ICIC Express Letters
Part B: Applications
Volume **7**, Number **2**, February **2016**

 ISSN 2185-2766
pp. **301–306**

RISK ANALYSIS OF NATURAL DISASTER IN SOUTH KOREA USING EXTREME VALUE THEORY

Tae-Jin Lim

Department of Industrial and Information Systems Engineering
Soongsil University
No. 369, Sangdo-Ro, Dongjak-Gu, Seoul 156-743, Korea
tjlim@ssu.ac.kr

Received July 2015; accepted September 2015

Abstract. *This paper investigates the societal risk due to natural disaster in South Korea from annually aggregated fatality data from 1916 to 2013. Extreme value theory (EVT) was employed to fit the fatality distribution, because outliers were found in the data. To analyze the risk induced by natural disaster, four types of fitting models were implemented: the empirical model, the Weibull model, the generalized extreme value (GEV) distribution, and the generalized Pareto distribution (GPD). Annual exceedance probability curves were constructed from the four models, and they were compared on the basis of the accuracy measure developed in this research. A trimmed data set as well as the original data set was analyzed in order to compare the effect of missing data. Even if the empirical cubic curve fits best for the original data, the GEV model fits best for the trimmed data and has the advantage of estimating the return level of fatality given a return period of interest. The confidence intervals for the return level were constructed from the GEV model. The proposed methodology may be applied to event based data sets, once the database of natural disasters will be developed in South Korea.*
Keywords: Risk profile, Natural disaster, Extreme value theory (EVT), Generalized extreme value (GEV) distribution, Exceedance probability (FN) curve, Return level

1. **Introduction.** With the increasing demand to manage risk within their domestic area, many countries have developed risk profiles for each risk source [1-3]. One of the major concerns would be the risk due to natural disaster, because their impact is getting bigger due to climate change.

One of the oldest studies on disaster risk is Chapter 6 of the *Reactor Safety Study* of 1979 [4], where exceedance frequency curves were introduced. Since then, similar types of curves have been utilized in many risk studies. Annual disaster statistical review [1] is published annually based on EM-DAT database [5]. Global assessment report on disaster risk reduction (GAR) [2] is also published annually based on DesInventar database [6].

However, there have not been sufficient researches on developing risk profiles in Korea, so little is known about the risk level caused by natural disasters. In order to investigate the societal risk due to natural disasters in South Korea, we perform risk analysis from annually aggregated fatality data from 1916 to 2013. An event based data set would be more suitable in risk analysis, but it is not available at this point.

We adopt the extreme value theory (EVT) to analyze the fatality data, because outliers are found in the data as shown in Figure 1. EVT is recognized to be useful in estimating the annual exceedance probability of a rare event.

To analyze the risk induced by natural disaster, four types of fitting methods were implemented: the empirical model, the Weibull model, the generalized extreme value (GEV) distribution, and the generalized Pareto distribution (GPD). In most risk analyses, only the empirical method has been utilized to construct the exceedance probability (FN) curves. In addition to this, we employ three parametric models to develop the FN curves.

We also provide a measure which assesses the accuracy of the models, and we compare the four models on the basis of the measure. Even if the empirical cubic curve fits best for the original data, the GEV model fits best for the trimmed data and provides the return level of fatality given a return period of interest. We finally construct the confidence interval for the return level from the GEV model.

This paper is organized as follows. The background of this research is described in Section 2. Descriptive statistical approaches for analyzing the fatality data are explained in Section 3. Risk analyses by fitting alternative models are performed in Section 4, and concluding remarks are presented in Section 5.

2. Background.

A huge effort has been made developing disaster database such as EM-DAT [5], DesInventar [6], NatCat [7], SIGMA [8], GLIDE [9], and BASICS [10], in order to manage the risks due to disasters striking all over the world. Even though annual reports on disaster statistics have been published in Korea since the late 1970s, little is known about the risk levels induced from Korean domestic disasters.

2.1. Data collection.

In order to develop FN curves for major natural disasters in Korea, we investigated famous databases abroad.

