Science & Technology Development Journal – Engineering and Technology, 3(SI3):SI45-SI51

Open Access Full Text Article

Research article

Using bootstrap to increase data in predictive analytics with extreme value distribution

Cuong K. Dang[1,*], Dam T. Duong[2], Duong T. T. Duong[3]

Use your smartphone to scan this QR code and download this article

ABSTRACT

The bootstrap is one of the method of studying statistical math which this article uses it but is a major tool for studying and evaluating the values of parameters in probability distribution. Overview of the theory of infinite distribution functions. The tool to deal with the problems raised in the paper is the mathematical methods of random analysis by theory of random process and multivariate statistics. Observations (realisations of a stationary process) are not independent, but dependence in time series is relatively simple example of dependent data. Through a simulation study we found that the pseudo data generated from the bootstrap method always showed a weaker dependence among the observations than the time series they were sampled from, hence we can draw the conclusion that even by re-sampling blocks instead of single observations we will lose some of structural from of the original sample. A potential difficulty by the using of likelihood methods for the GEV concerns the regularity conditions that are required for the usual asymptotic properties associated with the maximum likelihood estimator to be valid. To estimate the value of a parameter in GEV we can use classical methods of mathematical statistics such as the maximum likelihood method or the least squares method, but they all require a certain number samples for verification. For the bootstrap method, this is obviously not needed; here we use the limit theorems of probability theory and multivariate statistics to solve the problem even if there is only one sample data. That is the important practical significance that our paper wants to convey. In predictive analysis problems, in case the actual data is incomplete, not long enough, we can use bootstrap to add data.

Key words: Bootstrap, Time series, Bootstrap Jackknife, Generalized Extreme Value distributions, Predictive Analytics

[1]*Nong Lam University, HCMC, Vietnam*

[2]*University of Information Technology VNU-HCM, Vietnam*

[3]*Vietnam National University Ho Chi Minh City, Vietnam*

Correspondence

Cuong K. Dang, Nong Lam University, HCMC, Vietnam

Email: dkcuong@hcmuaf.edu.vn

History

- Received: 01-10-2019
- Accepted: 30-12-2020
- Published: 31-12-2020

DOI : 10.32508/stdjet.v3iSI3.608

Check for updates

VNU-HCM Press

BASIC FRAMEWORK

Block bootstrap methods for time series

Observations (realisations of a stationary process) are not independent, dependence in time series is relatively simple example of dependent data. Block bootstrap methods by time series data have been put forward by Hall, 1985[1], Kunsch[2], Singh[3], Politis and Romano[4], Lahiri[5].

Let X_1, X_2, X_n be independent and identically random variables with distribution function F. The basic step in extreme value theory is to investigate the distribution of $M_n = \max(X_1, X_2, X_n)$ as $n \to \infty$. Suppose there is sequence of constants $a_n > 0$, $b_n \in R$ such that:

$$P(M_n \le a_n x + b_n) = \lim_{n\to\infty} \lim F^n (a_n x + b_n) \underset{n\to\infty}{=} G(x), \ \forall x \in C(G) \quad (1)$$

with $G(x)$ is a non-degenerate distribution function, $C(G)$ is the set of all continuity points of $G(x)$. Limit distribution functions $G(x)$ satisfying equation. The function (1) is the well known extreme value of three types of distributions (Frechet, Weibull, and Gumbel distributions)[6,7].

The generalized extreme value (GEV) family of distribution is:

$$G(X) = e^{-\left(1+\xi\left(\frac{x-\mu}{\sigma}\right)\right)^{-1/\xi}}, \quad \left\{x : 1+\xi\left(\frac{x-\mu}{\sigma}\right) > 0\right\} \quad (2)$$

where μ is a location parameter $\mu \in R$, σ is a scale parameter (with $\sigma > 0$), and ξ is the extreme value shape parameter.

Moving Block Bootstrap (MBB)

Blocks length l, starting at X_i : $B_i = (X_i, X_{i+1}, ..., X_{i+l-1})$. To get a bootstrap sample we do:

(i) Draw with replacement $B_1^*, B_{1+1}^*, ..., B_k^*$ from $B_1, B_2, ..., B_{n-l+1}$.

(ii) Concatenate the blocks $B_1^*, B_{1+1}^*, ..., B_k^*$ to give the bootstrap sample $X_1^*, X_2^*, ..., X_{kl}^*, l = 1$, corresponds to the clacical i.i.d bootstrap.

