Extreme Value Modeling of Minnesota River Flood

Deepak Sanjel,[1*] You-Gan Wang[2]

[1]Minnesota State University, Mankato, USA
[2] University of Queensland, Brisbane, Australia

[*] Corresponding author; E-mail: deepak.sanjel@mnsu.edu

Abstract

Accurately modeling extreme events such as flood, fire, hurricane etc. has become more and more important. Several methods of analyzing extreme values are proposed in literature however, most of them are based on the extreme value limit distributions or some related family of distributions. In this paper we review these techniques and proposed to use Bayesian method to analyze such extreme value events. These ideas are illustrated with an analysis of Minnesota river flood data. Comparisons between different models such as Block Maxima model, Picks Over Threshold (POT) model and Bayesian approach have been made.

Key Words: Picks over threshold method, Bayesian, extreme values, Generalized Extreme Value distribution, return levels.

Introduction

During September 2010 heavy rainfall ranging more than 10 inches caused severe flooding across Minnesota. June 2012 the most damaging flood in Duluth's history began when heavy rains fell over already saturated ground, making the situation worse. Widespread flooding that occurred as a result of the heavy rainfall caused evacuations of hundreds of residents, and damages in excess of 100s of million dollars to residences, businesses, and infrastructure. In this paper we discuss if appropriate model is fitted on such extreme events using historical data, we can predict these events more precisely.

The traditional and best known method of analyzing extreme values is based on the extreme value limiting distributions first introduced by Fréchet (1927) arises as limiting distribution of maxima (or minima) in samples of independent and identically distributed random variables.

Results on Extreme Value Analysis (EVA) dates back to Fréchet (1927), who obtained the asymptotic distribution of the maximum. Fisher and Tippet (1928) and von Mises (1936) presented the first studies on the extremal limit problems. However Gnedenko (1943) was the first who gave conditions for the existence of sequences.

The first book on modeling the extremal values was *Statistics of Extremes* by Gumbel (1958) which provides detail discussion on Extreme Value Analysis. There are more recent publications on applications of Generalized Extreme Value (GEV) model on various areas such as oceanography: Haan (1990), Robinson & Twan (1997), sports data: Henery (1984), insurgence and finance: Smith(2003), just to name a few.

In recent years a number of alternative approaches have been studied. Other recent works published in extreme value analysis are Coles (2001), Embrechts et al. (2003), Beirlant et al. (2004), Castillo et al. (2005) and Reiss and Thomas (2007), among others.

In this paper, we used Minnesota river flood data and utilized three approaches. Block Maxima Model, Pick Over Threshold (POT) Model and Bayesian approach.

Block Maxima Model

Traditional and best known method of analyzing extreme values is based on the extreme value limiting distributions originally introduced by Fisher and Tippet (1928) which was derived using block maxima (or minima).

Let $X_1, \cdots X_n \overset{i.i.d.}{\sim} F$ with mean μ and finite variance, σ^2. The central limit theorem states as $n \to \infty, \quad \bar{X} \sim N(\mu, \sigma^2/n)$ or $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1) \ \ as \ \ n \to \infty$

Extremes values data are often rare and normal distribution is inappropriate. We are interested in modeling the tail of the underlying distribution.

Consider the iid sample $(X_1, \cdots, X_n)$ whose common distribution function $\sim F$. We want to know the distribution of the maximum $M_n = max(X_1, \cdots, X_n)$.

