e-ISSN 1607-3274 Радіоелектроніка, інформатика, управління. 2020. № 2
p-ISSN 2313-688X Radio Electronics, Computer Science, Control. 2020. № 2

UDC 004.94:658

RISKS ESTIMATION METHOD BY CLUSTERED EXTREME DATA OF PROCESS COVARIATES

Tereshchenko I. V. – PhD, Associate Professor of the Department of Info-Communication Engineering, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine.

Tereshchenko A. I. – PhD, Assistant of the Department of Information and Cyber Security Management, State University of Telecommunications, Kyiv, Ukraine.

Shtangey S. V. – PhD, Associate Professor of the Department of Info-Communication Engineering, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine.

ABSTRACT

Context. This paper presents a method for solving the problem of detecting and taking into account the influence of various (external and/or internal) factors on extreme and risky values of the multivariate observed parameters (covariates) of technological and/or diagnostic processes. Taking into account external and internal influence factors on covariates, by analogy with critical process parameters, is a significant addition to the extreme values statistics and the estimations the influence of the variability of process's covariates on the expected losses, i.e. value at risk. Risk-oriented analysis is an actual tool for the data behavior investigation of the multivariate observations of process's parameters.

Objective. To disclose a method for detecting and taking into account the factors influence on the distribution functions parameters of the observed extreme values of process's covariates and determine the influence of these distribution functions parameters on estimates of risks values.

Method. The method consistently uses: the procedures of multivariate statistical cluster analysis, transformation the matrix of observed extreme values of process's covariates into data frame with factor variables, estimation the extremal index and distribution functions parameters of nonclustered and clustered the observed extreme data of covariates and estimation the risk values on the calculated values of distribution functions parameters. The proposed sequence of actions is aimed at implementing the information technology of statistical causal analysis of the influence of factors on the variability of process's covariates and their risk values due to the application of the clustering procedure for observed multivariate extreme values of covariates. The method is implementing the R-language packages software.

Results. Clustering of the multivariate observed extreme values of process's covariates allows to identifying the influence of environmental (manufacturing) factors and estimates the covariates' risky values taking into account of this influence.

Conclusions. The method is an information technology of statistical causal analysis of factors influence on the variability of process's covariates and theirs risk values due to the application of the clustering procedure of covariates' multivariate values. The prospect of further research is to improve the methods of causal multivariate statistical analysis of the various factors influence on the exogenous and endogenous parameters of manufacturing and other processes in order to reduce the variability of these parameters and, as a result, minimize the risks.

KEYWORDS: extreme value theory, generalized extreme value distribution, generalized Pareto distributions, value at risk, extreme value index, cluster analysis, process approach

ABBREVIATIONS

AMS is a annual maximum series;
ACER is a average conditional exceedance rate;
BSS is a between sum of squares;
c.d.f. is a cumulative distribution function;
Cl.dist. is a clustered distance;
CPPs is a critical process parameters;
DC is a determination coefficient;
d.f. is a distribution function;
EVA is a extreme value analysis;
EVI is a extreme value index;
EVS is a extreme value statistics;
EVT is a extreme value theory;
GD is a Gaussian distribution;
GEVD is a generalized extreme value distribution;
GPD is a generalized Pareto distribution;
i.i.d. is a independent and identically distribution;
MLE is a maximum likelihood estimation;
PD is a Pareto distribution;
p.d.f. is a probability density function;
POT is a peak over threshold;
ProNEVA is a process-informed nonstationary EVA;
Q-Q plot is a quantile-quantile plot;
TSS is a total sum of squares;
UK is a United Kingdom;
VaR is a value at risk;
CVaR is a conditional value at risk;
VAT is a visual assessment of cluster tendency.

NOMENCLATURE

ξ is a extremal index or shape parameter (shape);
μ is a location parameter (location);
σ is a scale parameter (scale);
α is a tail index;
X is a $n \times m$ array of maximums statistical data;
X^j is a j-th m-dimension object of observations;
n is a n-th component of X (is a n-th observation);
m is a m-th component of X (is a m-th covariate);
x_{jk} – is a j-th value of the k-th covariate;
X_k is a vector of the n observations of the k-th covariate;
x_{nk} is the n-th observations of the k-th covariate;
X_q^v is a $q \times m$ sample from the m-dimensional general population of the v-th sign (factor);

DOI 10.15588/1607-3274-2020-2-6

x_{qk}^{v} is a vector of q observations of the k-th covariate for the v-th sign;

q is a quantity of observations in the cluster that corresponds to the v-th factor;

X_F is a data table (data frame) with factor variables;

L_k^v is a vector of extreme values of covariate k by the v-th factor;

$F_{\xi,\mu,\sigma}^{v}()$ is a generalized distribution function of Fisher-Tippett-Gnedenko extremes values;

$l()$ is a maximum likelihood estimation;

M_n^i is a maximum value of X_i in block $i = 1,...,m$;

W_{jk} is a j-th residual for each of k-th covariate;

$OpCVaR$ is a operational conditional value at risk;

$OpVaR$ is a operational value at risk;

$OpVaR_\beta$ is a maximum value of $OpVaR$ that will not be exceeded with probability β;

$OpCVaR_\beta$ is a maximum value of $OpCVaR$ that will not be exceeded with probability β;

$OpCVaR_\beta^v$ is a maximum value of $OpCVaR$ that will not be exceeded with probability β_v for a given factor v;

$OpVaR_\beta^v$ is a maximum value of $OpVaR$ that will not be exceeded with probability β_v for a given factor v;

β_v is a probability of exceeded the maximum value for a given factor v;

$N()$ is a Gaussian distribution function;

η is a Gaussian expected value;

σ_N is a Gaussian standard deviation;

$q_{0.99}$ is a 0.99-quantile of normal distribution;

$f()$ is a standard normal p.d.f.;

$\Phi()$ is a standard normal c.d.f.;

w is a quantile probability;

u is a threshold;

r is a quantity of threshold exceedances.

INTRODUCTION

The rapid change in market conditions and the increased hazard of crisis situations nowadays set the task of improving methods that allow to assessing the factors influence on the VaR of manufacture and diagnostic processes parameters, based on observed extreme data of this parameters (covariates).

Thus, **the object of study** is the risk estimation process using the EVS for factorized multivariate observations of covariates' values of technological and/or diagnostic processes.

The observed extreme data of process's covariates are the values which equal to, close to or exceed the limit values to acceptable for certain requirements (for example, the requirements of branch standards for product's quality, environmental standards, etc.). Therefore, mentally and functionally, investigations are based on the general concept of risk-oriented quality management ISO 9001: 2015/ISO 31000: 2018 and the methodology of the process approach [1–3].

At the instrumental level, the study used the EVT tools, which is an important part of statistical science for the development of information technologies practical applications [4–6]. So in particular, EVT offers a mathematical apparatus for estimating $OpVaR$ and $OpCVaR$, as well as predicting the results of extreme, rare, but dangerous consequences of events [5].

The main cause of the covariates' risky values is the various factors influence on manufacture/diagnostic processes [3], as shown in Fig. 1.

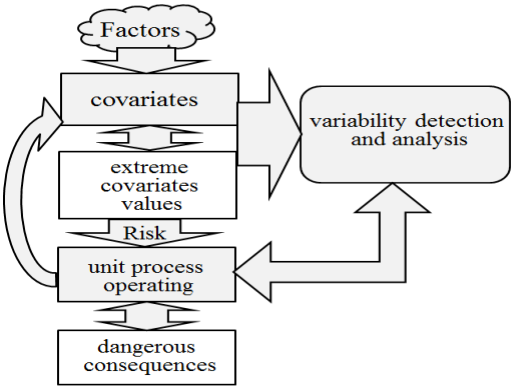

Figure 1 – The structure of the formation the extreme and risk values of process's covariates

The factors influence is showed in the variability of the process's exogenous parameters values (covariates) and/or the reaching/exceedance the permissible thresholds of these covariates.

Therefore, **the subject of study** are methods for detecting and estimating the various factors influences on VaR based on variability analysis of multivariate observations of covariates' extremes values of technological and/or diagnostic processes.

The goal of the paper is a developing and investigation the method of multivariate statistical causal analysis of the factors influences on variability of multivariate observations of process's covariates extremes and their values at risks.

The method used the clustering procedure of multivariate observations of covariates' extremes values and the EVT tools for estimate of the distribution functions parameters of covariates.

According to EVT and well-known approaches [7–14], risks are estimated by the parameters of the generalized extreme values distributions family – GEVD: ξ – EVI, μ and σ ($\sigma > 0$). The GEVD family includes the Gumbel, Frechet and Weibull distributions. After the corresponding normalization of the values and according to the Fisher-Tippett-Gnedenko theorem, the distributions of extremes' samples for the i.i.d. random variables (i.e.

DOI 10.15588/1607-3274-2020-2-6

observations of the covariates) converge to these distributions.

For the risks values of covariates, estimating as a high quantiles (0.95 and higher), are used the generalized Pareto or Gaussian distributions as distribution functions for covariates values that exceed a given sufficiently large threshold [4, 5, 11].

One of the GEVD parameters is the EVI, which used for decompose the researched sequence (observation series) into detailed independent clusters in order to identify and evaluate the GEVD/GPD parameters for maximum values of such clusters from the initial observation series [7–9]. EVI is also used to classify d.f. by so-called "tail index" – $\alpha = 1/\xi$.

