The Korean Journal of Applied Statistics (2020) DOI: http://dx.doi.org/10.5351/KJAS.2020.33.1.107
33(1), 107–114

Comparison of log-logistic and generalized extreme value distributions for predicted return level of earthquake

Nak Gyeong Ko[a] · Il Do Ha[a,1] · Dae Heung Jang[a]

[a]Department of Statistics, Pukyong National University

(Received January 9, 2020; Revised January 11, 2020; Accepted January 11, 2020)

Abstract

Extreme value distributions have often been used for the analysis (e.g., prediction of return level) of data which are observed from natural disaster. By the extreme value theory, the block maxima asymptotically follow the generalized extreme value distribution as sample size increases; however, this may not hold in a small sample case. For solving this problem, this paper proposes the use of a log-logistic (LLG) distribution whose validity is evaluated through goodness-of-fit test and model selection. The proposed method is illustrated with data from annual maximum earthquake magnitudes of China. Here, we present the predicted return level and confidence interval according to each return period using LLG distribution.

Keywords: earthquake, log-logistic distribution, goodness-of-fit test, model selection, predicted return level

1. 서론

극단값 이론에 따르면 (Fisher와 Tippett, 1928), 표본의 수가 충분히 큰 경우 연속적인 블록 최댓값(block maxima)들은 점근적으로 일반화 극단값(generalized extreme value; GEV) 분포를 따른다. 따라서 지진이나 홍수와 같은 자연 재해로부터 관측되는 자료를 대상으로 재현 수준(return level) 예측 등과 같은 자료 분석을 위해 GEV 분포가 자주 사용되어 왔다 (Nadarajah와 Choi, 2007; Pisarenko 등, 2010; Lee 등, 2014; Bae 등, 2018). 여기서 Nadarajah와 Choi (2007) 그리고 Lee 등 (2014)는 강우량 자료에 대한 재현수준 예측을 위해 GEV 분포를 사용하였고, Pisarenko 등 (2010)와 Bae 등 (2018)은 지진자료에 대해 이와 같은 분석을 위해 GEV 분포를 사용 하였다. 하지만 소 표본인 경우 이러한 블록 최댓값들은 GEV 분포를 따르지 않을 수도 있다. 본 논문에서는 이러한 문제점을 해결하기 위해 모형 적합도 검정 및 모형 선택을 통해 로그-로지스틱(log-logistic; LLG) 분포의 사용을 제안한다.

LLG 분포는 양의 실수 값을 가지는 자료(예: 실패시간)를 모형화 하는데 사용된다. 예를 들어, 기존 문헌에 의하면 수문 자료(hydrological data) (Ashkar와 Mahdi, 2003) 및 신뢰수명 자료(reliability data) (Akhtar와 Khan, 2014)의 분석에 LLG 분포를 적용하여 왔다. 제안된 방법의 한 예증을 위해 중국의

This work was supported by the Korea Institute of Energy Technology Evaluation and Planning (KETEP) and the Ministry of Trade, Industry & Energy (MOTIE) of the Republic of Korea (No. 20171510101960).
[1]Corresponding author: Department of Statistics, Pukyong National University, 45, Yongso-ro, Nam-Gu, Busan 48513, South Korea. E-mail: idha1204@gmail.com

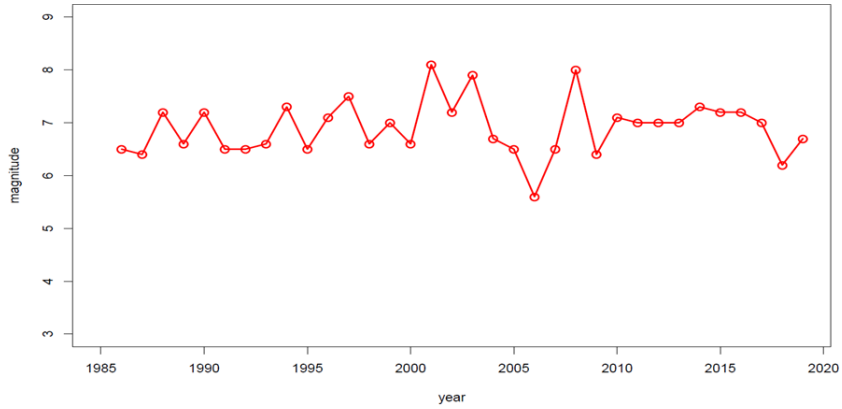

Figure 2.1. Time series plot for annual maximum earthquake magnitudes of China.