Blocks in the MBB may overlap[8,9].

Cite this article : Dang C K, Duong D T, Duong D T T. **Using bootstrap to increase data in predictive analytics with extreme value distribution**. *Sci. Tech. Dev. J. – Engineering and Technology;* 3(SI3):SI45-SI50.

Non overlap Block Bootstrap (NBB)

Blocks of length l: $B_1 = (X_1, X_2, ..., X_l)$; $B_2 = (X_l, X_{l+1}, ..., X_{2l})$ ...

$B_{\left[\frac{n}{l}\right]} = (X_{n-l+1}, X_{n-l+1}, ..., x_n)$. To get a bootstrap sample we do:

Resample blocks $B_1^*, B_2^*, ...B_{\left[\frac{n}{l}\right]}^*$ with replacement,

Concantenate to get bootstrap sample $X_1^*, X_2^*, ...X_{\left[\frac{n}{l}\right]}^*$

$b = \left[\frac{n}{l}\right]$ blocks ($\left[\frac{n}{l}\right]$ is the largest number less than or equal to $\frac{n}{l}$).

$B_i = (X_{il}, X_{il+1}, ..., X_{il+l-1}), i = 1, 2, ..., \left[\frac{n}{l}\right]$.

NBB fewer blocks than in the MBB [10].

Circular block bootstrap (CBB)

Figure 1 show the blocks of length l, with sample replacement from $\{B_1, B_2, ..., B_m\}$;

$lb = m \approx n$, every observation reseives equal weight:

$B_i = (X_i, X_{i+1}, ..., X_{i+l-1}), i = 1, 2, ..., m$.

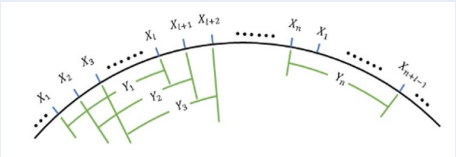

Figure 1: Blocks of length *l*.

Stationary block bootstrap (SB)

Blocks are no longer of equal size [11]. The bootstrap sample is chosen according to some probability measure on the sequences. The static bootstrap method involves sampling blocks of random lengths by each block with a geometrical distribution:

$(H_1, K_1), (H_2, K_2), ...H \sim \text{Uniform}(1, 2, ..., n)$, $K \sim \text{Geometric}(p)$ for some $p > 0$.

Properties of Block Bootstrap Methods

Through a simulation study we found that pseudo data achieved from the bootstrap method always displayed a weaker dependence among the observations than the time series they were sampled from, hence we can draw the conclusion that even by resampling blocks instead of single observations we will lose some of structural from of the original sample.

The pseudo time series produced by the moving block method is not stationary, even if the original series X_t is stationary.

The pseudo time series produced by the stationary bootstrap method is actually a stationary time series

The mean $\bar{X}_N^*$ of the moving block boostrap is biased in the sense that:

$$E\left(\bar{X}_N^* | X_1, X_2, ..., X_n\right) - \bar{X}_n \neq 0$$

The MBB estimator of the variance of $\sqrt{n}\bar{X}_n$ is biased. This situation creates problems in using the percentile method with the MBB.

The usual estimator:

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \square \left(X_i - \bar{X}_n\right)^2$$

Should be modified to:

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \square \{\left(X_i - \bar{X}_n\right)^2 + \sum_{k=1}^{i-1} \square \sum_{i=1}^{n-k} \left(X_i - \bar{X}_n\right)\left(X_{i+k} - \bar{X}_n\right)\}$$

By the modification the bootstrap will be able to improve substantially on the normal approximation.

Comparison the block bootstrap methods: We find that overall the MBB and CBB methods give the estimators with the smallest standard error and the SB method the largest. (Random block length leads to a larger variance of the parameter estimates than for the other methods when block length is fixed). The b

The block bootstrap procedure

Assume that, the statistic $\widehat{\theta}$ estimates is a θ functional, depending on the m-dimensional marginal distribution of the time series data. Now, build vectors of consecutive obsevations

$Y_t = (X_{t-m+1}, ..., X_t)$, $t = m, ..., n$.

