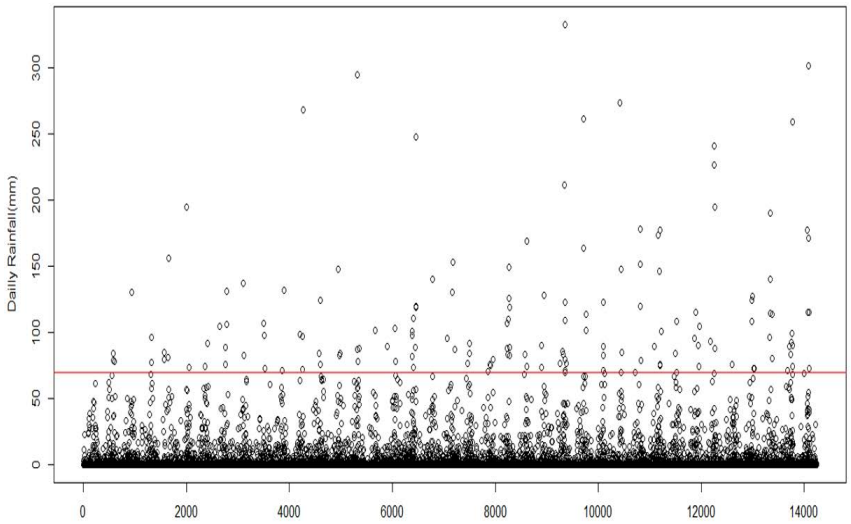

Figure 2.1 Daily rainfall in Seoul (1973 - 2010)

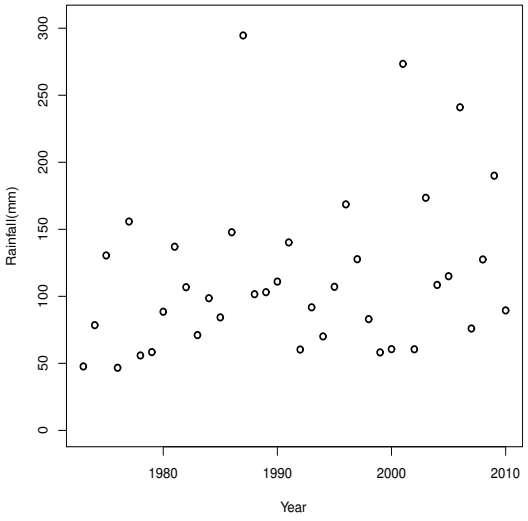

Figure 2.2 Maximum daily rainfall distribution for every July

극단분포는 굼벨, 프레셰와 와이블 분포로 알려져 있으며, 확률밀도함수 f는 아래와 같은 형태로 존재한다.

$$\begin{aligned} Gumbel: f(x) &= \exp\left\{-\exp\left[-\left(\frac{x-\mu}{\sigma}\right)\right]\right\}, -\infty < x < \infty; \\ Frèchet: f(x) &= \begin{cases} 0, & x \geq \mu, \\ \exp\left\{-\left(\frac{x-\mu}{\sigma}\right)^{-\alpha}\right\}, & x > \mu; \end{cases} \\ Weibull: f(x) &= \begin{cases} \exp\left\{-\left[-\left(\frac{x-\mu}{\sigma}\right)^{\alpha}\right]\right\}, & x < \mu, \\ 1, & x \geq \mu. \end{cases} \end{aligned}$$

위의 세 가지 형태의 극단분포와 더불어 확률변수 X가 다음의 확률밀도함수를 가질 때 일반화 극단분포 (generalized extreme value distribution)를 따른다고 하며, 이를 이를 $X \sim GEV(\mu, \sigma, \xi)$로 나타내자.

$$f(x \mid \mu, \sigma, \xi) = \begin{cases} \frac{1}{\sigma}\left\{1+\xi\left(\frac{x-\mu}{\sigma}\right)\right\}^{-\frac{1}{\xi}-1} \exp\left\{-\left(1+\xi\left(\frac{x-\mu}{\sigma}\right)\right)\right\}^{-\frac{1}{\xi}}, & 1+\xi\left(\frac{x-\mu}{\sigma}\right) > 0, \\ 0, & \text{그 외.} \end{cases}$$

위의 세가지 형태의 극단분포를 결합한 분포로 μ는 위치모수이고, σ는 척도모수이고, ξ는 형상모수이다.