연별 최대 지진규모 자료를 대상으로 하여 LLG 분포와 GEV 분포에 대해 적합도 검정과 모형 선택을 실시한 후, LLG 분포가 보다 적절한 모형임을 보인다. 이를 통해 LLG 분포를 이용하여 재현 기간별 지진의 재현수준 예측 및 신뢰구간을 제시한다.

본 논문의 구성은 다음과 같다. 2절에서는 자료소개 및 기본적 분석을 제시한다. 3절에서는 연구방법을 위한 GEV 분포와 LLG 분포를 기술하고, 각 분포의 모수를 추정한다. 4절에서는 두 분포의 비교를 위해 적합도 검정 및 모형 선택을 실시하고, 지진규모 예측을 제시한다. 마지막으로 5절에서는 분석 결과에 대해 토론한다. 본 분석에 사용된 모든 계산은 R 프로그램을 이용하였다.

2. 연구자료 및 기초분석

2.1. 연구 자료의 설명

분석에 사용되는 자료는 총 34개의 소표본 자료로 1986년 1월에서 2019년 7월까지 중국에서 발생한 연별 최대 규모 데이터(annual maximum earthquake magnitudes)이다. 중국 지진국 홈페이지(http://data.earthquake.cn/gcywfl/index.html)는 2009년부터 발생한 규모 3이상의 지진 자료만을 제공한다. 1986년부터 2008년의 자료는 "중국지진역사기록표"(http://www.docin.com/p-1168791466.html)에서 얻었다.

2.2. 기초 분석

Figure 2.1은 연구대상 자료에 대한 시계열 그래프이다. 연별 최대지진이 규모 5에서 8사이에서 주로 발생하여 왔음을 알 수 있다. 특히 2000년에서 2010년 사이에 규모 8정도의 매우 큰 규모의 지진이 세 번 발생하였음을 보여준다. 특히 이 세 곳은 중국의 칭하이 (2001년 규모 8.1), 알타이 (2003년 규모 7.9), 그리고 쓰촨성 (2008년 규모 8.0) 지역이다.

2.3. 정상성 경향 검토

본 자료는 시계열 자료 (Figure 2.1)이므로 정상성(stationarity) 여부를 조사하기 위해 경향성 분석 및 단위근 검정(unit root test)을 실시하였다. 먼저 경향성 분석을 위해 만-켄달 검정법(Mann-Kendall test)을 사용하였다 (H_0: 단조 추세가 존재하지 않는다). Table 2.1의 결과에 의하면, 본 자료는 유의수준 5%에서 단조 추세가 존재하지 않음을 알 수 있다 (p-value = 0.464). 다음으로 AR(1) 모형하에

Table 2.1. Mann-Kendall trend test and Dickey-Fuller's unit root test for annual maximum earthquake magnitudes of China

Test	Test statistic	p-value
Mann-Kendall	0.093	0.464
Dickey-Fuller	−5.484	0.010

Table 3.1. Estimation of parameters in the generalized extreme value distribution of annual maximum earthquake magnitudes of China

$\hat{\mu}$		$\hat{\sigma}$		$\hat{\xi}$	
Est	SE	Est	SE	Est	SE
6.71	0.094	0.50	0.063	−0.23	0.098

Est = estimate; SE = estimated standard error.