Build overlapping block of consecutive vectors,

$(Y_m, ..., Y_{m+l-1}), (Y_{m+1}, ..., Y_{m+l}), ..., (Y_{m+l-1}, ..., Y_n)$

where $l \in N$ is the block length parameter. Simplicity, first assume that

$m + l - 1 = kl$ with $k \in N$. So that, resample block independently,

$Y_{S_1+1}, ..., Y_{S_1+l}, Y_{S_2+1}, ...Y_{S_2+l}, ..., Y_{S_k+1}, ..., Y_{S_k+l}$,

where the starting points of blocks $S_1, ..., S_k$ are i.i.d. Uniform $(\{m-1, ..., n-l\})$ These resampled blocks of m-vectors could be referred to the block bootstrap sample. On the other hand, as we will concern the bootstrapped block estimator is not simply defined by the plug-in rule and the concept of the bootstrap sample is not clear. If $n - m + 1$ is not a multiple of l, we resample $k = [(n = m + 1)/l] + 1$ blocks, but we use only a portion of the k-th block to get $n - m + 1$ resampled vectors in total.

Assume that $\widehat{\theta} = T\left(F_n^{(m)}\right)$,

where $F_n^{(m)}(x) = \frac{1}{n-m+1}\sum_{i=m}^{n} \square 1_{[Y_t \leq x]}$ is a empirical distribution function of the m-dimensional marginal distribution of $(X_t)_{t \in Z}$, and T is a smooth functional. The block bootstrapped estimator is defined as

$\widehat{\theta}^{*B} = T\left(F_n^{(m)*B}\right)$,

$F_n^{(m)*B}(x) = \frac{1}{n-m+1}\sum_{i=m}^{n} \square \sum_{t=S_i+1}^{S_i+l} \square 1_{[Y_t \leq x]}$

we have $E^{*B}\left[\widehat{\theta}^{*B}\right] \neq \widehat{\theta}$.

This definition of the block bootstrapped estimator, can be interpreted as

$$\widehat{\theta}^{*B} = \varphi_{n-m+1}(Y_{S_1},...,Y_{S_1+l}, Y_{S_2},...Y_{S_2+l},..., Y_{S_k},..., Y_{S_k+l})$$
$$\widehat{\theta} = \varphi_{n-m+1}(Y_m,...,Y_n),$$

So, we can say that it employs a plus-in rule based on vectorized observations.

Choosing an optimal block length

The orders of magnitude of the optimal block size are known in some inference problems. According to those authors three different settings of practical importance can be identified; estimation of the bias or variance, estimation of a one-sided distribution function and estimation of a two-sided distribution function.

The optimal block length in the above situations of different size being $b \sim Cn^{\frac{1}{k}}$, with $k = 3,4,...$ respectively where n is the sample size.

This result, will be used here as the basis for the choosing the optimal block length. Two main approaches can be pointed out:

A cross validation method proposed by Hall et. al.[12] and a plug-in method based on a recent work of Lahiri[5].

Based on a research of Lahiri, a nonparametric plug-in (NPPI) method for selecting the optimal block length in order to reduce the bias will be considered. Unlike traditional plug-in method this method employs nonparametric resampling procedures to estimate the relevant constants in the leading term of the optimal block length.

The variance of block bootstrap estimator is an increasing function of the block length l while its bias is a decreasing function of block length. As a result, for each block bootstrap estimator, there is critical value, l_n^0 that minimizes mean-square error (MSE). The value of l that minimizes the leading term in the expension of the MSE is denominated MSE-optimal block length.

The following notation will be used:

$\triangle_n = \theta(m_n) - \theta(\mu)$;
$\beta = E\triangle_n$; $\sigma_\infty^2 = Var(\sqrt{n}\triangle_n)$;
$\widehat{\triangle}_n = \theta(\widehat{m}_n) - \theta(\widehat{E}\widehat{m}_n)$;
$\widehat{\beta} = \widehat{E}\widehat{\triangle}_n$; $\widehat{\sigma}_\infty^2 = \widehat{Var}(\sqrt{n}\widehat{\triangle}_n)$;

The block bootstrap will be able to use either overlapping or non-overlapping blocks. Define one-sided and symmetrical distribution functions of the normalized statistic $\sqrt{n}\frac{\triangle_n}{\sigma_\infty}$ by $F_1(z) = P\left(\sqrt{n}\frac{\triangle_n}{\sigma_\infty} < z\right)$ and $F_2(z) = P\left(\sqrt{n}\frac{|\triangle_n|}{\sigma_\infty} \leq z\right)$