Since,

$P[M_n \leq x] = P[X_1 \leq x; \cdots; X_n \leq x]$

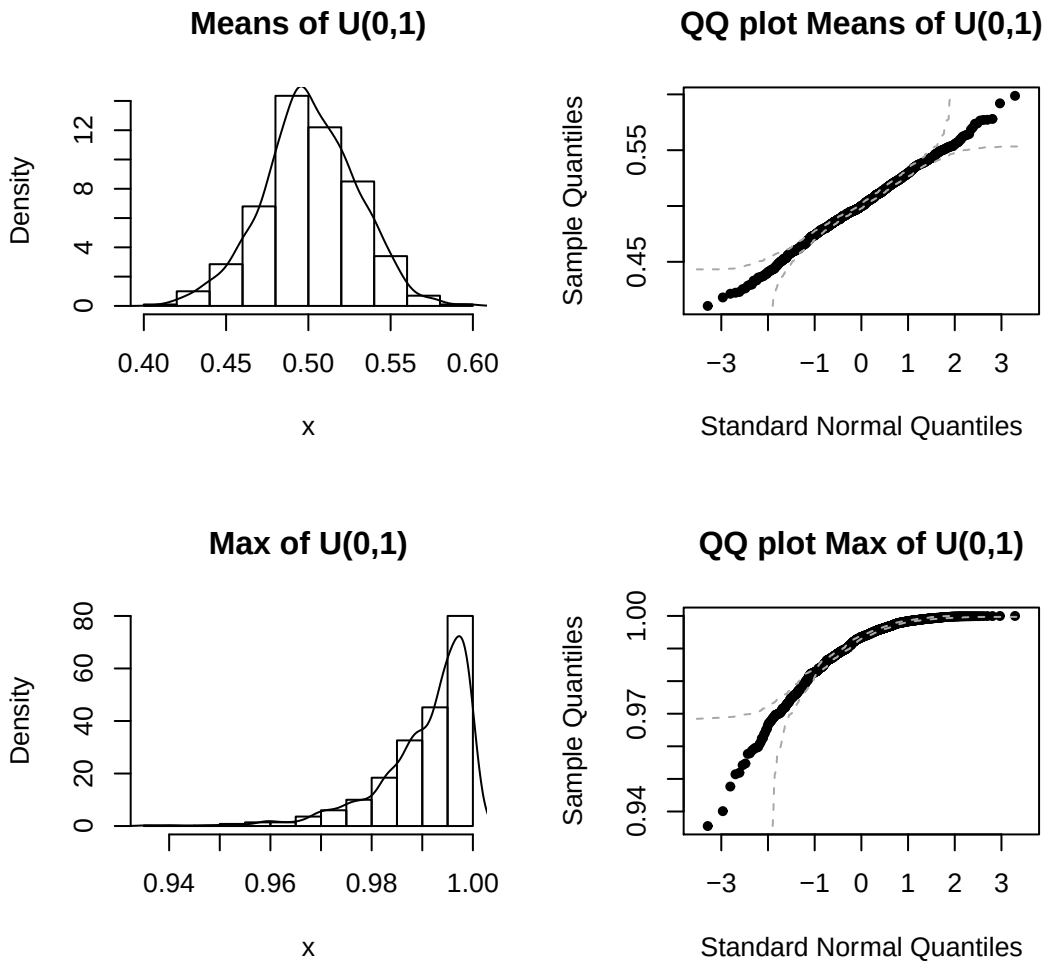

Figure 1: Normal fit and Q-Q plots of the sample average and maximum of random samples (n = 1000) from an uniform distribution.

$= P[X_1 \leq x] \cdots P[X_n \leq x]$ (i.i.d. sample)

$= F^n(x)$

Because F is unknown, this is not very helpful and small discrepancies in the estimate of F can lead to large discrepancies for F^n.

To derive the limiting distribution of extreme values. If there exist sequences of constants $\{a_n > 0\}$ and $\{b_n\}$ such that

$P\left\{\frac{M_n - b_n}{a_n} \leq x\right\} \to G(x)$ as $n \to \infty$.

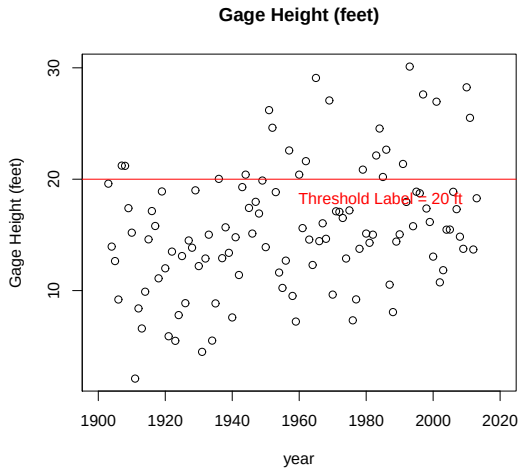

Figure 2: Annual peak gage height from 1903 through 2013 for the Minnesota River near Mankato.

where G is a non-degenerate distribution function. Then G belongs to one of the following three types extreme value distributions: type I, II or III. Originally stated without detailed mathematical proof by Fisher and Tippett (1928), and later rigorously derived by Gnedenko (1943).

