UDC 519.766.4:004.942
DOI: 10.20535/SRIT.2308-8893.2024.1.04

OPERATIONAL RISK ESTIMATION USING SYSTEM ANALYSIS METHODOLOGY

P.I. BIDYUK, O.L. TYMOSHCHUK, L.B. LEVENCHUK

Abstract. Financial risks are considered today as popular research topics due to the existing practical necessity for the use of their mathematical models, estimates of possible loss in many areas of human activities, forecasting, and respective managerial decisions in financial and other spheres where capital, obligations, stocks, bonds, and other activities are circulating successfully. Financial processes today exhibit sophisticated forms of evolution in time that require the application of sophisticated modeling, risk estimating, forecasting, and decision-making/support methods, techniques, and procedures. The system analysis approach is applied to solving such problems as a unique and universal research methodology. The financial risks, specifically the operational ones in the study considered, are classified as nonlinear and nonstationary processes that require appropriate methods for analysis and a rather sophisticated analytical description to estimate and forecast possible loss. The results of operational risk analysis are achieved in the form of systemic methodology, models constructed with statistical data, regression analysis, and Bayesian techniques, and estimated loss with the models. The models and system analysis approach proposed for analyzing financial processes are suitable for practical applications, provided the users have appropriate statistical data and expert estimates.

Keywords: financial operational risk, mathematical model, statistical data, system analysis methodology, loss estimation, decision support system.

INTRODUCTION

Financial risks are related to very popular research topics due to existing practical necessity for the use of their mathematical models, estimates of possible loss in many areas of human activities, forecasting and respective managerial decisions in financial and many other spheres where capital, obligations, stocks, bonds and other activities are circulating successfully [1]. Generally financial processes often exhibit rather sophisticated forms of evolution in time what requires application of modern modeling, forecasting and decision making/support methods, technics and procedures. The effective system analysis approach should be applied to solving such problems successfully [2]. The most known financial risks can be classified as nonlinear and nonstationary processes (NNP) that require appropriate attention for analysis and rather sophisticated analytical description to estimate and forecast possible loss.

© *P.I. Bidyuk, O.L. Tymoshchuk, L.B. Levenchuk*, 2024

The possible process nonstationarities appear due to availability of deterministic and stochastic trends, changing in time variance, and possible nonlinearities can be divided into two wide classes: nonlinearities with respect to variables and nonlinearities with respect to parameters. The nonlinearities with respect to variables are easier to cope with because the models (parameters) containing them can be estimated easier than models nonlinear with respect to parameters. For example, the model describing process with nonlinear trend (say, polynomial trend of second or higher order) can be correctly estimated using ordinary or nonlinear LS (NLS) or maximum likelihood (ML) methods. On the other side, the models nonlinear with respect to parameters may require for their estimation application of sophisticated optimization procedures like NLS, including Bayesian techniques such as Monte Carlo for Markov Chain (MCMC) or Bayesian optimization approach.

Thus, most of the financial processes that researchers have to cope with today are nonlinear and nonstationary what requires paying special attention to modeling their structure and parameter estimation. Generally, most of the processes around us (in economy, finances, industrial production, weather forecasting, description of natural phenomena) are related to NNPs, and they should be modeled, forecasted and governed correctly using respective theoretical basis provided by the estimation theory [3]. Usually specialized intellectual decision support systems (IDSS), based upon system analysis principles, are constructed to analyze the processes that help substantially to improve constructed model adequacy, quality of forecasts, respective decisions, and increase general understanding of practical situations related to the modeling and forecast estimation problems. This is especially true regarding financial risk analysis where different model types are to be applied to reach high quality results of possible loss estimation/forecasting.

Among often met financial risks that researchers today have to cope with the following ones should be mentioned: market risk, operational risk, risk of a country, credit and liquidity risks, risk of structured products sale, Exchange-traded products (ETP) risk, Over-the-counter (OTC) derivatives risk and some others. In this study we provide for a short review of modern methods for modeling selected types of financial risks, and estimation of possible loss with special attention to active usage of IDSS. Comparison of some approaches to risk estimation will also be given. The studies [4, 5, 6] provide for grounded notion of "risk", consider possible instruments for operational risk analysis, and describe basic problems met on the way of development and practical implementation of risk estimation and prediction systems in financial institutions. The study [7] touches upon analysis of key risks inherent to Ukrainian banking system. Many studies are devoted to analysis and use of loss distributions approach (LDA) because it provides the possibility for estimating the volume of capital necessary for covering operational loss. The paper [8] is completely dedicated to analysis and practical use of the popular LDA approach.

PROBLEM STATEMENT

The purposes of the study are as follows: to provide a short review of operational financial risk modeling methods and respective loss estimation; to consider possibility of financial risk estimation by making use of regression techniques (generalized linear models) and Bayesian approach to data analysis; to compare quality of risk estimation by the methods selected; to give necessary illustrative examples of financial operational risk estimation; to stress necessity of constructing and practical application of IDSS to estimation/prediction of possible financial loss.

SOME GENERAL APPROACHES TO FINANCIAL RISK ESTIMATION

Here we start analysis with operational risk that is available practically in every activity when we consider capital circulation problems on different levels of economy and finances. The term “operational risk” does not have clearly established definition. Some financial institutions define it as non-measurable risk (what is also correct). In 2001 the Basel Committee on Banking Supervision formulated general supervision standards and governing principles for banking system, and gave definition for operational risk: “operational is the risk of loss due to inadequate or erroneous internal processes, actions of co-workers or systems, or influence of external events” [1].

The Basel Committee also proposed some methods for operational risk estimation, namely: the basic indicator approach (BIA); standardized indicator approach (SIA); internal estimation approach (IMA); an approach based upon loss distribution curve (LDA); scoring cards, and scenario analysis techniques. The first three methods are relatively simple, and can be used in various financial organizations though they do not take into consideration special features of an enterprise, its size and principles if risk management implemented within its structure. Using the methods mentioned “good” and “bad” financial company will be keeping the same amount of capital to cover operational risks. The last three methods are considered as advanced approaches to operational risk analysis and estimation. They are based upon mathematical models and advanced information technologies, and provide the possibility for decreasing the capital necessary for covering operational risk loss. Now consider some types of models used in practice of operational and some other risks analysis.