Note the importance of determining EVI, which is evaluated before declustering and supports the bootstrap procedure for assessing variability of estimates [15].

Next, the problem areas of the paper are formulated and attention is drawn to possible limitations.

Software implementations of approach on R language, described above, give adequate results for extremes from a some number of observations blocks for certain time intervals (i.e. AMS), as well as for extremes from a continuous sequence of observations values that exceed a certain threshold during any the observation interval (i.e. POT) [10, 14].

It is significant that the values in the observations blocks AMS and POT- values are data that structured exclusively by time. That is, the GEVD parameters for this approach mainly depend on the influence of the time factor in the behavior of an observed data.

This creates conflicting requirements when choosing the length of time intervals and the distribution of extremes along them when the block AMS method is implementing. A similar problem arises when implementing the POT method, when the threshold value will depend on the length of the time interval and the amount of observed data [14].

Also note that the estimation of the extreme index and the subsequent declustering of the observed sequence according to the time attribute (exceedance times) leads to a large number of clusters with a small number of objects (observations), which is a problem for interpretation the results of these studies.

For example, the "Daily BMW Stock Returns" data (bmwRet) [7], which are used for comparison, in the initial series of observations contains 6146 points (objects). Evaluation of the extremal index of observations series and subsequent declustering of 1180 maximums gives 867 clusters with the number of elements no more than four [9].

Fig. 2 shows the data of exceeding a certain subjective threshold, for example, the average value of quarterly observations that are not programmatically divided into clusters in Fig. 2, but for which, according to Fig. 2, can be visually determined the groups of extremes (gray color in Fig. 2.) in an amount much less than 867.

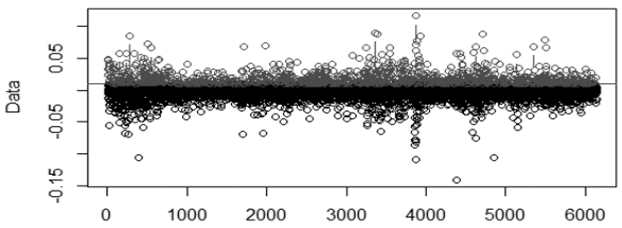

Figure 2 – Threshold (horizontal line) excess graph for bmwRet data

Obviously, the results [7–9] are difficult to interpretation. For solving the problem of more valid data grouping, it is advisable to consider shorter time intervals, for example, a month or a quarter, however, may be lost the monthly or quarterly dependencies , i.e. lost the trends for over the longer time periods.

Existing approaches to the classification of distribution functions (trends) are often used the limitation argumentation, when all components of a multivariate variable reach extremes with the same speed [16]. In this case, the investigations are synthetic because it is unlikely that extreme values of all variables (covariates) are observed simultaneously. Thus, in order to detect the homogeneity of multivariate observed objects of and the influence of external factors, it becomes necessary to group the objects of observations not only by the time factor, but also by some measure of distances in the multivariate area of independent signs i.e. clustering [17–19].

Taking into account external and internal influence factors on covariates, by analogy with CPPs [20, 21], is a significant addition to the analysis using EVS and the estimating the influence of process's covariates variability on the expected losses, i.e. VaR. By analogy, CPPs can be identified, for example, with the parameters of environmental, hydrological observations and etc.

The presented method sequentially implements the clustering procedures of extreme observed data, estimates the parameters of the covariates' d.f. for nonclustered and clustered data observations, and also estimates the impact of these parameters at the VaR.

Clustering allows to identify the factors of the manufacturing environment influence on the values of process's covariates and to estimate the VaR for covariates taking into account the influence of these factors.

Thus, the method eliminates the drawback of analysis the observed extreme values only by the time factor as an action that has already taken place and adding the influence of the manufacturing environment in the form of external and internal factors influences on the observed covariates values.

1 PROBLEM STATEMENT

Let the X is a $n \times m$ maximums (extremes) array of n statistical m-dimensional observations be known for process's m covariates (parameters, predictors, exogens).

DOI 10.15588/1607-3274-2020-2-6

$X = \left\{ X^j \right\}$, where $j = 1, 2, ..., n$, $X^j = \left\{ x_{j1}, ..., x_{jm} \right\}$ – rows of X array.

For the X array, it is necessary to conduct an exploratory data analysis that to detect the factors influences on the m-dimensional values of covariates' observations.

For the covariates array, taking into account the influence of external factors, to calculate the $OpVaR$ and $OpCVaR$ as the high quantiles of distribution functions of the observed extreme values of covariates, which will not be exceeded with some probability (for example, 0.95, 0.99).

To compare the results of calculating the $OpVaR$ and $OpCVaR$ for different factors that will be determined during the research.

Let assumed that the n observed values of each k-th covariate – X_k have asymptotically close to one of the GEVD [11]. $X_k = \left\{ x_{1k}, ..., x_{nk} \right\}$ – columns of X array, where $k = 1, 2, ..., m$.

2 REVIEW OF THE LITERATURE

Scientific and engineering-applied topics of publications, which are analyzed in a review [22], indicate that the EVT methodology is increasingly used by practitioners for various branches of research.

The requirements have increased for the accuracy of modeling and determining the parameters of extreme events (in particular, the index of extreme values) for the statistical control of product's quality, biostatistics, environmental monitoring etc. [23], where the extreme values of the observed parameters are indicators of risk situations. Thus, EVS requires tools that should be easy to use, but also must to implement the complex productive statistical models and provide an adequate interpretation of the results. This contradiction is necessitates the development of understandable methods for solving the problem of detecting and taking into account the influence of various factors on extreme and risk values of observed parameters of manufacturing and others processes.

Depending on the data view and its content and also following the deductive approach to research, it is necessary to take into account the method of extremes determine (AMS and/or POT) for further statistical analysis of observed extreme values of process's covariates [24]. The AMS and POT procedures are associated with the collection, sorting, and initial processing (exploratory analysis) of observed data that preceding the methods for improving the accuracy of VaR and EVI estimates [7, 25]. For example, for an adequate EVI estimate, a class of kernel estimates with a reduced bias and a parameterized tuning tool was proposed that allows changing the standard asymptotic error [26].

The paper [27] presents a ProNEVA, as a generalized tool for integrating a various types of physical factors (that is basic processes), stationary and nonstationary concepts and methods of extreme values analysis (i.e. AMS and POT).

For the analysis of nonstationary time series, the ACER method is also applicable [28]. The idea of the ACER is that a sequence of nonparametric distribution functions is constructing in order to approximate the distribution of the collected history of extreme values.

At statistical estimates the extremes of the multivariate data the problem of the dependence (in particular, correlation) of the components (covariates) of multivariate observations inevitably arises. For this currently relevant problem, the review [29] considers the most interesting examples of strongly correlated variables for which there are very few exact results of the EVS.

In paper [30] a test for detecting sequential correlations in multivariate time series was proposed. This test uses Spearman's rank correlation properties and extreme value theory.

For multivariate theories and assessment methods of extreme values is indicative the paper [16]. In this paper for characterize, evaluate and extrapolate the distribution of a multivariate random variable is developed a semi-parametric approach that overcomes the limitations to arguments when all components of the multivariate variable become large at the same rate.

Hazard assessment at a regional scale may be performed using a spatial model for maxima that can be obtained by combining the GEVD for the univariate marginal distributions with extreme-value copulas to describe their dependence structure, as justified by the theory of multivariate extreme values [31].

The paper [32] describes a multivariate statistical dependency model for hydrological observations, which used to estimate flood losses in a large and heterogeneous region.

Note that the sources [29–32] are closest to the subject of our paper, where the relationship of observed extreme values objects together with the influence of exogenous factors is investigated. The implementations of adequate approaches with examples of numerical methods for the corresponding sections of EVT are presented in [12].

Classification and reviews of risks calculation methods [33, 34] and methods for calculating the parameters of GEVD are described in [7, 13, 14]. Based on these works, as described below, generalizations were made for the main approaches to risks estimates (section 3). Software reviews [35, 36] for statistical analysis of extreme values are useful for research.

Review of the literature allows to conclusions:

– identification of factors (dependent excesses in observation groups), as a rule, occurs on background of independent clusters of the observed objects formation;

– as already mentioned, clustering the observations according to the time attribute [7, 8], evaluating the extreme index and subsequent declustering of dependent sequences using the method of estimating the time intervals of Ferro & Segers [7, 9, 15] is difficult for interpretation and, therefore, an expedient method is clustering the objects of observations by the ratios of a

DOI 10.15588/1607-3274-2020-2-6

certain measure of multivariate distances in the area of independent signs [17, 18];

– EVT applied research is evolving in the direction of increasing the accuracy and non-bias of EVI and VaR estimates, and they also offer options for solving the problem of evaluating the impact on EVI and VaR the time and/or covariate dependences of the observed values in multivariate applications.

It is also important that the statement of the problem and the method for solving it are performed from the perspective of an object-oriented analysis, which corresponds to modern tendencies of the process approach to support the product quality management system [37].

3 MATERIALS AND METHODS

Consider the content of the implementation stages for proposed method: the overall scheme is shown in Fig. 3.