From the advanced search function of EM-DAT [5], we found statistics for Korea from 1900 to 2014, but the total number of disasters recorded during that period was only 109 which is far less than the number of typhoons that affected Korea. The situation is worse in other databases. For example, GLIDE [9], managed by Asian Disaster Reduction Center, has only 33 listings for Korea. This is because worldwide databases list only the major events reported from each country.

In order to find domestic databases, we searched the websites of the Statistics Korea [11], the Korean Meteorological Agency [12], the Ministry of Public Safety and Security [13], and the National Disaster Information Center [14]. Finally, annually aggregated fatality data from 1916 to 2013 could be collected from the annual report of disasters [15].

2.2. Extreme value theory.

Generally, there are two practical approaches of EVT: block maxima (BM) and peaks over threshold (POT) [16]. The BM method uses the maxima within blocks of equal length. For the BM method, the GEV distribution is employed to describe the maxima. With location parameter μ, scale parameter σ, and shape parameter ξ, the GEV distribution is given by

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

The GEV distribution is flexible, because it comprises the Weibull ($\xi < 0$), the Gumbel ($\xi \to 0$), and the Frechet ($\xi > 0$) distributions. The return level is very important in the EVT, and is defined as the point x_p for which $G(x_p) = 1 - p$ with a return period $T = 1/p$. Then the return level of the EVT is obtained as:

$$x_p = \begin{cases} \mu - \dfrac{\sigma}{\xi}\left[1 - \left\{-\log(1-p)\right\}^{-\xi}\right], & \xi \neq 0 \\ \mu - \sigma\log\left\{-\log(1-p)\right\}, & \xi = 0 \end{cases} \tag{2}$$

The POT method deals with observations that exceed a selected threshold, and provides the GPD as the limiting distribution. The GPD with a threshold u is defined as:

$$H(y) = 1 - \left[1+\xi y/\left\{\sigma+\xi(u-\mu)\right\}\right]^{-1/\xi}, \quad y > 0 \tag{3}$$

Like the GEV, the GPD can be expressed by three extreme distributions: the Pareto ($\xi > 0$), the exponential ($\xi \to 0$), and the Beta distributions ($\xi < 0$). Let ζ_u denote the probability that an observation exceeds the threshold u, and n_y denote the number of observations per year. Then the N-year return level x_N from the GPD can be obtained as:

$$x_N = u + \left\{\sigma+\xi(u-\mu)\right\}/\xi \times \left[(Nn_y\zeta_u)^{\xi} - 1\right] \tag{4}$$

2.3. **Risk classification.** Even if there are many quantitative risk measures [17], risk measures can be mainly classified into individual risk and societal risk. Let P_f denote the probability of disaster, and $P_{d|f}$ denote the conditional probability of death, given a disaster occurs. Then the individual risk can be defined as:

$$IR = P_f \times P_{d|f} \tag{5}$$

The societal risk measures have many classes. The aggregated weighted risk is calculated by multiplying the number of houses $h(x, y)$ inside a certain area A with their individual risk level $IR(x, y)$:

$$AWR = \int \int_A IR(x, y)h(x, y)dxdy \tag{6}$$

Similarly, the number of fatalities can be determined by integrating the individual risk levels $IR(x, y)$ and the population density $m(x, y)$:

$$E(N) = \int \int_A IR(x, y)m(x, y)dxdy \tag{7}$$

Societal risk is usually represented graphically in the FN curve which displays the probability of exceedance as a function of the number of fatalities, on a double logarithmic scale.

$$1 - F_N(x) = P(N > x) = \int_x^\infty f_N(y)dy \tag{8}$$

Then the potential loss of life can be obtained from the FN curve:

$$E(N) = \int_0^\infty x f_N(x)\, dx = \int_0^\infty [1 - F_N(x)]dx \tag{9}$$

3. **Descriptive Statistical Analysis of the Annual Fatality Data.** We collected data on annual fatalities caused by natural disasters in South Korea from 1916 to 2013. Unfortunately, the type of disasters and the event records are unavailable, and the data from 1945 to 1957 are missing due to societal chaos. The time series plot and the box plot shown in Figure 1 reveal some extreme values in the annual fatalities.