GENERALIZED LINEAR MODELS

Generalized linear models (GLM) create to some extent universal approach to modeling various processes including linear and nonlinear, stationary and nonstationary ones. This is possible thanks to the fact that this group of models includes the models of different structures: pure linear Bayesian multiple regression; Poisson regression; variance and covariance analysis; nonlinear structures including nonlinear polynomial elements; nonlinear structures like logit and probit models that are capable to capture nonlinear dependences in the processes under study. Combination of linear and nonlinear structural elements in one model is directed to describing complex nonlinear nonstationary dependences inherent to many financial processes. For example, nonlinear logit and probit models can be combined with paired or multiple linear regression, Bayesian networks (static and/or dynamic), Bayesian data filtering procedures that can be linear or nonlinear.

ESTIMATION OF REGRESSION MODEL STRUCTURE

The basics of regression model structure estimation were proposed by Box and Junkins in the 1960s. Today the modeling methodology proposed, based on system analysis approach, includes the following steps:

- systemic analysis of a process under study on the basis of available basic theory and expert knowledge, analysis of inputs and outputs behavior, represented by the time series information, study of earlier existing model structures, and other accessible information;
- preliminary processing of available data that is based upon application of appropriate normalizing and filtering procedures, appropriate analysis of possible extreme values, filling in missing values, structuring the data etc;
- analysis of statistical data for stationarity and nonlinearity using available statistical tests;
- generation and estimation of candidate model structures with taking into consideration possible structure of GLM: identification of stochastic and systemic components and link function; computing descriptive characteristics; determining statistical significance of independent variables using Wald statistic; estimate statistical characteristics of other structural elements of constructed mathematical model (for example, residuals);
- selection/development and application of model parameter estimation method (or methods); it can be linear or nonlinear least squares (NLS) method, maximum likelihood (ML) or Monte Carlo based Bayesian approach;
- selection of the best fitting model in the set of estimated candidates using statistical quality criteria.

Now consider some details of the stages mentioned.

PRELIMINARY DATA PROCESSING

The time series theory considers observations as random variables containing deterministic component. Generally we have practically always process the data measurements influenced by some random external disturbances and measurement errors. The purpose of preliminary data processing is in eliminating possible errors from available measurements and improving conditions for determining their distribution and further modeling. According to current requirements when GLMs are hired for formal describing the processes under study the data should correspond to the family of exponential distributions [9; 10].

Preliminary data processing usually includes the following operations:

- normalization and possible correcting of measurements; normalization means application of logarithmic operator or transforming the data to acceptable and convenient amplitude interval;
- possible data correction may include imputation of missing values, application of filtering techniques, and processing extreme values;
- computing, when necessary, first and higher order differences that are necessary for analysis of corresponding effects of a time series under study.

Sample mean is subtracted sometimes from the observations to get a possibility for constructing a model for deviations. It is clear that application of specific data preprocessing technique depends on specific features of the data available and established purpose of modeling (forecasting, deeper analysis of a process or control).

Nonlinear deterministic trend available in data can be identified by estimating the following time polynomial:

$$y(k) = a_0 + c_1 k + c_2 k^2 + \ldots + c_m k^m ,$$

where $k=0,1,2,...$ is discrete time ($t=kT_s$), where t is continuous time, and T_s is sampling interval for measurements. If one of the coefficients c_i, $i=2,...,m$ is statistically significant, then hypothesis on trend linearity is rejected. In the case when trend is quickly changing it time and its functional description is not adequate enough, then model of stochastic trend should be constructed that is based upon combinations of random processes.

ESTIMATION OF OTHER ELEMENTS OF GLM MODEL STRUCTURE

Taking into consideration possible GLM distributions there are several GLM types presented in Table 1.

Table 1. GLM types considered in the study

Type of model	Link function	Distribution of dependent variable
GLM	$g(\mu)=\mu$	Normal
Log-linear model	$g(\mu)=\ln(\mu)$	Poisson
Logistic model	$g(\mu)=\ln\left(\frac{\mu}{1-\mu}\right)$	Binomial
Probit-analysis	$g(\mu)=\Phi^{-1}\mu$	Binomial
"Survival" analysis	$g(\mu)=\mu^{-1}$	Gamma-distribution, exponential

Considering GLM, model structure is estimated from the point of view of its components and includes the following elements:

- stochastic component: this is independent variable that is characterized by distribution related to the exponential family;
- systematic component that includes p independent variables creating so called "linear predictor":

$$\eta = X\cdot\beta\,;$$

- estimating features of the link function with taking into consideration their classification.

Besides the definition mentioned the model structure includes the following elements: model order (maximum order of equations creating the model); model dimension or number of equations creating the model; selection of a link function, variance function and elements of the linear predictor.

MODEL PARAMETER ESTIMATION

After estimating model structure the next problem is to compute model parameters using available statistical/experimental data. Usually, the model structure estimated provides information on a number of parameters to be estimated and the estimation method to be applied in concrete case. In some cases, it is reasonable to use the saving principle that supposes the following: number of parameters to be estimated should not exceed their necessary quantity. Here "necessity" can be viewed as necessity to preserve in the model constructed basic statistical characteristics of the process being studied: mean value, variance and covariance.

Very often parameters of a generalized linear model can be estimated using ordinary LS (OLS) method. In a case of normal data distribution this is equivalent to application of maximum likelihood (ML) method. Generally, widely used methods for estimation of GLM parameters are the following: OLS, weighted LS (WLS), ML, method of moments and Monte Carlo techniques (iterative and non-iterative). As far as the OLS in some cases results in biased parameter estimates due to its sensitivity to data outliers an alternative procedure for estimation is ML.

In the case of normal residuals logarithmic likelihood function, l, can be represented for n measurements as follows:

$$-2l = n\log(2\pi\sigma^2) + \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{\sigma^2}.$$

For a fixed, σ^2, maximization of l is equivalent to minimization of quadratic deviations from mean: $\sum [y(k) - \mu]^2$; thus, for linear model we have:

$$\eta_i = \mu_i = \sum_{j=1}^{p} x_{ij}\beta_j .$$

Very popular today method for GLM parameter estimation is Markov Chain Monte Carlo (MCMC) procedure that can be applied to linear and nonlinear models. The estimation algorithm based on this method is functioning as follows [11]:

- sampling X_u^{i+1} from the distribution $P(X_u \mid X_0, \theta^i)$, where X_0 is experimental data;
- sampling parameter estimate, θ^{i+1}, from distribution $P(\theta \mid X_0, X_u^i)$.

The statistical series generated this way under weak regularity conditions has marginal stationary distribution, $P(X_u, \theta \mid X_0)$, that can be trivially transformed into $P(\theta \mid X_0)$. The MCMC techniques require rather high computational resources but they are very popular due to their universality, good scaling characteristics, capability to take into account non-observable variables, low estimation errors, and possibilities for parallel computing. Correctness of the conditions necessary for parameter estimation can be analyzed after computing the parameters. After computing the parameters, we get estimates of the random process influencing the process under study as follows:

$$\hat{\varepsilon}(k) = e(k) = y(k) - \hat{y}(k),$$

and analyze its statistical characteristics indicating correctness of the parameter values.