At the first stage of the initial data exploratory analysis, the compatibility of the data order (equivalence class) and clustering trends are assessed. This helps to understand the general data background and the direction of further research.

At the second stage, clustering of random observed multivariate extreme values of covariates – X was carried out. Data clustering is seen as the result of the manufacturing environment factor impacted. This operation allows to get a certain number of m-dimensional observed objects that are combined into clusters. Moreover, it is clear which observations formed a particular cluster. It is significant that clustering allows identifying the influence factors that led to the grouping of m-dimensional objects (points).

For clustering operations, non-hierarchical object partitioning algorithms implemented in the R language were used and according to which the data of n observations were decomposed into s clusters with previously unknown factors/signs – v [20].

These algorithms allows to find the centroids, i.e. the centers of objects concentration which located as far as possible from each other and with the minimum scatter of objects within each cluster and also with the verification of clustering quality [19].

The set of clustered objects can be interpreted as a $q \times m$ sample – $X_q^v = \left\{ x_{qk}^v \right\}$ from the m-dimensional general population of the v-th sign ($v = 1, ..., s$, $q = 1, ..., n$), and each sign v can be identified with a specific cluster i.e. factor.

Clustering is characterized by the ratio of the intercluster (constrained) variance of observed objects scattering (i.e. BSS) to the total variance of observed objects scattering (i.e. TSS) [12]: BSS/TSS. This ratio is known as the DC.

The result of the second stage is the transformation of a homogeneous data array X into a data table (data frame) with factor variables – X_F, as shown in Fig.3.

At the third stage, from the X_F data frame for each v-th cluster, the number of q values is selected for each k-th covariate – $\left\{ x_{qk}^v \right\}$, which is interpreted as a vector of extreme values of covariate k by the v-th sign (factor):

$$L_k^v = \max\{x_{qk}^v\} . \quad (1)$$

This operation is also justified due to the fact that the values of the covariates observations do not achieve extremes simultaneously.

This fact is taken into account in the d.f. of the sample for each k-th covariate of each v-th cluster. For interpretation, it is important that the various extremes of each k-th covariate will be related with the various v-th factors.

It is proofed that the quantity L_k^v has a d.f. close to the generalized distribution function of extreme values of Fisher-Tippett-Gnedenko [6, 12]. This generalized function has the view:

$$F_{\xi,\mu,\sigma}^v(x) = \begin{cases} \exp\left[-\left(1 + \dfrac{\xi\left(x-\mu\right)^{-1/\xi}}{\sigma} \right) \right], & if\ \xi \neq 0, \\ \exp\left[\exp\left(-\dfrac{x-\mu}{\sigma} \right) \right], & if\ \xi = 0, \end{cases} \quad (2)$$

where $x = x_{qk}^v$ and ξ, μ, σ – GEVD parameters.

The result of the third stage is the decomposition of the clustered multivariate values of the process's covariates – $X = \{X^j\}$ i.e. formation the samples of q clustered data for each k-th parameter $\left\{ x_{qk}^v \right\}$, as shown in Fig. 3.

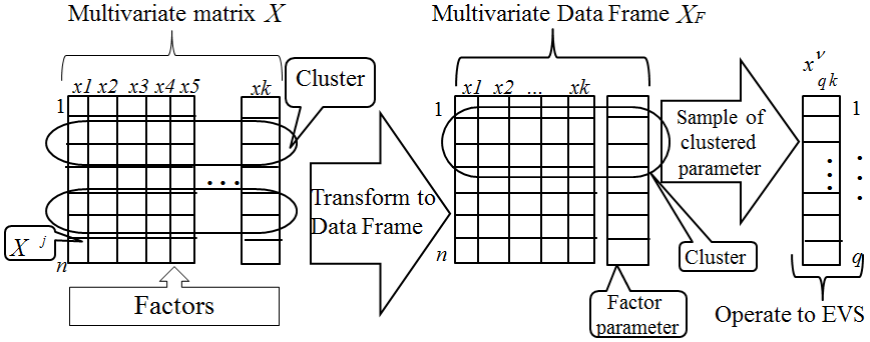

Figure 3 – The overall scheme of method implementation

DOI 10.15588/1607-3274-2020-2-6

At the fourth stage, for each *k*-th covariate of nonclustered and clustered multivariate data, EVI is estimated by the Ferro & Segers method with further calculation of the GEVD/GPD and Gaussian d.f. parameters using the MLE.

Note, that for further risk analysis EVI estimating is a great importance. In paper, EVI values were estimated using the Ferro & Segers and Heffernan methods for dependent extremes [15, 16]. According to these methods, EVI value is determined for covariate, above which EVI is stable for higher values of covariates [13, 14].

It is advisable to choose these covariates values as the values of "high enough threshold" *u*. Then, the GEV/GPD and Gaussian d.f. parameters can be estimated using method of moments, method of probability-weighted moments [12, 38] or using the MLE [12–14].

MLE offers the advantage of simultaneous estimation of the three parameters (ξ, μ, σ), and it applies well to the series of maxima per block. Also, MLE gives valid estimates for the case $\xi > 1/2$ [13, 15, 16].

The log likelihood function assuming i.i.d. observations from the GEV d.f. with $\xi \neq 0$ is [13]:

$$l(\xi,\mu,\sigma) = -m\cdot\ln(\sigma) - (1+1/\xi)\cdot\sum_{i=1}^{m}\ln\left[1+\xi\cdot\left(\frac{M_n^i-\mu}{\sigma}\right)\right] - \\ -\sum_{i=1}^{m}\left[1+\xi\cdot\left(\frac{M_n^i-\mu}{\sigma}\right)\right]^{-1/\xi}, \quad (3)$$

where $1+\xi\cdot(\frac{M_n^i-\mu}{\sigma})>0$, M_n^i is defined as the maximum value of X_i in block $i=1,...,m$. In our case, covariates are represented by single-block observations.

Studies shows that Ferro & Segers and Heffernan EVI estimates for dependent excesses are consistent with the EVI estimates by MLE for the GEVD/GPD. The parameters ξ, σ for the Pareto distribution were calculated using the MLE.

As already said: for high threshold values that are given by high quantiles (0.95 and higher), the distribution of random extremes values is close to the generalized Gaussian or Pareto distributions [13, 14]. The data distribution functions were checked for compliance with Pareto/Gaussian distributions by the graphical method using Q-Q plot. In Q-Q plot diagrams the standardized residuals are compared with the values expected from the reference distribution.

In particular for GEVD, the graphs show aspects of the so-called "crude residuals" [13]:

$$W_{jk} = \left(1+\hat{\xi}\cdot\frac{x_{jk}-\hat{\mu}}{\hat{\sigma}}\right)^{-1/\hat{\xi}}. \quad (4)$$

Formula (4) uses MLEs of ξ, μ, σ – $\hat{\xi}, \hat{\mu}, \hat{\sigma}$.

If the assumption of the proposed distribution is fulfilled, then the points on the Q-Q diagram should lie close to a straight line with an inclined angle of 45°.

The results of the fourth stage are EVI estimates for non-clustered and clustered multivariate covariates values and the calculated parameters of the GEVD/GPD and Gaussian distributions of covariates by using the MLE.

At the fifth stage, taking into account the obtained EVI estimates for Pereto and the Gaussian d.f. of the nonclustered and clustered data observations of covariates, the values $OpVaR_\beta$ and $OpCVaR_\beta$ are determined.

For the maximum values of the *k*-th covariate clustered observations the measure of operational risk $OpVaR_\beta$ is interpreted as the maximum value that will not be exceeded with probability β_v for a given factor *v* – $OpVaR_\beta^v$. For many applications, measure $OpCVaR_\beta^v$ is relevant i.e. the expected value at risk, provided that it exceeds the value of $OpVaR_\beta^v$. General approaches to the calculation of $OpVaR_\beta$ and $OpCVaR_\beta$ are follows:

– to use the analytical formulas for calculating $OpVaR_\beta$ and $OpCVaR_\beta$, when the GEVD parameters ξ, μ, σ for the of the covariates are calculated by MLE, method of moments, or method of probability-weighted moments, taking into account the limitations on EVI (for example $0 < \xi < 1$) [5, 10, 12];

– to use the numerical computer methods for estimating the parameters ξ, μ, σ, $OpVaR_\beta$ and $OpCVaR_\beta$ taking into account the GEVD, Gaussian and/or Pareto distributions for the random variable *X* (covariates array) [13];

The second approach was used for research, since it is flexible for VaR calculation methods depending on the distribution functions of covariates, limitations on the parameters of these distribution functions, and methods for calculating these parameters. So, for the Gaussian distribution [13]:

$$OpVaR_{0.99} = \eta + \sigma_N \cdot q_{0.99}, \quad (5)$$

where $X \sim N(\eta, \sigma_N)$.

$$OpCVaR_{0.99} = \eta + \sigma_N \cdot \frac{f(z)}{1-\Phi(z)}, \quad (6)$$

where $z = (OpVaR_{0.99} - \eta)/\sigma_N$.