I. Gumbel:

$$G(x) = exp\left\{-exp\left[\left(\frac{\mu - x}{\sigma}\right)\right]\right\}, \quad -\infty < x < \infty$$

II. Fréchet:

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

III. Weibull

$$G(x) = exp\left[-\left(\frac{\mu - x}{\sigma}\right)^{\xi}\right], \quad x < \mu$$

Combining all the above three distribution we get the Generalized Extremal Value Distribution (GEVD) given as

$$G(x;\mu,\sigma,\xi) = \begin{cases} exp\left\{-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi}\right\} & for\ 1+\xi\left(\frac{x-\mu}{\sigma}\right) \geq 0 \\ & ,\xi \neq 0 \\ exp\left[-exp\left\{-\left(\frac{x-\mu}{\sigma}\right)\right\}\right] & for\ -\infty < x < \infty \\ & ,\xi = 0 \end{cases}$$

The model has three parameters: a location parameter μ; a scale parameter σ and shape

parameter ξ. Gumbel (shape $\rightarrow 0$, light tail), Fréchet (shape > 0, heavy tail) and Weibull (shape < 0, bounded upper tail at location - scale/shape).

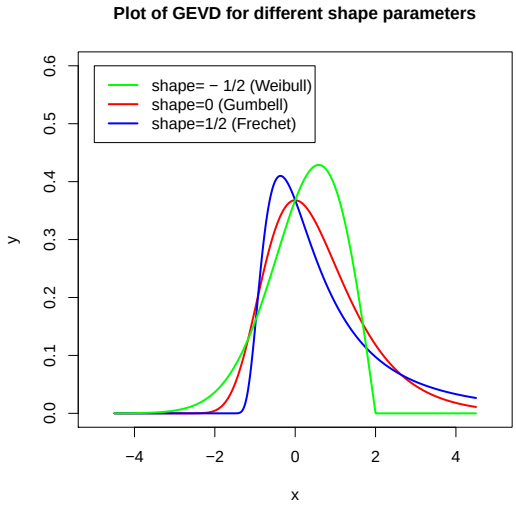

Figure 3: GEVD density plots for Shape parameter -ve (Weibull), 0 (Gumbell) and +ve (Fréchet). Location=0 and Scale=1.

Return level:

Another common measure of extreme events is the p-year return level. The p-year return level, x_p, is the level so extreme it is expected to occur once every p time-units (year, day, hour...).

(Let p=100) The inverse problem of return levels gives the stream flow level that exceeded with probability 1/100 in a given year. This quantity is called the 100-year return level.

Inverting the GEVD distribution formula,

$$x_p = \begin{cases} \mu - \frac{\sigma}{\xi}\left[1 - \{-log(1-p)\}^{-\xi}\right] & for\ \xi \neq 0 \\ \\ \mu - \sigma log\{-log(1-p)\} & for \quad \xi = 0 \end{cases}$$

where $G(x_p) = 1 - p$ and x_p is called the return level associated with the return period $1/p$.

Pick Over Threshold (POT) Model:

Let $X_1, X_2, ...$ be an iid sequence of random variables with marginal distribution, F. Interest is now in the conditional probability of $X's$ exceeding a certain value $u + y$, given that X already exceeds a threshold value, u.

$Pr\{X > u + y | X > u\} = \frac{1-F(u+y)}{1-F(u)}$, $y > 0$

F is not known, the distribution of $[(X - u)|X > u]$, is approximated by Generalized Pareto Distribution (GPD). (for more see Todorovic and Zelenhasic (1970)).

