Journal of Mathematics and System Science 6 (2016) 129-138
doi: 10.17265/2159-5291/2016.04.001

Application of Extreme Value Theory to Generation and Analysis of Pseudorandom Samples

Svitlana Trukhan[1] and Petro Bidyuk[2]

1. National Technical University of Ukraine 'KPI', Institute for Applied System Analysis, Mathematical methods of System Analysis Department, Peremohy Avenue, 37, Kyiv, 03056, Ukraine.

2. National Technical University of Ukraine 'KPI', Institute for Applied System Analysis, Mathematical methods of System Analysis Department, Peremohy Avenue, 37, Kyiv, 03056, Ukraine.

Received: January 19, 2016 / Accepted: February 16, 2016 / Published: April 25, 2016.

Abstract: The article deals with the methodology of pseudorandom data analysis. As a mathematical tool for carrying out the research the extreme value theory was used that creates one of the directions in mathematical statistics, and is related to investigating the extreme deviations from the median values in probability distributions. Also, the methods for estimating unknown parameters and algorithm of random-number generation are discussed. The models of treatment the extreme values are constructed which are based on machine generated sample and approach is proposed for their future application for constructing forecasting models.

Keywords: Extreme value theory, extreme value threshold, simulation and modeling, maximum likelihood estimator, pseudorandom sample generation.

1. Introduction

According to the present knowledge the simulation modeling is a powerful tool for research and creation of complex system models and forecasting processes of various origin including those associated with management and decision making under risk. Comparing with other approaches the simulation modeling allows us to generate a large number of alternatives, and thus improve the quality of managerial decisions and more accurately predict their consequences. The aim of the simulation modeling is in constructing of a simulation model of a system (object) under study and using the results of simulated experiment for studying the law(s) of functioning (for example, the probabilities distribution law of a random variable), the system behavior with regard to defined limits, and the target functions in terms of interaction with specific environment.

However, practical usage of this method for solving various problems is not very popular. Primarily, because there are a lot of difficulties with combined application of respective mathematical tools and due to the necessity of processing large sophisticated data sets. In general, there are also cases with empty data sets and that is why it is necessary to generate the modeling data using appropriate mathematical tools for approximate reproduction of a real world of the process of interest.

Due to the necessity to solve new tasks of modeling and forecasting the processes based on large data samples, which couldn't be accomplished with existing methods and techniques, we come to the next conclusion. The field of development modern integrated data processing systems, methods and approaches for treating such data sets should be carefully studied. One of these approaches is extreme value analysis (EVA) or extreme value theory. It is widely used to solve such tasks as a regulation of the

Corresponding author: Svitlana Trukhan, master degree, postgraduate student at the Institute for Applied System Analysis, research fields: mathematical methods of system analysis, applied statistics, time series analysis, software engineering. E-mail: Svetlana.trukhan@gmail.com.

structure of portfolio assets in finances (e.g. in insurance and investments), analysis of occurrence of the risky situations in financial organizations, traffic prediction in telecommunication, weather forecasting etc.

The extreme value theory is focused on purposeful analysis and evaluation of the probability of random variables occurrence associated with extreme events, and considers it in as a rare event. Generally, extreme values are not fixed they are new random variables which are related to the type of source distribution and sample size. For example, in property insurance a rare but probable event is the occurrence of insured event, which must be accompanied by payment of a large insurance premium.

Therefore, the problem of the machine-generated pseudorandom data analysis should be treated as a probability model, which is constructed using the extreme value theory. One of the key points related to the process of development of an adequate model is a reasonable choice of the method for estimating unknown model parameters. Very often the problem of evaluating the unknown model parameters could be resolved by maximum likelihood method and Bayesian approach. An advantage of the latter method is in its possibility for application to small data samples and for the cases with degenerate data [4, 5]. Today, maximum likelihood method is a popular and relatively universal approach, that could be applied for parameter estimation of wide class of linear and nonlinear models [6-8].

2. Materials and Methods

An objective of the research is in application of extreme value theory for analysis and estimation of unknown model parameters using generated pseudorandom data. The necessary experimental data has been obtained by the algorithm of random number generation. The following tasks should be completed for achieving the stated goal: to investigate the properties of the distributions and methods of evaluating unknown parameters of extreme values; to study the algorithms for generating random numbers; to develop an effective model for analyzing pseudo-random numbers and estimating unknown parameters of selected models; to provide examples of analysis the generated values using the methodology of extreme value and programming environment R.

