The Journal of the Convergence on Culture Technology (JCCT)
Vol. 8, No. 1, pp.531-536, January 31, 2022. pISSN 2384-0358, eISSN 2384-0366

http://dx.doi.org/10.17703/JCCT.2022.8.1.531

JCCT 2022-1-60

극단값 분포 추정을 위한 모수적 비모수적 방법

Parametric nonparametric methods for estimating extreme value distribution

우승현*, 강기훈**

Seunghyun Woo*, Kee-Hoon Kang**

요 약 본 논문은 꼬리가 두꺼운 분포의 꼬리부분에 대한 분포를 추정할 경우 모수적 방법과 비모수적 방법의 성능에 대해 비교하였다. 모수적 방법으로는 일반화 극단값 분포와 일반화 파레토 분포를 이용하였고, 비모수적 방법은 커널형 확률밀도함수 추정방법을 적용하였다. 두 접근법의 비교를 위해 2014년부터 2018년까지 서울시 관측소별 일일 미세먼지 공공데이터를 이용하여 블록 최댓값 모형과 분계점 초과치 모형을 적용하여 함수 추정한 결과를 함께 보이고 2년, 5년, 10년의 재현수준을 통해 고농도의 미세먼지가 일어날 지역을 예측하였다.

주요어 : 분계점 초과치 모형, 블록 최댓값 모형, 비모수적 추정법, 일반화 극단값 분포, 커널 함수 추정, 평활량 선택

Abstract This paper compared the performance of the parametric method and the nonparametric method when estimating the distribution for the tail of the distribution with heavy tails. For the parametric method, the generalized extreme value distribution and the generalized Pareto distribution were used, and for the nonparametric method, the kernel density estimation method was applied. For comparison of the two approaches, the results of function estimation by applying the block maximum value model and the threshold excess model using daily fine dust public data for each observatory in Seoul from 2014 to 2018 are shown together. In addition, the area where high concentrations of fine dust will occur was predicted through the return level.

Key words : Bandwidth selection, Block maximum model, Generalized extreme value distribution, Kernel function estimation, Nonparametric estimation, Peak over threshold model

Ⅰ. 서 론

변동성이 큰 자료에서 극단적인 변화 또는 예측에 관한 연구들이 활발히 진행되고 있고 극단값에 해당되는 확률밀도함수의 꼬리확률을 추정하는 것에 최근에 관심이 많아지고 있다. 특히, 환경 분야에서 재해 방지 및 관리를 위한 일환으로 [1], [2]와 같이 여러 연구들이 진행되었으며 극단값을 모형화하는 것은 매우 중요한 연구 분야 중 하나로 알려져 있다. [3]은 1914년에서 1962년의 영국 서부지역 일일 강수량의 데이터를 이용하여 일반화 극단값 분포를 모델링한 사례를 제시하였고, 인위의 블록의 데이터의 최대치를 추출하는 것보다 일정 수준의 높은 값 이상인 강수량 30mm이상인 사건을 극한으로 정의하는 것이 더 효율적이라는 것을 소개하였다. 이렇게 하면 임계값을 초과한다는 의미에서 극단값인 모든 관측치를 모델 적합에 사용할 수 있기 때문에

*준회원, 한국외국어대학교 통계학과 석사 (제1저자)
**정회원, 한국외국어대학교 통계학과 교수 (교신저자)
접수일: 2021년 12월 29일, 수정완료일: 2022년 1월 1일
게재확정일: 2022년 1월 8일

Received: December 29, 2021 / Revised: January 1, 2022
Accepted: January 8, 2022
*Corresponding Author: khkang@hufs.ac.kr
Dept. of Statistics, Hankuk University of Foreign Studies, Korea

효율성이 향상된다.

분포의 꼬리 부분의 모형을 추정하는 모수적인 기법에는 주로 일반화 극단값 분포를 이용한다. 또한 극단값의 정의에 대한 접근법에 따라 블록 최댓값 모형 또는 분계점 초과치 모형 접근법은 [4]에서 일반화 극단값 분포 또는 일반화 파레토 분포로 근사됨이 알려져 있다. 일반화 극단값 분포는 3가지의 모수를 가지게 되는데, 그 중 형상 모수의 부호에 따라 분포의 꼬리 성질, 얼마나 두꺼운지 또는 분포의 형태를 결정하기 때문에 이를 적절히 추정하는 것이 극단값 분포를 추정하는데 있어서 매우 중요한 문제라고 할 수 있다.