Two data sets are analyzed under two separate assumptions. The first uses the original data, and the second uses the data in which the annual fatalities are greater than or equal to 10. If the trimmed data fit as good as the original data, then we do not have to worry about missing data that may have insignificant consequences. The descriptive statistics for the two data sets are given in Table 1. Three parametric models are fitted for each data set by employing R-package and library 'MASS' and 'ismev' [18], and the estimates are shown in Table 2. Diagnostic plots for the GEV model are illustrated in Figure 2. The probability plot and the quantile plot show that the trimmed data set fits better under the GEV model, because the data points from the trimmed data look closer to a straight line.

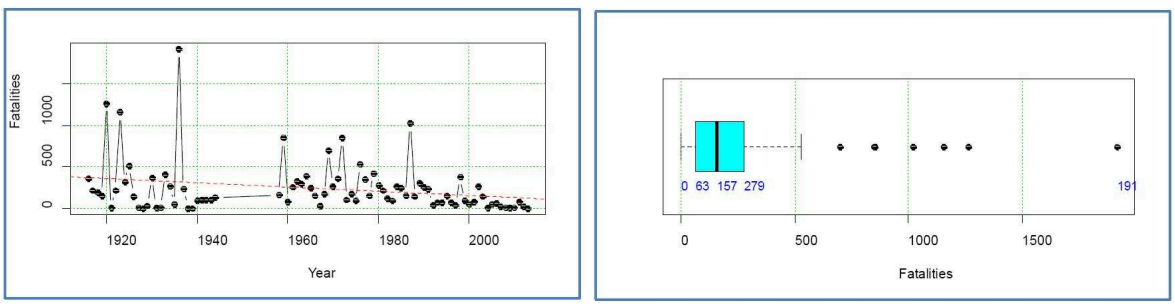


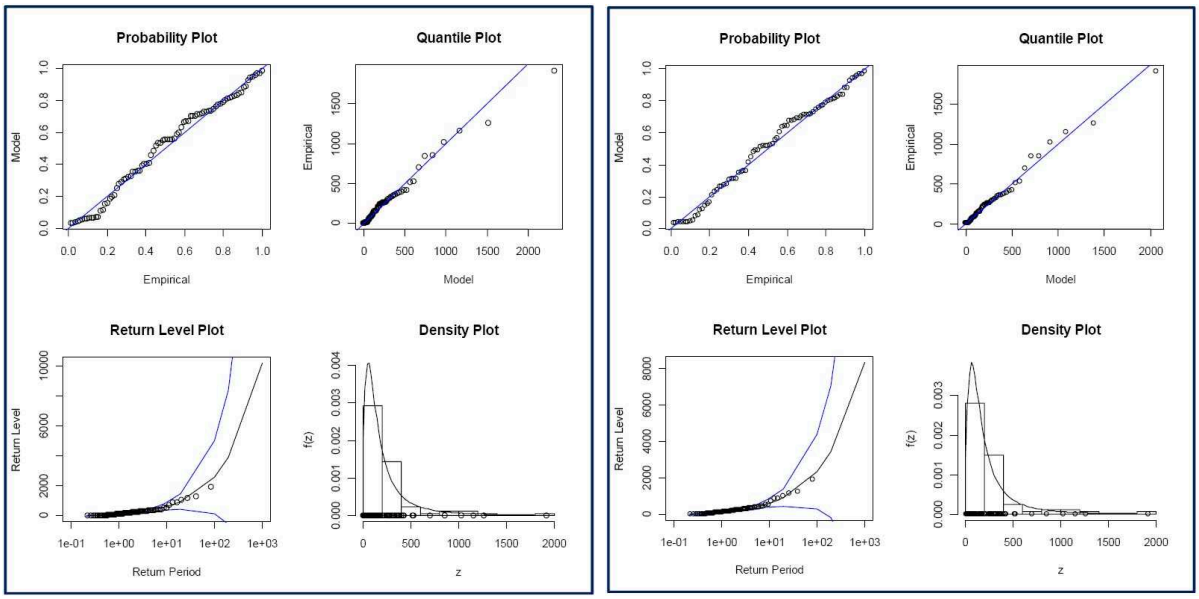
FIGURE 1. Time series plot and box plot of the annual fatality data