Define bootstrap analogs of F_1 and F_2 by $\widehat{F}_1(z) = \widehat{P}\left(\sqrt{n}\frac{\widehat{\triangle}_n}{\widehat{\sigma}_\infty}\right)$ and $\widehat{F}_2(z) = \widehat{P}\left(\sqrt{n}\frac{\left|\widehat{\triangle}_n\right|}{\widehat{\sigma}_\infty} < z\right)$. Let $(\psi, \widehat{\psi})$ denote either $\left(\beta, \widehat{\beta}\right)$ or $\left(\frac{1}{n}\sigma_\infty^2, \frac{1}{n}\widehat{\sigma}_\infty^2\right)$, and ϕ denote the standard normal density function: $\phi(x) = \frac{1}{\sqrt{2\pi}} exp\left\{-\frac{1}{2}x^2\right\}$.

Hall et. al. show that there are constants $C_i\,(i = 1,2,...,6)$ such that, in addition, $n^{-1} + n^{-1}l = o(1)$ as $n \to \infty$ then

$$E(\psi - \widehat{\psi})^2 \sim \frac{1}{n^2}\left(C_1\frac{1}{l^2} + C_2\frac{1}{l}\right), \quad (3)$$

$$E\left[F_1(z) - \widehat{F}_1(z)\right]^2 \sim \frac{1}{n^2}\left(C_3\frac{1}{l^2} + C_4\frac{l^2}{n}\right)\phi(z)^2, \quad (4)$$

$$E\left[F_2(z) - \widehat{F}_2(z)\right]^2 \sim \frac{1}{n^2}\left(C_5\frac{1}{l^2} + C_6\frac{l^3}{n}\right)\phi(z)^2, \quad (5)$$

Where the symbol ~ indicates that the quantiny on the right-hand side is the leading term of an asymptotic expansion. The constants C_i do not depend on n or l. The terms involving C_2, C_4 and C_6 correspond to the variance. The variance terms smaller if the bocks are overlapping than if they are non-overlapping.

Following the expressions (3), (4), and (5), so that the asymptotically optimal block length (in the case of minimizing the AMSE) is $l = A_1 n^{1/3}$ for bias or variance estimation, $l = A_2 n^{1/4}$ for estimating a one-sided distribution function, and $l = A_3 n^{1/5}$ for estimating a symmetrical distribution function ($A_j > 0$; $j = 1,2,3$ are suitable constants that dependents on certain population parametrs).

JACKKNIFE METHOD

We assume a vector of parameters such as θ. The bias of θ as an estimate of an estimat or $\widehat{\theta}_0$ of θ_0 is defined by $\triangle = E\widehat{\theta}_0 - \theta_0$.

A large bias is often an undesirable factor in the estimator's performance. We will able to use the bootstrap to estimate the bias of any estimator θ as an estimate of an estimator $\widehat{\theta}_0$. We generate B independent bootstrap samples X^{*1},X^{*2},...,X^{*B}, each consisting of n data value drawn with replication corresponding to each bootstrap sample from X, as $X^{*1} = X_{i1}$, $X^{*2} = X_{i2}$,..., $X^{*} = X_{in}$.

We can select the sample of B in the range 25 – 1000. Then evaluate of the bootstrap application corresponding to each bootstrap sample, it may be an indication that the statistic. $\widehat{X} * (b) = SX^{*b}$, $b = 1,2,...,B$. The bootstrap estimate of bias is defined by $\triangle_B = \widehat{\theta}_0^* - \widehat{\theta}_0$ where $\widehat{\theta}_0^* = \frac{1}{n}\sum_{b=1}^{B}\widehat{\theta}_b^*$.

We have focused on the standard error as a measure of accuracy for an estimator $\widehat{\theta}_0$. Estimate the standart error $se_B\widehat{\theta}_0$ by the sample standard deviation of the B replications,

$$\widehat{se}_B = \left[\frac{1}{B-1}\Sigma_{b=1}^{B}\left(\widehat{\theta}(b) - \widehat{\theta}_0^*\right)^2\right]^{\frac{1}{2}}. \quad (6)$$

The jackknife estimate of bias is a alternative method to find out the bias which it was original computer based method for estimating biases and standard errors. If we have a sample set $x = (x_1, x_2, ..., x_n)$ and an estimator $\widehat{\theta}_0 = S(X)$. The i^{th} jackknife sample $x_{(i)}$, is defined to be x with i^{th} data point removed, $x_{(i)} = (x_1, x_2, ..., x_{i-1}, x_{i+1}, ...x_n)$.