본 연구에서 사용한 비모수적 밀도함수 추정에서는 평활량 선택 방법으로 묵시적 추정량 (rule of thumbs estimator, ROT), 최대가능도 교차확인법 (maximum likelihood cross-validation, MLCV), 비편향 교차확인법 (unbiased cross-validation, UCV) 세 가지를 소개한다. 각 추정법을 비교하는 기준으로는 [5]에서 사용된 평균절대편차 (mean absolute deviation, MAD)를 이용하여 경험적 누적분포함수와 모수적, 비모수적 방법으로 추정된 누적분포함수를 비교한다. 응용 사례로는 서울시 관측소 별 일별 평균 대기오염도 정보 데이터를 이용하여 어떤 추정방법이 좋은지 비교해보고 2년, 5년, 10년 재현수준을 통해 특정 지역에서 일어날 미세먼지 농도를 예상해보고자 한다.

II. 모수적 추정법

1. 블록 최댓값 (Block Maximum) 모형

일반화 극단값 분포는 일정한 기간마다 극댓값을 구해 데이터의 확률분포를 모형화하는 방법이며 일반화 파레토 분포는 일정한 임계치를 초과하는 극단값에 대해 확률분포를 모형화하는 방법이다.[6] 즉, 분포의 꼬리부분에 대한 통계적 이론으로 전체 분포가 아닌 양쪽 꼬리 부분을 강조함으로써 데이터 전체보다는 극단적인 데이터의 특징을 관찰하는 데 매우 유용하다. 극단의 영역을 추정하기 위해서 극단값 또는 고분위수의 자료를 이용하여 일반화 극단값 분포를 사용하여 모형을 추정한다.

일반화 극단값 이론에서 사용되는 [4]의 Fisher-Tippet-Gnedenko 정리를 소개하기 위해 임의의 분포로부터 얻은 서로 독립이며 동일한 분포를 따르는 확률변수 $\{X_i,\ 1 \le i \le n\}$를 고려하자. 이 확률변수들의 최댓값

$$M_n = \max\{X_1, \cdots, X_n\}$$

의 분포함수는 어떤 상수열 $\{a_n > 0\}$ 과 $\{b_n\}$에 대해 n이 충분이 크면, 분포함수 G로 수렴한다.

$$P\{(M_n - b_n)/a_n \le z\} \to G(z) \qquad (1)$$

이를 일반화 극단값(generalized extreme value, GEV) 분포라고 정의한다. 이는 표본평균의 분포를 중심극한정리에 의해 정규분포로 근사하는 것과 같은 맥락에서 이해될 수 있다.

(1)의 일반화 극단값 분포의 형태를 표현한다면 다음과 같다.

$$G(z) = \exp\left\{-\left[1+\gamma\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\gamma}\right\},\ 1+\gamma\left(\frac{z-\mu}{\sigma}\right) > 0$$

이 식에서 μ는 위치 (location) 모수, σ는 척도 (scale) 모수, γ는 형상 (shape) 모수이다. 여기서 형상모수는 극단값을 나타내는 지수이고, 극단값 분포의 형태와 모형을 결정하는 가장 중요한 값이다. GEV 분포는 (i) 형상모수 $\gamma = 0$인 경우 Gumbel 분포, (ii) $\gamma > 0$인 경우 Frèchet 분포, (iii) $\gamma < 0$인 경우 Weibull 분포가 된다.[3]

$$\text{(i)}: G(z) = \exp\left\{-\exp\left[-\left(\frac{z-\mu}{\sigma}\right)\right]\right\},\ -\infty < z < \infty \quad (2)$$

$$\text{(ii)}: G(z) = \begin{cases} 0, & z \le \mu \\ \exp\left\{-\left[1+\gamma\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\gamma}\right\}, & z > \mu \end{cases} \quad (3)$$

$$\text{(iii)}: G(z) = \begin{cases} \exp\left\{-\left[1+\gamma\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\gamma}\right\}, & z < \mu \\ 1, & z \ge \mu \end{cases} \quad (4)$$

GEV 분포를 이용하여 극단적인 사건이 일어날 경우 얼마나 큰 사건이 일어날지에 대한 확률을 추정하기 위해서 [2]에서 소개한 재현수준 (return level)을 통해 알 수 있다. 일반적으로 어떤 폭풍우가 평균 r년 동안 한 번 이상 발생할 경우에 재현기간 (return period)을 r년 가진다고 표현한다. 재현수준은 GEV 분포에서 주어진 재현기간에 대응되는 분위수를 의미한다. GEV 분포의 함수를 $G(z)$라고 할 때 재현기간이 r년인 재현수준은

$G(z^*)=P(Z\le z^*)=1-1/r$을 만족하는 z^*이다.

