UNIVERSITI
TEKNOLOGI
MARA

What's what PSPM

EISSN: 2756-7729

SEPTEMBER 2023/ VOL 2

FACTORIAL!

Build a custom mobile apps using Thunkable

Extreme Value Analysis:
A better way to analyse rare datasets

FFEATURE EXTRACTION AND MATCHING FROM IMAGES

EXTREME VALUE ANALYSIS:
A BETTER WAY TO ANALYSE RARE DATASETS

Zuraida binti Jaafar
Pengajian Sains Matematik
Kolej Pengajian Pengkomputeran, Informatik dan Matematik, Universiti Teknologi MARA (UiTM),
Cawangan Negeri Sembilan, Kampus Kuala Pilah, 72000, Negeri Sembilan Darul Khusus, Malaysia.
zuraida@uitm.edu.my

Introduction

Imagine being in the centre of Malaysia, amidst the vibrant cities, lush rainforests, and breathtaking scenery that make up this lovely country. The weather in Malaysia often has sunny days, occasional rain showers, and light breezes. This predictable weather is what you would consider the "normal" situation.

However, once in a blue moon, a powerful monsoon passes through a few locations especially, on the shores of the South China Sea with heavy rain and strong winds. Many places are in danger of being submerged by rivers that overflow and the sea that surges. This extraordinary weather event represents extreme value data.

Extreme value data refers to observations or data points that represent exceptionally rare or extreme events within a dataset. These events are situated in the tails of the probability distribution and are typically characterized by their infrequent occurrence and significant deviation from the typical or expected values. Extreme value data can be found in various fields, from finance (market crashes) to meteorology (severe storms), and from healthcare (rare diseases) to engineering (catastrophic failures).

Figure 1: Various Field of Extreme Events

Extreme Value Theory (EVT)

The foundation of studying the probability of the occurrence of extreme values uses the Extreme Value Theory (EVT); a branch of statistics that focuses on modelling the extreme tails of probability distributions. The frequency and magnitude of extreme events are described by the upper part of the distribution, which is typically known as the "tail". Probability distributions can be classified into heavy-tailed, and light-tailed, depending on their tail behaviour (Figure 2).

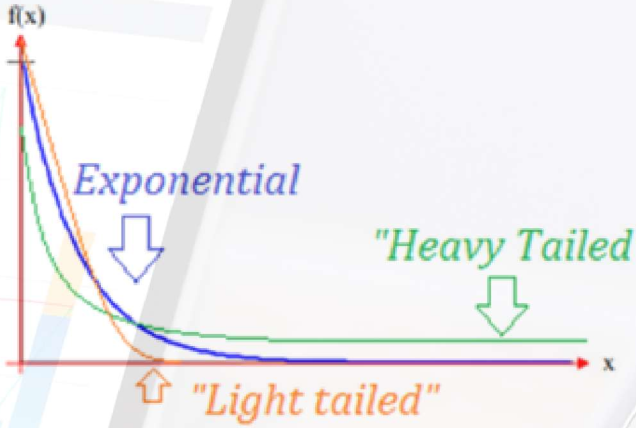

Figure 2: Graph of Extreme Events

Block Maxima (BM) approach

The field of EVT was pioneered by Fréchet (197), Fisher & Tippett (1928), Gumbel (1935) and Von Mises (1936) introduced the asymptotic theory of extreme value distributions. While Gnedenko (1943) and de Haan (1941) provided mathematical proof of the fact that under certain conditions, three families of distributions (Gumbel, Fréchet and Weibull) can arise as limiting distributions of extreme values in random samples. The three limiting distributions were unified by Jenkinson (1955) into a single expression known as the Generalized Extreme Value (GEV) distribution. This approach also can be defined as Block Maxima (BM). A single expression known as the Generalized Extreme Value distribution is given by:

$$F(x) = \exp\left(-\left(1+\gamma\left(\frac{x-\alpha}{\beta}\right)\right)^{-\frac{1}{\gamma}}\right), 1+\gamma\left(\frac{x-\alpha}{\beta}\right) \geq 0, \alpha > 0$$

where α ,β and γ are location, scale and shape parameters respectively.

The block maxima (BM) approach consists of dividing the observations into subsets (usually months or years) and considering the maximum value of each subset to be analysed (Figure 3).