For the Pareto distribution [17]:

$$OpVaR_{0.99} = u + \frac{\sigma}{\xi}\cdot((\frac{n}{r}(1-w))^{-\xi} - 1), \quad (7)$$

DOI 10.15588/1607-3274-2020-2-6

$$OpCVaR_{0.99} = \frac{\sigma + \xi \cdot (OpVaR_{0.99} - u)}{1 - \xi}. \quad (8)$$

Thus, the result of the fifth stage is the calculated values of $OpVaR_{\beta}$ and $OpCVaR_{\beta}$ for both the Gaussian and Pereto d.f. of the nonclustered and clustered data observations.

At the sixth stage, the values of $OpVaR_{\beta}$ and $OpCVaR_{\beta}$ are compared for both the Gaussian and Pareto distribution functions of nonclustered and clustered data, as well as for the same name covariates from the various clusters.

Thus, the influence level of manufacturing environment factors on the VaR for each observed covariate is visible. It is proposed to explain this influence depending on the sample size of the covariates in the cluster and the numerical values of the parameters $OpVaR_{\beta}$ and $OpCVaR_{\beta}$.

The result of the sixth stage is a comparative assessment of the VaR values for observed covariates, depending on certain manufacturing environment factors.

Thus, the important result of the method application is the estimation of VaR depending on factors that interpret the influence of the manufacturing environment as the clustering of process's extreme multivariate values of covariates and variation of covariates distribution functions parameters.

4 EXPERIMENTS

The experiments are aimed at investigating the functioning of the developed method for detecting and taking into account the influence of various factors on the clusterisation and variation of extreme values distribution functions parameters of multivariate observations of process's covariates and further determine the influence of these parameters on risky covariates' values.

The paper uses five-dimensional air quality monitoring data comprising the measurements series of ground level ozone (O_3), nitrogen dioxide (NO_2), nitrogen oxide (NO), sulphur dioxide (SO_2) and particulate matter (PM_{10}) in Leeds (UK) city center, during the years 1994–1998 inclusively.

In particular, was used the daily maxima of O_3 and NO_2 variables during winter periods 1994–1998 inclusively. The gases are recorded in parts per billion and the particulate matter in micrograms per cubic meter [9, 16].

The scheme of experiment is proposed as a sequence of statistical computational procedures of processing the values of covariates' multivariate observations as process's exogenous data:

1. Conducting exploratory analysis of covariates values: determining the compatibility of the data order (equivalence class), assessing the clustering trend. The results of this operation determine the strategy for further research.

2. Clustering the initial data matrix and determining: number of clusters, number of observations (multivariate objects) in the clusters, withincluster sums of squared distances and the coefficient of determination. Creating the multivariate table of covariates values with a factor variable.

3. Decomposition the multivariate values (observations) of the process's covariates into clusters and formation the clustered data tables (data frames).

4. Estimation of EVI for nonclustered and clustered multivariate observations of covariates and determine the Gaussian/Pareto functions parameters for covariates distributions using the MLE.

5. Calculation of VaR (i.e. $OpVaR_{\beta}$ and $OpCVaR_{\beta}$) using the distribution functions parameters, which are defined in the fourth stage.

6. Comparative analysis of the obtained VaR for the Gaussian/Pareto distribution functions of nonclustered and clustered multivariate observations of covariates.

The scheme of experiment corresponds to the steps of the proposed method. For the experiment, software packages of the R language are used.

5 RESULTS

Researches consist in synergy of parametric and nonparametric procedures as well as descriptive and inferential statistics tools.

These procedures use: exploratory analysis of empirical data, cluster data analysis, statistical estimation of sample data distribution functions parameters, obtaining the quantitative characteristics of VaR, and also presenting the results of multivariate statistical analysis in the form of tables and graphs.

At the first method's stage (exploratory analysis): observational data, prone to grouping, are represented by dark squares along the main diagonal of the VAT diagram (hopkins stat = 0.8396), as shown in Fig. 4.

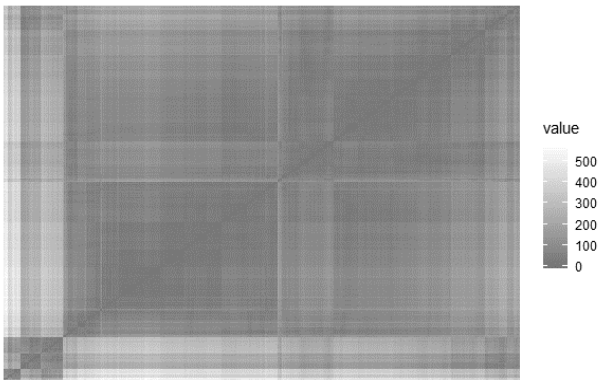

Figure 4 – Visual assessment of cluster tendency diagram

At the second stage: the implementation of the search procedure for the optimal scheme for combining covariate values into clusters determined the number of clusters equal to three, as shown in Fig. 5.

DOI 10.15588/1607-3274-2020-2-6

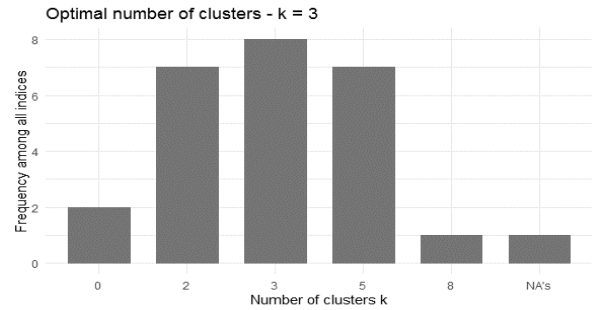

Figure 5 – Diagram for estimating the optimal quantity of clusters

At the third stage: the separation of the observed values of the process's covariates into clusters is presented in the form of an ordination diagram in the space of two main components (Dim1, Dim2), as shown in Fig. 6. The ordination diagram on Fig. 6 visualizes the reduction of three clusters data to the two principal components. Clustering results provide relevant factorized data for further analysis.

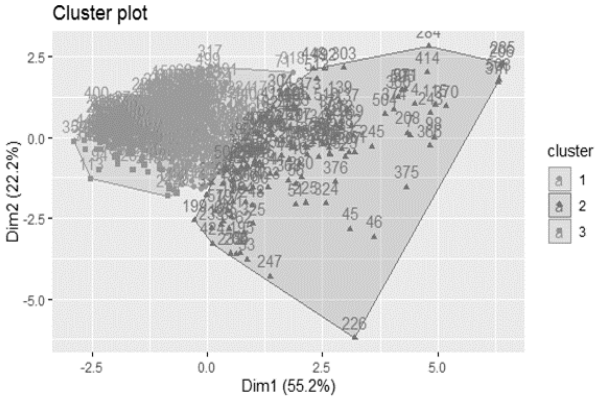

Figure 6 – Diagram of partitioning the observed covariates values into clusters

So, in the fourth stage for distribution functions of nonclustered and clustered covariate data, EVI is estimated using the Ferro & Segers method with further calculation and analysis of the GEVD/GPD parameters by MLE.

Fig. 7 shows graphs of the EVI dependence on noncluster (a) and cluster (b) values of surface ozone. O_3.

The vertical dotted lines show the covariates values for which the EVI estimates are stable for higher covariates values. It is advisable to choose these covariates values as thresholds, for example, for the implementation of POT/AMS analysis, and EVI estimates correspond to these values. Since estimates of $OpVaR_\beta$ and $OpCVaR_\beta$ as the values of the high quantiles of the excesses distribution function are made under the assumption that the covariates extremes have generalized Gaussian or Pareto distribution functions, presented diagrams of graphical verification of the accordance the GEVD/GPD to distribution functions of nonclustered (a) and clustered (b) surface ozone data (O_3), as shown in Fig. 8.

A graphical verification of the accordance of the Gaussian distribution to the distribution functions of nonclustered (a) and clustered (b) surface ozone data (O_3) is show in Fig. 9.

The Q-Q plots visualize the result of comparing standardized residues (4) with the values of the reference distributions.

At the fifth stage, $OpVaR_\beta$ and $OpCVaR_\beta$ values are determined for both the Gaussian and Pareto distribution functions of the nonclustered and clustered observations data of covariates.

For calculations, numerical computer methods were used that are adapted to different VaR calculation tools depending on the distribution functions, limitations for the parameters of these distribution functions, and methods for calculating of these parameters (evir, ReIns R-Packages).

At the sixth stage, a comparative analysis of $OpVaR_\beta$ and $OpCVaR_\beta$ values was performed, which were calculated for different clusters of O_3 covariate, as well as Gaussian and Pareto distribution functions of O_3 covariate.