서 Dickey-Fuller의 단위근 검정법을 실시 (H_0: 단위근이 존재한다. 즉 비정상성이다)한 결과 유의수준 5%에서 정상성으로 나타났다 (p-value = 0.01). 따라서 본 논문에서는 정상성에 기초하여 자료 분석을 실시하고자 한다.

3. 일반화 극단값 분포와 로그-로지스틱 분포

3.1. 일반화 극단값 분포

블록 최댓값 모형은 일정한 단위 기간 동안의 최댓값과 같은 극단값을 모형화하는 대표적인 방법이다. Fisher-Tippett 정리 (1928)에 따르면 연속적인 블록에서 나타나는 최댓값들의 분포는 표본의 크기가 증가함에 따라 점근적으로 GEV 분포로 수렴한다. GEV 분포의 누적분포함수 (Jenkinson, 1955)는 다음과 같이 정의된다.

$$G(x) = \exp\left[-\left\{1+\xi\left(\frac{x-\mu}{\sigma}\right)\right\}^{-\frac{1}{\xi}}\right], \quad x : 1+\xi\left(\frac{x-\mu}{\sigma}\right) > 0,$$

여기서 μ는 실수 값을 갖는 위치(location)모수, σ는 양의 실수 값을 갖는 척도(scale)모수, 그리고 ξ는 실수 값을 갖는 형상(shape)모수이다. GEV 분포에서 x의 서포트(support)는 $1+\xi((x-\mu)/\sigma) > 0$의 제약조건 하에서 정의된다. 특히 형상모수 ξ는 GEV 분포의 형태를 결정하는 모수로서, 그 크기에 따라 $\xi = 0$일 때 Gumbel 분포, $\xi > 0$일 때 Frechet 분포, 그리고 $\xi < 0$일 때 Weibull 분포가 된다. 본 논문에서는 GEV 분포의 모수를 추정하기 위해 최대가능도 추정법(maximum likelihood estimation)을 사용하였다. 블록의 개수가 m일 때 GEV 분포의 로그 가능도함수(log-likelihood function)는 다음과 같이 주어진다.

$$\ell(\mu,\sigma,\xi) = -m\log\sigma - \left(1+\frac{1}{\xi}\right)\sum_{i=1}^{m}\log\left[1+\xi\left(\frac{x_i-\mu}{\sigma}\right)\right] - \sum_{i=1}^{m}\left[1+\xi\left(\frac{x_i-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}},$$

여기서 $i = 1,\ldots,m$에 대해 $1+\xi((x_i-\mu)/\sigma) > 0$이다. 따라서 세 모수의 추정에 대한 결과는 Table 3.1과 같다. 여기서 $\hat{\xi} = -0.23$으로 음수이므로 본 연구 자료에 대해 GEV 분포를 적합 하는 경우 Weibull 분포 형태로 나타남을 알 수 있다.

3.2. 로그-로지스틱 분포

로그-로지스틱(LLG) 분포는 $\log(X)$의 분포가 로지스틱 분포(logistic distribution)일 때 X는 로그-로지스틱 분포를 따르는 것으로서 LLG 분포는 로그-정규(log-normal)분포와 유사하지만 하나의 차이는

Table 3.2. Estimation of parameters in the log-logistic distribution of annual maximum earthquake magnitudes of China

$\hat{\alpha}$		$\hat{\beta}$	
Est	SE	Est	SE
6.88	0.085	24.1	3.420

Est = estimate; SE = estimated standard error.