TABLE 1. Descriptive statistics for the annual fatality data

Original Data ($n = 85$)					Trimmed Data ($n = 80$)				
mean	stdev	Q_1	Q_2	Q_3	mean	stdev	Q_1	Q_2	Q_3
241.1	310.6	63.0	157.0	279.0	255.9	314.2	77.5	158.5	301.5

TABLE 2. Parameter estimates for the annual fatality data

Original Data ($n = 85$)				Trimmed Data ($n = 80$)			
model	shape	scale	location	model	shape	scale	location
Weibull	0.861	227.272		Weibull	0.816	222.238	
GEV	0.588	104.230	93.456	GEV	0.542	108.311	106.538
GPD	0.222	189.394	1.0	GPD	0.219	192.529	9.9



(a) Original data (b) Trimmed data

FIGURE 2. GEV diagnostic plots from two data sets

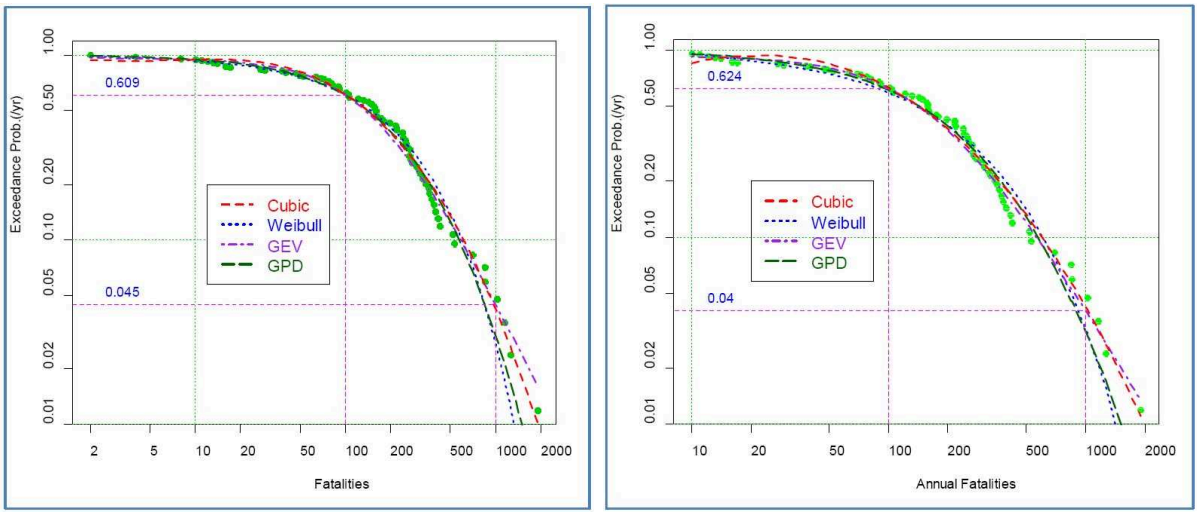
4. **Risk Analysis of the Annual Fatality Data.** For each data set, the exceedance probability curves are drawn from three parametric models and an empirical cubic polynomial model, as shown in Figure 3. The empirical cubic model looks best for the original data, but the GEV looks good for both data sets. The return probabilities of annual fatality level 100 and 1000 are shown in the graph. They look similar for both data sets.