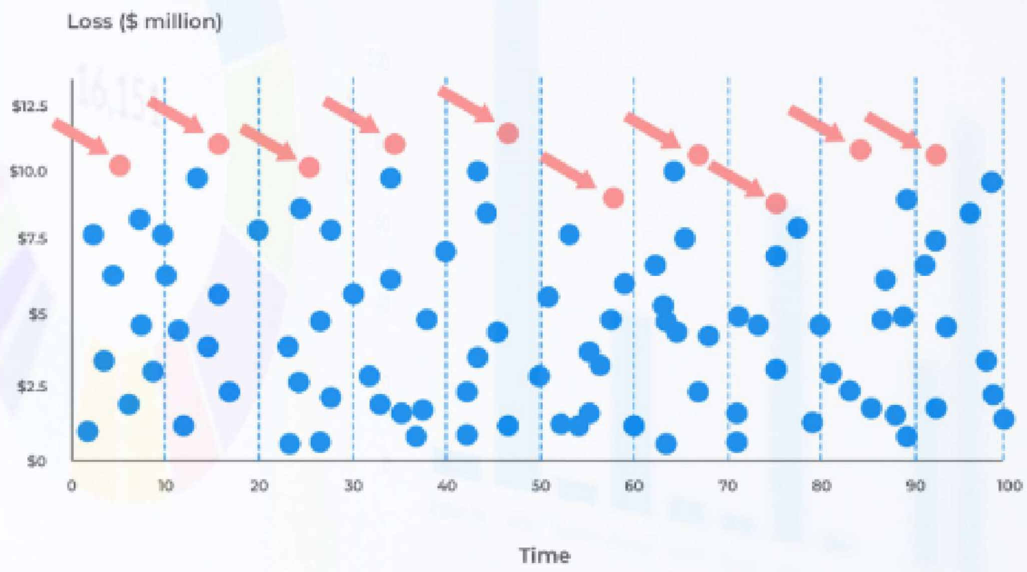

Figure 3: Block Maxima of Finance Losses

Peak-Over-Threshold (POT) approach

Instead of just considering the maximum value of each block as an extreme value, the observations above a high level or threshold can be considered extreme values as well. It is the so-called Peak-Over-Threshold method (POT) which is based on exceedances above thresholds. The threshold approach is the analogue of the Generalized Extreme Value distribution for the annual maxima, but it leads to a distribution called the Generalized Pareto Distribution (GPD) which is proven to be more flexible than the annual maxima (Goldstein et al., 2003; Smith & Shively, 1995). The mathematical form of the Generalized Pareto distribution is the following:

$$F(x;\mu,\sigma,\xi)=\begin{cases}1-\left(1+\xi\left(\frac{x-\mu}{\sigma}\right)\right)^{-\frac{1}{\xi}},\xi\neq 0\\ 1-\exp\left(-\left(\frac{x-\mu}{\sigma}\right)\right),\xi=0\end{cases}$$

where ξ , μ and σ are shape, location, and scale parameters respectively.

The fundamentals of extreme value analysis based on threshold values were established by Balkema & de Haan (1974) and Pickands (1975) whereby the mean number of exceedances above a high threshold in a cluster was given as the key parameter (Figure 4).

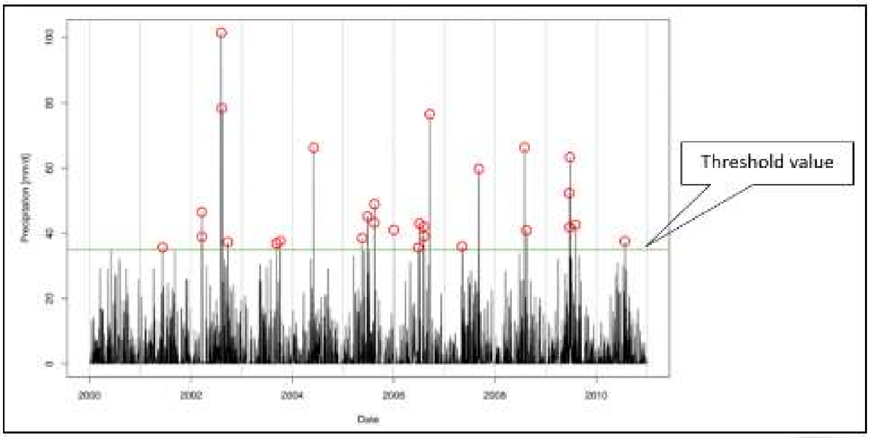

Figure 4: Peak-Over-Threshold of Precipitation

Practically, POT approach is more efficient for measuring the tail behaviour than the BM approach since it focuses on all observations which over the threshold. While BM approach involve loss of information as some blocks might have more than one extreme in them. However, the POT approach always facing a challenge in the choice of threshold, beyond which data can be considered as extreme data or more formally where the asymptotically justified extreme value models will provide an adequate approximation to the tail of the distribution.

Conclusion

In summary, analysing extreme value data is not only a matter of statistical interest but also a practical necessity in addressing risks, ensuring safety, protecting the environment, and making informed decisions in a wide range of domains. It allows us to prepare for and respond to rare but high-impact events, ultimately enhancing the safety, security, and well-being of individuals, communities, and societies.

References:

[1] A. A. Balkema, & L. de Haan. (1974). Residual Life Time At Great Age. Statistics, 2(5), 347–370. http://projecteuclid.org/euclid.aop/1176996548
[2] de Haan, L. (1941). On Regular Variation and Its Application to the Weak Convergence of Sample Extremes. In Mathematical Centre Tract (Vol. 32). https://doi.org/10.2307/1402855
[3] Fisher R.A., & Tippett L. H. C. (1928). Limiting forms of the frequency distribution in the largest particle size and smallest member of a sample. Proc. Camb. Phil. Soc., 24(x), 180–190.
[4] Fréchet, M. (1927). Sur la loi de Probabilité de l'écart Maximum. Ann.Soc.Polon.Math, 6, 93.
[5] Gnedenko, B. . (1943). Sur la distribution limite du terme maximum of d'unesérie Aléatorie. Annals of Mathematics, 44, 423–453. http://dx.doi.org/10.2307/1968974
[6] Gumbel, E. (1935). Les valeurs extrêmes des distributions statistiques. Annales de l'institut Henri Poincaré, 2, 115–158. https://eudml.org/doc/78993
[7] 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(348), 158–171. https://doi.org/10.1002/qj.49708134804
[8] Pickands, J. (1975). Statistical Inference Using Extreme Order Statistics. The Annals of Statistics, 3(1), 119–131. http://www.jstor.org/stable/2958083
[9] Von Mises, R. (1936). 'La distribution de la plus grandede nvaleurs.' American Mathematical Society, Reproduced, Selected Papers Volumen II, Providence, R.I (1954), 271–294.