그럼 사후예측분포에 대해 알아보자. 우선, θ에 대한 사전분포가 $\pi(\theta)$로 주어지면, θ에 대한 사후분포는 베이즈 정리에 의해 사전 정보와 우도함수의 곱으로 다음과 같이 표현된다.

$$\pi(\theta|\underline{x}) = \frac{\pi(\theta)L(\theta|\underline{x})}{\int_\theta \pi(\theta)L(\theta|\underline{x})d\theta.}$$

여기서 $f(\underline{x})$는 $\underline{x}$의 주변분포이다. 한편, θ에 의존하지 않는 정규화 상수인 $f(\underline{x})$를 생략하면 사후분포는 다음 식과 같다.

$$\pi(\theta|\underline{x}) \propto f(\underline{x}|\theta) \times \pi(\theta).$$

또한, 관측된 자료 $x_1, \cdots, x_n$가 주어졌을 때 미래 관측값 y의 사후예측분포 $f(y|\underline{x})$를 다음과 같이 구할 수 있다.

$$f(y|\underline{x}) = \int f(y|\mu, \sigma, \xi)\pi(\mu, \sigma, \xi|\underline{x})d\mu d\sigma d\xi.$$

하지만 복잡한 수식으로 인해 적분이 불가능할 때, 우리가 사후분포로부터 임의추출한 모수들로부터 사후예측분포를 다음과 같이 추정할 수 있다.

$$\hat{f}(y|\underline{x}) = \frac{1}{M}\sum_{i}^{M} f(y|\mu_i, \sigma_i, \xi_i).$$

마지막으로 MCMC (Markov chain Monte Carlo)방법에 대해 알아보자. 만약 θ의 사후분포로부터 랜덤표본 $\theta_1, \cdots, \theta_k$를 생성할 수 있다면, 적절한 함수 $g(\theta)$ 에 대해 대수의 법칙에 의해 다음과 같이 통계적으로 추정할 수 있다.

$$\lim_{m\to\infty} \frac{1}{m}\sum_{i=1}^{m} g(\theta_i) = E[g(\theta)|x_i] \approx \frac{1}{n}\sum_{i=1}^{m} g(\theta_i).$$

이와 같이 적분을 통계적으로 추정하는 방법을 몬테칼로 기법이라 한다.

깁스 표집 (Gibbs Sampler)은 다차원 결합확률분포가 복잡하여 직접 랜덤표본을 생성하기 어려운 경우 각 변수의 조건부확률분포로부터 랜덤표본을 반복적으로 생성하면 적절한 조건 하에 이들의 극한분포가 결합확률밀도함수가 된다는 사실에 근거하여 난수를 생성하는 방법이다. 이 때 개개의 조건부확률분포로부터의 난수 발생이 쉬워야 한다.

메트로폴리스-헤스팅스 알고리즘 (Metropolis-Hastings algorithm)은 난수를 발생하고자 하는 목표확률분포를 적당한 후보생성밀도함수로부터 후보난수를 추출한다. 그 후 이동확률을 계산하여 난수를 생성하는 방법이다.

그럼 이제까지 알아본 베이지안 추정을 통해 모형을 예측하자. 관측값들이 일반화 극단 분포 $X_i \sim GEV(\mu, \sigma, \xi)$를 따른다고 가정하자. 이 분포의 우도함수는 결합 확률밀도함수와 같으며, Coles와 Pericchi (2003)는 사후분포를 구하기 위해 다음과 같은 사전분포를 가정했다.

$$\mu|\sigma \sim N(0, \sigma^2),$$
$$\sigma \sim IG(\alpha_1, \beta_1),$$
$$\xi \sim G(\alpha_2, \beta_2).$$

이 분포들은 각각 독립이며 큰 분산을 가지고 있다. 이 것은 큰 분산을 가짐으로써 사전분포의 정보를 무시할 수 있는 평평한 분포를 만들기 위함이다. 결합 사전분포와 관심있는 사후분포는 다음과 같다.