2.1 Algorithm for Generating Random Numbers

Practically, in many cases generation of a sequence of random digits with Gaussian distribution is required for solving the prediction task and many others. The most frequently for generating pseudorandom sequences is used central limit theorem. According to this theorem distribution of a sum of N identically distributed random variables is converging to normal distribution law with $N \to \infty$. In view of the last statement, independence of the uniformly distributed sequence $\{x_n\}$ has been rearranged to the sequence of numbers with Gaussian distribution $\{y_n\}$ with the following expression:

$$y_n = \frac{1}{N}\sum_{i=0}^{N-1} x(n \cdot N - i) \qquad (1)$$

Here the parameter N should be fairly large number [12].

As a variation of this method is Reider suggestion, where L of independent uniformly distributed sequences is modified to the new sequence of L non-correlated random variables with Gaussian distribution using Hadamard matrix. Each of the L Gaussian variables is obtained by adding (subtraction) of L numbers with uniformly probability distribution. The case when $L > 16$ is evidenced with the best approximation of normal distribution. This approach is effective because from L uniformly distributed variables we obtain L normally distributed random variables, but not L / N as after using equation (1).

An additional point is that there is a direct method of transformation the pair of uniformly distributed random variable to pair of pseudorandom variable with

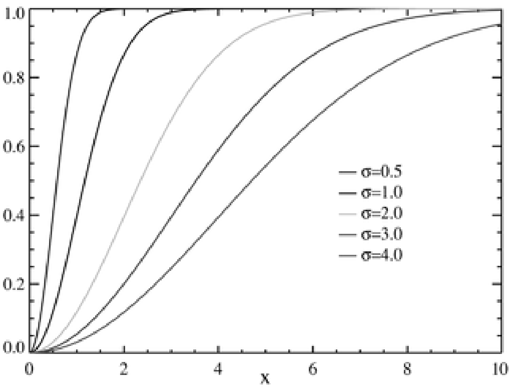


Fig. 1 Raleigh distribution function.

required distribution. Suppose that $\{x_n\}$ is uniformly distributed on interval $(0;1)$ sequence of random values and $\{y_n\}$ is defined as:

$$y_n = \sqrt{2\sigma^2 \ln\left[\frac{1}{x_n}\right]}. \tag{2}$$

Then in this case $\{y_n\}$ will have Raleigh distribution, i.e.:

$$P_y(y_0) = \frac{y_0}{\sigma^2}\exp\left(\frac{-y_0^{\ 2}}{2\sigma^2}\right). \tag{3}$$

Figure 1 illustrates graphical view of Raleigh distribution.

Now let's generate two new random variables $\{w_n\}$ and $\{w_{n+1}\}$ using the following equations:

$$w_n = y_n \cos[2\pi x_{n+1}] \tag{4}$$

$$w_{n+1} = y_n \sin[2\pi x_{n+1}] \tag{5}$$

These variables will be normally distributed with zero mean and variance equal to σ^2. It should be noticed that $\{w_n\}$ and $\{w_{n+1}\}$ should be non-correlated variables. It is equivalent to independence for normally distributed random variables [12].

So, in practice, the method discussed allows receiving efficient results, but it needs some additional calculating for the logarithms, sine, and cosine [12].

2.2 Model of Extreme Values Processing

The mathematical model of extreme values processing is presented [1]:

$$M_n = \max\{X_1,...X_n\}, \tag{6}$$

where: $X_1,...,X_n$ is a sequence of independent random variables having a common distribution function F.

In equation (6) the value of M_n represents the maximum of the process over n time units of observation. The distribution of M_n can be derived exactly for all values of n [1]:

$$\begin{aligned}\Pr\{M_n \le z\} &= \Pr\{X_1 \le z,...,X_n \le z\} = \\ &= \Pr\{X_1 \le z\}\times...\times\Pr\{X_n \le z\} = \{F(z)\}^n\end{aligned} \tag{7}$$

However, the function F is unknown. One possibility is to use standard statistical technique to estimate F based on observed data and then to substitute this estimate into (7). Unfortunately, very small discrepancies in the estimate of F can lead to substantial discrepancies for F^n. An alternative approach is to accept that F is unknown and to look for approximate families of models for F^n, which can be estimated on the basis of the extreme data only. This is similar to the usual practice of approximating the distribution of sample means by the normal distribution as justified by the central limit theorem.