MODEL DIAGNOSTICS: SELECTION OF THE BEST MODEL OF THE CANDIDATES CONSTRUCTED

At this stage model adequacy is analyzed that includes the steps given below.

a) Visual study of residual graph, $e(k) = y(k) - \hat{y}(k)$, where $\hat{y}(k)$ is dependent variable estimate computed with the model constructed. The graph should not exhibit outliers and long periods of time with large values of residuals (i.e., long periods of model inadequacy). In the case of constructing GLM such model characteristics can result from random selection of distribution law for dependent variable without substantial argumentation. It can result in substantial deviations of forecast estimates computed with the model from actual observations.

b) The estimated model residuals should not be auto-correlated. To analyze the possible autocorrelation level it is necessary to compute autocorrelation function (ACF) and partial ACF (PACF) for the series, $\{e(k)\}$, and analyze statistical significance of the functions with Q-statistics. Besides, the level of residuals autocorrelation can be additionally tested with Durbin–Watson statistic, $DW = 2 - 2\rho$, where, $\rho = E[e(k)e(k-1)]/\sigma_e^2$, is correlation between neighboring values of residuals; σ_e^2 is variance of the residual process; if $\rho = 0$, autocorrelation between residuals is absent, and ideal value for the DW statistic is 2.

c) Statistical significance of the model parameter estimates can be tested with Student $t-$ statistic, $t = [\hat{a} - a^0]/SE_{\hat{a}}$, where $\hat{a}$ is parameter estimate; a^0 is zero-hypothesis regarding the estimate; $SE_{\hat{a}}$ is standard error for the estimate. Usually all the computer systems that include functionality for time series analysis provide all necessary information for a user regarding analysis of statistical significance of parameter estimates.

d) Determination coefficient, $R^2 = \text{var}[\hat{y}]/\text{var}[y]$, where, $\text{var}(\hat{y})$ is variance of dependent variable estimated via the model constructed; $\text{var}(y)$ is actual sample variance of dependent variable. The ideal value of the coefficient is, $R^2 = 1$, when the values of variance in nominator and denominator are the same. This statistical parameter can be interpreted as a measure of information contained in a sample and in a model. From this point of view R^2 compares volume of information represented by a model to the volume of information represented by data sample used for model constructing.

e) The sum of squared errors should take minimum value for the best candidate model constructed:

$$\sum_{k=1}^{N} e^2(k) = \sum_{k=1}^{N} [\hat{y}(k) - y(k)]^2 \rightarrow \min_{\hat{\theta}}.$$

f) For estimating constructed model adequacy Akaike information criterion (AIC) is used:

$$AIC = N \ln\left(\sum_{k=1}^{N} e^2(k)\right) + 2n,$$

and Bayes–Schwarz criterion:

$$BSC = N \ln\left(\sum_{k=1}^{N} e^2(k)\right) + n\ln(N),$$

where $n = p + q + 1$, is a number of estimated model parameters (p is a number of parameters for auto-regression part; q is a number of parameters for the moving average part; 1 is added in a case when the bias, a_0, is added). Both criteria exhibit minimum values for the best model among the candidates estimated.

g) Besides the statistics mentioned, Fisher statistic $F \sim R^2/(1 - R^2)$ is applied that may show adequacy of a model as a whole after testing it for statistical significance.

To formulate statistical inference regarding the model constructed Bayesian factor, $BF(i, j)$, is also used that is represented by the ratio of posterior probabilities to prior [12; 13; 14]. The criteria of maximum marginal density of the distribution, $p(x \mid M_i)$, is used that corresponds to the following condition: $BF(i, j) > 1$.

Correct application of the methodology presented provides a possibility for constructing adequate mathematical model in the class of generalized linear models if the statistical/experimental data hired corresponds to the requirement of system representation and contain necessary information.

EXAMPLE 1. GLM MODEL CONSTRUCTING

Structure of statistical data hired for model constructing is shown in Table 2. The data shows loss due to car insurance for three selected regions of Ukraine.

Table 2. The structure of statistical data

N	Characteristic of data	Value of the characteristic
1	Power of sample	9546
2	Dependent variable	Loss due to insurance
3	Region where police was sold	Kyiv, Crimea, Odesa
4	Year of a car production	Starting from 2006
5	Car brend	Mitsubishi, Toyota, VAZ

For the data described in Table 2, suppose that dependent variable is normally distributed, and accept as a link log-function. The histogram for dependent variable “loss” is presented in Fig. 1.

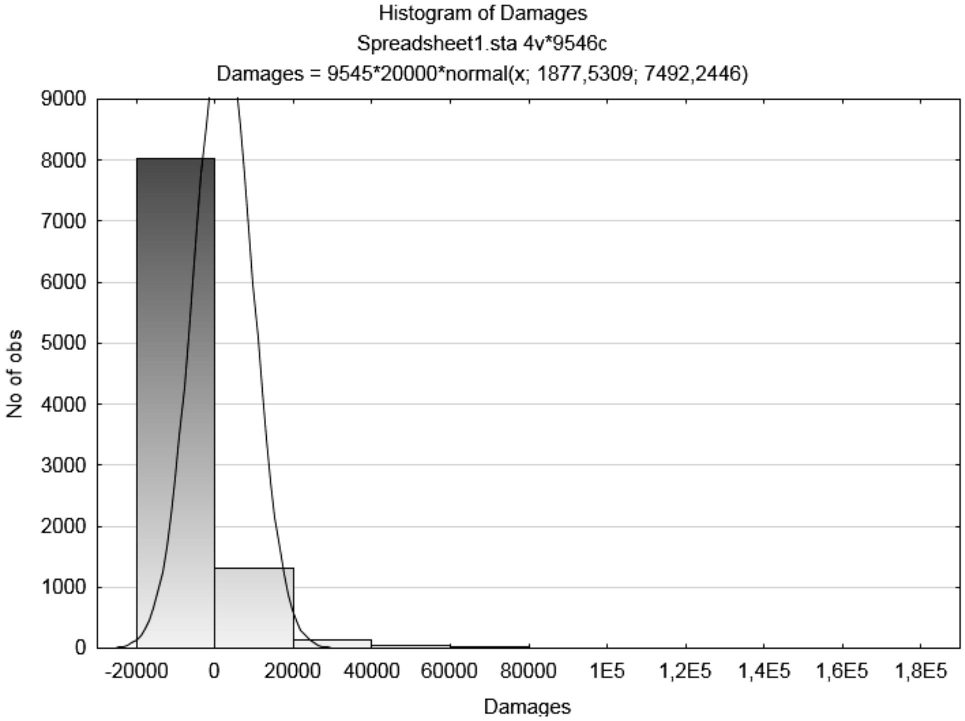

Fig. 1. The histogram for normally distributed dependent variable

The mean value of dependent variable “Loss” is 1877.531. To select the best model from the set of possible models constructed Akaike information criterion was used. Its minimum value corresponds to the best model. For log-normal model value of the criterion is 20.666. The drawback of the criterion is that its estimate asymptotically overestimates true value with non-zero probability. An alternative is Hannan–Quinn criterion that is based upon minimizing of corresponding sum instead of value itself.