$$\pi(\mu, \log(\sigma), \xi) = \pi_\mu(\mu) \times \pi_{\log(\sigma)}(\log(\sigma)) \times \pi_\xi(\xi),$$
$$\pi(\mu, \sigma, \xi|\underline{x}) \propto \pi_\mu(\mu) \times \pi_{\log(\sigma)}(\log(\sigma)) \times \pi_\xi(\xi) \times L(\mu, \sigma, \xi \mid \underline{x}).$$

깁스 표집과 메트로폴리스-헤스팅스 알고리즘을 사용하기 위해서 각 모수의 조건부 사전분포는 다음과 같다.

$$\pi(\mu|\sigma, \xi, \underline{x}) = \pi_\mu(\mu) \times L(\mu, \sigma, \xi \mid \underline{x}),$$
$$\pi(\sigma|\mu, \xi, \underline{x}) = \pi_{\log(\sigma)}(\log(\sigma)) \times L(\mu, \sigma, \xi \mid \underline{x}),$$
$$\pi(\xi|\mu, \sigma, \underline{x}) = \pi_\xi(\xi) \times L(\mu, \sigma, \xi \mid \underline{x}).$$

이 논문에서는 Coles와 Pericchi (2003)와 달리 사전분포들이 계층적모형을 따르는 다음과 같은 분포를 설정했다.

$$\mu|\sigma \sim N(0, \sigma^2),$$
$$\sigma \sim IG(\alpha_1, \beta_1),$$
$$\xi \sim G(\alpha_2, \beta_2).$$

이 때, 결합 사전분포는 다음과 같이 간단히 표현 할 수 있다.

$$\pi(\mu, \sigma, \xi) \propto \Big(\frac{1}{\sigma}\Big)^{\alpha_1+2} \xi^{\alpha_2-1} exp\Big(-\frac{\mu^2}{2\sigma^2} - \frac{\beta_1}{\sigma} - \frac{\xi}{\beta_2}\Big). \tag{2.1}$$

또한 관심 있는 사후분포는 다음과 같이 정의된다.

$$\begin{aligned}\pi(\mu, \sigma, \xi|\underline{x}) \propto \Big(\frac{1}{\sigma}\Big)^{n+\alpha_1+2} \xi^{\alpha_2-1} \prod_{i=1}^{n} \Big\{1 + \xi\frac{(x_i-\mu)}{\sigma}\Big\}^{-1/\xi-1} \\ \times exp\Big[-\sum_{i=1}^{n}\Big\{1 + \xi\frac{(x_i-\mu)}{\sigma}\Big\}^{-1/\xi-1} - \frac{\mu^2}{2\sigma^2} - \frac{\beta_1}{\sigma} - \frac{\xi}{\beta_2}\Big].\end{aligned} \tag{2.2}$$

이 분포는 우리가 알고 있는 일반적인 분포의 모양이 아니지만 MCMC 알고리즘을 통해서 이 분포를 따르는 난수를 만들 수 있다. 깁스 표집과 메트로폴리스-헤스팅스 알고리즘은 각 모수의 조건부 확률분포로부터 임의 추출된다. 사후분포로부터 각 모수의 조건부 확률분포는 다음과 같이 표현된다.

$$\pi(\mu|\sigma,\xi,\underline{x}) \propto \prod_{i=1}^{n}\Big\{1+\xi\frac{(x_i-\mu)}{\sigma}\Big\}^{-\frac{1}{\xi}-1} \times exp\Big[-\sum_{i=1}^{n}\Big\{1+\xi\frac{(x_i-\mu)}{\sigma}\Big\}^{-\frac{1}{\xi}-1} - \frac{\mu^2}{2\sigma^2}\Big]. \quad (2.3)$$