If there exists the sequence of constants $\{a_n > 0\}$ and $\{b_n > 0\}$ such that

$$\Pr\left\{\frac{M_n - b_n}{a_n} \le z\right\} \to F(a_n x + b_n x)^n \to G(z), \tag{8}$$

where: G is a non-degenerate distribution function, then G belongs to one of the following families:

Gumbel distribution:

$$G(z) = \exp\left\{-\exp\left(-\left(\frac{z-b}{a}\right)\right)\right\}, \quad -\infty < z < \infty;$$

Frechet distribution:

$$G(z) = \begin{cases} 0, \ z \le b; \\ \exp\left(-\left(\frac{z-b}{a}\right)^{-\xi}\right), & z > b; \end{cases}$$

Weibull distribution:

$$G(z)=\begin{cases}\exp\left(-\left(-\left(\dfrac{z-b}{a}\right)\right)^{\xi}\right), & z<b;\\ 1, \ z\geq b\end{cases}$$

for $a>0,\ b-$ real number. In case of families 3 and 4, the parameter $\alpha>0$.

These three families of distributions could be combined into a single family of models having distribution function of the form:

$$G(z)=\exp\left\{-\left[1+\xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi}\right\}, \qquad (9)$$

This is the generalized extreme value (GEV) family of distributions. The model has three parameters: $\mu-$ location parameter; $\sigma-$ scale parameter; $\xi-$ shape parameter [2]. The distribution functions for each one from the GEV families are presented on the Figure 3.

Figure 2 shows that the three types of distributions have different forms of tail behavior. Weibull distribution has infinite end point $z_{\sup}=\dfrac{\mu-\sigma}{\xi}$, but for Frechet and Gumbel distribution $z_{\sup}=\infty$. However, Gumbel density function is damping out exponentially, whereas Frechet density function is polynomial. Gumbel distribution is similarly to normal, log-normal, gamma-normal distibutions. Frechet distribution has heavy tail, that could be denoted as $E(X^r)=\infty$ for $r\geq\dfrac{1}{\xi}$ (it means that it has infinite variance if $\xi\geq 1/2$).

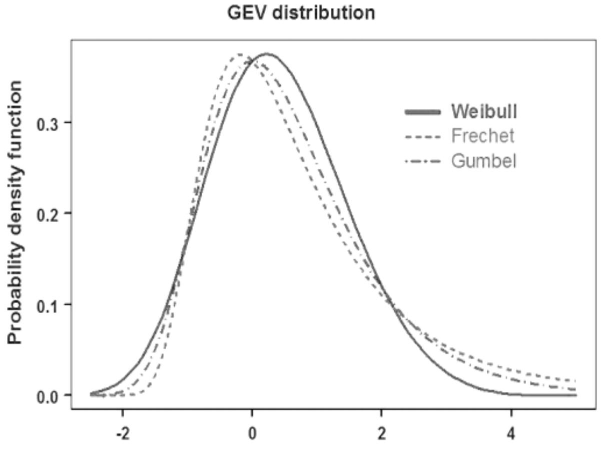


Fig. 2 Density functions for three types of distribution.

Generalized Pareto distribution (GPD) is identified to separate class, which can be derived from the next equation:

$$F_u(y)=\frac{F(u+y)-F(u)}{1-F(u)}, \qquad (10)$$

where: $u\rightarrow w_F=\sup\{x:F(x)<1\}$.

This parameter can be found by calculating the limit: $F_u(y)\approx G(y,\sigma_u,\xi)$, where: G is GPD. It is equivalent to

$$G(y,\sigma,\xi)=1-\left(1+\xi\frac{y}{\sigma}\right)_+^{-1/\xi}, \qquad (11)$$

if $\xi>0$ then heavy tail appears $x^{-1/\xi}$, which is equivalent to Pareto distribution;

if $\xi=0$ and $\xi\rightarrow 0,$ then finally $G(y,\sigma,0)=1-\exp\left(-\dfrac{y}{\sigma}\right)$. It's exponential distribution with mean σ;

if $\xi<0$ that finite upper point locates at the level of $-\dfrac{\sigma}{\xi}$.

The advantage of GEV is invariance of each distribution form which belongs to this class of distributions.

Consider briefly the method of extreme value processing. For this purpose consider statistical sample with n independent identically distributed values $X_1,...,X_n$. The following steps should be applied for processing extreme values.

(1) Grouping the data into sequences of n observations. The data sample should include from 50 to 100 values.

(2) The maxima Z_i of each block i should be calculated.

(3) Finally, the GEV-distribution will be fitted to this series of block maxima $Z_1,Z_2,...,Z_n$.

Practically, the length of blocks are assumed as one year sample or as the annual maxima Z_i of year i.