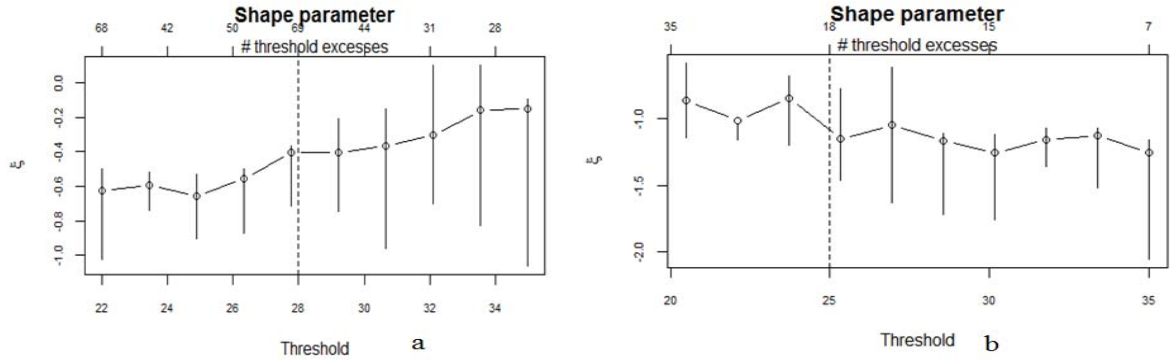

Figure 7 – Graphs of the EVI dependence on nonclustered (a) and clustered (b) values of surface ozone (covariate O_3) and thresholds values (vertical dotted lines)

DOI 10.15588/1607-3274-2020-2-6

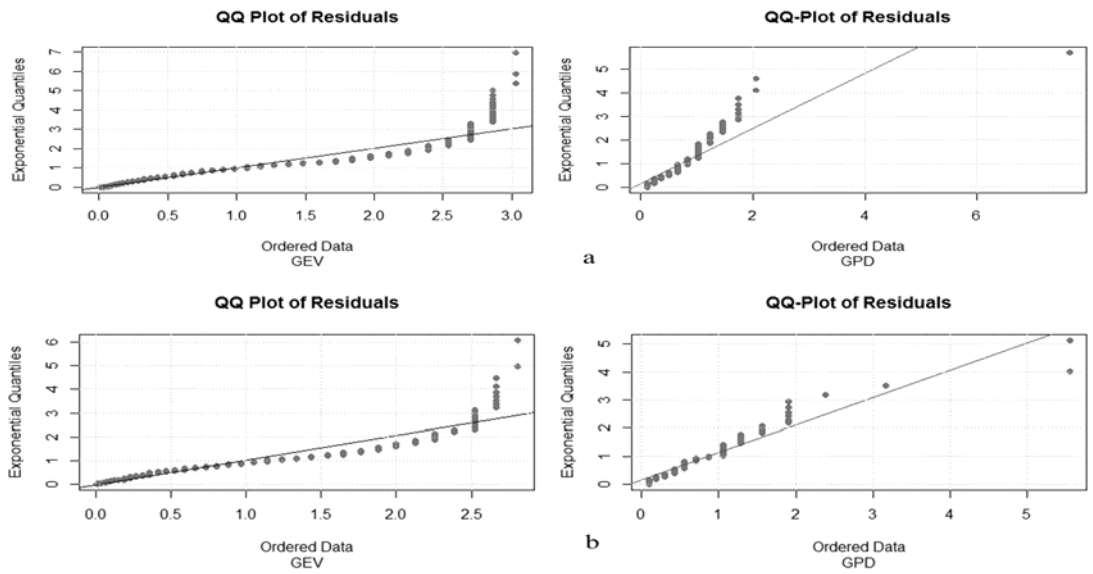

Figure 8 – Graphical verification of the accordance the GEVD/GPD to distribution functions of nonclustered (a) and clustered (b) surface ozone data (O_3) (GEVD/GPD Q-Q plot)

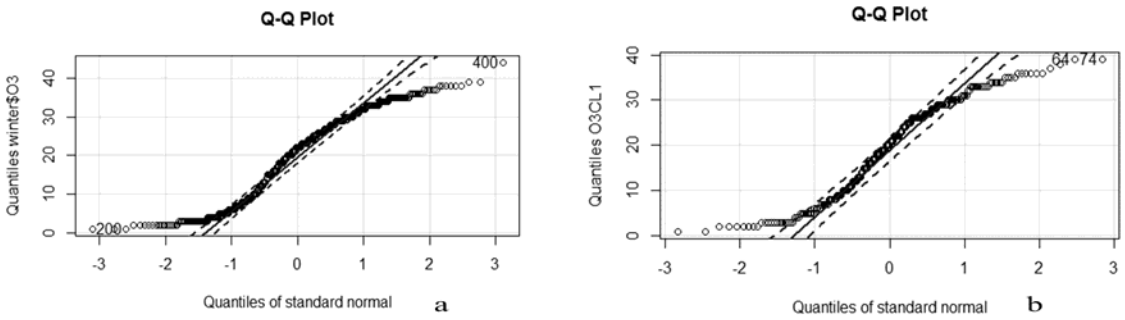

Figure 9 – Graphical verification of the accordance of the Gaussian distribution to the distribution functions of nonclustered (a) and clustered (b) surface ozone data (O_3)

The generalized result of the sixth stage is comparative VaR estimates for observed covariates depending on certain environmental factors.

That, the calculations results allows to identifying the factor influences on the observed covariates values and estimating the VaR values for the according clusters.

6 DISCUSSION

Consider the cause-effect relationships the results of the sequential implementation of the method's stages, presented in the paper (item 3), which lead to the solution of the problem (item 1).

First stage. For understanding the general essence of the initial data behavior and the direction of further research, need to evaluate the data equivalence class and the data tendency to clustering.

Since the covariates data (gas variables) are the same order values and represented as extremes, the procedures of reduction to the same order, smoothing, and data normalization were not performed.

Visual assessment of cluster tendency, where potential groups are represented by dark squares along the main diagonal of the VAT-diagram, showed a strong tendency for clustering with Hopkins statistics equal to 0.8396 for diagram which shown in Fig. 4.

Second stage. Clustering was performed using the R language "NbClust" package, which implements a method for checking the quality of clustering by 30 indices. This method allows to finding the optimal scheme for combining covariates into clusters by rearranging multisets at various combinations of the number of groups, distance metrics, and clustering methods. Based on the calculations, the optimal number of clusters was chosen equal to three, as shown in Fig. 5. The quantity of multivariate objects in the clusters was: 218, 68, and 246 respectively.

The effectiveness of clustering was evaluated by the Zagoruiko criterion of compactness (Zagoruiko test). A search was made for such a partition of objects into groups when the distances between objects from one group (Cl.dist.) in table $v \times v$ (intra-cluster distances, i.e. on-diagonal table elements) are less than a certain value of $\varepsilon > 0$, and between objects from different groups – more than ε, as shown in Table 1.

DOI 10.15588/1607-3274-2020-2-6

Table 1 – The quality data of clustering estimation by the Zagoruiko test

Cl. dist / № Cl	Cl. dist. 1	Cl. dist. 2	Cl. dist. 3
1	60.86545	213.6285	102.17053
2	213.62847	112.9572	301.93825
3	102.17053	301.9382	48.19131

The Cl.dist. values in Table 1 confirm the correctness of the partition of the array of observations into 3 clusters according to the Zagoruiko criterion: on-diagonal table elements < off-diagonal table elements.

Third stage. Visualization the result of dividing the multivariate observed objects into clusters is shown in Fig. 6.

The ratio of the intercluster variance of the observations scatter (BSS) to the total variance of the observations scatter (TSS) gave a value of 76.3%.

This means that the proportion of the scatter variance of the observed objects equal to 76.3% can be explained by the impact of factors that cause the clustering of multivariate observations.

To visualize clustering used the projection (unconstrained ordination) of three data clusters on the axis of the principal components (Dim.1, Dim.2), as shown in Fig. 6. This diagram is accompanied by Table 2, which shows the percentage of data variance for the principal components.

From Table 2 it can be seen that the two main components (Dim. 1, 2) account for 55% and 22% of the scatter values respectively. Table 2 data confirms the correctness of visualization in Fig. 6.

Table 2 – Percentage of data variances at unconstrained ordination of the covariates array on the principal components axis

Parameter / №	Eigenvalue	Percentage of variance	Cumulative percentage of variance
Dim. 1	2.7580	55.1608	55.1608
Dim. 2	1.1110	22.2208	77.3817
Dim. 3	0.5560	11.1216	88.5033
Dim. 4	0.3645	7.2908	95.7941
Dim. 5	0.2102	4.2058	100.0

Fourth stage. The graphs in Fig. 7 visualizes the dependence of EVI on the covariates' values, which was estimated by the Ferro & Segers method for covariate of surface ozone O_3 taking into account nonclustered (a) and clustered (b) data.

The EVI value and the corresponding value of covariate's threshold are selected from the graphs based on the recommendation of the curve part linearity and stability for higher covariates' values (item 3). The selected EVI values correspond to the values of covariates 28 and 25, which are indicated by vertical lines, as shown in Fig. 7. EVI estimates which correspond to these covariate values: about $\xi = -0.4$ for 28 and about $\xi = -1.2$ for 25. EVI values indicate the appropriateness of estimating the parameters of the covariates distribution functions using the MLE.

The MLE of EVI for the GEVD and GPD distribution functions of the nonclustered and clustered surface ozone data O_3 are consistent with EVI estimates by Ferro & Segers method, as shown in Table 3.

Table 3 – MLE values for the parameters of the distribution functions GEVD and GPD of the nonclustered and clustered surface ozone data O_3

Clusters	d. f.	ξ	μ	σ	ξ Ferro& Segers
nonclustered data	GEV	–0.405	16.837	11.320	–0.4
	GPD	–0.550	NA*	8.9375	
clustered data	GEV	–0.483	16.606	11.664	–1.2
	GPD	–0.783	NA*	11.117	

NA* (not available). For GPD are determined parameters ξ, σ .