For i = 1,2,...,n, the jackknife estimate of bias is defined by

$\widehat{se}_{jack} = \left[\frac{n-1}{n}\Sigma_{i=1}^{n}\square\right]$, where $\widehat{\theta}_{(\cdot)} = \frac{1}{n}\Sigma_{i=1}^{n}\square\widehat{\theta}_{(i)}$. The jackknife estimate of standard error is $\widehat{se}_{jack} = \left[\frac{n-1}{n}\Sigma_{i=1}^{n}\square\left(\widehat{\theta}_{(i)} - \widehat{\theta}_{(\cdot)}\right)^2\right]^{\frac{1}{2}}$

The jackknife usually provides a nice and simple approximation to the bootstrap, for estimation of standard error and bias.

As a basic rule, if a bias of less than 0.25, the standard errors can be ignored, unless we are trying to do careful confidence interval calculations. The root mean square error of an estimator $\widehat{\theta}$ for θ is $\sqrt{E\left(\widehat{\theta} - \theta\right)^2}$ a measure of accuracy that takes into regards both bias and standard error. It can be shown that the square root of the square is equal,

$$\begin{aligned}\sqrt{E\left(\widehat{\theta} - \theta\right)^2} &= \widehat{se}\sqrt{1 + \left(\frac{\triangle}{\widehat{se}}\right)} \\ &\cong \widehat{se}\left[1 + 0.5\left(\frac{\triangle}{\widehat{se}}\right)^2\right]\end{aligned} \quad (7)$$

If $\Delta = 0$ then the root mean square equals its minimum value of standard error. Reverse, rate $\left|\frac{\triangle}{\widehat{se}}\right| = 0.25$, then the root mean square error is no more that 0.031 greater than value of standard error.

The obvious bias corrected estimator is, $\theta_{corr} = \widehat{\theta}_0 - \triangle = 2\widehat{\theta}_0 - \widehat{\theta}_0^*$ where, $\triangle = \triangle_B$. When bias is small compared to the estimated standard error *se*; then it is b safer to use $\widehat{\theta}_0$ than θ_{corr}. Reverse, bias is large compared to standard error, then it may be an indication that the statistic $\widehat{\theta}_0 = S(X)$ is not an appropriate estimate of the parameter θ.

Quantifying the accuracy of an estimation tool can often be clearer by calculating confidence intervals. A standard result claims that $\widehat{\theta}_0$ is the maximum likelihood estimator has a limiting multivariate normal distribution with mean θ_0 and variance covariance matrix $H_{\theta_0} = I(\theta_0)^{-1}$,

Where $I(\theta) = \left[e_{i,j}(\theta)\right]_{dxd}$ with $e_{i,j}(\theta) = -E\frac{\partial^2 L(\theta)}{\partial\theta_i\partial\theta_j}$, and $L(\theta) = \Sigma_{i=1}^{n}\square\log\log f_0(x_i)$ is likehood function. The matrix $I(\theta)$ is "Fischer's information matrix". Because the true value of θ_0 is generally unknown, it is usual to approximate the term of I with those of the "Fischer's information matrix" defined by $I(\theta) = \left[-\frac{\partial^2 L(\theta)}{\partial\theta_i\partial\theta_j}\right]_{dxd}$ and evaluated at $\theta = \widehat{\theta}$. Denoting an arbitrary term in the inverse of $I_O(\theta)$ by $\widetilde{\sigma}_{i,j}$, it follows that an approximate $(1-\tau)$ with $0 < \tau < 1$, confidence interval for θ_0 is $\widehat{\theta}_i \pm z_{\tau/2}\sqrt{\widetilde{\sigma_{i,j}}}$

Let $\widehat{\theta}_0$ be the maximum likelihood estimator of the k-dimensional parameter θ_0 with approximate variance covariance matrix H_{θ_0}.