When GEV distribution has been fitted then it is possible to calculate the quantile function z_p for the annual maximum distribution as [3, 4]:

$$z_p = \begin{cases} \mu - (\sigma/\xi)\left(1-(-\log(1-p))^{-\xi}\right), & \xi \neq 0; \\ \mu - \sigma\log(-\log(1-p)), & \xi = 0; \end{cases} \tag{12}$$

The quantile function will be changed if accept the next suggestion $y_p = -\log(1-p)$:

$$z_p = \begin{cases} \mu - (\sigma/\xi)\left(1-(y_p)^{-\xi}\right), & \xi \neq 0; \\ \mu - \sigma\log(y_p), & \xi = 0; \end{cases} \tag{13}$$

This function can be plotted against $\log(y_p)$, but the plot will be linear in the case when $\xi = 0$; the plot will be convex in case of $\xi < 0$ with asymptotic limit equal to $(\mu-\sigma)/\xi$ as $p \to 0$; finally, the plot will be concave for $\xi > 0$ and will not have finite bound. This graph is named a return level plot and it is usually used as validation tool and as a way for presenting the best fitted model [3].

(4) Estimation of model parameters and choice of optimum length of blocks.

The task of choice of the optimum length of blocks implies a tradeoff between bias and variance. In the case when the length of the blocks is small, then the approximation of the distribution by the limit value is quite poor. So, this is leading to the bias in estimation and extrapolation. On the other side, long blocks will generate only a few data leading to large estimated variance.

The likelihood method is the most commonly used for estimation of model parameters. There is one difficulty with this approach, which means that regularity conditions for its application are not satisfied by the GEV distribution. That’s because the end-point of the distribution depends on the parameter values. This violation means that the standard asymptotic likelihood results are not automatically applicable. Smith studied the above problem in detail. As a result he found the following [3]:

- when $\xi > -0{,}5$ the maximum likelihood estimators have usual asymptotic properties;
- when $-1 < \xi < 0{,}5$ the maximum likelihood estimators can be obtained in general but they do not have the standard asymptotic properties;
- when $\xi < -1$ the maximum likelihood estimators are unlikely to be obtainable.

Observe that the case of $\xi < -0{,}5$ corresponds to distributions with a very short bounded upper tail, which is rarely the case in real applications of extreme value modeling [5].

The log-likelihood for the GEV distribution, when $\xi \neq 0$, should be defined as:

$$\begin{aligned} l(\mu,\sigma,\xi) &= -m\log\sigma - \\ &-(1+1/\xi)\sum_{i=1}^{m}\log\left(1+\xi\frac{z_i-\mu}{\sigma}\right) - \\ &-\sum_{i=1}^{m}\log\left(1+\xi\frac{z_i-\mu}{\sigma}\right)^{-1/\xi}, \end{aligned} \tag{14}$$

provided that $\left(1+\xi\frac{z_i-\mu}{\sigma}\right) > 0$ for $i = 1,\ldots,m.$

When this condition is not satisfied then the likelihood is zero and the log-likelihood is minus infinity.

In the Gumbel case ($\xi = 0$), the log-likelihood is as follows

$$\begin{aligned} l(\mu,\sigma) &= -m\log\sigma - \sum_{i=1}^{m}\left(\frac{z_i-\mu}{\sigma}\right) - \\ &-\sum_{i=1}^{m}\left(-\frac{z_i-\mu}{\sigma}\right) \end{aligned} \tag{15}$$

By maximizing these log-likelihood functions we obtain the maximum likelihood estimates $\left(\hat{\mu},\hat{\sigma},\hat{\xi}\right)$. The optimization is made using numerical optimization algorithms [3, 5].

(5) Graphical GEV-model checking.

It is impossible to check the validity of an extrapolation based on the GEV model, and assessment can be done with reference to the observed data. There are the following types of plots that could be used for graphical checking.

Probability plot is a comparison of empirical and fitted distribution functions. The empirical distribution function evaluated in the i th ordered block maximum, Z_i , is $\widetilde{G}_i(Z_i) = i/(m+1)$ and the fitted distribution function in the same point is

$$\hat{G}(Z_i) = \exp\left\{-\left(1+\hat{\xi}\left(\frac{z_{(i)}-\hat{\mu}}{\hat{\sigma}}\right)\right)^{-1/\hat{\xi}}\right\}. \quad (16)$$

In order to get a good model it is necessary that $\widetilde{G}(Z_i) = \hat{G}(Z_i)$. In practice the plot of points $\left\{\widetilde{G}(Z_i),\ \hat{G}(Z_i)\right\},\ i = 1,...,m$ – should lie close to the first diagonal. But because both functions are bounded to approach 1 as the values of z increase, the plot is the least informative in this region. The following graph avoids this deficiency.