$H(y) = 1 - \left(1 + \frac{\xi y}{\tilde{\sigma}}\right)^{-1/\xi}$ defined on $\{y > 0, \ and \ \left(1 + \frac{\xi y}{\tilde{\sigma}}\right) > 0\}$

where $\tilde{\sigma} = \sigma + \xi(u - \mu)$ and μ, σ, ξ are as defined in GEVD.

Bayesian Parameter Estimation

Let $\mathbf{y} = (y_1, \cdots, y_m)$ be observed data with pdf $f(y|\boldsymbol{\theta})$ where $\boldsymbol{\theta}$ denotes the vector of parameters in parameter space Ω.

According to Bayes' theorem, posterior distribution $\pi(\boldsymbol{\theta}|y)$ is proportional to the product of likelihood function $f(y|\boldsymbol{\theta})$ and and prior distribution for $\boldsymbol{\theta}$ denoted by $\pi(\boldsymbol{\theta})$

$\pi(\boldsymbol{\theta}|y) = \frac{f(y|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int_{\Omega} f(y|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}} \ \propto \ f(y|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$

The estimate of $\boldsymbol{\theta}$ is obtained by calculating the mean or median of the posterior distribution.

Bayesian Prediction: To predict the future observation Y_{m+1} with density function $f(y_{m+1}|\boldsymbol{\theta})$ obtain the posterior predictive density function of future observation Y_{m+1} given $\mathbf{y}$ is given by

$f(y_{m+1}|y) = \int_{\Omega} f(y_{m+1}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}$

The posterior predictive probability of Y_{m+1} exceeding some high threshold y is given by

$P(Y_{m+1} > y|\mathbf{y}) = \int_{\Omega} P(Y_{m+1} > y|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}$

Posterior prediction distribution is often difficult to solve analytically. We can use MCMC simulation to estimate it. A posterior predictive $(1 - p)$ quantile is then obtained by solving

$P(Y_{m+1} > y|\mathbf{y}) = p$

Analysis and Result

The data used for illustration is yearly maximum gage height for 110 years recorded in Minnesota river at Mankato 1903 to 2013. Taken from U.S. Geological Survey.

http://nwis.waterdata.usgs.gov/mn/nwis/peak

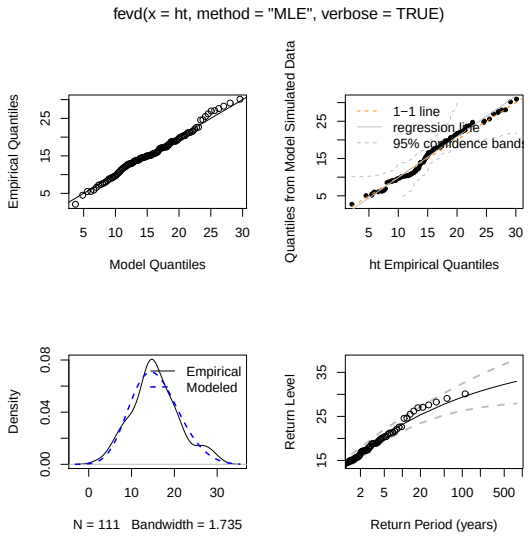

Figure 4: Plots of goodness of fit and return level using block maxima model.

Paramet	MLE	Bayesian
Loc (μ)	13.37 (0.557)	13.43 (0.677)
Scale (σ)	5.31 (0.386)	5.42 (0.456)
Shape (ξ)	-0.20 (0.060)	-0.189 (0.072)

Table 1: Estimated model parameters using MLE method for GEVD and standard errors (in parenthesis) compared with Bayesian estimation.

Method	Return Level	95% CI
GEVD (MLE)	29.247	(26.264, 32.430)
Bayesian	29.35	(27.248, 35.643)

Table 2: 100 year return level.

Model	5 yr	25 yr	50yr	100yr	200yr
GEVD	20.26	25.93	27.76	29.35	30.72
Bayesian	20.43	26.35	28.29	29.94	31.45

Table 3: Return levels for 5, 25, 50, 100 and 200 years gage height in feet.

Model	1965	1969	1993	2001	2010
Gage ht	29.09	27.07	30.11	26.96	28.25

Table 4: Actual historical data from MN river (Source: USGS Water Resources).