Moreover, a confidence interval can derived from the likelihood function, using approximation

$$L(\theta_0) = 2\left(L\left(\widehat{\theta}_0\right) - L(\theta_0)\right) \sim \chi_2^2$$

It follows that an approximate $(1-\tau)$ confidence region for θ_0 is given by $C_\tau = \{\theta: L(\theta) \le c_\tau\}$, where c_τ is the $(1-\tau)$ quantile of the χ_d^2 distributtion. This approximation is usually more accurate than that based on the asymptotic normality of the maximum likelihood estimator.

The log likelihood for θ can be formally written as $L_p\left(\theta^{(1)}, \theta^{(2)}\right)$ where θ have two components.

The profile log likelihood for $\theta^{(1)}$ is define as

$L_p\left(\theta^{(1)}\right) = L\left(\theta^{(1)}, \theta^{(2)}\right)$

and similarly,

$L_p\left(\theta^{(2)}\right) = L\left(\theta^{(1)}, \theta^{(2)}\right)$.

So, under suitable regularly conditions, for large n,

$L_p\left(\theta^{(1)}\right) = 2\{L\left(\widehat{\theta}_0\right) - L_p\left(\theta^{(1)}\right)\} \sim \chi_k^2$.

For a single component θ_i, $C_r = \left\{\theta_i : L_p(\theta_i \le c_r)\right\}$ is a $(1-\tau)$ confidence interval, where c_r is the $(1-\tau)$ quantile of the χ_1^2 distribution.

Another method of model selection is the Akaike Information Criterion (AIC). The AIC has played a significant role in solving problems in a wide variety of fields as a model selection criterion for analyzing actual data. The AIC is defined by

AIC = -2 (maximum log likelihood) +2 (number of free parameters).

The amount of free parameters in a model refers to the dimensions of the parameter vector θ contained in the specified model $f(x|\theta)$.

ANALYSIS FOR GEV DISTRIBUTIONS

An implicit difficulty with the using of likelihood methods for the GEV concerns the regularity conditions that are required for the usual asymptotic properties associated with the maximum likelihood estimator to be valid. Those conditions are not satisfied by the GEV model because the end points of the GEV distribution are functions of the parameter values, $\mu = -\frac{\sigma}{\xi}$ is an upper end-point of the distribution when $\xi < 0$; and a lower end-point when $\xi > 0$. These offend of the usual regularity conditions means that the standard asymptotic likelihood results are not automatically applicable[13,14]. This problem have studied in our details and obtained the following results

(i). if $\xi > -0.5$ maximum likelihood estimators are regular, in the impression of having the usual, asymptotic properties

(ii). when $-1 < \xi < -0.5$, maximum likelihood estimators are generally obtainable, but do not have the standard asymptotic properties, and

(iii). with $\xi < -1$, maximum likelihood estimators are unlikely to be obtainable.

Under the assumption that X_1, X_2 ,..., X_m are independent random variables having the GEV distribution, the log likelihood for the GEV parameters when $\xi \neq 0$ is

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

where $1+\xi\left(\frac{x_i-\mu}{\sigma}\right) > 0$, with i =1, 2, …, m.

The case $\xi \to 0$ requires separate treatment using the Gumbel limit of the GEV distribution. This leads to the log likelihood

$$L(x,\mu,\sigma,\xi) = -m\log\sigma - \left(1+\frac{1}{\xi}\right)\sum_{i=1}^{m} \left(\frac{x_i-\mu}{\sigma}\right) - \sum_{i=1}^{m} exp\left[-\left(\frac{x_i-\mu}{\sigma}\right)\right] \quad (9)$$

There is no analytical solution, but for any given dataset the maximization is straightforward using standard numerical optimization algorithms.

Estimates of extreme quantile of the maximum distribution under linear normalization are obtained by inverting equation (1)

$$x_p = \begin{cases} \mu - \frac{\sigma}{\xi}\left(1-(-\log(1-p))^{-\xi}\right); & \xi \neq 0 \\ \mu - \sigma\log\log(1-p); & \xi = 0 \end{cases} \quad (10)$$

The return levels are exceeded by the annual maximum in any particular time with probability $(1-p)$.