In order to compare the models, a measure of accuracy is developed as the area of the residuals around the horizontal axis. Let $P^*(N > x)$ denote the empirical FN curve, and $\hat{P}(N > x)$ denote a parametric FN curve. Then the accuracy measure is defined as the integration of residuals in the FN curve:

$$\int_{x_{\min}}^{x_{\max}} \left| \hat{P}(N > x) - P^*(N > x) \right| dx \tag{10}$$

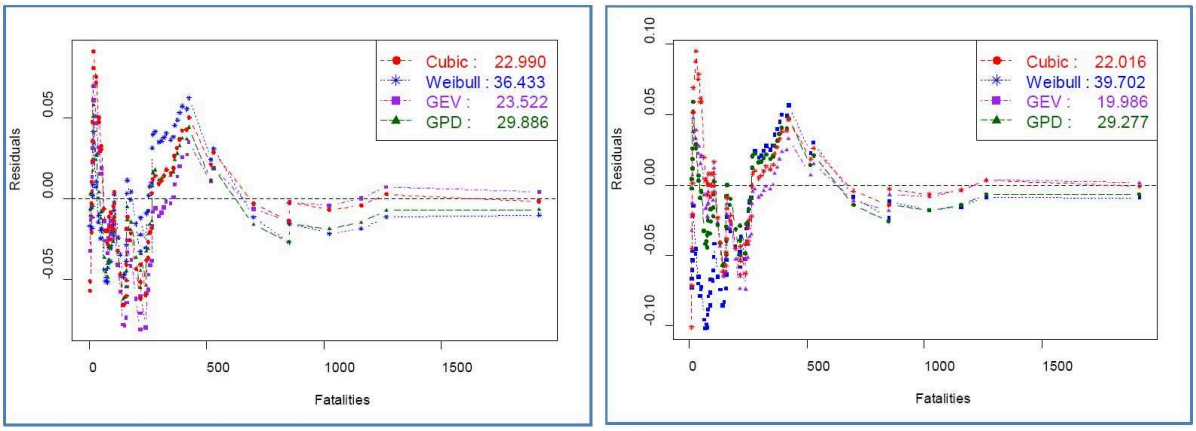
The estimated measure of fit for each model is given in Figure 4. Each value is calculated by integrating the absolute value of residual curves for each model. The empirical cubic model fits best for the original data, but the GEV fits best for the trimmed data.

The confidence intervals for the return level of period 10 under the GEV model are given in Table 3. The profile log-likelihood curves are also shown in Figure 5. Surprisingly, the



(a) Original data (b) Trimmed data

FIGURE 3. Exceedance probability curves from two data sets



(a) Original data (b) Trimmed data

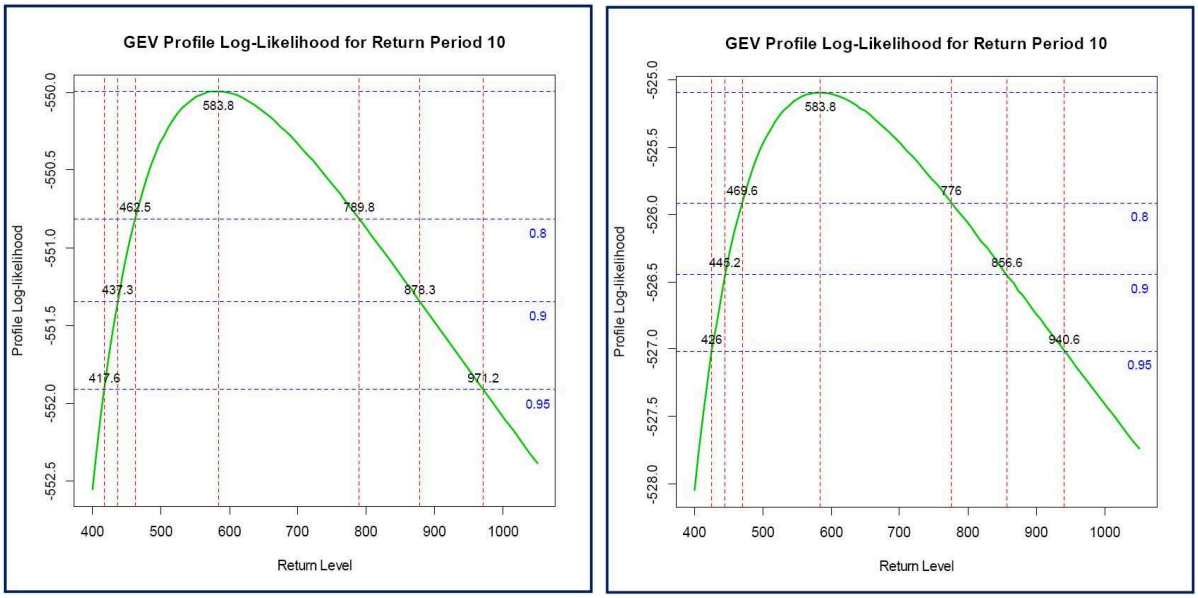
FIGURE 4. Residual plots and the accuracy measure from two data sets