Q-Q plot is a probability plot, which is applied for comparing two probability distributions by plotting their quantiles against each other. Additionally it helps to compare the shapes of distributions, providing a graphical view of how properties of model (e.g. location, scale, and skewness) are similar or different in two observed distributions. The approximation by a normal distribution remains a basic assumption in most of the Value-at-Risk (VaR) techniques. However, the most financial and actuarial series are fat-tailed. The graph of the quantiles makes it possible to assess the goodness of the fit of a series to the parametric model.

The graph of quantiles (Q-Q plots) is a representation of a set of points and is defined by this set of points [4]:

$$\left\{X_{k,n},\ F^{-1}\left(\frac{n-k+1}{n}\right),\ k = 1,...,n\right\} \quad (17)$$

The graph will have a linear form in the case when the parametric model fits the data well. Thus, the graph makes it possible to compare various estimated models and choose the best one; to assess how well the selected model fits the tail of the assumed empirical distribution. For example, if the series is approximated by a normal distribution and if the empirical data are fat-tailed, the graph will show a curve to the top at the right end or to the bottom at the left end. Besides the plots mentioned above, there are also exist return level plots, and mean excess function [3, 4].

The return level plot represents the set of points $(\log y_p, \hat{z}_p),\ 0 < p < 1.$ The confidence intervals are usually added to this plot to increase content of its information. The most important side of return periods in actuarial analysis is due to the fact that the return period could be used as a design assumption.

The mean excess function is a graphical tool which is widely used in studying of risk, insurance data and extreme values. The definition for it is as follows: suppose that X is a random variable and given threshold x_F , then [4]:

$$e(u) = E(X-u \mid X>u),\ where\ 0 \le u \le x_F; \quad (18)$$

where: $e(\cdot)$ is called the mean excess function; $e(u)$ is the mean excess over the threshold u.

If X follows an exponential distribution with a parameter λ, the function is equal to $e(u) = \lambda^{-1}$ for any $u > 0$. For the GPD we have

$$e(u) = \frac{\beta+\xi u}{1-\xi},\ \ where\ (\beta+\xi u) > 0. \quad (19)$$

The mean excess function for a fat-tailed series is located between the constant mean excess function of an exponential distribution $e(u) = \lambda^{-1}$ and GPD, which is linear and tends towards infinity for high thresholds as u tends towards infinity (Embrechts, Kluppelberg and Mikosch) [4].

(6) The choice of the threshold.

The threshold models use to provide effective result of fitting model to GEV-distributions. Let’s suggest that data exceeds the threshold level u; stochastic behavior of these values over u should be calculated and analyzed; $X_1,...,X_n$ is a sequence of independent and identically distributed random variables, having distribution F. Then conditional probability could be defined as follows:

$$F_u(y) = P(X \le u+y \mid X>u),$$

or

$$F_u(y) = \frac{F(u+y)-F(u)}{1-F(u)}. \quad (20)$$

The following result gives an approximation to this probability for high values of the threshold u.

The issue of how to choose the threshold is similar to that of selecting the size of block of maximums in the sense that both imply a balance between the bias and variance. A low level leads to failure in the asymptotic approximation of the model and a high level provides for few observations and then high variance.

A method of choice the threshold is based on the mean of the GPD: if Y is a random variable following GPD with required parameters σ and ξ, then mean value $E(Y) = \sigma/(1-\xi), \xi < 1$. Otherwise the mean will have infinite value.

If the model is valid for a threshold u_0 then it's also valid for all thresholds u that are greater than u_0. It means in both cases that [5]:

$$\begin{aligned} &e(u_0) = E(X - u_0 / X > u_0) = \widetilde{\sigma}_{u_0} /(1-\xi); \\ &e(u) = E(X - u / X > u) = \widetilde{\sigma}_u /(1-\xi) = \\ &= (\widetilde{\sigma}_{u_0} + \xi(u - u_0))/(1-\xi); \end{aligned} \tag{21}$$

Thus, $e(u) = E(X - u / X > u)$ is a linear function of u. Based on equation (21), the procedure to estimate the threshold is as follows [3, 10]:

- construct the mean residual life plot, by representing the points

$$\left(u, \sum_{i=1}^{n_u} (x_i - u) / n_u \right), \quad u < x_{\max}, \tag{22}$$

where: n_u is the number of observations exceeding u, and $x_{\max}$ is the maximum observation in the data set;

- choose as threshold the value above which the plot is approximately linear in u. The representation of confidence intervals can help to determine this point.