꼬리가 더 두터운(heavy-tailed) 분포이다. 특히 로그-정규분포와 달리 LLG의 누적분포함수는 아래와 같이 명확한 형태(closed form)로 주어진다.

$$G(x) = \frac{\left(\frac{x}{\alpha}\right)^{\beta}}{1 + \left(\frac{x}{\alpha}\right)^{\beta}}, \quad x > 0,$$

여기서 α와 β는 각각 양의 실수 값을 갖는 척도모수와 형상모수이다. 본 논문에서는 GEV 분포에서와 같이 최대 가능도 추정법을 LLG 분포의 모수 추정에 사용한다. 대응하는 로그 가능도함수는 다음과 같이 주어진다.

$$\ell(\alpha, \beta) = n \log\left(\frac{\beta}{\alpha}\right) + (\beta - 1) \sum_{i=1}^{n} \log\left(\frac{x_i}{\alpha}\right) - 2 \sum_{i=1}^{n} \log\left[1 + \left(\frac{x_i}{\alpha}\right)^{\beta}\right].$$

이를 통해 LLG 분포의 모수 추정의 결과는 Table 3.2와 같다.

4. 비교 분석

본 절에서는 제안된 방법의 예증을 위해 2절의 중국의 지진자료를 대상으로 하여 적합도 검정 및 모형 선택을 실시하여 LLG 분포 사용의 타당성을 보이고자 한다.

4.1. 적합도 검정 및 모형 선택

4.1.1. 모형의 적합도 검정 적합도 검정(goodness-of-fit tests)이란 주어진 표본자료가 모집단의 특정 확률분포에 얼마나 적합 하는지를 판단하는 방법이다. 이를 위해 본 논문에서는 대상 연구 자료에 대해 얻어진 경험적 분포(empirical distribution)와 우리가 가정한 모집단의 확률분포(즉 LLG, GEV)가 얼마나 잘 일치하는지를 검정하기 위해, 다음의 세 가지 유용한 적합도 검정법을 사용한다.

i) Kolmogorov-Smirnov (KS) 검정

ii) Anderson-Darling (AD) 검정

iii) Cramer-von Mises (CVM) 검정

Table 4.1의 검정 결과에 의하면, 유의수준 5%를 기준으로 GEV 분포에 대해 KS와 CVM 검정은 만족하지만 AD 검정에서는 만족하지 않음을 알 수 있다 (p-value = 0.043). 하지만 LLG 분포는 세 가지 검정법 모두 만족함을 확인할 수 있다. 뿐만 아니라 Figure 4.1에서도 GEV 분포보다 LLG 분포가 대상 자료에 더 잘 적합함을 확인할 수 있다. 특히 Figure 4.1의 왼쪽 그림(a)의 경우 LLG 분포는 경험적 누적분포함수(empirical cumulative distribution function; ECDF)를 전반적으로 잘 따라가지만, GEV 분포는 규모 값이 큰 경우 눈에 띄게 벗어남을 알 수 있다.

Table 4.1. The p-values of goodness-of-fit tests for GEV and LLG distributions of annual maximum earthquake magnitudes of China

p-value	KS	AD	CVM
GEV	0.686	0.043	0.065
LLG	0.520	0.065	0.056

GEV = generalized extreme value; LLG = log-logistic; KS = Kolmogorov-Smirnov; AD = Anderson-Darling; CVM = Cramer-von Mises.