2. 분계점 초과치 (Peak over Threshold) 모형

블록 최댓값 모형은 특정 분포나 데이터의 최댓값만 이용하기 때문에 데이터가 많이 손실된다는 단점이 있고, 어떤 블록에서의 최댓값이 아닌 값들이 다른 블록의 최댓값보다 훨씬 큼에도 불구하고 극단값이라고 정의되지 않는 상황이 발생할 수 있다. 따라서 어떤 임계점 (threshold) 이상의 값들을 이용하여 분포를 추정할 때, 더 많은 극단값들을 사용함으로써 추정의 효율성을 극대화할 수 있다.

마찬가지로 임의의 분포 F를 따르는 $\{X_i,\ 1\le i\le n\}$에 대해서 충분히 큰 실수 u에 대해 임계점 u를 초과하는 X_i 값들을 극단 사건이라고 할 수 있다. 독립이면서 동일한 분포를 따르는 임의의 확률변수 $X_1, X_2, \cdots, X_n$에 대하여 $Y=X-u$ 라고 할 때, Y의 분포는 (5)와 같이 정의할 수 있다.

$$P(Y>y|X>u)=\frac{1-F(u+y)}{1-F(u)},\ y>0,\ \sigma>0 \quad (5)$$

[3]은 (5)에 대응하는 분포함수가 다음의 일반화 파레토 (Generalized Pareto, GP) 분포의 분포함수 H로 근사됨을 증명하였다.

$$H(y)=1-\left(1+\frac{\gamma y}{\tilde{\sigma}}\right)^{-1/\gamma},\ \tilde{\sigma}=\sigma+\gamma(u-\mu).$$

III. 비모수적 추정법

1. 커널 확률밀도함수 추정

모집단의 분포함수를 가정하는 것이 힘들다고 판단될 경우, 모집단의 분포를 아예 설정하지 않는 것이 오류의 가능성을 줄일 수 있다. 이런 경우 사용되는 방법을 비모수적 방법이라고 한다. 비모수적 함수 추정이란 주어진 자료에 근거하여 모집단의 분포를 추정하는 것이고, 비모수적 함수 추정 방법에는 일반적으로 커널함수 (kernel function)를 사용한다. 일반적으로 커널함수로 Uniform, Gaussian, Epanechnikov 등등을 사용하고, 본문에서는 가우시안 커널을 사용하여 분포를 추정하였다. 이 밖에 다양한 커널함수와 그에 해당하는 커널함수의 특징들은 [7]을 참고하기 바란다. 커널함수를 이용한 함수추정량을 커널밀도함수추정량이라고 부르며, 연속확률 변수 X에 대해 독립이고 동일한 분포의 임의의 함수추정량은 아래와 같다.

$$\hat{f}_h(x)=\frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{x-X_i}{h}\right)$$

여기서, h는 추정량의 부드러운 정도를 결정하는 평활량 (bandwidth)이고, K는 커널함수로 $\int K(x)dx=1$, $\int x^2K(x)dx<\infty$를 만족한다. 평활량의 크기에 따라 함수의 모형과 형태가 결정되므로 평활량을 선택하는 것은 비모수적 함수 추정에 매우 중요하다. 가장 적절한 h를 찾기 위해서 최적의 성능을 내는 $\hat{f}_h(x)$를 찾는 방법의 하나로 적분제곱오차 (intergrated sqaured error, ISE)를 최소화하는 커널추정량을 찾는 것을 생각해 볼 수 있다.