$$\pi(\sigma|\mu,\xi,\underline{x}) \propto \Big(\frac{1}{\sigma}\Big)^{n+\alpha_1+2}\prod_{i=1}^{n}\Big\{1+\xi\frac{(x_i-\mu)}{\sigma}\Big\}^{-\frac{1}{\xi}-1} \times exp\Big[-\sum_{i=1}^{n}\Big\{1+\xi\frac{(x_i-\mu)}{\sigma}\Big\}^{-\frac{1}{\xi}-1} - \frac{\mu^2}{2\sigma^2} - \frac{\beta_1}{\sigma}\Big]. \quad (2.4)$$

$$\pi(\xi|\mu,\sigma,\underline{x}) \propto \xi^{\alpha_2-1}\prod_{i=1}^{n}\Big\{1+\xi\frac{(x_i-\mu)}{\sigma}\Big\}^{-\frac{1}{\xi}-1} \times exp\Big[-\sum_{i=1}^{n}\Big\{1+\xi\frac{(x_i-\mu)}{\sigma}\Big\}^{-\frac{1}{\xi}-1} - \frac{\xi}{\beta_2}\Big]. \quad (2.5)$$

위와 같은 조건부 사후분포로부터 깁스 표집과 메트로폴리스-헤스팅스 알고리즘을 활용하여 임의추출된 난수로 사후예측분포를 생성할 수 있다.

3. 사례 연구

2010년도를 기준으로 1991년부터 20년동안, 1990년부터 21년동안, 최대 1973년부터 38년 동안의 7월 중 최대 강우량 자료를 가지고 19개의 사후예측분포를 생성 시켰다. 사전분포의 모수의 변화에 따른 분포의 차이는 차이는 Figure 3.1과 같다. Figure 3.1은 1991년부터 20년 동안의 사후예측분포를 모수의 변화에 따라 나타내본 것이다. ξ의 모수변화보다 σ의 모수변화에 따라 민감한 것 처럼 보인다. 여기서는 σ의 사전분포로 역감마 분포 α_1이 3이고 β_1이 2인 분포, ξ는 감마 분포 α_2가 3이고 β_2가 2인 분포로 임의 지정하고, 깁스 표집의 반복횟수는 10,000번이고, 그 중 수렴할때까지의 기간을 고려하여 초기값 5,000번은 버렸다.