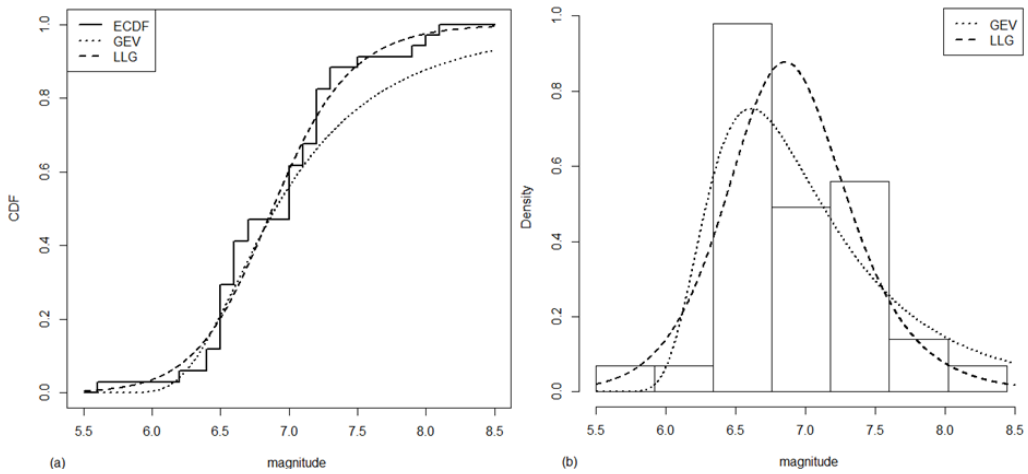

Figure 4.1. (a): Plots of empirical cumulative distribution function (ECDF) versus fitted GEV and LLG distributions for the China data; (b): histogram of the China data versus graphs of fitted GEV and LLG distributions. GEV = generalized extreme value; LLG = log-logistic.

Table 4.2. AIC and BIC for GEV and LLG distributions of annual maximum earthquake magnitudes of China

p-value	AIC	BIC
GEV	57.08	61.66
LLG	54.48	57.53

AIC = Akaike's information criterion; BIC = Bayesian information criterion; GEV = generalized extreme value; LLG = log-logistic.

4.1.2. 모형 선택 모형 선택을 위해 본 논문에서는 일반적으로 자주 사용되는 두 가지 기준인 다음의 Akaike's information criterion (AIC)와 Bayesian information criterion (BIC)를 사용한다.

i) $\text{AIC} = -2\ell + 2k$,

ii) $\text{BIC} = -2\ell + k\log(n)$

여기서 ℓ은 로그가능도함수, k는 적합된 모수의 개수, n은 자료의 개수이다. 이러한 두 가지 모형 선택 기준들은 적합된 모수의 개수가 증가함에 따라 벌점(penalty)을 주는 방식으로 그 값이 작은 모형을 적절한 모형으로 선택해 준다. Table 4.2에서 보는 바와 같이 GEV와 LLG 분포에 두 모형선택 방법을 적용한 결과, AIC와 BIC 모두 LLG 분포에서 더 작은 값을 가지므로 모형 선택에서도 LLG가 더 적절한 분포임을 확인 할 수 있다. 따라서 다음 절의 지진규모의 예측에 LLG 분포를 사용하고자 한다.

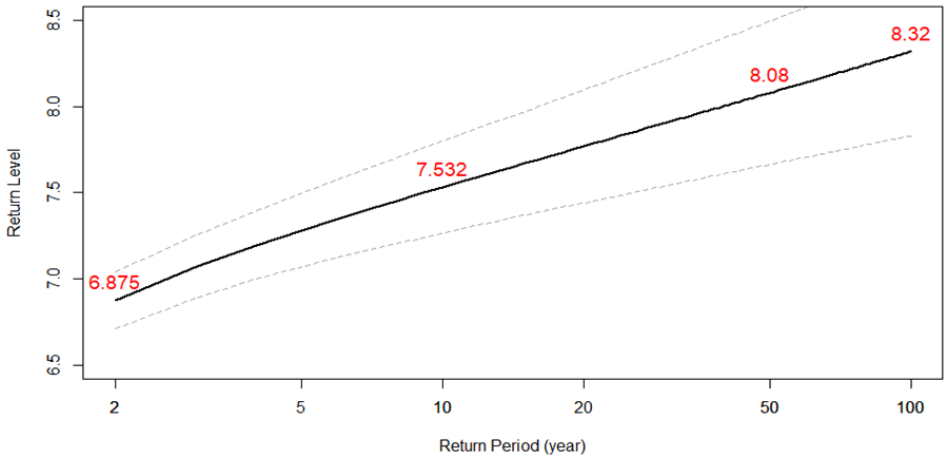

Figure 4.2. Predicted return level and confidence interval according to return period for the China data.