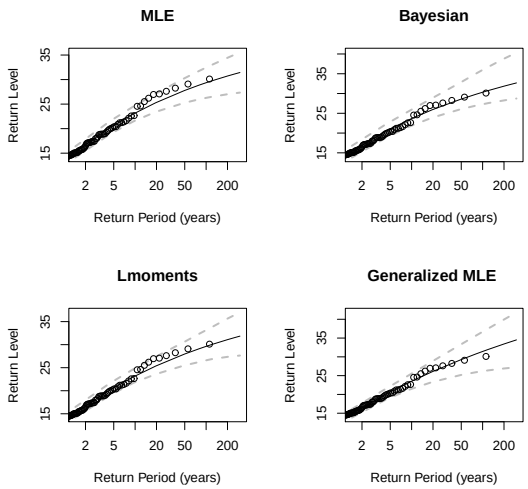

Figure 5: Return level plots of MLE, L-moment, Bayesian and Generalized MLE methods (Solid line represents the GEV model estimates with dashed lines indicate the 95% confidence interval,black dots denote the empirical estimates).

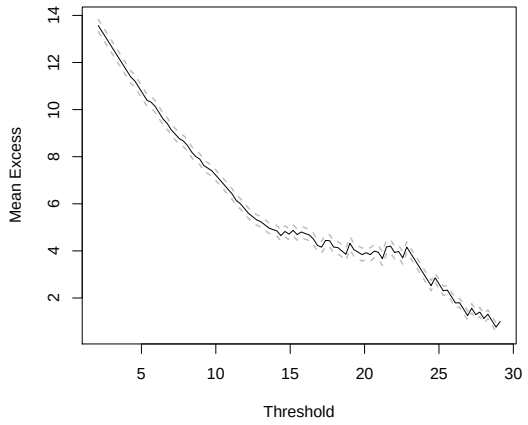

Figure 6: Mean Residual Life Plot of River Flow Data. Used to set the threshold.

Based on the mean residual life plot (Figure 6) we selected threshold $u = 15$ as the curve is linear from 0 to 15 after that it become non-linear. Maximum likelihood estimates in this case using Generalized Pareto Distribution (GPD) approximation for Pick Over Threshold

(POT) model is given in Table 7.

Model	5yr	25yr	50yr	100yr	200yr
GEVD	20.26	25.93	27.76	29.35	30.72
Bayesian	20.43	26.35	28.29	29.94	31.45
POT	33.33	34.48	34.83	35.12	35.36

Table 5: Return levels for 5, 25, 50, 100 and 200 years

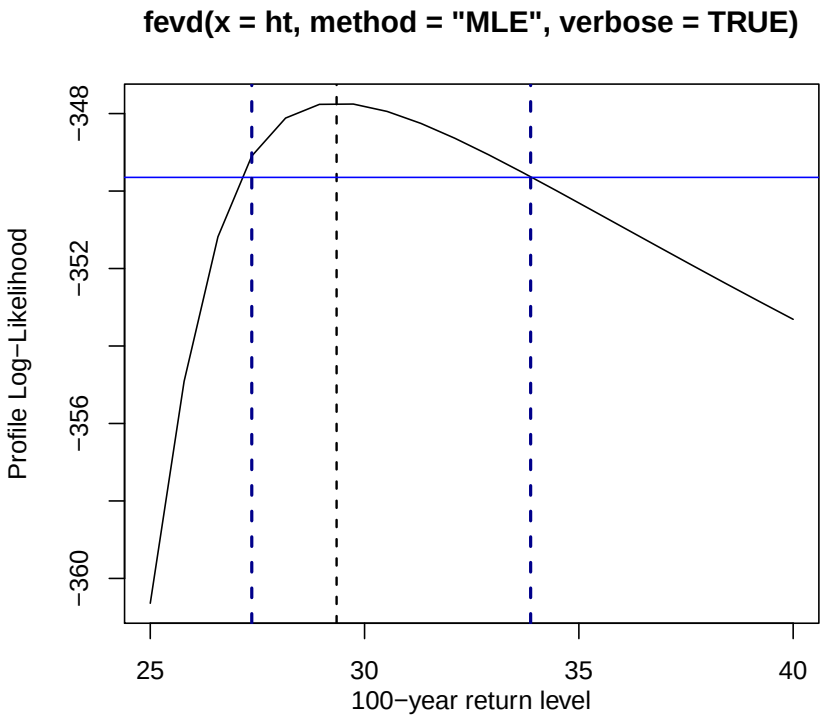

Figure 7: Profile log likelihood plot shows the similar estimates of 100 year return level. 95 % Confidence interval for 100 year return level is approx. (27,34) which is close to the estimate by other methods given in Table 8.