If x_p are plotted against $\frac{1}{1-p}$ the plots are linear. By substituting the maximum likelihood estimates of the GEV parameters into (10), the maximum likelihood estimate of x_p for $0 < p < 1$, is obtained as

$$\hat{x}_p = \begin{cases} \hat{\mu} - \frac{\hat{\sigma}}{\hat{\xi}}\left(1-y_p^{-\xi}\right), & \hat{\xi} \neq 0 \\ \hat{\mu} - \hat{\sigma}\log\left(y_p\right), & \hat{\xi} = 0 \end{cases} \quad (11)$$

where $y_p = -\log\log(1-p)$. By the delta method, we get

$$Var\left(x_p\right) \cong \nabla x_p^T H_0 \nabla x_p,$$

where $\theta = [\mu, \sigma, \xi]$, H_0 is variance covariance matrix, and

$$\nabla x_p^T = \left[\frac{\partial x_p}{\partial \mu}; \frac{\partial x_p}{\partial \sigma}; \frac{\partial x_p}{\partial \xi}\right] = [1, -\xi^{-1}\left(1-y_p^{-\xi}\right), \sigma\xi^{-2}(1-y_p^{-\xi}) - \sigma\xi^{-1}y_p^{-\xi}\log\left(y_p\right)]$$

corresponding value is $\left(\hat{\mu}, \hat{\sigma}, \hat{\xi}\right)$.

When the $\hat{\xi} \to 0$ and equation (12) is still valid with

$$\nabla x_p^T = \left[1, \log y_p\right], \quad (12)$$

corresponding value is $(\hat{\mu}, \hat{\sigma})$.

Profile likelihood. Numerical evaluation of the profile likelihood for any of the individual parameters μ, σ or ξ is straightforward. For example, to obtain the profile likelihood for ξ, we fix $\xi = \xi_0$, and maximize the log likelihood (8) with respect to the remaining parameters, μ and σ. This is repeated for a range of values of ξ_0. This methodology can also be applied when inference is required on some combination of parameters. In particular, we can obtain the profile likelihood for any specified return level xp: This requires a re-parameterization of the GEV model, so that xp is one of the model parameters, after which the profile log likelihood is obtained by maximization with respect to the remaining parameters in the usual way. Re-parameterization is straightforward,

$$\mu = \begin{cases} x_p - \frac{\sigma}{\xi}\left(1-y_p^{-\xi}\right), & \xi \neq 0, \\ x_p - \sigma\log\left(y_p\right), & \xi = 0. \end{cases} \quad (13)$$

so that replacement of μ in (8),(9) with (13) has the desired effect of expressing the GEV model in terms of the parameters (x_p,σ,ξ).

Model validity. A probability plot is a comparison of the empirical and fitted distribution functions. With ordered block maximum data $x_1 \leq x_2 \leq \ldots \leq x_m$, the empirical distribution function evaluated at x_i is given by

$$G\left(x_{(i)}\right) = \frac{i}{m+1}$$

By substitution of parameter estimates into (2), the corresponding model based estimates are

$$\widehat{G}(x_i) = \{exp\left[-\left(1+\widehat{\xi}\left(\frac{x_{(i)}-\widehat{\mu}}{\widehat{\sigma}}\right)\right)^{-\frac{1}{\xi}}\right]; \xi \neq 0,$$

$$exp\left[-exp\left(-\left(\frac{x_{(i)}-\widehat{\mu}}{\widehat{\sigma}}\right)\right)\right]; \xi \neq 0.$$

We then construct plot consisting of the points.

$$\left\{\left(G^{-1}\left(x_{(i)}\right), \widehat{G}\left(x_{(i)}\right)\right), i = 1,2,...,m\right\}$$

A weakness of the probability plot for extreme value models is that both $G^{-1}\left(x_{(i)}\right)$ and $\widehat{G}\left(x_{(i)}\right)$ are bound to approach 1 as $x_{(i)}$ increases, while it is usually the accuracy of the model for large values of x that is of greatest concern. That is, the probability plot provides the least information in the region of most interest. This deficiency is avoided by the quantile plot, consisting of the points

$$\left\{\left(\widehat{G}^{-1}\left(\frac{i}{m+1}\right), x_{(i)}\right), i = 1,2,...,m\right\}$$

If $\widehat{G}$ is a reasonable estimate of G, then the quantile plot should also consist of points close to the unit diagonal.