4.2. 지진 규모 예측

재현수준이란 극단적인 사건이 발생할 위험성을 정량적으로 표현한 값이다 (Ryu 등, 2016). 어떤 사건이 발생할 초과확률을 p라 할 때, 초과확률의 역수를 재현기간(T)이라고 정의한다. 따라서 어떤 확률변수 X에 대해 재현기간이 T년인 재현수준은 $\Pr(X \geq x_T) = p\ (= 1/T)$를 만족하는 분위수(quantile) x_T이다. 따라서 확률변수 X가 LLG 분포를 따르는 경우 재현수준 x_T는 다음과 같이 쉽게 얻어진다.

$$x_T = \alpha(T-1)^{\frac{1}{\beta}}.$$

4.2.1. 최대가능도 방법 재현 수준에 따른 신뢰구간은 최대가능도 추정량(maximum likelihood estimator; MLE)의 점근 정규성과 델타방법(delta method)에 의해 다음과 같이 계산 된다 (Ashkar과 Mahdi, 2003).

$$\mathrm{Var}(\hat{x}_T) = \left(\frac{\partial x_T}{\partial \alpha}\right)^2 \mathrm{Var}(\hat{\alpha}) + \left(\frac{\partial x_T}{\partial \beta}\right)^2 \mathrm{Var}\left(\hat{\beta}\right) + 2\left(\frac{\partial x_T}{\partial \alpha}\right)\left(\frac{\partial x_T}{\partial \beta}\right)\mathrm{Cov}\left(\hat{\alpha}, \hat{\beta}\right),$$

여기서 $\partial x_T/\partial \alpha = (T-1)^{1/\beta}$와 $\partial x_T/\partial \beta = -\left(\alpha/\beta^2\right)(T-1)^{1/\beta}\log(T-1)$이다. 이를 이용하여 재현기간에 따른 지진 규모를 예측한 결과는 Figure 4.2와 같으며, 재현 기간에 따라 증가하는 추세임을 확인할 수 있다.

4.2.2. 붓스트랩 방법 최대가능도(maximum likelihood; ML) 방법으로 얻은 재현 수준의 신뢰구간 결과를 비모수적 붓스트랩(Bootstrap)방법 ("fitdistrplus" R package)의 결과와 비교하였다. Table 4.3의 결과에 의하면 두 방법 (ML과 Bootstrap의 방법)이 거의 동일한 결과를 보여줌을 알 수 있다. 특히 Table 4.3의 ML방법에 의하면 100년의 재현기간에서 지진규모의 재현수준의 95% 신뢰구간이 (7.830, 8.809)로 지진 위험성이 매우 높은 편으로 나타난다.

5. 토론 및 향후과제

본 논문에서는 소표본의 경우 블록 최댓값들에 대한 GEV 분포의 한 대안으로 LLG분포의 사용을 제안하였다. 이를 위해 중국 연별 최대 규모 자료를 이용하여 적합된 두 분포를 적합도 검정과 모형 선택방

Table 4.3. Estimation of return level and confidence interval according to return period for the China data by ML and bootstrap methods

Return period	Return level (Confidence interval)	
	ML method	Bootstrap method
2 year	6.875 (6.708, 7.043)	6.875 (6.702, 7.049)
10 year	7.532 (7.264, 7.799)	7.532 (7.254, 7.788)
50 year	8.080 (7.664, 8.497)	8.080 (7.676, 8.475)
100 year	8.320 (7.830, 8.809)	8.320 (7.848, 8.782)

ML = maximum likelihood.

법을 실시하여 LLG 분포의 적절성을 보였다. 이에 따라 4절에서 LLG 분포을 이용하여 지진의 재현수준 예측을 제시하였다.

2.1절에서 소개한 데이터가 비록 시계열자료이지만 블록(block)이 매우 긴 연별 최대값 자료이기 때문에 자료 간 독립일 가능성이 높다 (Coles 등, 2001, p.54). 따라서 본 논문에서는 자기상관성(autocorrelation)을 고려하지 않고 자료 분석을 실시하였다.