$$ISE(h)=ISE(\hat{f}_h(x))=\int[\hat{f}_h(x)-f(x)]^2dx \quad (6)$$

2. 평활량의 선택

1) 묵시적 선택법

다양한 평활량 선택방법들이 있지만, 그 중 대표적으로 정규분포를 이용한 방법 중 하나인 [7]의 묵시적 (Rule of Thumbs, RT) 추정량을 소개한다. 여기서 묵시적이란 과학적인 규칙이 아닌 경험에 의한 규칙이라는 뜻으로, 일반적인 상황에서 편히 쓰일 수 있는 평활량이라고 생각하면 이해하기 쉽다. 이것은 모집단의 분포를 정규분포로 가정하고 점근적인 평균제곱오차를 최소로하는 평활량 h^*를 의미한다.

$$h^*=\left(\frac{4\hat{\sigma}^5}{3n}\right)^{1/5}\approx 1.059\hat{\sigma}n^{1/5}$$

따라서 표본의 개수 n과 표본표준편차 $\hat{\sigma}$를 알고 있다면 쉽게 구할 수 있는 방법이다.

2) 최대가능도 교차확인법

최대가능도 교차확인법 (maximum likelihood cross-validation, MLCV)은 [8]에 의해 제안된 평활량 선택

방법으로, 로그가능도함수를 최대화하는 방법이다.

$$MLCV(h) = \frac{1}{n}\sum_{i=1}^{n}\log[\hat{f}_{-i}(x;h)]$$
$$= \frac{1}{n}\sum_{i=1}^{n}\log\left[\sum_{j\neq i}K\left(\frac{X_i - X_j}{h}\right)\right] - \log[(n-1)h]$$

이 방법에 의한 평활량은 $h_{mlcv} = \arg\max_{h>0} MLCV(h)$로 구해질 수 있다. [9]는 함수의 꼬리 부분으로 갈수록 꼬리 효과 (tail effect)를 받아 극단치와 꼬리가 긴 분포를 추정하는데 있어서 어려움을 겪는다고 설명하였다.

3) 비편향 교차확인법

최소제곱 교차확인법이라고도 불리며 (6)의 ISE를 최소화하는 데서 시작한다. ISE 식을 풀어서 쓰면,

$$ISE(h) = \int \hat{f}_h^2(x)dx - \int \hat{f}_h(x)f(x)dx + \int \hat{f}^2(x)dx$$

이 되고, 여기서 우변의 마지막 항은 h에 의존하지 않으므로 제거하고, 우변 앞의 두 개항에 대한 비편향 추정량에 대해 교차확인을 실시하는 것에서 비롯된 아이디어이다. 즉, 목적함수는

$$UCV(h) = \frac{R(K)}{nh} + \sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}\left[\int \frac{1}{n^2h^2}K\left(\frac{x-X_i}{h}\right)K\left(\frac{x-X_j}{h}\right)dx - \frac{2}{n(n-1)h}K\left(\frac{x_i - X_j}{h}\right)\right]$$

으로 하고, 이것을 최소화하는 평활량 h를 찾아 추정량 $h_{ucv} = \arg\min_{h>0} UCV(h)$로 한다.

[9]는 MLCV와 반대로 UCV는 높은 변동성과 과소평활한 데이터에 적합하다고 하였다. 다만 계산이 오래 걸리고 복잡하다는 단점을 가지고 있다.

위에서 소개한 방법들 이외에도 미지의 함수 추정을 위한 기준과 목표함수에 따라 다양한 교차확인법 (biased cross-validation), 플러그-인 방법 (plug-in bandwidth), 혼합형 방법 (mixing method) 등 평활량 선택에 관한 여러 가지 접근법이 있고, 이에 관한 대표적인 참고문헌들로는 [10], [11], [12] 등이 있다.

Ⅳ. 데이터 분석

실제 데이터에서의 분포 추정의 결과를 비교하기 위해서 서울시 평균 대기 오염도 데이터에 대해서 시뮬레이션을 수행하였다. 본 연구에서 사용된 데이터는 서울 열린 데이터광장에서 제공한 2014년부터 2018년도 까지 서울시 39개의 관측소에서 관측된 대기 오염도 데이터이고 총 69,882개의 관측 데이터 중 미세먼지(PM10) 농도의 변수를 사용하였고, R 패키지 extRemes를 이용하여 분석을 실시하였다. 극단값의 꼬리부분의 분포 함수 추정을 위해 블록 최댓값 모형 (BM)과 분계점 초과치 모형 (POT) 접근법을 통해 극단치에 대한 표본을 추출하였고, 각각 매 월의 최댓값과 미세먼지 농도가 100μg 이상의 자료들을 추출하여 분포를 추정하였다.