CONCLUSION

To estimate the value of a parameter in GEV we can use classical methods of mathematical statistics such as the maximum likelihood method or the least squares method but they all require a certain number. samples for verification. For the bootstrap method, this is obviously not needed; here we use the limit theorems of probability theory and multivariate statistics to solve the problem even if there is only one sample data. That is the important practical significance that our paper wants to convey. We used the bootstrap method to process statistical data in hydrological and used random calculations, R software for analysis data.

In the research of Cuong at. el.[15]. Regarding water, salinity and flood peaks of the Mekong Delta, we have forecasted for the period up to 2020 based on data from 1976 to 2016. But we still see that the data is not long enough for more accurate forecasts, therefore we will use bootstrap to increase data for Predictive Analytics problems.

LIST OF ABBREVIATION

AIC: Akaike Information Criterion
CBB: Circular Block Bootstrap
EVD: Extreme Value Distribution
GEV: General Extreme Value
MBB: Moving Block Bootstrap
MSE: Mean-Square Error
NBB: Non-Overlapping Block Bootstrap
NPPI: Non-Parametric Plug-In
SB: Stationary Block Bootstrap

CONFLICT OF INTEREST

The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

AUTHORS' CONTRIBUTION

Cuong K. Dang, and Dam T. Duong conceived the idea of the paper, developed the research methodology and implemented formal analysis. Duong T. T. Duong, collected and manipulated input data, prepared the manuscript. All authors have read and agreed to the published version of the manuscript.

REFERENCES

1. Peter Hall. The Bootstrap and Edgewood Expansion. Springer Series in Statistics. 1992;Available from: https://doi.org/10.1007/978-1-4612-4384-7.
2. Kunsch HR. The jackknife and the bootstrap for general stationary obsersations. Ann. Statist. 1989;17(09):1217–1241. Available from: https://doi.org/10.1214/aos/1176347265.
3. Singh S. A computational method of forecasting based on fuzzy time series. Mathematics and Computers in Simulation. 2009;79:539–554. Available from: https://doi.org/10.1016/j.matcom.2008.02.026.
4. Politis DN, Romano JP. The stationary bootstrap. Journal of Americal Statistical Association. 1994;89:1303–1313. Available from: https://doi.org/10.1080/01621459.1994.10476870.
5. Lahiri S. Resampling Methods for Dependent Data. Springer. 2014;.
6. Coles S. An introduction statistical modeling of extreme values. Springer. 2001;Available from: https://doi.org/10.1007/978-1-4471-3675-0.
7. Castollo E, Hadi AS, Balakrishnan N, Sarabia JM. Extreme Value and Related Models with Application in Engineering and Science. Wiley-Intercience. 2004;.
8. Davison AC, Hinkley DV. Bootstrap method and their application. Cambridge University Press. 1997;Available from: https://doi.org/10.1017/CBO9780511802843.
9. Efron B, Tibshirani R. An introduction to the bootstrap. Chapman & Hall/CRC. 1993;Available from: https://doi.org/10.1007/978-1-4899-4541-9.
10. Radovanov B, Marcikie A. A comparison of four different block bootstrap methods. Croatian Operational Research Review. CRORR 5. 2014;p. 189–202. Available from: https://doi.org/10.17535/crorr.2014.0007.
11. Manfred Mudelsee. Climate Time Series Analysis. Classical Statistical and Bootstrap Method. Second Edition, Springer. 2014;.
12. Hall P, Horowitz JL, Jing BY. On blocking rules for the bootstrap with dependent data. Biometrika. 82. 1995;651(574). Available from: https://doi.org/10.1093/biomet/82.3.561.
13. Dam DT, Tai VV, Truc PM, Cuong DK. Forecasting crest of salinity at the main stations of Ca Mau province by fuzzy time series model. Can Tho University Journal of Science. 2016;74:86–78.
14. Dam DT, Cuong DK, Linh HM. Solving the Problem of Hydrometeorological Data Analysis by Random Process Theory. Science & Technology Development Journal. 2017;20(K2-2017):101–106. Available from: https://doi.org/10.32508/stdj.v20iK2.455.

15. Cuong DK, Dam DT, Duong DTT, Loi NK, Vo NS, Kortun A. Extreme Value Distributions In Hydrological Analysis In The Mekong Delta: Case Study In Ca Mau, An Giang Provinces. EAI Endorsed Transactions on Industrial Networks and Intelligent Systems Journal. 2019;6. Available from: https://doi.org/10.4108/eai.13-6-2019.159122.