향후 과제로는 지진규모의 예측에 유용한 또 다른 확률분포로서 일반화 파레토 분포(generalized Pareto distribution) (Bae 등, 2018)를 이용하여 LLG 분포 및 GEV 분포와 함께 모형검토 및 재현수준 예측을 비교하고자 한다. 나아가 한국을 비롯한 일본의 지진자료도 이와 같은 방법으로 분석을 실시하는 것도 흥미 있는 향후 연구과제가 될 것으로 사료된다.

References

Ashkar, F. and Mahdi, S. (2003). Comparison of two fitting methods for the log-logistic distribution, *Water Resources Research*, **39**, 1–8.

Akhtar, M. T. and Khan, A. A. (2014). Log-logistic distribution as a reliability model: a Bayesian analysis, *American Journal of Mathematics and Statistics*, **4**, 162–170.

Bae, E., Kim, S. Y., and Kim, C. (2018). Extreme value theory based earthquake risk modelling and management, *The Journal of Risk Management*, **29**, 1–24.

Coles, S., Bawa, J., Trenner, L., and Dorazio, P. (2001). *An Introduction to Statistical Modeling of Extreme Values*, Springer, London.

Fisher, R. A. and Tippett, L. H. C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. In , *Mathematical Proceedings of the Cambridge Philosophical Society*, **24**, 180–190.

Jenkinson, A. F. (1955). The frequency distribution of the annual maximum (or minimum) values of meteorological elements, *Quarterly Journal of the Royal Meteorological Society*, **81**, 158–171.

Lee, J. J., Kim, N. H., Kwon, H. J., and Kim, Y. (2014). A Bayesian analysis of return level for extreme precipitation in Korea, *Journal of the Korean Data and Information Science Society*, **27**, 947–958.

Nadarajah, S. and Choi, D. (2007). Maximum daily rainfall in South Korea, *Journal of Earth System Science*, **116**, 311–320.

Pisarenko, V. F., Sornette, D., and Rodkin, M. V. (2010). Distribution of maximum earthquake magnitudes in future time intervals: application to the seismicity of Japan (1923–2007). *Earth, Planets and Space*, **62**, 567.

Ryu, S., Eom, E., Kwon, T., and Yoon, S. (2016). The estimation of CO concentration in Daegu-Gyeongbuk area using GEV distribution, *Journal of the Korean Data and Information Science Society*, **27**, 1001–1012.

지진 재현수준 예측에 대한 로그-로지스틱 분포와 일반화 극단값 분포의 비교

고낙경[a] · 하일도[a,1] · 장대흥[a]

[a]부경대학교 통계학과

(2020년 1월 9일 접수, 2020년 1월 11일 수정, 2020년 1월 11일 채택)

요 약

자연 재해로부터 관측되는 자료를 대상으로 재현 수준 예측 등과 같은 자료 분석을 위해 일반화 극단값 분포(generalized extreme value)가 자주 사용되어 왔다. 표본 수가 충분히 큰 경우 연속적인 블록 최댓값들은 점근적으로 일반화 극단값 분포를 따른다. 하지만 소표본인 경우 이러한 사실은 성립되지 않을 수도 있다. 본 논문에서는 이러한 문제점을 해결하기 위해 모형 적합도 검정 및 모형 선택을 통해 로그-로지스틱(log-logistic) 분포의 사용을 제안한다. 하나의 예증으로서 중국 지진 자료를 대상으로 하여 로그-로지스틱 분포를 이용하여 재현 기간별 재현 수준 예측 및 신뢰구간을 제시한다.

주요용어: 지진, 로그-로지스틱 분포, 적합도 검정, 모형 선택, 재현 수준 예측

본 연구는 산업통상자원부(MOTIE)와 한국에너지기술평가원 (KETEP)의 지원을 받아 수행한 연구 과제임 (No. 20171510101960).

[1]교신저자: (48513) 부산광역시 남구 용소로 45, 부경대학교 통계학과. E-mail: idha1204@gmail.com