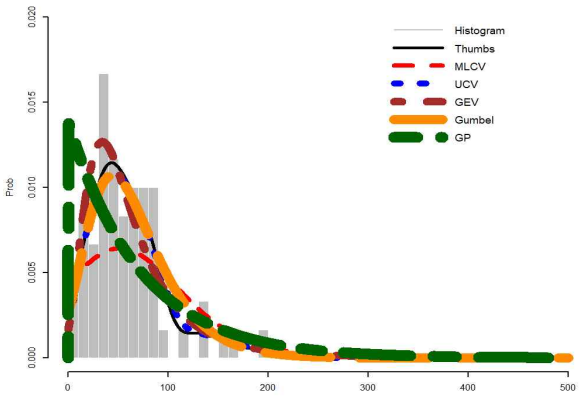

그림 1. 강남구 자료에 블록 최댓값 모형을 적용한 분포 추정
Figure 1. Distribution estimation by applying the block maximum value model to Gangnam-gu data

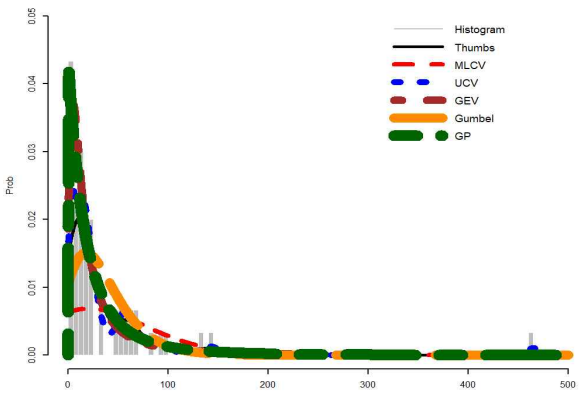

그림 2. 강남구 자료에 분계점 초과치 모형을 적용한 분포 추정
Figure 2. Distribution estimation by applying the peak over threshold model to Gangnam-gu data

선행 연구로서 [5]는 서울시 초미세먼지(PM2.5)에 대하여 일반화 극단값 이론을 적용하여 재현수준과 일반화 극단값 분포의 모수를 추정하였다. 그림 1과 그림 2는 강남구의 2014년부터 2018년 까지의 일별 미세먼지 데이터를 이용하여 블록 최댓값 모형과 분계점 초과치 모형을 적용하여 모수적 비모수적 방법으로 추정한 밀도함수의 그림이다. 강남구의 데이터로 미루어 보았을 때, $100\mu g$ 부근의 미세먼지 농도가 많은 것을 볼 수 있고, 실제 데이터에서도 GP분포는 극단적으로 왼쪽으로 치우쳐 진 추정 결과를 볼 수 있고 MLCV는 평활량을 크게 추정하는 경향을 볼 수 있다.

MAD를 구하여 모수적 비모수적 방법으로 추정된 누적확률분포와 경험적 누적확률분포를 비교하였고, 그 중에 10개소의 결과를 표 1에 제시하였고, 39개소 중 가장 MAD가 작게 나온 추정 방법을 표 2에 수록하였다. 표 1과 2로 보았을 때, 비모수적 방법으로는 ROT 또는 UCV가 MAD를 작게 추정하는 경향이 있고 그에 못지 않게 접근법과 관측소에 따라 모수적인 분포 추정 방법인 GEV가 MAD를 작게 추정하는 방법이라고 할 수 있다.

표 1. 관측소별 자료의 추정법에 따른 MAD 결과
Table 1. MAD result according to the estimation method of data for each measurement station