TABLE 3. Confidence intervals for the return level of period 10

Original Data ($n = 85$)			Trimmed Data ($n = 80$)		
Conf. level	LCL	UCL	Conf. level	LCL	UCL
80%	462.5	789.8	80%	469.5	776.0
90%	437.3	878.3	90%	445.2	856.6
95%	417.6	971.2	95%	426.0	940.6

confidence intervals for the return level of period 10 are tighter in the trimmed data than in the original data. This implies that we may get better results with the trimmed data.

5. **Concluding Remarks.** To analyze the risk induced by natural disaster in South Korea, we introduce parametric models in addition to the empirical model. We develop parametric FN curves and propose a measure of fit by integrating the residual of the curves. We compare the measure of fit from Weibull, GEV, GPD, and cubic models. The GEV model based on the EVT provided good fitting results both in the original data and in the trimmed data. We also construct confidence intervals for the return level of period 10, by drawing profile log-likelihood curves. The results show that we may get tighter



(a) Original data (b) Trimmed data

FIGURE 5. Profile log-likelihood curves for the return level of period 10

confidence intervals from the trimmed data than from the original data. This implies that the parametric EVT approach may be applied to a data set with missing data that may have insignificant consequences. The proposed methodology may be applied to an event based data set classified by each disaster type, once the database of natural disasters is constructed in South Korea.

Acknowledgment. This work was supported by National Research Foundation of Korea. The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the article.

REFERENCES

[1] CRED/UCL, *Annual Disaster Statistical Review: Numbers and Trends*, 2013.
[2] UNISDR, *Global Assessment Report on Disaster Risk Reduction*, GAR, 2013.
[3] UNISDR, *Probabilistic Modelling of Natural Risks at the Global Level*, 2013.
[4] WASH-1400, *Reactor Safety Study, Chapter 6*, 1979.
[5] *EM-DAT Homepage*, http://www.em-dat.net/index.htm.
[6] *DesInventar Homepage*, http://www.desinventar.net/DesInventar/.
[7] *NatCat Homepage*, http://www.munichre.com.
[8] *Sigma Homepage*, http://www.swissre.com.
[9] *GLIDE Homepage*, http://www.glidenumber.net/glide/public/search/search.jsp.
[10] *BASICS Homepage*, http://www.basics.org.uk/data/searchPage.php.
[11] *The Statistics Korea Homepage*, http://www.kostat.go.kr.
[12] *Korean Meteorological Agency Homepage*, http://www.kma.go.kr.
[13] *Ministry of Public Safety and Security Homepage*, www.mpss.go.kr.
[14] *National Disaster Information Center Homepage*, http://www.safekorea.go.kr.
[15] *Annual Report of Disasters*, National Emergency Management Agency, 2013.
[16] S. G. Coles, *An Introduction to Statistical Modeling of Extreme Values*, Springer, London, 2001.
[17] S. N. Jonkmana, P. Gelder and J. K. Vrijling, An overview of quantitative risk measures for loss of life and economic damage, *Journal of Hazardous Materials*, vol.99, pp.1-30, 2003.
[18] A. Stephenson, Package 'ismev', *Manual*, pp.1-87, 2015.