Information Schwarz criterion usually selects the best model with a number of parameters that does not exceed number of parameters in the model selected by the Akaike criterion. The Schwarz criterion is asymptotically more reliable, and the Akaike criterion has a tendency to bias closer to selection of parametric models. However, for the example under consideration the values of Akaike, Hannan–Quinn, and Schwarz criteria are practically equal; that is why any of them is suitable for use. The standard deviation for dependent variable, “Loss”, is 7492.245.

The Likelihood Ratio Test is used for testing restrictions regarding parameters of statistical model, estimated with sample statistical data. If the value of *LR*-statistic exceeds critical value from χ-squared distribution at given significance level then the restrictions are neglected, and the model without restrictions is hired; otherwise, the model with restrictions is used.

The variance calculated shows how much the mean random value deviates from mathematical expectation; here it is important that this is quadratic value. At the same time the variance itself is not very convenient for practical analysis because it has quadratic dimensionality of a random value. That is why standard deviation is used further on as a measure of risk. From the point of view of financial analysis, the standard deviation is more important because the mean deviation of insurance results from expected return shows actual financial results and insurance risk. As a risk measure can also be used mean absolute deviation (mad). In practice the value of standard deviation is greater than mean absolute deviation but these values have the same order, and here the following relation holds: $mad = 0.7979 \cdot S$. The result of forecasting loss and risk is presented in Table 3. The relative error of forecasting with log-normal model amounted to 1.06%, what is high quality result for the model hired. The total forecasted loss amounts to 18111231.380, and actual loss was 17921032.581, what supports the fact that the variable “Loss” is normally distributed. The proposal to use logarithmic link function is acceptable for subsequent analysis of alternative models.

Table 3. Forecasting results with log-normal model

Value	Total loss	Mean	Std. deviance	Max	Min	Variance
Actual	17921032.58	18577.53	7492.24	151771.4	0.00	
Forecasted	18111231.38	18976.45	9379.910	14010.97	634.0	49.535

Thus, the log-normal model is acceptable but it is not the best for the dataset used. That is why the search for the better model was continued. Other modeling results are given in Table 4.

Table 4. Other modeling results with application of alternative data distributions and link functions

N	Model characteristics		Total forecasted loss	Actual total loss	Deviation of actual data from forecasts	Risk of loss
	Dependent variable distribution	Link function				
1	Gamma	LOG	102008320.905	17921032.581	84087288.32	1.003
2	Normal	LOG	18111231.380		190198.799	0.495
3	Poisson	LOG	17921032.574		0.007	0.547
4	Normal	identity	17921032.589		0.009	0.532

Results of model parameters estimation with the use of classic and Bayesian approach are presented in Table 5.

Table 5. Results of model parameters estimation with the use of classic and Bayesian approach

N	Classic approach			Bayesian approach			
	Mean	Std. deviance	Variance, %	Mean	Std. deviance	Variance, %	R-squared
1	11805.7	15358.1	130.091	11804.34	15247.23	128.669	0.897
2	1897.45	939.91	49.535	1897.294	939.94	49.4	0.998
3	1877.53	1027.56	54.73	1877.301	1027.552	55.679	0.998
4	1877.53	999.302	53.224	1876.909	999.751	53.809	1.0

Thus, after constructing models using various proposals regarding initial distributions of dependent variable and link function the following results of computing were achieved:

1) the best data approximating model turned out to be the model with initial Poisson distribution of dependent variable and logarithmic link function what is proved by the forecasted total loss of 17921032.574 with practically zero errors of forecasting;

2) the value at risk for the models constructed was in the range 40–60%, what requires extra measures for minimizing this value;

3) the model based on normal distribution of data showed the risk about 49–50 %, but there were observed substantial deviations of forecasts from actual data;

4) comparing the models based upon normal data distribution with logarithmic and identity link functions it was established that Akaike criterion accepts the same value of about 20.66; that is why the model selection should be based upon forecasts of total loss;

5) the model based upon gamma-distribution with logarithmic link function showed maximum deviation from actual data (84087288.324) and maximum error: about 100%;

6) it is clear from results in Table 4 that the quality of model estimation with Bayesian approach are close to classic estimation with the maximum likelihood method but with better estimates of variance and standard deviation.

Thus, it can be concluded that the model based upon Poisson distribution and exponential link function is better for practical use from the point of view of forecasting, risk estimation, and parameter estimation.

BAYESIAN NETWORK APPLICATION TO OPERATIONAL RISK ANALYSIS

The next approach to operational risk estimation that we consider in the study is based upon application of Bayesian networks (BN) that is constructed for a commercial enterprise. The bases for the model create expert estimates though statistical data can also be hired successfully. Comparing to regression approach, BN provide a possibility for taking into consideration not only risk dependence on risk factors, but also the dependences of between risk factors [15]. The probabilistic inference also can be computed using incomplete data.

Mathematically, BN is directed acyclic graph nodes of which correspond to risk factors and environment variables, and edges correspond to possible relations between nodes. Each BN is also described by the set of conditional distributions characterizing risk factors and environment variables.

Formally, BN is a triple $\mathbf{N} = <\mathbf{V}, \mathbf{G}, \mathbf{J}>$, where $\mathbf{V}$ stands for a set of network variables; $\mathbf{G}$ is directed acyclic graph nodes of which represent variables of the process being modeled; $\mathbf{J}$ represents joint probability distribution for the network variables, $\mathbf{V} = \{X_1, X_2, ..., X_n\}$ [15; 16]. Here also Markov condition should be fulfilled: each network variable does not depend on all other variables except for the parent nodes of the variable. To construct the BN model first mutual information between all nodes is computed to discover possible dependences, then optimal network structure is estimated using appropriate criterion, say minimum description length (MDL) that is analyzed and re-computed at each iteration of the structure learning algorithm.