관측소	추정법	ROT	MLCV	UCV	GEV	GP
강남구	BM	0.0197	0.0877	0.0254	0.0208	0.0720
	POT	0.0391	0.1231	0.0230	0.0236	0.0415
공항대로	BM	0.0182	0.0839	0.0292	0.0289	0.0998
	POT	0.0238	0.1128	0.0165	0.0236	0.0178
광진구	BM	0.0177	0.0912	0.0231	0.0246	0.0724
	POT	0.0376	0.1227	0.0234	0.0403	0.0368
구로구	BM	0.0250	0.0956	0.0248	0.0236	0.0705
	POT	0.0300	0.1178	0.0209	0.0201	0.0216
마포구	BM	0.0204	0.0993	0.0244	0.0320	0.0834
	POT	0.0359	0.1164	0.0228	0.0251	0.0307
신촌로	BM	0.0236	0.0836	0.0234	0.0363	0.1075
	POT	0.0290	0.1099	0.0137	0.0316	0.0470
영등포구	BM	0.0233	0.0995	0.0278	0.0197	0.0645
	POT	0.0302	0.1297	0.0205	0.0247	0.0268
종로	BM	0.0196	0.1055	0.0258	0.0246	0.0881
	POT	0.0456	0.1234	0.0199	0.0254	0.0625
천호대로	BM	0.0225	0.0972	0.0235	0.0293	0.1077
	POT	0.0319	0.1337	0.0229	0.0270	0.0205
화랑로	BM	0.0190	0.0885	0.0274	0.0201	0.0710
	POT	0.0232	0.1200	0.0245	0.0235	0.0371

표 2. 각 추정법에 따른 최소 MAD 관측소 개수
Table 2. Number of minimum MAD stations according to each estimation method

추정법	ROT	MLCV	UCV	GEV	GP	계
BM	23	-	5	11	-	39
POT	1	-	32	4	2	39

표 3에는 GEV 분포를 이용하여 극단값 자료의 분포를 추정했을 때 사건의 2년, 5년, 10년의 재현수준을 10개의 측정소에 대해 제시하였다. 2년 재현수준을 살펴보면 공항대로, 신촌로, 영등포로, 천호대로에서 다른 관측소에 비해 높게 나오는 것을 확인할 수 있다. 5년 재현수준은 공항대로, 신촌로, 영등포로, 천호대로가 $100\mu g$ 이상의 미세먼지 수준을 보이는 것으로 보아 5년 이내에 한 번 이상은 미세먼지 기준을 초과하는 나쁨 수준이 발생할 수 있는 위험지역으로 볼 수 있다. 마지막으로 10년 재현수준을 확인한 결과, 공항대로, 신촌로, 영등포로가 타 지역에 비해 10년 재현수준이 높은 값을 가지는 것을 볼 수 있다. 모든 년도의 재현수준을 비교해보았을 때, 공항대로, 신촌로, 영등포로가 다른 지역에 비해 높은 재현수준을 가지고 있기 때문에 향후 미세먼지 대책을 세울 경우 더욱 관심을 기울여야 할 지역으로 분류될 수 있다.

표 3. 관측소별 GEV 분포를 이용한 재현수준
Table 3. Return level using GEV distribution for each measurement station

기간 / 측정소	2년	5년	10년
강남구	51.270	93.106	127.415
공항대로	64.687	109.293	142.809
광진구	47.480	88.037	122.175
구로구	45.308	84.617	117.259
마포구	47.674	86.953	119.089
신촌로	71.402	119.155	156.077
영등포구	42.189	78.996	110.040
영등포로	64.349	111.930	148.705
천호대로	64.595	102.842	131.132
청계천로	38.234	68.810	94.248

V. 결 론

본 연구에서는 꼬리가 긴 분포에 대하여 블록 최댓값 모형과 분계점 초과치 모형을 이용하여 극단값에 대한 모수적 비모수적인 방법으로 분포를 추정하고 실제 데이터를 이용하였을 때 MAD를 비교하여 모수적 방법과 비모수적 방법의 성능 차이를 비교하였다. 모수적

분포 추정 방법 같은 경우 최대가능도추정법을 이용하여 모수를 추정하였다. 서울시에서 제공한 서울시 39개의 관측소에서 관측된 대기 오염도 데이터를 이용하여 월 별 단위 블록 최댓값 자료와 분계점 초과치 모형을 적용한 미세먼지 농도가 100μg 이상의 자료가 이용되었다.