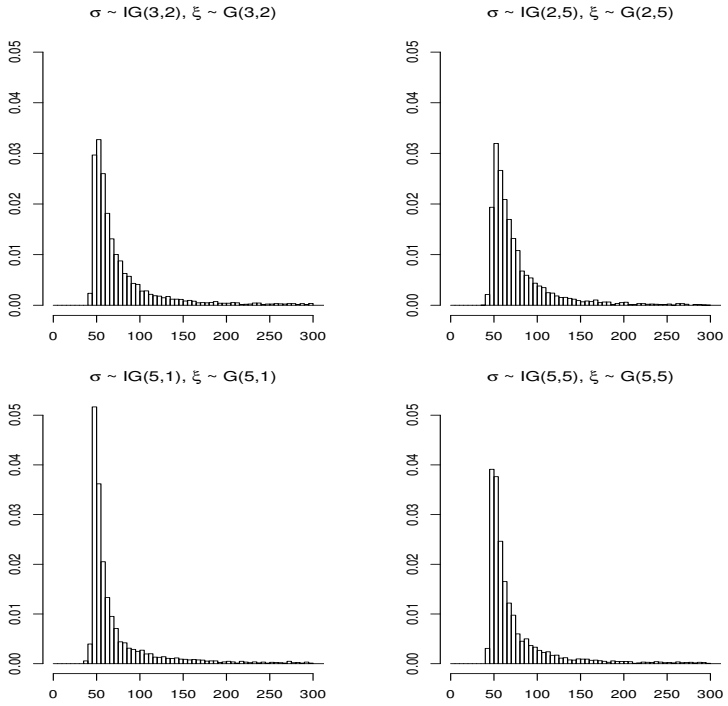

Figure 3.1 Posterior predicted densities (1991-2010) for each parameter

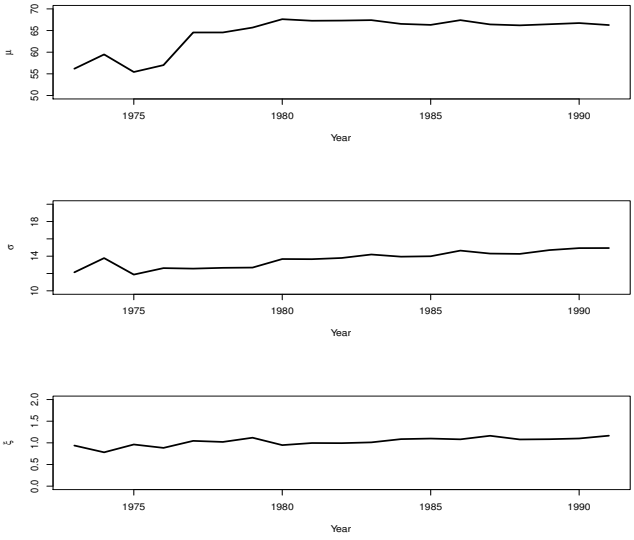

Figure 3.2 Posterior means of parameters

Figure 3.2는 각 모수의 기간별 사후평균이다. μ의 사후 평균의 변화를 보면 38년 동안의 분포에서 56.19였으나 크고 작은 변동을 거쳐 20년 동안의 66.27로 증가하였으며, σ는 12.14에서 14.94으로 증가했다. 이것을 통해 최근 20년 동안의 자료의 사후예측분포가 38년 동안의 자료의 분포보다 강우량이 증가하고 변동이 커질 것으로 보인다. 또한 1977년이 포함되는 34년간의 분포와 1976년 이전을 포함하는 분포와 차이가 큰데, 모수의 변화 시점이 아닌가 생각한다.

Figure 3.3 Point estimations of posterior predictive distribution

Figure 3.3을 통해서 분포들의 점추정값에 대해 알아보자. 왜도가 큰 일반화 극단분포에서 영향력이 큰 관측값과 이상치에 민감한 사후 평균의 경우 좋은 통계량이라 하기에 부족하다. 사후분포에서 MAP와 같은 개념의 사후예측분포의 MAPP (most a posteriori predictive probability)와 사후예측 중앙값에 대해 살펴보자. μ가 증가함에 따라 MAPP에 증가함을 볼 수 있다. 사후예측 중앙값 또한 지속적으로 증가하였다. 이 추정값들을 통해 강우량이 과거에 비해 최근에 상대적으로 크다고 할 수 있다.

그럼 기간별 모형을 가지고 2011년도의 7월 최대값의 포함 확률을 알아보자. $100(1-\alpha)\%$ 포함 확률을 가지는 미래 관측값 y의 신용구간 중 크기가 가장 작은 신뢰구간을 선택하자. y에 대한 $100(1-\alpha)\%$ HPPD (highest posterior predictive density) 신뢰구간은 다음과 같다.

$$C = \{y | f(y|\underline{x}) \geq k(\alpha)\}.$$

여기서 $k(\alpha)$는 $P(y \in C|\underline{x}) \geq 1-\alpha$를 만족하는 가장 큰 상수이다. 이것은 사후분포의 HPD 신용구간과 같은 의미이다.

각 분포의 95% HPPD 신뢰구간을 구해보면 Figure 3.4과 같다. 2011년도 7월의 최대값은 301.5mm임을 고려하였을 때, 95%내에 포함하기 시작하는 1977년도를 기점으로, 신뢰구간이 넓어짐을 알 수 있다. HPPD 신뢰 구간의 넓어짐은 집중호우의 확률이 증가하였다고 할 수 있고, 이 것을 통해 최근 강우량의 분포가 변화되었음을 알 수 있다.