As an approach for choosing a threshold the conditionally acceptable method is used. This method is based on the following rule: the threshold will be set in the region which tail equals to 5-10% against all samples. The main assumption is that it should not include more than 10-15% of values. For example, Rocco (2011), McNeil and Frey (2000) used the percentage of a tail equal to 10% [10].

(7) Parameter estimation.

The most commonly used method to estimate the parameters is the maximum likelihood method. Having determined a threshold, the parameters of the generalized Pareto distribution can be estimated by maximum likelihood. Assume that $y_1,..., y_k$ are the k excesses of a threshold u. For $\xi \neq 0$ the log-likelihood is derived from the following equation [1]:

$$\begin{aligned} &l(\sigma, \xi) = -k \log \sigma - \\ &- (1 + 1/\xi) \sum_{i=1}^{k} \log(1 + \xi \; y_i / \sigma), \end{aligned} \tag{23}$$

provided $(1 + \xi \; y_i / \sigma) > 0$ for $i = 1,..., k$; otherwise, $l(\sigma, \xi) = -\infty$.

In the case when $\xi = 0$ the log-likelihood is defined as [1]:

$$l(\sigma) = -k(\log \sigma - \sigma^{-1} \sum_{i=1}^{k} y_i). \tag{24}$$

Another method of parameters estimation is Bayesian approach. Contrary to the maximum likelihood method it has an advantage that it is independent on regularity assumptions regarding initial distribution. The Bayesian approach was successfully applied for estimating unknown parameters of generalized linear models [9, 10].

3. Results and Discussion

3.1 Results

An experimental investigation of the efficiency of the proposed method was performed using machine random number generation according to the reviewed algorithm. Dimensionality of the sample was 250 points, which includes the following input variables: location parameter, scale parameter, shape parameter, dimension of the sample. Two samples with the same dimensions were examined. Well-reasoned parameter mismatch is provided in connection with the last results from our previous research [13].

The R package was used of 2.9.2 version as a programming language and the software environmental for statistical computing and graphics, implementation the algorithm of random number generation, presumptive analysis of data etc. In this environment was used an alliance of Rcmdr, extRemes, evdBayes, mcmPack modules and ‘Simulate Data’ function from extRemes module.

Figure 3 depicts the result of random number generation with the input data. The presumptive analysis of received sample shows that array of data is degenerate and the values are erratic. There is no convergence of series to earlier defined interval.

As a result of analysis the descriptive statistics give a chance to assume that input data could be approximated to GEV- or GPD- distribution (Figure 4).

The graphs in Figure 5 illustrate estimated GEV-model such as probability and quantile plots, a return-level plot, and a density estimate plot. In the case of perfect fit the data would line up on the diagonal of the probability and quantile plots. Briefly, the quantile plot compares the model quantiles against the data (empirial) quantiles. A quantile plot which deviates greatly from a straight line suggests that the model assumptions may be invalid for the data plotted. The return level plot shows the return period against the return level, and shows an estimated 95% confidence interval. The return level is the level that is expected to be exceeded, on average, once every m time points (in this case conceptional years). The return period is the amount of time expected for waiting for the exceeding of a particular return level. The return level period is the period of time expected for waiting for the exceeding of a particular return level. For example, in Figure 6 one would expect the maximum value of sample to exceed about 10 points on average every 100 times.

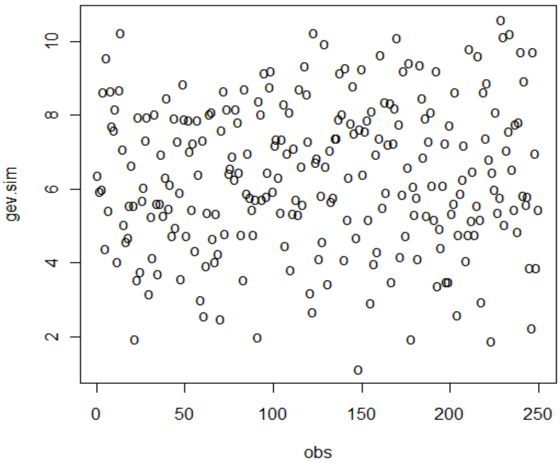
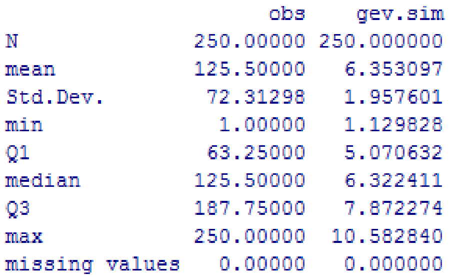


Fig. 3 The results of machine random number generation with dimension of sample 250 points and the next input data: $\mu = 5.86;$ $\sigma = 1.97;$ $\xi = -0.36;$



```
                         obs    gev.sim
N                  250.00000 250.000000
mean               125.50000   6.353097
Std.Dev.            72.31298   1.957601
min                  1.00000   1.129828
Q1                  63.25000   5.070632
median             125.50000   6.322411
Q3                 187.75000   7.872274
max                250.00000  10.582840
missing values       0.00000   0.000000
```

Fig. 4 Descriptive statistics of already generated sample.