For two independent discrete events D and S Bayes theorem can be formulated as follows:

$$p(D \mid S) = \frac{p(D)\, p(S \mid D)}{p(S)}.$$

Suppose that state variable D can accept two values: D_t is true probability that system under study accepted one of possible state; D_f its opposite state. These two probabilities sum to 1 independently on the value that accepts variable S:

$$p(D_t \mid S) + p(D_f \mid S) = 1.$$

Apply to this equality Bayes theorem:

$$\frac{p(D_t)\, p(S \mid D_t)}{p(S)} + \frac{p(D_f)\, p(S \mid D_f)}{p(S)} = 1,$$

or

$$p(S) = p(D_t)\, p(S \mid D_t) + p(D_f)\, p(S \mid D_f).$$

Thus, known estimate $p(S)$ can be eliminated from consideration. In this example variable D has two states only, however, it is clear that $p(S)$ can be eliminated with arbitrary number of states for D.

Procedure of constructing Bayesian network. To estimate the degree of dependency between two random variables x^i and x^j the following expression was proposed in [17]:

$$MI(x^i,x^j)=\sum_{x^i,x^j} p(x^i,x^j)\cdot\log\left(\frac{p(x^i,x^j)}{p(x^i)\cdot P(x^j)}\right).$$

This is mutual information that shows volume of information contained in the variable x^i about variable x^j. It accepts nonnegative values only, $MI(x^i,x^j)\geq 0$, and in the case of complete independence between the variables it accepts zero value $MI(x^i,x^j)=0$, because $p(x^i,x^j)=p(x^i)\cdot P(x^j)$ and

$$\log\left(\frac{p(x^i,x^j)}{p(x^i)\cdot P(x^j)}\right)=\log\left(\frac{p(x^i)\cdot P(x^j)}{p(x^i)\cdot P(x^j)}\right)=\log(1)=0.$$

In the case BN contains N nodes, estimation of $MI(x^i,x^j)$ for all possible pairs of x^i and x^j it is necessary to perform $\frac{N\cdot(N-1)}{2}$ computations under condition $MI(x^i,x^j)=MI(x^j,x^i)$.

The use of minimum description length (MDL) principle to estimate BN structure. According to the Shannon coding theory, for known distribution $P(X)$ for variable X the optimal length of code for transmitting specific value of random variable X tends to the value of $L(x)=-\log P(x)$ [18]. The source entropy $S(P)=-\sum_x P(x)\cdot\log P(x)$ is minimum expected length of the coded message. Any other code that is based upon incorrect representation of information source will result in longer expected length of message. In other words, better model of message source results in more compact code of data.

In the problem of BN learning as a data source is functioning some unknown distribution function $P(D\,|h_0)$, where $D=\{d_1,...,d_N\}$ is dataset; h is hypothesis regarding probabilistic origin of the data; $L(D|\,h)=-\log P(D|\,h)$ is empirical risk that is additive regarding the number of observations and proportional to empirical errors [19]. The difference between $P(D|\,h_0)$ and modeled distribution, $P(D|\,h)$, according to the Kullbak–Leibler measure is determined as follows:

$$\left|P(D|\,h)-P(D\,|h_0)\right|=\sum_D P(D|\,h_0)\cdot\log\frac{P(D|\,h_0)}{P(D|\,h)}=$$

$$=\sum_D P(D|\,h_0)\cdot\left|L(D|\,h)-L(D|\,h_0)\right|\geq 0.$$

In fact, this is the difference between hypothetically expected data code length and minimum possible length. This difference is always nonnegative and it equals to zero in the case of complete equality between distributions only. The MDL principle in its general formulation declares: from the set of possible candidate models it is necessary to select the one that describes data in short and completely (without loss of information) [19].

Thus, in general form the problem of forming MDL is formulated as follows: for the set of learning data $D=\{d_1,...,d_n\}$, $d_i=\{x_i^{(1)}x_i^{(2)}...x_i^{(N)}\}$ (here upper index is variable number); n is a number of observations; each observation includes N ($N \ge 2$) variables $X^{(1)}, X^{(2)}, ..., X^{(N)}$. Each j-th variable ($j = 1, ..., n$) has $A^{(j)} = \{0, 1, ..., \alpha^{(j)} - 1\}$ ($\alpha^{(j)} \ge 2$) states, and each structure, $g \in G$, of BN is represented by N sets of parents, ($\Pi^{(1)}, ..., \Pi^{(N)}$), i.e., for each node, $j = 1, ..., n$, $\Pi^{(j)}$, this is a set of parent variables such, that $\Pi^{(j)} \subseteq \{X^{(1)}, ..., X^{(N)}\} \setminus \{X^{(j)}\}$ (any node cannot be a parent for itself, what means that graph does contain cycles). Thus, MDL of the structure $g \in G$ for n observations, $x^n = d_1 d_2 ... d_n$, is computed as follows:

$$L(g, x^n) = H(g, x^n) + \frac{k(g)}{2} \cdot \log(n),$$

where $k(g)$ is a number of independent conditional probabilities in the net structure, g, and $H(g, x^n)$ is empirical entropy:

$$H(g, x^n) = \sum_{j \in J} H(j, g, x^n), \quad k(g) = \sum_{j \in J} k(j, g),$$

where MDL for j-th node is computed as follows:

$$L(j, g, x^n) = H(j, g, x^n) + \frac{k(j, g)}{2} \cdot \log(n);$$

$k(j, g)$ is a number of conditional probabilities for j-th node:

$$k(j, g) = (\alpha^{(j)} - 1) \cdot \prod_{k \in \phi(j)} \alpha^k,$$

where $\phi(j) \subseteq \{1, .., j-1, j+1, ..., N\}$ is such a set that $\Pi^{(j)} = \{X^{(k)} : k \in \phi^{(j)}\}$.

The empirical entropy of j-th node is computed as follows:

$$H(j, g, x^n) = \sum_{s \in S(j,g)} \sum_{q \in A^{(j)}} -n[q, s, j, g] \cdot \log \frac{n[q, s, j, g]}{n[s, j, g]},$$

$$n(s, j, g) = \sum_{i=1}^{n} I(\pi_i^{(j)} = s); \quad n[q, s, j, g] = \sum_{i=1}^{n} I(x_i = q, \pi_i^{(j)} = s),$$

where $\pi^{(j)} = \Pi^{(j)}$ means that $X^{(k)} = x^{(k)}, \forall k \in \phi^{(j)}$; the function $I(E) = 1$, if the predicate, $E = true$, otherwise $I(E) = 0$. The simplified BN structure learning algorithm is shown in Fig. 2; it is performing in a cycle analysis of all possible acyclic network structures. The structure g^* contains optimal network structure. The optimal structure accepts minimum value for the function $L(g, x^n)$.