There is a close connection between the limiting GEVD for block maxima and the limiting GPD for threshold excesses [13]. For a given value of u (threshold), the parameters ξ, μ and σ of the GEVD determine the parameters ξ and σ for GPD. If $\xi < 0$ the distribution functions is in the Weibull family and GPD is a Pareto type II distribution [13]. Note, that for $\xi > -0.5$ the MLEs for μ, σ and ξ are consistent and asymptotically normally distributed with asymptotic variance given by the inverse of the observed information matrix.

Thus, aforementioned allows the use the MLEs for the Gauss and Pareto distribution functions parameters when the excesses are determined for high quantiles of distributions (0.95 and higher).

By Q-Q-plot diagrams are compared the standardized residuals to values that are expected as the reference distribution. So, in Fig. 8, the Q-Q plot is nearly linear in the ordered data area, which confirms the validity of using GEVD and GPD to estimating the statistics of observed data. Note a more exactly corresponding with the GPD for clustered data, as shown in Fig. 8 b.

Graphical verification of the accordance to Gaussian distribution is visualized in Fig.9. From Fig.9 shows that the observed extremes of O_3 distribution have thicker tails than the Gaussian distribution. In the area of 95% confidence interval (dashed lines), the nonclustered (a) and clustered (b) data covariates O_3 coincide with the normal distribution within the interval $]-1.1[$ for quantiles of standard Gaussian distribution. Note a more exactly corresponding with the Gaussian distribution for clustered data, as shown in Fig. 9 b.

Fifth stage. For the Gaussian and Pareto distribution functions of nonclustered and clustered observed data, parameters $OpVaR_\beta$ and $OpCVaR_\beta$ were calculated in accordance with formulas (5), (6) and (7), (8). The calculations were performed for the data of the first and third clusters (Cl. 1, Cl. 3) of surface ozone O_3 with the quantity of observations 218 and 246, respectively.

Table 4 summarizes the calculation data for the parameters $OpVaR_\beta$ (VaR) and $OpCVaR_\beta$ (CVaR) for the Gaussian and Pareto distributions of nonclustered and clustered observed data.

DOI 10.15588/1607-3274-2020-2-6

Table 4 – $OpVaR_\beta$ and $OpCVaR_\beta$ values for the Gaussian and Pareto distribution functions of nonclustered and clustered observed data of surface ozone O_3

Nonclustered (GD)			Clustered Gaussian Distribution (GD)					
			Cl. 1			Cl. 3		
	VaR	CVaR		VaR	CVaR		VaR	CVaR
0.95	37.978	42.53	0.95	37.327	41.887	0.95	39.329	43.414
0.99	45.403	49.09	0.99	44.764	48.462	0.99	45.991	49.303
Nonclustered (PD)			Clustered Pareto Distribution (PD)					
			Cl. 1			Cl. 3		
	VaR	CVaR		VaR	CVaR		VaR	CVaR
0.95	37.994	40.21	0.95	36.321	37.578	0.95	37.728	39.926
0.99	41.665	42.57	0.99	38.372	38.728	0.99	41.376	42.401

Sixth stage. Table 4 allows to compare the estimates of $OpVaR_\beta$ and $OpCVaR_\beta$ values for the Gaussian and Pareto distribution functions of nonclustered and clustered observed data of covariate O_3. Table 4 shows the difference in risk estimates for non-clustered and clustered observational data.

The ratio $OpVaR_\beta < OpCVaR_\beta$ is explainable by the definition of the operational conditional value at risk $OpCVaR$ as the expected value provided that the operational value at risk is exceeded.

It can be argued that VaR assessments are more reliable for nonclustered and clustered data, which more precisely coincide with the Pareto and Gaussian distribution functions, as shown in Fig. 8 and Fig. 9. Suppose also that VaR is influenced by the withincluster sum of distances as measure of the multivariate objects scattering within groups.

Thus, for each sample of observed data and in accordance with the proposed method: from the beginning it is necessary to evaluate the data distribution function, the parameters of this function and EVI, and then calculate the VaR values.

CONCLUSIONS

The method is an information technology of statistical causal analysis of the influence of factors on the variability of the processes covariates and values of their VaR due to the application of the clustering procedure for the multivariate observed extreme values of covariates.

The proposed approach is based on the determination of the initial mutual ordering of the observed data (clustering) and allows to associate this specific homogeneity in the data variation with external causes.

The mutual ordering of data in homogeneous observation groups is characterized by intercluster (constrained) variance, and the withincluster variance of data is characterized by withincluster variance associated with the action of random (unconstrained) factors.

The basic point is the determination of EVI and the location and scale parameters for the distribution functions of nonclustered and clustered covariates values. The parameters of the distribution functions are data for determining VaRs, which are then compared for nonclustered and clustered covariates in order to interpret the factors influence. So, for certain processes, the quantity and degree of factor influence on covariates' multivariate observed data and covariates' VaRs are identified.

Thus, **the scientific novelty** of the obtained results consists in the fact that is firstly proposed the risks estimation method by extreme data of the process covariates, which are clustered at the homogeneous signs of scattering the multivariate objects in the multidimensional area of their random values.

The research results are the compilation consequence of multivariate cluster analysis tools with tools of the extreme values theory, which allows to establishing a causal relationship and giving an adequate **practical interpretation** of the influence of environmental (manufacturing) factors on risky values of the process's covariates.

The **prospect of further research** is to improve the methods of causal multivariate statistical analysis of the various factors influence on the exogenous and endogenous parameters of manufacturing and other processes in order to reduce the variability of these parameters and, as a result, minimize the risks.

ACKNOWLEDGEMENTS

The paper is submitted as a competition framework of scientific and technological developments "Science for the safety of man and society" in accordance with the Resolution of the scientific council of the National Research Foundation of Ukraine (protocol №. 7 of May 11, 2020).

REFERENCES

1. ISO/TC 176/SC 2/N 544R3. ISO 9000 Introduction and Support Package: Guidance on the Concept and Use of the Process Approach for management systems [Electronic resource]. Access mode: https://www.iso.org/files/live/sites/isoorg/files/archive/pdf/en/04_concept_and_use_of_the_process_approach_for_management_systems.pdf
2. ISO 9001: 2015. Quality management systems [Electronic resource]. Access mode: https://www.iso.org/standard/62085.html
3. ISO 31000 Risk management [Electronic resource]. Access mode: https://www.iso.org/iso-31000-risk-management.html
4. Beirlant J., Goegebeur Y., Segers J. et al. Statistics of Extremes: Theory and Applications, Wiley Series in Probability and Statistics, Chichester, John Wiley & Sons, 2004, 522 p.
5. Castillo E., Hadi A. S., Balakrishnan N. et al. Extreme Value and Related Models with Applications in Engineering and Science. New York, Wiley, 2004, 362 p.
6. Kreinovich V., Nguyen H. T., Sriboonchitta S., Kosheleva O. Modeling Extremal Events Is Not Easy: Why the Extreme Value Theorem Cannot Be as General as the Central Limit Theorem [Electronic resource]. Access mode: http://digitalcommons.utep.edu/cs_techrep/923
7. Repository CRAN [Electronic resource]. Access mode: https://cran.r-project.org/web/packages/fExtremes/fExtremes.pdf
8. Coles S. An Introduction to Statistical Modeling of Extreme Values. London, Springer, 2001, 209 p.
9. Repository CRAN [Electronic resource]. Access mode: https://cran.r-project.org/web/packages/texmex/index.html.