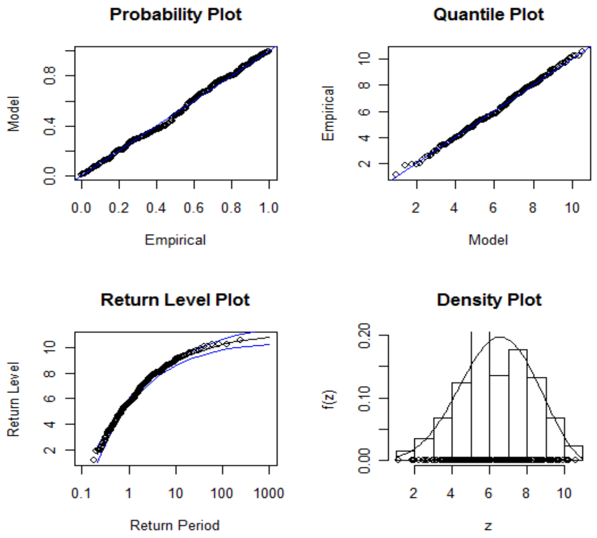
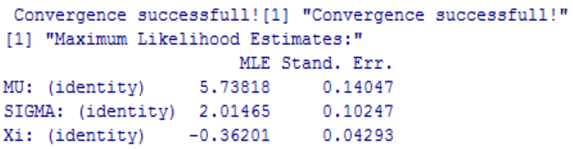


Fig. 5 GEV-fit diagnostics for the machine random-number generated dataset in the first experiment.

Figure 6 shows results obtained for parameter estimation of constructed model using maximum likelihood technique.



```
 Convergence successfull![1] "Convergence successfull!"
[1] "Maximum Likelihood Estimates:"
                         MLE Stand. Err.
MU: (identity)       5.73818     0.14047
SIGMA: (identity)  2.01465     0.10247
Xi: (identity)      -0.36201     0.04293
```

Fig. 6 Results of parameter estimation for GEV-model.

Table 1 Comparative analysis of parameters for GEV-distribution.

			σ - scale parameter		ξ - shape parameter		μ - location parameter		
N	Results	Distribution	Maximum likelihood estimation	Std. Error	Maximum likelihood estimation	Std. Error	Maximum likelihood estimation	Std. Error	Log-likelihood
1.	Empirical	GEV-distribution	1,953	0,712	-0,650	0,095	5,738	0,140	518,829
	Theoretical		1,97		-0,36		$\mu = 5{,}86$		-
2.	Empirical	GEV- distribution	0,955	0,054	0,209	0,051	0,093	0,068	413,192
	Theoretical		1		0,2		$\mu = 0$		-

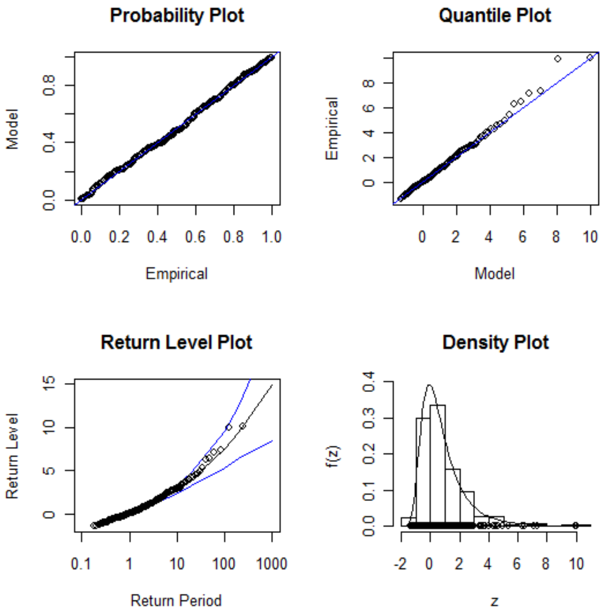


Fig. 7 GEV-fit diagnostics for the random number generated dataset in the second experiment.