BN learning algorithm using MDL principle

1. $g^* \leftarrow g_0 (\in G)$;
2. for $\forall g \in G - \{g_0\}$: if $L(g, x^n) < L(g^*, x^n)$ then $g^* \leftarrow g$;
3. solution is contained in g^*.

EXAMPLE 2. BN MODEL CONSTRUCTION FOR OPERATIONAL RISK ESTIMATION AT COMMERCIAL ENTERPRISE

Consider commercial enterprise that provides for the services in entertainment sphere, and has obligations to other commercial enterprises for established possible services. These enterprises are banks, owners of cloud technologies and servers, and outsourcing companies that perform some separate functions for the entertainment enterprise. The obligations suppose transfer of money to other enterprises. That is why the main enterprise income should cover all obligations. Suppose that the main enterprise income is equal to one million euros.

Certainly, during the period of the main enterprise functioning various risk may arise that influence negatively its income. A substantial part of the risks create operational risks. To have a possibility for covering the risks the enterprise should direct a part of its capital (income) on covering operational risks, this is so called Capital at Risk (CaR). The use of statistical data in such case is practically impossible due to impossibility of separating operational loss from other risks. That is why the basic source of information for modeling operational risks provide the expert estimates.

According to recommendations provided for researchers by Basel there exist four basic factors influencing generation of operational risks:

- Risk related to working personnel: this type of risk is adhered to errors and incorrect actions of personnel, their insufficient qualification, sometimes overload at work place, possible irrational organization of working activities etc.
- Risk related to systems and work technologies. This type of risk is related to insufficiently high information technologies used at enterprise, the hardware system characteristics can be inadequate to performing operations, data processing techniques can be inadequate or data itself may exhibit unaccepted characteristics.
- Processes risk: this is risk of loss due to errors in the processes of performing various financial operations and corresponding calculations, forming reports, price forming operations etc.
- Risk induced by external influences: this type of risk is generated by events from external environment; for example, by legislative changes, general policy of state, economy, as well as the risks of external physical influences into activities of financial organization.

Each of the four named factors is also influenced by some other reasons of coming to being operational risk that can be analyzed by the experts of specific commercial enterprise. Table 6 illustrates the nodes (variables) selected for estimating the structure of Bayesian network. The nodes and respective values were selected using expert estimates.

The central point regarding the Bayesian network is touching upon variable Loss that is described by the quantitative values in euros as follows: $\{0, 1000, 10000, 50000, 100000\}$. The specific numbers are used to define the variable because of necessity to perform linear interpolation. The numbers were selected by an expert taking into consideration the loss at enterprise due to operational risks that took place earlier. Using the analysis being performed the enterprise managers can estimate the loss say within a week/month and accumulate necessary capital to cover the loss.

Table 6. Nodes of Bayesian network and their possible values

N	Variable name	Name of node in network	Possible values
1	Possible loss	Loss	0, 1000, 10000, 50000, 10000 ($)
2	Risk of external environment	Risk of External Environment	Low, Medium, High
3	Risk of working personnel	Risk of Staff	Low, Medium, High
4	Risk of hardware system used	Risk of System	Low, Medium, High
5	Risk of information processes	Risk of Process	Low, Medium, High
6	Staff qualification level	Qualified staff	Low, Medium, High
7	Compliance of	Compliance to staff	Low, Medium, High
8	Possible changes in legislation	Changes in legislation	Low, Medium, High
9	Reliability of payment systems used and partner banks	Reliability of payment systems and partner banks	Low, Medium, High
10	Possible hacker attacks and viruses	Hacker attacks and viruses	Low, Medium, High
11	Loss of information	Loss of information	0, 50, 100 (%)
12	Using workflow automation	Using workflow automation	Low, Medium, High
13	Possible crashes of servers	Server crashes	Low, Medium, High
14	Network failures	Network failures	Low, Medium, High
15	Using of firewall in the system	Using of Firewall	Yes, No
16	Using of UPS to support system	Using of UPS	Yes, No

The Loss variable is influenced by the four basic factors mentioned above: *Risk of Staff, Risk of System, Risk of Processes* and *Risk of External Environment.* These factors have the following levels: *Low*, *Medium*, and *High*, that can be interpreted as probability of realizing specific type of risk.

All other nodes and connections between them were identified on the basis of expert data, and reflect real reasons that influence specific type of risk:

- *Qualified staff* — describes the level of competence and qualification of personnel working at the enterprise. This node influences Risk of Staff, and Loss of information.

- *Compliance staff* — is related to correspondence of actually working staff to the necessary staff for functioning of an enterprise. This node has substantial influence on the variable Risk of Staff.

- *Changes in legislation* — describes possible level of legislative changes in a country. This is substantial reason for possible material loss because the possibilities for entertainment at any moment can be restricted or forbidden at all. This node influences the Risk of External Environment.

- *Reliability of payment systems and partner banks* — this factor can be a substantial reason for possible loss due to low quality of financial transactions (prolonged transaction time and possible errors). The factor also influences the Risk of External Environment.

- *Hacker attacks and viruses* — describes the level of hacker and viruses attacks to the enterprise servers. This influence can result in server reloading and system rejects. The node influences the Risk of External Environment.

- *Loss of information* — the loss is usually described in percentage form and reflects the portion of information that could be lost in the process of functioning of an enterprise. The node influences the Risk of Processes.

- *Using workflow automation* — this factor describes the level of usage specialized software that provides the possibility for performing workflow automation. This approach influences the Risk of Processes.
- *Network failures* — describes the level of computer network reliability that usually does not depend on enterprise (under study) itself. The level can be changed by changing network provider. This factor influences the Risk of External Environment.
- *Using of Firewall* — this factor has two states: Yes and No, and is described by probabilities that are related to the time intervals when firewall is used or not. It influences on Hacker attacks and viruses and Server crashes.
- *Using of UPS* — this factor is also characterized by two states and it is also described by the probabilities related to time intervals when uninterrupted power source was used or not. In its turn the factor influences on possible Server crashes.
- *Server Crashes* — the factor describes possible level of server crashes of an enterprise and depends on the Use of firewall and Uninterrupted power supplies. The reason is that loss of power or virus/hacker/ddos attack may result in server turnoff. The factor influences on the Risk of External Environment.