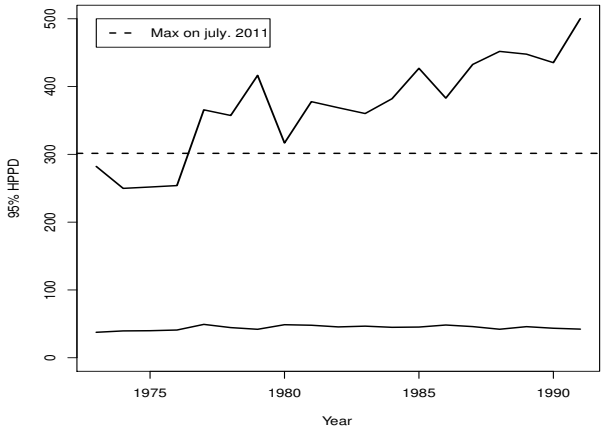

Figure 3.4 95% HPPD credible intervals with maximum rainfall in July, 2011

4. 결 론

강우량 자료에 대해 일반화 극단분포를 모형으로 하고 사전분포는 계층적 모형을 적용 사후예측분포를 구하고자 하였다. 실제 서울특별시의 38년 동안의 자료를 통해 생성한 사후예측분포와 2011년도의 실제자료를 비교하여 모형이 얼마나 적합한지, 기후변화가 어떻게 일어났는지에 대해 알아보았다. 7월 강우량의 예측 모형에서 사전분포의 모수의 변화에 따른 모형의 변화는 미비하였다. 그리고 최근 20년간의 자료만으로도 충분한 포함 확률을 확보하였으며, 이를 통해서 집중호우의 확률이 증가한 것을 알 수 있다.

차후에는 모수의 사전분포를 다양하게 변화시켜서 포함 확률을 최적화할 수 있는 방안을 검토하겠으며, 강수량이 변화한 시점을 찾아보는 것도 시도할 것이다.

References

Aitchison, J. and Dunsmore, I. R. (1975). *Statistical prediction analysis*, Cambridge University Press, Cambridge, UK.

Coles, S. G. and Pericchi, L. (2003). Anticipating catastrophes through extreme value modelling. *Journal of Applied Statistics*, **52**, 405-416.

Coles, S. G. and Tawn, J. A. (1996). A Bayesian analysis of extreme rainfall data. *Journal of Applied Statistics*, **45**, 463-478.

Engelund, S. and Rackwitz, R. (1992). On predictive distribution functions for the three asymptotic extreme value distriibutions. *Journal of Structural Safety*, **11**, 255-258.

Hosking, J. R. M. (1990). Ł-moments: Analysis and estimation of distributions using linear combinations of order statistics. *Journal of the Royal Statistical Society B*, **52**, 105-124.

Kotz, S. and Nadarajah, S. (2001). *Extreme value distributions: Theory and applications*, Imperial College Press, London.

National Emergency Management Agency (2012). *The annals of disasters for 2011*, Recovery support divison, National Emergency Agency, Seoul.

Smith, E. (2005). *Bayesian modelling of extreme rainfall data*, Ph. D. Thesis, School of Philosophy, University of Newcastle, Newcastle, UK.

Smith, R. L. (1985). Maximum likelihood estimation in a class of non-regular cases. *Biometrika*, **72**, 67-92.

Journal of the Korean Data & Information Science Society
2013, **24**(4), 857–865

http://dx.doi.org/10.7465/jkdi.2013.24.4.857
한국데이터정보과학회지

Prediction of extreme rainfall with a generalized extreme value distribution

Yong Kyu Sung[1] · Joong K. Sohn[2]

[12]Department of Statistics, Kyungpook National University

Received 21 April 2013, revised 5 July 2013, accepted 18 July 2013

Abstract

Extreme rainfall causes heavy losses in human life and properties. Hence many works have been done to predict extreme rainfall by using extreme value distributions. In this study, we use a generalized extreme value distribution to derive the posterior predictive density with hierarchical Bayesian approach based on the data of Seoul area from 1973 to 2010. It becomes clear that the probability of the extreme rainfall is increasing for last 20 years in Seoul area and the model proposed works relatively well for both point prediction and predictive interval approach.

Keywords: Extreme rainfall, generalized extreme value distribution, hierarchical Bayes, highest posterior predictive density credible region, Markov chain Monte Carlo, most a posteriori predictive probability, predictive posterior density.

[1] Master, Department of Statistics, Kyungpook National University, Daegu 702-701, Korea.
[2] Corresponding author: Professor, Department of Statistics, Kyungpook National University, Daegu 702-701, Korea. E-mail: jsohn@knu.ac.kr