A comparative analysis of parameters for GEV-distribution on the basis of two generated samples is given in the Table 1. It should be noticed that the more accurate is the set of the scale and shape parameters, with selected zero as the location parameter, the better will be approximation of the practical to the theoretical plot (Figure 7).

So, as a result of comparing the plots of density distribution for the models constructed (Figure 5 and Figure 7) it should be emphasized that the best is GEV model with zero values of location parameter.

In view of the last result the minimum deviation between experimental and theoretical parameters was presented in Table 1. This is an evidence for high accuracy of random number generation algorithm and correct choice of initial parameters.

3.1 Discussions

Thus, application of the proposed model for processing extreme statistical samples helps to successfully solve the problem of degenerated data using the of extreme value theory and the random number generation algorithm.

To estimate unknown parameters of constructed models which belong to the class of GEV-distribution maximum likelihood method was used. Thus, the theory of extreme data is a powerful tool for working with actual statistical data and machine generated samples of random numbers. The application of the new combined methods for the problem of extreme data processing and unknown parameters estimation, the choice of the best model based on approach of input parameters selection, and studying the efficient algorithms for pseudorandom data generation opens up the new directions of their use in mathematical modeling of complex systems.

4. Conclusions

The study directed to finding of an effective algorithm for pseudorandom data generation and the technique for extreme values processing was performed. It was proposed and experimentally justified the efficiency of established multistep approach based on the extreme value theory and random number generation algorithm. The example considered shows that the proposed comprehensive approach for the extreme values processing is an effective and convenient tool for analysis of the degenerated samples in actual statistical data and the newly machine-generated.

The involvement of the new combined methods which will handle poorly structured, degenerated samples opens up the new fields for their use in various applied mathematics studies. In the future it is necessary to study the possibility of using the results of extreme values processing for analysis of predictive generalized linear models with different data origins (economic, engineering systems and machine-generated samples). Application of the proposed procedures for extreme values processing ensures highly accurate approximation of a sample to the class of defined distributions and avoidance of statistical noise. The comparative analysis of parameters estimation for models with different input values showed that more accurate choice of the shape and scale parameters will provide faster convergence of the series. Also, the field of mathematical modeling and forecasting processes in conditions of application of effective mathematical methods and algorithms for data processing, and parameter estimation, the forecasts generated could be a worthy argument for economic stabilization as whole.

Acknowledgment

The research has been conducted in the frames of the grant of Ukrainian Ministry for Education № 2813p.: “Development of systemic methodology for modeling and estimating of financial risks (2015 – 2016)”.

References

[1] S. Coles, Introduction to Statistical Modeling of Extreme Values, Springer-Verlag, London, 2001, pp. 45-73.

[2] R.L. Smith, An overview of Extreme value theory, Bernoulli Center, Lausanne, 2009, pp. 7-28.

[3] F. Mallor, E. Nualart, E. Omey, An introduction to statistical modeling of extreme values, Hub research paper, 36 (2009) 5-31.

[4] R.H. Shumway, D.S. Stoffer, R.H. Shumway, Time series analysis and its applications, Springer, New York, 2006, pp. 47-78.

[5] A. Romano, G. Secundo, Dynamic learning methods, Springer, New York, 2009, pp. 32-45.

[6] P. Mccullagh, J. Nelder, Generalized Linear Models, Chapman & Hall, New York, 1989, pp. 21-44.

[7] R.S. Tsay, Analysis of financial time series, John Wiley & Sons, New Jersey, 2010, pp. 325-356.

[8] J. Besag, Markov Chain Monte Carlo for Statistical Inference, Center for Statistics and the Social Sciences, 9 (2001) 24-25.

[9] P.I. Bidyuk, S.V. Trukhan, Estimation of generalized linear models using Bayesian approach in actuarial modeling, Naukovi Visti NTUU ‘KPI’, 6 (2014), 49-55.

[10] J. Beirlant, Statistics of extremes: Theory and application, John Wiley & Sons, New York, 2004, pp. 3-123.

[11] D.W. Kelton, J.S. Smith, D.T. Sturrock, Simio and simulation: Modeling, Analysis, Applications, Simio LLC, Cincinnati, 2013, pp. 6-69.

[12] L.R. Rabiner, B. Gold, Theory and application of digital signal processing, Englewood Cliffs, New Jersey, 1975, pp. 42-78.

[13] P.I. Bidyuk, S.V. Trukhan, Methodology of extreme values analysis and its application for parameter estimation Generalized Linear Models, “Radio Electronics, Computer Science, Control”, Zaporizhzhya National Technical University, 1 (2016) – accepted and will be published.