The next step of estimating Bayesian network structure is filling in probability tables for nodes. All probabilities in this example were estimated using expert estimates. On the other side, availability of necessary statistical information at the enterprise (i.e., storing the information during its functioning) will give the possibility for the fast replacing expert estimates by actual data. This way work adaptive risk analyzing systems.

For the nodes that do not have parents it will be unconditional probabilities for each possible state. The nodes that have parents require estimating conditional probabilities. The problem is that number of required conditional probabilities is growing fast with growing number of parents and their possible states. In case of discrete variables the number of values in conditional probability tables (CPT) can be estimated via the expression:

$$Count = NodeStates \cdot \prod_{i}^{Parents} ParentStates_i\ ,$$

where *Count* is the number of values necessary for CPTs; *NodeStates* is number of states for a specific node under consideration; *Parents* is the number of parents for a specific node being analyzed; $ParentStates_i$ represents the number of states for $-i$-th parent. Using this formula it was estimated that for the node Loss it were required about 400 values. Example of CPT for the node of “Loss of information” is given below in Table 7.

Table 7. Conditional probabilities for the node “Loss of Information”

Using workflow automation	**Low**	**Low**	**Low**	**Medium**	**Medium**	**Medium**	**High**	**High**	**High**
Loss of information	0%	50%	100%	0%	50%	100%	0%	50%	100%
Low	0.25	0.1	0.03	0.55	0.42	0.31	0.95	0.8	0.6
Medium	0.3	0.22	0.1	0.25	0.3	0.38	0.04	0.12	0.23
High	0.45	0.68	0.87	0.2	0.28	0.31	0.01	0.08	0.17

As a result of the model parameter estimation the BN was constructed for operational risk analysis for commercial enterprise given in Fig. 2.

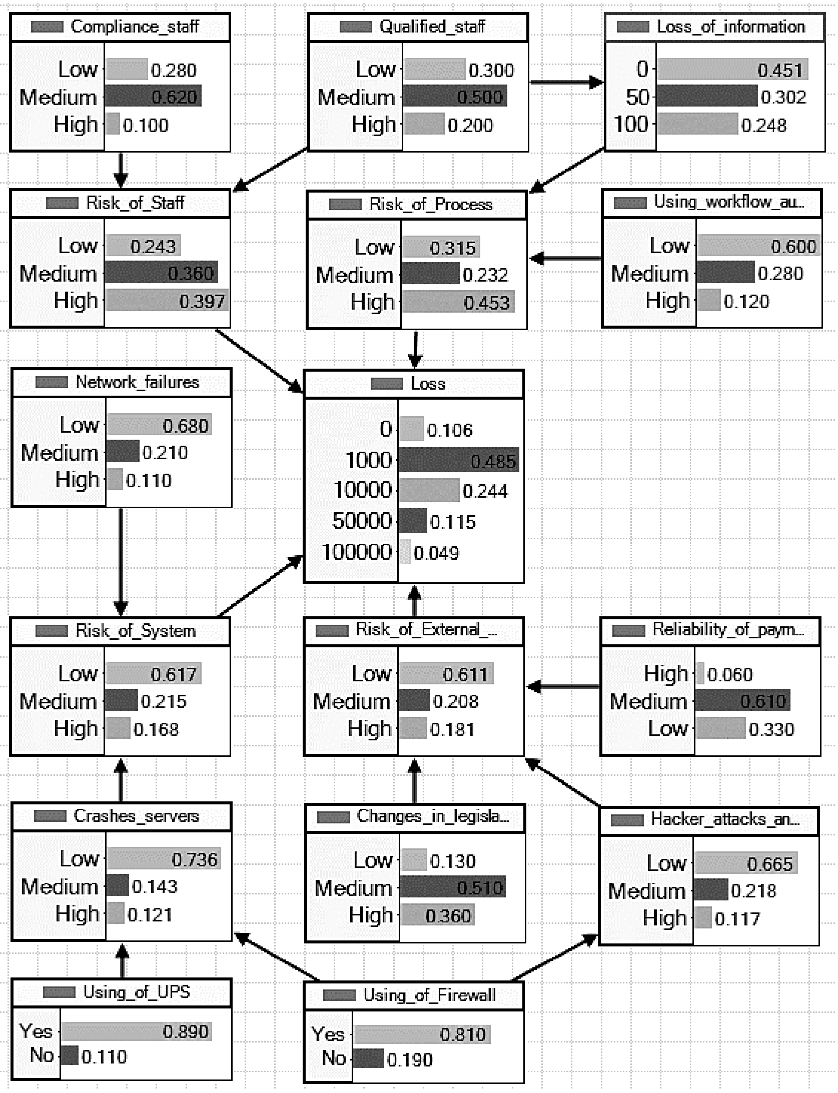

Fig. 2. Possible probabilistic inference estimation using the BN constructed

ANALYSIS OF THE RESULTS

Table 8 contains the capital necessary for covering the operational risks: the 90, 95 i 99 quintiles reflect the probabilities 0.9, 0.95 and 0.99, correspondingly.

Table 8. Capital required for covering the risks at selected level of confidence

Quintile, %	90	95	99
CaR, $	38476	56833	97554

It can be easily tracked that Capital-at-Risk (CaR) is substantially growing with growing confidence level from 90 to 99. According to the proposal by the Basel Basic Indicator Approach (BIA) acceptable level of risk for banking sphere is 15%. Thus, the enterprise with monthly income of about one million euros should keep CaR of 150000 euros. This is about three times as much than the result achieved in the example above. This result stresses that enterprises in banking sphere should use the Advanced Measurement Approaches (AMA) for analysis of operational risks. The volume of CaR can be reduced substantially this way and available extra capital could be used more rationally to achieve higher income.

CONCLUSIONS

The reasons for emerging financial operational risks in financial organizations were presented. It was shown that an urgent task for the organizations is development and practical use of the risk analysis and management systems on the basis of modern system analysis approach, providing the possibility for constructing mathematical models, more specifically probabilistic models of Bayesian type that provide the possibility for identification and taking into consideration uncertainties related to the random nature of multiple events in financial analysis.

Improved systemic multistep methodology of constructing models for financial processes and financial risks of arbitrary origin was proposed. An example of the methodology application is given that shows effectiveness of the methodology in application to operational risk analysis. Application of the methodology proposed for modeling financial processes with the use of generalized linear models and Bayesian parameter estimation guaranties high quality of risk estimation with minimum errors.