DOI 10.15588/1607-3274-2020-2-6

10. Novak S. Y. Extreme Value Methods with Applications to Finance. Florida, CRC Press, 2011, 399 p.
11. Yan J. Extreme Value Modeling and Risk Analysis: Methods and Applications / J. Yan, D. K. Dey. – Florida: CRC Press, 2016. – 540 p.
12. Tiemann T. K. Introductory Business Statistics with Interactive Spreadsheets [Electronic resource]. Access mode: http://people.wcsu.edu/lightwoods/mat120%20s19/IBStats%20CanEd.pdf
13. Zivot E., Wang J. Modeling Financial Time Series with S-PLUS: Second Edition. New York, Springer-Verlag, 2006, 705 p.
14. Bensalah Y. Steps in Applying Extreme Value Theory to Finance: A Review [Electronic resource]. Access mode: https://www.banqueducanada.ca/wp-content/uploads/2010/01/wp00-20.pdf
15. Ferro C. A. T., Segers J. Inference for clusters of extreme values, *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 2003, Vol. 65, No. 2, pp. 545–556.
16. Heffernan J. E., Tawn J. A. A conditional approach for multivariate extreme values, *Journal of the Royal Statistical Society. Series B: Statistical Methodology,* 2004, Vol. 66, No. 3, pp. 497–546.
17. An overview of non-parametric clustering and computer vision [Electronic resource]. Access mode: https://web.archive.org/web/20080113181815/http://www.nerd-cam.com/cluster-results/
18. Tutorial with introduction of Clustering Algorithms [Electronic resource]. Access mode: http://home.deib.polimi.it/matteucc/Clustering/tutorial_html/
19. Coghlan A. A Little Book of R For Multivariate Analysis, Release 0.1 [Electronic resource]. Access mode: http://a-little-book-of-r-for-time-series.readthedocs.org/
20. Havrylko Ye., Kurchenko O., Tereshchenko I. et al. The method of multivariate statistical analysis of the time multivariate critical quality attributes of manufacture process with the data factorization, *Radio Electronics, Computer Science, Control,* 2019, № 1, pp. 167–177.
21. Branch Information Technologies of Quality Management / [V. Nakonechnyi, S. Toliupa, I. Tereshchenko et al.] // Problems of Infocommunications. Science and Technology (PIC S&T): International Scientific-Practical Conference, Kharkov, 9–12 Oct. 2018: proceedings. – Kharkov : IEEE, 2018. – P. 783–788.
22. Gomes M. I., Guillou A. Extreme Value Theory and Statistics of Univariate Extremes: A Review, *International Statistical Review*, 2015, Vol. 83, No. 2, pp. 263–292.
23. Gomes M. I. Generalized Means in Statistical EVT [Electronic resource]. Access mode: https://www.researchgate.net/publication/337635760_Generalized_Means_in_Statistical_EVT
24. Bezak N., Brilly M., Šraj M. Comparison between the peaks-over-threshold method and the annual maximum method for flood frequency analysis, *Hydrological Sciences Journal*, 2014,Vol. 59, No. 5, pp. 959–977.
25. Gomes M. I., Caeiro F., Figueiredo F. et al. Corrected-Hill versus partially reduced-bias value-at-risk estimation, *Journal Communications in Statistics – Simulation and Computation*, 2020, Vol. 49, No. 4, pp. 867–885.
26. Caeiro F., Henriques-Rodrigues L., Gomes P. D. A simple class of reduced bias kernel estimators of extreme value parameters, *Computational and Mathematical Methods*, 2019, Vol. 1, No. 3, pp. 1–12.
27. [Ragno E., Kouchak A. A., Cheng L. et al. A generalized framework for process-informed nonstationary extreme value analysis, *Advances in Water Resources*, 2019-08, Vol. 130, pp. 270–282.
28. Chai W. A., Leira B. J., Naess A. Probabilistic methods for estimation of the extreme value statistics of ship ice loads, *Cold Regions Science and Technology*, 2018-02, Vol. 146, pp. 87–97.
29. Majumdar S. N., Pal A., Schehr G. Extreme value statistics of correlated random variables: A pedagogical review, *Physics Reports,* 2020-01-22, Vol. 840, pp. 1–32.
30. Tsay R. S. Testing serial correlations in high-dimensional time series via extreme value theory, *Journal of Econometrics*, 2020, Vol. 216, No. 1, pp. 106–117.
31. Carreau J., Toulemonde G. Extra-parametrized extreme value copula: Extension to a spatial framework [Electronic resource]. Access mode: https://hal.inria.fr/hal-02419118/document
32. Quinn N., Bates P., Neal J. et al. The spatial dependence of flood hazard and risk in the United States, *Water Resources Research,* 2019, Vol. 55, No. 3, pp. 1890–1911.
33. Abad P., Benito S., López C. A comprehensive review of Value at Risk methodologies, *The Spanish Review of Financial Economics,* 2014-01, Vol. 12, No. 1, pp. 15–32.
34. Zrazhevska N. G., Zrazhevsky A. G. Classification methods for risk measures VaR and CVaR calculation and estimation, *System Research & Information Technologies*, 2016, No. 3, pp. 126–141.
35. Stephenson A., Gilleland E. Software for the analysis of extreme events: The current state and future directions, *Extremes*, 2005-01, Vol. 8, No. 3, pp. 87–109.
36. Gilleland E., Ribatet M., Stephenson A. G. A software review for extreme value analysis, *Extremes*, 2013-03-01, Vol. 16, No. 1, pp. 103–119.
37. Natarajan D. ISO 9001 Quality Management Systems (Management and Industrial Engineering). Cham, Springer, 2017-03-31, 181 p.
38. Greenwood J. A., Landwehr J. M., Matalas N. C. et al. Probability weighted moments: definition and relation to parameters of several distributions expressible in inverse form, *Water Resources Research*, 1979, Vol. 15, No. 5, pp. 1049–1054.

Received 21.04.2020.
Accepted 17.06.2020.

УДК 004.94:658

МЕТОД ОЦІНКИ РИЗИКІВ ЗА КЛАСТЕРНИМИ ЕКСТРЕМАЛЬНИМИ ДАНИМИ КОВАРІАТ ПРОЦЕСУ

Терещенко І. В. – канд. техн. наук, доцент, доцент кафедри інфокомунікаційної інженерії, Харківський національний університет радіоелектроніки, Харків, Україна.

Терещенко А. І. – канд. техн. наук, асистент кафедри управління інформаційною та кібернетичною безпекою, Державний університет телекомунікацій, Київ, Україна.

Штангей С. В. – канд. техн. наук, доцент, доцент кафедри інфокомунікаційної інженерії, Харківський національний університет радіоелектроніки, Харків, Україна.

DOI 10.15588/1607-3274-2020-2-6

АНОТАЦІЯ

Актуальність. У даній роботі представлений метод вирішення проблеми виявлення та врахування впливу різних (зовнішніх та/або внутрішніх) факторів на екстремальні та ризикові значення багатовимірних спостережуваних параметрів (коваріат) технологічних та/або діагностичних процесів. Врахування факторів зовнішнього та внутрішнього впливу на коваріати за аналогією з критичними параметрами процесу є суттєвим доповненням до статистики екстремальних значень та оцінок впливу змінності коваріат процесу на очікувані втрати, тобто значення ризику. Ризик орієнтований аналіз, є актуальним інструментом для дослідження поведінки даних багатовимірних спостережень параметрів процесу.

Метод. Метод послідовно використовує: процедури багатовимірного статистичного кластерного аналізу, перетворення матриці спостережуваних екстремальних значень коваріат процесу в фрейм даних з факторними змінними, оцінку екстремального індексу та параметрів функцій розподілу некластеризованих та кластеризованих спостережуваних екстремальних даних коваріат та оцінки значення ризику для обчислених значень параметрів функцій розподілу. Пропонована послідовність дій спрямована на впровадження інформаційної технології статистичного причинно-наслідкового аналізу впливу факторів на змінність коваріат процесу і значень їх ризиків за рахунок застосування процедури кластеризації для спостережуваних багатовимірних екстремальних значень коваріат. Метод використовує програмні пакети мови R.

Результати. Кластеризація багатовимірних спостережуваних екстремальних значень коваріат процесу дозволяє виявити вплив екологічних (виробничих) факторів та оцінити ризикові значення коваріат з урахуванням цього впливу.

Висновки. Метод являє собою інформаційну технологію статистичного причинно-наслідкового аналізу впливу факторів на змінність коваріат процесу і значень їх ризиків за рахунок застосування процедури кластеризації багатовимірних значень коваріат. Перспектива подальших досліджень полягає в удосконаленні методів причинно-наслідкового багатовимірного статистичного аналізу впливу різних факторів на екзогенні та ендогенні параметри виробничих та інших процесів з метою зниження змінності цих параметрів і, як наслідок, мінімізації ризиків.

КЛЮЧОВІ СЛОВА: теорія екстремальних значень, узагальнений розподіл екстремальних значень, узагальнений розподіл Парето, величина ризику, індекс екстремальних значень, кластерний аналіз, процесний підхід.

УДК 004.94:658

МЕТОД ОЦЕНКИ РИСКОВ ПО КЛАСТЕРНЫМ ЭКСТРЕМАЛЬНЫМ ДАННЫМ КОВАРИАТ ПРОЦЕССА

Терещенко И. В. – канд. техн. наук, доцент, доцент кафедры инфокоммуникационной инженерии, Харьковский национальный университет радиоэлектроники, Харьков, Украина.

Терещенко А. И. – канд. техн. наук, ассистент кафедры управления информационной и кибернетической безопасностью, Государственный университет телекоммуникаций, Киев, Украина.

Штангей С. В. – канд. техн. наук, доцент, доцент кафедры инфокоммуникационной инженерии, Харьковский национальный университет радиоэлектроники, Харьков, Украина.

АННОТАЦИЯ

Актуальность. В данной работе представлен метод решения проблемы выявления и учета влияния различных (внешних и/или внутренних) факторов на экстремальные и рисковые значение многомерных наблюдаемых параметров (ковариат) технологических и/или диагностических процессов. Учет факторов внешнего и внутреннего влияния на ковариаты по аналогии с критическими параметрами процесса является существенным дополнением к статистике экстремальных значений и оценкам влияния изменчивости ковариат процесса на ожидаемые потери, то есть значение риска. Риск ориентированный анализ, является актуальным инструментом для исследования поведения данных многомерных наблюдений параметров процесса.

Метод. Метод последовательно использует процедуры многомерного статистического кластерного анализа, трансформации матрицы экстремальных данных наблюдений ковариат процесса в таблицы данных с факторными переменными, оценки экстремального индекса и параметров функций распределения некластеризованных и кластеризованных экстремальных данных наблюдений ковариат, а также оценки величин рисков по рассчитанным значениям параметров функций распределения. Предлагаемая последовательность действий направлена на внедрение информационной технологии статистического причинно-следственного анализа влияния факторов на изменчивость ковариат процесса и значения их рисков за счет применения процедуры кластеризации для наблюдаемых многомерных экстремальных значений ковариат. Метод использует программные пакеты языка R.