Model	1965	1969	1993	2001	2010
Gage ht	29.09	27.07	30.11	26.96	28.25

Table 6: Actual historical data from MN river

Paramet	MLE	Bayesian	POT model
Loc (μ)	13.37 (0.557)	13.43 (0.677)	-
Scale (σ)	5.31 (0.386)	5.42 (0.456)	6.006 (1.216)
Shape (ξ)	-0.20 (0.060)	-0.189 (0.072)	-0.2795 (0.158)

Table 7: Estimated model parameters and standard errors (in parenthesis)

Method	Return Level	95% CI
MLE	29.247	(26.2645, 32.4305)
Bayesian	30.236	(27.248, 35.643)
POT	35.124	(14.687, 55.559)

Table 8: 100 year return level (normal approx)

Conclusions and Summary

Classical extreme value theory has suggested a number of techniques for statistical modeling however, these methods are not appropriate in some situations as extreme data are rare and assumption of limiting distribution may not be correct. The analysis of Minnesota river flood level data has been performed using traditional Block Maxima Model, relatively new Pick Over Threshold (POT) model, and nonparametric Bayesian MCMC technique. Comparison of estimates of different approaches have been made. Actual recorded values of flood level shows the model is well fitted to the data.

The analysis has been done considering only single long time series without incorporating other possible covariates such as rainfall, temperature etc. or by considering values at different sites. Davison and Smith (1989) has discussed this type of analysis concerning multivariate extreme.

Acknowledgment

This paper was completed during my sabbatical leave from Minnesota State University on Spring 2014. I would like to thank Minnesota State University, Mankato and Center for Applications in Natural Resource Mathematics (CARM), University of Queensland, Australia for providing me the support during this research work.

References

Balkema, A. and Haan, L. 1990. A Convergence Rate in Extreme-Value Theory, Journal of Applied Probability, 27:577-585.

Castillo, E. Hadi, S. Balakrishnan, N. and Sarabia, J. 2005. Extreme Value and Related Models with Applications in Engineering and Science, Willey.

Coles, S. 2001. *An Introduction to Statistical Modeling of Extreme Values.* Springer.

Davison and Smith, 1989. Models for exceedances over high threshold. Journal of Royal Statistical Society B. 393-442.

Gumbel, E.J. 1958. Statistics of Extremes. Columbia University Press, New York.

Fisher, R. A. and Tippet, L.H.C. 1928. On the estimation of the frequency distributions of the largest or smallest member of a sample. Proceeding of the Cambridge philosophical society, 24, 180-190.

Fréchet, M., 1927. "Sur la loi de probabilit de l'cart maximum." Ann. Soc. Polon. Math. 6, 93.

Gnedenko B.V., 1943. Sur la distribution limite du terme maximum d'une serie aleatoire, Annals of Mathematics, 44, 423453.

Kotz, S.; Nadarajah, S., 2000. Extreme value distributions: theory and applications, World Scientific.

Reiss, R.-D. and Thomas, M. 2007. Statistical Analysis of Extreme Values: with applications to insurance, finance, hydrology and other fields. Birkhäuser, 530pp., 3rd edition.

Katz, R. W., Parlange, M. B. and Naveau, P. 2002. Statistics of extremes in hydrology. Advances in Water Resources, 25, 1287-1304.

Todorovic, P. Zelenhasic, E. 1970. A stochastic model for flood analysis. Water Resour. Res., 6 , pp. 16411658

von Mises R. 1936. La distribution da la plus grande de n valeurs, Revue de l'Union Interbalkanique vol. 1 pp. 1-20