Also simplified methodology was proposed for constructing models of risk in the form of Bayesian networks using the notion of mutual information between variables of network, and criteria of estimation the quality of model structure on the basis of minimum description length. An example of BN model constructing was given for a commercial enterprise on the basis of expert estimates. The model was used for estimating distribution of loss in the case of operational risk availability. Analysis of estimation results provided by Bayesian network proves that models of the type selected can be successfully applied for estimating possible loss in financial organizations.

In future studies it is reasonable to create specialized commercial intellectual decision support system for performing mathematical modeling and financial risks estimation. The system should provide a possibility for the use of statistical data, expert estimates and generated (simulated) variables of continuous type convenient for the use in Bayesian framework. Also possible data uncertainties should be identified and taken into consideration to improve the results of all the stages of expert estimation and statistical data processing. All the stages of data and expert estimates processing should be controlled by appropriate sets of statistical quality analysis criteria to reach high quality of intermediate and final results of computing.

REFERENCES

1. *International Convergence of Capital Measurement and Capital Standards. A Revised Framework. Comprehensive Version*. Basel: Basel Committee on Banking Supervision, Bank for International Settlements, 2006, 158 p.

2. M.Z. Zgurovsky, N.D. Pankratova, *System Analysis: Theory and Applications*. Berlin: Springer-Verlag, 2007, 446 p.
3. M. Denuit, J. Dhaene, M. Goovaerts, and R. Kaas, *Actuarial Theory for Dependent Risks: Measures, Orders and Models*. Chichester (West Sussex): John Wiley & Sons, Inc., 2005, 440 p.
4. A. Cruz, *Modeling, Measuring and Hedging Operational Risk*. London: Wiley, 2002, 346 p.
5. *Operational Risk Regulation, Analysis and Management*; edited by Carol Alexander. New York: Pearson Education Limited, 2003, 369 p.
6. Pavel Shevchenko, *Modelling Operational Risk Using Bayesian Inference*. New York: Springer, 2011, 302 p.
7. S.O. Dmitrov, *Modeling operational risk of commercial bank*. Sumy: The State Higher Educational Institution “UABS NBU”, 2010, 264 p.
8. A. Frachot, P. Georges, and T. Roncalli, *Loss Distribution Approach for operational risk*. Available: http://thierry-roncalli.com/download/lda.pdf
9. P.I. Bidyuk, S.V. Trukhan, “Development and Usage of Information System for Analysis and Forecasting Financial Processes,” *System Sciences and Cybernetics*, no. 1, pp. 26–37, 2013.
10. P.I. Bidyuk, M.S. Rubets, “Information System for Modeling and Estimation of Operational Risks Using Artificial Intelligence Methods,” *System Sciences and Cybernetics*, no. 1, pp. 5–29, 2015.
11. W.R. Gilks, S. Richardson, and D.J. Spiegelhalter, *Markov Chain Monte Carlo in Practice*. Boca Raton (Florida): Chapman & Hall/CRC, 1998, 486 p.
12. C. Alexander, *Bayesian Methods for Measuring Operational Risk*. Available: http://www.icmacentre.ac.uk/pdf/bayesian.pdf
13. X. Hao, *Operational Risk Control of Commercial Banks Based on Bayesian Network*. Atlantis Press, 2013, pp. 913–918.
14. Y.K. Yoon, *Modelling operational risk in financial institutions using bayesian networks*. London: University of London, 2003, 83 p.
15. M. Neil, N.E. Fenton, and M. Tailor, “Using bayesian networks to model expected and unexpected operational losses,” *Risk Analysis*, pp. 34–57, 2005.
16. Alan Stuart, Keith Ord, *Kendall's Advanced Theory of Statistics: Volume 1, Distribution Theory*. Wiley, 1994, 700 p.
17. C.K. Chow, C.N. Liu, “Approximating discrete probability distributions with dependence trees,” *IEEE Transactions on information theory*, vol. 14, issue 3, 6 p., May 1968.
18. T. Minka et al., “Infer.NET 2.6.,” *Microsoft Research*, Cambridge, 2014. Available: http://research.microsoft.com/internet
19. M.Z. Zgurovsky, P.I. Bidyuk, O.M. Terentiev, and T.I. Prosyankina-Zharova, *Bayesian Networks in Decision Support Systems*. Kyiv: Publishing Company “Edelweiss” LLC, 2015, 300 p.

Received 15.10.2023

INFORMATION ON THE ARTICLE

Petro I. Bidyuk, ORCID: 0000-0002-7421-3565, Educational and Research Institute for Applied System Analysis of the National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”, Ukraine, e-mail: pbidyuke_00@ukr.net

Oxana L. Tymoshchuk, ORCID: 0000-0003-1863-3095, Educational and Research Institute for Applied System Analysis of the National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”, Ukraine, e-mail: oxana.tim@gmail.com

Liudmyla B. Levenchuk, ORCID: 0000-0002-8600-0890, Educational and Research Institute for Applied System Analysis of the National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Ukraine, e-mail: lusi.levenchuk@gmail.com

ОЦІНЮВАННЯ ОПЕРАЦІЙНОГО РИЗИКУ З ВИКОРИСТАННЯМ МЕТОДОЛОГІЇ СИСТЕМНОГО АНАЛІЗУ / П.І. Бідюк, О.Л. Тимощук, Л.Б. Левенчук

Анотація. Фінансові ризики — популярна тема досліджень практичної необхідності використання їх математичних моделей, оцінювання можливих втрат у багатьох напрямах людської діяльності, прогнозуванні і відповідних управлінських рішеннях у фінансовій та інших сферах, де капітал, облігації, біржові акції та інші активи успішно використовуються. Фінансові процеси характеризуються складною формою розвитку у часі, що потребує застосування відповідних методів, прийомів та процедур моделювання, оцінювання ризиків, прогнозування і підтримання/прийняття рішень. Для вирішення цих проблем застосовано підхід на основі системного аналізу як унікальну та універсальну методологію дослідження. Фінансові ризики, зокрема операційні, які розглядаються у роботі, класифікуються як нелінійні нестаціонарні процеси, що потребують застосування належних методів для аналізу і досить складного аналітичного опису для оцінювання і прогнозування можливих втрат. Результати аналізу операційних ризиків отримано у формі системної методології, нових моделей, побудованих з використанням статистичних даних, регресійного аналізу та байєсівських методів, а також оцінок втрат, отриманих за допомогою створених моделей. Побудовано моделі і запропоновано системний підхід до аналізу фінансових процесів, придатний для практичного використання за умови, що користувачі матимуть належні статистичні дані та експертні оцінки.

Ключові слова: фінансовий операційний ризик, математична модель, статистичні дані, методологія системного аналізу, оцінювання втрат, система підтримання прийняття рішень.