월별 단위 블록 최댓값 같은 경우 2014년부터 2018년까지 관측소 별 총 60개의 자료가 이용되었으므로 소표본으로 볼 수 있다. 분계점 초과치 모형의 경우, 100μg 이상의 데이터를 모두 사용하기 때문에 블록 최댓값 모형보다 훨씬 많은 데이터를 사용한다는 점이 다르다. 또한 GEV 분포를 이용하였을 경우 2년, 5년, 10년의 재현수준을 확인하였을 때, 공항대로, 신촌로, 영등포로 관측소에서 측정된 미세먼지의 재현수준이 타 관측소보다 높게 나온 것을 확인할 수 있었다. 즉, 타 지역보다 고농도의 미세먼지에 노출될 가능성이 높다고 볼 수 있어서 세 지역 부근에서의 미세먼지 농도에 대해 주의깊게 관찰하고 대책을 세울 필요성이 있다.

또한 GEV 분포를 이용하였을 경우에는 재현수준을 구하여 향후 발생할 사건에 대해 예상하고 결과를 도출해낼 수 있지만 GP 분포를 이용하였을 경우 재현수준을 구하는 것이 매우 복잡하고 임계점에 따른 결과가 크게 다른 경우가 있기 때문에 이 경우를 극복하여 GP 분포를 이용한 재현수준을 구하는 것이 향후 과제가 되겠다. GEV 분포와 GP 분포에서 사용하는 데이터의 개수도 다르다는 점이 하나의 차이점이다. 대부분의 분포 추정 결과 극단값에 대한 분포 추정의 경우에도 비모수적 함수 추정의 결과가 좋게 나온 것을 확인할 수 있었다. 모수적 방법의 한계점을 뛰어넘어 극단값에 대한 비모수적 방법을 적절히 활용할 수 있다면 보다 더 좋은 결과와 함께 데이터의 또 다른 의미있는 내용을 파악할 수 있는 기회가 될 것이다.

References

[1] D. Weon, “Service platform design for smart environment disaster management”, *The Journal of the Convergence on Culture Technology (JCCT)*, Vol. 4, No. 3, pp.247-252, August 31, 2018. https://doi.org/10.17703/JCCT.2018.4.3.247

[2] C.Y. Lim and S.H. Lee, “Wireless network design for construction of atmospheric and marine environment monitoring system using buoy”, *The International Journal of Advanced Culture Technology (IJACT)*, Vol. 8, No. 3, pp. 269-274, September 31, 2020. https://doi.org/10.17703/IJACT.2020.8.3.269

[3] S. Coles, “An introduction to statistical modeling of extreme values”. *Springer Series in Statistics*. Springer Verlag London, 2001.

[4] R.A. Fisher, and L.H.C. Tippett, “Limiting forms of the frequency distribution of the largest or smallest member of a sample”. *Mathematical Proceedings of the Cambridge Philosophical Society*, 24, pp.180-190, 1928.

[5] J.W. Oh and T.J. Lim, “Analysis of extreme values of ultrafine dust (PM2.5) by region in Seoul”, *Journal of the Korean Society for Quality Management*, Vol. 47, No. 1, pp.47-57, 2019.

[6] J. Beirlant, Y. Goegebeur, J. Segers, and L. Teugels, “Statistics of Extremes : Theory and Applications”, *Wiley*, 2004.

[7] B.W. Silverman, “Density estimation for statistics and data analysis”, *Monographs on Statistics and Applied Probability*, Chapman and Hall, London, 1986.

[8] J.D.F. Habbema, J. Hermans, and K. Van den Broek, “A stepwise discrimination analysis program using density estimation”, *Proceedings in Computational Statistics 1974*. Physica Verlag, Vienna, 1974.

[9] X. Wu, “Robust likelihood cross validation for kernel density estimation”, *Journal of Business & Economic Statistics*, Vol. 37, No. 4, pp. 761-770, 2018. https://doi.org/10.1080/07350015.2018.1424633

[10]W. Stute, “Modified cross validation in density estimation”, *Journal of Statistical Planning and Inference*, 30, pp.293 - 305, 1992.

[11]S.J. Sheather and M.C. Jones, “A reliable data-based bandwidth selection method for kernel density estimation”. *Journal of the Royal Statistical Society* series B, 53, pp.683 - 690, 1991.

[12]E. Mammen, M.D. Martínez-Miranda, J.P. Nielsen, and S. Sperlich, “Do-validation for kernel density estima-tion”, *Journal of American Statistical Association*, 106, pp.651 - 660, 2011.

※ 이 연구는 2021학년도 한국외국어대학교 교내학술연구비 지원에 의하여 이루어진 것임