Результаты. Кластеризация многомерных наблюдаемых экстремальных значений ковариат процесса позволяет выявить влияние экологических (производственных) факторов и оценить рисковые значение ковариат с учетом этого влияния.

Выводы. Метод представляет собой информационную технологию статистического причинно-следственного анализа влияния факторов на изменчивость ковариат процесса и значений их рисков за счет применения процедуры кластеризации многомерных значений ковариат. Перспектива дальнейших исследований заключается в совершенствовании методов причинно-следственного многомерного статистического анализа влияния различных факторов на экзогенные и эндогенные параметры производственных и других процессов с целью снижения изменчивости этих параметров и, как следствие, минимизации рисков.

КЛЮЧЕВЫЕ СЛОВА: теория экстремальных значений, обобщенное распределение экстремальных значений, обобщенное распределение Парето, величина риска, индекс экстремальных значений, кластерный анализ, процессный подход.

ЛІТЕРАТУРА / ЛИТЕРАТУРА

1. ISO/TC 176/SC 2/N 544R3. ISO 9000 Introduction and Support Package: Guidance on the Concept and Use of the Process Approach for management systems [Electronic resource]. – Access mode: https://www.iso.org/files/live/sites/isoorg/files/archive/pdf/en/04_concept_and_use_of_the_process_approach_for_management_systems.pdf

DOI 10.15588/1607-3274-2020-2-6

2. ISO 9001: 2015. Quality management systems [Electronic resource]. – Access mode: https://www.iso.org/standard/62085.html
3. ISO 31000 Risk management [Electronic resource]. – Access mode: https://www.iso.org/iso-31000-risk-management.html
4. Statistics of Extremes: Theory and Applications / [J. Beirlant, Y. Goegebeur, J. Segers et al.]. – Wiley Series in Probability and Statistics. – Chichester : John Wiley & Sons, 2004. – 522 p.
5. Extreme Value and Related Models with Applications in Engineering and Science / [E. Castillo, A. S. Hadi, N. Balakrishnan et al.]. – New York: Wiley, 2004. – 362 p.
6. Kreinovich V. Modeling Extremal Events Is Not Easy: Why the Extreme Value Theorem Cannot Be as General as the Central Limit Theorem [Electronic resource] / V. Kreinovich, H. T. Nguyen, S. Sriboonchitta, O. Kosheleva. – Access mode: http://digitalcommons.utep.edu/cs_techrep/923
7. Repository CRAN [Electronic resource]. – Access mode: https://cran.r-project.org/web/packages/fExtremes/fExtremes.pdf
8. Coles S. An Introduction to Statistical Modeling of Extreme Values / S. Coles. – London: Springer, 2001. – 209 p.
9. Repository CRAN [Electronic resource]. – Access mode: https://cran.r-project.org/web/packages/texmex/index.html
10. Novak S. Y. Extreme Value Methods with Applications to Finance / S. Y. Novak. – Florida: CRC Press, 2011. – 399 p.
11. Yan J. Extreme Value Modeling and Risk Analysis: Methods and Applications / J. Yan, D. K. Dey. – Florida : CRC Press, 2016. – 540 p.
12. Tiemann T. K. Introductory Business Statistics with Interactive Spreadsheets [Electronic resource] / T. K. Tiemann. – Access mode: http://people.wcsu.edu/lightwoods/mat120%20s19/IBStats%20CanEd.pdf
13. Zivot E. Modeling Financial Time Series with S-PLUS: Second Edition / E. Zivot, J. Wang. – New York: Springer-Verlag, 2006. – 705 p.
14. Bensalah Y. Steps in Applying Extreme Value Theory to Finance: A Review [Electronic resource] / Y. Bensalah. – Access mode: https://www.banqueducanada.ca/wp-content/uploads/2010/01/wp00-20.pdf
15. Ferro C. A. T. Inference for clusters of extreme values / C. A. T. Ferro, J. Segers // Journal of the Royal Statistical Society. Series B: Statistical Methodology. – 2003. – Vol. 65, № 2. – P. 545–556.
16. Heffernan J. E. A conditional approach for multivariate extreme values / J. E. Heffernan, J. A. Tawn // Journal of the Royal Statistical Society. Series B: Statistical Methodology. – 2004. – Vol. 66, № 3. – P. 497–546.
17. An overview of non-parametric clustering and computer vision [Electronic resource]. – Access mode: https://web.archive.org/web/20080113181815/http://www.nerd-cam.com/cluster-results/
18. Tutorial with introduction of Clustering Algorithms [Electronic resource]. – Access mode: http://home.deib.polimi.it/matteucc/Clustering/tutorial_html/
19. Coghlan A. A Little Book of R For Multivariate Analysis, Release 0.1 [Electronic resource] / A. Coghlan. – Access mode: http://a-little-book-of-r-for-time-series.readthedocs.org/
20. The method of multivariate statistical analysis of the time multivariate critical quality attributes of manufacture process with the data factorization / [Ye. Havrylko, O. Kurchenko, I. Tereshchenko et al.] // Radio Electronics, Computer Science, Control. – 2019. – № 1. – P. 167–177.
21. Branch Information Technologies of Quality Management / [V. Nakonechnyi, S. Toliupa, I. Tereshchenko et al.] // Problems of Infocommunications. Science and Technology (PIC S&T): International Scientific-Practical Conference, Kharkov, 9–12 Oct. 2018: proceedings. – Kharkov: IEEE, 2018. – P. 783–788.
22. Gomes M. I. Extreme Value Theory and Statistics of Univariate Extremes: A Review / M. I. Gomes, A. Guillou // International Statistical Review. – 2015. – Vol. 83, № 2. – P. 263–292.
23. Gomes M. I. Generalized Means in Statistical EVT [Electronic resource] / M. I. Gomes. – Access mode: https://www.researchgate.net/publication/337635760_Generalized_Means_in_Statistical_EVT
24. Bezak N. Comparison between the peaks-over-threshold method and the annual maximum method for flood frequency analysis / N. Bezak, M. Brilly, M. Šraj // Hydrological Sciences Journal. – 2014. – Vol.59, № 5. – P. 959–977.
25. Corrected-Hill versus partially reduced-bias value-at-risk estimation / [M. I. Gomes, F. Caeiro, F. Figueiredo et al.] // Journal Communications in Statistics – Simulation and Computation. – 2020. – Vol. 49, № 4. – P. 867–885.
26. Caeiro F., Henriques-Rodrigues L., Gomes P. D. A simple class of reduced bias kernel estimators of extreme value parameters / F. Caeiro, L. Henriques-Rodrigues, D. P. Gomes // Computational and Mathematical Methods. – 2019. – Vol. 1, № 3. – P. 1–12.
27. A generalized framework for process-informed nonstationary extreme value analysis / [E. Ragno, A. A. Kouchak, L. Cheng et al.] // Advances in Water Resources. – 2019-08. – Vol. 130 – P. 270–282.
28. Chai W. A. Probabilistic methods for estimation of the extreme value statistics of ship ice loads / W. Chai, B. J. Leira, A. Naess // Cold Regions Science and Technology. – 2018-02. – Vol. 146. – P. 87–97.
29. Majumdar S. N. Extreme value statistics of correlated random variables: A pedagogical review / S. N. Majumdar, A. Pal, G. Schehr // Physics Reports. – 2020-01-22. – Vol. 840. – P. 1-32.
30. Tsay R. S. Testing serial correlations in high-dimensional time series via extreme value theory / R. S. Tsay // Journal of Econometrics. – 2020. – Vol. 216, № 1. – P. 106–117.
31. Carreau J. Extra-parametrized extreme value copula: Extension to a spatial framework [Electronic resource] / J. Carreau, G. Toulemonde. – Access mode: https://hal.inria.fr/hal-02419118/document
32. The spatial dependence of flood hazard and risk in the United States / [N. Quinn, P. Bates, J. Neal et al.] // Water Resources Research. – 2019. – Vol.55, № 3. – P. 1890-1911.
33. Abad P. A comprehensive review of Value at Risk methodologies / P. Abad, S. Benito, C. López // The Spanish Review of Financial Economics. – 2014-01. – Vol. 12, № 1. – P. 15-32.
34. Zrazhevska N. G. Classification methods for risk measures VaR and CVaR calculation and estimation / N. G. Zrazhevska, A. G. Zrazhevsky // System Research & Information Technologies. – 2016. – № 3. – P. 126-141.
35. Stephenson A. Software for the analysis of extreme events: The current state and future directions / A. Stephenson, E. Gilleland // Extremes. – 2005-01. – Vol. 8, № 3. – P. 87–109.
36. Gilleland E. A software review for extreme value analysis / E. Gilleland, M. Ribatet, A. G. Stephenson // Extremes. – 2013-03-01. – Vol. 16, № 1. – P. 103–119.
37. Natarajan D. ISO 9001 Quality Management Systems (Management and Industrial Engineering) / D. Natarajan. – Cham : Springer, 2017-03-31. – 181 p.
38. Probability weighted moments: definition and relation to parameters of several distributions expressible in inverse form / [J. A. Greenwood, J. M. Landwehr, N. C. Matalas et al.] // Water Resources Research. – 1979. – Vol. 15, № 5. – P. 1049–1054.

DOI 10.15588/1607-3274-2020-2-6